

**cloudera<sup>®</sup>**

**CDS Powered by Apache  
Spark**

## **Important Notice**

© 2010-2019 Cloudera, Inc. All rights reserved.

Cloudera, the Cloudera logo, and any other product or service names or slogans contained in this document are trademarks of Cloudera and its suppliers or licensors, and may not be copied, imitated or used, in whole or in part, without the prior written permission of Cloudera or the applicable trademark holder. If this documentation includes code, including but not limited to, code examples, Cloudera makes this available to you under the terms of the Apache License, Version 2.0, including any required notices. A copy of the Apache License Version 2.0, including any notices, is included herein. A copy of the Apache License Version 2.0 can also be found here: <https://opensource.org/licenses/Apache-2.0>

Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation. All other trademarks, registered trademarks, product names and company names or logos mentioned in this document are the property of their respective owners. Reference to any products, services, processes or other information, by trade name, trademark, manufacturer, supplier or otherwise does not constitute or imply endorsement, sponsorship or recommendation thereof by us.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Cloudera.

Cloudera may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Cloudera, the furnishing of this document does not give you any license to these patents, trademarks copyrights, or other intellectual property. For information about patents covering Cloudera products, see <http://tiny.cloudera.com/patents>.

The information in this document is subject to change without notice. Cloudera shall not be liable for any damages resulting from technical errors or omissions which may be present in this document, or from use of this document.

### **Cloudera, Inc.**

**395 Page Mill Road  
Palo Alto, CA 94306  
info@cloudera.com  
US: 1-888-789-1488  
Intl: 1-650-362-0488  
www.cloudera.com**

### **Release Information**

Version: CDS 2.4.x Powered By Apache Spark  
Date: May 6, 2019

# Table of Contents

<b>CDS Powered by Apache Spark Overview.....</b>	<b>6</b>
--	----------

<b>CDS Powered by Apache Spark Release Notes.....</b>	<b>7</b>
---	----------

CDS Powered by Apache Spark Requirements.....	7
<i>CDH Versions.....</i>	7
<i>Cloudera Manager Versions.....</i>	8
<i>Scala 2.11 Requirement.....</i>	9
<i>Python Requirement.....</i>	9
<i>JDK 8 Requirement.....</i>	9
CDS Powered by Apache Spark New Features and Changes.....	9
<i>What's New in CDS 2.4 Release 2.....</i>	9
<i>What's New in CDS 2.4 Release 1.....</i>	9
<i>What's New in CDS 2.3 Release 4.....</i>	9
<i>What's New in CDS 2.3 Release 3.....</i>	9
<i>What's New in CDS 2.3 Release 2.....</i>	10
<i>What's New in CDS 2.3 Release 1.....</i>	10
<i>What's New in CDS 2.2 Release 4.....</i>	10
<i>What's New in CDS 2.2 Release 3.....</i>	10
<i>What's New in CDS 2.2 Release 2.....</i>	10
<i>What's New in CDS 2.2 Release 1.....</i>	10
<i>What's New in CDS 2.1 Release 4.....</i>	10
<i>What's New in CDS 2.1 Release 3.....</i>	10
<i>What's New in CDS 2.1 Release 2.....</i>	10
<i>What's New in CDS 2.1 Release 1.....</i>	10
<i>What's New in CDS 2.0 Release 2.....</i>	11
<i>What's New in CDS 2.0 Release 1.....</i>	11
CDS Powered by Apache Spark Known Issues.....	11
<i>Structured Streaming exactly-once fault tolerance constraints.....</i>	11
<i>DecimalType push-down to Parquet data sources has been disabled.....</i>	11
<i>JOB_SUMMARY_LEVEL Parquet flag is not supported.....</i>	11
<i>LZ4, BROTLI, and ZSTD codecs are not supported.....</i>	11
<i>SparkSQL StringStartsWith filter is not supported.....</i>	12
<i>Apache Spark XSS vulnerability in UI CVE-2018-8024.....</i>	12
<i>CDS 2.3 Release 3 requires additional configuration to run PySpark on Cloudera Data Science Workbench versions 1.3.x (and lower).....</i>	12
<i>UnsatisfiedLinkError observed when using Snappy compression in the spark2-shell.....</i>	12
<i>Spark jobs fail when lineage collection is enabled.....</i>	13

<i>Spark on Kubernetes is not supported</i> .....	13
<i>ORC file format is not supported</i> .....	13
<i>Accessing multiple clusters simultaneously is not supported</i> .....	13
<i>Parquet logical type <code>TIMESTAMP_MICROS</code> unavailable</i> .....	13
<i>Spark SQL does not respect Sentry ACLs when communicating with Hive metastore</i> .....	13
<i>Empty result when reading Parquet table created by <code>saveAsTable()</code></i> .....	14
<i>Spark 2 Version requirement for clusters managed by Cloudera Manager</i> .....	14
<i>Spark Standalone is not supported</i> .....	14
<i>HiveOnSpark is not supported with CDS</i> .....	14
<i>SparkOnHBase is not supported with CDS</i> .....	14
<i>Using the JDBC Datasource API to access Hive or Impala is not supported</i> .....	15
<i>Dynamic allocation and Spark Streaming</i> .....	15
<i>Oozie Spark2 action is not supported</i> .....	15
<i>SparkR is not supported</i> .....	15
<i>GraphX is not supported</i> .....	15
<i>Thrift server is not supported</i> .....	15
<i>Spark SQL CLI is not supported</i> .....	15
<i>Rolling upgrades are not supported</i> .....	15
<i>Package-based installation is not supported</i> .....	15
<i>Spark-Avro Library is not supported; Use Built-In Avro Data Source</i> .....	15
<i>Hardware acceleration for MLLib is not supported</i> .....	15
<i>Cost based optimization is not supported</i> .....	15
<i>Running <code>spark2-submit</code> with <code>--principal</code> and <code>--keytab</code> arguments does not work in client mode</i> .....	15
<i>Long-running apps on a secure cluster might fail if driver is restarted</i> .....	16
<i>CDS Powered by Apache Spark Incompatible Changes</i> .....	16
<i>CDS Powered by Apache Spark Fixed Issues</i> .....	16
<i>Issues Fixed in CDS 2.4 Release 2</i> .....	17
<i>Issues Fixed in CDS 2.4 Release 1</i> .....	17
<i>Issues Fixed in CDS 2.3 Release 4</i> .....	17
<i>Issues Fixed in CDS 2.3 Release 3</i> .....	17
<i>Issues Fixed in CDS 2.3 Release 2</i> .....	18
<i>Issues Fixed in CDS 2.3 Release 1</i> .....	19
<i>Issues Fixed in CDS 2.2 Release 4</i> .....	19
<i>Issues Fixed in CDS 2.2 Release 3</i> .....	19
<i>Issues Fixed in CDS 2.2 Release 2</i> .....	23
<i>Issues Fixed in CDS 2.2 Release 1</i> .....	24
<i>Issues Fixed in CDS 2.1 Release 4</i> .....	42
<i>Issues Fixed in CDS 2.1 Release 3</i> .....	42
<i>Issues Fixed in CDS 2.1 Release 2</i> .....	44
<i>Issues Fixed in CDS 2.1 Release 1</i> .....	48
<i>Issues Fixed in CDS 2.0 Release 2</i> .....	49
<i>Issues Fixed in CDS 2.0 Release 1</i> .....	49
<i>CDS Powered by Apache Spark Version, Packaging, and Download Information</i> .....	59
<i>CDS Versions Available for Download</i> .....	59

<i>CDS Maven Artifacts</i> .....	60
Using the CDS Powered by Apache Spark Maven Repository.....	60
<i>CDS 2.4 Powered by Apache Spark Maven Artifacts</i> .....	61
<i>CDS 2.3 Powered by Apache Spark Maven Artifacts</i> .....	64
<i>CDS 2.2 Powered by Apache Spark Maven Artifacts</i> .....	68
<i>CDS 2.1 Powered by Apache Spark Maven Artifacts</i> .....	74
<i>CDS 2.0 Powered by Apache Spark Maven Artifacts</i> .....	80
<b>Installing or Upgrading CDS Powered by Apache Spark</b> .....	<b>83</b>
Install CDS Powered by Apache Spark.....	83
Upgrading to CDS 2.4 Powered By Apache Spark.....	84
<b>Administering CDS Powered by Apache Spark</b> .....	<b>86</b>
Configuring Spark 2 Tools as the Default.....	86
<b>Security Considerations for CDS Powered by Apache Spark</b> .....	<b>87</b>
Configuration Settings for Encryption.....	87
<b>Running Applications with CDS Powered by Apache Spark</b> .....	<b>88</b>
The Spark 2 Job Commands.....	88
Canary Test for pyspark2 Command.....	88
Fetching Spark 2 Maven Dependencies.....	88
Adapting the Spark WordCount App for Spark 2.....	89
Accessing the Spark 2 History Server.....	89
<b>Integrating CDS Powered by Apache Spark with Apache Kafka</b> .....	<b>90</b>
Requirements.....	90
Running Spark Jobs that Integrate with Kafka.....	90
Building Applications.....	91
Reading from Authorized Kafka.....	91
<b>Troubleshooting CDS Powered by Apache Spark</b> .....	<b>92</b>
<b>Frequently Asked Questions about CDS Powered by Apache Spark</b> .....	<b>93</b>
<b>Appendix: Apache License, Version 2.0</b> .....	<b>94</b>

## CDS Powered by Apache Spark Overview



**Note:**

This documentation refers to CDS 2.4 Powered by Apache Spark. This component is generally available and is supported on CDH 5.9 and higher.

A Hive compatibility issue in CDS 2.0 Release 1 affects CDH 5.10.1 and higher, CDH 5.9.2 and higher, CDH 5.8.5 and higher, and CDH 5.7.6 and higher. If you are using one of these CDH versions, you must upgrade to the CDS 2.0 Release 2 or higher parcel, to avoid Spark 2 job failures when using Hive functionality.

Apache Spark is a general framework for distributed computing that offers high performance for both batch and interactive processing. It exposes APIs for Java, Python, and Scala.

For detailed API information, see the [Apache Spark project site](#).



**Note:** Although this document makes some references to the external Spark site, not all the features, components, recommendations, and so on are applicable to Spark when used on CDH. Always cross-check the Cloudera documentation before building a reliance on some aspect of Spark that might not be supported or recommended by Cloudera. In particular, see [CDS Powered by Apache Spark Known Issues](#) for components and features to avoid.

CDS Powered by Apache Spark is an [add-on service](#) for CDH, distributed as a parcel and custom service descriptor, consisting of Apache Spark 2 core and several related projects:

- **Spark SQL:** Module for working with structured data. Allows you to seamlessly mix SQL queries with Spark programs.
- **Spark Streaming:** API that allows you to build scalable fault-tolerant streaming applications.
- **MLlib:** API that implements common machine learning algorithms.

Cloudera products include these versions of Apache Spark: 1.6, 2.0, 2.1, 2.2, 2.3, and 2.4.

Spark 1.6 is included as part of CDH 5 in Cloudera Enterprise 5.7.x and higher. The latest documentation is available at [Cloudera Enterprise documentation](#).

This document describes the separately released CDS 2.4 Powered by Apache Spark. It is shipped separately for ease of use and convenience of consumption. It enables customers to install and upgrade the features of Apache Spark 2 without going through a full upgrade of the CDH cluster.

On CDH 5, a Spark 1.6 service can coexist with a Spark 2 service. The configurations of the two services do not conflict and both services use the same YARN service. The port of the Spark History Server is 18088 for Spark 1.6 and 18089 for Spark 2.

### Unsupported Features

Consult [CDS Powered by Apache Spark Known Issues](#) on page 11 for a comprehensive list of features that are not supported with CDS Powered by Apache Spark.

### Related Information

- [Cloudera Community Spark forum](#)
- [Apache Spark documentation](#)

# CDS Powered by Apache Spark Release Notes

The release notes provide information on requirements, new and changed features, known issues, and fixed issues, and version and packaging information for CDS Powered by Apache Spark.

## CDS Powered by Apache Spark Requirements

The following sections describe software requirements for CDS Powered by Apache Spark.

### CDH Versions



**Important:** CDS Powered by Apache Spark is available in parcel format only, and not packages. Because Cloudera Manager does not support using parcels and packages in the same cluster, you cannot use CDS if you are using a package-based installation of CDH.

The CDS parcel version displayed in Cloudera Manager, which is also part of the [parcel file name](#), is structured as follows:

**<CDS\_version>-1.<cdh\_build\_version>.p<patch\_version>.<build\_number>**

The *<cdh\_build\_version>* portion is the version of CDH upon which the release was built. It is *not* the minimum supported CDH version. For example, although CDS 2.1 Release 3 was built on CDH 5.13.3, it is still supported on CDH 5.7 and higher CDH 5 versions.

Supported versions of CDH are described below.

A Hive compatibility issue in CDS 2.0 Release 1 affects CDH 5.10.1 and higher, CDH 5.9.2 and higher, CDH 5.8.5 and higher, and CDH 5.7.6 and higher. If you are using one of these CDH versions, you must upgrade to the CDS 2.0 Release 2 or higher parcel, to avoid Spark 2 job failures when using Hive functionality.


CDS Powered by Apache Spark Version	Supported CDH Versions
2.4 Release 2	CDH 5.10 and any higher CDH 5.x versions
2.4 Release 1	CDH 5.10 and any higher CDH 5.x versions
2.3 Release 4	CDH 5.9 and any higher CDH 5.x versions
2.3 Release 3	
2.3 Release 2	
2.3 Release 1	Not released due to late bug discovered. If downloaded do not use.
2.2 Release 4	CDH 5.8 and any higher CDH 5.x versions
2.2 Release 3	
2.2 Release 2	
2.2 Release 1	CDH 5.8 - CDH 5.13
2.1 Release 4	CDH 5.7 and any higher CDH 5.x versions
2.1 Release 3	CDH 5.7 and any higher CDH 5.x versions
2.1 Release 2	CDH 5.7 and any higher CDH 5.x versions
2.1 Release 1	CDH 5.7 - CDH 5.12

CDS Powered by Apache Spark Version	Supported CDH Versions
2.0 Release 2	CDH 5.7 - 5.11
2.0 Release 1	CDH 5.7 up to 5.7.5, CDH 5.8 up to 5.8.4, CDH 5.9 up to 5.9.1, CDH 5.10.0.  CDS 2.0 Release 2 is required for any higher maintenance releases of these CDH versions.

A Spark 1.6 service (included in CDH 5.7 and higher) can co-exist on the same cluster as Spark 2 (installed as a separate parcel). The two services are configured to not conflict, and both run on the same YARN service. Spark 2 uses the [external shuffle service](#) from the CDH installation if Spark 1 is already installed, or installs the shuffle service itself if necessary. Only the external shuffle service classes from the CDH installation can be used.

Although Spark 1 and Spark 2 can coexist in the same CDH cluster, you cannot use multiple Spark 2 versions simultaneously in the same Cloudera Manager instance. All CDH clusters managed by the same Cloudera Manager Server must use exactly the same version of CDS Powered by Apache Spark. For example, you cannot use the built-in CDH Spark service, a CDS 2.1 service, and a CDS 2.2 service. You must choose only one CDS 2 Powered by Apache Spark release. Make sure to install or upgrade the CDS 2 [service descriptor](#) and parcels across all machines of all clusters at the same time.

### Cloudera Manager Versions

 **Important:** Because CDS Powered by Apache Spark is only installable using the parcel mechanism, it can only be used on clusters managed by Cloudera Manager. Additionally, because Cloudera Manager does not support using parcels and packages in the same cluster, you cannot use CDS if you are using a package-based installation of CDH.

Applicable versions of Cloudera Manager for CDS Powered by Apache Spark are described below.

CDS Powered by Apache Spark Version	Supported Cloudera Manager Versions
2.4 Release 2	Cloudera Manager 5.11 and any higher Cloudera Manager 5.x versions
2.4 Release 1	
2.3 Release 4	
2.3 Release 3	
2.3 Release 2	
2.3 Release 1	Never officially released; if downloaded, do not use
2.2 Release 4	Cloudera Manager 5.8.3, 5.9 and any higher Cloudera Manager 5.x versions
2.2 Release 3	
2.2 Release 2	
2.2 Release 1	
2.1 Release 4	
2.1 Release 3	
2.1 Release 2	
2.1 Release 1	
2.0 Release 2	



CDS Powered by Apache Spark Version	Supported Cloudera Manager Versions
2.0 Release 1	

## Scala 2.11 Requirement

Spark 2 does not work with Scala 2.10. Use Scala 2.11 only.

## Python Requirement

CDS Powered by Apache Spark requires one of the following Python versions:

- Python 2.7 or higher, when using Python 2.
- Python 3.4 or higher, when using Python 3. (CDS 2.0 only supports Python 3.4 and 3.5; CDS 2.1 and higher include support for Python 3.6 and higher.)

## JDK 8 Requirement

CDS 2.2 and higher require JDK 8 only. If you are using CD 2.2 or higher, you must remove JDK 7 from all cluster and gateway hosts to ensure proper operation.

Check [the supported JDK versions](#) and see [Java Development Kit Installation](#) for the installation steps.

## CDS Powered by Apache Spark New Features and Changes

The following sections describe what's new and changed in each CDS Powered by Apache Spark release.

### What's New in CDS 2.4 Release 2

This is purely a maintenance release. See [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.4 Release 1

- Added support for [Structured Streaming](#). However, note that the following features of Structured Streaming *are not* supported:
  - Continuous processing, which is still experimental, is not supported.
  - Stream static joins with HBase have not been tested and therefore are not supported.
  - Also note the associated Known Issue here: [Structured Streaming exactly-once fault tolerance constraints](#) on page 11
- Added support for built-in Apache Avro data source. For details, refer: [SPARK-24768](#), [Apache Avro Data Source Guide](#).

Also see [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.3 Release 4

This is purely a maintenance release. See [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.3 Release 3

This is purely a maintenance release. See [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.3 Release 2

- More flexibility to interpret `TIMESTAMP` values written by Impala. Setting the `spark.sql.parquet.int96TimestampConversion` configuration setting to `true` makes Spark interpret `TIMESTAMP` values, when reading from Parquet files written by Impala, without applying any adjustment from the UTC to the local time zone of the server. This behavior provides better interoperability for Parquet data written by Impala, which does not apply any time zone adjustment to `TIMESTAMP` values when reading or writing them.

Also see [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.3 Release 1

CDS 2.3 Release 1 was never officially released; if downloaded, do not use.

### What's New in CDS 2.2 Release 4

This is purely a maintenance release. See [CDS Powered By Apache Spark Fixed Issues](#) for the list of fixed issues.

### What's New in CDS 2.2 Release 3

This is purely a maintenance release. See [CDS Powered By Apache Spark Fixed Issues](#) for the list of fixed issues.

### What's New in CDS 2.2 Release 2

This is purely a maintenance release. See [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.2 Release 1

- Support for CDH 5.12 and associated features.
- Support for using Spark 2 jobs to read and write data on the Azure Data Lake Store (ADLS) cloud service.
- CDS 2.2 and higher require JDK 8 only. If you are using CD 2.2 or higher, you must remove JDK 7 from all cluster and gateway hosts to ensure proper operation.

Also see [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.1 Release 4

This is purely a maintenance release. See [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.1 Release 3

This is purely a maintenance release. See [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.1 Release 2

This is purely a maintenance release. See [CDS Powered by Apache Spark Fixed Issues](#) on page 16 for the list of fixed issues.

### What's New in CDS 2.1 Release 1

- New direct connector to Kafka that uses the new Kafka consumer API. See [Integrating CDS Powered by Apache Spark with Apache Kafka](#) on page 90 for details.

## What's New in CDS 2.0 Release 2

- A Hive compatibility issue in CDS 2.0 Release 1 affects CDH 5.10.1 and higher, CDH 5.9.2 and higher, CDH 5.8.5 and higher, and CDH 5.7.6 and higher. If you are using one of these CDH versions, you must upgrade to the CDS 2.0 Release 2 or higher parcel, to avoid Spark 2 job failures when using Hive functionality.

## What's New in CDS 2.0 Release 1

- New `SparkSession` object replaces `HiveContext` and `SQLContext`.
  - Most of the Hive logic has been reimplemented in Spark.
  - Some Hive dependencies still exist:
    - SerDe support.
    - UDF support.
- Added support for the unified Dataset API.
- Faster Spark SQL achieved with whole stage code generation.
- More complete SQL syntax now supports subqueries.
- Adds the `spark-csv` library.
- Backport of SPARK-5847. The root for metrics is now the app name (`spark.app.name`) instead of the app ID. The app ID requires investigation to match to the app name, and changes when streaming jobs are stopped and restarted.

## CDS Powered by Apache Spark Known Issues

The following sections describe the current known issues and limitations in CDS Powered by Apache Spark. In some cases, a feature from the upstream Apache Spark project is currently not considered reliable enough to be supported by Cloudera. For a number of integration features in CDH that rely on Spark, the feature does not work with CDS Powered by Apache Spark because CDH 5 components are not introducing dependencies on Spark 2.

### Structured Streaming exactly-once fault tolerance constraints

In Spark Structured Streaming, the exactly-once fault tolerance for file sink is valid only for files that are in the manifest. These files are located in the `_spark_metadata` subdirectory of the file sink output directory. Only process files that have file names starting with digits. Other temporary files can also appear in this directory, but they should not be processed. Typically, these temporary files have names starting with a period (".").

You can list the valid manifest files, excluding the temporary files, by using a command like the following, which assumes your output directory is located at `/tmp/output`. As the appropriate user, run the following command to list the valid manifest files:

```
hadoop fs -ls /tmp/output/_spark_metadata/[0-9]*
```

### DecimalType push-down to Parquet data sources has been disabled

The support for `DecimalType` push-down to Parquet data sources, introduced in upstream Apache Spark 2.4 ([SPARK-24549](#)), has been disabled in CDS 2.4 release 1.

### JOB\_SUMMARY\_LEVEL Parquet flag is not supported

The `JOB_SUMMARY_LEVEL` Parquet flag is not supported in CDS 2.4. When writing to Parquet files, users should use the `ENABLE_JOB_SUMMARY` flag instead.

### LZ4, BROTLI, and ZSTD codecs are not supported

CDS 2.4 does not support the LZ4, BROTLI, and ZSTD codecs.

## SparkSQL StringStartsWith filter is not supported

The SparkSQL `StringStartsWith` filter is not supported with CDS 2.4.

## Apache Spark XSS vulnerability in UI CVE-2018-8024

A malicious user can construct a URL pointing to a Spark UI's job and stage info pages that can be used to execute arbitrary scripts and expose user information, if this URL is accessed by another user who is unaware of the malicious intent.

**Products affected:** CDS Powered By Apache Spark

**Releases affected:**

- CDS 2.1.0 release 1 and release 2
- CDS 2.2.0 release 1 and release 2
- CDS 2.3.0 release 2

**Users affected:** Potentially any user who uses the Spark UI.

**Date/time of detection:** May 28, 2018

**Detected by:** Spencer Gietzen (Rhino Security Labs)

**Severity (Low/Medium/High):** High

**Impact:** XSS vulnerabilities can be used to steal credentials or to perform arbitrary actions as the targeted user.

**CVE:** CVE-2018-8024

**Immediate action required:** Upgrade to a version of CDS Powered by Apache Spark where this issue is fixed, or as a workaround, disable the Spark UI for jobs and the Spark History Server.

**Addressed in release/refresh/patch:** CDS 2.3.0 release 3.

## CDS 2.3 Release 3 requires additional configuration to run PySpark on Cloudera Data Science Workbench versions 1.3.x (and lower)

Due to a security fix in CDS 2.3 release 3, there is now a mismatch between the versions of `py4j` that ship with the two products:

- Cloudera Data Science Workbench 1.3.x (and lower) includes `py4j 0.10.4`, and,
- CDS 2.3 release 3 includes `py4j 0.10.7`.

This version mismatch results in PySpark session/job failures on Cloudera Data Science Workbench.

**Workaround:** The Cloudera Data Science Workbench documentation includes more details about [this known issue and its workarounds](#).

**Cloudera Bug:** CDH-69733, DSE-4316

## UnsatisfiedLinkError observed when using Snappy compression in the spark2-shell

In CDS 2.3 release 2, when you use `spark2-shell` to read or write to a parquet table with snappy compression, the following `UnsatisfiedLinkError` occurs:

```
java.lang.UnsatisfiedLinkError:
org.xerial.snappy.SnappyNative.uncompressedLength(Ljava/nio/ByteBuffer;II)I
at org.xerial.snappy.SnappyNative.uncompressedLength(Native Method)
at org.xerial.snappy.Snappy.uncompressedLength(Snappy.java:561)
at parquet.hadoop.codec.SnappyDecompressor.decompress(SnappyDecompressor.java:62)
```

This happens because the 1.0.4.1 version of snappy needs to access the top-most class loader which collides with a change related to fixing `userClassPathFirst` (SPARK-18646).

Applications run with `spark2-submit` are not affected by this issue.

**Workaround:** Copy `snappy-java-1.1.4.jar` to `/opt/cloudera/parcels/SPARK2/lib/spark2/jars/` on every node in the cluster. You can download `snappy-java-1.1.4.jar` from:

<https://repository.cloudera.com/cloudera/list/repo1/org/xerial/snappy/snappy-java/1.1.4/snappy-java-1.1.4.jar>

**Cloudera Bug:** CDH-67889

**Resolution:** Upgrade to CDS 2.3 Release 3, which contains the fix.

## Spark jobs fail when lineage collection is enabled

In CDS 2.3 release 2, Spark jobs fail when lineage is enabled because Cloudera Manager does not automatically create the associated lineage log directory (`/var/log/spark2/lineage`) on all required cluster hosts. Note that this feature is enabled by default in CDS 2.3 release 2.

Implement one of the following workarounds to continue running Spark jobs.

### Workaround 1 - Deploy the Spark gateway role on all hosts that are running the YARN NodeManager role

Cloudera Manager only creates the lineage log directory on hosts with Spark 2 roles deployed on them. However, this is not sufficient because the Spark driver can run on any host that is running a YARN NodeManager. To ensure Cloudera Manager creates the log directory, add the Spark 2 gateway role to every cluster host that is running the YARN NodeManager role.

For instructions on how to add a role to a host, see the Cloudera Manager documentation: [Adding a Role Instance](#)

### Workaround 2 - Disable Spark Lineage Collection

To disable the feature, log in to Cloudera Manager and go to the Spark 2 service. Click **Configuration**. Search for the **Enable Lineage Collection** property and uncheck the checkbox to disable lineage collection. Click **Save Changes**.



**Important:** This issue also affects Spark jobs deployed through Cloudera Data Science Workbench. For details, see [the associated Known Issue](#) in the Cloudera Data Science Workbench documentation.

**Cloudera Bug:** CDH-67643, CDH-68832

**Partial Resolution:** Upgrade to CDS 2.3 Release 3, which contains a partial fix. Spark jobs do not fail, but lineage is not collected.

## Spark on Kubernetes is not supported

The “Spark On Kubernetes” feature of Apache Spark 2.3 (and higher) is currently not supported. This feature is currently still designated as experimental within Apache Spark.

## ORC file format is not supported

Currently, Cloudera does not support reading and writing Hive tables containing data files in the Apache ORC (Optimized Row Columnar) format from Spark applications. Cloudera recommends using Apache Parquet format for columnar data. That file format can be used with Spark, Hive, and Impala.

## Accessing multiple clusters simultaneously is not supported

Spark does not support accessing multiple clusters in the same application.

## Parquet logical type `TIMESTAMP_MICROS` unavailable

Although SPARK-10365 introduces the Parquet logical type `TIMESTAMP_MICROS`, this logical type is not available in the Parquet support libraries included with CDS Powered by Apache Spark.

## Spark SQL does not respect Sentry ACLs when communicating with Hive metastore

Even if user is configured via Sentry to not have read permission to a Hive table, a Spark SQL job running as that user can still read the table's metadata directly from the Hive metastore.

**Cloudera Bug:** CDH-33658

### Empty result when reading Parquet table created by `saveAsTable()`

After a Parquet table is created by the `saveAsTable()` function, Spark SQL queries against the table return an empty result set. The issue is caused by the “path” property of the table not being written to the Hive metastore during the `saveAsTable()` call.

**Cloudera Bug:** CDH-60037

**Affects:** CDS 2.2 Release 1

**Severity:** High

**Workaround:** You can set the path manually before the call to `saveAsTable()`:

```
val options = Map("path" -> "/path/to/hdfs/directory/containing/table")
df.write.options(options).saveAsTable("db_name.table_name")
```

Or you can add the path to the metastore when the table already exists, for example:

```
spark.sql("alter table db_name.table_name set SERDEPROPERTIES
('path'='hdfs://host.example.com:8020/warehouse/path/db_name.db/table_name' )")
spark.catalog.refreshTable("db_name.table_name")
```

**Resolution:** Upgrade to CDS 2.2 Release 2, which contains the fix.

### Spark 2 Version requirement for clusters managed by Cloudera Manager



**Important:**

Because CDS Powered by Apache Spark is only installable using the parcel mechanism, it can only be used on clusters managed by Cloudera Manager. Additionally, because Cloudera Manager does not support using parcels and packages in the same cluster, you cannot use CDS if you are using a package-based installation of CDH.

Although Spark 1 and Spark 2 can coexist in the same CDH cluster, you cannot use multiple Spark 2 versions simultaneously in the same Cloudera Manager instance. All CDH clusters managed by the same Cloudera Manager Server must use exactly the same version of CDS Powered by Apache Spark. For example, you cannot use the built-in CDH Spark service, a CDS 2.1 service, and a CDS 2.2 service. You must choose only one CDS 2 Powered by Apache Spark release. Make sure to install or upgrade the CDS 2 [service descriptor](#) and parcels across all machines of all clusters at the same time.

### Spark Standalone is not supported

Spark Standalone is not supported for CDS Powered by Apache Spark.

### HiveOnSpark is not supported with CDS

The HiveOnSpark module is a CDH 5 component that has a dependency on Apache Spark 1.6. Because CDH 5 components do not have any dependencies on Spark 2, the HiveOnSpark module does not work with CDS Powered by Apache Spark. You can still use Spark 2 with Hive using other methods.

### SparkOnHBase is not supported with CDS

The SparkOnHBase module is a CDH 5 component that has a dependency on Apache Spark 1.6. Because CDH 5 components do not have any dependencies on Spark 2, the SparkOnHBase module does not work with CDS Powered by Apache Spark. You can still use Spark 2 with HBase using other methods.

## Using the JDBC Datasource API to access Hive or Impala is not supported

## Dynamic allocation and Spark Streaming

If you are using Spark Streaming, Cloudera recommends that you disable dynamic allocation by setting `spark.dynamicAllocation.enabled` to `false` when running streaming applications.

## Oozie Spark2 action is not supported

The Oozie Spark action is a CDH component that has a dependency on Spark 1.6. Because CDH components do not have any dependencies on Spark 2, the Oozie Spark action does not work with CDS Powered by Apache Spark.

## SparkR is not supported

SparkR is not supported for CDS Powered by Apache Spark. (SparkR is also not supported in CDH with Spark 1.6.)

## GraphX is not supported

GraphX is not supported for CDS Powered by Apache Spark. (GraphX is also not supported in CDH with Spark 1.6.)

## Thrift server is not supported

The Thrift JDBC/ODBC server is not supported for CDS Powered by Apache Spark. (The Thrift server is also not supported in CDH with Spark 1.6.)

## Spark SQL CLI is not supported

The Spark SQL CLI is not supported for CDS Powered by Apache Spark. (The Spark SQL CLI is also not supported in CDH with Spark 1.6.)

## Rolling upgrades are not supported

Rolling upgrades are not possible from Spark 1.6 bundled with CDH, to CDS Powered by Apache Spark.

## Package-based installation is not supported

CDS Powered by Apache Spark is only installable as a parcel.

## Spark-Avro Library is not supported; Use Built-In Avro Data Source

The `spark-avro` library is not integrated into the CDS Powered by Apache Spark parcel. Starting with CDS 2.4 release 1, you can use the built-in Avro data source instead. For documentation, see [Apache Avro Data Source Guide](#).

## Hardware acceleration for MLlib is not supported

This feature, part of the GPL Extras package for CDH, is not supported with the CDS 2 Powered By Apache Spark. This feature is supported for Spark 1.6.

## Cost based optimization is not supported

The cost based optimization feature is not supported in CDS Powered by Apache Spark. Do NOT set the `spark.sql.cbo.enabled` configuration option to `true`.

## Running `spark2-submit` with `--principal` and `--keytab` arguments does not work in client mode

The `spark2-submit` script's `--principal` and `--keytab` arguments do not work with Spark-on-YARN's `client` mode. Use `cluster` mode instead.

### Long-running apps on a secure cluster might fail if driver is restarted

If you submit a long-running app on a secure cluster using the `--principal` and `--keytab` options in cluster mode, and a failure causes the driver to restart after 7 days (the default maximum HDFS delegation token lifetime), the new driver fails with an error similar to the following:

```
Exception in thread "main"  
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.security.token.SecretManager$InvalidToken):  
token <token_info> can't be found in cache
```

**Workaround:** None

**Affected Versions:** All CDS 2.0, 2.1, and 2.2 releases

**Fixed Versions:** CDS 2.3 Release 2

**Apache Issue:** [SPARK-23361](#)

**Cloudera Issue:** CDH-64865

### CDS Powered by Apache Spark Incompatible Changes

The following sections describe changes in CDS Powered by Apache Spark that might require special handling during upgrades, or code changes within existing applications.

#### Incompatible Changes in CDS 2.4

No new incompatible changes in this release.

#### Incompatible Changes in CDS 2.3

No new incompatible changes in this release.

#### Incompatible Changes in CDS 2.2

No new incompatible changes in this release.

#### Incompatible Changes in CDS 2.1

No new incompatible changes in this release.

#### Incompatible Changes in CDS 2.0

- `HiveContext` and `SQLContext` have been removed.
- `DataFrames` have been removed from the Scala API. `DataFrame` is now a special case of `Dataset`.
- Spark 2.0 and higher do not use an assembly JAR for standalone applications.

### CDS Powered by Apache Spark Fixed Issues

The following sections describe the issues fixed in each CDS Powered by Apache Spark release.



## Issues Fixed in CDS 2.4 Release 2



**Important:** The fix for [SPARK-25250](#) in CDS 2.4 Release 1 introduced an issue where jobs can hang indefinitely. SPARK-25250 is reverted in CDS 2.4 Release 2. If you are installing or upgrading to CDS 2.4, use Release 2 or higher.

For more information, see the discussions in the Apache Spark Git pull requests [24359](#) and [24375](#).

The following list includes issues fixed in CDS 2.4 Release 2. Test-only changes have been omitted. In addition to the fixes listed here, this release also includes all the fixes that are in the Apache Spark 2.4.2 upstream release. For more information, see the [Apache Spark 2.4.2 upstream release notes](#).

- [\[SPARK-13704\]](#) Reduce rack resolution time

## Issues Fixed in CDS 2.4 Release 1



**Important:** The fix for [SPARK-25250](#) in CDS 2.4 Release 1 introduced an issue where jobs can hang indefinitely. SPARK-25250 is reverted in CDS 2.4 Release 2. If you are installing or upgrading to CDS 2.4, use Release 2 or higher.

For more information, see the discussions in the Apache Spark Git pull requests [24359](#) and [24375](#).

The following list includes issues fixed in CDS 2.4 Release 1. Test-only changes have been omitted. In addition to the fixes listed here, this release also includes all the fixes that are in the Apache Spark 2.4.1 upstream release. For more information, see the [Apache Spark 2.4.1 upstream release notes](#).

- [\[SPARK-26089\]](#) Handle large corrupt shuffle blocks
- [\[SPARK-26349\]](#) PySpark should not accept insecure p4j gateways
- [\[SPARK-26688\]](#) Provide configuration of initially blacklisted YARN nodes
- [\[SPARK-27094\]](#) Thread interrupt being swallowed while launching executors in YarnAllocator
- [\[SPARK-24865\]](#) AnalysisBarrier removed in Apache Spark 2.4. This feature was introduced in Apache Spark 2.3 but has now been removed due to performance issues.
- **[Security Fix]** Even though [SPARK-26019](#) was a part of the upstream Spark 2.4.1 release, it has been omitted from CDS 2.4 release 1 to prevent users from passing in insecure `py4j` connections on secure servers.

## Issues Fixed in CDS 2.3 Release 4

The following list includes issues fixed in CDS 2.3 Release 4. Test-only changes are omitted. This release includes all fixes that are in the Apache Spark 2.3.2 upstream release. For more information, see the [Apache Spark 2.3.2 upstream release notes](#).

- [\[SPARK-25454\]](#)[SQL] add a new config for picking minimum precision for integral literals
- [\[SPARK-24918\]](#)[CORE] Executor Plugin API

## Issues Fixed in CDS 2.3 Release 3

The following list includes issues fixed in CDS 2.3 Release 3. Test-only changes have been omitted. In addition to the fixes listed here, this release also includes all fixes that are in the Apache Spark 2.3.1 upstream release. For more information, see the [Apache Spark 2.3.1 upstream release notes](#).

- [\[SPARK-16451\]](#)[REPL] Spark-shell / pyspark should finish gracefully when "SaslException: GSS initiate failed" is hit
- [\[SPARK-17756\]](#)[PYTHON][STREAMING] java.lang.ClassCastException returned when using 'cartesian' with DStream.transform
- [\[SPARK-24029\]](#) Set the "reuse address" flag on listen sockets
- [\[SPARK-24216\]](#)[SQL] Spark TypedAggregateExpression uses getSimpleName this is not safe in Scala
- [\[SPARK-24369\]](#)[SQL] Correct handling for multiple distinct aggregations that have the same argument set

- [\[SPARK-24468\]](#)[SQL] DecimalType 'adjustPrecisionScale' might fail when scale is negative
- [\[SPARK-24495\]](#)[SQL] SortMergeJoin with duplicate keys produces wrong results
- [\[SPARK-24506\]](#)[UI] Add UI filters to tabs added after binding
- [\[SPARK-24542\]](#)[SQL] Hive UDF series UDFXPathXXXX allows users to pass carefully crafted XML to access arbitrary files
- [\[SPARK-24548\]](#)[SQL] JavaPairRDD to Dataset<Row> in Spark generates ambiguous results
- [\[SPARK-24552\]](#) Task attempt numbers are reused when stages are retried
- [\[SPARK-24578\]](#)[CORE] Reading remote cache block behavior changes and causes timeout issue
- [\[SPARK-24583\]](#)[SQL] Wrong schema type in InsertIntoDataSourceCommand
- [\[SPARK-24589\]](#)[CORE] OutputCommitCoordinator might allow duplicate commits

### Issues Fixed in CDS 2.3 Release 2

The following list includes issues fixed in CDS 2.3 Release 2. Test-only changes have been omitted. In addition to the fixes listed here, this release also includes all the fixes that are in the Apache Spark 2.3.0 upstream release. For more information, see the [Apache Spark 2.3.0 upstream release notes](#).

- [\[SPARK-23361\]](#)[YARN] Allow AM to restart after initial tokens expire
- [\[SPARK-23644\]](#)[CORE][UI][BACKPORT-2.3] Use absolute path for REST call in SHS
- [\[SPARK-23623\]](#)[SS] Avoid concurrent use of cached consumers in CachedKafkaConsumer (branch-2.3)
- [\[SPARK-23670\]](#)[SQL] Fix memory leak on SparkPlanGraphWrapper
- [\[SPARK-23608\]](#)[CORE][WEBUI] Add synchronization in SHS between attachSparkUI and detachSparkUI functions to avoid concurrent modification issue to Jetty Handlers
- [\[SPARK-23671\]](#)[CORE] Fix condition to enable the SHS thread pool
- [\[SPARK-23658\]](#)[LAUNCHER] InProcessAppHandle uses the wrong class in getLogger
- [\[SPARK-23020\]](#)[CORE][BRANCH-2.3] Fix another race in the in-process launcher test
- [\[SPARK-23551\]](#)[BUILD] Exclude `hadoop-mapreduce-client-core` dependency from `orc-mapreduce`
- [\[SPARK-23475\]](#)[UI][BACKPORT-2.3] Show also skipped stages
- [\[SPARK-23729\]](#)[CORE] Respect URI fragment when resolving globs
- [\[SPARK-23660\]](#) Fix exception in yarn cluster mode when application ended fast
- [\[SPARK-23438\]](#)[DSTREAMS] Fix DStreams data loss with WAL when driver crashes
- [\[SPARK-23630\]](#)[YARN] Allow user's hadoop conf customizations to take effect.
- [\[SPARK-23476\]](#)[CORE] Generate secret in local mode when authentication on

### Apache Spark XSS vulnerability in UI CVE-2018-8024

A malicious user can construct a URL pointing to a Spark UI's job and stage info pages that can be used to execute arbitrary scripts and expose user information, if this URL is accessed by another user who is unaware of the malicious intent.

**Products affected:** CDS Powered By Apache Spark

**Releases affected:**

- CDS 2.1.0 release 1 and release 2
- CDS 2.2.0 release 1 and release 2
- CDS 2.3.0 release 2

**Users affected:** Potentially any user who uses the Spark UI.

**Date/time of detection:** May 28, 2018

**Detected by:** Spencer Gietzen (Rhino Security Labs)

**Severity (Low/Medium/High):** High

**Impact:** XSS vulnerabilities can be used to steal credentials or to perform arbitrary actions as the targeted user.

**CVE:** CVE-2018-8024

**Immediate action required:** Upgrade to a version of CDS Powered by Apache Spark where this issue is fixed, or as a workaround, disable the Spark UI for jobs and the Spark History Server.

**Addressed in release/refresh/patch:** CDS 2.3.0 release 3.

### Issues Fixed in CDS 2.3 Release 1

CDS 2.3 Release 1 was never officially released; if downloaded, do not use.

### Issues Fixed in CDS 2.2 Release 4

The following list includes issues fixed in CDS 2.2 Release 4. Test-only changes are omitted.

- [\[SPARK-25402\]](#)[SQL][BACKPORT-2.2] Null handling in BooleanSimplification
- [\[SPARK-24809\]](#)[SQL] Serializing LongToUnsafeRowMap in executor may result in data error
- [\[SPARK-24957\]](#)[SQL][BACKPORT-2.2] Average with decimal followed by aggregation returns wrong result
- [\[SPARK-24918\]](#)[CORE] Executor Plugin API
- [PYSARK][SQL] Updates to RowQueue
- [PYSARK] Updates to pyspark broadcast
- [\[SPARK-25253\]](#)[PYSARK] Refactor local connection & auth code
- [\[SPARK-23243\]](#)[\[SPARK-20715\]](#)[CORE][2.2] Fix RDD.repartition() data correctness issue

### Issues Fixed in CDS 2.2 Release 3

The following list includes issues fixed in CDS 2.2 Release 3. Test-only changes are omitted.

- [PYSARK] Updates to Accumulators
- [\[SPARK-21525\]](#)[STREAMING] Check error code from supervisor RPC.
- [\[SPARK-24552\]](#)[CORE] Use unique id instead of attempt number for writes.
- [\[SPARK-22897\]](#)[CORE] Expose stageAttemptId in TaskContext
- [\[SPARK-24589\]](#)[CORE] Correctly identify tasks in output commit coordinator.
- [\[SPARK-24506\]](#)[UI] Add UI filters to tabs added after binding
- [WEBUI] Avoid possibility of script in query param keys
- [MINOR] Add port SSL config in toString and scaladoc
- [\[SPARK-24257\]](#)[SQL] LongToUnsafeRowMap calculate the new size may be wrong
- [\[SPARK-23850\]](#)[SQL] Add separate config for SQL options redaction.
- [PYSARK] Update py4j to version 0.10.7.
- [\[SPARK-21278\]](#)[PYSARK] Upgrade to Py4J 0.10.6
- [\[SPARK-23697\]](#)[CORE] LegacyAccumulatorWrapper should define isZero correctly
- [\[SPARK-23941\]](#)[MESOS] Mesos task failed on specific spark app name
- [\[SPARK-23963\]](#)[SQL] Properly handle large number of columns in query on text-based Hive table
- [\[SPARK-24007\]](#)[SQL] EqualNullSafe for FloatType and DoubleType might generate a wrong result by codegen.
- [\[SPARK-23759\]](#)[UI] Unable to bind Spark UI to specific host name / IP
- [\[SPARK-23649\]](#)[SQL] Skipping chars disallowed in UTF-8
- [\[SPARK-23525\]](#)[BACKPORT][SQL] Support ALTER TABLE CHANGE COLUMN COMMENT for external hive table
- [\[SPARK-23434\]](#)[SQL] Spark should not warn `metadata directory` for a HDFS file path
- [\[SPARK-23508\]](#)[CORE] Fix BlockmanagerId in case blockManagerIdCache cause oom
- [\[SPARK-22700\]](#)[ML] Bucketizer.transform incorrectly drops row containing NaN - for branch-2.2
- [\[SPARK-23230\]](#)[SQL] When hive.default.fileformat is other kinds of file types, create textfile table cause a serde error
- [\[SPARK-23053\]](#)[CORE] taskBinarySerialization and task partitions calculate in DagScheduler.submitMissingTasks should keep the same RDD checkpoint status
- [\[SPARK-23391\]](#)[CORE] It may lead to overflow for some integer multiplication
- [\[SPARK-23376\]](#)[SQL] creating UnsafeKVExternalSorter with BytesToBytesMap may fail
- [\[SPARK-23186\]](#)[SQL] Initialize DriverManager first before loading JDBC Drivers

- [\[SPARK-23358\]](#)[CORE] When the number of partitions is greater than  $2^{28}$ , it will result in an error result
- [\[SPARK-23281\]](#)[SQL] Query produces results in incorrect order when a composite order by clause refers to both original columns and aliases
- [\[SPARK-23095\]](#)[SQL] Decorrelation of scalar subquery fails with java.util.NoSuchElementException
- [\[SPARK-22982\]](#) Remove unsafe asynchronous close() call from FileDownloadChannel
- [\[SPARK-23001\]](#)[SQL] Fix NullPointerException when DESC a database with NULL description
- [\[SPARK-22972\]](#) Couldn't find corresponding Hive SerDe for data source provider org.apache.spark.sql.hive.org
- [\[SPARK-22984\]](#) Fix incorrect bitmap copying and offset adjustment in GenerateUnsafeRowJoiner
- [\[SPARK-22983\]](#) Don't push filters beneath aggregates with empty grouping expressions
- [\[SPARK-22862\]](#) Docs on lazy elimination of columns missing from an encoder
- [\[SPARK-22574\]](#)[MESOS][SUBMIT] Check submission request parameters
- [\[SPARK-22289\]](#)[ML] Add JSON support for Matrix parameters (LR with coefficients bound)
- [\[SPARK-22688\]](#)[SQL] Upgrade Janino version to 3.0.8
- [\[SPARK-22686\]](#)[SQL] DROP TABLE IF EXISTS should not show AnalysisException
- [\[SPARK-22635\]](#)[SQL][ORC] FileNotFoundException while reading ORC files containing special characters
- [\[SPARK-22601\]](#)[SQL] Data load is getting displayed successful on providing non existing nonlocal file path
- [\[SPARK-22653\]](#) executorAddress registered in CoarseGrainedSchedulerBac...
- [\[SPARK-22373\]](#) Bump Janino dependency version to fix thread safety issue...
- [\[SPARK-22637\]](#)[SQL] Only refresh a logical plan once.
- [\[SPARK-22603\]](#)[SQL] Fix 64KB JVM bytecode limit problem with FormatString
- [\[SPARK-23852\]](#)[SQL] Add test that fails if PARQUET-1217 is not fixed.
- [\[SPARK-23991\]](#)[DSTREAMS] Fix data loss when WAL write fails in allocateBlocksToBatch
- [\[SPARK-24309\]](#)[CORE] AsyncEventQueue should stop on interrupt.
- [\[SPARK-19181\]](#)[CORE] Fixing flaky "SparkListenerSuite.local metrics"
- [\[SPARK-23433\]](#)[CORE] Late zombie task completions update all tasksets
- [\[SPARK-23816\]](#)[CORE] Killed tasks should ignore FetchFailures.
- [\[SPARK-22864\]](#)[CORE] Disable allocation schedule in ExecutorAllocationManagerSuite.
- [\[SPARK-22495\]](#) Fix setup of SPARK\_HOME variable on Windows
- [\[SPARK-22595\]](#)[SQL] fix flaky test: CastSuite.SPARK-22500: cast for struct should not generate codes beyond 64KB
- [\[SPARK-22591\]](#)[SQL] GenerateOrdering shouldn't change CodegenContext.INPUT\_ROW
- [\[SPARK-17920\]](#)[SQL] [FOLLOWUP] Backport PR 19779 to branch-2.2
- [\[SPARK-17920\]](#)[\[SPARK-19580\]](#)[\[SPARK-19878\]](#)[SQL] Backport PR 19779 to branch-2.2 - Support writing to Hive table which uses Avro schema url 'avro.schema.url'
- [\[SPARK-22548\]](#)[SQL] Incorrect nested AND expression pushed down to JDBC data source
- [\[SPARK-22500\]](#)[SQL] Fix 64KB JVM bytecode limit problem with cast
- [\[SPARK-22550\]](#)[SQL] Fix 64KB JVM bytecode limit problem with elt
- [\[SPARK-22508\]](#)[SQL] Fix 64KB JVM bytecode limit problem with GenerateUnsafeRowJoiner.create()
- [\[SPARK-22549\]](#)[SQL] Fix 64KB JVM bytecode limit problem with concat\_ws
- [\[SPARK-22498\]](#)[SQL] Fix 64KB JVM bytecode limit problem with concat
- [\[SPARK-22538\]](#)[ML] SQLTransformer should not unpersist possibly cached input dataset
- [\[SPARK-22540\]](#)[SQL] Ensure HighlyCompressedMapStatus calculates correct avgSize
- [\[SPARK-22535\]](#)[PYSPARK] Sleep before killing the python worker in PythRunner.MonitorThread (branch-2.2)
- [\[SPARK-22501\]](#)[SQL] Fix 64KB JVM bytecode limit problem with in
- [\[SPARK-22494\]](#)[SQL] Fix 64KB limit exception with Coalesce and AtLeastNNonNulls
- [\[SPARK-22499\]](#)[SQL] Fix 64KB JVM bytecode limit problem with least and greatest
- [\[SPARK-22479\]](#)[SQL] Exclude credentials from SaveintoDataSourceCommand.simpleString
- [\[SPARK-22469\]](#)[SQL] Accuracy problem in comparison with string and numeric
- [\[SPARK-22471\]](#)[SQL] SQLListener consumes much memory causing OutOfMemoryError
- [\[SPARK-22442\]](#)[SQL][FOLLOWUP] ScalaReflection should produce correct field names for special characters
- [\[SPARK-22442\]](#)[SQL] ScalaReflection should produce correct field names for special characters

- [\[SPARK-22464\]](#)[BACKPORT-2.2][SQL] No pushdown for Hive metastore partition predicates containing null-safe equality
- [\[SPARK-22488\]](#)[BACKPORT-2.2][SQL] Fix the view resolution issue in the SparkSession internal table() API
- [\[SPARK-21720\]](#)[SQL] Fix 64KB JVM bytecode limit problem with AND or OR
- [\[SPARK-19606\]](#)[MESOS] Support constraints in spark-dispatcher
- [\[SPARK-21667\]](#)[STREAMING] ConsoleSink should not fail streaming query with checkpointLocation option
- [\[SPARK-19644\]](#)[SQL] Clean up Scala reflection garbage after creating Encoder (branch-2.2)
- [\[SPARK-22284\]](#)[SQL] Fix 64KB JVM bytecode limit problem in calculating hash for nested structs
- [\[SPARK-22294\]](#)[DEPLOY] Reset spark.driver.bindAddress when starting a Checkpoint
- [\[SPARK-22243\]](#)[DSTREAM] spark.yarn.jars should reload from config when checkpoint recovery
- [\[SPARK-22472\]](#)[SQL] add null check for top-level primitive values
- [\[SPARK-22287\]](#)[MESOS] SPARK\_DAEMON\_MEMORY not honored by MesosClusterD...
- [\[SPARK-22417\]](#)[PYTHON][FOLLOWUP] Fix for createDataFrame from pandas.DataFrame with timestamp
- [\[SPARK-22417\]](#)[PYTHON] Fix for createDataFrame from pandas.DataFrame with timestamp
- [\[SPARK-22429\]](#)[STREAMING] Streaming checkpointing code does not retry after failure
- [\[SPARK-22211\]](#)[SQL] Remove incorrect FOJ limit pushdown
- [\[SPARK-22333\]](#)[SQL][BACKPORT-2.2] timeFunctionCall(CURRENT\_DATE, CURRENT\_TIMESTAMP) has conflicts with columnReference
- [\[SPARK-22291\]](#)[SQL] Conversion error when transforming array types of uuid, inet and cidr to StingType in PostgreSQL
- [\[SPARK-19727\]](#)[SQL][FOLLOWUP] Fix for round function that modifies original column
- [\[SPARK-22356\]](#)[SQL] data source table should support overlapped columns between data and partition schema
- [\[SPARK-22355\]](#)[SQL] Dataset.collect is not threadsafe
- [\[SPARK-22328\]](#)[CORE] ClosureCleaner should not miss referenced superclass fields
- [\[SPARK-22227\]](#)[CORE] DiskBlockManager.getAllBlocks now tolerates temp files
- [\[SPARK-22319\]](#)[CORE][BACKPORT-2.2] call loginUserFromKeytab before accessing hdfs
- [\[SPARK-22249\]](#)[FOLLOWUP][SQL] Check if list of value for IN is empty in the optimizer
- [\[SPARK-22271\]](#)[SQL] mean overflows and returns null for some decimal variables
- [\[SPARK-22249\]](#)[SQL] isin with empty list throws exception on cached DataFrame
- [\[SPARK-22223\]](#)[SQL] ObjectHashAggregate should not introduce unnecessary shuffle
- [\[SPARK-22273\]](#)[SQL] Fix key/value schema field names in HashMapGenerators.
- [\[SPARK-14387\]](#) [\[SPARK-16628\]](#) [\[SPARK-18355\]](#)[SQL] Use Spark schema to read ORC table instead of ORC file schema
- [\[SPARK-22217\]](#)[SQL] ParquetFileFormat to support arbitrary OutputCommitters
- [\[SPARK-22218\]](#) spark shuffle services fails to update secret on app re-attempts
- [\[SPARK-20466\]](#)[CORE] HadoopRDD#addLocalConfiguration throws NPE
- [\[SPARK-22178\]](#)[SQL] Refresh Persistent Views by REFRESH TABLE Command
- [\[SPARK-22158\]](#)[SQL] convertMetastore should not ignore table property
- [\[SPARK-22146\]](#) FileNotFoundException while reading ORC files containing special characters
- [\[SPARK-22161\]](#)[SQL] Add Impala-modified TPC-DS queries
- [\[SPARK-22129\]](#)[SPARK-22138] Release script improvements
- [\[SPARK-22143\]](#)[SQL] Fix memory leak in OffHeapColumnVector
- [\[SPARK-22135\]](#)[MESOS] metrics in spark-dispatcher not being registered properly
- [\[SPARK-22140\]](#) Add TPCDSQuerySuite
- [\[SPARK-22141\]](#)[BACKPORT][SQL] Propagate empty relation before checking Cartesian products
- [\[SPARK-22120\]](#)[SQL] TestHiveSparkSession.reset() should clean out Hive warehouse directory
- [\[SPARK-22107\]](#) Change as to alias in python quickstart
- [\[SPARK-22109\]](#)[SQL] Resolves type conflicts between strings and timestamps in partition column
- [\[SPARK-22092\]](#) Reallocation in OffHeapColumnVector.reserveInternal corrupts struct and array data
- [\[SPARK-18136\]](#) Fix SPARK\_JARS\_DIR for Python pip install on Windows
- [\[SPARK-21384\]](#)[YARN] Spark + YARN fails with LocalFileSystem as default FS

- [\[SPARK-22076\]](#)[SQL] Expand.projections should not be a Stream
- [\[SPARK-22052\]](#) Incorrect Metric assigned in MetricsReporter.scala
- [\[SPARK-22043\]](#)[PYTHON] Improves error message for show\_profiles and dump\_profiles
- [\[SPARK-21953\]](#) Show both memory and disk bytes spilled if either is present
- [\[SPARK-21985\]](#)[PYSPARK] PairDeserializer is broken for double-zipped RDDs
- [\[SPARK-18608\]](#)[ML][FOLLOWUP] Fix double caching for PySpark OneVsRest.
- [\[SPARK-21980\]](#)[SQL] References in grouping functions should be indexed with semanticEquals
- [\[SPARK-18608\]](#)[ML] Fix double caching
- [\[SPARK-20098\]](#)[PYSPARK] dataType's typeName fix
- [\[SPARK-21954\]](#)[SQL] JacksonUtils should verify MapType's value type instead of key type
- [\[SPARK-21915\]](#)[ML][PYSPARK] Model 1 and Model 2 ParamMaps Missing
- [\[SPARK-21925\]](#) Update trigger interval documentation in docs with behavior change in Spark 2.2
- [\[SPARK-21418\]](#)[SQL] NoSuchElementException: None.get in DataSourceScanExec with sun.io.serialization.extendedDebugInfo=true
- [\[SPARK-21884\]](#)[\[SPARK-21477\]](#)[BACKPORT-2.2][SQL] Mark LocalTableScanExec's input data transient
- [\[SPARK-21714\]](#)[CORE][BACKPORT-2.2] Avoiding re-uploading remote resources in yarn client mode
- [\[SPARK-21798\]](#) No config to replace deprecated SPARK\_CLASSPATH config for launching daemons like History Server
- [\[SPARK-21818\]](#)[ML][MLLIB] Fix bug of MultivariateOnlineSummarizer.variance generate negative result
- [\[SPARK-21681\]](#)[ML] fix bug of MLOR do not work correctly when featureStd contains zero (backport PR for 2.2)
- [\[SPARK-21826\]](#)[SQL] outer broadcast hash join should not throw NPE
- [\[SPARK-21721\]](#)[SQL][FOLLOWUP] Clear FileSystem deleteOnExit cache when paths are successfully removed
- [\[SPARK-21739\]](#)[SQL] Cast expression should initialize timeZoneId when it is called statically to convert something into TimestampType
- [\[SPARK-18464\]](#)[SQL][BACKPORT] support old table which doesn't store schema in table properties
- [\[SPARK-21656\]](#)[CORE] spark dynamic allocation should not idle timeout executors when tasks still to run
- [\[SPARK-21723\]](#)[ML] Fix writing LibSVM (key not found: numFeatures)
- [\[SPARK-21721\]](#)[SQL] Clear FileSystem deleteOnExit cache when paths are successfully removed
- [\[SPARK-21563\]](#)[CORE] Fix race condition when serializing TaskDescriptions and adding jars
- [\[SPARK-21595\]](#) Separate thresholds for buffering and spilling in ExternalAppendOnlyUnsafeRowArray
- [\[SPARK-21699\]](#)[SQL] Remove unused getTableOption in ExternalCatalog
- [\[SPARK-21503\]](#)[UI] Spark UI shows incorrect task status for a killed Executor Process
- [\[SPARK-21648\]](#)[SQL] Fix confusing assert failure in JDBC source when parallel fetching parameters are not properly provided.
- [\[SPARK-21374\]](#)[CORE] Fix reading globbed paths from S3 into DF with disabled FS cache
- [\[SPARK-21647\]](#)[SQL] Fix SortMergeJoin when using CROSS
- [\[SPARK-21621\]](#)[CORE] Reset numRecordsWritten after DiskBlockObjectWriter.commitAndGet called
- [\[SPARK-21588\]](#)[SQL] SQLContext.getConf(key, null) should return null
- [\[SPARK-21580\]](#)[SQL] Integers in aggregation expressions are wrongly taken as group-by ordinal
- [\[SPARK-21330\]](#)[SQL] Bad partitioning does not allow to read a JDBC table with extreme values on the partition column
- [\[SPARK-12717\]](#)[PYTHON] Adding thread-safe broadcast pickle registry
- [\[SPARK-21339\]](#)[CORE] spark-shell --packages option does not add jars to classpath on windows
- [\[SPARK-21555\]](#)[SQL] RuntimeReplaceable should be compared semantically by its canonicalized child
- [\[SPARK-21306\]](#)[ML] OneVsRest should support setWeightCol
- [\[SPARK-21538\]](#)[SQL] Attribute resolution inconsistency in the Dataset API
- [\[SPARK-21494\]](#)[NETWORK] Use correct app id when authenticating to external service.
- [\[SPARK-21447\]](#)[WEB UI] Spark history server fails to render compressed
- [\[SPARK-21383\]](#)[YARN] Fix the YarnAllocator allocates more Resource
- [\[SPARK-21243\]](#)[CORE] Limit no. of map outputs in a shuffle fetch

- [\[SPARK-21446\]](#)[SQL] Fix setAutoCommit never executed
- [\[SPARK-21441\]](#)[SQL] Incorrect Codegen in SortMergeJoinExec results failures in some cases
- [\[SPARK-21414\]](#) Refine SlidingWindowFunctionFrame to avoid OOM.
- [\[SPARK-21457\]](#)[SQL] ExternalCatalog.listPartitions should correctly handle partition values with dot
- [\[SPARK-21445\]](#) Make IntWrapper and LongWrapper in UTF8String Serializable
- [\[SPARK-21332\]](#)[SQL] Incorrect result type inferred for some decimal expressions
- [\[SPARK-21344\]](#)[SQL] BinaryType comparison does signed byte array comparison
- [\[SPARK-21219\]](#)[CORE] Task retry occurs on same executor due to race co...
- [\[SPARK-21369\]](#)[CORE] Don't use Scala Tuple2 in common/network-\*
- [\[SPARK-21272\]](#) SortMergeJoin LeftAnti does not update numOutputRows
- [\[SPARK-21342\]](#) Fix DownloadCallback to work well with RetryingBlockFetcher.
- [\[SPARK-21083\]](#)[SQL] Store zero size and row count when analyzing empty table
- [\[SPARK-21343\]](#) Refine the document for spark.reducer.maxReqSizeShuffleToMem.
- [\[SPARK-20342\]](#)[CORE] Update task accumulators before sending task end event.
- [\[SPARK-21228\]](#)[SQL] InSet incorrect handling of structs
- [\[SPARK-21312\]](#)[SQL] correct offsetInBytes in UnsafeRow.writeToStream
- [\[SPARK-21300\]](#)[SQL] ExternalMapToCatalyst should null-check map key prior to converting to internal value.
- [\[SPARK-20256\]](#)[SQL] SessionState should be created more lazily
- [\[SPARK-21170\]](#)[CORE] Utils.tryWithSafeFinallyAndFailureCallbacks throws IllegalArgumentException: Self-suppression not permitted
- [\[SPARK-21936\]](#)[SQL][2.2] backward compatibility test framework for HiveExternalCatalog
- [\[SPARK-22306\]](#)[SQL][2.2] alter table schema should not erase the bucketing metadata at hive side
- [\[SPARK-21617\]](#)[SQL] Store correct table metadata when altering schema in Hive metastore.
- [\[SPARK-19611\]](#)[SQL][FOLLOWUP] set dataSchema correctly in HiveMetastoreCatalog.convertToLogicalRelation
- [\[SPARK-23660\]](#) Fix exception in yarn cluster mode when application ended fast
- [\[SPARK-21834\]](#) Incorrect executor request in case of dynamic allocation
- [\[SPARK-22252\]](#)[SQL][2.2] FileFormatWriter should respect the input query schema
- [\[SPARK-23438\]](#)[DSTREAMS] Fix DStreams data loss with WAL when driver crashes
- [\[SPARK-17788\]](#)[\[SPARK-21033\]](#)[SQL] fix the potential OOM in UnsafeExternalSorter and ShuffleExternalSorter
- [\[SPARK-21319\]](#)[SQL] Fix memory leak in sorter
- [\[SPARK-21551\]](#)[PYTHON] Increase timeout for PythonRDD.serveIterator
- [\[SPARK-25164\]](#)[SQL] Avoid rebuilding column and path list for each column in parquet reader
- [\[SPARK-23207\]](#)[\[SPARK-22905\]](#)[\[SPARK-24564\]](#)[\[SPARK-25114\]](#)[SQL][BACKPORT-2.2] Shuffle+Repartition on a DataFrame could lead to incorrect answers
- [\[SPARK-25081\]](#)[CORE] Nested spill in ShuffleExternalSorter should not access released memory page (branch-2.2)

## Issues Fixed in CDS 2.2 Release 2

The following list includes issues fixed in CDS 2.2 Release 2. Test-only changes are omitted.

- [\[SPARK-22850\]](#)[CORE] Ensure queued events are delivered to all event queues.
- [\[SPARK-14228\]](#)[CORE][YARN] Lost executor of RPC disassociated, and occurs exception: Could not find CoarseGrainedScheduler or it has been stopped
- [\[SPARK-21321\]](#)[SPARK CORE] Spark very verbose on shutdown
- Add secure link to avoid redirect when SSL is enabled.
- [\[SPARK-22162\]](#)[BRANCH-2.2] Executors and the driver should use consistent JobIDs in the RDD commit protocol
- [\[SPARK-21188\]](#)[CORE] releaseAllLocksForTask should synchronize the whole method
- [\[SPARK-21549\]](#)[CORE] Respect OutputFormats with no/invalid output directory provided
- [\[SPARK-21549\]](#)[CORE] Respect OutputFormats with no output directory provided
- [\[SPARK-21907\]](#)[CORE][BACKPORT 2.2] oom during spill
- Correct database location for namenode HA.
- [\[SPARK-22341\]](#)[YARN] Impersonate correct user when preparing resources.

- [\[SPARK-22290\]](#)[CORE] Avoid creating Hive delegation tokens when not necessary.
- [\[SPARK-21376\]](#)[YARN] Fix yarn client token expire issue when cleaning the staging files in long running scenario
- [\[SPARK-21254\]](#)[WEBUI] History UI performance fixes
- [\[SPARK-21890\]](#) Credentials not being passed to add the tokens
- Respect configured kerberos name mapping in SHS.
- Not able to read hive table from Cloudera version of Spark 2.2
- [\[SPARK-22083\]](#)[CORE] Release locks in MemoryStore.evictBlocksToFreeSpace
- [\[SPARK-18838\]](#)[HOTFIX][YARN] Check internal context state before stopping it.
- [\[SPARK-18838\]](#)[CORE] Add separate listener queues to LiveListenerBus.
- [\[SPARK-21928\]](#)[CORE] Set classloader on SerializerManager's private kryo
- [\[SPARK-21523\]](#)[ML] update breeze to 0.13.2 for an emergency bugfix in strong wolfe line search
- Relax upper C5 compatibility limit.
- Use correct FileSystem when getting qualified paths.
- [\[SPARK-13669\]](#)[SPARK-20898][CORE] Improve the blacklist mechanism to handle external shuffle service unavailable situation
- Preserve original permissions of SHS config file.
- Add Spark config for choosing locality of the YARN AM.
- [\[SPARK-21522\]](#)[CORE] Fix flakiness in LauncherServerSuite.
- [\[SPARK-20904\]](#)[CORE] Don't report task failures to driver during shutdown.
- Write encryption configs to config file.

### Issues Fixed in CDS 2.2 Release 1

The following list includes issues fixed in CDS 2.2 Release 1. Test-only changes are omitted.

- [\[SPARK-18992\]](#)[SQL] Move spark.sql.hive.thriftServer.singleSession to SQLConf
- [\[SPARK-10364\]](#)[SQL] Support Parquet logical type TIMESTAMP\_MILLIS
- [\[SPARK-10643\]](#)[CORE] Make spark-submit download remote files to local in client mode
- [\[SPARK-10748\]](#)[MESOS] Log error instead of crashing Spark Mesos dispatcher when a job is misconfigured
- [\[SPARK-10849\]](#)[SQL] Adds option to the JDBC data source write for user to specify database column type for the create table
- [\[SPARK-11569\]](#)[ML] Fix StringIndexer to handle null value properly
- [\[SPARK-11968\]](#)[MLLIB] Optimize MLLIB ALS recommendForAll
- [\[SPARK-12334\]](#)[SQL][PYSPARK] Support read from multiple input paths for orc file in DataFrameReader.orc
- [\[SPARK-12552\]](#)[CORE] Correctly count the driver resource when recovering from failure for Master
- [\[SPARK-12757\]](#)[CORE] lower "block locks were not released" log to info level
- [\[SPARK-12837\]](#)[CORE] Do not send the name of internal accumulator to executor side
- [\[SPARK-12868\]](#)[SQL] Allow adding jars from hdfs
- [\[SPARK-13330\]](#)[PYSPARK] PYTHONHASHSEED is not propagated to python worker
- [\[SPARK-13369\]](#) Add config for number of consecutive fetch failures
- [\[SPARK-13446\]](#)[SQL] Support reading data from Hive 2.0.1 metastore
- [\[SPARK-13450\]](#) Introduce ExternalAppendOnlyUnsafeRowArray. Change CartesianProductExec, SortMergeJoin, WindowExec to use it
- [\[SPARK-13568\]](#)[ML] Create feature transformer to impute missing values
- [\[SPARK-13721\]](#)[SQL] Support outer generators in DataFrame API
- [\[SPARK-13747\]](#)[CORE] Add ThreadUtils.awaitReady and disallow Await.ready
- [\[SPARK-13931\]](#) Stage can hang if an executor fails while speculated tasks are running
- [\[SPARK-14049\]](#)[CORE] Add functionality in spark history sever API to query applications by end time
- [\[SPARK-14272\]](#)[ML] Add Loglikelihood in GaussianMixtureSummary
- [\[SPARK-14352\]](#)[SQL] approxQuantile should support multi columns
- [\[SPARK-14471\]](#)[SQL] Aliases in SELECT could be used in GROUP BY
- [\[SPARK-14489\]](#)[ML][PYSPARK] ALS unknown user/item prediction strategy



- [\[SPARK-14503\]](#)[ML] spark.ml API for FPGrowth
- [\[SPARK-14536\]](#)[SQL] fix to handle null value in array type column for postgres.
- [\[SPARK-14709\]](#)[ML] spark.ml API for linear SVM
- [\[SPARK-14772\]](#)[PYTHON][ML] Fixed Params.copy method to match Scala implementation
- [\[SPARK-14804\]](#)[SPARK][GRAPHX] Fix checkpointing of VertexRDD/EdgeRDD
- [\[SPARK-14958\]](#)[CORE] Failed task not handled when there's error deserializing failure reason
- [\[SPARK-14975\]](#)[ML] Fixed GBClassifier to predict probability per training instance and fixed interfaces
- [\[SPARK-15040\]](#)[ML][PYSPARK] Add Imputer to PySpark
- [\[SPARK-15288\]](#)[MESOS] Mesos dispatcher should handle gracefully when any thread gets UncaughtException
- [\[SPARK-15354\]](#)[CORE] Topology aware block replication strategies
- [\[SPARK-15355\]](#)[CORE] Proactive block replication
- [\[SPARK-15463\]](#)[SQL] Add an API to load DataFrame from Dataset[String] storing CSV
- [\[SPARK-15555\]](#)[MESOS] Driver with --supervise option cannot be killed in Mesos mode
- [\[SPARK-15615\]](#)[SQL] Add an API to load DataFrame from Dataset[String] storing JSON
- [\[SPARK-16101\]](#)[HOTFIX] Fix the build with Scala 2.10 by explicit typed argument
- [\[SPARK-16101\]](#)[SQL] Refactoring CSV write path to be consistent with JSON data source
- [\[SPARK-16122\]](#)[CORE] Add rest api for job environment
- [\[SPARK-16213\]](#)[SQL] Reduce runtime overhead of a program that creates an primitive array in DataFrame
- [\[SPARK-16440\]](#)[MLLIB] Ensure broadcasted variables are destroyed even in case of exception
- [\[SPARK-16473\]](#)[MLLIB] Fix BisectingKMeans Algorithm failing in edge case
- [\[SPARK-16475\]](#)[SQL] broadcast hint for SQL queries - disallow space as the delimiter
- [\[SPARK-16548\]](#)[SQL] Inconsistent error handling in JSON parsing SQL functions
- [\[SPARK-16554\]](#)[CORE] Automatically Kill Executors and Nodes when they are Blacklisted
- [\[SPARK-16599\]](#)[CORE] java.util.NoSuchElementException: None.get at at org.apache.spark.storage.BlockInfoManager.releaseAllLocksForTask
- [\[SPARK-16609\]](#) Add to\_date/to\_timestamp with format functions
- [\[SPARK-16654\]](#)[CORE] Add UI coverage for Application Level Blacklisting
- [\[SPARK-16792\]](#)[SQL] Dataset containing a Case Class with a List type causes a CompileException (converting sequence to list)
- [\[SPARK-16845\]](#)[SQL] `GeneratedClass\$SpecificOrdering` grows beyond 64 KB
- [\[SPARK-16848\]](#)[SQL] Check schema validation for user-specified schema in jdbc and table APIs
- [\[SPARK-16929\]](#) Improve performance when check speculatable tasks.
- [\[SPARK-17019\]](#)[CORE] Expose on-heap and off-heap memory usage in various places
- [\[SPARK-17075\]](#)[SQL] implemented filter estimation
- [\[SPARK-17076\]](#)[SQL] Cardinality estimation for join based on basic column statistics
- [\[SPARK-17077\]](#)[SQL] Cardinality estimation for project operator
- [\[SPARK-17078\]](#)[SQL] Show stats when explain
- [\[SPARK-17080\]](#)[SQL] join reorder
- [\[SPARK-17137\]](#)[ML][WIP] Compress logistic regression coefficients
- [\[SPARK-17161\]](#)[PYSPARK][ML] Add PySpark-ML JavaWrapper convenience function to create Py4J JavaArrays
- [\[SPARK-17204\]](#)[CORE] Fix replicated off heap storage
- [\[SPARK-17237\]](#)[SQL] Remove backticks in a pivot result schema
- [\[SPARK-17424\]](#) Fix unsound substitution bug in ScalaReflection.
- [\[SPARK-17455\]](#)[MLLIB] Improve PAVA implementation in IsotonicRegression
- [\[SPARK-17471\]](#)[ML] Add compressed method to ML matrices
- [\[SPARK-17495\]](#)[SQL] Support date, timestamp, decimal and interval types in Hive hash
- [\[SPARK-17498\]](#)[ML] StringIndexer enhancement for handling unseen labels
- [\[SPARK-17568\]](#)[CORE][DEPLOY] Add spark-submit option to override ivy settings used to resolve packages/artifacts
- [\[SPARK-17629\]](#)[ML] methods to return synonyms directly

- [\[SPARK-17645\]](#)[MLLIB][ML] add feature selector method based on: False Discovery Rate (FDR) and Family wise error rate (FWE)
- [\[SPARK-17647\]](#)[SQL] Fix backslash escaping in 'LIKE' patterns.
- [\[SPARK-17663\]](#)[CORE] SchedulableBuilder should handle invalid data access via scheduler.allocation.file
- [\[SPARK-17685\]](#)[SQL] Make SortMergeJoinExec's currentVars is null when calling createJoinKey
- [\[SPARK-17714\]](#)[CORE][TEST-MAVEN][TEST-HADOOP2.6] Avoid using ExecutorClassLoader to load Netty generated classes
- [\[SPARK-17724\]](#)[STREAMING][WEBUI] Unevaluated new lines in tooltip in DAG Visualization of a job
- [\[SPARK-17747\]](#)[ML] WeightCol support non-double numeric datatypes
- [\[SPARK-17755\]](#)[CORE] Use workerRef to send RegisterWorkerResponse to avoid the race condition
- [\[SPARK-17791\]](#)[SQL] Join reordering using star schema detection
- [\[SPARK-17847\]](#)[ML] Reduce shuffled data size of GaussianMixture & copy the implementation from mllib to ml
- [\[SPARK-17874\]](#)[CORE] Add SSL port configuration.
- [\[SPARK-17912\]](#) [SQL] Refactor code generation to get data for ColumnVector/ColumnarBatch
- [\[SPARK-17913\]](#)[SQL] compare atomic and string type column may return confusing result
- [\[SPARK-17914\]](#)[SQL] Fix parsing of timestamp strings with nanoseconds
- [\[SPARK-17931\]](#) Eliminate unnecessary task (de) serialization
- [\[SPARK-17979\]](#)[SPARK-14453] Remove deprecated SPARK\_YARN\_USER\_ENV and SPARK\_JAVA\_OPTS
- [\[SPARK-18020\]](#)[STREAMING][KINESIS] Checkpoint SHARD\_END to finish reading closed shards
- [\[SPARK-18036\]](#)[ML][MLLIB] Fixing decision trees handling edge cases
- [\[SPARK-18055\]](#)[SQL] Use correct mirror in ExpressionEncoder
- [\[SPARK-18080\]](#)[ML][PYTHON] Python API & Examples for Locality Sensitive Hashing
- [\[SPARK-18112\]](#)[SQL] Support reading data from Hive 2.1 metastore
- [\[SPARK-18113\]](#) Use ask to replace askWithRetry in canCommit and make receiver idempotent.
- [\[SPARK-18120\]](#)[SPARK-19557][SQL] Call QueryExecutionListener callback methods for DataFrameWriter methods
- [\[SPARK-18123\]](#)[SQL] Use db column names instead of RDD column ones during JDBC Writing
- [\[SPARK-18127\]](#) Add hooks and extension points to Spark
- [\[SPARK-18194\]](#)[ML] Log instrumentation in OneVsRest, CrossValidator, TrainValidationSplit
- [\[SPARK-18206\]](#)[ML] Add instrumentation for MLP,NB,LDA,AFT,GLM,Isotonic,LiR
- [\[SPARK-18218\]](#)[ML][MLLIB] Reduce shuffled data size of BlockMatrix multiplication and solve potential OOM and low parallelism usage problem By split middle dimension in matrix multiplication
- [\[SPARK-18243\]](#)[SQL] Port Hive writing to use FileFormat interface
- [\[SPARK-18278\]](#)[SCHEDULER] Documentation to point to Kubernetes cluster scheduler
- [\[SPARK-18285\]](#)[SPARKR] SparkR approxQuantile supports input multiple columns
- [\[SPARK-18335\]](#)[SPARKR] createDataFrame to support numPartitions parameter
- [\[SPARK-18352\]](#)[SQL] Support parsing multiline json files
- [\[SPARK-18389\]](#)[SQL] Disallow cyclic view reference
- [\[SPARK-18406\]](#)[CORE] Race between end-of-task and completion iterator read lock release
- [\[SPARK-18495\]](#)[UI] Document meaning of green dot in DAG visualization
- [\[SPARK-18537\]](#)[WEB UI] Add a REST api to serve spark streaming information
- [\[SPARK-18541\]](#)[PYTHON] Add metadata parameter to pyspark.sql.Column.alias()
- [\[SPARK-18567\]](#)[SQL] Simplify CreateDataSourceTableAsSelectCommand
- [\[SPARK-18579\]](#)[SQL] Use ignoreLeadingWhiteSpace and ignoreTrailingWhiteSpace options in CSV writing
- [\[SPARK-18589\]](#)[SQL] Fix Python UDF accessing attributes from both side of join
- [\[SPARK-18601\]](#)[SQL] Simplify Create/Get complex expression pairs in optimizer
- [\[SPARK-18609\]](#)[SPARK-18841][SQL] Fix redundant Alias removal in the optimizer
- [\[SPARK-18613\]](#)[ML] make spark.mllib LDA dependencies in spark.ml LDA private
- [\[SPARK-18682\]](#)[SS] Batch Source for Kafka
- [\[SPARK-18687\]](#)[PYSARK][SQL] Backward compatibility - creating a Dataframe on a new SQLContext object fails with a Derby error

- [\[SPARK-18698\]](#)[ML] Adding public constructor that takes uid for IndexToString
- [\[SPARK-18699\]](#)[SQL] Put malformed tokens into a new field when parsing CSV data
- [\[SPARK-18726\]](#)[SQL] resolveRelation for FileFormat DataSource don't need to listFiles twice
- [\[SPARK-18750\]](#)[YARN] Avoid using "mapValues" when allocating containers.
- [\[SPARK-18772\]](#)[SQL] Avoid unnecessary conversion try for special floats in JSON
- [\[SPARK-18788\]](#)[SPARKR] Add API for getNumPartitions
- [\[SPARK-18800\]](#)[SQL] Correct the assert in UnsafeKVExternalSorter which ensures array size
- [\[SPARK-18801\]](#)[SQL] Support resolve a nested view
- [\[SPARK-18808\]](#)[ML][MLLIB] ml.KMeansModel.transform is very inefficient
- [\[SPARK-18817\]](#)[SPARKR][SQL] change derby log output to temp dir
- [\[SPARK-18821\]](#)[SPARKR] Bisecting k-means wrapper in SparkR
- [\[SPARK-18823\]](#)[SPARKR] add support for assigning to column
- [\[SPARK-18828\]](#)[SPARKR] Refactor scripts for R
- [\[SPARK-18837\]](#)[WEBUI] Very long stage descriptions do not wrap in the UI
- [\[SPARK-18847\]](#)[GRAPHX] PageRank gives incorrect results for graphs with sinks
- [\[SPARK-18857\]](#)[SQL] Don't use `Iterator.duplicate` for `incrementalCollect` in Thrift Server
- [\[SPARK-18862\]](#)[SPARKR][ML] Split SparkR mllib.R into multiple files
- [\[SPARK-18863\]](#)[SQL] Output non-aggregate expressions without GROUP BY in a subquery does not yield an error
- [\[SPARK-18871\]](#)[SQL] New test cases for IN/NOT IN subquery
- [\[SPARK-18874\]](#)[SQL] Fix 2.10 build after moving the subquery rules to optimization
- [\[SPARK-18877\]](#)[SQL] `CSVInferSchema.inferField` on DecimalType should find a common type with `typeSoFar`
- [\[SPARK-18885\]](#)[SQL] unify CREATE TABLE syntax for data source and hive serde tables
- [\[SPARK-18901\]](#)[ML] Require in LR LogisticAggregator is redundant
- [\[SPARK-18905\]](#)[STREAMING] Fix the issue of removing a failed jobset from JobScheduler.jobSets
- [\[SPARK-18909\]](#)[SQL] The error messages in `ExpressionEncoder.toRow/fromRow` are too verbose
- [\[SPARK-18911\]](#)[SQL] Define CatalogStatistics to interact with metastore and convert it to Statistics in relations
- [\[SPARK-18917\]](#)[SQL] Remove schema check in appending data
- [\[SPARK-18929\]](#)[ML] Add Tweedie distribution in GLM
- [\[SPARK-18932\]](#)[SQL] Support partial aggregation for collect\_set/collect\_list
- [\[SPARK-18936\]](#)[SQL] Infrastructure for session local timezone support.
- [\[SPARK-18937\]](#)[SQL] Timezone support in CSV/JSON parsing
- [\[SPARK-18939\]](#)[SQL] Timezone support in partition values.
- [\[SPARK-18943\]](#)[SQL] Avoid per-record type dispatch in CSV when reading
- [\[SPARK-18952\]](#) Regex strings not properly escaped in codegen for aggregations
- [\[SPARK-18958\]](#)[SPARKR] R API toJSON on DataFrame
- [\[SPARK-18960\]](#)[SQL][SS] Avoid double reading file which is being copied.
- [\[SPARK-18961\]](#)[SQL] Support `SHOW TABLE EXTENDED ... PARTITION` statement
- [\[SPARK-18963\]](#) o.a.s.unsafe.types.UTF8StringSuite.writeToOutputStreamIntArray test
- [\[SPARK-18966\]](#)[SQL] NOT IN subquery with correlated expressions may return incorrect result
- [\[SPARK-18967\]](#)[SCHEDULER] compute locality levels even if delay = 0
- [\[SPARK-18969\]](#)[SQL] Support grouping by nondeterministic expressions
- [\[SPARK-18971\]](#)[CORE] Upgrade Netty to 4.0.43.Final
- [\[SPARK-18972\]](#)[CORE] Fix the netty thread names for RPC
- [\[SPARK-18973\]](#)[SQL] Remove SortPartitions and RedistributeData
- [\[SPARK-18975\]](#)[CORE] Add an API to remove SparkListener
- [\[SPARK-18980\]](#)[SQL] implement Aggregator with TypedImperativeAggregate
- [\[SPARK-18985\]](#)[SS] Add missing @InterfaceStability.Evolving for Structured Streaming APIs
- [\[SPARK-18986\]](#)[CORE] ExternalAppendOnlyMap shouldn't fail when forced to spill before calling its iterator
- [\[SPARK-18989\]](#)[SQL] DESC TABLE should not fail with format class not found
- [\[SPARK-18991\]](#)[CORE] Change ContextCleaner.referenceBuffer to use ConcurrentHashMap to make it faster

- [\[SPARK-18997\]](#)[CORE] Recommended upgrade libthrift to 0.9.3
- [\[SPARK-18998\]](#)[SQL] Add a cbo conf to switch between default statistics and estimated statistics
- [\[SPARK-19004\]](#)[SQL] Fix `JDBCWriteSuite.testH2Dialect` by removing `getCatalystType`
- [\[SPARK-19008\]](#)[SQL] Improve performance of Dataset.map by eliminating boxing/unboxing
- [\[SPARK-19010\]](#)[CORE] Include Kryo exception in case of overflow
- [\[SPARK-19012\]](#)[SQL] Fix `createViewCommand` to throw AnalysisException instead of ParseException
- [\[SPARK-19017\]](#)[SQL] NOT IN subquery with more than one column may return incorrect results
- [\[SPARK-19019\]](#) [PYTHON] Fix hijacked `collections.namedtuple` and port cloudpickle changes for PySpark to work with Python 3.6.0
- [\[SPARK-19020\]](#)[SQL] Cardinality estimation of aggregate operator
- [\[SPARK-19021\]](#)[YARN] Generalize HDFSCredentialProvider to support non HDFS security filesystems
- [\[SPARK-19024\]](#)[SQL] Implement new approach to write a permanent view
- [\[SPARK-19025\]](#)[SQL] Remove SQL builder for operators
- [\[SPARK-19026\]](#) SPARK\_LOCAL\_DIRS(multiple directories on different disks) cannot be deleted
- [\[SPARK-19028\]](#)[SQL] Fixed non-thread-safe functions used in SessionCatalog
- [\[SPARK-19029\]](#)[SQL] Remove databaseName from SimpleCatalogRelation
- [\[SPARK-19033\]](#)[CORE] Add admin acls for history server
- [\[SPARK-19038\]](#)[YARN] Avoid overwriting keytab configuration in yarn-client
- [\[SPARK-19041\]](#)[SS] Fix code snippet compilation issues in Structured Streaming Programming Guide
- [\[SPARK-19042\]](#) spark executor can't download the jars when uber jar's http url contains any query strings
- [\[SPARK-19048\]](#)[SQL] Delete Partition Location when Dropping Managed Partitioned Tables in InMemoryCatalog
- [\[SPARK-19054\]](#)[ML] Eliminate extra pass in NB
- [\[SPARK-19055\]](#)[SQL][PYSPARK] Fix SparkSession initialization when SparkContext is stopped
- [\[SPARK-19058\]](#)[SQL] fix partition related behaviors with DataFrameWriter.saveAsTable
- [\[SPARK-19059\]](#)[SQL] Unable to retrieve data from parquet table whose name startswith underscore
- [\[SPARK-19060\]](#)[SQL] remove the supportsPartial flag in AggregateFunction
- [\[SPARK-19062\]](#) Utils.writeByteBuffer bug fix
- [\[SPARK-19064\]](#)[PYSPARK] Fix pip installing of sub components
- [\[SPARK-19065\]](#)[SQL] Don't inherit expression id in dropDuplicates
- [\[SPARK-19066\]](#)[SPARKR] SparkR LDA doesn't set optimizer correctly
- [\[SPARK-19067\]](#)[SS] Processing-time-based timeout in MapGroupsWithState
- [\[SPARK-19069\]](#)[CORE] Expose task 'status' and 'duration' in spark history server REST API.
- [\[SPARK-19070\]](#) Clean-up dataset actions
- [\[SPARK-19072\]](#)[SQL] codegen of Literal should not output boxed value
- [\[SPARK-19073\]](#) LauncherState should be only set to SUBMITTED after the application is submitted
- [\[SPARK-19080\]](#)[SQL] simplify data source analysis
- [\[SPARK-19082\]](#)[SQL] Make ignoreCorruptFiles work for Parquet
- [\[SPARK-19083\]](#) sbin/start-history-server.sh script use of \$@ without quotes
- [\[SPARK-19084\]](#)[SQL] Ensure context class loader is set when initializing Hive.
- [\[SPARK-19085\]](#)[SQL] cleanup OutputWriterFactory and OutputWriter
- [\[SPARK-19088\]](#)[SQL] Optimize sequence type deserialization codegen
- [\[SPARK-19089\]](#)[SQL] Add support for nested sequences
- [\[SPARK-19092\]](#)[SQL] Save() API of DataFrameWriter should not scan all the saved files
- [\[SPARK-19093\]](#)[SQL] Cached tables are not used in SubqueryExpression
- [\[SPARK-19104\]](#)[SQL] Lambda variables in ExternalMapToCatalyst should be global
- [\[SPARK-19107\]](#)[SQL] support creating hive table with DataFrameWriter and Catalog
- [\[SPARK-19110\]](#)[ML][MLLIB] DistributedLDAModel returns different logPrior for original and loaded model
- [\[SPARK-19115\]](#)[SQL] Supporting Create Table Like Location
- [\[SPARK-19118\]](#)[SQL] Percentile support for frequency distribution table
- [\[SPARK-19120\]](#) Refresh Metadata Cache After Loading Hive Tables

- [\[SPARK-19129\]](#)[SQL] SessionCatalog: Disallow empty part col values in partition spec
- [\[SPARK-19130\]](#)[SPARKR] Support setting literal value as column implicitly
- [\[SPARK-19132\]](#)[SQL] Add test cases for row size estimation and aggregate estimation
- [\[SPARK-19133\]](#)[SPARKR][ML] fix glm for Gamma, clarify glm family supported
- [\[SPARK-19134\]](#)[EXAMPLE] Fix several sql, mllib and status api examples not working
- [\[SPARK-19137\]](#)[SQL] Fix `withSQLConf` to reset `OptionalConfigEntry` correctly
- [\[SPARK-19139\]](#)[CORE] New auth mechanism for transport library.
- [\[SPARK-19140\]](#)[SS] Allow update mode for non-aggregation streaming queries
- [\[SPARK-19142\]](#)[SPARKR] spark.kmeans should take seed, initSteps, and tol as parameters
- [\[SPARK-19146\]](#)[CORE] Drop more elements when stageData.taskData.size > retainedTasks
- [\[SPARK-19148\]](#)[SQL] do not expose the external table concept in Catalog
- [\[SPARK-19149\]](#)[SQL] Unify two sets of statistics in LogicalPlan
- [\[SPARK-19151\]](#)[SQL] DataFrameWriter.saveAsTable support hive overwrite
- [\[SPARK-19152\]](#)[SQL] DataFrameWriter.saveAsTable support hive append
- [\[SPARK-19153\]](#)[SQL] DataFrameWriter.saveAsTable work with create partitioned table
- [\[SPARK-19155\]](#)[ML] MLib GeneralizedLinearRegression family and link should case insensitive
- [\[SPARK-19157\]](#)[SQL] should be able to change spark.sql.runSQLOnFiles at runtime
- [\[SPARK-19158\]](#)[SPARKR][EXAMPLES] Fix ml.R example fails due to lack of e1071 package.
- [\[SPARK-19160\]](#)[PYTHON][SQL] Add udf decorator
- [\[SPARK-19161\]](#)[PYTHON][SQL] Improving UDF Docstrings
- [\[SPARK-19162\]](#)[PYTHON][SQL] UserDefinedFunction should validate that func is callable
- [\[SPARK-19163\]](#)[PYTHON][SQL] Delay \_judf initialization to the \_\_call\_\_
- [\[SPARK-19164\]](#)[PYTHON][SQL] Remove unused UserDefinedFunction.\_broadcast
- [\[SPARK-19168\]](#)[STRUCTURED STREAMING] StateStore should be aborted upon error
- [\[SPARK-19178\]](#)[SQL] convert string of large numbers to int should return null
- [\[SPARK-19179\]](#)[YARN] Change spark.yarn.access.namenodes config and update docs
- [\[SPARK-19180\]](#) [SQL] the offset of short should be 2 in OffHeapColumn
- [\[SPARK-19182\]](#)[DSTREAM] Optimize the lock in StreamingJobProgressListener to not block UI when generating Streaming jobs
- [\[SPARK-19183\]](#)[SQL] Add deleteWithJob hook to internal commit protocol API
- [\[SPARK-19185\]](#)[DSTREAM] Make Kafka consumer cache configurable
- [\[SPARK-19207\]](#)[SQL] LocalSparkSession should use Slf4JLoggerFactory.INSTANCE
- [\[SPARK-19211\]](#)[SQL] Explicitly prevent Insert into View or Create View As Insert
- [\[SPARK-19218\]](#)[SQL] Fix SET command to show a result correctly and in a sorted order
- [\[SPARK-19219\]](#)[SQL] Fix Parquet log output defaults
- [\[SPARK-19220\]](#)[UI] Make redirection to HTTPS apply to all URIs.
- [\[SPARK-19221\]](#)[PROJECT INFRA][R] Add winutils binaries to the path in AppVeyor tests for Hadoop libraries to call native codes properly
- [\[SPARK-19223\]](#)[SQL][PYSPARK] Fix InputFileBlockHolder for datasources which are based on HadoopRDD or NewHadoopRDD
- [\[SPARK-19227\]](#)[SPARK-19251] remove unused imports and outdated comments
- [\[SPARK-19229\]](#)[SQL] Disallow Creating Hive Source Tables when Hive Support is Not Enabled
- [\[SPARK-19231\]](#)[SPARKR] add error handling for download and untar for Spark release
- [\[SPARK-19232\]](#)[SPARKR] Update Spark distribution download cache location on Windows
- [\[SPARK-19236\]](#)[SQL][BACKPORT-2.2] Added createOrReplaceGlobalTempView method
- [\[SPARK-19237\]](#)[SPARKR][CORE] On Windows spark-submit should handle when java is not installed
- [\[SPARK-19239\]](#)[PYSPARK] Check parameters whether equals None when specify the column in jdbc API
- [\[SPARK-19244\]](#)[CORE] Sort MemoryConsumers according to their memory usage when spilling
- [\[SPARK-19246\]](#)[SQL] CatalogTable's partitionSchema order and exist check
- [\[SPARK-19247\]](#)[ML] Save large word2vec models

- [\[SPARK-19254\]](#)[SQL] Support Seq, Map, and Struct in functions.lit
- [\[SPARK-19257\]](#)[SQL] location for table/partition/database should be java.net.URI
- [\[SPARK-19260\]](#) Spaces or "%20" in path parameter are not correctly handled with...
- [\[SPARK-19261\]](#)[SQL] Alter add columns for Hive serde and some datasource tables
- [\[SPARK-19263\]](#) DAGScheduler should avoid sending conflicting task set.
- [\[SPARK-19265\]](#)[SQL] make table relation cache general and does not depend on hive
- [\[SPARK-19267\]](#)[SS] Fix a race condition when stopping StateStore
- [\[SPARK-19268\]](#)[SS] Disallow adaptive query execution for streaming queries
- [\[SPARK-19271\]](#)[SQL] Change non-cbo estimation of aggregate
- [\[SPARK-19272\]](#)[SQL] Remove the param `viewOriginalText` from `CatalogTable`
- [\[SPARK-19276\]](#)[CORE] Fetch Failure handling robust to user error handling
- [\[SPARK-19279\]](#)[SQL] Infer Schema for Hive Serde Tables and Block Creating a Hive Table With an Empty Schema
- [\[SPARK-19281\]](#)[PYTHON][ML] spark.ml Python API for FPGrowth
- [\[SPARK-19282\]](#)[ML][SPARKR] RandomForest Wrapper and GBT Wrapper return param "maxDepth" to R models
- [\[SPARK-19284\]](#)[SQL] append to partitioned datasource table should without custom partition location
- [\[SPARK-19290\]](#)[SQL] add a new extending interface in Analyzer for post-hoc resolution
- [\[SPARK-19291\]](#)[SPARKR][ML] spark.gaussianMixture supports output log-likelihood.
- [\[SPARK-19292\]](#)[SQL] filter with partition columns should be case-insensitive on Hive tables
- [\[SPARK-19295\]](#)[SQL] IsolatedClientLoader's downloadVersion should log the location of downloaded metastore client jars
- [\[SPARK-19296\]](#)[SQL] Deduplicate url and table in JdbcUtils
- [\[SPARK-19304\]](#)[STREAMING][KINESIS] fix kinesis slow checkpoint recovery
- [\[SPARK-19305\]](#)[SQL] partitioned table should always put partition columns at the end of table schema
- [\[SPARK-19306\]](#)[CORE] Fix inconsistent state in DiskBlockObject when exception occurred
- [\[SPARK-19307\]](#)[PYSPARK] Make sure user conf is propagated to SparkContext.
- [\[SPARK-19309\]](#)[SQL] disable common subexpression elimination for conditional expressions
- [\[SPARK-19311\]](#)[SQL] fix UDT hierarchy issue
- [\[SPARK-19313\]](#)[ML][MLLIB] GaussianMixture should limit the number of features
- [\[SPARK-19314\]](#)[SS][CATALYST] Do not allow sort before aggregation in Structured Streaming plan
- [\[SPARK-19318\]](#)[SQL] Fix to treat JDBC connection properties specified by the user in case-sensitive manner.
- [\[SPARK-19319\]](#)[SPARKR] SparkR Kmeans summary returns error when the cluster size doesn't equal to k
- [\[SPARK-19324\]](#)[SPARKR] Spark VJM stdout output is getting dropped in SparkR
- [\[SPARK-19329\]](#)[SQL] Reading from or writing to a datasource table with a non pre-existing location should succeed
- [\[SPARK-19330\]](#)[DSTREAMS] Also show tooltip for successful batches
- [\[SPARK-19333\]](#)[SPARKR] Add Apache License headers to R files
- [\[SPARK-19334\]](#)[SQL] Fix the code injection vulnerability related to Generator functions.
- [\[SPARK-19336\]](#)[ML][PYSPARK] LinearSVC Python API
- [\[SPARK-19338\]](#)[SQL] Add UDF names in explain
- [\[SPARK-19342\]](#)[SPARKR] bug fixed in collect method for collecting timestamp column
- [\[SPARK-19347\]](#) ReceiverSupervisorImpl can add block to ReceiverTracker multiple times because of askWithRetry.
- [\[SPARK-19348\]](#)[PYTHON] PySpark keyword\_only decorator is not thread-safe
- [\[SPARK-19350\]](#)[SQL] Cardinality estimation of Limit and Sample
- [\[SPARK-19359\]](#)[SQL] renaming partition should not leave useless directories
- [\[SPARK-19365\]](#)[CORE] Optimize RequestMessage serialization
- [\[SPARK-19372\]](#)[SQL] Fix throwing a Java exception at df.fliter() due to 64KB bytecode size limit
- [\[SPARK-19373\]](#)[MESOS] Base spark.scheduler.minRegisteredResourceRatio on registered cores rather than accepted cores
- [\[SPARK-19377\]](#)[WEBUI][CORE] Killed tasks should have the status as KILLED
- [\[SPARK-19378\]](#)[SS] Ensure continuity of stateOperator and eventTime metrics even if there is no new data in trigger

- [\[SPARK-19382\]](#)[ML] Test sparse vectors in LinearSVCSuite
- [\[SPARK-19384\]](#)[ML] forget unpersist input dataset in IsotonicRegression
- [\[SPARK-19385\]](#)[SQL] During canonicalization, `NOT(...(l, r))` should not expect such cases that l.hashCode > r.hashCode
- [\[SPARK-19387\]](#)[SPARKR] Tests do not run with SparkR source package in CRAN check
- [\[SPARK-19391\]](#)[SPARKR][ML] Tweedie GLM API for SparkR
- [\[SPARK-19395\]](#)[SPARKR] Convert coefficients in summary to matrix
- [\[SPARK-19397\]](#)[SQL] Make option names of LIBSVM and TEXT case insensitive
- [\[SPARK-19398\]](#) Change one misleading log in TaskSetManager.
- [\[SPARK-19399\]](#)[SPARKR] Add R coalesce API for DataFrame and Column
- [\[SPARK-19400\]](#)[ML] Allow GLM to handle intercept only model
- [\[SPARK-19403\]](#)[PYTHON][SQL] Correct pyspark.sql.column.\_\_all\_\_ list.
- [\[SPARK-19405\]](#)[STREAMING] Support for cross-account Kinesis reads via STS
- [\[SPARK-19406\]](#)[SQL] Fix function to\_json to respect user-provided options
- [\[SPARK-19407\]](#)[SS] defaultFS is used FileSystem.get instead of getting it from uri scheme
- [\[SPARK-19408\]](#)[SQL] filter estimation on two columns of same table
- [\[SPARK-19409\]](#)[SPARK-17213] Cleanup Parquet workarounds/hacks due to bugs of old Parquet versions
- [\[SPARK-19411\]](#)[SQL] Remove the metadata used to mark optional columns in merged Parquet schema for filter predicate pushdown
- [\[SPARK-19413\]](#)[SS] MapGroupsWithState for arbitrary stateful operations
- [\[SPARK-19421\]](#)[ML][PYSPARK] Remove numClasses and numFeatures methods in LinearSVC
- [\[SPARK-19425\]](#)[SQL] Make ExtractEquiJoinKeys support UDT columns
- [\[SPARK-19427\]](#)[PYTHON][SQL] Support data type string as a returnType argument of UDF
- [\[SPARK-19429\]](#)[PYTHON][SQL] Support slice arguments in Column.\_\_getitem\_\_
- [\[SPARK-19432\]](#)[CORE] Fix an unexpected failure when connecting timeout
- [\[SPARK-19435\]](#)[SQL] Type coercion between ArrayTypes
- [\[SPARK-19436\]](#)[SQL] Add missing tests for approxQuantile
- [\[SPARK-19437\]](#) Rectify spark executor id in HeartbeatReceiverSuite.
- [\[SPARK-19441\]](#)[SQL] Remove IN type coercion from PromoteStrings
- [\[SPARK-19446\]](#)[SQL] Remove unused findTightestCommonType in TypeCoercion
- [\[SPARK-19447\]](#) Make Range operator generate "recordsRead" metric
- [\[SPARK-19448\]](#)[SQL] optimize some duplication functions between HiveClientImpl and HiveUtils
- [\[SPARK-19450\]](#) Replace askWithRetry with askSync.
- [\[SPARK-19452\]](#)[SPARKR] Fix bug in the name assignment method
- [\[SPARK-19454\]](#)[PYTHON][SQL] DataFrame.replace improvements
- [\[SPARK-19456\]](#)[SPARKR] Add LinearSVC R API
- [\[SPARK-19459\]](#)[SQL] Add Hive datatype (char/varchar) to StructField metadata
- [\[SPARK-19460\]](#)[SPARKR] Update dataset used in R documentation, examples to reduce warning noise and confusions
- [\[SPARK-19463\]](#)[SQL] refresh cache after the InsertIntoHadoopFsRelationCommand
- [\[SPARK-19464\]](#)[CORE][YARN][TEST-HADOOP2.6] Remove support for Hadoop 2.5 and earlier
- [\[SPARK-19466\]](#)[CORE][SCHEDULER] Improve Fair Scheduler Logging
- [\[SPARK-19467\]](#)[ML][PYTHON] Remove cyclic imports from pyspark.ml.pipeline
- [\[SPARK-19472\]](#)[SQL] Parser should not mistake CASE WHEN(...) for a function call
- [\[SPARK-19481\]](#) [REPL] [MAVEN] Avoid to leak SparkContext in Signaling.cancelOnInterrupt
- [\[SPARK-19488\]](#)[SQL] fix csv infer schema when the field is Nan/Inf etc
- [\[SPARK-19495\]](#)[SQL] Make SQLConf slightly more extensible
- [\[SPARK-19496\]](#)[SQL] to\_date udf to return null when input date is invalid
- [\[SPARK-19497\]](#)[SS] Implement streaming deduplication
- [\[SPARK-19499\]](#)[SS] Add more notes in the comments of Sink.addBatch()
- [\[SPARK-19500\]](#) [SQL] Fix off-by-one bug in BytesToBytesMap

- [\[SPARK-19501\]](#)[YARN] Reduce the number of HDFS RPCs during YARN deployment
- [\[SPARK-19505\]](#)[PYTHON] AttributeError on Exception.message in Python3
- [\[SPARK-19506\]](#)[ML][PYTHON] Import warnings in pyspark.ml.util
- [\[SPARK-19508\]](#)[CORE] Improve error message when binding service fails
- [\[SPARK-19512\]](#)[SQL] codegen for compare structs fails
- [\[SPARK-19514\]](#) Making range interruptible.
- [\[SPARK-19517\]](#)[SS] KafkaSource fails to initialize partition offsets
- [\[SPARK-19518\]](#)[SQL] IGNORE NULLS in first / last in SQL
- [\[SPARK-19520\]](#)[STREAMING] Do not encrypt data written to the WAL.
- [\[SPARK-19525\]](#)[CORE] Add RDD checkpoint compression support
- [\[SPARK-19529\]](#) TransportClientFactory.createClient() shouldn't call awaitUninterruptibly()
- [\[SPARK-19533\]](#)[EXAMPLES] Convert Java tests to use lambdas, Java 8 features
- [\[SPARK-19535\]](#)[ML] RecommendForAllUsers RecommendForAllItems for ALS on Dataframe
- [\[SPARK-19537\]](#) Move pendingPartitions to ShuffleMapStage.
- [\[SPARK-19539\]](#)[SQL] Block duplicate temp table during creation
- [\[SPARK-19540\]](#)[SQL] Add ability to clone SparkSession wherein cloned session has an identical copy of the SessionState
- [\[SPARK-19542\]](#)[SS] Delete the temp checkpoint if a query is stopped without errors
- [\[SPARK-19543\]](#) from\_json fails when the input row is empty
- [\[SPARK-19544\]](#)[SQL] Improve error message when some column types are compatible and others are not in set operations
- [\[SPARK-19545\]](#)[YARN] Fix compile issue for Spark on Yarn when building against Hadoop 2.6.0~2.6.3
- [\[SPARK-19548\]](#)[SQL] Support Hive UDFs which return typed Lists/Maps
- [\[SPARK-19549\]](#) Allow providing reason for stage/job cancelling
- [\[SPARK-19554\]](#)[UI,YARN] Allow SHS URL to be used for tracking in YARN RM.
- [\[SPARK-19556\]](#)[CORE] Do not encrypt block manager data in memory.
- [\[SPARK-19560\]](#) Improve DAGScheduler tests.
- [\[SPARK-19561\]](#)[SQL] add int case handling for TimestampType
- [\[SPARK-19563\]](#)[SQL] avoid unnecessary sort in FileFormatWriter
- [\[SPARK-19564\]](#)[SPARK-19559][SS][KAFKA] KafkaOffsetReader's consumers should not be in the same group
- [\[SPARK-19567\]](#)[CORE][SCHEDULER] Support some Schedulable variables immutability and access
- [\[SPARK-19570\]](#)[PYSPARK] Allow to disable hive in pyspark shell
- [\[SPARK-19571\]](#)[R] Fix SparkR test break on Windows via AppVeyor
- [\[SPARK-19572\]](#)[SPARKR] Allow to disable hive in sparkR shell
- [\[SPARK-19573\]](#)[SQL] Make NaN/null handling consistent in approxQuantile
- [\[SPARK-19583\]](#)[SQL] CTAS for data source table with a created location should succeed
- [\[SPARK-19587\]](#)[SQL] bucket sorting columns should not be picked from partition columns
- [\[SPARK-19589\]](#)[SQL] Removal of SQLGEN files
- [\[SPARK-19590\]](#)[PYSPARK][ML] Update the document for QuantileDiscretizer in pyspark
- [\[SPARK-19594\]](#)[STRUCTURED STREAMING] StreamingQueryListener fails to handle QueryTerminatedEvent if more than one listeners exists
- [\[SPARK-19595\]](#)[SQL] Support json array in from\_json
- [\[SPARK-19597\]](#)[CORE] test case for task deserialization errors
- [\[SPARK-19598\]](#)[SQL] Remove the alias parameter in UnresolvedRelation
- [\[SPARK-19599\]](#)[SS] Clean up HDFSMetadataLog
- [\[SPARK-19601\]](#)[SQL] Fix CollapseRepartition rule to preserve shuffle-enabled Repartition
- [\[SPARK-19603\]](#)[SS] Fix StreamingQuery explain command
- [\[SPARK-19607\]](#) Finding QueryExecution that matches provided executionId
- [\[SPARK-19610\]](#)[SQL] Support parsing multiline CSV files
- [\[SPARK-19611\]](#)[SQL] Introduce configurable table schema inference



- [\[SPARK-19616\]](#)[SPARKR] weightCol and aggregationDepth should be improved for some SparkR APIs
- [\[SPARK-19617\]](#)[SS] Fix the race condition when starting and stopping a query quickly
- [\[SPARK-19618\]](#)[SQL] Inconsistency wrt max. buckets allowed from Dataframe API vs SQL
- [\[SPARK-19620\]](#)[SQL] Fix incorrect exchange coordinator id in the physical plan
- [\[SPARK-19622\]](#)[WEBUI] Fix a http error in a paged table when using a `Go` button to search.
- [\[SPARK-19626\]](#)[YARN] Using the correct config to set credentials update time
- [\[SPARK-19631\]](#)[CORE] OutputCommitCoordinator should not allow commits for already failed tasks
- [\[SPARK-19633\]](#)[SS] FileSource read from FileSink
- [\[SPARK-19635\]](#)[ML] DataFrame-based API for chi square test
- [\[SPARK-19636\]](#)[ML] Feature parity for correlation statistics in MLlib
- [\[SPARK-19637\]](#)[SQL] Add to\_json in FunctionRegistry
- [\[SPARK-19639\]](#)[SPARKR][EXAMPLE] Add spark.svmLinear example and update vignettes
- [\[SPARK-19641\]](#)[SQL] JSON schema inference in DROPMALFORMED mode produces incorrect schema for non-array/object JSONs
- [\[SPARK-19646\]](#)[CORE][STREAMING] binaryRecords replicates records in scala API
- [\[SPARK-19650\]](#) Commands should not trigger a Spark job
- [\[SPARK-19652\]](#)[UI] Do auth checks for REST API access.
- [\[SPARK-19654\]](#)[SPARKR][SS] Structured Streaming API for R
- [\[SPARK-19658\]](#)[SQL] Set NumPartitions of RepartitionByExpression In Parser
- [\[SPARK-19659\]](#) Fetch big blocks to disk when shuffle-read.
- [\[SPARK-19660\]](#)[CORE][SQL] Replace the configuration property names that are deprecated in the version of Hadoop 2.6
- [\[SPARK-19664\]](#)[SQL] put hive.metastore.warehouse.dir in hadoopconf to overwrite its original value
- [\[SPARK-19666\]](#)[SQL] Skip a property without getter in Java schema inference and allow empty bean in encoder creation
- [\[SPARK-19669\]](#)[SQL] Open up visibility for sharedState, sessionState, and a few other functions
- [\[SPARK-19673\]](#)[SQL] "ThriftServer default app name is changed wrong"
- [\[SPARK-19674\]](#)[SQL] Ignore driver accumulator updates don't belong to the execution when merging all accumulator updates
- [\[SPARK-19677\]](#)[SS] Committing a delta file atop an existing one should not fail on HDFS
- [\[SPARK-19678\]](#)[SQL] remove MetastoreRelation
- [\[SPARK-19679\]](#)[ML] Destroy broadcasted object without blocking
- [\[SPARK-19682\]](#)[SPARKR] Issue warning (or error) when subset method "[" takes vector index
- [\[SPARK-19688\]](#)[STREAMING] Not to read `spark.yarn.credentials.file` from checkpoint.
- [\[SPARK-19691\]](#)[SQL] Fix ClassCastException when calculating percentile of decimal column
- [\[SPARK-19693\]](#)[SQL] Make the SET mapreduce.job.reduces automatically converted to spark.sql.shuffle.partitions
- [\[SPARK-19694\]](#)[ML] Add missing 'setTopicDistributionCol' for LDAModel
- [\[SPARK-19695\]](#)[SQL] Throw an exception if a `columnNameOfCorruptRecord` field violates requirements in json formats
- [\[SPARK-19701\]](#)[SQL][PYTHON] Throws a correct exception for 'in' operator against column
- [\[SPARK-19702\]](#)[MESOS] Increase default refuse\_seconds timeout in the Mesos Spark Dispatcher
- [\[SPARK-19704\]](#)[ML] AFTSurvivalRegression should support numeric censorCol
- [\[SPARK-19706\]](#)[PYSPARK] add Column.contains in pyspark
- [\[SPARK-19707\]](#)[CORE] Improve the invalid path check for sc.addJar
- [\[SPARK-19709\]](#)[SQL] Read empty file with CSV data source
- [\[SPARK-19715\]](#)[STRUCTURED STREAMING] Option to Strip Paths in FileSource
- [\[SPARK-19716\]](#)[SQL] support by-name resolution for struct type elements in array
- [\[SPARK-19718\]](#)[SS] Handle more interrupt cases properly for Hadoop
- [\[SPARK-19719\]](#)[SS] Kafka writer for both structured streaming and batch queires
- [\[SPARK-19720\]](#)[CORE] Redact sensitive information from SparkSubmit console

- [\[SPARK-19721\]](#)[SS] Good error message for version mismatch in log files
- [\[SPARK-19723\]](#)[SQL] create datasource table with an non-existent location should work
- [\[SPARK-19727\]](#)[SQL] Fix for round function that modifies original column
- [\[SPARK-19733\]](#)[ML] Removed unnecessary castings and refactored checked casts in ALS.
- [\[SPARK-19734\]](#)[PYTHON][ML] Correct OneHotEncoder doc string to say dropLast
- [\[SPARK-19735\]](#)[SQL] Remove HOLD\_DDLTIME from Catalog APIs
- [\[SPARK-19736\]](#)[SQL] refreshByPath should clear all cached plans with the specified path
- [\[SPARK-19737\]](#)[SQL] New analysis rule for reporting unregistered functions without relying on relation resolution
- [\[SPARK-19739\]](#)[CORE] propagate S3 session token to cluser
- [\[SPARK-19740\]](#)[MESOS] Add support in Spark to pass arbitrary parameters into docker when running on mesos with docker containerizer
- [\[SPARK-19745\]](#)[ML] SVCAGgregator captures coefficients in its closure
- [\[SPARK-19746\]](#)[ML] Faster indexing for logistic aggregator
- [\[SPARK-19748\]](#)[SQL] refresh function has a wrong order to do cache invalidate and regenerate the inmemory var for InMemoryFileIndex with FileStatusCache
- [\[SPARK-19749\]](#)[SS] Name socket source with a meaningful name
- [\[SPARK-19751\]](#)[SQL] Throw an exception if bean class has one's own class in fields
- [\[SPARK-19757\]](#)[CORE] DriverEndpoint#makeOffers race against CoarseGrainedSchedulerBackend#killExecutors
- [\[SPARK-19758\]](#)[SQL] Resolving timezone aware expressions with time zone when resolving inline table
- [\[SPARK-19761\]](#)[SQL] create InMemoryFileIndex with an empty rootPaths when set PARALLEL\_PARTITION\_DISCOVERY\_THRESHOLD to zero failed
- [\[SPARK-19763\]](#)[SQL] qualified external datasource table location stored in catalog
- [\[SPARK-19765\]](#)[SPARK-18549][SQL] UNCACHE TABLE should un-cache all cached plans that refer to this table
- [\[SPARK-19766\]](#)[SQL] Constant alias columns in INNER JOIN should not be folded by FoldablePropagation rule
- [\[SPARK-19774\]](#) StreamExecution should call stop() on sources when a stream fails
- [\[SPARK-19775\]](#)[SQL] Remove an obsolete `partitionBy().insertInto()` test case
- [\[SPARK-19777\]](#) Scan runningTasksSet when check speculatable tasks in TaskSetManager.
- [\[SPARK-19779\]](#)[SS] Delete needless tmp file after restart structured streaming job
- [\[SPARK-19786\]](#)[SQL] Facilitate loop optimizations in a JIT compiler regarding range()
- [\[SPARK-19787\]](#)[ML] Changing the default parameter of regParam.
- [\[SPARK-19791\]](#)[ML] Add doc and example for fpgrowth
- [\[SPARK-19792\]](#)[WEBUI] In the Master Page,the column named "Memory per Node" ,I think it is not all right
- [\[SPARK-19793\]](#) Use clock.getTimeMillis when mark task as finished in TaskSetManager.
- [\[SPARK-19795\]](#)[SPARKR] add column functions to \_json, from\_json
- [\[SPARK-19796\]](#)[CORE] Fix serialization of long property values in TaskDescription
- [\[SPARK-19806\]](#)[ML][PYSPARK] PySpark GeneralizedLinearRegression supports tweedie distribution.
- [\[SPARK-19807\]](#)[WEB UI] Add reason for cancellation when a stage is killed using web UI
- [\[SPARK-19812\]](#) YARN shuffle service fails to relocate recovery DB acro...
- [\[SPARK-19813\]](#) maxFilesPerTrigger combo latestFirst may miss old files in combination with maxFileAge in FileStreamSource
- [\[SPARK-19817\]](#)[SQL] Make it clear that `timeZone` option is a general option in DataFrameReader/Writer.
- [\[SPARK-19818\]](#)[SPARKR] rbind should check for name consistency of input data frames
- [\[SPARK-19820\]](#)[CORE] Add interface to kill tasks w/ a reason
- [\[SPARK-19825\]](#)[R][ML] spark.ml R API for FPGrowth
- [\[SPARK-19828\]](#)[R] Support array type in from\_json in R
- [\[SPARK-19830\]](#)[SQL] Add parseTableSchema API to ParserInterface
- [\[SPARK-19831\]](#)[CORE] Reuse the existing cleanupThreadExecutor to clean up the directories of finished applications to avoid the block
- [\[SPARK-19832\]](#)[SQL] DynamicPartitionWriteTask get partitionPath should escape the partition name
- [\[SPARK-19841\]](#)[SS] watermarkPredicate should filter based on keys

- [\[SPARK-19843\]](#)[SQL] UTF8String => (int / long) conversion expensive for invalid inputs
- [\[SPARK-19846\]](#)[SQL] Add a flag to disable constraint propagation
- [\[SPARK-19849\]](#)[SQL] Support ArrayType in to\_json to produce JSON array
- [\[SPARK-19850\]](#)[SQL] Allow the use of aliases in SQL function calls
- [\[SPARK-19853\]](#)[SS] uppercase kafka topics fail when startingOffsets are SpecificOffsets
- [\[SPARK-19857\]](#)[YARN] Correctly calculate next credential update time.
- [\[SPARK-19858\]](#)[SS] Add output mode to flatMapGroupsWithState and disallow invalid cases
- [\[SPARK-19859\]](#)[SS] The new watermark should override the old one
- [\[SPARK-19861\]](#)[SS] watermark should not be a negative time.
- [\[SPARK-19865\]](#)[SQL] remove the view identifier in SubqueryAlias
- [\[SPARK-19868\]](#) conflict TasksetManager lead to spark stopped
- [\[SPARK-19872\]](#) [PYTHON] Use the correct deserializer for RDD construction for coalesce/repartition
- [\[SPARK-19873\]](#)[SS] Record num shuffle partitions in offset log and enforce in next batch.
- [\[SPARK-19876\]](#)[SS][WIP] OneTime Trigger Executor
- [\[SPARK-19877\]](#)[SQL] Restrict the nested level of a view
- [\[SPARK-19882\]](#)[SQL] Pivot with null as a distinct pivot value throws NPE
- [\[SPARK-19886\]](#) Fix reportDataLoss if statement in SS KafkaSource
- [\[SPARK-19887\]](#)[SQL] dynamic partition keys can be null or empty string
- [\[SPARK-19889\]](#)[SQL] Make TaskContext callbacks thread safe
- [\[SPARK-19891\]](#)[SS] Await Batch Lock notified on stream execution exit
- [\[SPARK-19893\]](#)[SQL] should not run DataFrame set operations with map type
- [\[SPARK-19896\]](#)[SQL] Throw an exception if case classes have circular references in toDS
- [\[SPARK-19899\]](#)[ML] Replace featuresCol with itemsCol in ml.fpm.FPGrowth
- [\[SPARK-19905\]](#)[SQL] Bring back Dataset.inputFiles for Hive SerDe tables
- [\[SPARK-19911\]](#)[STREAMING] Add builder interface for Kinesis DStreams
- [\[SPARK-19912\]](#)[SQL] String literals should be escaped for Hive metastore partition pruning
- [\[SPARK-19915\]](#)[SQL] Exclude cartesian product candidates to reduce the search space
- [\[SPARK-19916\]](#)[SQL] simplify bad file handling
- [\[SPARK-19918\]](#)[SQL] Use TextFileFormat in implementation of TextInputJsonDataSource
- [\[SPARK-19919\]](#)[SQL] Defer throwing the exception for empty paths in CSV datasource into `DataSource`
- [\[SPARK-19922\]](#)[ML] small speedups to findSynonyms
- [\[SPARK-19923\]](#)[SQL] Remove unnecessary type conversions per call in Hive
- [\[SPARK-19924\]](#)[SQL] Handle InvocationTargetException for all Hive Shim
- [\[SPARK-19925\]](#)[SPARKR] Fix SparkR spark.getSparkFiles fails when it was called on executors.
- [\[SPARK-19931\]](#)[SQL] InMemoryTableScanExec should rewrite output partitioning and ordering when aliasing output attributes
- [\[SPARK-19933\]](#)[SQL] Do not change output of a subquery
- [\[SPARK-19944\]](#)[SQL] Move SQLConf from sql/core to sql/catalyst
- [\[SPARK-19945\]](#)[SQL] add test suite for SessionCatalog with HiveExternalCatalog
- [\[SPARK-19946\]](#)[TESTING] DebugFilesystem.assertNoOpenStreams should report the open streams to help debugging
- [\[SPARK-19948\]](#) Document that saveAsTable uses catalog as source of truth for table existence.
- [\[SPARK-19949\]](#)[SQL] unify bad record handling in CSV and JSON
- [\[SPARK-19953\]](#)[ML] Random Forest Models use parent UID when being fit
- [\[SPARK-19955\]](#)[PYSPARK] Jenkins Python Conda based test.
- [\[SPARK-19959\]](#)[SQL] Fix to throw NullPointerException in df[java.lang.Long].collect
- [\[SPARK-19960\]](#)[CORE] Move `SparkHadoopWriter` to `internal/io/`
- [\[SPARK-19965\]](#)[SS] DataFrame batch reader may fail to infer partitions when reading FileStreamSink's output
- [\[SPARK-19967\]](#)[SQL] Add from\_json in FunctionRegistry
- [\[SPARK-19968\]](#)[SS] Use a cached instance of `KafkaProducer` instead of creating one every batch.
- [\[SPARK-19969\]](#)[ML] Imputer doc and example

- [\[SPARK-19970\]](#)[SQL] Table owner should be USER instead of PRINCIPAL in kerberized clusters
- [\[SPARK-19980\]](#)[SQL] Add NULL checks in Bean serializer
- [\[SPARK-19985\]](#)[ML] Fixed copy method for some ML Models
- [\[SPARK-19987\]](#)[SQL] Pass all filters into FileIndex
- [\[SPARK-19990\]](#)[SQL][TEST-MAVEN] create a temp file for file in test.jar's resource when run mvn test accross different modules
- [\[SPARK-19991\]](#)[CORE][YARN] FileSegmentManagedBuffer performance improvement
- [\[SPARK-19993\]](#)[SQL] Caching logical plans containing subquery expressions does not work.
- [\[SPARK-19994\]](#)[SQL] Wrong outputOrdering for right/full outer smj
- [\[SPARK-19995\]](#)[YARN] Register tokens to current UGI to avoid re-issuing of tokens in yarn client mode
- [\[SPARK-19998\]](#)[BLOCK MANAGER] Change the exception log to add RDD id of the related the block
- [\[SPARK-19999\]](#) Workaround JDK-8165231 to identify PPC64 architectures as supporting unaligned access
- [\[SPARK-20003\]](#)[ML] FPGrowthModel setMinConfidence should affect rules generation and transform
- [\[SPARK-20009\]](#)[SQL] Support DDL strings for defining schema in functions.from\_json
- [\[SPARK-20010\]](#)[SQL] Sort information is lost after sort merge join
- [\[SPARK-20017\]](#)[SQL] change the nullability of function 'StringToMap' from 'false' to 'true'
- [\[SPARK-20018\]](#)[SQL] Pivot with timestamp and count should not print internal representation
- [\[SPARK-20020\]](#)[SPARKR] DataFrame checkpoint API
- [\[SPARK-20021\]](#)[PYSPARK] Miss backslash in python code
- [\[SPARK-20023\]](#)[SQL] Output table comment for DESC FORMATTED
- [\[SPARK-20024\]](#)[SQL][TEST-MAVEN] SessionCatalog reset need to set the current database of ExternalCatalog
- [\[SPARK-20030\]](#)[SS] Event-time-based timeout for MapGroupsWithState
- [\[SPARK-20038\]](#)[SQL] FileFormatWriter.ExecuteWriteTask.releaseResources() implementations to be re-entrant
- [\[SPARK-20039\]](#)[ML] rename ChiSquare to ChiSquareTest
- [\[SPARK-20040\]](#)[ML][PYTHON] pyspark wrapper for ChiSquareTest
- [\[SPARK-20042\]](#)[WEB UI] Fix log page buttons for reverse proxy mode
- [\[SPARK-20043\]](#)[ML] DecisionTreeModel: ImpurityCalculator builder fails for uppercase impurity type Gini
- [\[SPARK-20046\]](#)[SQL] Facilitate loop optimizations in a JIT compiler regarding sqlContext.read.parquet()
- [\[SPARK-20047\]](#)[ML] Constrained Logistic Regression
- [\[SPARK-20048\]](#)[SQL] Cloning SessionState does not clone query execution listeners
- [\[SPARK-20051\]](#)[SS] Fix StreamSuite flaky test - recover from v2.1 checkpoint
- [\[SPARK-20057\]](#)[SS] Renamed KeyedState to GroupState in mapGroupsWithState
- [\[SPARK-20059\]](#)[YARN] Use the correct classloader for HBaseCredentialProvider
- [\[SPARK-20064\]](#)[PYSPARK] Bump the PySpark verison number to 2.2
- [\[SPARK-20067\]](#)[SQL] Unify and Clean Up Desc Commands Using Catalog Interface
- [\[SPARK-20070\]](#)[SQL] Redact DataSourceScanExec treeString
- [\[SPARK-20076\]](#)[ML][PYSPARK] Add Python interface for ml.stats.Correlation
- [\[SPARK-20078\]](#)[MESOS] Mesos executor configurability for task name and labels
- [\[SPARK-20084\]](#)[CORE] Remove internal.metrics.updatedBlockStatuses from history files.
- [\[SPARK-20085\]](#)[MESOS] Configurable mesos labels for executors
- [\[SPARK-20086\]](#)[SQL] CollapseWindow should not collapse dependent adjacent windows
- [\[SPARK-20088\]](#) Do not create new SparkContext in SparkR createSparkContext
- [\[SPARK-20092\]](#)[R][PROJECT INFRA] Add the detection for Scala codes dedicated for R in AppVeyor tests
- [\[SPARK-20094\]](#)[SQL] Preventing push down of IN subquery to Join operator
- [\[SPARK-20097\]](#)[ML] Fix visibility discrepancy with numInstances and degreesOfFreedom in LR and GLR
- [\[SPARK-20100\]](#)[SQL] Refactor SessionState initialization
- [\[SPARK-20102\]](#) Fix nightly packaging and RC packaging scripts w/ two minor build fixes
- [\[SPARK-20104\]](#)[SQL] Don't estimate IsNull or IsNotNull predicates for non-leaf node
- [\[SPARK-20119\]](#)[TEST-MAVEN] Fix the test case fail in DataSourceScanExecRedactionSuite
- [\[SPARK-20120\]](#)[SQL] spark-sql support silent mode

- [\[SPARK-20121\]](#)[SQL] simplify NullPropagation with NullIntolerant
- [\[SPARK-20124\]](#)[SQL] Join reorder should keep the same order of final project attributes
- [\[SPARK-20125\]](#)[SQL] Dataset of type option of map does not work
- [\[SPARK-20126\]](#)[SQL] Remove HiveSessionState
- [\[SPARK-20127\]](#)[CORE] few warning have been fixed which IntelliJ IDEA reported IntelliJ IDEA
- [\[SPARK-20131\]](#)[CORE] Don't use `this` lock in StandaloneSchedulerBackend.stop
- [\[SPARK-20134\]](#)[SQL] SQLMetrics.postDriverMetricUpdates to simplify driver side metric updates
- [\[SPARK-20136\]](#)[SQL] Add num files and metadata operation timing to scan operator metrics
- [\[SPARK-20140\]](#)[DSTREAM] Remove hardcoded kinesis retry wait and max retries
- [\[SPARK-20143\]](#)[SQL] DataType.fromJson should throw an exception with better message
- [\[SPARK-20145\]](#) Fix range case insensitive bug in SQL
- [\[SPARK-20146\]](#)[SQL] fix comment missing issue for thrift server
- [\[SPARK-20148\]](#)[SQL] Extend the file commit API to allow subscribing to task commit messages
- [\[SPARK-20151\]](#)[SQL] Account for partition pruning in scan metadataTime metrics
- [\[SPARK-20156\]](#)[CORE][SQL][STREAMING][MLLIB] Java String toLowerCase "Turkish locale bug" causes Spark problems
- [\[SPARK-20159\]](#)[SPARKR][SQL] Support all catalog API in R
- [\[SPARK-20160\]](#)[SQL] Move ParquetConversions and OrcConversions Out Of HiveSessionCatalog
- [\[SPARK-20164\]](#)[SQL] AnalysisException not tolerant of null query plan.
- [\[SPARK-20165\]](#)[SS] Resolve state encoder's deserializer in driver in FlatMapGroupsWithStateExec
- [\[SPARK-20166\]](#)[SQL] Use XXX for ISO 8601 timezone instead of ZZ (FastDateFormat specific) in CSV/JSON timeformat options
- [\[SPARK-20172\]](#)[CORE] Add file permission check when listing files in FsHistoryProvider
- [\[SPARK-20173\]](#)[SQL][HIVE-THRIFTSERVER] Throw NullPointerException when HiveThriftServer2 is shutdown
- [\[SPARK-20175\]](#)[SQL] Exists should not be evaluated in Join operator
- [\[SPARK-20177\]](#) Document about compression way has some little detail ch...
- [\[SPARK-20183\]](#)[ML] Added outlierRatio arg to MLTestingUtils.testOutliersWithSmallWeights
- [\[SPARK-20186\]](#)[SQL] BroadcastHint should use child's stats
- [\[SPARK-20189\]](#)[DSTREAM] Fix spark kinesis testcases to remove deprecated createStream and use Builders
- [\[SPARK-20190\]](#)[APP-ID] applications//jobs' in rest api,status should be [running]s...
- [\[SPARK-20191\]](#)[YARN] Crate wrapper for RackResolver so tests can override it.
- [\[SPARK-20194\]](#) Add support for partition pruning to in-memory catalog
- [\[SPARK-20195\]](#)[SPARKR][SQL] add createTable catalog API and deprecate createExternalTable
- [\[SPARK-20196\]](#)[PYTHON][SQL] update doc for catalog functions for all languages, add pyspark refreshByPath API
- [\[SPARK-20197\]](#)[SPARKR] CRAN check fail with package installation
- [\[SPARK-20198\]](#)[SQL] Remove the inconsistency in table/function name conventions in SparkSession.Catalog APIs
- [\[SPARK-20204\]](#)[SQL] remove SimpleCatalystConf and CatalystConf type alias
- [\[SPARK-20209\]](#)[SS] Execute next trigger immediately if previous batch took longer than trigger interval
- [\[SPARK-20211\]](#)[SQL][BACKPORT-2.2] Fix the Precision and Scale of Decimal Values when the Input is BigDecimal between -1.0 and 1.0
- [\[SPARK-20214\]](#)[ML] Make sure converted csc matrix has sorted indices
- [\[SPARK-20217\]](#)[CORE] Executor should not fail stage if killed task throws non-interrupted exception
- [\[SPARK-20223\]](#)[SQL] Fix typo in tpcds q77.sql
- [\[SPARK-20224\]](#)[SS] Updated docs for streaming dropDuplicates and mapGroupsWithState
- [\[SPARK-20229\]](#)[SQL] add semanticHash to QueryPlan
- [\[SPARK-20231\]](#)[SQL] Refactor star schema code for the subsequent star join detection in CBO
- [\[SPARK-20232\]](#)[PYTHON] Improve combineByKey docs
- [\[SPARK-20233\]](#)[SQL] Apply star-join filter heuristics to dynamic programming join enumeration
- [\[SPARK-20239\]](#)[CORE] Improve HistoryServer's ACL mechanism
- [\[SPARK-20244\]](#)[CORE] Handle incorrect bytesRead metrics when using PySpark

- [\[SPARK-20246\]](#)[SQL] should not push predicate down through aggregate with non-deterministic expressions
- [\[SPARK-20250\]](#)[CORE] Improper OOM error when a task been killed while spilling data
- [\[SPARK-20253\]](#)[SQL] Remove unnecessary nullchecks of a return value from Spark runtime routines in generated Java code
- [\[SPARK-20254\]](#)[SQL] Remove unnecessary data conversion for Dataset with primitive array
- [\[SPARK-20255\]](#) Move listLeafFiles() to InMemoryFileIndex
- [\[SPARK-20260\]](#)[MLLIB] String interpolation required for error message
- [\[SPARK-20262\]](#)[SQL] AssertNotNull should throw NullPointerException
- [\[SPARK-20264\]](#)[SQL] asm should be non-test dependency in sql/core
- [\[SPARK-20265\]](#)[MLLIB] Improve Prefix'span pre-processing efficiency
- [\[SPARK-20270\]](#)[SQL] na.fill should not change the values in long or integer when the default value is in double
- [\[SPARK-20273\]](#)[SQL] Disallow Non-deterministic Filter push-down into Join Conditions
- [\[SPARK-20274\]](#)[SQL] support compatible array element type in encoder
- [\[SPARK-20275\]](#)[UI] Do not display "Completed" column for in-progress applications
- [\[SPARK-20278\]](#)[R] Disable 'multiple\_dots\_linter' lint rule that is against project's code style
- [\[SPARK-20280\]](#)[CORE] FileStatusCache Weigher integer overflow
- [\[SPARK-20281\]](#)[SQL] Print the identical Range parameters of SparkContext APIs and SQL in explain
- [\[SPARK-20283\]](#)[SQL] Add preOptimizationBatches
- [\[SPARK-20284\]](#)[CORE] Make {Des,S}erializationStream extend Closeable
- [\[SPARK-20289\]](#)[SQL] Use StaticInvoke to box primitive types
- [\[SPARK-20291\]](#)[SQL] NaNv(FloatType, NullType) should not be cast to NaNv(DoubleType, DoubleType)
- [\[SPARK-20300\]](#)[ML][PYSPARK] Python API for ALSModel.recommendForAllUsers,Items
- [\[SPARK-20301\]](#)[FLAKY-TEST] Fix Hadoop Shell.runCommand flakiness in Structured Streaming tests
- [\[SPARK-20302\]](#)[SQL] Short circuit cast when from and to types are structurally the same
- [\[SPARK-20303\]](#)[SQL] Rename createTempFunction to registerFunction
- [\[SPARK-20304\]](#)[SQL] AssertNotNull should not include path in string representation
- [\[SPARK-20316\]](#)[SQL] Val and Var should strictly follow the Scala syntax
- [\[SPARK-20318\]](#)[SQL] Use Catalyst type for min/max in ColumnStat for ease of estimation
- [\[SPARK-20329\]](#)[SQL] Make timezone aware expression without timezone unresolved
- [\[SPARK-20335\]](#)[SQL] Children expressions of Hive UDF impacts the determinism of Hive UDF
- [\[SPARK-20341\]](#)[SQL] Support BigInt's value that does not fit in long value range
- [\[SPARK-20344\]](#)[SCHEDULER] Duplicate call in FairSchedulableBuilder.addTaskSetManager
- [\[SPARK-20345\]](#)[SQL] Fix STS error handling logic on HiveSQLException
- [\[SPARK-20349\]](#)[SQL] ListFunctions returns duplicate functions after using persistent functions
- [\[SPARK-20350\]](#) Add optimization rules to apply Complementation Laws.
- [\[SPARK-20354\]](#)[CORE][REST-API] When I request access to the 'http://ip:port/api/v1/applications' link, return 'sparkUser' is empty in REST API.
- [\[SPARK-20356\]](#)[SQL] Pruned InMemoryTableScanExec should have correct output partitioning and ordering
- [\[SPARK-20358\]](#)[CORE] Executors failing stage on interrupted exception thrown by cancelled tasks
- [\[SPARK-20359\]](#)[SQL] Avoid unnecessary execution in EliminateOuterJoin optimization that can lead to NPE
- [\[SPARK-20360\]](#)[PYTHON] reprs for interpreters
- [\[SPARK-20364\]](#)[SQL] Disable Parquet predicate pushdown for fields having dots in the names
- [\[SPARK-20365\]](#)[YARN] Remove local scheme when add path to ClassPath.
- [\[SPARK-20366\]](#)[SQL] Fix recursive join reordering: inside joins are not reordered
- [\[SPARK-20367\]](#) Properly unescape column names of partitioning columns parsed from paths.
- [\[SPARK-20373\]](#)[SQL][SS] Batch queries with 'Dataset/DataFrame.withWatermark()' does not execute
- [\[SPARK-20377\]](#)[SS] Fix JavaStructuredSessionization example
- [\[SPARK-20381\]](#)[SQL] Add SQL metrics of numOutputRows for ObjectHashAggregateExec
- [\[SPARK-20385\]](#)[WEB-UI] Submitted Time' field, the date format needs to be formatted, in running Drivers table or Completed Drivers table in master web ui.

- [\[SPARK-20386\]](#)[SPARK CORE] modify the log info if the block exists on the slave already
- [\[SPARK-20391\]](#)[CORE] Rename memory related fields in ExecutorSummary
- [\[SPARK-20393\]](#)[WEBUI] Strengthen Spark to prevent XSS vulnerabilities
- [\[SPARK-20397\]](#)[SPARKR][SS] Fix flaky test: test\_streaming.R.Terminated by error
- [\[SPARK-20398\]](#)[SQL] range() operator should include cancellation reason when killed
- [\[SPARK-20399\]](#)[SQL] Add a config to fallback string literal parsing consistent with old sql parser behavior
- [\[SPARK-20403\]](#)[SQL] Modify the instructions of some functions
- [\[SPARK-20404\]](#)[CORE] Using Option(name) instead of Some(name)
- [\[SPARK-20405\]](#)[SQL] Dataset.withNewExecutionId should be private
- [\[SPARK-20409\]](#)[SQL] fail early if aggregate function in GROUP BY
- [\[SPARK-20410\]](#)[SQL] Make sparkConf a def in SharedSQLContext
- [\[SPARK-20412\]](#) Throw ParseException from visitNonOptionalPartitionSpec instead of returning null values.
- [\[SPARK-20420\]](#)[SQL] Add events to the external catalog
- [\[SPARK-20421\]](#)[CORE] Mark internal listeners as deprecated.
- [\[SPARK-20423\]](#)[ML] fix MLOR coeffs centering when reg == 0
- [\[SPARK-20426\]](#) Lazy initialization of FileSegmentManagedBuffer for shuffle service.
- [\[SPARK-20430\]](#)[SQL] Initialise RangeExec parameters in a driver side
- [\[SPARK-20435\]](#)[CORE] More thorough redaction of sensitive information
- [\[SPARK-20439\]](#)[SQL] Fix Catalog API listTables and getTable when failed to fetch table metadata
- [\[SPARK-20441\]](#)[SPARK-20432][SS] Within the same streaming query, one StreamingRelation should only be transformed to one StreamingExecutionRelation
- [\[SPARK-20449\]](#)[ML] Upgrade breeze version to 0.13.1
- [\[SPARK-20451\]](#) Filter out nested mapType datatypes from sort order in randomSplit
- [\[SPARK-20452\]](#)[SS][KAFKA] Fix a potential ConcurrentModificationException for batch Kafka DataFrame
- [\[SPARK-20459\]](#)[SQL] JdbcUtils throws IllegalStateException: Cause already initialized after getting SQLException
- [\[SPARK-20461\]](#)[CORE][SS] Use UninterruptibleThread for Executor and fix the potential hang in CachedKafkaConsumer
- [\[SPARK-20464\]](#)[SS] Add a job group and description for streaming queries and fix cancellation of running jobs using the job group
- [\[SPARK-20465\]](#)[CORE] Throws a proper exception when any temp directory could not be got
- [\[SPARK-20471\]](#) Remove AggregateBenchmark testsuite warning: Two level hashmap is disabled but vectorized hashmap is enabled
- [\[SPARK-20473\]](#) Enabling missing types in ColumnVector.Array
- [\[SPARK-20474\]](#) Fixing OnHeapColumnVector reallocation
- [\[SPARK-20476\]](#)[SQL] Block users to create a table that use commas in the column names
- [\[SPARK-20482\]](#)[SQL] Resolving Casts is too strict on having time zone set
- [\[SPARK-20483\]](#) Mesos Coarse mode may starve other Mesos frameworks
- [\[SPARK-20487\]](#)[SQL] Display `serde` for `HiveTableScan` node in explained plan
- [\[SPARK-20492\]](#)[SQL] Do not print empty parentheses for invalid primitive types in parser
- [\[SPARK-20496\]](#)[SS] Bug in KafkaWriter Looks at Unanalyzed Plans
- [\[SPARK-20501\]](#)[ML] ML 2.2 QA: New Scala APIs, docs
- [\[SPARK-20505\]](#)[ML] Add docs and examples for ml.stat.Correlation and ml.stat.ChiSquareTest.
- [\[SPARK-20514\]](#)[CORE] Upgrade Jetty to 9.3.11.v20160721
- [\[SPARK-20517\]](#)[UI] Fix broken history UI download link
- [\[SPARK-20529\]](#)[CORE] Allow worker and master work with a proxy server
- [\[SPARK-20534\]](#)[SQL] Make outer generate exec return empty rows
- [\[SPARK-20537\]](#)[CORE] Fixing OffHeapColumnVector reallocation
- [\[SPARK-20540\]](#)[CORE] Fix unstable executor requests.
- [\[SPARK-20541\]](#)[SPARKR][SS] support awaitTermination without timeout
- [\[SPARK-20544\]](#)[SPARKR] skip tests when running on CRAN

- [\[SPARK-20546\]](#)[DEPLOY] spark-class gets syntax error in posix mode
- [\[SPARK-20548\]](#)[FLAKY-TEST] share one REPL instance among REPL test cases
- [\[SPARK-20549\]](#) java.io.CharConversionException: Invalid UTF-32' in JsonToStructs
- [\[SPARK-20553\]](#)[ML][PYSPARK] Update ALS examples with recommend-all methods
- [\[SPARK-20555\]](#)[SQL] Fix mapping of Oracle DECIMAL types to Spark types in read path
- [\[SPARK-20558\]](#)[CORE] clear InheritableThreadLocal variables in SparkContext when stopping it
- [\[SPARK-20567\]](#) Lazily bind in GenerateExec
- [\[SPARK-20569\]](#)[SQL] RuntimeReplaceable functions should not take extra parameters
- [\[SPARK-20571\]](#)[SPARKR][SS] Flaky Structured Streaming tests
- [\[SPARK-20574\]](#)[ML] Allow Bucketizer to handle non-Double numeric column
- [\[SPARK-20576\]](#)[SQL] Support generic hint function in Dataset/DataFrame
- [\[SPARK-20584\]](#)[PYSPARK][SQL] Python generic hint support
- [\[SPARK-20585\]](#)[SPARKR] R generic hint support
- [\[SPARK-20587\]](#)[ML] Improve performance of ML ALS recommendForAll
- [\[SPARK-20588\]](#)[SQL] Cache TimeZone instances.
- [\[SPARK-20590\]](#)[SQL] Use Spark internal datasource if multiples are found for the same shorten name
- [\[SPARK-20594\]](#)[SQL] The staging directory should be a child directory starts with "." to avoid being deleted if we set hive.exec.stagingdir under the table directory.
- [\[SPARK-20600\]](#)[SS] KafkaRelation should be pretty printed in web UI
- [\[SPARK-20613\]](#) Remove excess quotes in Windows executable
- [\[SPARK-20616\]](#) RuleExecutor logDebug of batch results should show diff to start of batch
- [\[SPARK-20621\]](#)[DEPLOY] Delete deprecated config parameter in 'spark-env.sh'
- [\[SPARK-20626\]](#)[SPARKR] address date test warning with timezone on windows
- [\[SPARK-20627\]](#)[PYSPARK] Drop the hadoop distribution name from the Python version
- [\[SPARK-20630\]](#)[WEB UI] Fixed column visibility in Executor Tab
- [\[SPARK-20631\]](#)[PYTHON][ML] LogisticRegression.\_checkThresholdConsistency should use values not Params
- [\[SPARK-20665\]](#)[SQL] Bround" and "Round" function return NULL
- [\[SPARK-20669\]](#)[ML] LoR.family and LDA.optimizer should be case insensitive
- [\[SPARK-20674\]](#)[SQL] Support registering UserDefinedFunction as named UDF
- [\[SPARK-20678\]](#)[SQL] Ndv for columns not in filter condition should also be updated
- [\[SPARK-20685\]](#) Fix BatchPythonEvaluation bug in case of single UDF w/ repeated arg.
- [\[SPARK-20686\]](#)[SQL] PropagateEmptyRelation incorrectly handles aggregate without grouping
- [\[SPARK-20687\]](#)[MLLIB] mllib.Matrices.fromBreeze may crash when converting from Breeze sparse matrix
- [\[SPARK-20688\]](#)[SQL] correctly check analysis for scalar sub-queries
- [\[SPARK-20700\]](#)[SQL] InferFiltersFromConstraints stackoverflows for query (v2)
- [\[SPARK-20702\]](#)[CORE] TaskContextImpl.markTaskCompleted should not hide the original error
- [\[SPARK-20704\]](#)[SPARKR] change CRAN test to run single thread
- [\[SPARK-20705\]](#)[WEB-UI] The sort function can not be used in the master page when you use Firefox or Google Chrome.
- [\[SPARK-20707\]](#)[ML] ML deprecated APIs should be removed in major release.
- [\[SPARK-20710\]](#)[SQL] Support aliases in CUBE/ROLLUP/GROUPING SETS
- [\[SPARK-20714\]](#)[SS] Fix match error when watermark is set with timeout = no timeout / processing timeout
- [\[SPARK-20716\]](#)[SS] StateStore.abort() should not throw exceptions
- [\[SPARK-20717\]](#)[SS] Minor tweaks to the MapGroupsWithState behavior
- [\[SPARK-20718\]](#)[SQL] FileSourceScanExec with different filter orders should be the same after canonicalization
- [\[SPARK-20725\]](#)[SQL] partial aggregate should behave correctly for sameResult
- [\[SPARK-20727\]](#) Skip tests that use Hadoop utils on CRAN Windows
- [\[SPARK-20741\]](#)[SPARK SUBMIT] Added cleanup of JARs archive generated by SparkSubmit
- [\[SPARK-20756\]](#)[YARN] yarn-shuffle jar references unshaded guava
- [\[SPARK-20759\]](#) SCALA\_VERSION in \_config.yml should be consistent with pom.xml



- [\[SPARK-20763\]](#)[SQL] The function of `month` and `day` return the value which is not we expected.
- [\[SPARK-20764\]](#)[ML][PYSPARK] Fix visibility discrepancy with numInstances and degreesOfFreedom in LR and GLR - Python version
- [\[SPARK-20768\]](#)[PYSPARK][ML] Expose numPartitions (expert) param of PySpark FPGrowth.
- [\[SPARK-20773\]](#)[SQL] ParquetWriteSupport.writeFields is quadratic in number of fields
- [\[SPARK-20776\]](#) Fix perf. problems in JobProgressListener caused by TaskMetrics construction
- [\[SPARK-20781\]](#) the location of Dockerfile in docker.properties.templat is wrong
- [\[SPARK-20788\]](#)[CORE] Fix the Executor task reaper's false alarm warning logs
- [\[SPARK-20790\]](#)[MLLIB] Correctly handle negative values for implicit feedback in ALS
- [\[SPARK-20792\]](#)[SS] Support same timeout operations in mapGroupsWithState function in batch queries as in streaming queries
- [\[SPARK-20796\]](#) the location of start-master.sh in spark-standalone.md is wrong
- [\[SPARK-20798\]](#) GenerateUnsafeProjection should check if a value is null before calling the getter
- [\[SPARK-20801\]](#) Record accurate size of blocks in MapStatus when it's above threshold.
- [\[SPARK-20813\]](#)[WEB UI] Fixed Web UI executor page tab search by status not working
- [\[SPARK-20814\]](#)[MESOS] Restore support for spark.executor.extraClassPath.
- [\[SPARK-20815\]](#)[SPARKR] NullPointerException in RPackageUtils#checkManifestForR
- [\[SPARK-20831\]](#)[SQL] Fix INSERT OVERWRITE data source tables with IF NOT EXISTS
- [\[SPARK-20843\]](#)[CORE] Add a config to set driver terminate timeout
- [\[SPARK-20844\]](#) Remove experimental from Structured Streaming APIs
- [\[SPARK-20848\]](#)[SQL] Shutdown the pool after reading parquet files
- [\[SPARK-20854\]](#)[SQL] Extend hint syntax to support expressions
- [\[SPARK-20857\]](#)[SQL] Generic resolved hint node
- [\[SPARK-20861\]](#)[ML][PYTHON] Delegate looping over paramMap to estimators
- [\[SPARK-20862\]](#)[MLLIB][PYTHON] Avoid passing float to ndarray.reshape in LogisticRegressionModel
- [\[SPARK-20867\]](#)[SQL] Move hints from Statistics into HintInfo class
- [\[SPARK-20868\]](#)[CORE] UnsafeShuffleWriter should verify the position after FileChannel.transferTo
- [\[SPARK-20872\]](#)[SQL] ShuffleExchange.nodeName should handle null coordinator
- [\[SPARK-20874\]](#)[EXAMPLES] Add Structured Streaming Kafka Source to examples project
- [\[SPARK-20876\]](#)[SQL][BACKPORT-2.2] If the input parameter is float type for ceil or floor,the result is not we expected
- [\[SPARK-20877\]](#)[SPARKR][WIP] add timestamps to test runs
- [\[SPARK-20897\]](#)[SQL] cached self-join should not fail
- [\[SPARK-20908\]](#)[SQL] Cache Manager: Hint should be ignored in plan matching
- [\[SPARK-20920\]](#)[SQL] ForkJoinPool pools are leaked when writing hive tables with many partitions
- [\[SPARK-20922\]](#)[CORE] Add whitelist of classes that can be deserialized by the launcher.
- [\[SPARK-20924\]](#)[SQL] Unable to call the function registered in the not-current database
- [\[SPARK-20926\]](#)[SQL] Removing exposures to guava library caused by directly accessing SessionCatalog's tableRelationCache
- [\[SPARK-20929\]](#)[ML] LinearSVC should use its own threshold param
- [\[SPARK-20940\]](#)[CORE] Replace IllegalAccessError with IllegalStateException
- [\[SPARK-20942\]](#)[WEB-UI] The title style about field is error in the history server web ui.
- [\[SPARK-20954\]](#)[SQL][BRANCH-2.2][EXTENDED] DESCRIBE ` result should be compatible with previous Spark
- [\[SPARK-20955\]](#)[CORE] Intern "executorId" to reduce the memory usage
- [\[SPARK-20967\]](#)[SQL] SharedState.externalCatalog is not really lazy
- [\[SPARK-20979\]](#)[SS] Add RateSource to generate values for tests and benchmark
- [\[SPARK-20980\]](#)[SQL] Rename `wholeFile` to `multiLine` for both CSV and JSON
- [\[SPARK-20986\]](#)[SQL] Reset table's statistics after PruneFileSourcePartitions rule.
- [\[SPARK-21041\]](#)[SQL] SparkSession.range should be consistent with SparkContext.range
- [\[SPARK-21042\]](#)[SQL] Document Dataset.union is resolution by position
- [\[SPARK-21050\]](#)[ML] Word2vec persistence overflow bug fix

- [\[SPARK-21059\]](#)[SQL] LikeSimplification can NPE on null pattern
- [\[SPARK-21060\]](#)[WEB-UI] Css style about paging function is error in the executor page. Css style about paging function is error in the executor page. It is different of history server ui paging function css style.
- [\[SPARK-21072\]](#)[SQL] TreeNode.mapChildren should only apply to the children node.
- [\[SPARK-21079\]](#)[SQL] Calculate total size of a partition table as a sum of individual partitions
- [\[SPARK-21085\]](#)[SQL] Failed to read the partitioned table created by Spark 2.1
- [\[SPARK-21089\]](#)[SQL] Fix DESC EXTENDED/FORMATTED to Show Table Properties
- [\[SPARK-21090\]](#)[CORE] Optimize the unified memory manager code
- [\[SPARK-21126\]](#) The configuration which named "spark.core.connection.auth.wait.timeout" hasn't been used in spark
- [\[SPARK-21129\]](#)[SQL] Arguments of SQL function call should not be named expressions
- [\[SPARK-21132\]](#)[SQL] DISTINCT modifier of function arguments should not be silently ignored
- [\[SPARK-21133\]](#)[CORE] Fix HighlyCompressedMapStatus#writeExternal throws NPE
- [\[SPARK-21138\]](#)[YARN] Cannot delete staging dir when the clusters of "spark.yarn.stagingDir" and "spark.hadoop.fs.defaultFS" are different
- [\[SPARK-21144\]](#)[SQL] Print a warning if the data schema and partition schema have the duplicate columns
- [\[SPARK-21150\]](#)[SQL] Persistent view stored in Hive metastore should be case preserving
- [\[SPARK-21159\]](#)[CORE] Don't try to connect to launcher in standalone cluster mode.
- [\[SPARK-21165\]](#) [SQL] [2.2] Use executedPlan instead of analyzedPlan in INSERT AS SELECT [WIP]
- [\[SPARK-21167\]](#)[SS] Decode the path generated by File sink to handle special characters
- [\[SPARK-21176\]](#)[WEB UI] Limit number of selector threads for admin ui proxy servlets to 8
- [\[SPARK-21181\]](#) Release byteBuffers to suppress netty error messages
- [\[SPARK-21203\]](#)[SQL] Fix wrong results of insertion of Array of Struct
- [\[SPARK-21253\]](#)[CORE] Disable spark.reducer.maxReqSizeShuffleToMem
- [\[SPARK-21258\]](#)[SQL] Fix WindowExec complex object aggregation with spilling
- [\[SPARK-5484\]](#)[GRAPHX] Periodically do checkpoint in Pregel
- [\[SPARK-6227\]](#)[MLLIB][PYSPARK] Implement PySpark wrappers for SVD and PCA (v2)
- [\[SPARK-8184\]](#)[SQL] Add additional function description for weekofyear
- [\[SPARK-9002\]](#)[CORE] KryoSerializer initialization does not include 'Array[Int]'
- [\[SPARK-9435\]](#)[SQL] Reuse function in Java UDF to correctly support expressions that require equality comparison between ScalaUDF

### Issues Fixed in CDS 2.1 Release 4

The following list includes issues fixed in CDS 2.1 Release 4. Test-only changes are omitted.

- [\[SPARK-23243\]](#)[\[SPARK-20715\]](#)[CORE][2.2] Fix RDD.repartition() data correctness issue
- [\[SPARK-17769\]](#)[CORE][SCHEDULER] Some FetchFailure refactoring
- [\[SPARK-26201\]](#) Fix python broadcast with encryption
- [\[SPARK-24918\]](#)[CORE] Executor Plugin API
- [PYSPARK][SQL] Updates to RowQueue
- [PYSPARK] Updates to pyspark broadcast
- [\[SPARK-25253\]](#)[PYSPARK] Refactor local connection & auth code
- CDH-74338. Upgrade jackson-databind to Cloudera version
- [spark] CDH-57150. Exit spark-shell/spark-submit/pyspark with correct error message if no client configuration found

### Issues Fixed in CDS 2.1 Release 3

The following list includes issues fixed in CDS 2.1 Release 3. Test-only changes are omitted.

- [\[SPARK-23207\]](#)[\[SPARK-22905\]](#)[\[SPARK-24564\]](#)[\[SPARK-25114\]](#)[SQL][BACKPORT-2.1] Shuffle+Repartition on a DataFrame could lead to incorrect answers

- [\[SPARK-24950\]](#)[SQL] DateTimeUtilsSuite daysToMillis and millisToDays fails w/java 8 181-b13
- [PYSARK] Updates to Accumulators
- [\[SPARK-24809\]](#)[SQL] Serializing LongToUnsafeRowMap in executor may result in data error
- [\[SPARK-20223\]](#)[SQL] Fix typo in tpcds q77.sql
- [\[SPARK-24589\]](#)[CORE] Correctly identify tasks in output commit coordinator [branch-2.1].
- [\[SPARK-22897\]](#)[CORE] Expose stageAttemptId in TaskContext
- [\[SPARK-23732\]](#)[DOCS] Fix source links in generated scaladoc.
- [WEBUI] Avoid possibility of script in query param keys
- Fix compilation caused by SPARK-24257
- [\[SPARK-24257\]](#)[SQL] LongToUnsafeRowMap calculate the new size may be wrong
- [R][BACKPORT-2.2] backport lint fix
- [SPARKR] Match pyspark features in SparkR communication protocol.
- [PYSARK] Update py4j to version 0.10.7.
- [\[SPARK-21278\]](#)[PYSARK] Upgrade to Py4J 0.10.6
- [\[SPARK-23697\]](#)[CORE] LegacyAccumulatorWrapper should define isZero correctly
- [\[SPARK-23053\]](#)[CORE][BRANCH-2.1] taskBinarySerialization and task partitions calculate in DagScheduler.submitMissingTasks should keep the same RDD checkpoint status
- [\[SPARK-22862\]](#) Docs on lazy elimination of columns missing from an encoder
- [\[SPARK-22688\]](#)[SQL] Upgrade Janino version to 3.0.8
- [\[SPARK-22373\]](#)[BUILD][FOLLOWUP][BRANCH-2.1] Updates other dependency lists too for Janino
- [\[SPARK-22373\]](#) Bump Janino dependency version to fix thread safety issue...
- [\[SPARK-22548\]](#)[SQL] Incorrect nested AND expression pushed down to JDBC data source
- [\[SPARK-22377\]](#)[BUILD] Use /usr/sbin/lsif if lsif does not exists in release-build.sh
- [\[SPARK-22327\]](#)[SPARKR][TEST][BACKPORT-2.1] check for version warning
- [\[SPARK-22429\]](#)[STREAMING] Streaming checkpointing code does not retry after failure
- [MINOR][DOC] automatic type inference supports also Date and Timestamp
- [\[SPARK-21991\]](#)[LAUNCHER][FOLLOWUP] Fix java lint
- [\[SPARK-21991\]](#)[LAUNCHER] Fix race condition in LauncherServer#acceptConnections
- [\[SPARK-22273\]](#)[SQL] Fix key/value schema field names in HashMapGenerators.
- [\[SPARK-22206\]](#)[SQL][SPARKR] gaply in R can't work on empty grouping columns
- [\[SPARK-20466\]](#)[CORE] HadoopRDD#addLocalConfiguration throws NPE
- [\[SPARK-22167\]](#)[R][BUILD] sparkr packaging issue allow zinc
- [\[SPARK-22129\]](#)[\[SPARK-22138\]](#) Release script improvements
- [\[SPARK-18136\]](#) Fix SPARK\_JARS\_DIR for Python pip install on Windows
- [\[SPARK-22072\]](#)[\[SPARK-22071\]](#)[BUILD] Improve release build scripts
- [\[SPARK-19318\]](#)[\[SPARK-22041\]](#)[\[SPARK-16625\]](#)[BACKPORT-2.1][SQL] Docker test case failure: `: General data types to be mapped to Oracle`
- [\[SPARK-22052\]](#) Incorrect Metric assigned in MetricsReporter.scala
- [\[SPARK-22043\]](#)[PYTHON] Improves error message for show\_profiles and dump\_profiles
- [\[SPARK-21953\]](#) Show both memory and disk bytes spilled if either is present
- [\[SPARK-21985\]](#)[PYSARK] PairDeserializer is broken for double-zipped RDDs
- [\[SPARK-21976\]](#)[DOC] Fix wrong documentation for Mean Absolute Error.
- [\[SPARK-21950\]](#)[SQL][PYTHON][TEST] pyspark.sql.tests.SQLTests2 should stop SparkContext.
- [\[SPARK-21826\]](#)[SQL][2.1][2.0] outer broadcast hash join should not throw NPE
- [MINOR] Correct validateAndTransformSchema in GaussianMixture and AFTSurvivalRegression
- [\[SPARK-18752\]](#)[SQL] Follow-up: add scaladoc explaining isSrcLocal arg.
- [\[SPARK-18752\]](#)[HIVE] isSrcLocal" value should be set from user query.
- [CDH-70445] Executor Plugin Api.
- [\[SPARK-23852\]](#)[SQL] Add test that fails if PARQUET-1217 is not fixed.
- [CDH-68516] Check for null when writing decimal.

- [CDH-69165] Handle file names with spaces in classpath.
- [SPARK-23991][DSTREAMS] Fix data loss when WAL write fails in allocateBlocksToBatch
- [SPARK-24309][CORE] AsyncEventQueue should stop on interrupt.
- [SPARK-22850][CORE] Ensure queued events are delivered to all event queues.
- [CDH-68051] Try to fetch tokens for all KMS servers.
- [SPARK-23433][CORE] Late zombie task completions update all tasksets
- [SPARK-23660] Fix exception in yarn cluster mode when application ended fast
- [SPARK-23438][DSTREAMS] Fix DStreams data loss with WAL when driver crashes
- [SPARK-18971][CORE] Upgrade Netty to 4.0.43.Final
- [SPARK-21551][PYTHON] Increase timeout for PythonRDD.serveIterator

### Issues Fixed in CDS 2.1 Release 2

The following list includes issues fixed in CDS 2.1 Release 2. Test-only changes are omitted.

- [SPARK-19554][UI,YARN] Allow SHS URL to be used for tracking in YARN RM.
- [SPARK-22083][CORE] Release locks in MemoryStore.evictBlocksToFreeSpace
- [SPARK-18838][HOTFIX][YARN] Check internal context state before stopping it.
- [SPARK-18838][CORE] Add separate listener queues to LiveListenerBus.
- [SPARK-21928][CORE] Set classloader on SerializerManager's private kryo
- [SPARK-21254][WEBUI] History UI performance fixes
- [SPARK-21135][WEB UI] On history server page<sup>2</sup> **caution** of incompleting applications should be hidden instead of showing up as 0
- [SPARK-20942][WEB-UI] The title style about field is error in the history server web ui.
- [SPARK-20603][SS][TEST] Set default number of topic partitions to 1 to reduce the load
- [SPARK-18682][SS] Batch Source for Kafka
- [SPARK-19155][ML] MLlib GeneralizedLinearRegression family and link should case insensitive
- [SPARK-19542][HOTFIX][SS] Fix the missing import in DataStreamReaderWriterSuite
- [SPARK-20280][CORE] FileStatusCache Weigher integer overflow
- [SPARK-19646][BUILD][HOTFIX] Fix compile error from cherry-pick of SPARK-19646 into branch 2.1
- [SPARK-21834] Incorrect executor request in case of dynamic allocation
- [SPARK-21721][SQL][BACKPORT-2.1] Clear FileSystem deleteOnExit cache when paths are successfully removed
- [SPARK-21588][SQL] SQLContext.getConf(key, null) should return null
- [SPARK-21330][SQL] Bad partitioning does not allow to read a JDBC table with extreme values on the partition column
- [SPARK-12717][PYTHON][BRANCH-2.1] Adding thread-safe broadcast pickle registry
- [SPARK-21555][SQL] RuntimeReplaceable should be compared semantically by its canonicalized child
- [SPARK-21306][ML] For branch 2.1, OneVsRest should support setWeightCol
- [SPARK-21446][SQL] Fix setAutoCommit never executed
- [SPARK-21441][SQL] Incorrect Codegen in SortMergeJoinExec results failures in some cases
- [SPARK-21332][SQL] Incorrect result type inferred for some decimal expressions
- [SPARK-21344][SQL] BinaryType comparison does signed byte array comparison
- [SPARK-21083][SQL][BRANCH-2.1] Store zero size and row count when analyzing empty table
- [SPARK-21345][SQL][TEST][TEST-MAVEN][BRANCH-2.1] SparkSessionBuilderSuite should clean up stopped sessions.
- [SPARK-21312][SQL] correct offsetInBytes in UnsafeRow.writeToStream
- [SPARK-20256][SQL][BRANCH-2.1] SessionState should be created more lazily
- [SPARK-19104][BACKPORT-2.1][SQL] Lambda variables in ExternalMapToCatalyst should be global
- [SPARK-21203][SQL] Fix wrong results of insertion of Array of Struct
- [SPARK-20555][SQL] Fix mapping of Oracle DECIMAL types to Spark types in read path
- [SPARK-21181] Release byteBuffers to suppress netty error messages
- [SPARK-21167][SS] Decode the path generated by File sink to handle special characters

- [\[SPARK-21138\]](#)[YARN] Cannot delete staging dir when the clusters of "spark.yarn.stagingDir" and "spark.hadoop.fs.defaultFS" are different
- [\[SPARK-19688\]](#)[STREAMING] Not to read `spark.yarn.credentials.file` from checkpoint.
- [\[SPARK-21114\]](#)[TEST][2.1] Fix test failure in Spark 2.1/2.0 due to name mismatch
- [\[SPARK-21072\]](#)[SQL] TreeNode.mapChildren should only apply to the children node.
- [\[SPARK-16251\]](#)[SPARK-20200][CORE][TEST] Flaky test: org.apache.spark.rdd.LocalCheckpointSuite.missing checkpoint block fails with informative message
- [\[SPARK-20211\]](#)[SQL][BACKPORT-2.2] Fix the Precision and Scale of Decimal Values when the Input is BigDecimal between -1.0 and 1.0
- [\[SPARK-21064\]](#)[CORE][TEST] Fix the default value bug in NettyBlockTransferServiceSuite
- [\[SPARK-20920\]](#)[SQL] ForkJoinPool pools are leaked when writing hive tables with many partitions
- [\[SPARK-20940\]](#)[CORE] Replace IllegalAccessError with IllegalStateException
- [\[SPARK-20275\]](#)[UI] Do not display "Completed" column for in-progress applications
- [\[SPARK-20868\]](#)[CORE] UnsafeShuffleWriter should verify the position after FileChannel.transferTo
- [\[SPARK-20250\]](#)[CORE] Improper OOM error when a task been killed while spilling data
- [\[SPARK-20848\]](#)[SQL] Shutdown the pool after reading parquet files
- [\[SPARK-20862\]](#)[MLLIB][PYTHON] Avoid passing float to ndarray.reshape in LogisticRegressionModel
- [\[SPARK-18406\]](#)[CORE][BACKPORT-2.1] Race between end-of-task and completion iterator read lock release
- [\[SPARK-20687\]](#)[MLLIB] mllib.Matrices.fromBreeze may crash when converting from Breeze sparse matrix
- [\[SPARK-20798\]](#) GenerateUnsafeProjection should check if a value is null before calling the getter
- [\[SPARK-17424\]](#) Fix unsound substitution bug in ScalaReflection.
- [\[SPARK-20665\]](#)[SQL] Bround" and "Round" function return NULL
- [\[SPARK-20685\]](#) Fix BatchPythonEvaluation bug in case of single UDF w/ repeated arg.
- [\[SPARK-20688\]](#)[SQL] correctly check analysis for scalar sub-queries
- [\[SPARK-19933\]](#)[SQL] Do not change output of a subquery
- [\[SPARK-20631\]](#)[PYTHON][ML] LogisticRegression.\_checkThresholdConsistency should use values not Params
- [\[SPARK-20686\]](#)[SQL] PropagateEmptyRelation incorrectly handles aggregate without grouping
- [\[SPARK-17685\]](#)[SQL] Make SortMergeJoinExec's currentVars is null when calling createJoinKey
- [\[SPARK-20615\]](#)[ML][TEST] SparseVector.argmax throws IndexOutOfBoundsException
- [\[SPARK-20616\]](#) RuleExecutor logDebug of batch results should show diff to start of batch
- [\[SPARK-20546\]](#)[DEPLOY] spark-class gets syntax error in posix mode
- [\[SPARK-20558\]](#)[CORE] clear InheritableThreadLocal variables in SparkContext when stopping it
- [\[SPARK-20540\]](#)[CORE] Fix unstable executor requests.
- [\[SPARK-20517\]](#)[UI] Fix broken history UI download link
- [\[SPARK-20404\]](#)[CORE] Using Option(name) instead of Some(name)
- [\[SPARK-20451\]](#) Filter out nested mapType datatypes from sort order in randomSplit
- [\[SPARK-20450\]](#)[SQL] Unexpected first-query schema inference cost with 2.1.1
- [\[SPARK-20439\]](#)[SQL][BACKPORT-2.1] Fix Catalog API listTables and getTable when failed to fetch table metadata
- [\[SPARK-20407\]](#)[TESTS][BACKPORT-2.1] ParquetQuerySuite 'Enabling/disabling ignoreCorruptFiles' flaky test
- [\[SPARK-20243\]](#)[TESTS] DebugFilesystem.assertNoOpenStreams thread race
- [\[SPARK-20409\]](#)[SQL] fail early if aggregate function in GROUP BY
- [\[SPARK-20359\]](#)[SQL] Avoid unnecessary execution in EliminateOuterJoin optimization that can lead to NPE
- [\[SPARK-17647\]](#)[SQL] Fix backslash escaping in 'LIKE' patterns.
- [\[SPARK-20335\]](#)[SQL][BACKPORT-2.1] Children expressions of Hive UDF impacts the determinism of Hive UDF
- [\[SPARK-20131\]](#)[CORE] Don't use `this` lock in StandaloneSchedulerBackend.stop
- [\[SPARK-20304\]](#)[SQL] AssertNotNull should not include path in string representation
- [\[SPARK-20291\]](#)[SQL] NaNv(FloatType, NullType) should not be cast to NaNv(DoubleType, DoubleType)
- [\[SPARK-17564\]](#)[TESTS] Fix flaky RequestTimeoutIntegrationSuite.furtherRequestsDelay
- [\[SPARK-20270\]](#)[SQL] na.fill should not change the values in long or integer when the default value is in double
- [\[SPARK-20264\]](#)[SQL] asm should be non-test dependency in sql/core

- [\[SPARK-20260\]](#)[MLLIB] String interpolation required for error message
- [\[SPARK-20262\]](#)[SQL] AssertNotNull should throw NullPointerException
- [\[SPARK-20246\]](#)[SQL] should not push predicate down through aggregate with non-deterministic expressions
- [\[SPARK-20214\]](#)[ML] Make sure converted csc matrix has sorted indices
- [\[SPARK-20191\]](#)[YARN] Crate wrapper for RackResolver so tests can override it.
- [\[SPARK-20190\]](#)[APP-ID] applications//jobs' in rest api,status should be [running]s...
- [\[SPARK-20164\]](#)[SQL] AnalysisException not tolerant of null query plan.
- [\[SPARK-20059\]](#)[YARN] Use the correct classloader for HBaseCredentialProvider
- [\[SPARK-20134\]](#)[SQL] SQLMetrics.postDriverMetricUpdates to simplify driver side metric updates
- [\[SPARK-20043\]](#)[ML] DecisionTreeModel: ImpurityCalculator builder fails for uppercase impurity type Gini
- [\[SPARK-20125\]](#)[SQL] Dataset of type option of map does not work
- [\[SPARK-19995\]](#)[YARN] Register tokens to current UGI to avoid re-issuing of tokens in yarn client mode
- [\[SPARK-20086\]](#)[SQL] CollapseWindow should not collapse dependent adjacent windows
- [\[SPARK-19959\]](#)[SQL] Fix to throw NullPointerException in df[java.lang.Long].collect
- [\[SPARK-20017\]](#)[SQL] change the nullability of function 'StringToMap' from 'false' to 'true'
- [\[SPARK-19912\]](#)[SQL] String literals should be escaped for Hive metastore partition pruning
- [\[SPARK-17204\]](#)[CORE] Fix replicated off heap storage
- [\[SPARK-19980\]](#)[SQL][BACKPORT-2.1] Add NULL checks in Bean serializer
- [\[SPARK-19970\]](#)[SQL][BRANCH-2.1] Table owner should be USER instead of PRINCIPAL in kerberized clusters
- [\[SPARK-19994\]](#)[SQL] Wrong outputOrdering for right/full outer smj
- [\[SPARK-19946\]](#)[TESTS][BACKPORT-2.1] DebugFilesystem.assertNoOpenStreams should report the open streams to help debugging
- [\[SPARK-19872\]](#) [PYTHON] Use the correct deserializer for RDD construction for coalesce/repartition
- [\[SPARK-19887\]](#)[SQL] dynamic partition keys can be null or empty string
- [\[SPARK-19944\]](#)[SQL] Move SQLConf from sql/core to sql/catalyst (branch-2.1)
- [\[SPARK-19924\]](#)[SQL][BACKPORT-2.1] Handle InvocationTargetException for all Hive Shim
- [\[SPARK-19893\]](#)[SQL] should not run DataFrame set operations with map type
- [\[SPARK-19891\]](#)[SS] Await Batch Lock notified on stream execution exit
- [\[SPARK-19861\]](#)[SS] watermark should not be a negative time.
- [\[SPARK-19813\]](#) maxFilesPerTrigger combo latestFirst may miss old files in combination with maxFileAge in FileStreamSource
- [\[SPARK-18055\]](#)[SQL] Use correct mirror in ExpressionEncoder
- [\[SPARK-19348\]](#)[PYTHON] PySpark keyword\_only decorator is not thread-safe
- [\[SPARK-19859\]](#)[SS][FOLLOW-UP] The new watermark should override the old one.
- [\[SPARK-19859\]](#)[SS] The new watermark should override the old one
- [\[SPARK-19561\]](#)[SQL] add int case handling for TimestampType
- [\[SPARK-19774\]](#) StreamExecution should call stop() on sources when a stream fails
- [\[SPARK-19779\]](#)[SS] Delete needless tmp file after restart structured streaming job
- [\[SPARK-19748\]](#)[SQL] refresh function has a wrong order to do cache invalidate and regenerate the inmemory var for InMemoryFileIndex with FileStatusCache
- [\[SPARK-19594\]](#)[STRUCTURED STREAMING] StreamingQueryListener fails to handle QueryTerminatedEvent if more than one listeners exists
- [\[SPARK-19707\]](#)[CORE] Improve the invalid path check for sc.addJar
- [\[SPARK-19674\]](#)[SQL] Ignore driver accumulator updates don't belong to ...
- [\[SPARK-19691\]](#)[SQL][BRANCH-2.1] Fix ClassCastException when calculating percentile of decimal column
- [\[SPARK-19646\]](#)[CORE][STREAMING] binaryRecords replicates records in scala API
- [\[SPARK-19500\]](#) [SQL] Fix off-by-one bug in BytesToBytesMap
- [\[SPARK-19622\]](#)[WEBUI] Fix a http error in a paged table when using a `Go` button to search.
- [\[SPARK-19603\]](#)[SS] Fix StreamingQuery explain command

- [\[SPARK-19329\]](#)[SQL][BRANCH-2.1] Reading from or writing to a datasource table with a non pre-existing location should succeed
- [\[SPARK-19501\]](#)[YARN] Reduce the number of HDFS RPCs during YARN deployment
- [\[SPARK-17714\]](#)[CORE][TEST-MAVEN][TEST-HADOOP2.6] Avoid using ExecutorClassLoader to load Netty generated classes
- [\[SPARK-19542\]](#)[SS] Delete the temp checkpoint if a query is stopped without errors
- [\[SPARK-19543\]](#) from\_json fails when the input row is empty
- [\[SPARK-19509\]](#)[SQL] Grouping Sets do not respect nullable grouping columns
- [\[SPARK-18609\]](#)[SPARK-18841][SQL][BACKPORT-2.1] Fix redundant Alias removal in the optimizer
- [\[SPARK-19407\]](#)[SS] defaultFS is used FileSystem.get instead of getting it from uri scheme
- [\[SPARK-19472\]](#)[SQL] Parser should not mistake CASE WHEN(...) for a function call
- [\[SPARK-19432\]](#)[CORE] Fix an unexpected failure when connecting timeout
- [\[SPARK-19377\]](#)[WEBUI][CORE] Killed tasks should have the status as KILLED
- [\[SPARK-19378\]](#)[SS] Ensure continuity of stateOperator and eventTime metrics even if there is no new data in trigger
- [\[SPARK-19406\]](#)[SQL] Fix function to\_json to respect user-provided options
- [\[SPARK-19338\]](#)[SQL] Add UDF names in explain
- [\[SPARK-18863\]](#)[SQL] Output non-aggregate expressions without GROUP BY in a subquery does not yield an error
- [\[SPARK-19330\]](#)[DSTREAMS] Also show tooltip for successful batches
- [\[SPARK-19017\]](#)[SQL] NOT IN subquery with more than one column may return incorrect results
- [\[SPARK-16473\]](#)[MLLIB] Fix BisectingKMeans Algorithm failing in edge case
- [\[SPARK-9435\]](#)[SQL] Reuse function in Java UDF to correctly support expressions that require equality comparison between ScalaUDF
- [\[SPARK-19306\]](#)[CORE] Fix inconsistent state in DiskBlockObject when exception occurred
- [\[SPARK-19155\]](#)[ML] Make family case insensitive in GLM
- [\[SPARK-14536\]](#)[SQL][BACKPORT-2.1] fix to handle null value in array type column for postgres.
- [\[SPARK-19267\]](#)[SS] Fix a race condition when stopping StateStore
- [\[SPARK-19168\]](#)[STRUCTURED STREAMING] StateStore should be aborted upon error
- [\[SPARK-19065\]](#)[SQL] Don't inherit expression id in dropDuplicates
- [\[SPARK-18905\]](#)[STREAMING] Fix the issue of removing a failed jobset from JobScheduler.jobSets
- [\[SPARK-17237\]](#)[SQL] Remove backticks in a pivot result schema
- [\[SPARK-19140\]](#)[SS] Allow update mode for non-aggregation streaming queries
- [\[SPARK-19137\]](#)[SQL] Fix `withSQLConf` to reset `OptionalConfigEntry` correctly
- [\[SPARK-19765\]](#)[SPARK-18549][SPARK-19093][SPARK-19736][BACKPORT-2.1][SQL] Backport Three Cache-related PRs to Spark 2.1
- [\[SPARK-18877\]](#)[SQL][BACKPORT-2.1] CSVInferSchema.inferField` on DecimalType should find a common type with `typeSoFar`
- [\[SPARK-18837\]](#)[WEBUI] Very long stage descriptions do not wrap in the UI
- [\[SPARK-18972\]](#)[CORE] Fix the netty thread names for RPC
- [\[SPARK-18985\]](#)[SS] Add missing @InterfaceStability.Evolving for Structured Streaming APIs
- [\[SPARK-18973\]](#)[SQL] Remove SortPartitions and RedistributeData
- [\[SPARK-18947\]](#)[SQL] SQLContext.tableNames should not call Catalog.listTables
- [\[SPARK-18927\]](#)[SS] MemorySink for StructuredStreaming can't recover from checkpoint if location is provided in SessionConf
- [\[SPARK-18899\]](#)[SPARK-18912][SPARK-18913][SQL] refactor the error checking when append data to an existing table
- [\[SPARK-18921\]](#)[SQL] check database existence with Hive.databaseExists instead of getDatabase
- [\[SPARK-18108\]](#)[SQL] Fix a schema inconsistent bug that makes a parquet reader fail to read data
- [\[SPARK-18892\]](#)[SQL] Alias percentile\_approx approx\_percentile
- [\[SPARK-18555\]](#)[MINOR][SQL] Fix the @since tag when backporting from 2.2 branch into 2.1 branch
- [\[SPARK-18555\]](#)[SQL] DataFrameNaFunctions.fill miss up original values in long integers

- [\[SPARK-18535\]](#)[SPARK-19720][CORE][BACKPORT-2.1] Redact sensitive information
- [\[SPARK-19506\]](#)[ML][PYTHON] Import warnings in pyspark.ml.util
- [\[SPARK-13669\]](#)[SPARK-20898][CORE] Improve the blacklist mechanism to handle external shuffle service unavailable situation
- [\[SPARK-13747\]](#)[CORE] Add ThreadUtils.awaitReady and disallow Await.ready
- [\[SPARK-13747\]](#)[CORE] Fix potential ThreadLocal leaks in RPC when using ForkJoinPool
- [\[SPARK-21522\]](#)[CORE] Fix flakiness in LauncherServerSuite.
- [\[SPARK-20904\]](#)[CORE] Don't report task failures to driver during shutdown.
- [\[SPARK-20393\]](#)[WEBUI] Strengthen Spark to prevent XSS vulnerabilities
- [\[SPARK-19146\]](#)[CORE] Drop more elements when stageData.taskData.size > retainedTasks
- [\[SPARK-20084\]](#)[CORE] Remove internal.metrics.updatedBlockStatuses from history files.
- [\[SPARK-18991\]](#)[CORE] Change ContextCleaner.referenceBuffer to use ConcurrentHashMap to make it faster
- [\[SPARK-19185\]](#)[DSTREAM] Make Kafka consumer cache configurable
- [\[SPARK-20756\]](#)[YARN] yarn-shuffle jar references unshaded guava
- [\[SPARK-20922\]](#)[CORE][HOTFIX] Don't use Java 8 lambdas in older branches.
- [\[SPARK-20922\]](#)[CORE] Add whitelist of classes that can be deserialized by the launcher.
- Add health aggregation to custom service descriptor.
- Update Spark2 service descriptor to not provide "cdh-plugin" since no one in CDH (yarn, in particular) uses Spark2 bits.
- Don't localize topology.py location.
- Relax compatibility to all C5 versions.

### Issues Fixed in CDS 2.1 Release 1

The following list includes issues fixed in CDS 2.1 Release 1. Test-only changes are omitted.

- Preview of: [\[SPARK-19554\]](#)[UI,YARN] Allow SHS URL to be used for tracking in YARN RM.
- [\[SPARK-16554\]](#)[CORE] Automatically Kill Executors and Nodes when they are Blacklisted
- [\[SPARK-16654\]](#)[CORE] Add UI coverage for Application Level Blacklisting
- [\[SPARK-8425\]](#)[CORE] Application Level Blacklisting
- [\[SPARK-18117\]](#)[CORE] Add test for TaskSetBlacklist
- [\[SPARK-18949\]](#)[SQL][BACKPORT-2.1] Add recoverPartitions API to Catalog
- [\[SPARK-19459\]](#)[SQL][BRANCH-2.1] Support for nested char/varchar fields in ORC
- [\[SPARK-19611\]](#)[SQL] Introduce configurable table schema inference
- [\[SPARK-19082\]](#)[SQL] Make ignoreCorruptFiles work for Parquet
- [\[SPARK-19766\]](#)[SQL] Constant alias columns in INNER JOIN should not be folded by FoldablePropagation rule
- [\[SPARK-19677\]](#)[SS] Committing a delta file atop an existing one should not fail on HDFS
- [\[SPARK-19038\]](#)[YARN] Avoid overwriting keytab configuration in yarn-client
- [\[SPARK-19617\]](#)[SS] Fix the race condition when starting and stopping a query quickly (branch-2.1)
- [\[SPARK-19599\]](#)[SS] Clean up HDFSMetadataLog
- [\[SPARK-19529\]](#) TransportClientFactory.createClient() shouldn't call awaitUninterruptibly()
- [\[SPARK-18717\]](#)[SQL] Make code generation for Scala Map work with immutable.Map also
- [\[SPARK-19512\]](#)[BACKPORT-2.1][SQL] codegen for compare structs fails #16852
- [\[SPARK-19481\]](#) [REPL] [MAVEN] Avoid to leak SparkContext in Signaling.cancelOnInterrupt
- [\[SPARK-18750\]](#)[YARN] Follow up: move test to correct directory in 2.1 branch.
- [\[SPARK-18750\]](#)[YARN] Avoid using "mapValues" when allocating containers.
- [\[SPARK-19268\]](#)[SS] Disallow adaptive query execution for streaming queries
- [\[SPARK-18850\]](#)[SS] Make StreamExecution and progress classes serializable
- [\[SPARK-18589\]](#)[SQL] Fix Python UDF accessing attributes from both side of join
- [\[SPARK-19314\]](#)[SS][CATALYST] Do not allow sort before aggregation in Structured Streaming plan
- [\[SPARK-19129\]](#)[SQL] SessionCatalog: Disallow empty part col values in partition spec
- [\[SPARK-19048\]](#)[SQL] Delete Partition Location when Dropping Managed Partitioned Tables in InMemoryCatalog



- [\[SPARK-19019\]](#) [PYTHON] Fix hijacked `collections.namedtuple` and port cloudpickle changes for PySpark to work with Python 3.6.0
- [\[SPARK-19092\]](#) [SQL][BACKPORT-2.1] Save() API of DataFrameWriter should not scan all the saved files #16481
- [\[SPARK-19120\]](#) Refresh Metadata Cache After Loading Hive Tables
- [\[SPARK-19180\]](#) [SQL] the offset of short should be 2 in OffHeapColumn
- [\[SPARK-19178\]](#) [SQL] convert string of large numbers to int should return null
- [\[SPARK-18687\]](#) [PYSPARK][SQL] Backward compatibility - creating a Dataframe on a new SQLContext object fails with a Derby error
- [\[SPARK-19055\]](#) [SQL][PYSPARK] Fix SparkSession initialization when SparkContext is stopped
- [\[SPARK-16845\]](#) [SQL] `GeneratedClass\$SpecificOrdering` grows beyond 64 KB
- [\[SPARK-18952\]](#) [BACKPORT] Regex strings not properly escaped in codegen for aggregations
- [\[SPARK-17807\]](#) [CORE] split test-tags into test-JAR
- [\[SPARK-18908\]](#) [SS] Creating StreamingQueryException should check if logicalPlan is created
- [\[SPARK-18528\]](#) [SQL] Fix a bug to initialise an iterator of aggregation buffer
- [\[SPARK-18234\]](#) [SS] Made update mode public
- [\[SPARK-18588\]](#) [SS][KAFKA] Create a new KafkaConsumer when error happens to fix the flaky test
- [\[SPARK-18894\]](#) [SS] Fix event time watermark delay threshold specified in months or years
- [\[SPARK-18281\]](#) [SQL] [PYSPARK] Remove timeout for reading data through socket for local iterator
- [\[SPARK-18761\]](#) [CORE] Introduce "task reaper" to oversee task killing in executors
- [\[SPARK-18928\]](#) Check TaskContext.isInterrupted() in FileScanRDD, JDBC RDD & UnsafeSorter
- [\[SPARK-18700\]](#) [SQL] Add StripedLock for each table's relation in cache
- [\[SPARK-18703\]](#) [SPARK-18675][SQL][BACKPORT-2.1] CTAS for hive serde table should work for all hive versions AND Drop Staging Directories and Data Files
- [\[SPARK-18827\]](#) [CORE] Fix cannot read broadcast on disk
- [\[SPARK-19520\]](#) [STREAMING] Do not encrypt data written to the WAL.
- [\[SPARK-19857\]](#) [YARN] Correctly calculate next credential update time.
- [\[SPARK-19626\]](#) [YARN] Using the correct config to set credentials update time
- Preview of: [\[SPARK-4105\]](#) retry the fetch or stage if shuffle block is corrupt
- [\[SPARK-19307\]](#) [PYSPARK] Make sure user conf is propagated to SparkContext.

## Issues Fixed in CDS 2.0 Release 2

- [\[SPARK-4563\]](#) [CORE] Allow driver to advertise a different network address.
- [\[SPARK-18993\]](#) Unable to build/compile Spark in IntelliJ due to missing Scala deps in spark-tags
- [\[SPARK-19314\]](#) Do not allow sort before aggregation in Structured Streaming plan
- [\[SPARK-18762\]](#) Web UI should be http:4040 instead of https:4040
- [\[SPARK-18745\]](#) java.lang.IndexOutOfBoundsException running query 68 Spark SQL on (100TB)
- [\[SPARK-18703\]](#) Insertion/CTAS against Hive Tables: Staging Directories and Data Files Not Dropped Until Normal Termination of JVM
- [\[SPARK-18091\]](#) Deep if expressions cause Generated SpecificUnsafeProjection code to exceed JVM code size limit

## Issues Fixed in CDS 2.0 Release 1

- [\[SPARK-4563\]](#) [CORE] Allow driver to advertise a different network address.
- [\[SPARK-18685\]](#) [TESTS] Fix URI and release resources after opening in tests at ExecutorClassLoaderSuite
- [\[SPARK-18677\]](#) Fix parsing ['key'] in JSON path expressions.
- [\[SPARK-18617\]](#) [SPARK-18560][TESTS] Fix flaky test: StreamingContextSuite. Receiver data should be deserialized properly
- [\[SPARK-18617\]](#) [SPARK-18560][TEST] Fix flaky test: StreamingContextSuite. Receiver data should be deserialized properly
- [\[SPARK-18274\]](#) [ML][PYSPARK] Memory leak in PySpark JavaWrapper
- [\[SPARK-18674\]](#) [SQL] improve the error message of using join

- [\[SPARK-18617\]](#)[CORE][STREAMING] Close "kryo auto pick" feature for Spark Streaming
- [\[SPARK-17843\]](#)[WEB UI] Indicate event logs pending for processing on h...
- [\[SPARK-17783\]](#)[SQL][BACKPORT-2.0] Hide Credentials in CREATE and DESC FORMATTED/EXTENDED a PERSISTENT/TEMP Table for JDBC
- [\[SPARK-18640\]](#) Add synchronization to TaskScheduler.runningTasksByExecutors
- [\[SPARK-18553\]](#)[CORE] Fix leak of TaskSetManager following executor loss
- [\[SPARK-18597\]](#)[SQL] Do not push-down join conditions to the left side of a Left Anti join [BRANCH-2.0]
- [\[SPARK-18118\]](#)[SQL] fix a compilation error due to nested JavaBeans
- [\[SPARK-17251\]](#)[SQL] Improve `OuterReference` to be `NamedExpression`
- [\[SPARK-18436\]](#)[SQL] isin causing SQL syntax error with JDBC
- [\[SPARK-18519\]](#)[SQL][BRANCH-2.0] map type can not be used in EqualTo
- [\[SPARK-18053\]](#)[SQL] compare unsafe and safe complex-type values correctly
- [\[SPARK-18504\]](#)[SQL] Scalar subquery with extra group by columns returning incorrect result
- [\[SPARK-18477\]](#)[SS] Enable interrupts for HDFS in HDFSMetadataLog
- [\[SPARK-18546\]](#)[CORE] Fix merging shuffle spills when using encryption.
- [\[SPARK-18547\]](#)[CORE] Propagate I/O encryption key when executors register.
- [\[SPARK-16625\]](#)[SQL] General data types to be mapped to Oracle
- [\[SPARK-18462\]](#) Fix ClassCastException in SparkListenerDriverAccumUpdates event
- [\[SPARK-18459\]](#)[SPARK-18460][STRUCTUREDSTREAMING] Rename triggerId to batchId and add triggerDetails to json in StreamingQueryStatus (for branch-2.0)
- [\[SPARK-18430\]](#)[SQL][BACKPORT-2.0] Fixed Exception Messages when Hitting an Invocation Exception of Function Lookup
- [\[SPARK-18400\]](#)[STREAMING] NPE when resharding Kinesis Stream
- [\[SPARK-18300\]](#)[SQL] Do not apply foldable propagation with expand as a child [BRANCH-2.0]
- [\[SPARK-18337\]](#) Complete mode memory sinks should be able to recover from checkpoints
- [\[SPARK-16808\]](#)[CORE] History Server main page does not honor APPLICATION\_WEB\_PROXY\_BASE
- [\[SPARK-17348\]](#)[SQL] Incorrect results from subquery transformation
- [\[SPARK-18416\]](#)[STRUCTURED STREAMING] Fixed temp file leak in state store
- [\[SPARK-18432\]](#)[DOC] Changed HDFS default block size from 64MB to 128MB
- [\[SPARK-18010\]](#)[CORE] Reduce work performed for building up the application list for the History Server app list UI page
- [\[SPARK-18382\]](#)[WEBUI] "run at null:-1" in UI when no file/line info in call site info
- [\[SPARK-18426\]](#)[STRUCTURED STREAMING] Python Documentation Fix for Structured Streaming Programming Guide
- [\[SPARK-17982\]](#)[SQL][BACKPORT-2.0] SQLBuilder should wrap the generated SQL with parenthesis for LIMIT
- [\[SPARK-18387\]](#)[SQL] Add serialization to checkEvaluation.
- [\[SPARK-18368\]](#)[SQL] Fix regexp replace when serialized
- [\[SPARK-18342\]](#) Make rename failures fatal in HDFSBackedStateStore
- [\[SPARK-18280\]](#)[CORE] Fix potential deadlock in `StandaloneSchedulerBackend.dead`
- [\[SPARK-17703\]](#)[SQL][BACKPORT-2.0] Add unnamed version of addReferenceObj for minor objects.
- [\[SPARK-18137\]](#)[SQL] Fix RewriteDistinctAggregates UnresolvedException when a UDAF has a foldable TypeCheck
- [\[SPARK-18283\]](#)[STRUCTURED STREAMING][KAFKA] Added test to check whether default starting offset in latest
- [\[SPARK-18125\]](#)[SQL][BRANCH-2.0] Fix a compilation error in codegen due to splitExpression
- [\[SPARK-17849\]](#)[SQL] Fix NPE problem when using grouping sets
- [\[SPARK-17693\]](#)[SQL][BACKPORT-2.0] Fixed Insert Failure To Data Source Tables when the Schema has the Comment Field
- [\[SPARK-17981\]](#)[SPARK-17957][SQL][BACKPORT-2.0] Fix Incorrect Nullability Setting to False in FilterExec
- [\[SPARK-18189\]](#)[SQL][FOLLOWUP] Move test from RepI Suite to prevent java.lang.ClassCircularityError
- [\[SPARK-17337\]](#)[SPARK-16804][SQL][BRANCH-2.0] Backport subquery related PRs
- [\[SPARK-18200\]](#)[GRAPHX][FOLLOW-UP] Support zero as an initial capacity in OpenHashSet
- [\[SPARK-18200\]](#)[GRAPHX] Support zero as an initial capacity in OpenHashSet

- [\[SPARK-18111\]](#)[SQL] Wrong approximate quantile answer when multiple records have the minimum value(for branch 2.0)
- [\[SPARK-18160\]](#)[CORE][YARN] spark.files & spark.jars should not be passed to driver in yarn mode
- [\[SPARK-16796\]](#)[WEB UI] Mask spark.authenticate.secret on Spark environ...
- [\[SPARK-18133\]](#)[BRANCH-2.0][EXAMPLES][ML] Python ML Pipeline Examl...
- [\[SPARK-18144\]](#)[SQL] logging StreamingQueryListener\$QueryStartedEvent
- [\[SPARK-18114\]](#)[HOTFIX] Fix line-too-long style error from backport of SPARK-18114
- [\[SPARK-18148\]](#)[SQL] Misleading Error Message for Aggregation Without Window/GroupBy
- [\[SPARK-18189\]](#)[SQL] Fix serialization issue in KeyValueGroupedDataset
- [\[SPARK-18114\]](#)[MESOS] Fix mesos cluster scheduler generage command option error
- [\[SPARK-18030\]](#)[TESTS] Fix flaky FileStreamSourceSuite by not deleting the files
- [\[SPARK-18143\]](#)[SQL] Ignore Structured Streaming event logs to avoid breaking history server (branch 2.0)
- [\[SPARK-16312\]](#)[FOLLOW-UP][STREAMING][KAFKA][DOC] Add java code snippet for Kafka 0.10 integration doc
- [\[SPARK-18164\]](#)[SQL] ForeachSink should fail the Spark job if `process` throws exception
- [\[SPARK-16963\]](#)[SQL] Fix test "StreamExecution metadata garbage collection"
- [\[SPARK-17813\]](#)[SQL][KAFKA] Maximum data per trigger
- [\[SPARK-18132\]](#) Fix checkstyle
- [\[SPARK-18009\]](#)[SQL] Fix ClassCastException while calling toLocalIterator() on dataframe produced by RunnableCommand
- [\[SPARK-16963\]](#)[STREAMING][SQL] Changes to Source trait and related implementation classes
- [\[SPARK-13747\]](#)[SQL] Fix concurrent executions in ForkJoinPool for SQL (branch 2.0)
- [\[SPARK-18063\]](#)[SQL] Failed to infer constraints over multiple aliases
- [\[SPARK-16304\]](#) LinkageError should not crash Spark executor
- [\[SPARK-17733\]](#)[SQL] InferFiltersFromConstraints rule never terminates for query
- [\[SPARK-18022\]](#)[SQL] java.lang.NullPointerException instead of real exception when saving DF to MySQL
- [\[SPARK-16988\]](#)[SPARK SHELL] spark history server log needs to be fixed to show https url when ssl is enabled
- [\[SPARK-18070\]](#)[SQL] binary operator should not consider nullability when comparing input types
- [\[SPARK-17624\]](#)[SQL][STREAMING][TEST] Fixed flaky StateStoreSuite.maintenance
- [\[SPARK-18044\]](#)[STREAMING] FileStreamSource should not infer partitions in every batch
- [\[SPARK-17153\]](#)[SQL] Should read partition data when reading new files in filestream without globbing
- [\[SPARK-18093\]](#)[SQL] Fix default value test in SQLConfSuite to work rega...
- [\[SPARK-17810\]](#)[SQL] Default spark.sql.warehouse.dir is relative to local FS but can resolve as HDFS path
- [\[SPARK-18058\]](#)[SQL] [BRANCH-2.0]Comparing column types ignoring Nullability in Union and SetOperation
- [\[SPARK-17123\]](#)[SQL][BRANCH-2.0] Use type-widened encoder for DataFrame for set operations
- [\[SPARK-17698\]](#)[SQL] Join predicates should not contain filter clauses
- [\[SPARK-17986\]](#)[ML] SQLTransformer should remove temporary tables
- [\[SPARK-16606\]](#)[MINOR] Tiny follow-up to , to correct more instances of the same log message typo
- [\[SPARK-17853\]](#)[STREAMING][KAFKA][DOC] make it clear that reusing group.id is bad
- [\[SPARK-16312\]](#)[STREAMING][KAFKA][DOC] Doc for Kafka 0.10 integration
- [\[SPARK-17812\]](#)[SQL][KAFKA] Assign and specific startingOffsets for structured stream
- [\[SPARK-17929\]](#)[CORE] Fix deadlock when CoarseGrainedSchedulerBackend reset
- [\[SPARK-17926\]](#)[SQL][STREAMING] Added json for statuses
- [\[SPARK-17811\]](#) SparkR cannot parallelize data.frame with NA or NULL in Date columns
- [\[SPARK-18034\]](#) Upgrade to MiMa 0.1.11 to fix flakiness
- [\[SPARK-17999\]](#)[KAFKA][SQL] Add getPreferredLocations for KafkaSourceRDD
- [\[SPARK-18003\]](#)[SPARK CORE] Fix bug of RDD zipWithIndex & zipWithUniqueld index value overflowing
- [\[SPARK-17989\]](#)[SQL] Check ascendingOrder type in sort\_array function rather than throwing ClassCastException
- [\[SPARK-17675\]](#)[CORE] Expand Blacklist for TaskSets
- [\[SPARK-17623\]](#)[CORE] Clarify type of TaskEndReason with a failed task.
- [\[SPARK-17304\]](#) Fix perf. issue caused by TaskSetManager.abortIfCompletelyBlacklisted

- [\[SPARK-15865\]](#)[CORE] Blacklist should not result in job hanging with less than 4 executors
- [\[SPARK-15783\]](#)[CORE] Fix Flakiness in BlacklistIntegrationSuite
- [\[SPARK-15783\]](#)[CORE] still some flakiness in these blacklist tests so ignore for now
- [\[SPARK-15714\]](#)[CORE] Fix flaky o.a.s.scheduler.BlacklistIntegrationSuite
- [\[SPARK-10372\]](#) [CORE] basic test framework for entire spark scheduler
- [\[SPARK-16106\]](#)[CORE] TaskSchedulerImpl should properly track executors added to existing hosts
- [\[SPARK-18001\]](#)[DOCUMENT] fix broke link to SparkDataFrame
- [\[SPARK-17711\]](#)[TEST-HADOOP2.2] Fix hadoop2.2 compilation error
- [\[SPARK-17731\]](#)[SQL][STREAMING][FOLLOWUP] Refactored StreamingQueryListener APIs for branch-2.0
- [\[SPARK-17841\]](#)[STREAMING][KAFKA] drain commitQueue
- [\[SPARK-17711\]](#) Compress rolled executor log
- [\[SPARK-17751\]](#)[SQL][BACKPORT-2.0] Remove spark.sql.eagerAnalysis and Output the Plan if Existed in AnalysisException
- [\[SPARK-17731\]](#)[SQL][STREAMING] Metrics for structured streaming for branch-2.0
- [\[SPARK-17892\]](#)[SQL][2.0] Do Not Optimize Query in CTAS More Than Once #15048
- [\[SPARK-17819\]](#)[SQL][BRANCH-2.0] Support default database in connection URIs for Spark Thrift Server
- [\[SPARK-17953\]](#)[DOCUMENTATION] Fix typo in SparkSession scaladoc
- [\[SPARK-17863\]](#)[SQL] should not add column into Distinct
- [\[SPARK-17387\]](#)[PYSPARK] Creating SparkContext() from python without spark-submit ignores user conf
- [\[SPARK-17834\]](#)[SQL] Fetch the earliest offsets manually in KafkaSource instead of counting on KafkaConsumer
- [\[SPARK-17876\]](#) Write StructuredStreaming WAL to a stream instead of materializing all at once
- [\[SPARK-16827\]](#)[BRANCH-2.0] Avoid reporting spill metrics as shuffle metrics
- [\[SPARK-17782\]](#)[STREAMING][KAFKA] alternative eliminate race condition of poll twice
- [\[SPARK-17790\]](#)[SPARKR] Support for parallelizing R data.frame larger than 2GB
- [\[SPARK-17884\]](#)[SQL] To resolve Null pointer exception when casting from empty string to interval type.
- [\[SPARK-17808\]](#)[PYSPARK] Upgraded version of Pyrolite to 4.13
- [\[SPARK-17880\]](#)[DOC] The url linking to `AccumulatorV2` in the document is incorrect.
- [\[SPARK-17816\]](#)[CORE][BRANCH-2.0] Fix ConcurrentModificationException issue in BlockStatusesAccumulator
- [\[SPARK-17346\]](#)[SQL][TESTS] Fix the flaky topic deletion in KafkaSourceStressSuite
- [\[SPARK-17738\]](#)[TEST] Fix flaky test in ColumnTypeSuite
- [\[SPARK-17417\]](#)[CORE] Fix # of partitions for Reliable RDD checkpointing
- [\[SPARK-17832\]](#)[SQL] TableIdentifier.quotedString creates un-parseable names when name contains a backtick
- [\[SPARK-17806\]](#) [SQL] fix bug in join key rewritten in HashJoin
- [\[SPARK-17782\]](#)[STREAMING][BUILD] Add Kafka 0.10 project to build modules
- [\[SPARK-17346\]](#)[SQL][TEST-MAVEN] Add Kafka source for Structured Streaming (branch 2.0)
- [\[SPARK-17707\]](#)[WEBUI] Web UI prevents spark-submit application to be finished
- [\[SPARK-17805\]](#)[PYSPARK] Fix in sqlContext.read.text when pass in list of paths
- [\[SPARK-17612\]](#)[SQL][BRANCH-2.0] Support `DESCRIBE table PARTITION` SQL syntax
- [\[SPARK-17792\]](#)[ML] L-BFGS solver for linear regression does not accept general numeric label column types
- [\[SPARK-17750\]](#)[SQL][BACKPORT-2.0] Fix CREATE VIEW with INTERVAL arithmetic
- [\[SPARK-17803\]](#)[TESTS] Upgrade docker-client dependency
- [\[SPARK-17780\]](#)[SQL] Report Throwable to user in StreamExecution
- [\[SPARK-17798\]](#)[SQL] Remove redundant Experimental annotations in sql.streaming
- [\[SPARK-17643\]](#) Remove comparable requirement from Offset (backport for branch-2.0)
- [\[SPARK-17758\]](#)[SQL] Last returns wrong result in case of empty partition
- [\[SPARK-17778\]](#)[TESTS] Mock SparkContext to reduce memory usage of BlockManagerSuite
- [\[SPARK-17773\]](#)[BRANCH-2.0] Input/Output] Add VoidObjectInspector
- [\[SPARK-17549\]](#)[SQL] Only collect table size stat in driver for cached relation.
- [\[SPARK-17559\]](#)[MLLIB] persist edges if their storage level is non in PeriodicGraphCheckpoint
- [\[SPARK-17112\]](#)[SQL] "select null" via JDBC triggers IllegalArgument Exception in Thriftserver

- [\[SPARK-17753\]](#)[SQL] Allow a complex expression as the input a value based case statement
- [\[SPARK-17587\]](#)[PYTHON][MLLIB] SparseVector \_\_getitem\_\_ should follow \_\_getitem\_\_ contract
- [\[SPARK-17736\]](#)[DOCUMENTATION][SPARKR] Update R README for rmarkdown,...
- [\[SPARK-17721\]](#)[MLLIB][ML] Fix for multiplying transposed SparseMatrix with SparseVector
- [\[SPARK-17672\]](#) Spark 2.0 history server web Ui takes too long for a single application
- [\[SPARK-17712\]](#)[SQL] Fix invalid pushdown of data-independent filters beneath aggregates
- [\[SPARK-16343\]](#)[SQL] Improve the PushDownPredicate rule to pushdown predicates correctly in non-deterministic condition.
- [\[SPARK-17641\]](#)[SQL] Collect\_list/Collect\_set should not collect null values.
- [\[SPARK-17673\]](#)[SQL] Incorrect exchange reuse with RowDataSourceScan (backport)
- [\[SPARK-17644\]](#)[CORE] Do not add failedStages when abortStage for fetch failure
- [\[SPARK-17666\]](#) Ensure that RecordReaders are closed by data source file scans (backport)
- [\[SPARK-17056\]](#)[CORE] Fix a wrong assert regarding unroll memory in MemoryStore
- [\[SPARK-17618\]](#) Guard against invalid comparisons between UnsafeRow and other formats
- [\[SPARK-17652\]](#) Fix confusing exception message while reserving capacity
- [\[SPARK-17649\]](#)[CORE] Log how many Spark events got dropped in LiveListenerBus
- [\[SPARK-17650\]](#) malformed url's throw exceptions before bricking Executors
- [\[SPARK-10835\]](#)[ML] Word2Vec should accept non-null string array, in addition to existing null string array
- [\[SPARK-15703\]](#)[SCHEDULER][CORE][WEBUI] Make ListenerBus event queue size configurable (branch 2.0)
- [\[SPARK-4563\]](#)[CORE] Allow driver to advertise a different network address.
- [\[SPARK-17577\]](#)[CORE][2.0 BACKPORT] Update SparkContext.addFile to make it work well on Windows
- [\[SPARK-17210\]](#)[SPARKR] sparkr.zip is not distributed to executors when running sparkr in RStudio
- [\[SPARK-17640\]](#)[SQL] Avoid using -1 as the default batchSize for FileStreamSource.FileEntry
- [\[SPARK-16240\]](#)[ML] ML persistence backward compatibility for LDA - 2.0 backport
- [\[SPARK-17502\]](#)[17609][SQL][BACKPORT][2.0] Fix Multiple Bugs in DDL Statements on Temporary Views
- [\[SPARK-17599\]](#)[\[SPARK-17569\]](#) Backport and to Spark 2.0 branch
- [\[SPARK-17616\]](#)[SQL] Support a single distinct aggregate combined with a non-partial aggregate
- [\[SPARK-17638\]](#)[STREAMING] Stop JVM StreamingContext when the Python process is dead
- [\[SPARK-17613\]](#) S3A base paths with no '/' at the end return empty DataFrames
- [\[SPARK-17421\]](#)[DOCS] Documenting the current treatment of MAVEN\_OPTS.
- [\[SPARK-17494\]](#)[SQL] changePrecision() on compact decimal should respect rounding mode
- [\[SPARK-17627\]](#) Mark Streaming Providers Experimental
- [\[SPARK-17512\]](#)[CORE] Avoid formatting to python path for yarn and mesos cluster mode
- [\[SPARK-17418\]](#) Prevent kinesis-asl-assembly artifacts from being published
- [\[SPARK-17617\]](#)[SQL] Remainder(%) expression.eval returns incorrect result on double value
- [\[SPARK-15698\]](#)[SQL][STREAMING] Add the ability to remove the old MetadataLog in FileStreamSource (branch-2.0)
- [\[SPARK-17051\]](#)[SQL] we should use hadoopConf in InsertIntoHiveTable
- [\[SPARK-17513\]](#)[SQL] Make StreamExecution garbage-collect its metadata
- [\[SPARK-17160\]](#) Properly escape field names in code-generated error messages
- [\[SPARK-17100\]](#) [SQL] fix Python udf in filter on top of outer join
- [\[SPARK-16439\]](#) [SQL] bring back the separator in SQL UI
- [\[SPARK-17611\]](#)[yarn][test] Make shuffle service test really test auth.
- [\[SPARK-17433\]](#) YarnShuffleService doesn't handle moving credentials levelDb
- [\[SPARK-17438\]](#)[WEBUI] Show Application.executorLimit in the application page
- [\[SPARK-17473\]](#)[SQL] fixing docker integration tests error due to different versions of jars.
- [\[SPARK-17589\]](#)[TEST][2.0] Fix test case `create external table` in MetastoreDataSourcesSuite
- [\[SPARK-17297\]](#)[DOCS] Clarify window/slide duration as absolute time, not relative to a calendar
- [\[SPARK-17571\]](#)[SQL] AssertOnQuery.condition should always return Boolean value
- [\[SPARK-16462\]](#)[\[SPARK-16460\]](#)[\[SPARK-15144\]](#)[SQL] Make CSV cast null values properly
- [\[SPARK-17586\]](#)[BUILD] Do not call static member via instance reference

- [\[SPARK-17546\]](#)[DEPLOY] start-\* scripts should use hostname -f
- [\[SPARK-17541\]](#)[SQL] fix some DDL bugs about table management when same-name temp view exists
- [\[SPARK-17480\]](#)[SQL][FOLLOWUP] Fix more instances which calls List.length/size which is O(n)
- [\[SPARK-17491\]](#) Close serialization stream to fix wrong answer bug in putIteratorAsBytes()
- [\[SPARK-17575\]](#)[DOCS] Remove extra table tags in configuration document
- [\[SPARK-17548\]](#)[MLLIB] Word2VecModel.findSynonyms no longer spuriously rejects the best match when invoked with a vector
- [\[SPARK-17561\]](#)[DOCS] DataFrameWriter documentation formatting problems
- [\[SPARK-17567\]](#)[DOCS] Use valid url to Spark RDD paper
- [\[SPARK-17558\]](#) Bump Hadoop 2.7 version from 2.7.2 to 2.7.3
- [\[SPARK-17484\]](#) Prevent invalid block locations from being reported after put() exceptions
- [\[SPARK-17364\]](#)[SQL] Antlr lexer wrongly treats full qualified identifier as a decimal number token when parsing SQL string
- [\[SPARK-17483\]](#) Refactoring in BlockManager status reporting and block removal
- [\[SPARK-17114\]](#)[SQL] Fix aggregates grouped by literals with empty input
- [\[SPARK-17547\]](#) Ensure temp shuffle data file is cleaned up after error
- [\[SPARK-17521\]](#) Error when I use sparkContext.makeRDD(Seq())
- [\[SPARK-17465\]](#)[SPARK CORE] Inappropriate memory management in `org.apache.spark.storage.MemoryStore` may lead to memory leak
- [\[SPARK-17463\]](#)[CORE] Make CollectionAccumulator and SetAccumulator's value can be read thread-safely
- [\[SPARK-17511\]](#) Yarn Dynamic Allocation: Avoid marking released container as Failed
- [\[SPARK-17514\]](#) df.take(1) and df.limit(1).collect() should perform the same in Python
- [\[SPARK-17445\]](#)[DOCS] Reference an ASF page as the main place to find third-party packages
- [\[SPARK-16711\]](#) YarnShuffleService doesn't re-init properly on YARN rolling upgrade
- [\[SPARK-15074\]](#)[SHUFFLE] Cache shuffle index file to speedup shuffle fetch
- [\[SPARK-17480\]](#)[SQL] Improve performance by removing or caching List.length which is O(n)
- [\[SPARK-17525\]](#)[PYTHON] Remove SparkContext.clearFiles() from the PySpark API as it was removed from the Scala API prior to Spark 2.0.0
- [\[SPARK-17531\]](#) Don't initialize Hive Listeners for the Execution Client
- [\[SPARK-17515\]](#) CollectLimit.execute() should perform per-partition limits
- [\[SPARK-17474\]](#) [SQL] fix python udf in TakeOrderedAndProjectExec
- [\[SPARK-17485\]](#) Prevent failed remote reads of cached blocks from failing entire job
- [\[SPARK-14818\]](#) Post-2.0 MiMa exclusion and build changes
- [\[SPARK-17503\]](#)[CORE] Fix memory leak in Memory store when unable to cache the whole RDD in memory
- [\[SPARK-17486\]](#) Remove unused TaskMetricsUIData.updatedBlockStatuses field
- [\[SPARK-17336\]](#)[PYSARK] Fix appending multiple times to PYTHONPATH from spark-config.sh
- [\[SPARK-17439\]](#)[SQL] Fixing compression issues with approximate quantiles and adding more tests
- [\[SPARK-17396\]](#)[CORE] Share the task support between UnionRDD instances.
- [\[SPARK-17354\]](#) [SQL] Partitioning by dates/timestamps should work with Parquet vectorized reader
- [\[SPARK-17456\]](#)[CORE] Utility for parsing Spark versions
- [\[SPARK-17339\]](#)[CORE][BRANCH-2.0] Do not use path to get a filesystem in hadoopFile and newHadoopFile APIs
- [\[SPARK-16533\]](#)[CORE] - backport driver deadlock fix to 2.0
- [\[SPARK-17370\]](#) Shuffle service files not invalidated when a slave is lost
- [\[SPARK-17296\]](#)[SQL] Simplify parser join processing [BACKPORT 2.0]
- [\[SPARK-17372\]](#)[SQL][STREAMING] Avoid serialization issues by using Arrays to save file names in FileStreamSource
- [\[SPARK-17279\]](#)[SQL] better error message for exceptions during ScalaUDF execution
- [\[SPARK-17316\]](#)[CORE] Fix the 'ask' type parameter in 'removeExecutor'
- [\[SPARK-17110\]](#) Fix StreamCorruptionException in BlockManager.getRemoteValues()
- [\[SPARK-17299\]](#) TRIM/LTRIM/RTRIM should not strips characters other than spaces

- [\[SPARK-16334\]](#) [BACKPORT] Reusing same dictionary column for decoding consecutive row groups shouldn't throw an error
- [\[SPARK-16922\]](#) [\[SPARK-17211\]](#) [SQL] make the address of values portable in LongToUnsafeRowMap
- [\[SPARK-17356\]](#)[SQL] Fix out of memory issue when generating JSON for TreeNode
- [\[SPARK-17369\]](#)[SQL][2.0] MetastoreRelation toJSON throws AssertionError due to missing otherCopyArgs
- [\[SPARK-17358\]](#)[SQL] Cached table(parquet/orc) should be shard between beelines
- [\[SPARK-17353\]](#)[\[SPARK-16943\]](#)[\[SPARK-16942\]](#)[BACKPORT-2.0][SQL] Fix multiple bugs in CREATE TABLE LIKE command
- [\[SPARK-17391\]](#)[TEST][2.0] Fix Two Test Failures After Backport
- [\[SPARK-17335\]](#)[SQL] Fix ArrayType and MapType CatalogString.
- [\[SPARK-16663\]](#)[SQL] desc table should be consistent between data source and hive serde tables
- [\[SPARK-16959\]](#)[SQL] Rebuild Table Comment when Retrieving Metadata from Hive Metastore
- [\[SPARK-17347\]](#)[SQL][EXAMPLES] Encoder in Dataset example has incorrect type
- [\[SPARK-17230\]](#) [SQL] Should not pass optimized query into QueryExecution in DataFrameWriter
- [\[SPARK-17261\]](#) [PYSARK] Using HiveContext after re-creating SparkContext in Spark 2.0 throws "Java.lang.IllegalStateException: Cannot call methods on a stopped sparkContext"
- [\[SPARK-16935\]](#)[SQL] Verification of Function-related ExternalCatalog APIs
- [\[SPARK-17352\]](#)[WEBUI] Executor computing time can be negative-number because of calculation error
- [\[SPARK-17342\]](#)[WEBUI] Style of event timeline is broken
- [\[SPARK-17355\]](#) Workaround for HIVE-14684 / HiveResultSetMetaData.isSigned exception
- [\[SPARK-16926\]](#) [SQL] Remove partition columns from partition metadata.
- [\[SPARK-17271\]](#)[SQL] Planner adds un-necessary Sort even if child orde...
- [\[SPARK-17318\]](#)[TESTS] Fix RepSuite replicating blocks of object with class defined in repl again
- [\[SPARK-17180\]](#)[\[SPARK-17309\]](#)[\[SPARK-17323\]](#)[SQL][2.0] create AlterViewAsCommand to handle ALTER VIEW AS
- [\[SPARK-17316\]](#)[TESTS] Fix MesosCoarseGrainedSchedulerBackendSuite
- [\[SPARK-17316\]](#)[CORE] Make CoarseGrainedSchedulerBackend.removeExecutor non-blocking
- [\[SPARK-17243\]](#)[WEB UI] Spark 2.0 History Server won't load with very large application history
- [\[SPARK-17318\]](#)[TESTS] Fix RepSuite replicating blocks of object with class defined in repl
- [\[SPARK-17264\]](#)[SQL] DataStreamWriter should document that it only supports Parquet for now
- [\[SPARK-17301\]](#)[SQL] Remove unused classTag field from AtomicType base class
- [\[SPARK-17063\]](#) [SQL] Improve performance of MSCK REPAIR TABLE with Hive metastore
- [\[SPARK-16216\]](#)[SQL][FOLLOWUP][BRANCH-2.0] Bacoport enabling timestamp type tests for JSON and verify all unsupported types in CSV
- [\[SPARK-17216\]](#)[UI] fix event timeline bars length
- [ML][MLLIB] The require condition and message doesn't match in SparseMatrix.
- [\[SPARK-15382\]](#)[SQL] Fix a bug in sampling with replacement
- [\[SPARK-17274\]](#)[SQL] Move join optimizer rules into a separate file
- [\[SPARK-17270\]](#)[SQL] Move object optimization rules into its own file (branch-2.0)
- [\[SPARK-17269\]](#)[SQL] Move finish analysis optimization stage into its own file
- [\[SPARK-17244\]](#) Catalyst should not pushdown non-deterministic join conditions
- [\[SPARK-17235\]](#)[SQL] Support purging of old logs in MetadataLog
- [\[SPARK-17246\]](#)[SQL] Add BigDecimal literal
- [\[SPARK-17165\]](#)[SQL] FileStreamSource should not track the list of seen files indefinitely
- [\[SPARK-17242\]](#)[DOCUMENT] Update links of external dstream projects
- [\[SPARK-17231\]](#)[CORE] Avoid building debug or trace log messages unless the respective log level is enabled
- [\[SPARK-17205\]](#) Literal.sql should handle Infinity and NaN
- [\[SPARK-15083\]](#)[WEB UI] History Server can OOM due to unlimited TaskUIData
- [\[SPARK-16700\]](#)[PYSARK][SQL] create DataFrame from dict/Row with schema
- [\[SPARK-17167\]](#)[2.0][SQL] Issue Exceptions when Analyze Table on In-Memory Cataloged Tables
- [\[SPARK-16991\]](#)[\[SPARK-17099\]](#)[\[SPARK-17120\]](#)[SQL] Fix Outer Join Elimination when Filter's isNotNull Constraints Unable to Filter Out All Null-supplying Rows

- [\[SPARK-17061\]](#)[\[SPARK-17093\]](#)[SQL][BACKPORT] MapObjects should make copies of unsafe-backed data
- [\[SPARK-17193\]](#)[CORE] HadoopRDD NPE at DEBUG log level when getLocationInfo == null
- [\[SPARK-17228\]](#)[SQL] Not infer/propagate non-deterministic constraints
- [\[SPARK-16216\]](#)[SQL][BRANCH-2.0] Backport Read/write dateFormat/timestampFormat options for CSV and JSON
- [\[SPARK-16781\]](#)[PYSPARK] java launched by PySpark as gateway may not be the same java used in the spark environment
- [\[SPARK-17086\]](#)[ML] Fix InvalidArgumentException issue in QuantileDiscretizer when some quantiles are duplicated
- [\[SPARK-17186\]](#)[SQL] remove catalog table type INDEX
- [\[SPARK-17194\]](#) Use single quotes when generating SQL for string literals
- [\[SPARK-13286\]](#) [SQL] add the next expression of SQLException as cause
- [\[SPARK-17182\]](#)[SQL] Mark Collect as non-deterministic
- [\[SPARK-16550\]](#)[\[SPARK-17042\]](#)[CORE] Certain classes fail to deserialize in block manager replication
- [\[SPARK-17162\]](#) Range does not support SQL generation
- [\[SPARK-16320\]](#)[DOC] Document G1 heap region's effect on spark 2.0 vs 1.6
- [\[SPARK-17085\]](#)[STREAMING][DOCUMENTATION AND ACTUAL CODE DIFFERS - UNSUPPORTED OPERATIONS]
- [\[SPARK-17115\]](#)[SQL] decrease the threshold when split expressions
- [\[SPARK-17098\]](#)[SQL] Fix `NullPropagation` optimizer to handle `COUNT(NULL) OVER` correctly
- [\[SPARK-12666\]](#)[CORE] SparkSubmit packages fix for when 'default' conf doesn't exist in dependent module
- [\[SPARK-17124\]](#)[SQL] RelationalGroupedDataset.agg should preserve order and allow multiple aggregates per column
- [\[SPARK-17104\]](#)[SQL] LogicalRelation.newInstance should follow the semantics of MultiInstanceRelation
- [\[SPARK-17150\]](#)[SQL] Support SQL generation for inline tables
- [\[SPARK-17158\]](#)[SQL] Change error message for out of range numeric literals
- [\[SPARK-17149\]](#)[SQL] array.sql for testing array related functions
- [\[SPARK-17113\]](#) [SHUFFLE] Job failure due to Executor OOM in offheap mode
- [\[SPARK-16686\]](#)[SQL] Remove PushProjectThroughSample since it is handled by ColumnPruning
- [\[SPARK-11227\]](#)[CORE] UnknownHostException can be thrown when NameNode HA is enabled.
- [\[SPARK-16994\]](#)[SQL] Whitelist operators for predicate pushdown
- [\[SPARK-16961\]](#)[CORE] Fixed off-by-one error that biased randomizeInPlace
- [\[SPARK-16947\]](#)[SQL] Support type coercion and foldable expression for inline tables
- [\[SPARK-17069\]](#) Expose spark.range() as table-valued function in SQL
- [\[SPARK-17117\]](#)[SQL] 1 / NULL should not fail analysis
- [\[SPARK-16391\]](#)[SQL] Support partial aggregation for reduceGroups
- [\[SPARK-16995\]](#)[SQL] TreeNodeException when flat mapping RelationalGroupedDataset created from DataFrame containing a column created with lit/expr
- [\[SPARK-17096\]](#)[SQL][STREAMING] Improve exception string reported through the StreamingQueryListener
- [\[SPARK-17102\]](#)[SQL] bypass UserDefinedGenerator for json format check
- [\[SPARK-15285\]](#)[SQL] Generated SpecificSafeProjection.apply method grows beyond 64 KB
- [\[SPARK-17084\]](#)[SQL] Rename ParserUtils.assert to validate
- [\[SPARK-17089\]](#)[DOCS] Remove api doc link for mapReduceTriplets operator
- [\[SPARK-16964\]](#)[SQL] Remove private[sql] and private[spark] from sql.execution package [Backport]
- [\[SPARK-17065\]](#)[SQL] Improve the error message when encountering an incompatible DataSourceRegister
- [\[SPARK-16508\]](#)[SPARKRR] Split docs for arrange and orderBy methods
- [\[SPARK-17027\]](#)[ML] Avoid integer overflow in PolynomialExpansion.getPolySize
- [\[SPARK-16966\]](#)[SQL][CORE] App Name is a randomUUID even when "spark.app.name" exists
- [\[SPARK-17013\]](#)[SQL] Parse negative numeric literals
- [\[SPARK-16975\]](#)[SQL] Column-partition path starting '\_' should be handled correctly
- [\[SPARK-17022\]](#)[YARN] Handle potential deadlock in driver handling messages
- [\[SPARK-17018\]](#)[SQL] literals.sql for testing literal parsing
- [\[SPARK-17015\]](#)[SQL] group-by/order-by ordinal and arithmetic tests



- [\[SPARK-15899\]](#)[SQL] Fix the construction of the file path with hadoop Path for Spark 2.0
- [\[SPARK-17011\]](#)[SQL] Support testing exceptions in SQLQueryTestSuite
- [\[SPARK-17007\]](#)[SQL] Move test data files into a test-data folder
- [\[SPARK-17008\]](#)[\[SPARK-17009\]](#)[SQL] Normalization and isolation in SQLQueryTestSuite.
- [\[SPARK-16866\]](#)[SQL] Infrastructure for file-based SQL end-to-end tests
- [\[SPARK-17010\]](#)[MINOR][DOC] Wrong description in memory management document
- [\[SPARK-15639\]](#) [\[SPARK-16321\]](#) [SQL] Push down filter at RowGroups level for parquet reader
- [\[SPARK-16324\]](#)[SQL] regexp\_extract should doc that it returns empty string when match fails
- [\[SPARK-16522\]](#)[MESOS] Spark application throws exception on exit.
- [\[SPARK-16905\]](#) SQL DDL: MSCK REPAIR TABLE
- [\[SPARK-16956\]](#) Make ApplicationState.MAX\_NUM\_RETRY configurable
- [\[SPARK-16950\]](#) [PYSARK] fromOffsets parameter support in KafkaUtils.createDirectStream for python3
- [\[SPARK-16610\]](#)[SQL] Add `orc.compress` as an alias for `compression` option.
- [\[SPARK-16563\]](#)[SQL] fix spark sql thrift server FetchResults bug
- [\[SPARK-16953\]](#) Make requestTotalExecutors public Developer API to be consistent with requestExecutors/killExecutors
- [\[SPARK-16586\]](#)[CORE] Handle JVM errors printed to stdout.
- [\[SPARK-16936\]](#)[SQL] Case Sensitivity Support for Refresh Temp Table
- [\[SPARK-16457\]](#)[SQL] Fix Wrong Messages when CTAS with a Partition By Clause
- [\[SPARK-16939\]](#)[SQL] Fix build error by using `Tuple1` explicitly in StringFunctionsSuite
- [\[SPARK-16409\]](#)[SQL] regexp\_extract with optional groups causes NPE
- [\[SPARK-16911\]](#) Fix the links in the programming guide
- [\[SPARK-16870\]](#)[DOCS] Summary:add "spark.sql.broadcastTimeout" into docs/sql-programming-gu...
- [\[SPARK-16932\]](#)[DOCS] Changed programming guide to not reference old accumulator API in Scala
- [\[SPARK-16925\]](#) Master should call schedule() after all executor exit events, not only failures
- [\[SPARK-16772\]](#)[PYTHON][DOCS] Fix API doc references to UDFRegistration + Update "important classes"
- [\[SPARK-16750\]](#)[FOLLOW-UP][ML] Add transformSchema for StringIndexer/VectorAssembler and fix failed tests.
- [\[SPARK-16907\]](#)[SQL] Fix performance regression for parquet table when vectorized parquet record reader is not being used
- [\[SPARK-16863\]](#)[ML] ProbabilisticClassifier.fit check thresholds' length
- [\[SPARK-16877\]](#)[BUILD] Add rules for preventing to use Java annotations (Deprecated and Override)
- [\[SPARK-16880\]](#)[ML][MLLIB] make ann training data persisted if needed
- [\[SPARK-16875\]](#)[SQL] Add args checking for DataSet randomSplit and sample
- [\[SPARK-16802\]](#) [SQL] fix overflow in LongToUnsafeRowMap
- [\[SPARK-16873\]](#)[CORE] Fix SpillReader NPE when spillFile has no data
- [\[SPARK-14204\]](#)[SQL] register driverClass rather than user-specified class
- [\[SPARK-16714\]](#)[\[SPARK-16735\]](#)[\[SPARK-16646\]](#) array, map, greatest, least's type coercion should handle decimal type
- [\[SPARK-16796\]](#)[WEB UI] Visible passwords on Spark environment page
- [\[SPARK-16831\]](#)[PYTHON] Fixed bug in CrossValidator.avgMetrics
- [\[SPARK-16787\]](#) SparkContext.addFile() should not throw if called twice with the same file
- [\[SPARK-16850\]](#)[SQL] Improve type checking error message for greatest/least
- [\[SPARK-16836\]](#)[SQL] Add support for CURRENT\_DATE/CURRENT\_TIMESTAMP literals
- [\[SPARK-16062\]](#) [\[SPARK-15989\]](#) [SQL] Fix two bugs of Python-only UDTs
- [\[SPARK-16837\]](#)[SQL] TimeWindow incorrectly drops slideDuration in constructors
- [\[SPARK-15541\]](#) Casting ConcurrentHashMap to ConcurrentMap (master branch)
- [\[SPARK-16558\]](#)[EXAMPLES][MLLIB] examples/mllib/LDAExample should use MLVector instead of MLlib Vector
- [\[SPARK-16734\]](#)[EXAMPLES][SQL] Revise examples of all language bindings
- [\[SPARK-16818\]](#) Exchange reuse incorrectly reuses scans over different sets of partitions
- [\[SPARK-15869\]](#)[STREAMING] Fix a potential NPE in StreamingJobProgressListener.getBatchUIData

- [\[SPARK-16774\]](#)[SQL] Fix use of deprecated timestamp constructor & improve timezone handling
- [\[SPARK-16791\]](#)[SQL] cast struct with timestamp field fails
- [\[SPARK-16778\]](#)[SQL][TRIVIAL] Fix deprecation warning with SQLContext
- [\[SPARK-16805\]](#)[SQL] Log timezone when query result does not match
- [\[SPARK-16813\]](#)[SQL] Remove private[sql] and private[spark] from catalyst package
- [\[SPARK-16812\]](#) Open up SparkILoop.getAddedJars
- [\[SPARK-16800\]](#)[EXAMPLES][ML] Fix Java examples that fail to run due to exception
- [\[SPARK-16748\]](#)[SQL] SparkExceptions during planning should not wrapped in TreeNodeException
- [\[SPARK-16761\]](#)[DOC][ML] Fix doc link in docs/ml-guide.md
- [\[SPARK-16750\]](#)[ML] Fix GaussianMixture training failed due to feature column type mistake
- [\[SPARK-16664\]](#)[SQL] Fix persist call on Data frames with more than 200...
- [\[SPARK-16772\]](#) Correct API doc references to PySpark classes + formatting fixes
- [\[SPARK-16764\]](#)[SQL] Recommend disabling vectorized parquet reader on OutOfMemoryError
- [\[SPARK-16740\]](#)[SQL] Fix Long overflow in LongToUnsafeRowMap
- [\[SPARK-16639\]](#)[SQL] The query with having condition that contains grouping by column should work
- [\[SPARK-15232\]](#)[SQL] Add subquery SQL building tests to LogicalPlanToSQLSuite
- [\[SPARK-16730\]](#)[SQL] Implement function aliases for type casts
- [\[SPARK-16729\]](#)[SQL] Throw analysis exception for invalid date casts
- [\[SPARK-16621\]](#)[SQL] Generate stable SQLs in SQLBuilder
- [\[SPARK-16633\]](#)[\[SPARK-16642\]](#)[\[SPARK-16721\]](#)[SQL] Fixes three issues related to lead and lag functions
- [\[SPARK-16724\]](#) Expose DefinedByConstructorParams
- [\[SPARK-16672\]](#)[SQL] SQLBuilder should not raise exceptions on EXISTS queries
- [\[SPARK-16722\]](#)[TESTS] Fix a StreamingContext leak in StreamingContextSuite when eventually fails
- [\[SPARK-14131\]](#)[STREAMING] SQL Improved fix for avoiding potential deadlocks in HDFSMetadataLog
- [\[SPARK-16715\]](#)[TESTS] Fix a potential ExprId conflict for SubexpressionEliminationSuite."Semantic equals and hash"
- [\[SPARK-16485\]](#)[DOC][ML] Fixed several inline formatting in ml features doc
- [\[SPARK-16703\]](#)[SQL] Remove extra whitespace in SQL generation for window functions
- [\[SPARK-16698\]](#)[SQL] Field names having dots should be allowed for datasources based on FileFormat
- [\[SPARK-16648\]](#)[SQL] Make ignoreNullsExpr a child expression of First and Last
- [\[SPARK-16699\]](#)[SQL] Fix performance bug in hash aggregate on long string keys
- [\[SPARK-16515\]](#)[SQL][FOLLOW-UP] Fix test `script` on OS X/Windows...
- [\[SPARK-16690\]](#)[TEST] rename SQLTestUtils.withTempTable to withTempView
- [\[SPARK-16380\]](#)[EXAMPLES] Update SQL examples and programming guide for Python language binding
- [\[SPARK-16651\]](#)[PYSARK][DOC] Make `withColumnRenamed/drop` description more consistent with Scala API
- [\[SPARK-16650\]](#) Improve documentation of spark.task.maxFailures
- [\[SPARK-16287\]](#)[HOTFIX][BUILD][SQL] Fix annotation argument needs to be a constant
- [\[SPARK-16287\]](#)[SQL] Implement str\_to\_map SQL function
- [\[SPARK-16334\]](#) Maintain single dictionary per row-batch in vectorized parquet reader
- [\[SPARK-16656\]](#)[SQL] Try to make CreateTableAsSelectSuite more stable
- [\[SPARK-16644\]](#)[SQL] Aggregate should not propagate constraints containing aggregate expressions
- [\[SPARK-16440\]](#)[MLLIB] Destroy broadcasted variables even on driver
- [\[SPARK-5682\]](#)[CORE] Add encrypted shuffle in spark
- [\[SPARK-16901\]](#) Hive settings in hive-site.xml may be overridden by Hive's default values
- [\[SPARK-16272\]](#)[CORE] Allow config values to reference conf, env, system props.
- [\[SPARK-16632\]](#)[SQL] Use Spark requested schema to guide vectorized Parquet reader initialization
- [\[SPARK-16634\]](#)[SQL] Workaround JVM bug by moving some code out of ctor.
- [\[SPARK-16505\]](#)[YARN] Optionally propagate error during shuffle service startup.
- [\[SPARK-14963\]](#)[MINOR][YARN] Fix typo in YarnShuffleService recovery file name
- [\[SPARK-14963\]](#)[YARN] Using recoveryPath if NM recovery is enabled

- [\[SPARK-16349\]](#)[SQL] Fall back to isolated class loader when classes not found.
- [\[SPARK-16119\]](#)[sql] Support PURGE option to drop table / partition.

## CDS Powered by Apache Spark Version, Packaging, and Download Information

The following sections provide links to the parcel and service descriptor files for the different CDS versions.

### CDS Versions Available for Download



**Note:** The parcel version displayed in Cloudera Manager, which is also part of the parcel file name, is structured as follows:

`<CDS_version>.cloudera<release_number>-1.<cdh_build_version>.p<patch_version>.<build_number>`

For example:

```
2.1.0.cloudera3-1.cdh5.13.3.p0.569822
```

The `<CDS_version>.cloudera<release_number>` portion tells you the complete version number of the release. It tells you the major and minor versions and the release number. In the example above, the major version is CDS **2**, the minor version is **.1**, and the release number is **3**. The complete version number is therefore CDS 2.1 Release 3.

The `<cdh_build_version>` portion only signifies the version of CDH upon which the release was built. It is **not** the minimum supported CDH version. For example, although CDS 2.1 Release 3 was built on CDH 5.13.3, it is still supported on CDH 5.7 and higher CDH 5 versions.

To view the supported CDH versions and other requirements, see [CDS Powered by Apache Spark Requirements](#) on page 7.

**Table 1: Available CDS Versions**

Version	Custom Service Descriptor	Parcel Repository
2.4 Release 2	<a href="#">SPARK2_ON_YARN-2.4.0.cloudera2.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.4.0.cloudera2/">http://archive.cloudera.com/spark2/parcels/2.4.0.cloudera2/</a>
2.4 Release 1	<a href="#">SPARK2_ON_YARN-2.4.0.cloudera1.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.4.0.cloudera1/">http://archive.cloudera.com/spark2/parcels/2.4.0.cloudera1/</a>
2.3 Release 4	<a href="#">SPARK2_ON_YARN-2.3.0.cloudera4.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.3.0.cloudera4/">http://archive.cloudera.com/spark2/parcels/2.3.0.cloudera4/</a>
2.3 Release 3	<a href="#">SPARK2_ON_YARN-2.3.0.cloudera3.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.3.0.cloudera3/">http://archive.cloudera.com/spark2/parcels/2.3.0.cloudera3/</a>
2.3 Release 2	<a href="#">SPARK2_ON_YARN-2.3.0.cloudera2.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.3.0.cloudera2/">http://archive.cloudera.com/spark2/parcels/2.3.0.cloudera2/</a>
2.3 Release 1	Never officially released; if downloaded, do not use	Never officially released; if downloaded, do not use
2.2 Release 4	<a href="#">SPARK2_ON_YARN-2.2.0.cloudera4.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera4/">http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera4/</a>
2.2 Release 3	<a href="#">SPARK2_ON_YARN-2.2.0.cloudera3.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera3/">http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera3/</a>
2.2 Release 2	<a href="#">SPARK2_ON_YARN-2.2.0.cloudera2.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera2/">http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera2/</a>

Version	Custom Service Descriptor	Parcel Repository
2.2 Release 1	<a href="#">SPARK2_ON_YARN-2.2.0.cloudera1.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera1/">http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera1/</a>
2.1 Release 4	<a href="#">SPARK2_ON_YARN-2.1.0.cloudera4.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera4/">http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera4/</a>
2.1 Release 3	<a href="#">SPARK2_ON_YARN-2.1.0.cloudera3.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera3/">http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera3/</a>
2.1 Release 2	<a href="#">SPARK2_ON_YARN-2.1.0.cloudera2.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera2/">http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera2/</a>
2.1 Release 1	<a href="#">SPARK2_ON_YARN-2.1.0.cloudera1.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera1/">http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera1/</a>
2.0 Release 2	<a href="#">SPARK2_ON_YARN-2.0.0.cloudera2.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.0.0.cloudera2/">http://archive.cloudera.com/spark2/parcels/2.0.0.cloudera2/</a>
2.0 Release 1	<a href="#">SPARK2_ON_YARN-2.0.0.cloudera1.jar</a>	<a href="http://archive.cloudera.com/spark2/parcels/2.0.0.cloudera1/">http://archive.cloudera.com/spark2/parcels/2.0.0.cloudera1/</a>

## CDS Maven Artifacts

For information about using CDS Maven artifacts, see [Using the CDS Powered by Apache Spark Maven Repository](#) on page 60.

## Using the CDS Powered by Apache Spark Maven Repository



**Important:** CDS 2 no longer includes an assembly JAR. When you build an application JAR, *do not* include CDH or CDS JARs, because they are already provided. If you do, upgrading CDH or CDS can break your application. To avoid this situation, set the Maven dependency `scope` to `provided`. If you have already built applications which include the CDH or CDS JARs, update the dependency to set `scope` to `provided` and recompile.

For information on how to use CDH artifacts, see [Using the CDH 5 Maven Repository](#).

If you want to build applications or tools for use with CDS Powered by Apache Spark, and you are using Maven or Ivy for dependency management, you can pull the CDS artifacts from the Cloudera Maven repository. The repository is available at <https://repository.cloudera.com/artifactory/cloudera-repos/>.

The following is a sample POM (`pom.xml`) file:

```
<project xmlns="http://maven.apache.org/POM/4.0.0"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://maven.apache.org/POM/4.0.0
  http://maven.apache.org/maven-v4_0_0.xsd">
  <repositories>
    <repository>
      <id>cloudera</id>
      <url>https://repository.cloudera.com/artifactory/cloudera-repos/</url>
    </repository>
  </repositories>
</project>
```

For more information about the Maven artifacts for each CDS release, see the following topics:

## CDS 2.4 Powered by Apache Spark Maven Artifacts

The following tables lists the `groupId`, `artifactId`, and `version` required to access the artifacts for each CDS 2.4 Powered by Apache Spark release:

### CDS 2.4 Release 2 Maven Artifacts

The following pom fragment shows how to access a CDS 2.4 Release 2 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.4.0.cloudera2</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.4.0.cloudera2
	org.apache.spark	spark-core_2.11	2.4.0.cloudera2
	org.apache.spark	spark-graphx_2.11	2.4.0.cloudera2
	org.apache.spark	spark-hive-exec_2.11	2.4.0.cloudera2
	org.apache.spark	spark-hive_2.11	2.4.0.cloudera2
	org.apache.spark	spark-kvstore_2.11	2.4.0.cloudera2
	org.apache.spark	spark-launcher_2.11	2.4.0.cloudera2
	org.apache.spark	spark-mllib-local_2.11	2.4.0.cloudera2
	org.apache.spark	spark-mllib_2.11	2.4.0.cloudera2
	org.apache.spark	spark-network-common_2.11	2.4.0.cloudera2
	org.apache.spark	spark-network-shuffle_2.11	2.4.0.cloudera2
	org.apache.spark	spark-network-yarn_2.11	2.4.0.cloudera2
	org.apache.spark	spark-repl_2.11	2.4.0.cloudera2
	org.apache.spark	spark-sketch_2.11	2.4.0.cloudera2
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.4.0.cloudera2

Project	groupId	artifactId	version
	org.apache.spark	spark-sql_2.11	2.4.0.cloudera2
	org.apache.spark	spark-streaming-flume-sink_2.11	2.4.0.cloudera2
	org.apache.spark	spark-streaming-flume_2.11	2.4.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.4.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.4.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.4.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.4.0.cloudera2
	org.apache.spark	spark-streaming_2.11	2.4.0.cloudera2
	org.apache.spark	spark-tags_2.11	2.4.0.cloudera2
	org.apache.spark	spark-unsafe_2.11	2.4.0.cloudera2
	org.apache.spark	spark-yarn_2.11	2.4.0.cloudera2

CDS 2.4 Release 1 Maven Artifacts

The following pom fragment shows how to access a CDS 2.4 Release 1 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.4.0.cloudera1</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.4.0.cloudera1
	org.apache.spark	spark-core_2.11	2.4.0.cloudera1
	org.apache.spark	spark-graphx_2.11	2.4.0.cloudera1
	org.apache.spark	spark-hive-exec_2.11	2.4.0.cloudera1
	org.apache.spark	spark-hive_2.11	2.4.0.cloudera1

Project	groupId	artifactId	version
	org.apache.spark	spark-kvstore_2.11	2.4.0.cloudera1
	org.apache.spark	spark-launcher_2.11	2.4.0.cloudera1
	org.apache.spark	spark-mllib-local_2.11	2.4.0.cloudera1
	org.apache.spark	spark-mllib_2.11	2.4.0.cloudera1
	org.apache.spark	spark-network-common_2.11	2.4.0.cloudera1
	org.apache.spark	spark-network-shuffle_2.11	2.4.0.cloudera1
	org.apache.spark	spark-network-yarn_2.11	2.4.0.cloudera1
	org.apache.spark	spark-repl_2.11	2.4.0.cloudera1
	org.apache.spark	spark-sketch_2.11	2.4.0.cloudera1
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.4.0.cloudera1
	org.apache.spark	spark-sql_2.11	2.4.0.cloudera1
	org.apache.spark	spark-streaming-flume-sink_2.11	2.4.0.cloudera1
	org.apache.spark	spark-streaming-flume_2.11	2.4.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.4.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.4.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.4.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.4.0.cloudera1
	org.apache.spark	spark-streaming_2.11	2.4.0.cloudera1
	org.apache.spark	spark-tags_2.11	2.4.0.cloudera1
	org.apache.spark	spark-unsafe_2.11	2.4.0.cloudera1
	org.apache.spark	spark-yarn_2.11	2.4.0.cloudera1

### CDS 2.3 Powered by Apache Spark Maven Artifacts

The following tables lists the `groupId`, `artifactId`, and `version` required to access the artifacts for each CDS 2.3 Powered by Apache Spark release:

#### CDS 2.3 Release 4 Maven Artifacts

The following pom fragment shows how to access a CDS 2.3 Release 4 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.3.0.cloudera4</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.3.0.cloudera4
	org.apache.spark	spark-core_2.11	2.3.0.cloudera4
	org.apache.spark	spark-graphx_2.11	2.3.0.cloudera4
	org.apache.spark	spark-hive-exec_2.11	2.3.0.cloudera4
	org.apache.spark	spark-hive_2.11	2.3.0.cloudera4
	org.apache.spark	spark-kvstore_2.11	2.3.0.cloudera4
	org.apache.spark	spark-launcher_2.11	2.3.0.cloudera4
	org.apache.spark	spark-mllib-local_2.11	2.3.0.cloudera4
	org.apache.spark	spark-mllib_2.11	2.3.0.cloudera4
	org.apache.spark	spark-network-common_2.11	2.3.0.cloudera4
	org.apache.spark	spark-network-shuffle_2.11	2.3.0.cloudera4
	org.apache.spark	spark-network-yarn_2.11	2.3.0.cloudera4
	org.apache.spark	spark-repl_2.11	2.3.0.cloudera4
	org.apache.spark	spark-sketch_2.11	2.3.0.cloudera4
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.3.0.cloudera4



Project	groupId	artifactId	version
	org.apache.spark	spark-sql_2.11	2.3.0.cloudera4
	org.apache.spark	spark-streaming-flume-sink_2.11	2.3.0.cloudera4
	org.apache.spark	spark-streaming-flume_2.11	2.3.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.3.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.3.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.3.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.3.0.cloudera4
	org.apache.spark	spark-streaming_2.11	2.3.0.cloudera4
	org.apache.spark	spark-tags_2.11	2.3.0.cloudera4
	org.apache.spark	spark-unsafe_2.11	2.3.0.cloudera4
	org.apache.spark	spark-yarn_2.11	2.3.0.cloudera4

### CDS 2.3 Release 3 Maven Artifacts

The following pom fragment shows how to access a CDS 2.3 Release 3 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.3.0.cloudera3</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.3.0.cloudera3
	org.apache.spark	spark-core_2.11	2.3.0.cloudera3
	org.apache.spark	spark-graphx_2.11	2.3.0.cloudera3
	org.apache.spark	spark-hive-exec_2.11	2.3.0.cloudera3
	org.apache.spark	spark-hive_2.11	2.3.0.cloudera3

Project	groupId	artifactId	version
	org.apache.spark	spark-kvstore_2.11	2.3.0.cloudera3
	org.apache.spark	spark-launcher_2.11	2.3.0.cloudera3
	org.apache.spark	spark-mllib-local_2.11	2.3.0.cloudera3
	org.apache.spark	spark-mllib_2.11	2.3.0.cloudera3
	org.apache.spark	spark-network-common_2.11	2.3.0.cloudera3
	org.apache.spark	spark-network-shuffle_2.11	2.3.0.cloudera3
	org.apache.spark	spark-network-yarn_2.11	2.3.0.cloudera3
	org.apache.spark	spark-repl_2.11	2.3.0.cloudera3
	org.apache.spark	spark-sketch_2.11	2.3.0.cloudera3
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.3.0.cloudera3
	org.apache.spark	spark-sql_2.11	2.3.0.cloudera3
	org.apache.spark	spark-streaming-flume-sink_2.11	2.3.0.cloudera3
	org.apache.spark	spark-streaming-flume_2.11	2.3.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.3.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.3.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.3.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.3.0.cloudera3
	org.apache.spark	spark-streaming_2.11	2.3.0.cloudera3
	org.apache.spark	spark-tags_2.11	2.3.0.cloudera3
	org.apache.spark	spark-unsafe_2.11	2.3.0.cloudera3
	org.apache.spark	spark-yarn_2.11	2.3.0.cloudera3

## CDS 2.3 Release 2 Maven Artifacts

The following pom fragment shows how to access a CDS 2.3 Release 2 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.3.0.cloudera2</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.3.0.cloudera2
	org.apache.spark	spark-core_2.11	2.3.0.cloudera2
	org.apache.spark	spark-graphx_2.11	2.3.0.cloudera2
	org.apache.spark	spark-hive-exec_2.11	2.3.0.cloudera2
	org.apache.spark	spark-hive_2.11	2.3.0.cloudera2
	org.apache.spark	spark-kvstore_2.11	2.3.0.cloudera2
	org.apache.spark	spark-launcher_2.11	2.3.0.cloudera2
	org.apache.spark	spark-mllib-local_2.11	2.3.0.cloudera2
	org.apache.spark	spark-mllib_2.11	2.3.0.cloudera2
	org.apache.spark	spark-network-common_2.11	2.3.0.cloudera2
	org.apache.spark	spark-network-shuffle_2.11	2.3.0.cloudera2
	org.apache.spark	spark-network-yarn_2.11	2.3.0.cloudera2
	org.apache.spark	spark-repl_2.11	2.3.0.cloudera2
	org.apache.spark	spark-sketch_2.11	2.3.0.cloudera2
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.3.0.cloudera2
	org.apache.spark	spark-sql_2.11	2.3.0.cloudera2
	org.apache.spark	spark-streaming-flume-sink_2.11	2.3.0.cloudera2

Project	groupId	artifactId	version
	org.apache.spark	spark-streaming-flume_2.11	2.3.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.3.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.3.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.3.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.3.0.cloudera2
	org.apache.spark	spark-streaming_2.11	2.3.0.cloudera2
	org.apache.spark	spark-tags_2.11	2.3.0.cloudera2
	org.apache.spark	spark-unsafe_2.11	2.3.0.cloudera2
	org.apache.spark	spark-yarn_2.11	2.3.0.cloudera2

#### CDS 2.3 Release 1 Maven Artifacts

CDS 2.3 Release 1 was never officially released; if downloaded, do not use.

### CDS 2.2 Powered by Apache Spark Maven Artifacts

The following tables lists the `groupId`, `artifactId`, and `version` required to access the artifacts for each CDS 2.2 Powered by Apache Spark release:

#### CDS 2.2 Release 4 Maven Artifacts

The following pom fragment shows how to access a CDS 2.2 Release 4 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.2.0.cloudera4</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.2.0.cloudera4
	org.apache.spark	spark-core_2.11	2.2.0.cloudera4
	org.apache.spark	spark-graphx_2.11	2.2.0.cloudera4
	org.apache.spark	spark-hive-exec_2.11	2.2.0.cloudera4

Project	groupId	artifactId	version
	org.apache.spark	spark-hive_2.11	2.2.0.cloudera4
	org.apache.spark	spark-launcher_2.11	2.2.0.cloudera4
	org.apache.spark	spark-mllib-local_2.11	2.2.0.cloudera4
	org.apache.spark	spark-mllib_2.11	2.2.0.cloudera4
	org.apache.spark	spark-network-common_2.11	2.2.0.cloudera4
	org.apache.spark	spark-network-shuffle_2.11	2.2.0.cloudera4
	org.apache.spark	spark-network-yarn_2.11	2.2.0.cloudera4
	org.apache.spark	spark-repl_2.11	2.2.0.cloudera4
	org.apache.spark	spark-sketch_2.11	2.2.0.cloudera4
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.2.0.cloudera4
	org.apache.spark	spark-sql_2.11	2.2.0.cloudera4
	org.apache.spark	spark-streaming-flume-sink_2.11	2.2.0.cloudera4
	org.apache.spark	spark-streaming-flume_2.11	2.2.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.2.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.2.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.2.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.2.0.cloudera4
	org.apache.spark	spark-streaming_2.11	2.2.0.cloudera4
	org.apache.spark	spark-tags_2.11	2.2.0.cloudera4
	org.apache.spark	spark-unsafe_2.11	2.2.0.cloudera4
	org.apache.spark	spark-yarn_2.11	2.2.0.cloudera4

CDS 2.2 Release 3 Maven Artifacts

The following pom fragment shows how to access a CDS 2.2 Release 3 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.2.0.cloudera3</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.2.0.cloudera3
	org.apache.spark	spark-core_2.11	2.2.0.cloudera3
	org.apache.spark	spark-graphx_2.11	2.2.0.cloudera3
	org.apache.spark	spark-hive-exec_2.11	2.2.0.cloudera3
	org.apache.spark	spark-hive_2.11	2.2.0.cloudera3
	org.apache.spark	spark-launcher_2.11	2.2.0.cloudera3
	org.apache.spark	spark-mllib-local_2.11	2.2.0.cloudera3
	org.apache.spark	spark-mllib_2.11	2.2.0.cloudera3
	org.apache.spark	spark-network-common_2.11	2.2.0.cloudera3
	org.apache.spark	spark-network-shuffle_2.11	2.2.0.cloudera3
	org.apache.spark	spark-network-yarn_2.11	2.2.0.cloudera3
	org.apache.spark	spark-repl_2.11	2.2.0.cloudera3
	org.apache.spark	spark-sketch_2.11	2.2.0.cloudera3
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.2.0.cloudera3
	org.apache.spark	spark-sql_2.11	2.2.0.cloudera3
	org.apache.spark	spark-streaming-flume-sink_2.11	2.2.0.cloudera3
	org.apache.spark	spark-streaming-flume_2.11	2.2.0.cloudera3

Project	groupId	artifactId	version
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.2.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.2.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.2.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.2.0.cloudera3
	org.apache.spark	spark-streaming_2.11	2.2.0.cloudera3
	org.apache.spark	spark-tags_2.11	2.2.0.cloudera3
	org.apache.spark	spark-unsafe_2.11	2.2.0.cloudera3
	org.apache.spark	spark-yarn_2.11	2.2.0.cloudera3

### CDS 2.2 Release 2 Maven Artifacts

The following pom fragment shows how to access a CDS 2.2 Release 2 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.2.0.cloudera2</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.2.0.cloudera2
	org.apache.spark	spark-core_2.11	2.2.0.cloudera2
	org.apache.spark	spark-graphx_2.11	2.2.0.cloudera2
	org.apache.spark	spark-hive-exec_2.11	2.2.0.cloudera2
	org.apache.spark	spark-hive_2.11	2.2.0.cloudera2
	org.apache.spark	spark-launcher_2.11	2.2.0.cloudera2
	org.apache.spark	spark-mllib-local_2.11	2.2.0.cloudera2
	org.apache.spark	spark-mllib_2.11	2.2.0.cloudera2

Project	groupId	artifactId	version
	org.apache.spark	spark-network-common_2.11	2.2.0.cloudera2
	org.apache.spark	spark-network-shuffle_2.11	2.2.0.cloudera2
	org.apache.spark	spark-network-yarn_2.11	2.2.0.cloudera2
	org.apache.spark	spark-repl_2.11	2.2.0.cloudera2
	org.apache.spark	spark-sketch_2.11	2.2.0.cloudera2
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.2.0.cloudera2
	org.apache.spark	spark-sql_2.11	2.2.0.cloudera2
	org.apache.spark	spark-streaming-flume-sink_2.11	2.2.0.cloudera2
	org.apache.spark	spark-streaming-flume_2.11	2.2.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.2.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.2.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.2.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.2.0.cloudera2
	org.apache.spark	spark-streaming_2.11	2.2.0.cloudera2
	org.apache.spark	spark-tags_2.11	2.2.0.cloudera2
	org.apache.spark	spark-unsafe_2.11	2.2.0.cloudera2
	org.apache.spark	spark-yarn_2.11	2.2.0.cloudera2

### CDS 2.2 Release 1 Maven Artifacts

The following pom fragment shows how to access a CDS 2.2 Release 1 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.2.0.cloudera1</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.



Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.2.0.cloudera1
	org.apache.spark	spark-core_2.11	2.2.0.cloudera1
	org.apache.spark	spark-graphx_2.11	2.2.0.cloudera1
	org.apache.spark	spark-hive-exec_2.11	2.2.0.cloudera1
	org.apache.spark	spark-hive_2.11	2.2.0.cloudera1
	org.apache.spark	spark-launcher_2.11	2.2.0.cloudera1
	org.apache.spark	spark-mllib-local_2.11	2.2.0.cloudera1
	org.apache.spark	spark-mllib_2.11	2.2.0.cloudera1
	org.apache.spark	spark-network-common_2.11	2.2.0.cloudera1
	org.apache.spark	spark-network-shuffle_2.11	2.2.0.cloudera1
	org.apache.spark	spark-network-yarn_2.11	2.2.0.cloudera1
	org.apache.spark	spark-repl_2.11	2.2.0.cloudera1
	org.apache.spark	spark-sketch_2.11	2.2.0.cloudera1
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.2.0.cloudera1
	org.apache.spark	spark-sql_2.11	2.2.0.cloudera1
	org.apache.spark	spark-streaming-flume-sink_2.11	2.2.0.cloudera1
	org.apache.spark	spark-streaming-flume_2.11	2.2.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.2.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.2.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.2.0.cloudera1
org.apache.spark	spark-streaming-kafka-0-8_2.11	2.2.0.cloudera1	

Project	groupId	artifactId	version
	org.apache.spark	spark-streaming_2.11	2.2.0.cloudera1
	org.apache.spark	spark-tags_2.11	2.2.0.cloudera1
	org.apache.spark	spark-unsafe_2.11	2.2.0.cloudera1
	org.apache.spark	spark-yarn_2.11	2.2.0.cloudera1

### CDS 2.1 Powered by Apache Spark Maven Artifacts

The following tables lists the `groupId`, `artifactId`, and `version` required to access the artifacts for each CDS 2.1 Powered by Apache Spark release:

#### CDS 2.1 Release 4 Maven Artifacts

The following pom fragment shows how to access a CDS 2.1 Release 4 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.1.0.cloudera4</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.1.0.cloudera4
	org.apache.spark	spark-core_2.11	2.1.0.cloudera4
	org.apache.spark	spark-graphx_2.11	2.1.0.cloudera4
	org.apache.spark	spark-hive-exec_2.11	2.1.0.cloudera4
	org.apache.spark	spark-hive_2.11	2.1.0.cloudera4
	org.apache.spark	spark-launcher_2.11	2.1.0.cloudera4
	org.apache.spark	spark-mllib-local_2.11	2.1.0.cloudera4
	org.apache.spark	spark-mllib_2.11	2.1.0.cloudera4
	org.apache.spark	spark-network-common_2.11	2.1.0.cloudera4
	org.apache.spark	spark-network-shuffle_2.11	2.1.0.cloudera4

Project	groupId	artifactId	version
	org.apache.spark	spark-network-yarn_2.11	2.1.0.cloudera4
	org.apache.spark	spark-repl_2.11	2.1.0.cloudera4
	org.apache.spark	spark-sketch_2.11	2.1.0.cloudera4
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.1.0.cloudera4
	org.apache.spark	spark-sql_2.11	2.1.0.cloudera4
	org.apache.spark	spark-streaming-flume-sink_2.11	2.1.0.cloudera4
	org.apache.spark	spark-streaming-flume_2.11	2.1.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.1.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.1.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.1.0.cloudera4
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.1.0.cloudera4
	org.apache.spark	spark-streaming_2.11	2.1.0.cloudera4
	org.apache.spark	spark-tags_2.11	2.1.0.cloudera4
	org.apache.spark	spark-unsafe_2.11	2.1.0.cloudera4
	org.apache.spark	spark-yarn_2.11	2.1.0.cloudera4

### CDS 2.1 Release 3 Maven Artifacts

The following pom fragment shows how to access a CDS 2.1 Release 3 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.1.0.cloudera3</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.1.0.cloudera3

Project	groupId	artifactId	version
	org.apache.spark	spark-core_2.11	2.1.0.cloudera3
	org.apache.spark	spark-graphx_2.11	2.1.0.cloudera3
	org.apache.spark	spark-hive-exec_2.11	2.1.0.cloudera3
	org.apache.spark	spark-hive_2.11	2.1.0.cloudera3
	org.apache.spark	spark-launcher_2.11	2.1.0.cloudera3
	org.apache.spark	spark-mllib-local_2.11	2.1.0.cloudera3
	org.apache.spark	spark-mllib_2.11	2.1.0.cloudera3
	org.apache.spark	spark-network-common_2.11	2.1.0.cloudera3
	org.apache.spark	spark-network-shuffle_2.11	2.1.0.cloudera3
	org.apache.spark	spark-network-yarn_2.11	2.1.0.cloudera3
	org.apache.spark	spark-repl_2.11	2.1.0.cloudera3
	org.apache.spark	spark-sketch_2.11	2.1.0.cloudera3
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.1.0.cloudera3
	org.apache.spark	spark-sql_2.11	2.1.0.cloudera3
	org.apache.spark	spark-streaming-flume-sink_2.11	2.1.0.cloudera3
	org.apache.spark	spark-streaming-flume_2.11	2.1.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.1.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.1.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.1.0.cloudera3
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.1.0.cloudera3
	org.apache.spark	spark-streaming_2.11	2.1.0.cloudera3

Project	groupId	artifactId	version
	org.apache.spark	spark-tags_2.11	2.1.0.cloudera3
	org.apache.spark	spark-unsafe_2.11	2.1.0.cloudera3
	org.apache.spark	spark-yarn_2.11	2.1.0.cloudera3

### CDS 2.1 Release 2 Maven Artifacts

The following pom fragment shows how to access a CDS 2.1 Release 2 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.1.0.cloudera2</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.1.0.cloudera2
	org.apache.spark	spark-core_2.11	2.1.0.cloudera2
	org.apache.spark	spark-graphx_2.11	2.1.0.cloudera2
	org.apache.spark	spark-hive-exec_2.11	2.1.0.cloudera2
	org.apache.spark	spark-hive_2.11	2.1.0.cloudera2
	org.apache.spark	spark-launcher_2.11	2.1.0.cloudera2
	org.apache.spark	spark-mllib-local_2.11	2.1.0.cloudera2
	org.apache.spark	spark-mllib_2.11	2.1.0.cloudera2
	org.apache.spark	spark-network-common_2.11	2.1.0.cloudera2
	org.apache.spark	spark-network-shuffle_2.11	2.1.0.cloudera2
	org.apache.spark	spark-network-yarn_2.11	2.1.0.cloudera2
	org.apache.spark	spark-repl_2.11	2.1.0.cloudera2
	org.apache.spark	spark-sketch_2.11	2.1.0.cloudera2

Project	groupId	artifactId	version
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.1.0.cloudera2
	org.apache.spark	spark-sql_2.11	2.1.0.cloudera2
	org.apache.spark	spark-streaming-flume-sink_2.11	2.1.0.cloudera2
	org.apache.spark	spark-streaming-flume_2.11	2.1.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.1.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.1.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.1.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.1.0.cloudera2
	org.apache.spark	spark-streaming_2.11	2.1.0.cloudera2
	org.apache.spark	spark-tags_2.11	2.1.0.cloudera2
	org.apache.spark	spark-unsafe_2.11	2.1.0.cloudera2
	org.apache.spark	spark-yarn_2.11	2.1.0.cloudera2

### CDS 2.1 Release 1 Maven Artifacts

The following pom fragment shows how to access a CDS 2.1 Release 1 artifact from a Maven POM.

```

<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.1.0.cloudera1</version>
  <scope>provided</scope>
</dependency>
    
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.1.0.cloudera1
	org.apache.spark	spark-core_2.11	2.1.0.cloudera1
	org.apache.spark	spark-graphx_2.11	2.1.0.cloudera1
	org.apache.spark	spark-hive-exec_2.11	2.1.0.cloudera1

Project	groupId	artifactId	version
	org.apache.spark	spark-hive_2.11	2.1.0.cloudera1
	org.apache.spark	spark-launcher_2.11	2.1.0.cloudera1
	org.apache.spark	spark-mllib-local_2.11	2.1.0.cloudera1
	org.apache.spark	spark-mllib_2.11	2.1.0.cloudera1
	org.apache.spark	spark-network-common_2.11	2.1.0.cloudera1
	org.apache.spark	spark-network-shuffle_2.11	2.1.0.cloudera1
	org.apache.spark	spark-network-yarn_2.11	2.1.0.cloudera1
	org.apache.spark	spark-repl_2.11	2.1.0.cloudera1
	org.apache.spark	spark-sketch_2.11	2.1.0.cloudera1
	org.apache.spark	spark-sql-kafka-0-10_2.11	2.1.0.cloudera1
	org.apache.spark	spark-sql_2.11	2.1.0.cloudera1
	org.apache.spark	spark-streaming-flume-sink_2.11	2.1.0.cloudera1
	org.apache.spark	spark-streaming-flume_2.11	2.1.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.11	2.1.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-10_2.11	2.1.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-8-assembly_2.11	2.1.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.1.0.cloudera1
	org.apache.spark	spark-streaming_2.11	2.1.0.cloudera1
	org.apache.spark	spark-tags_2.11	2.1.0.cloudera1
	org.apache.spark	spark-unsafe_2.11	2.1.0.cloudera1
	org.apache.spark	spark-yarn_2.11	2.1.0.cloudera1

## CDS 2.0 Powered by Apache Spark Maven Artifacts

The following tables lists the `groupId`, `artifactId`, and `version` required to access the artifacts for each CDS 2.0 Powered by Apache Spark release:

### CDS 2.0 Release 2 Maven Artifacts

The following pom fragment shows how to access a CDS 2.0 Release 2 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.0.0.cloudera2</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.0.0.cloudera2
	org.apache.spark	spark-core_2.11	2.0.0.cloudera2
	org.apache.spark	spark-graphx_2.11	2.0.0.cloudera2
	org.apache.spark	spark-hive-exec_2.11	2.0.0.cloudera2
	org.apache.spark	spark-hive_2.11	2.0.0.cloudera2
	org.apache.spark	spark-launcher_2.11	2.0.0.cloudera2
	org.apache.spark	spark-mllib-local_2.11	2.0.0.cloudera2
	org.apache.spark	spark-mllib_2.11	2.0.0.cloudera2
	org.apache.spark	spark-network-common_2.11	2.0.0.cloudera2
	org.apache.spark	spark-network-shuffle_2.11	2.0.0.cloudera2
	org.apache.spark	spark-network-yarn_2.11	2.0.0.cloudera2
	org.apache.spark	spark-repl_2.11	2.0.0.cloudera2
	org.apache.spark	spark-sketch_2.11	2.0.0.cloudera2
	org.apache.spark	spark-sql_2.11	2.0.0.cloudera2
	org.apache.spark	spark-streaming-flume-sink_2.11	2.0.0.cloudera2



Project	groupId	artifactId	version
	org.apache.spark	spark-streaming-flume_2.11	2.0.0.cloudera2
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.0.0.cloudera2
	org.apache.spark	spark-streaming_2.11	2.0.0.cloudera2
	org.apache.spark	spark-tags_2.11	2.0.0.cloudera2
	org.apache.spark	spark-unsafe_2.11	2.0.0.cloudera2
	org.apache.spark	spark-yarn_2.11	2.0.0.cloudera2

### CDS 2.0 Release 1 Maven Artifacts

The following pom fragment shows how to access a CDS 2.0 Release 1 artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.0.0.cloudera1</version>
  <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-catalyst_2.11	2.0.0.cloudera1
	org.apache.spark	spark-core_2.11	2.0.0.cloudera1
	org.apache.spark	spark-graphx_2.11	2.0.0.cloudera1
	org.apache.spark	spark-hive_2.11	2.0.0.cloudera1
	org.apache.spark	spark-launcher_2.11	2.0.0.cloudera1
	org.apache.spark	spark-mllib-local_2.11	2.0.0.cloudera1
	org.apache.spark	spark-mllib_2.11	2.0.0.cloudera1
	org.apache.spark	spark-network-common_2.11	2.0.0.cloudera1
	org.apache.spark	spark-network-shuffle_2.11	2.0.0.cloudera1
	org.apache.spark	spark-network-yarn_2.11	2.0.0.cloudera1

Project	groupId	artifactId	version
	org.apache.spark	spark-repl_2.11	2.0.0.cloudera1
	org.apache.spark	spark-sketch_2.11	2.0.0.cloudera1
	org.apache.spark	spark-sql_2.11	2.0.0.cloudera1
	org.apache.spark	spark-streaming-flume-sink_2.11	2.0.0.cloudera1
	org.apache.spark	spark-streaming-flume_2.11	2.0.0.cloudera1
	org.apache.spark	spark-streaming-kafka-0-8_2.11	2.0.0.cloudera1
	org.apache.spark	spark-streaming_2.11	2.0.0.cloudera1
	org.apache.spark	spark-tags_2.11	2.0.0.cloudera1
	org.apache.spark	spark-unsafe_2.11	2.0.0.cloudera1
	org.apache.spark	spark-yarn_2.11	2.0.0.cloudera1

## Installing or Upgrading CDS Powered by Apache Spark

**Minimum Required Role:** [Cluster Administrator](#) (also provided by **Full Administrator**)

CDS Powered by Apache Spark is distributed as two files: a [custom service descriptor](#) file and a parcel, both of which must be installed on the cluster.



**Note:** Due to the potential for confusion between *CDS Powered by Apache Spark* and the initialism *CSD*, references to the custom service descriptor (CSD) file in this documentation use the term *service descriptor*.

### Install CDS Powered by Apache Spark



**Note:**

Although Spark 1 and Spark 2 can coexist in the same CDH cluster, you cannot use multiple Spark 2 versions simultaneously in the same Cloudera Manager instance. All CDH clusters managed by the same Cloudera Manager Server must use exactly the same version of CDS Powered by Apache Spark. For example, you cannot use the built-in CDH Spark service, a CDS 2.1 service, and a CDS 2.2 service. You must choose only one CDS 2 Powered by Apache Spark release. Make sure to install or upgrade the CDS 2 [service descriptor](#) and parcels across all machines of all clusters at the same time.

CDS 2.2 and higher require JDK 8 only. If you are using CD 2.2 or higher, you must remove JDK 7 from all cluster and gateway hosts to ensure proper operation.

Follow these steps to install CDS Powered by Apache Spark:

1. Check that all the software prerequisites are satisfied. If not, you might need to upgrade or install other software components first. See [CDS Powered by Apache Spark Requirements](#) on page 7 for details.
2. Install the CDS Powered by Apache Spark service descriptor into Cloudera Manager.



**Important:**

Because CDS Powered by Apache Spark is only installable using the parcel mechanism, it can only be used on clusters managed by Cloudera Manager. Additionally, because Cloudera Manager does not support using parcels and packages in the same cluster, you cannot use CDS if you are using a package-based installation of CDH.

- a. To download the CDS Powered by Apache Spark service descriptor, in the Version Information table in [CDS Versions Available for Download](#) on page 59, click the service descriptor link for the version you want to install.
- b. Log on to the Cloudera Manager Server host, and copy the CDS Powered by Apache Spark service descriptor in the [location configured](#) for service descriptor files.
- c. Set the file ownership of the service descriptor to `cloudera-scm:cloudera-scm` with permission 644.
- d. Restart the Cloudera Manager Server with the following command:

**RHEL 7 Compatible, SLES 12, Ubuntu**

```
systemctl restart cloudera-scm-server
```


### RHEL 6 Compatible

```
service cloudera-scm-server restart
```

3. In the Cloudera Manager Admin Console, add the [CDS Powered by Apache Spark parcel repository](#) to the Remote Parcel Repository URLs in Parcel Settings as described in [Parcel Configuration Settings](#).



**Note:** If your Cloudera Manager Server does not have Internet access, you can use the CDS Powered by Apache Spark parcel files: put them into a [new parcel repository](#), and then configure the Cloudera Manager Server to target this newly created repository.

4. Download the CDS Powered by Apache Spark parcel, distribute the parcel to the hosts in your cluster, and activate the parcel. See [Managing Parcels](#).
5. [Add the Spark 2 service](#) to your cluster.
  - a. In step #1, select a dependency option:
    - HDFS, YARN, ZooKeeper: Choose this option if you do not need access to a Hive service.
    - HDFS, Hive, YARN, ZooKeeper: Hive is an optional dependency for the Spark service. If you have a Hive service and want to access Hive tables from your Spark applications, choose this option to include Hive as a dependency and have the Hive client configurations always available to Spark applications.
  - b. In step #2, when customizing the role assignments for CDS Powered by Apache Spark, add a [gateway role](#) to every host.
  - c. Note that the History Server port is 18089 instead of the usual 18088.
  - d. Complete the steps to add the Spark 2 service.
6. Return to the Home page by clicking the Cloudera Manager logo.
7. Click  to restart the cluster.

## Upgrading to CDS 2.4 Powered By Apache Spark



### Note:

Although Spark 1 and Spark 2 can coexist in the same CDH cluster, you cannot use multiple Spark 2 versions simultaneously in the same Cloudera Manager instance. All CDH clusters managed by the same Cloudera Manager Server must use exactly the same version of CDS Powered by Apache Spark. For example, you cannot use the built-in CDH Spark service, a CDS 2.1 service, and a CDS 2.2 service. You must choose only one CDS 2 Powered by Apache Spark release. Make sure to install or upgrade the CDS 2 [service descriptor](#) and parcels across all machines of all clusters at the same time.

CDS 2.2 and higher require JDK 8 only. If you are using CD 2.2 or higher, you must remove JDK 7 from all cluster and gateway hosts to ensure proper operation.

If you are already using CDS 2.0, 2.1, 2.2, or 2.3, here are the steps to upgrade to CDS 2.4 Powered by Apache Spark, while keeping any non-default configurations for Spark 2 that have already been applied:

- Remove the service descriptor JAR for the older version of CDS Powered by Apache Spark from `/opt/cloudera/csd`. Refer to [CDS Powered by Apache Spark Version, Packaging, and Download Information](#) on page 59 for the names of the JAR files corresponding to each version.
- Add the service descriptor JAR for CDS 2.4 to `/opt/cloudera/csd`. Set correct permissions and ownership.
- Restart the `cloudera-scm-server` service.
- In Cloudera Manager, deactivate the parcel corresponding to the older version of CDS.

- In Cloudera Manager, activate the parcel corresponding to CDS 2.4.
- Restart services and deploy the client configurations.
- If you are using Cloudera Data Science Workbench, note that Cloudera Data Science Workbench does not automatically detect configuration changes on the CDH cluster. Perform a full reset of Cloudera Data Science Workbench so that it can pick up any changes as a result of the upgrade. For instructions, see the [associated known issue](#) in the Cloudera Data Science Workbench documentation.

## Administering CDS Powered by Apache Spark

Most administration tasks are the same whether you are using Spark 1 or Spark 2. To configure and manage Spark, follow the procedures in the [Cloudera Enterprise Spark Guide](#).

In addition, follow these procedures that are specific to Spark 2:

### Configuring Spark 2 Tools as the Default



**Important:** If you configure Spark 2 as the default, modules such as SparkOnHBase and HiveOnSpark no longer work due to dependencies on Spark 1.6 in CDH. For more information, see [CDS Powered by Apache Spark Known Issues](#) on page 11.

When you start trying out Spark 2, you can do most of your testing by running the standard Spark 1 commands such as `pyspark` and `spark-shell` alongside their Spark 2 equivalents such as `pyspark2` and `spark2-shell`. All of these commands are represented as symbolic links in `/usr/bin`.

If you are testing a workflow that has the original command names hardcoded in other scripts, you might configure the system so that issuing the `pyspark` command really runs the `pyspark2` script, and so on for other Spark-related binaries. This change is done using the Linux `alternatives` mechanism, which keeps track of the appropriate target for each of the `/usr/bin` symlinks.

To use Spark 2 tools as the default, run the following script *on all hosts in the cluster*:

```
for binary in pyspark spark-shell spark-submit; do
  # Generate the name of the new binary e.g. pyspark2, spark2-shell, etc.
  new_binary=$(echo $binary | sed -e 's/spark/spark2/')
  # Update the old alternative to the client binary to the new client binary
  # Use priority 11 because the default priority with which these alternatives are
  # created is 10
  update-alternatives --install /usr/bin/${binary} ${binary} /usr/bin/${new_binary} 11
done
# For configuration, we need to have a separate command
# because the destination is under /etc/ instead of /usr/bin like for binaries.
# The priority is different - 52 because Cloudera Manager sets up configuration symlinks
# with priority 51.
update-alternatives --install /etc/spark/conf spark-conf /etc/spark2/conf 52
```

To remove this setting and return to using the Spark contained in CDH, run the following script *on all hosts in the cluster*. It removes the Spark 2 targets of the symlinks and points those symlinks back to the original Spark-related scripts:

```
for binary in pyspark spark-shell spark-submit; do
  new_binary=$(echo $binary | sed -e 's/spark/spark2/')
  update-alternatives --remove ${binary} /usr/bin/${new_binary}
done
update-alternatives --remove spark-conf /etc/spark2/conf
```

## Security Considerations for CDS Powered by Apache Spark

The following sections cover any special considerations for security in CDS Powered by Apache Spark, or differences in security settings between Spark 1.6 and Spark 2.

### Configuration Settings for Encryption

To enable encryption for the shuffle service with CDS Powered by Apache Spark, use the following configuration setting:

```
spark.io.encryption.enabled=true
```

This is a change from Spark 1.6, where the corresponding option is `spark.shuffle.encryption.enabled=true`.

## Running Applications with CDS Powered by Apache Spark

With CDS Powered by Apache Spark, you can run Apache Spark 2 applications locally or distributed across a cluster, either by using an interactive shell or by submitting an application. Running Spark applications interactively is commonly performed during the data-exploration phase and for ad hoc analysis.

### The Spark 2 Job Commands

With Spark 2, you use slightly different command names than in Spark 1, so that you can run both versions of Spark side-by-side without conflicts:

- `spark2-submit` instead of `spark-submit`.
- `spark2-shell` instead of `spark-shell`.
- `pyspark2` instead of `pyspark`.

For development and test purposes, you can also configure each host so that invoking the Spark 1 command name runs the corresponding Spark 2 executable. See [Configuring Spark 2 Tools as the Default](#) on page 86 for details.

### Canary Test for pyspark2 Command

The following example shows a simple `pyspark2` session that refers to the `SparkContext`, calls the `collect()` function which runs a Spark 2 job, and writes data to HDFS. This sequence of operations helps to check if there are obvious configuration issues that prevent Spark 2 jobs from working at all. For the HDFS path for the output directory, substitute a path that exists on your own system.

```
$ hdfs dfs -mkdir /user/jdoe/spark
$ pyspark2
...
SparkSession available as 'spark'.
>>> strings = ["one","two","three"]
>>> s2 = sc.parallelize(strings)
>>> s3 = s2.map(lambda word: word.upper())
>>> s3.collect()
['ONE', 'TWO', 'THREE']
>>> s3.saveAsTextFile('hdfs:///user/jdoe/spark/canary_test')
>>> quit()
$ hdfs dfs -ls /user/jdoe/spark
Found 1 items
drwxr-xr-x - jdoe spark-users 0 2016-08-26 14:41 /user/jdoe/spark/canary_test
$ hdfs dfs -ls /user/jdoe/spark/canary_test
Found 3 items
-rw-r--r-- 3 jdoe spark-users 0 2016-08-26 14:41 /user/jdoe/spark/canary_test/_SUCCESS
-rw-r--r-- 3 jdoe spark-users 4 2016-08-26 14:41
/user/jdoe/spark/canary_test/part-00000
-rw-r--r-- 3 jdoe spark-users 10 2016-08-26 14:41
/user/jdoe/spark/canary_test/part-00001
$ hdfs dfs -cat /user/jdoe/spark/canary_test/part-00000
ONE
$ hdfs dfs -cat /user/jdoe/spark/canary_test/part-00001
TWO
THREE
```

### Fetching Spark 2 Maven Dependencies

The Maven coordinates are a combination of `groupId`, `artifactId` and `version`. The `groupId` and `artifactId` are the same as for the upstream Apache Spark project. For example, for `spark-core`, `groupId` is `org.apache.spark`, and `artifactId`



is `spark-core_2.11`, both the same as the upstream project. The version is different for the Cloudera packaging: see [Using the CDS Powered by Apache Spark Maven Repository](#) on page 60 for the exact name depending on which release you are using.

## Adapting the Spark WordCount App for Spark 2

The following pom fragment shows how to access a CDS Powered by Apache Spark artifact from a Maven POM.

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-core_2.11</artifactId>
  <version>2.4.0.cloudera2</version>
  <scope>provided</scope>
</dependency>
```

Use this dependency definition to update `pom.xml` for the example described in [Developing and Running a Spark WordCount Application](#). If you are using a different CDS version, see [Using the CDS Powered by Apache Spark Maven Repository](#) on page 60.

To account for changes in the Spark 2 API, before building the example, make the following updates to `com.cloudera.sparkwordcount.JavaWordCount`:

- Add `import java.util.Iterator;`
- Replace all instances of `Iterable` with `Iterator`.
- Perform the following replacements:
  - `return Arrays.asList(s.split(" "));` to `return Arrays.asList(s.split(" ")).iterator();`
  - `return chars;` to `return chars.iterator();`

## Accessing the Spark 2 History Server

The Spark 2 history server is available on port 18089, rather than port 18088 as with the Spark 1 history server.

## Integrating CDS Powered by Apache Spark with Apache Kafka

**Minimum Required Role:** [Cluster Administrator](#) (also provided by **Full Administrator**)

Version 2.1 Release 1 and higher of CDS Powered by Apache Spark includes an Apache Kafka integration feature that uses the new Kafka consumer API. This new Kafka consumer API supports reading data from secure Kafka clusters. In this context, secure clusters are those that are authenticated by Kerberos, and optionally using TLS/SSL for wire encryption.

### Requirements

To read data securely from Kafka, or to use the new Spark-Kafka integration that uses the new Kafka consumer API, requires the following software versions:

- CDS 2.1 Release 1 or higher.
- CDK Powered by Apache Kafka 2.1 or higher. (Although the Kafka consumer API is available starting in CDK Powered by Apache Kafka 2.0, the Spark integration requires Kafka 2.1.)

### Running Spark Jobs that Integrate with Kafka

To run jobs that use the new Kafka integration, you can use one of the following two techniques.



**Note:**

If you do not intend to use the new Kafka consumer integration and are using the existing Kafka integration (either the Kafka receiver or the direct connector), you do not need to do anything regardless of what Kafka version you are using. If you want to use the new Kafka consumer API integration, use one of the following two techniques to make sure you are using the newer Kafka jars in the Spark classpath.

#### Technique #1: Set `SPARK_KAFKA_VERSION` environment variable

When running jobs that require the new Kafka integration, set `SPARK_KAFKA_VERSION=0.10` in the shell before launching `spark-submit`. Use the appropriate environment variable syntax for your shell, such as:

```
# Set the environment variable for the duration of your shell session:
export SPARK_KAFKA_VERSION=0.10
spark-submit arguments

# Or:

# Set the environment variable for the duration of a single command:
SPARK_KAFKA_VERSION=0.10 spark-submit arguments
```

#### Technique #2: Set `spark_kafka_version` setting through Cloudera Manager

Set `spark_kafka_version` configuration in Cloudera Manager's Spark 2 service to be 0.10 and redeploy the client configuration. No need to source any environment variables when launching `spark-submit`.

Technique #2 is preferable if you have upgraded your Kafka brokers to CDK Powered by Apache Kafka 2.1 or higher, and do not intend to have your Spark jobs communicate with brokers running a version of Kafka prior to CDK Powered by Apache Kafka 2.1.

If you modify the default Kafka version to 0.10 by using technique #2, you can connect to old Kafka brokers (for example, one based on CDK Powered by Apache Kafka 2.0) by setting `SPARK_KAFKA_VERSION=0.9` when running your application.



**Note:** To keep compatibility and not break any existing users, the default is to use older (CDK Powered by Apache Kafka 2.0-based) client jars in the Spark classpath. Those jars let users communicate with older (CDK Powered by Apache Kafka 2.0) brokers but also newer brokers (CDK Powered by Apache Kafka 2.1) brokers as long as the old Kafka consumer APIs are being used. However, to use the new kafka consumer API integration, one of the two steps are required. This is to ensure that the new Kafka client jars (from CDK Powered by Apache Kafka 2.1) are in the Spark classpath. It is not practical to put both Kafka 2.0 client jars and Kafka 2.1 clients jars on the Spark classpath due to incompatibilities; the engineering best practice is to only use one version of any jar file in the classpath.

## Building Applications

To build applications against the new Kafka integration, you can add the dependency by using the following Maven coordinates:

```
<dependency>
  <groupId>org.apache.spark</groupId>
  <artifactId>spark-streaming-kafka-0-10_2.11</artifactId>
  <version>2.4.0.cloudera2</version>
</dependency>
```

## Reading from Authorized Kafka

To read from a Kafka cluster authorized by Sentry, give privileges to your consumer group as described in [Configuring Kafka Security](#). You must also grant privileges to another consumer group `spark-executor-<your_consumer_group>` in the same way. This is because the driver uses the consumer group specified in your app, but the executors use a different consumer group, which is hardcoded to `spark-executor-<your_consumer_group>`.

## Troubleshooting CDS Powered by Apache Spark

Troubleshooting for CDS Powered by Apache Spark mainly involves checking configuration settings and application code to diagnose performance and scalability issues.

### Commercial support for GA version

Cloudera customers with commercial support can now use normal support channels for CDS Powered by Apache Spark.

### Error instantiating Hive metastore class

A Hive compatibility issue in CDS 2.0 Release 1 affects CDH 5.10.1 and higher, CDH 5.9.2 and higher, CDH 5.8.5 and higher, and CDH 5.7.6 and higher. If you are using one of these CDH versions, you must upgrade to the CDS 2.0 Release 2 or higher parcel, to avoid Spark 2 job failures when using Hive functionality.

When you encounter a problem due to the Hive compatibility issue, the error stack starts like this:

```
java.lang.RuntimeException: Unable to instantiate
  org.apache.hadoop.hive ql.metadata.SessionHiveMetaStoreClient
  at org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance
    (MetaStoreUtils.java:1545)
  at org.apache.hadoop.hive.metastore
    .RetryingMetaStoreClient.<init>(RetryingMetaStoreClient.
```

The solution is to upgrade to CDS 2.0 Release 2 or higher.

### Wrong version of Python

When you use CDS Powered by Apache Spark with Python 2.x, you must use Python 2.7 or higher. You might need to install a new version of Python on all hosts in the cluster, because some Linux distributions come with Python 2.6 by default. If the right level of Python is not picked up by default, set the `PYSPARK_PYTHON` and `PYSPARK_DRIVER_PYTHON` environment variables to point to the correct Python executable before running the `pyspark2` command.

### API changes that are not backward-compatible

Between Spark 1.6 and Spark 2.0, some APIs have changed in ways that are not backward compatible. Recompile all applications to take advantage of Spark 2 capabilities. For any compilation errors, check if the corresponding function has changed in Spark 2, and if so, change your code to use the latest function name, parameters, and return type.

### A Spark component does not work or is unstable

Certain components from the Spark ecosystem are explicitly not supported with CDS Powered by Apache Spark. Check against the compatibility matrix for Spark to make sure the components you are using are all intended to work with CDS Powered by Apache Spark and CDH.

## Frequently Asked Questions about CDS Powered by Apache Spark

**Note:**

This documentation refers to CDS 2.4 Powered by Apache Spark. This component is generally available and is supported on CDH 5.9 and higher.

A Hive compatibility issue in CDS 2.0 Release 1 affects CDH 5.10.1 and higher, CDH 5.9.2 and higher, CDH 5.8.5 and higher, and CDH 5.7.6 and higher. If you are using one of these CDH versions, you must upgrade to the CDS 2.0 Release 2 or higher parcel, to avoid Spark 2 job failures when using Hive functionality.

This Frequently Asked Questions (FAQ) page covers general information about CDS Powered by Apache Spark, coexistence with Spark 1, and other questions that are relevant for early adopters of the latest Spark 2 features.

### Running Spark 1 and Spark 2 Side-by-Side

The Spark 2 service does not conflict with Spark 1 if it is installed. The history server uses a different port. Spark 2 shares the Spark 1 shuffle service if already available, or installs the shuffle service if not.

Although Spark 1 and Spark 2 can coexist in the same CDH cluster, you cannot use multiple Spark 2 versions simultaneously in the same Cloudera Manager instance. All CDH clusters managed by the same Cloudera Manager Server must use exactly the same version of CDS Powered by Apache Spark. For example, you cannot use the built-in CDH Spark service, a CDS 2.1 service, and a CDS 2.2 service. You must choose only one CDS 2 Powered by Apache Spark release. Make sure to install or upgrade the CDS 2 [service descriptor](#) and parcels across all machines of all clusters at the same time.

### Why doesn't feature or library XYZ work?

A number of features, components, libraries, and integration points from Spark 1.6 are not supported with CDS Powered by Apache Spark. See [CDS Powered by Apache Spark Known Issues](#) on page 11 for details.

## Appendix: Apache License, Version 2.0

### SPDX short identifier: Apache-2.0

Apache License  
Version 2.0, January 2004  
<http://www.apache.org/licenses/>

#### TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

##### 1. Definitions.

"License" shall mean the terms and conditions for use, reproduction, and distribution as defined by Sections 1 through 9 of this document.

"Licensor" shall mean the copyright owner or entity authorized by the copyright owner that is granting the License.

"Legal Entity" shall mean the union of the acting entity and all other entities that control, are controlled by, or are under common control with that entity. For the purposes of this definition, "control" means (i) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (ii) ownership of fifty percent (50%) or more of the outstanding shares, or (iii) beneficial ownership of such entity.

"You" (or "Your") shall mean an individual or Legal Entity exercising permissions granted by this License.

"Source" form shall mean the preferred form for making modifications, including but not limited to software source code, documentation source, and configuration files.

"Object" form shall mean any form resulting from mechanical transformation or translation of a Source form, including but not limited to compiled object code, generated documentation, and conversions to other media types.

"Work" shall mean the work of authorship, whether in Source or Object form, made available under the License, as indicated by a copyright notice that is included in or attached to the work (an example is provided in the Appendix below).

"Derivative Works" shall mean any work, whether in Source or Object form, that is based on (or derived from) the Work and for which the editorial revisions, annotations, elaborations, or other modifications represent, as a whole, an original work of authorship. For the purposes of this License, Derivative Works shall not include works that remain separable from, or merely link (or bind by name) to the interfaces of, the Work and Derivative Works thereof.

"Contribution" shall mean any work of authorship, including the original version of the Work and any modifications or additions to that Work or Derivative Works thereof, that is intentionally submitted to Licensor for inclusion in the Work by the copyright owner or by an individual or Legal Entity authorized to submit on behalf of the copyright owner. For the purposes of this definition, "submitted" means any form of electronic, verbal, or written communication sent to the Licensor or its representatives, including but not limited to communication on electronic mailing lists, source code control systems, and issue tracking systems that are managed by, or on behalf of, the Licensor for the purpose of discussing and improving the Work, but excluding communication that is conspicuously marked or otherwise designated in writing by the copyright owner as "Not a Contribution."

"Contributor" shall mean Licensor and any individual or Legal Entity on behalf of whom a Contribution has been received by Licensor and subsequently incorporated within the Work.

##### 2. Grant of Copyright License.

Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form.

##### 3. Grant of Patent License.

Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable (except as stated in this section) patent license to make, have made, use, offer to sell, sell, import, and otherwise transfer the Work, where such license applies only to those patent claims

licensable by such Contributor that are necessarily infringed by their Contribution(s) alone or by combination of their Contribution(s) with the Work to which such Contribution(s) was submitted. If You institute patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Work or a Contribution incorporated within the Work constitutes direct or contributory patent infringement, then any patent licenses granted to You under this License for that Work shall terminate as of the date such litigation is filed.

#### 4. Redistribution.

You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

1. You must give any other recipients of the Work or Derivative Works a copy of this License; and
2. You must cause any modified files to carry prominent notices stating that You changed the files; and
3. You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and
4. If the Work includes a "NOTICE" text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.

You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

#### 5. Submission of Contributions.

Unless You explicitly state otherwise, any Contribution intentionally submitted for inclusion in the Work by You to the Licensor shall be under the terms and conditions of this License, without any additional terms or conditions. Notwithstanding the above, nothing herein shall supersede or modify the terms of any separate license agreement you may have executed with Licensor regarding such Contributions.

#### 6. Trademarks.

This License does not grant permission to use the trade names, trademarks, service marks, or product names of the Licensor, except as required for reasonable and customary use in describing the origin of the Work and reproducing the content of the NOTICE file.

#### 7. Disclaimer of Warranty.

Unless required by applicable law or agreed to in writing, Licensor provides the Work (and each Contributor provides its Contributions) on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied, including, without limitation, any warranties or conditions of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR A PARTICULAR PURPOSE. You are solely responsible for determining the appropriateness of using or redistributing the Work and assume any risks associated with Your exercise of permissions under this License.

#### 8. Limitation of Liability.

In no event and under no legal theory, whether in tort (including negligence), contract, or otherwise, unless required by applicable law (such as deliberate and grossly negligent acts) or agreed to in writing, shall any Contributor be liable to You for damages, including any direct, indirect, special, incidental, or consequential damages of any character arising as a result of this License or out of the use or inability to use the Work (including but not limited to damages for loss of goodwill, work stoppage, computer failure or malfunction, or any and all other commercial damages or losses), even if such Contributor has been advised of the possibility of such damages.

#### 9. Accepting Warranty or Additional Liability.

## Appendix: Apache License, Version 2.0

While redistributing the Work or Derivative Works thereof, You may choose to offer, and charge a fee for, acceptance of support, warranty, indemnity, or other liability obligations and/or rights consistent with this License. However, in accepting such obligations, You may act only on Your own behalf and on Your sole responsibility, not on behalf of any other Contributor, and only if You agree to indemnify, defend, and hold each Contributor harmless for any liability incurred by, or claims asserted against, such Contributor by reason of your accepting any such warranty or additional liability.

END OF TERMS AND CONDITIONS

### APPENDIX: How to apply the Apache License to your work

To apply the Apache License to your work, attach the following boilerplate notice, with the fields enclosed by brackets "[]" replaced with your own identifying information. (Don't include the brackets!) The text should be enclosed in the appropriate comment syntax for the file format. We also recommend that a file or class name and description of purpose be included on the same "printed page" as the copyright notice for easier identification within third-party archives.

```
Copyright [yyyy] [name of copyright owner]

Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License.
```