Spark Release 3.4.4

Spark 3.4.4 is the last maintenance release containing security and correctness fixes. This release is based on the branch-3.4 maintenance branch of Spark. We strongly recommend all 3.4 users to upgrade to this stable release.

Notable changes

- [SPARK-43242] (https://issues.apache.org/jira/browse/SPARK-43242): Fix throw 'Unexpected type of BlockId' in shuffle corruption diagnose
- [SPARK-45988] (https://issues.apache.org/jira/browse/SPARK-45988): Fix typehints to handle `list` GenericAlias in Python 3.11+
- [SPARK-46535] (https://issues.apache.org/jira/browse/SPARK-46535): Fix NPE when describe extended a column without col stats
- [SPARK-46957] (https://issues.apache.org/jira/browse/SPARK-46957): Decommission migrated shuffle files should be able to cleanup from executor
- [SPARK-47129] (https://issues.apache.org/jira/browse/SPARK-47129): Make ResolveRelations cache connect plan properly
- [SPARK-47172] (https://issues.apache.org/jira/browse/SPARK-47172): Add support for AES-GCM for RPC encryption
- [SPARK-47828] (https://issues.apache.org/jira/browse/SPARK-47828): DataFrameWriterV2.overwrite fails with invalid plan
- [SPARK-47895] (https://issues.apache.org/jira/browse/SPARK-47895): group by all should be idempotent
- [SPARK-47897] (https://issues.apache.org/jira/browse/SPARK-47897): Fix ExpressionSet performance regression in scala 2.12
- [SPARK-47927] (https://issues.apache.org/jira/browse/SPARK-47927): Fix nullability attribute in UDF decoder
- [SPARK-48016] (https://issues.apache.org/jira/browse/SPARK-48016): Fix a bug in try_divide function when with decimals
- [SPARK-48019] (https://issues.apache.org/jira/browse/SPARK-48019): Fix incorrect behavior in ColumnVector/ColumnarArray with dictionary and nulls
- [SPARK-48037] (https://issues.apache.org/jira/browse/SPARK-48037): Fix SortShuffleWriter lacks shuffle write related metrics resulting in potentially inaccurate data

- [SPARK-48081] (https://issues.apache.org/jira/browse/SPARK-48081): Fix ClassCastException in NTile.checkInputDataTypes() when argument is non-foldable or of wrong type
- [SPARK-48105] (https://issues.apache.org/jira/browse/SPARK-48105): Fix the race condition between state store unloading and snapshotting
- [SPARK-48128] (https://issues.apache.org/jira/browse/SPARK-48128): For BitwiseCount / bit_count expression, fix codegen syntax error for boolean type inputs
- [SPARK-48155] (https://issues.apache.org/jira/browse/SPARK-48155): AQEPropagateEmptyRelation for join should check if remain child is just BroadcastQueryStageExec
- [SPARK-48172] (https://issues.apache.org/jira/browse/SPARK-48172): Fix escaping issues in JDBCDialects
- [SPARK-48248] (https://issues.apache.org/jira/browse/SPARK-48248): Fix nested array to respect legacy conf of inferArrayTypeFromFirstElement
- [SPARK-48292] (https://issues.apache.org/jira/browse/SPARK-48292): Revert SPARK-39195 Spark OutputCommitCoordinator should abort stage when committed file not consistent with task status
- [SPARK-48484] (https://issues.apache.org/jira/browse/SPARK-48484): Fix: V2Write use the same TaskAttemptId for different task attempts
- [SPARK-48642] (https://issues.apache.org/jira/browse/SPARK-48642): False SparkOutOfMemoryError caused by killing task on spilling
- [SPARK-48710] (https://issues.apache.org/jira/browse/SPARK-48710): Limit NumPy version to supported range (>=1.15,<2)
- [SPARK-48759] (https://issues.apache.org/jira/browse/SPARK-48759): Add migration doc for CREATE TABLE AS SELECT behavior change behavior change since Spark 3.4
- [SPARK-48791] (https://issues.apache.org/jira/browse/SPARK-48791): Fix perf regression caused by the accumulators registration overhead using CopyOnWriteArrayList
- [SPARK-48930] (https://issues.apache.org/jira/browse/SPARK-48930): Redact `awsAccessKeyId` by including `accesskey` pattern
- [SPARK-48934] (https://issues.apache.org/jira/browse/SPARK-48934): Python datetime types converted incorrectly for setting timeout in applyInPandasWithState
- [SPARK-48965] (https://issues.apache.org/jira/browse/SPARK-48965): Use the correct schema in `Dataset#toJSON`

- [SPARK-48991] (https://issues.apache.org/jira/browse/SPARK-48991): Move path initialization into try-catch block in FileStreamSink.hasMetadata
- [SPARK-49000] (https://issues.apache.org/jira/browse/SPARK-49000): Fix "select count(distinct 1) from t" where t is empty table by expanding RewriteDistinctAggregates
- [SPARK-49005] (https://issues.apache.org/jira/browse/SPARK-49005): Use `17-jammy` tag instead of `17-jre` to prevent Python 3.12
- [SPARK-49038] (https://issues.apache.org/jira/browse/SPARK-49038): SQLMetric should report the raw value in the accumulator update event
- [SPARK-49039] (https://issues.apache.org/jira/browse/SPARK-49039): Reset checkbox when executor metrics are loaded in the Stages tab
- [SPARK-49094] (https://issues.apache.org/jira/browse/SPARK-49094): Fix ignoreCorruptFiles non-functioning for hive orc impl with mergeSchema off
- [SPARK-49176] (https://issues.apache.org/jira/browse/SPARK-49176): Fix `spark.ui.custom.executor.log.url` docs by adding K8s
- [SPARK-49179] (https://issues.apache.org/jira/browse/SPARK-49179): Fix v2 multi bucketed inner joins throw AssertionError
- [SPARK-49182] (https://issues.apache.org/jira/browse/SPARK-49182): Stop publish site/docs/{version}/api/python/_sources dir
- [SPARK-49193] (https://issues.apache.org/jira/browse/SPARK-49193): Improve the performance of RowSetUtils.toColumnBasedSet
- [SPARK-49197] (https://issues.apache.org/jira/browse/SPARK-49197): Redact `Spark Command` output in `launcher` module
- [SPARK-49261] (https://issues.apache.org/jira/browse/SPARK-49261): Don't replace literals in aggregate expressions with group-by expressions
- [SPARK-49352] (https://issues.apache.org/jira/browse/SPARK-49352): Avoid redundant array transform for identical expression
- [SPARK-49385] (https://issues.apache.org/jira/browse/SPARK-49385): Fix `getReusablePVCs` to use `podCreationTimeout` instead of `podAllocationDelay`
- [SPARK-49408] (https://issues.apache.org/jira/browse/SPARK-49408): Use IndexedSeq in ProjectingInternalRow
- [SPARK-49595] (https://issues.apache.org/jira/browse/SPARK-49595): Fix `DataFrame.unpivot/melt` in Spark Connect Scala Client

- [SPARK-49628] (https://issues.apache.org/jira/browse/SPARK-49628): ConstantFolding should copy stateful expression before evaluating
- [SPARK-49750] (https://issues.apache.org/jira/browse/SPARK-49750): Mention delegation token support in K8s mode
- [SPARK-49760] (https://issues.apache.org/jira/browse/SPARK-49760): Correct handling of `SPARK_USER` env variable override in app master
- [SPARK-49804] (https://issues.apache.org/jira/browse/SPARK-49804): Fix to use the exit code of executor container always
- [SPARK-49836] (https://issues.apache.org/jira/browse/SPARK-49836): Fix possibly broken query when window is provided to window/session_window fn
- [SPARK-49843] (https://issues.apache.org/jira/browse/SPARK-49843): Fix change comment on char/varchar columns
- [SPARK-49959] (https://issues.apache.org/jira/browse/SPARK-49959): Fix ColumnarArray.copy() to read nulls from the correct offset
- [SPARK-50021] (https://issues.apache.org/jira/browse/SPARK-50021): Fix `ApplicationPage` to hide App UI links when UI is disabled
- [SPARK-50022] (https://issues.apache.org/jira/browse/SPARK-50022): Fix `MasterPage` to hide App UI links when UI is disabled

Dependency Changes

While being a maintenance release we did still upgrade some dependencies in this release they are:

- [SPARK-43394] (https://issues.apache.org/jira/browse/SPARK-43394): Upgrade maven to 3.8.8
- [SPARK-45590] (https://issues.apache.org/jira/browse/SPARK-45590): Upgrade okio to 1.17.6 from 1.15.0

You can consult JIRA for the detailed changes (https://s.apache.org/spark-3.4.4).

We would like to acknowledge all community members for contributing patches to this release.

Latest News

Spark 3.4.4 released (/news/spark-3-4-4-released.html) (Oct 27, 2024)

Preview release of Spark 4.0 (/news/spark-4.0.0-preview2.html) (Sep 26, 2024)

(https://www.apache.org/events/current-event.html)

DOWNLOAD SPARK (/DOWNLOADS.HTML)

Built-in Libraries:

SQL and DataFrames (/sql/)
Spark Streaming (/streaming/)
MLlib (machine learning) (/mllib/)
GraphX (graph) (/graphx/)

Third-Party Projects (/third-party-projects.html)