Spark 1.0.1发布了

2014年7月11日，Spark 1.0.1已经发布了，原文如下：

> We are happy to announce the availability of Spark 1.0.1! This release includes contributions from 70 developers. Spark 1.0.0 includes fixes across several areas of Spark, including the core API, PySpark, and MLlib. It also includes new features in Spark's (alpha) SQL library, including support for JSON data and performance and stability fixes.
>
> Visit the release notes to read about this release or download the release today.

下面是Spark 1.0.1 Release Notes 。

Bug

- [SPARK-1097] - ConcurrentModificationException
- [SPARK-1112] - When spark.akka.frameSize > 10, task results bigger than 10MiB block execution
- [SPARK-1199] - Type mismatch in Spark shell when using case class defined in shell
- [SPARK-1468] - The hash method used by partitionBy in Pyspark doesn't deal with None correctly.
- [SPARK-1518] - Spark master doesn't compile against hadoop-common trunk
- [SPARK-1704] - Support EXPLAIN in Spark SQL
- [SPARK-1712] - ParallelCollectionRDD operations hanging forever without any error messages
- [SPARK-1715] - Ensure actor is self-contained in DAGScheduler
- [SPARK-1749] - DAGScheduler supervisor strategy broken with Mesos
- [SPARK-1826] - Some bad head notations in sparksql
- [SPARK-1831] - add the security guide to the "More" drop down menu
- [SPARK-1838] - On a YARN cluster, Spark doesn't run on local mode
- [SPARK-1850] - Bad exception if multiple jars exist when running PySpark
- [SPARK-1852] - SparkSQL Queries with Sorts run before the user asks them to
- [SPARK-1866] - Closure cleaner does not null shadowed fields when outer scope is referenced
- [SPARK-1869] - `spark-shell --help` fails if called from outside spark home
- [SPARK-1901] - Standalone worker update exector's state ahead of executor process exit
- [SPARK-1912] - Compression memory issue during reduce
- [SPARK-1913] - Parquet table column pruning error caused by filter pushdown
- [SPARK-1914] - Simplify CountFunction not to traverse to evaluate all child expressions.

- [SPARK-1915] - AverageFunction should not count if the evaluated value is null.
- [SPARK-1916] - SparkFlumeEvent with body bigger than 1020 bytes are not read properly
- [SPARK-1917] - PySpark fails to import functions from {{scipy.special}}
- [SPARK-1922] - hql query throws "RuntimeException: Unsupported dataType" if struct field of a table has a column with underscore in name
- [SPARK-1926] - Nullability of Max/Min/First should be true.
- [SPARK-1930] - The Container is running beyond physical memory limits, so as to be killed.
- [SPARK-1931] - Graph.partitionBy does not reconstruct routing tables
- [SPARK-1933] - FileNotFoundException when a directory is passed to SparkContext.addJar/addFile
- [SPARK-1935] - Explicitly add commons-codec 1.5 as a dependency
- [SPARK-1938] - ApproxCountDistinctMergeFunction should return Int value.
- [SPARK-1958] - Calling .collect() on a SchemaRDD should call executeCollect() on the underlying query plan.
- [SPARK-1959] - String "NULL" is interpreted as null value
- [SPARK-1964] - Timestamp missing from HiveMetastore types parser
- [SPARK-1976] - misleading streaming document
- [SPARK-1977] - mutable.BitSet in ALS not serializable with KryoSerializer
- [SPARK-1978] - In some cases, spark-yarn does not automatically restart the failed container
- [SPARK-1984] - Maven build requires SCALA_HOME to be set even though it's not needed
- [SPARK-1995] - Add native support for UPPER() LOWER() and MIN() MAX()
- [SPARK-1998] - SparkFlumeEvent with body bigger than 1020 bytes are not read properly
- [SPARK-1999] - UI : StorageLevel in storage tab and RDD Storage Info never changes
- [SPARK-2003] - SparkContext(SparkConf) doesn't work in pyspark
- [SPARK-2009] - Key not found exception when slow receiver starts
- [SPARK-2025] - EdgeRDD persists after pregel iteration
- [SPARK-2030] - Bump SparkBuild.scala version number of branch-1.0 to 1.0.1-SNAPSHOT.
- [SPARK-2034] - KafkaInputDStream doesn't close resources and may prevent JVM shutdown
- [SPARK-2036] - CaseConversionExpression should check if the evaluated value is null.
- [SPARK-2041] - Exception when querying when tableName == columnName
- [SPARK-2043] - ExternalAppendOnlyMap doesn't always find matching keys
- [SPARK-2050] - LIKE, RLIKE, IN, BETWEEN and DIV in HQL should not be case sensitive
- [SPARK-2057] - run-example can only be run within spark_home
- [SPARK-2059] - Unresolved Attributes should cause a failure before execution time
- [SPARK-2067] - Spark logo in application UI uses absolute path
- [SPARK-2075] - Hadoop1 distribution of 1.0.0 does not contain classes expected by the Maven 1.0.0 artifact
- [SPARK-2080] - Yarn: history UI link missing, wrong reported user
- [SPARK-2088] - NPE in toString when creationSiteInfo is null after deserialization

- [SPARK-2091] - pyspark/mllib is not compatible with numpy-1.4
- [SPARK-2093] - NullPropagation should use exact type value.
- [SPARK-2107] - FilterPushdownSuite imports Junit and leads to compilation error
- [SPARK-2108] - Mark SparkContext methods that return block information as developer API's
- [SPARK-2109] - Setting SPARK_MEM for bin/pyspark does not work.
- [SPARK-2128] - No plan for DESCRIBE
- [SPARK-2135] - InMemoryColumnarScan does not get planned correctly
- [SPARK-2137] - Timestamp UDFs broken
- [SPARK-2140] - yarn stable client doesn't properly handle MEMORY_OVERHEAD for AM
- [SPARK-2144] - SparkUI Executors tab displays incorrect RDD blocks
- [SPARK-2146] - Fix the takeOrdered doc
- [SPARK-2147] - Master UI forgets about Executors when application exits cleanly
- [SPARK-2151] - spark-submit issue (int format expected for memory parameter)
- [SPARK-2152] - the error of comput rightNodeAgg about Decision tree algorithm in Spark MLlib
- [SPARK-2156] - When the size of serialized results for one partition is slightly smaller than 10MB (the default akka.frameSize), the execution blocks
- [SPARK-2164] - Applying UDF on a struct throws a MatchError
- [SPARK-2172] - PySpark cannot import mllib modules in YARN-client mode
- [SPARK-2176] - Extra unnecessary exchange operator in the result of an explain command
- [SPARK-2177] - describe table result contains only one column
- [SPARK-2184] - AddExchange isn't idempotent
- [SPARK-2187] - Explain command should not run the optimizer twice
- [SPARK-2191] - Double execution with CREATE TABLE AS SELECT
- [SPARK-2195] - Parquet extraMetadata can contain key information
- [SPARK-2196] - Fix nullability of CaseWhen.
- [SPARK-2204] - Scheduler for Mesos in fine-grained mode launches tasks on wrong executors
- [SPARK-2209] - Cast shouldn't do null check twice
- [SPARK-2210] - cast to boolean on boolean value gets turned into NOT((boolean_condition) = 0)
- [SPARK-2218] - rename Equals to EqualTo in Spark SQL expressions
- [SPARK-2241] - EC2 script should handle quoted arguments correctly
- [SPARK-2251] - MLLib Naive Bayes Example SparkException: Can only zip RDDs with same number of elements in each partition
- [SPARK-2252] - mathjax doesn't work in HTTPS (math formulas not rendered)
- [SPARK-2257] - The algorithm of ALS in mlib lacks a parameter
- [SPARK-2259] - Spark submit documentation for --deploy-mode is highly misleading
- [SPARK-2263] - Can't insert Map<K, V> values to Hive tables
- [SPARK-2266] - Log page on Worker UI displays "Some(app-id)"
- [SPARK-2267] - Log exception when TaskResultGetter fails to fetch/deserialize task result
- [SPARK-2270] - Kryo cannot serialize results returned by asJavaIterable (and thus groupBy/cogroup are broken in Java APIs when Kryo is used)

- [SPARK-2282] - PySpark crashes if too many tasks complete quickly
- [SPARK-2283] - PruningSuite fails if run right after HiveCompatibilitySuite
- [SPARK-2284] - Failed tasks reported as success if the failure reason is not ExceptionFailure
- [SPARK-2289] - Remove use of spark.worker.instances
- [SPARK-2307] - SparkUI Storage page cached statuses incorrect
- [SPARK-2328] - Add execution of `SHOW TABLES` before `TestHive.reset()`.
- [SPARK-2342] - Evaluation helper's output type doesn't conform to input type
- [SPARK-2349] - Fix NPE in ExternalAppendOnlyMap
- [SPARK-2350] - Master throws NPE
- [SPARK-2404] - spark-submit and spark-class may overwrite the already defined SPARK_HOME
- [SPARK-2417] - Decision tree tests are failing

Documentation

- [SPARK-1944] - Document --verbose in spark-shell -h

Improvement

- [SPARK-937] - Executors that exit cleanly should not have KILLED status
- [SPARK-1293] - Support for reading/writing complex types in Parquet
- [SPARK-1461] - Support Short-circuit Expression Evaluation
- [SPARK-1487] - Support record filtering via predicate pushdown in Parquet
- [SPARK-1495] - support leftsemijoin for sparkSQL
- [SPARK-1508] - Add support for reading from SparkConf
- [SPARK-1516] - Yarn Client should not call System.exit, should throw exception instead.
- [SPARK-1519] - support minPartitions parameter of wholeTextFiles() in pyspark
- [SPARK-1669] - SQLContext.cacheTable() should be idempotent
- [SPARK-1677] - Allow users to avoid Hadoop output checks if desired
- [SPARK-1790] - Update EC2 scripts to support r3 instance types
- [SPARK-1907] - spark-submit: add exec at the end of the script
- [SPARK-1947] - Child of SumDistinct or Average should be widened to prevent overflows the same as Sum.
- [SPARK-1990] - spark-ec2 should only need Python 2.6, not 2.7
- [SPARK-1992] - Support for Pivotal HD in the Maven build
- [SPARK-2042] - Take triggers unneeded shuffle.
- [SPARK-2053] - Add Catalyst expression for CASE WHEN
- [SPARK-2058] - SPARK_CONF_DIR should override all present configs
- [SPARK-2076] - Push Down the Predicate & Join Filter for OuterJoin
- [SPARK-2079] - Support batching when serializing SchemaRDD to Python
- [SPARK-2081] - Undefine output() from the abstract class Command and implement it in concrete subclasses

- [SPARK-2148] - Document custom class as key needing equals() AND hashcode()
- [SPARK-2161] - UI should remember executors that have been removed
- [SPARK-2163] - Set ``setConvergenceTol'' with a parameter of type Double instead of Int
- [SPARK-2186] - Spark SQL DSL support for simple aggregations such as SUM and AVG
- [SPARK-2225] - Turn HAVING without GROUP BY into WHERE
- [SPARK-2254] - ScalaRefection should mark primitive types as non-nullable.
- [SPARK-2286] - Report exception/errors for failed tasks that are not ExceptionFailure
- [SPARK-2287] - Make ScalaReflection be able to handle Generic case classes.
- [SPARK-2295] - Make JavaBeans nullability stricter.
- [SPARK-2366] - Add column pruning for the right side of LeftSemi join.
- [SPARK-2388] - Streaming from multiple different Kafka topics is problematic

New Feature

- [SPARK-1741] - Add predict(JavaRDD) to predictive models
- [SPARK-1830] - Deploy failover, Make Persistence engine and LeaderAgent Pluggable.
- [SPARK-1968] - SQL commands for caching tables
- [SPARK-2060] - Querying JSON Datasets with SQL and DSL in Spark SQL