Spark 1.6.1正式发布

Spark 1.6.1于2016年3月11日正式发布，此版本主要是维护版本，主要涉及稳定性修复，并不涉及到大的修改。推荐所有使用1.6.0的用户升级到此版本。

Spark 1.6.1主要修复的bug包括：
1、当写入数据到含有大量分区表时出现的OOM：SPARK-12546
2、实验性Dataset API的许多bug修复：SPARK-12478, SPARK-12696, SPARK-13101, SPARK-12932

完整的Bug修改列表可见： http://s.apache.org/spark-1.6.1

Release Notes - Spark - Version 1.6.1

Sub-task

- [SPARK-11031] - SparkR str() method on DataFrame objects
- [SPARK-12393] - Add read.text and write.text for SparkR

Bug

- [SPARK-7615] - MLLIB Word2Vec wordVectors divided by Euclidean Norm equals to zero
- [SPARK-9844] - File appender race condition during SparkWorker shutdown
- [SPARK-10524] - Decision tree binary classification with ordered categorical features: incorrect centroid
- [SPARK-10847] - Pyspark - DataFrame - Optional Metadata with `None` triggers cryptic failure
- [SPARK-11394] - PostgreDialect cannot handle BYTE types
- [SPARK-11624] - Spark SQL CLI will set sessionstate twice
- [SPARK-11972] - [Spark SQL] the value of 'hiveconf' parameter in CLI can't be got after enter spark-sql session
- [SPARK-12006] - GaussianMixture.train crashes if an initial model is not None
- [SPARK-12010] - Spark JDBC requires support for column-name-free INSERT syntax
- [SPARK-12016] - word2vec load model can't use findSynonyms to get words
- [SPARK-12026] - ChiSqTest gets slower and slower over time when number of features is large
- [SPARK-12268] - pyspark shell uses execfile which breaks python3 compatibility
- [SPARK-12300] - Fix schema inferance on local collections
- [SPARK-12316] - Stack overflow with endless call of `Delegation token thread` when application end.

- [SPARK-12327] - lint-r checks fail with commented code
- [SPARK-12346] - GLM summary crashes with NoSuchElementException if attributes are missing names
- [SPARK-12363] - PowerIterationClustering test case failed if we deprecated KMeans.setRuns
- [SPARK-12399] - Display correct error message when accessing REST API with an unknown app Id
- [SPARK-12424] - The implementation of ParamMap#filter is wrong.
- [SPARK-12453] - Spark Streaming Kinesis Example broken due to wrong AWS Java SDK version
- [SPARK-12470] - Incorrect calculation of row size in o.a.s.sql.catalyst.expressions.codegen.GenerateUnsafeRowJoiner
- [SPARK-12477] - [SQL] Tungsten projection fails for null values in array fields
- [SPARK-12478] - Dataset fields of product types can't be null
- [SPARK-12486] - Executors are not always terminated successfully by the worker.
- [SPARK-12489] - Fix minor issues found by Findbugs
- [SPARK-12499] - make_distribution should not override MAVEN_OPTS
- [SPARK-12502] - Script /dev/run-tests fails when IBM Java is used
- [SPARK-12511] - streaming driver with checkpointing unable to finalize leading to OOM
- [SPARK-12517] - No default RDD name for ones created by sc.textFile
- [SPARK-12526] - `ifelse`, `when`, `otherwise` unable to take Column as value
- [SPARK-12546] - Writing to partitioned parquet table can fail with OOM
- [SPARK-12558] - AnalysisException when multiple functions applied in GROUP BY clause
- [SPARK-12562] - DataFrame.write.format("text") requires the column name to be called value
- [SPARK-12579] - User-specified JDBC driver should always take precedence
- [SPARK-12582] - IndexShuffleBlockResolverSuite fails in windows
- [SPARK-12589] - result row size is wrong in UnsafeRowParquetRecordReader
- [SPARK-12591] - NullPointerException using checkpointed mapWithState with KryoSerializer
- [SPARK-12598] - Bug in setMinPartitions function of StreamFileInputFormat
- [SPARK-12611] - test_infer_schema_to_local depended on old handling of missing value in row
- [SPARK-12617] - socket descriptor leak killing streaming app
- [SPARK-12624] - When schema is specified, we should give better error message if actual row length doesn't match
- [SPARK-12629] - SparkR: DataFrame's saveAsTable method has issues with the signature and HiveContext
- [SPARK-12638] - Parameter explaination not very accurate for rdd function "aggregate"
- [SPARK-12647] - 1.6 branch test failure o.a.s.sql.execution.ExchangeCoordinatorSuite.determining the number of reducers: aggregate operator
- [SPARK-12654] - sc.wholeTextFiles with spark.hadoop.cloneConf=true fails on secure Hadoop
- [SPARK-12662] - Add a local sort operator to DataFrame used by randomSplit

- [SPARK-12673] - Prepending base URI of job description is missing
- [SPARK-12678] - MapPartitionsRDD should clear reference to prev RDD
- [SPARK-12682] - Hive will fail if the schema of a parquet table has a very wide schema
- [SPARK-12685] - word2vec trainWordsCount gets overflow
- [SPARK-12690] - NullPointerException in UnsafeInMemorySorter.free()
- [SPARK-12696] - Dataset serialization error
- [SPARK-12708] - Sorting task error in Stages Page when yarn mode
- [SPARK-12711] - ML StopWordsRemover does not protect itself from column name duplication
- [SPARK-12734] - Fix Netty exclusions and use Maven Enforcer to prevent bug from being reintroduced
- [SPARK-12739] - Details of batch in Streaming tab uses two Duration columns
- [SPARK-12746] - ArrayType(_, true) should also accept ArrayType(_, false)
- [SPARK-12747] - Postgres JDBC ArrayType(DoubleType) 'Unable to find server array type'
- [SPARK-12755] - Spark may attempt to rebuild application UI before finishing writing the event logs in possible race condition
- [SPARK-12760] - inaccurate description for difference between local vs cluster mode in closure handling
- [SPARK-12780] - Inconsistency returning value of ML python models' properties
- [SPARK-12783] - Dataset map serialization error
- [SPARK-12784] - Spark UI IndexOutOfBoundsException with dynamic allocation
- [SPARK-12805] - Outdated details in doc related to Mesos run modes
- [SPARK-12807] - Spark External Shuffle not working in Hadoop clusters with Jackson 2.2.3
- [SPARK-12841] - UnresolvedException with cast
- [SPARK-12859] - Names of input streams with receivers don't fit in Streaming page
- [SPARK-12874] - ML StringIndexer does not protect itself from column name duplication
- [SPARK-12921] - Use SparkHadoopUtil reflection to access TaskAttemptContext in SpecificParquetRecordReaderBase
- [SPARK-12961] - Work around memory leak in Snappy library
- [SPARK-12989] - Bad interaction between StarExpansion and ExtractWindowExpressions
- [SPARK-13047] - Pyspark Params.hasParam should not throw an error
- [SPARK-13056] - Map column would throw NPE if value is null
- [SPARK-13082] - sqlCtx.real.json() doesn't work with PythonRDD
- [SPARK-13087] - Grouping by a complex expression may lead to incorrect AttributeReferences in aggregations
- [SPARK-13088] - DAG viz does not work with latest version of chrome
- [SPARK-13101] - Dataset complex types mapping to DataFrame (element nullability) mismatch
- [SPARK-13121] - java mapWithState mishandles scala Option
- [SPARK-13122] - Race condition in MemoryStore.unrollSafely() causes memory leak
- [SPARK-13142] - Problem accessing Web UI /logPage/ on Microsoft Windows
- [SPARK-13153] - PySpark ML persistence failed when handle no default value

parameter
- [SPARK-13195] - PairDStreamFunctions.mapWithState fails in case timeout is set without updating State[S]
- [SPARK-13265] - Refactoring of basic ML import/export for other file system besides HDFS
- [SPARK-13298] - DAG visualization does not render correctly for jobs
- [SPARK-13300] - Spark examples page gives errors : Liquid error: pygments
- [SPARK-13312] - ML Model Selection via Train Validation Split example uses incorrect data
- [SPARK-13355] - Replace GraphImpl.fromExistingRDDs by Graph
- [SPARK-13371] - TaskSetManager.dequeueSpeculativeTask compares Option[String] and String directly.
- [SPARK-13390] - Java Spark createDataFrame with List parameter bug
- [SPARK-13410] - unionAll AnalysisException with DataFrames containing UDT columns.
- [SPARK-13441] - NullPointerException when either HADOOP_CONF_DIR or YARN_CONF_DIR is not readable
- [SPARK-13454] - Cannot drop table whose name starts with underscore
- [SPARK-13473] - Predicate can't be pushed through project with nondeterministic field
- [SPARK-13475] - HiveCompatibilitySuite should still run in PR builder even if a PR only changes sql/core
- [SPARK-13482] - `spark.storage.memoryMapThreshold` has two kind of the value.
- [SPARK-13697] - TransformFunctionSerializer.loads doesn't restore the function's module name if it's '__main__'

Documentation

- [SPARK-12351] - Add documentation of submitting Mesos jobs with cluster mode
- [SPARK-12507] - Expose closeFileAfterWrite and allowBatching configurations for Streaming
- [SPARK-12722] - Typo in Spark Pipeline example
- [SPARK-12758] - Add note to Spark SQL Migration section about SPARK-11724
- [SPARK-12814] - Add deploy instructions for Python in flume integration doc
- [SPARK-12894] - Add deploy instructions for Python in Kinesis integration doc
- [SPARK-13214] - Fix dynamic allocation docs
- [SPARK-13274] - Fix Aggregator Links on GroupedDataset Scala API
- [SPARK-13350] - Configuration documentation incorrectly states that PYSPARK_PYTHON's default is "python"
- [SPARK-13439] - Document that spark.mesos.uris is comma-separated

Improvement

- [SPARK-5273] - Improve documentation examples for LinearRegression
- [SPARK-11780] - Provide type aliases in org.apache.spark.sql.types for backwards

compatibility
- [SPARK-12120] - Improve exception message when failing to initialize HiveContext in PySpark
- [SPARK-12411] - Reconsider executor heartbeats rpc timeout
- [SPARK-12450] - Un-persist broadcasted variables in KMeans
- [SPARK-12701] - Logging FileAppender should use join to ensure thread is finished
- [SPARK-12834] - Use type conversion instead of Ser/De of Pickle to transform JavaArray and JavaList
- [SPARK-12932] - Bad error message with trying to create Dataset from RDD of Java objects that are not bean-compliant
- [SPARK-13094] - No encoder implicits for Seq[Primitive]
- [SPARK-13279] - Scheduler does O(N^2) operation when adding a new task set (making it prohibitively slow for scheduling 200K tasks)

New Feature

- [SPARK-10359] - Enumerate Spark's dependencies in a file and diff against it for new pull requests

Task

- [SPARK-13474] - Update packaging scripts to stage artifacts to home.apache.org