# Apache Spark 2.2.0正式发布

关于 Apache Spark 2.2.0 的详细新功能介绍请参见：《Apache Spark 2.2.0新特性详细介绍》

Apache Spark 2.2.0 持续了半年的开发，从RC1 到 RC6 终于在今天正式发布了。本版本是 2.x 版本线的第三个版本。在这个版本 Structured Streaming 的实验性标记（experimental tag）已经被移除，这也意味着后面的 2.2.x 之后就可以放心在线上使用了。除此之外，这个版本的主要集中点是系统的可用性（usability）、稳定性（stability）以及代码的润色（polish），并没有什么其他重大更新。此版本的一些新功能：

- 支持 LATERAL VIEW OUTER explode()，
- 支持给表添加列(ALTER TABLE table_name ADD COLUMNS)；
- 支持从Hive metastore 2.0/2.1中读取数据；
- 支持解析多行的JSON 或 CSV 文件；
- Structured Streaming为R语言提供的API；
- R语言支持完整的Catalog API；
- R语言支持 DataFrame checkpointing

此外，这个版本移除了对 Java 7 以及 Hadoop 2.5及其之前版本的支持。详细的更新如下：

## Core and Spark SQL

- API updates
    - SPARK-19107: Support creating hive table with DataFrameWriter and Catalog
    - SPARK-13721: Add support for LATERAL VIEW OUTER explode()
    - SPARK-18885: Unify CREATE TABLE syntax for data source and hive serde tables
    - SPARK-16475: Added Broadcast Hints BROADCAST, BROADCASTJOIN, and MAPJOIN, for SQL Queries
    - SPARK-18350: Support session local timezone
    - SPARK-19261: Support ALTER TABLE table_name ADD COLUMNS
    - SPARK-20420: Add events to the external catalog
    - SPARK-18127: Add hooks and extension points to Spark
    - SPARK-20576: Support generic hint function in Dataset/DataFrame
    - SPARK-17203: Data source options should always be case insensitive
    - SPARK-19139: AES-based authentication mechanism for Spark
- Performance and stability
    - Cost-Based Optimizer
        - SPARK-17075 SPARK-17076 SPARK-19020 SPARK-17077 SPARK-19350: Cardinality estimation for filter, join, aggregate, project and limit/sample operators
        - SPARK-17080: Cost-based join re-ordering
        - SPARK-17626: TPC-DS performance improvements using star-schema heuristics

    - SPARK-17949: Introduce a JVM object based aggregate operator
    - SPARK-18186: Partial aggregation support of HiveUDAFFunction
    - SPARK-18362 SPARK-19918: File listing/IO improvements for CSV and JSON
    - SPARK-18775: Limit the max number of records written per file
    - SPARK-18761: Uncancellable / unkillable tasks shouldn't starve jobs of resources
    - SPARK-15352: Topology aware block replication
- Other notable changes
    - SPARK-18352: Support for parsing multi-line JSON files
    - SPARK-19610: Support for parsing multi-line CSV files
    - SPARK-21079: Analyze Table Command on partitioned tables
    - SPARK-18703: Drop Staging Directories and Data Files after completion of Insertion/CTAS against Hive-serde Tables
    - SPARK-18209: More robust view canonicalization without full SQL expansion
    - SPARK-13446: [SPARK-18112] Support reading data from Hive metastore 2.0/2.1
    - SPARK-18191: Port RDD API to use commit protocol
    - SPARK-8425:Add blacklist mechanism for task scheduling
    - SPARK-19464: Remove support for Hadoop 2.5 and earlier
    - SPARK-19493: Remove Java 7 support

Programming guides: Spark Programming Guide and Spark SQL, DataFrames and Datasets Guide.

## Structured Streaming

- General Availablity
    - SPARK-20844: The Structured Streaming APIs are now GA and is no longer labeled experimental
- Kafka Improvements
    - SPARK-19719: Support for reading and writing data in streaming or batch to/from Apache Kafka
    - SPARK-19968: Cached producer for lower latency kafka to kafka streams.
- API updates
    - SPARK-19067: Support for complex stateful processing and timeouts using [flat]MapGroupsWithState
    - SPARK-19876: Support for one time triggers
- Other notable changes
    - SPARK-20979: Rate source for testing and benchmarks

Programming guide: Structured Streaming Programming Guide.

## MLlib

- New algorithms in DataFrame-based API
    - SPARK-14709: LinearSVC (Linear SVM Classifier) (Scala/Java/Python/R)

    - SPARK-19635: ChiSquare test in DataFrame-based API (Scala/Java/Python)
    - SPARK-19636: Correlation in DataFrame-based API (Scala/Java/Python)
    - SPARK-13568: Imputer feature transformer for imputing missing values (Scala/Java/Python)
    - SPARK-18929: Add Tweedie distribution for GLMs (Scala/Java/Python/R)
    - SPARK-14503: FPGrowth frequent pattern mining and AssociationRules (Scala/Java/Python/R)
- Existing algorithms added to Python and R APIs
    - SPARK-18239: Gradient Boosted Trees ®
    - SPARK-18821: Bisecting K-Means ®
    - SPARK-18080: Locality Sensitive Hashing (LSH) (Python)
    - SPARK-6227: Distributed PCA and SVD for PySpark (in RDD-based API)
- Major bug fixes
    - SPARK-19110: DistributedLDAModel.logPrior correctness fix
    - SPARK-17975: EMLDAOptimizer fails with ClassCastException (caused by GraphX checkpointing bug)
    - SPARK-18715: Fix wrong AIC calculation in Binomial GLM
    - SPARK-16473: BisectingKMeans failing during training with “java.util.NoSuchElementException: key not found” for certain inputs
    - SPARK-19348: pyspark.ml.Pipeline gets corrupted under multi-threaded use
    - SPARK-20047: Box-constrained Logistic Regression

Programming guide: Machine Learning Library (MLlib) Guide.

## SparkR

The main focus of SparkR in the 2.2.0 release was adding extensive support for existing Spark SQL features:

- Major features
    - SPARK-19654: Structured Streaming API for R
    - SPARK-20159: Support complete Catalog API in R
    - SPARK-19795: column functions to_json, from_json
    - SPARK-19399: Coalesce on DataFrame and coalesce on column
    - SPARK-20020: Support DataFrame checkpointing
    - SPARK-18285: Multi-column approxQuantile in R

Programming guide: SparkR (R on Spark).

## GraphX

- Bug fixes
    - SPARK-18847: PageRank gives incorrect results for graphs with sinks
    - SPARK-14804: Graph vertexRDD/EdgeRDD checkpoint results ClassCastException

- Optimizations
    - SPARK-18845: PageRank initial value improvement for faster convergence
    - SPARK-5484: Pregel should checkpoint periodically to avoid StackOverflowError

Programming guide: GraphX Programming Guide.

## Deprecations

- MLlib
    - SPARK-18613: spark.ml LDA classes should not expose spark.mllib in APIs. In spark.ml.LDAModel, deprecated oldLocalModel and getModel.
- SparkR
    - SPARK-20195: deprecate createExternalTable

## Changes of behavior

- MLlib
    - SPARK-19787: DeveloperApi ALS.train() uses default regParam value 0.1 instead of 1.0, in order to match regular ALS API's default regParam setting.
- SparkR
    - SPARK-19291: This added log-likelihood for SparkR Gaussian Mixture Models, but doing so introduced a SparkR model persistence incompatibility: Gaussian Mixture Models saved from SparkR 2.1 may not be loaded into SparkR 2.2. We plan to put in place backwards compatibility guarantees for SparkR in the future.