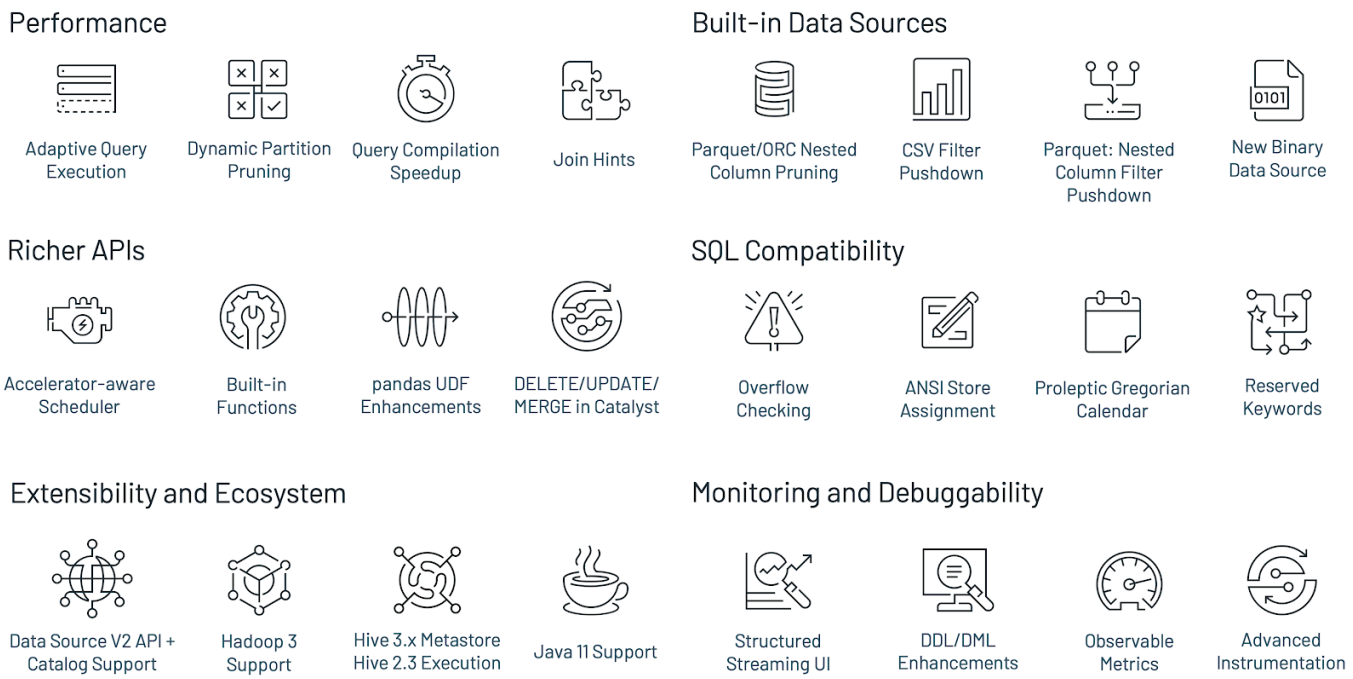
Apache Spark 3.0
第一个稳定版发布，终于可以在生产环境中使用啦！

Apache Spark 3.0.0 正式版是2020年6月18日发布的，其为我们带来大量新功能，很多功能加快了数据的计算速度。但是遗憾的是，这个版本并非稳定版。

不过就在昨天，Apache Spark 3.0.1
版本悄悄发布了（好像没看到邮件通知）！
值得大家高兴的是，这个版本是稳定版，官方推荐所有 3.0 的用户升级到这个版本。

Apache Spark 3.0 增加了很多令人兴奋的新特性，包括动态分区修剪（Dynamic Partition Pruning）、自适应查询执行（Adaptive Query Execution）、加速器感知调度（Accelerator-aware Scheduling）、支持 Catalog 的数据源API（Data Source API with Catalog Supports，参见 SPARK-31121）、SparkR 中的向量化（Vectorization in SparkR）、支持 Hadoop 3/JDK 11/Scala 2.12 等等。具体参见过往记忆大数据的 《历时近两年，Apache Spark 3.0.0 正式版终于发布了》 文章。



如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：iteblog_hadoop

Apache Spark 3.0.1 Release Note：https://spark.apache.org/releases/spark-release-3-0-1.html
所有修改的 ISSUE 参见：https://issues.apache.org/jira/secure/ReleaseNote.jspa?projectId=12315420&version=12347862
Apache Spark 3.0.1 下载地址：https://spark.apache.org/downloads.html

值得关注的修改

- [SPARK-26905]: Revisit reserved/non-reserved keywords based on the ANSI SQL standard
- [SPARK-31220]: repartition obeys spark.sql.adaptive.coalescePartitions.initialPartitionNum when spark.sql.adaptive.enabled
- [SPARK-31703]: Changes made by SPARK-26985 break reading parquet files correctly in BigEndian architectures (AIX + LinuxPPC64)
- [SPARK-31915]: Resolve the grouping column properly per the case sensitivity in grouped and cogrouped pandas UDFs
- [SPARK-31923]: Event log cannot be generated when some internal accumulators use unexpected types
- [SPARK-31935]: Hadoop file system config should be effective in data source options
- [SPARK-31968]: write.partitionBy() creates duplicate subdirectories when user provides duplicate columns
- [SPARK-31983]: Tables of structured streaming tab show wrong result for duration column
- [SPARK-32003]: Shuffle files for lost executor are not unregistered if fetch failure occurs after executor is lost
- [SPARK-32038]: Regression in handling NaN values in COUNT(DISTINCT)
- [SPARK-32073]: Drop R < 3.5 support
- [SPARK-32092]: CrossvalidatorModel does not save all submodels (it saves only 3)
- [SPARK-32136]: Spark producing incorrect groupBy results when key is a struct with nullable properties
- [SPARK-32148]: LEFT JOIN generating non-deterministic and unexpected result (regression in Spark 3.0)
- [SPARK-32220]: Cartesian Product Hint cause data error
- [SPARK-32310]: ML params default value parity
- [SPARK-32339]: Improve MLlib BLAS native acceleration docs
- [SPARK-32424]: Fix silent data change for timestamp parsing if overflow happens
- [SPARK-32451]: Support Apache Arrow 1.0.0 in SparkR
- [SPARK-32456]: Check the Distinct by assuming it as Aggregate for Structured Streaming
- [SPARK-32608]: Script Transform DELIMIT value should be formatted
- [SPARK-32646]: ORC predicate pushdown should work with case-insensitive analysis
- [SPARK-32676]: Fix double caching in KMeans/BiKMeans

已知的问题

- [SPARK-31511]: Make BytesToBytesMap iterator() thread-safe
- [SPARK-32779]: Spark/Hive3 interaction potentially causes deadlock
- [SPARK-32788]: non-partitioned table scan should not have partition filter
- [SPARK-32810]: CSV/JSON data sources should avoid globbing paths when inferring schema