Apache Hudi 0.6.0 版本发布，新功能介绍

本文英文原文：https://hudi.apache.org/releases.html

下载信息

- 源码：Apache Hudi 0.6.0 Source Release (asc, sha512)
- 二进制Jar包：nexus

如果想及时了
解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

2. 迁移指南

- 如果您从0.5.3以前的版本迁移至0.6.0，请仔细核对每个版本的迁移指南；
- 0.6.0版本从基于list的rollback策略变更为了基于marker文件的rollback策略，为进行平稳迁移，会在hoodie.properties文件中配置一个新属性hoodie.table.version；无论何时使用Hudi表新版本，如1（从0.6.0以前迁移到0.6.0），将会自动进行升级，并且只会对Hudi表升级一次，升级后hoodie.table.version属性将会自动更新。
- 类似也提供了一个降级命令行工具(-downgrade)，如用户想从0.6.0版本回退到之前的版本，此时hoodie.table.version将会从1变为0。
- 如果你在bulkInsert() RDD API中使用了自定义partitioner，注意0.6.0版本中该接口变为了BulkInsertPartitioner，需要对你的实现做适配。

3. 重点特性

3.1 写入端改进

- 对已有Parquet表进行迁移：支持通过Spark Datasource/DeltaStreamer引导已存在的Parquet表迁移至Hudi，同时可通过Hive，SparkSQL，AWS Athena进行查询（PrestoDB即将支持），技术细节请参考RFC-15。该特性暂时标记为experimental，在后续的0.6.x版本将持续进行完善。与传统重写方案相比资源消耗和耗时都有数据量的提升。
- bulk_insert支持原生写入：避免在bulk_insert写入路径中进行DataFrame - RDD转化，可显著提升bulk load的性能。后续的0.6.x版本将应用到其他的写操作以使得schema管理更为轻松，彻底避免spark-avro的转化。
- bulk_insert模式：Hudi bulk_insert对输入进行排序以便优化文件大小并避免在并发写入DFS多分区时的内存溢出问题，对于想在写入Hudi之前就已经准备好DataFrame的用户，Hudi也提供了hoodie.bulkinsert.sort.mode配置项。
- 支持Cleaning与写入并发执行，开启hoodie.clean.async=true以减少commit过程的耗时；
- Spark Streaming写入支持异步Compaction，可通过hoodie.datasource.compaction.async.enable进行配置。
- 支持通过marker文件进行Rollback，而不再对全表进行listing，设置hoodie.rollback.using.markers=true启用。
- 支持一种新的索引类型hoodie.index.type=SIMPLE，对于updates/deletes覆盖表大多数数据的场景，会比BLOOM_INDEX更快。
- 支持Azure Data Lake Storage V2， Alluxio 和 Tencent Cloud Object Storage
- HoodieMultiDeltaStreamer 支持在单个DeltaStreamer中消费多个Kafka流，降低使用DeltaStreamer作为数据湖摄取工具时的运维负担。
- 新增新的工具类InitialCheckPointProvider，以便在迁移至DeltaStreamer后设置Checkpoint。
- DeltaStreamer工具支持摄取CSV数据源，同时可chain多个transformers来构建更灵活的ETL作业。
- 引入新的Key生成器CustomKeyGenerator，对不同类型的Key、Partition路径提供更灵活

docs

3.2 查询端改进

- 从0.6.0版本开始，Spark DataSource支持MoR表的SNAPSHOT查询；
- 在之前版本中，对CoW表，Hudi仅仅支持HoodieCombineHiveInputFormat来确保对于任何查询都只会生成有限数量的mappers。Hudi现在对MoR表支持使用HoodieCombineInputFormat。
- 在HoodieROPathFilter中缓存MetaClient来加速Spark查询，这可以减少在S3上对Read-Optimized查询进行文件过滤的额外开销。

3.3 易用性提升

- 对Spark DAG赋名字以便更好的进行调试。

- 支持用户自定义可插拔指标报告者，另外内置Console，JMX，Prometheus，DataDog指标报告者。
- 新增Data Snapshot Exporter工具类，通过该工具类可将某一时刻的Hudi表导出为Parquet文件。
- 引入写入提交回调钩子，以便在Commit时可以通知增量pipelines，例如在新的commit到来后触发Apache Airflow作业。
- 支持通过CLI删除Savepoints。

- 新增命令 export instants来导出instant元数据。

4. 贡献者

感谢以下贡献者，排名不分先后

hddong, xushiyan, wangxianghu, shenh062326, prashantwason, bvaradar, vinothchandar, baobaoyeye, andreitaleanu, clocklear , linshan-ma, satishkotha, Trevor-zhang, pratyakshsharma, GuoPhilipse, nsivabalan, zhedoubushishi, umehrot2, lw309637554, DeyinZhong, zherenyu831, lamber-ken, garyli1019, bhasudha, n3nash, yihua, liujinhui1994, sreeram26, Yungthuis, cheshta2904, leesf

本文原文：Apache Hudi 0.6.0版本重磅发布