Apache Hive 0.13发布，新增ACID特性

4月16日在http://mirror.bit.edu.cn/apache/hive/hive-0.13.0/网址就可以下载Hive 0.13，这个版本在Hive执行速度、扩展性、SQL以及其他方面做了相当多的修改：

一、执行速度

用户可以选择基于Tez的查询，基于Tez的查询可以大大提高Hive的查询速度（官网上上可以提升100倍）。下面一些技术对查询速度的提升：

（1）、Broadcast Joins：和MapJoin类似，但是不需要在client端建立一个hashtable；
（2）、Dynamic Partitioned Hash Joins：基于大表bucketing的特性去分发小表；
（3）、Cardinality estimation-based decision： 主要是在Join算法和并行度上面；
（4）、Pre-launch of containers

目前Hive提供了矢量查询执行器模型，可以使得CPU的计算提高5-10倍，查询翻译速度提高2-3倍。

二、SQL方面

Hive 0.13支持与SQL标准为基础的授权功能，用户现在可以在SQL兼容的方式定义他们的授权策略。在实体上，扩展了SQL语言以便支持grant和revoke。而且Hive现在支持show roles、user privileges以及active privileges。支持可插拔的授权API。其他新特性：

（1）、支持DECIMAL和CHAR数据类型；
（2）、Unqualified的join条件；
（3）、基于标准的带引号标识符行为；
（4）、Common table expressions
（5）、IN, NOT IN, EXISTS and NOT EXISTS支持子查询；
（6）、Permanent functions
（7）、在WHERE后面支持JOIN条件。

当然在Hive 0.13中，事务的原子性、一致性和持久性在分区层得到保证，隔离性则通过开启ZooKeeper或内存中可用的锁机制来保证。通过在Hive 0.13中加入事务，实现在行级提供全部的ACID语义，这样的话，一个应用程序可以添加行，而另一个应用程序可以从同一分区中读取数据，互相之间不会产生干扰。配置事务Hive中引入了十个新的配置：hive.txn.manager、hive.txn.timeout、hive.txn.max.open.batch、hive.compactor.initiator.on、hive.compactor.worker.threads、hive.compactor.worker.timeout、hive.compactor.check.interval、hive.compactor.delta.num.threshold、hive.compactor.delta.pct.threshold和hive.compactor.abortedtxn.threshold，关于这些属性的默认值和含义可以去阅读官方文档。

三、其他方面的提升

主要是在HiveServer2, HCatalog and JDBC access等

（1）、Hive Server 2支持HTTP方式，SSL support for both binary and HTTP (HTTPS)、在HTTP（S）支持Kerberos 授权，Support for HTTP(S) through a trusted proxy。
（2）、HCatalog ：HCatalog parity for all Hive data types，Reconciliation of HCatalog and Hive “INSERT INTO” semantics。

（3）、JDBC：Support for JDBC job cancel，Async execution。
一张详细的表可以参照：

Notable Improvements in Apache Hive 0.13

Speed

- Interactive query through Hive on Tez
- Vectorized query execution engine
- Cost-based optimizer
- Split elimination for ORCFile
- Partition pruning for string and date datatypes
- Faster query planning

Scale

- More scalable dynamic partition loads
- Smaller hash tables, allowing more scalable MapJoins

SQL

- Subquery for IN / NOT IN
- Support for EXISTS and NOT EXISTS
- Common table expressions (CTEs)
- Support for CHAR datatype
- Scale and precision for DECIMAL datatype
- Authorization based on SQL standards
- JOIN conditions in the WHERE clause
- Support for non-ASCII column names
- Permanent UDFs
- Stream data into Hive from Flume (Experimental feature)

Additional Features

- HiveServer 2 improvements
  - PAM authentication support
  - SSL encryption
  - HTTP/HTTPS support
  - Cancel MR/Tez jobs via JDBC/ODBC
- HCatalog parity for all of Hive datatypes

hive 0.13