Spark SQL执行流程分析

可以使用Spark SQL，进行类SQL语法分析，使用命令如下：

*$ bin/spark-sql --master yarn*

*INFO HiveMetaStore: 0: Opening raw store with implemenation class:org.apache.hadoop.hive.metastore.ObjectStore*

*HiveClientImpl: Warehouse location for Hive client (version 1.2.1) is file:/usr/lib/spark/spark-warehouse*

*spark-sql> show databases;*

*INFO SparkSqlParser: Parsing command: show databases*

*INFO CodeGenerator: Code generated in 225.075773 ms*

*bigbench*

*default*

*Time taken: 1.659 seconds, Fetched 6 row(s)*

通过spark sql启动脚本，源码如下：

*exec "${SPARK\_HOME}"/bin/spark-submit --class org.apache.spark.sql.hive.thriftserver.SparkSQLCLIDriver "$@"*

脚本调用顺序为：*spark-sql —> spark-submit —>spark-class* 然后调用SparkSumit将任务提交到YARN中，Driver程序为SparkSQLCLIDriver：

*/usr/jdk64/jdk1.8.0\_77/bin/java -cp /usr/lib/spark/conf/:/usr/lib/spark/jars/\*:/usr/lib/hadoop/etc/hadoop/ -Xmx1g org.apache.spark.deploy.SparkSubmit --class org.apache.spark.sql.hive.thriftserver.SparkSQLCLIDriver spark-internal*

命令的执行流程如下：



1. 启动Spark SQL脚本后，会在YARN中以client模式启动Spark SQL Job，包括：Driver、ExecutorLauncher和Executor
2. 同时Driver即SparkSubmit进程，从控制台接收到用户输入的指令后，传给SQLDriver中执行
3. 在SparkSQLDriver中调用SQLContext，创建并提交查询任务，具体的查询解析及执行不再详细介绍

开启Kerberos后的认证过程: