一:Spark SQL下的Parquet使用最佳实践

1. 过去整个业界对大数据的分析的技术栈的Pipeline一般分为以下两种方式：
   1. Data Source -> HDFS ->MR/Hive/Spark(相当于ETL)->HDFS Parquet->Spark SQL/Impala->Result Service(可以放在DB中，也有可能被通过JDBC/ODBC来作为数据服务使用)；
   2. Data Source->Real time updata data to Hbase/DB->export to Parquet ->Spark SQL/Impala-> Result Service(可以放在DB中，也有可能被通过JDBC/ODBC来作为数据服务使用)；

上述的第二种方式完全可以通过Kafka+Spark Streaming+Spark SQL的方式取代 （内部也强烈建议采用Parquet的方式来存储数据）的方式取代。

2，期待的方式：

Data Source -> Kafka->Spark Streaming ->Parquet->Spark SQL(ML/Graphx等等)->Parquet->其它各种Data Mining等

二：Parquet的精要介绍

1. Parquet是列式存储格式的一种文件类型，列式存储有一下的核心优势：
   1. 可以跳过不符合条件的数据，只读取需要的数据，降低IO数据量；
   2. 压缩编码可以降低磁盘存储空间。由于同一列的数据类型是一样的，可以使用更高效的压缩编码（例如Run Length Encoding和Delta Encoding）进一步节约存储空间；
   3. 只读取需要的列，支出向量运算，能够获取更好的扫描性能；
2. 