BIG DATA
大数据时代

Apache Kylin初识

讲师:Jepson

www.dashujuedu.com

http://blog.itpub.net/30089851/abstract/1/

BIG DATA
大数据时代

一.Kylin是什么?主要特性

二.Kylin的Cube简单介绍

三.Kylin的架构图

四.Kylin的 安装(单节点)和测试案例

五.踩坑心得

BIG DATA
大数据时代

一.Kylin是什么?主要特性

- Apache Kylin™是一个开源的分布式分析引擎，提供Hadoop之上的SQL查询接口及多维分析（联机分析处理OLAP）能力以支持超大规模数据，最初由eBay Inc. 开发并贡献至开源社区。

- 当前流行的SQL-on-Hadoop方案需要扫描部分或者全部数据来完成查询， 查询延迟很大， 而kylin在SQL-on-Hadoop基础之上， 通过预计算cube方式， 以空间换时间的方案来大幅降低查询延时，从而弥补了现有方案的不足之处。

BIG DATA
大数据时代

Kylin主要特性

- **可扩展超快OLAP引擎:**
  Kylin是为减少在Hadoop上百亿规模数据查询延迟而设计

- **Hadoop ANSI SQL 接口:**
  Kylin为Hadoop提供标准SQL支持大部分查询功能

- **交互式查询能力:**
  通过Kylin，用户可以与Hadoop数据进行亚秒级交互，在同样的数据集上提供比Hive更好的性能

- **多维立方体（MOLAP Cube）:**
  用户能够在Kylin里为百亿以上数据集定义数据模型并构建立方体

- **与BI工具无缝整合:**
  Kylin提供与BI工具，如Tableau，的整合能力，即将提供对其他工具的整合

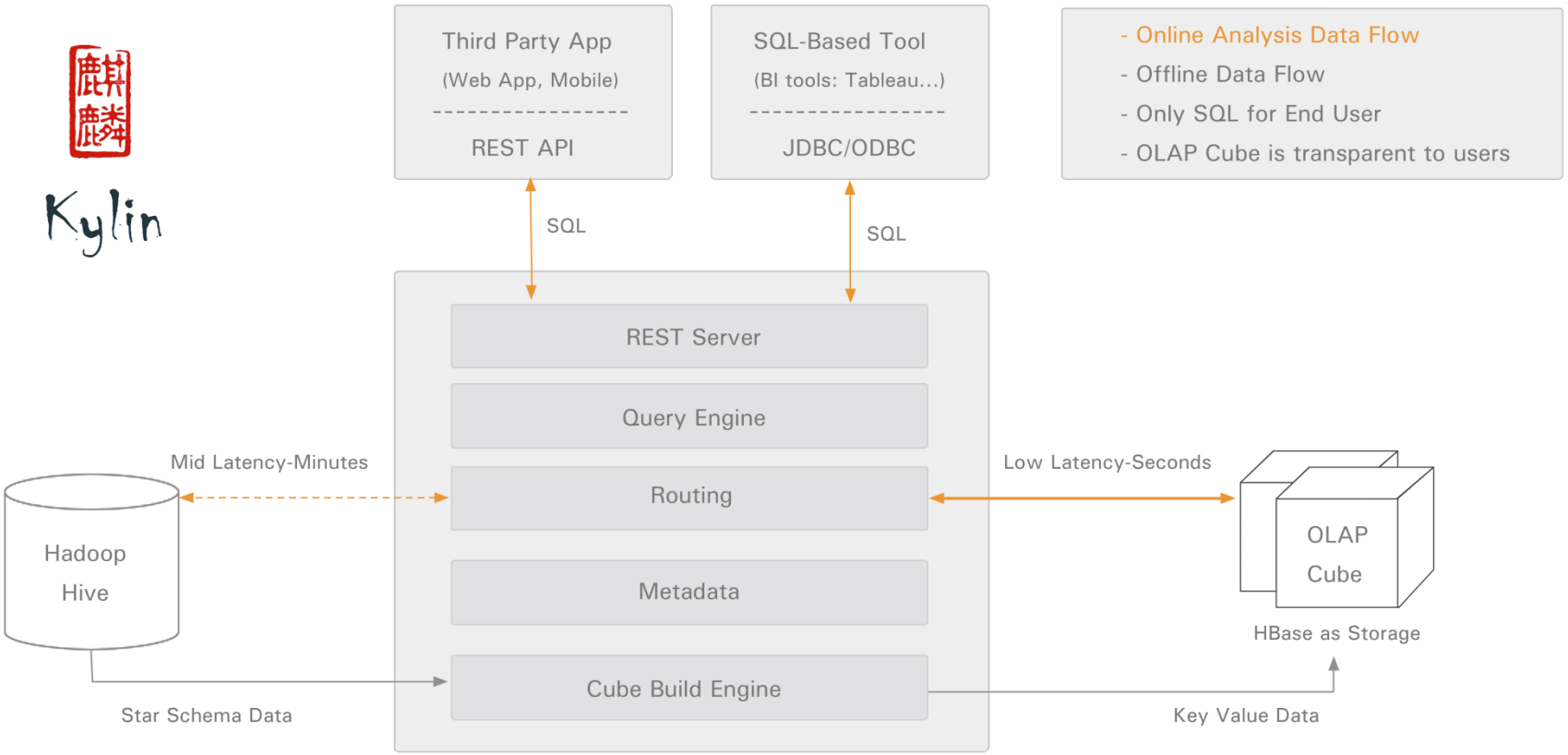
- **其他特性:**
  Job管理与监控 、压缩与编码、 增量更新 、 利用HBase Coprocessor
  基于HyperLogLog的Dinstinc Count近似算法 、友好的web界面以管理，监控和使用立方体
  项目及立方体级别的访问控制安全、支持LDAP

二.Kylin的Cube简单介绍

- Kylin使用Hadoop结合数据立方体（**Cube**）技术实现多维度快速OLAP分析能力

- OLAPCube是一种典型的多维数据分析技术，**Cube本身可以认为是不同维度数据组成的dataset**，一个OLAP Cube 可以拥有多个维度（Dimension），以及多个事实（Factor Measure）。用户通过OLAP工具从多个角度来进行数据的多维分析。通常认为OLAP包括三种基本的分析操作：上卷（rollup）、下钻（drilldown）、切片切块（slicingand dicing），原始数据经过聚合以及整理后变成一个或多个维度的视图。

三.Kylin的架构图



- 1.首先要求用户把数据放在Hadoop上，通过Hive管理，用户在Kylin中进行数据建模以后，Kylin会生成一系列的MapReduce任务来计算Cube，算好的Cube最后以K-V的方式存储在HBase中。
- 2.分析工具发送标准SQL查询,Kylin将它转换成对HBase的Scan，快速查到结果返回给请求方。

四.Kylin的 安装(单节点)和测试案例

- **1.准备Hadoop2.7.2+HBase1.1.5+Hive2.0.0集群环境及启动相关服务**

- **2. 添加hive_dependency和KYLIN_HOME环境变量**

hive_dependency=$HIVE_HOME/conf:/hadoop/hive/lib/*:$HIVE_HOME/hcatalog/share/hcatalog/hive-hcatalog-core-2.0.0.jar
export $hive_dependency

export KYLIN_HOME=/hadoop/kylin

PATH=.:$HADOOP_HOME/bin:$JAVA_HOME/bin:$ZOOKEEPER_HOME/bin:$HBASE_HOME/bin:$HIVE_HOME/bin:$KYLIN_HOME/bin:$PATH
export $PATH

- **3.下载安装**

wget https://dist.apache.org/repos/dist/release/kylin/apache-kylin-1.5.2.1/apache-kylin-1.5.2.1-HBase1.x-bin.tar.gz
tar -zxvf apache-kylin-1.5.1-HBase1.1.3-bin.tar.gz
ln -s /hadoop/apache-kylin-1.5.2.1-bin /hadoop/kylin

➢ 4.修改$KYLIN_HOME/bin/kylin.sh

export KYLIN_HOME=/hadoop/kylin # 改成绝对路径

export HBASE_CLASSPATH_PREFIX=${tomcat_root}/bin/bootstrap.jar:${tomcat_root}/bin/tomcat-juli.jar:${tomcat_root}/lib/*:$hive_dependency:$HBASE_CLASSPATH_PREFIX #在路径中添加$hive_dependency

➢ 5.修改$KYLIN_HOME/conf/kylin.properties

# List of web servers in use, this enables one web server instance to sync up with other servers.
kylin.rest.servers=localhost:7070
######新增
kylin.job.jar=$KYLIN_HOME/lib/kylin-job-1.5.2.1.jar
kylin.coprocessor.local.jar=$KYLIN_HOME /lib/kylin-coprocessor-1.5.2.1.jar

➢ 6.启动kylin

$KYLIN_HOME/bin/kylin.sh start

http://<ip>:7070/kylin 账号: **ADMIN**　密码:**KYLIN**

➢ 7.导入官网测试案例

$KYLIN_HOME/bin/kylin.sh stop

$KYLIN_HOME/bin/sample.sh

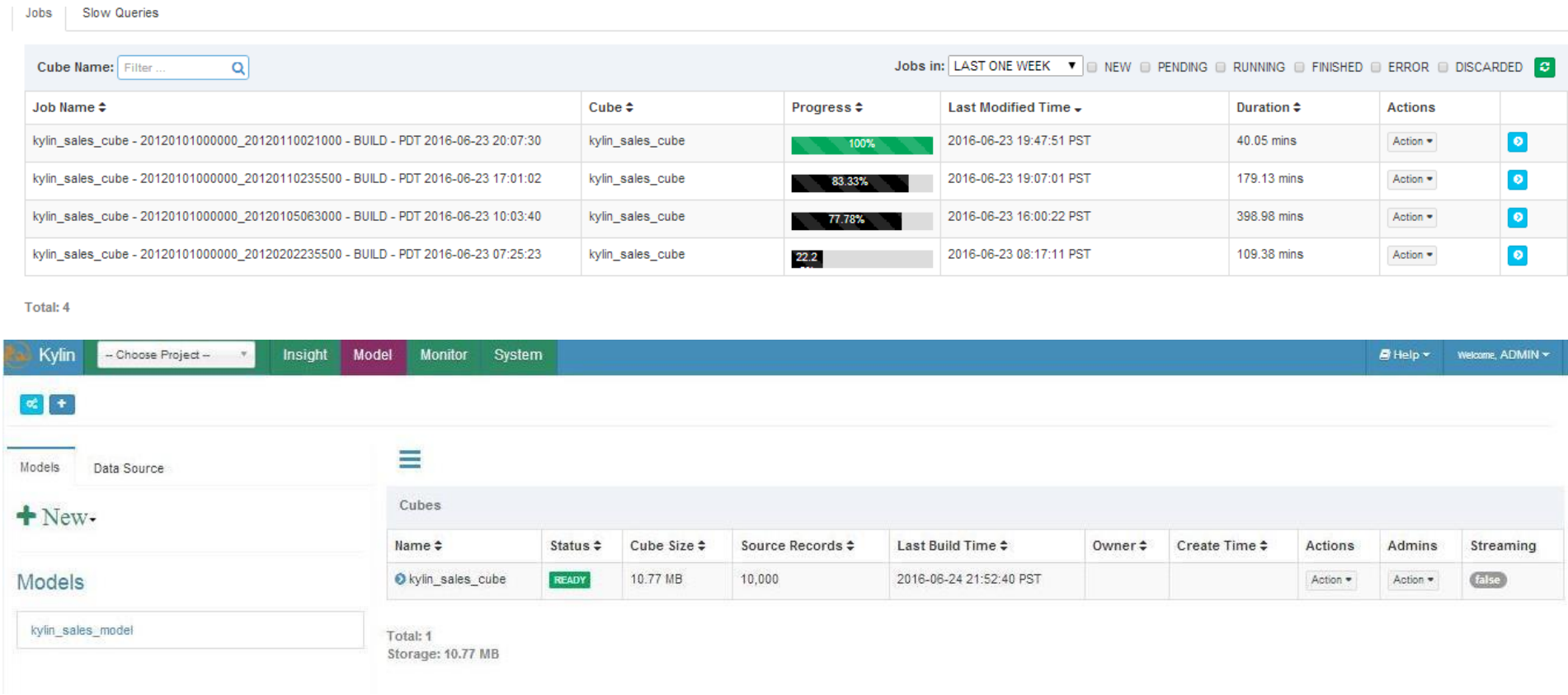
$KYLIN_HOME/bin/kylin.sh start

➢ 8.Build Cube

❖ 1>.选中'kylin_sales_cube'示例立方体，点击'Actions'->'Build'，选择一个截止日期，本试验中选择的是'2012-01-10'

(具体小时,分,秒随便选,因为最终hive sql的语句类似　WHERE (KYLIN_SALES.PART_DT >= '2012-01-01' AND KYLIN_SALES.PART_DT < '2012-01-10');)

❖ 2>.在'Monitor'标签中通过刷新页面检查进度条，直到100%



❖ 3>.在'Insight'标签中执行下面的SQL查询：

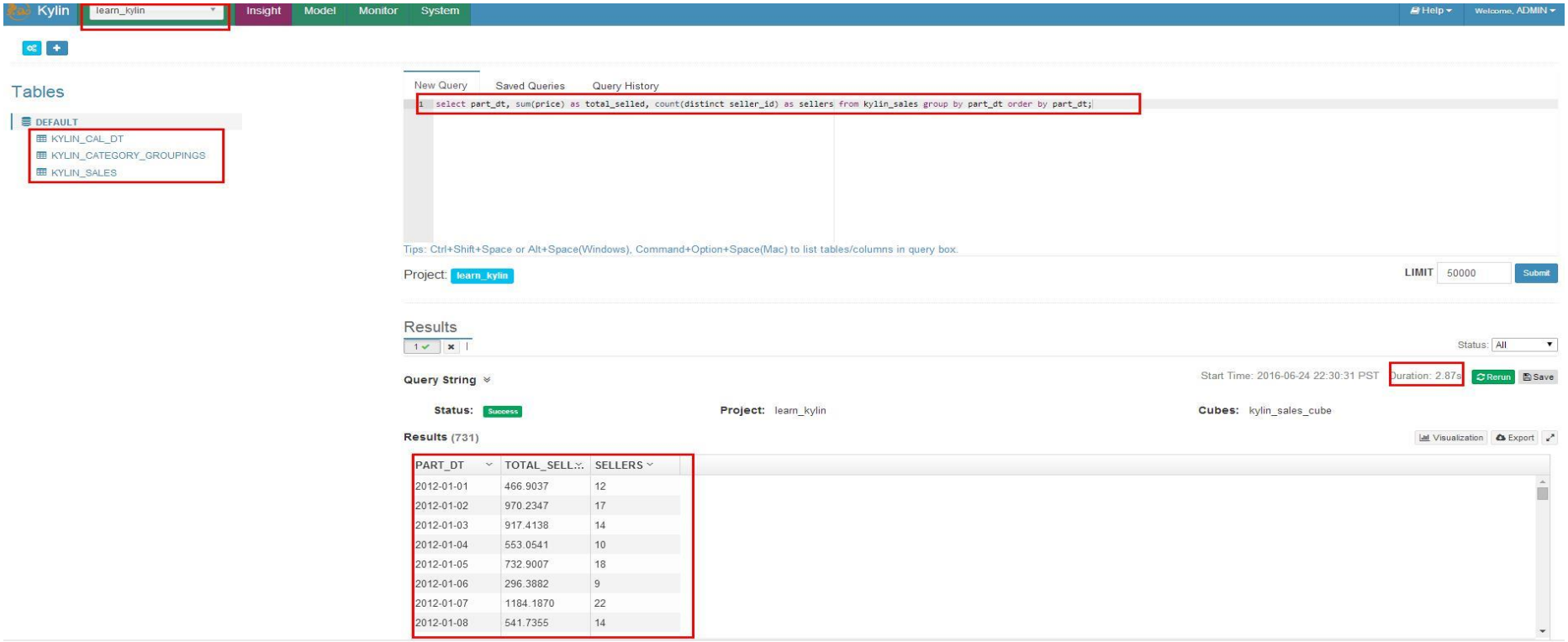
select part_dt, sum(price) as total_selled, count(distinct seller_id) as sellers from kylin_sales group by part_dt order by part_dt;

###耗时2.87s

❖ 4>.在hive中执行同一个SQL查询，验证kylin的查询结果(会开启MapReduce Job计算)

###耗时65.205s

查询执行和结果如图所示



❖ 5>.Build成功后，hive中建立了3+n个表，如图所示**(3个官网案例hive表,n个build的hive表)**

```
hive> show tables;
OK
kylin_cal_dt
kylin_category_groupings
kylin_intermediate_kylin_sales_cube_desc_20120101000000_20120105063000
kylin_intermediate_kylin_sales_cube_desc_20120101000000_20120110235500
kylin_intermediate_kylin_sales_cube_desc_20120101000000_20120202235500
kylin_sales
Time taken: 0.323 seconds, Fetched: 6 row(s)
hive>
```

❖ 6>.Build成功后，hbase中建立了1+n个表，如图所示**(1个元数据表,n个build的hbase表)**

```
hbase(main):029:0> list
TABLE
KYLIN_7932J1LDOI
KYLIN_JE63ERR0LR
KYLIN_KPVQ8J6VVK
KYLIN_O2EZ9WSOLO
KYLIN_TQVDJK5AP1
KYLIN_W1JS1OXZH4
kylin_metadata
tsnappy
8 row(s) in 0.0470 seconds

=> ["KYLIN_7932J1LDOI", "KYLIN_JE63ERR0LR", "KYLIN_KPVQ8J6VVK", "KYLIN_O2EZ9WSOLO", "KYLIN_TQVDJK5AP1", "KYLIN_W1JS1OXZH4", "kylin_metadata", "tsnappy"]
```

BIG DATA
大数据时代

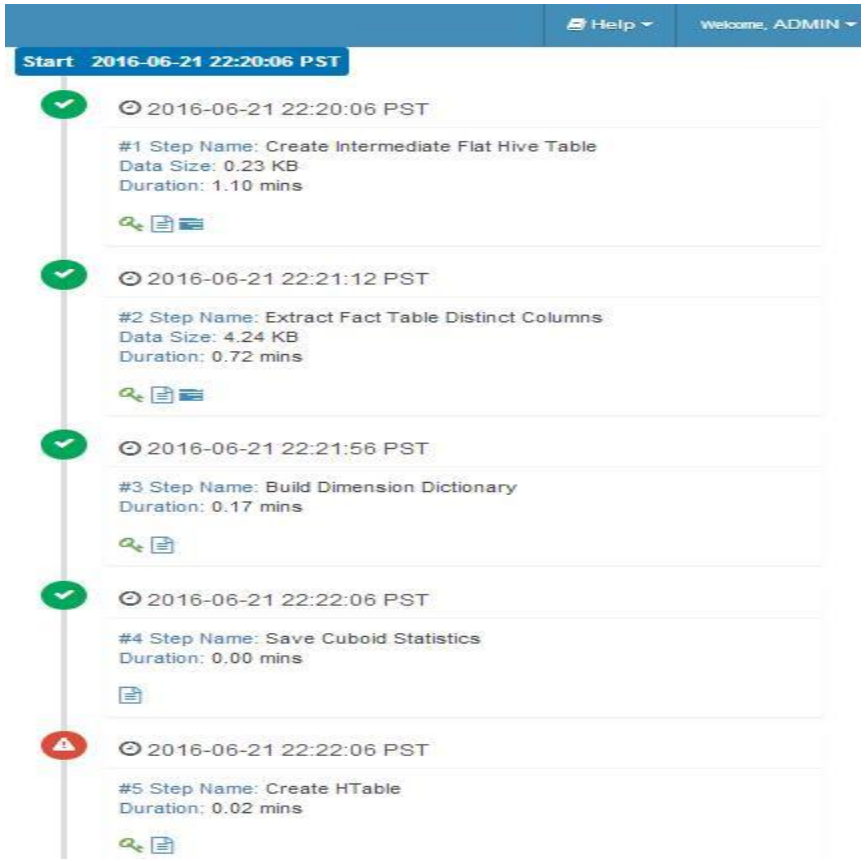
五.踩坑心得

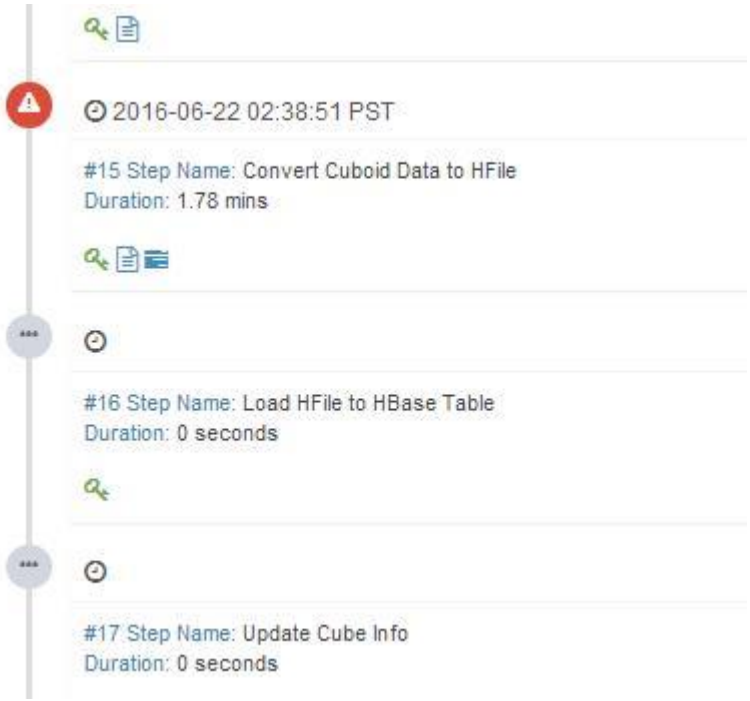
- **1.各个组件版本要兼容 --Step15**

  Hadoop2.7.2、HBase1.1.5、Hive2.0.0

  Kylin1.5.2.1(apache-kylin-1.5.1-HBase1.1.3-bin.tar.gz)

- **2.Hadoop和HBase要支持snappy库 --Step5**





BIG DATA
大数据时代

THANK YOU!