Sparak、Impala、Hive 对比测试

目录

一、测试概要

通过两组 SQL 对比测试，大致了解 Spark SQL 的性能水平，并横向与 Hive 及 Impala 做下对比，同时对 Sql 兼容性做下总结。

二、测试环境

1、硬件环境

名称	节点配置	数量	安装服务	备注
UDH 集群	CPU 8*CORE Memory 16G Disk 300G	4	HDFS YARN HIVE ZOOKEEPER IMPALA SPARK	机器是 UAP 云平台虚拟机

2、软件环境

UDH1.0.4

三、测试方法

1、Spark 测试方法

数据格式采用文本。

Spark 读取 HDFS 上的数据文件（文本），创建数据集（DataFrame），执行查询 SQL。

通过编写 Spark Java 程序进行测试，并对三种提交作业方式分(Stand-alone,yarn-clinet,yarn-cluster）别进行测试。

2、Impala 测试方法

数据格式采用 Parquet。

通过 Impala shell 直接提交查询 SQL。

分别对 Impala 两种运行方式进行测试（stand-alone,yarn)。

3、Hive 测试方法

数据格式采用文本。

通过 hive shell 直接提交查询 SQL。

4、测试 SQL

查询一：

SELECT ECRID,sum(AMTSUM) as total FROM COMMNAME_TMP_2014 group by ECRID

查询二：

select a.ECRID,b.storename,sum(a.AMTSUM) total from COMMNAME_TMP_2014 a left join bi_dim_storeinfo b on a.ECRID = b.pk_store group by a.ECRID,b.storename

5、数据量

分别对两组数据量进行测试：1000 万行、1 亿行，数据来源一真实项目数据。

	表名	列信息	数据量
事实表	commname_tmp_2014	hkey string,amtsum int,commname string,dsale string,saleno string,ecrid string,bcode string,qty int,foodtime int,payname int,halfhour int,pzbcode string,plbcode string	1000万行、一亿行
维表	bi_dim_storeinfo	pk_store string,storecode string,storename string,corpname string,pk_busmode string,pk_brand string,pk_zone string,pk_storetype string,pk_city string,storearea float,storetable float,busmodecode string,brandcode string,zonecode string,storetypecode string,citycode string,addr string	595

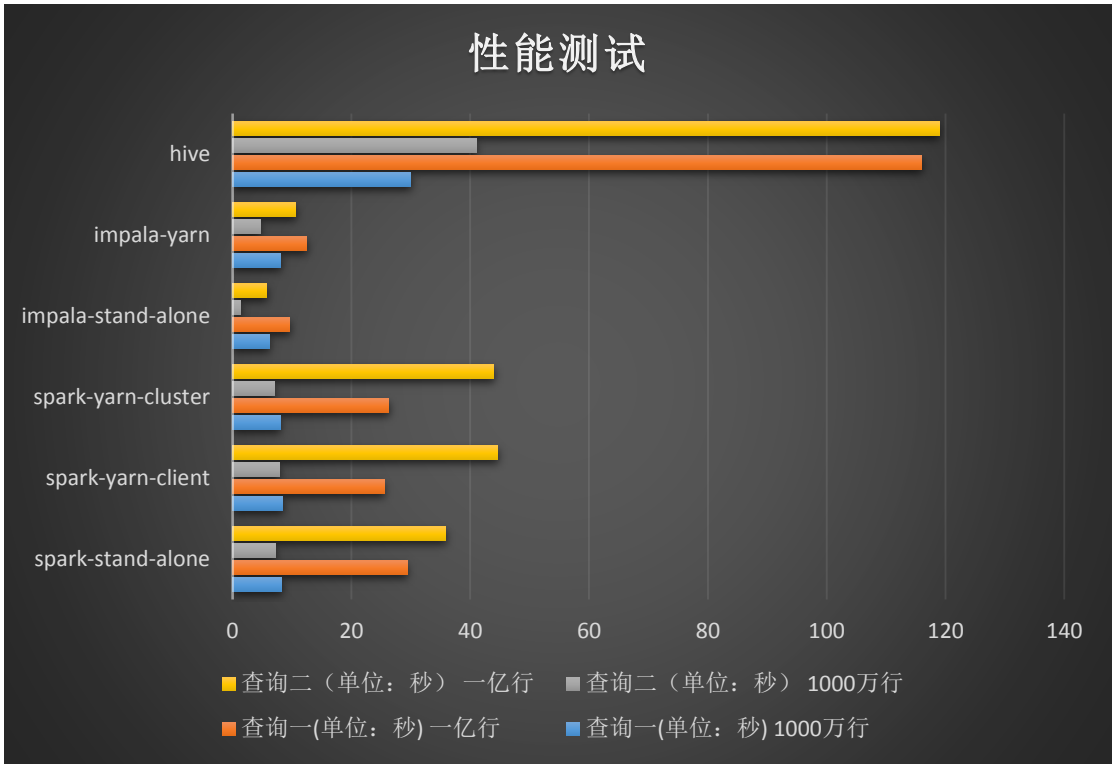
四、测试结果

测试数据：

调度方式	查询一(单位：秒)		查询二（单位：秒）	
	1000 万行	一亿行	1000 万行	一亿行
spark-stand-alone	8.3	29.4	7.2	35.9
spark-yarn-client	8.4	25.5	7.9	44.6
spark-yarn-cluster	8	26.3	7	44
impala-stand-alone	6.2	9.6	1.4	5.7

impala-yarn	8.1	12.5	4.7	10.6
hive	30	116	41	119

统计图：



五、SQL 兼容性

Spark SQL 几乎完全兼容 HIVE SQL 语法，只是 HIVE 特有的一些优化参数及极少用语法不支持。

Impala SQL 与 HIVE SQL 高度兼容，但不局限于 HIVE 已有的查询 SQL，同时 Impala 还支持 insert into。

六、小结

查询性能上 HIVE<SPARK<IMPALA

SQL 支持上 SPARK<HIVE<IMPALA