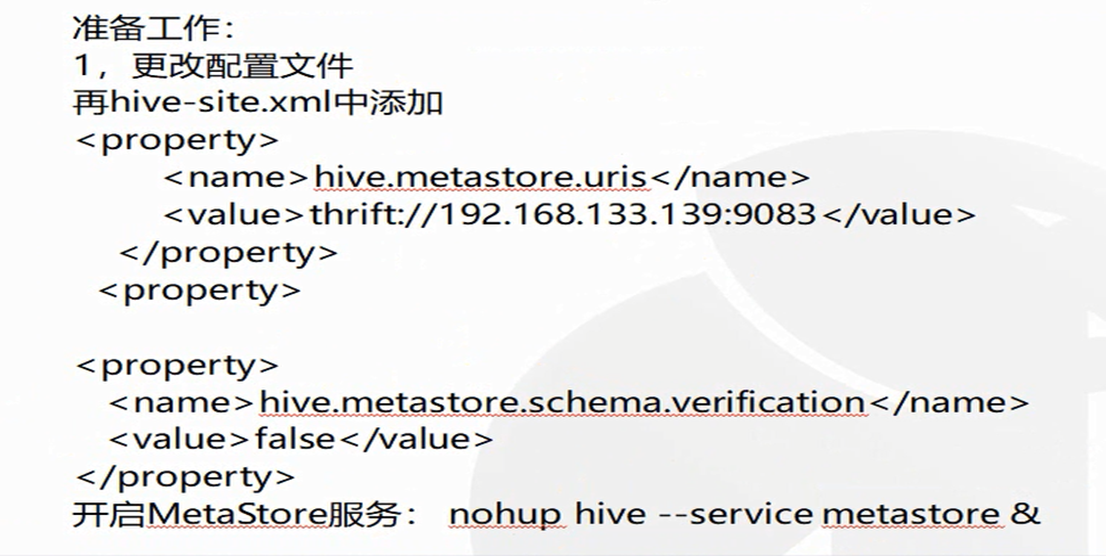
前期准备：

如果是原生集群则需要配置



**Spark On Hive，通过spark sql模块访问和使用Hive**

1、安装hvie版本1.2.1

2、将hive与mysql集成

3、将hdfs-site.xml、core-site.xml、hive-site.xml文件复制到spark/conf目录下

4、执行spark-sql命令

**通过代码操作spark on hive**

1、加入依赖：spark-hive\_2.10

2、将hive-site.xml、hdfs-site.xml、core-site.xml文件加入resource目录

3、编写代码案例

sqlContext.sql("CREATE TABLE IF NOT EXISTS src (key INT, value STRING)")

sqlContext.sql("LOAD DATA LOCAL INPATH kv1.txt' INTO TABLE src")

sqlContext.sql("FROM src SELECT key, value").collect().foreach(*println*)

1. **先按官网的参考代码，构建**

SPARK ON HIVE：让spark sql通过sql的方式去读取hive当中的数据

HIVE ON SPARK：让hive的计算引擎由MapReduce改为SPARK

|  |
| --- |
| val conf = new SparkConf().setMaster("local[\*]").setAppName("hotCount")  val sc = new SparkContext(conf)  //创建hive的实例  val hiveContext = new HiveContext(sc)  hiveContext.sql("CREATE TABLE IF NOT EXISTS src (key INT, value STRING)")  hiveContext.sql("LOAD DATA LOCAL INPATH 'E:/hive.txt' INTO TABLE src")  // Queries are expressed in HiveQL  hiveContext.sql("FROM src SELECT key, value").collect().foreach(println) |

出现：

|  |
| --- |
| Exception in thread "main" java.lang.UnsatisfiedLinkError: org.apache.hadoop.io.nativeio.NativeIO$Windows.createDirectoryWithMode0(Ljava/lang/String;I)V |

我们之后在windows当中配置的hadoop版本是从apache下载，但是我们现在所使用的hadoop版本是cdh的。所以出现的版本不一致的问题。

1. **需要去下载一个cdh版本的hadoop**

进入网站：<http://archive.cloudera.com/cdh5/cdh/5/>



配置HADOOP\_HOME的环境变量

将插件中的

中的hadoop.dll文件 复制到C:/windows/system32目录下

在插件中的其它文件，复制到hadoop目录的/bin目录下。

注意：有可能是插件问题



现在可以保证，HiveContext可以操作相应的SQL。

但是该SQL操作的内容是来自于windows本地，没有与hive表进行连接。

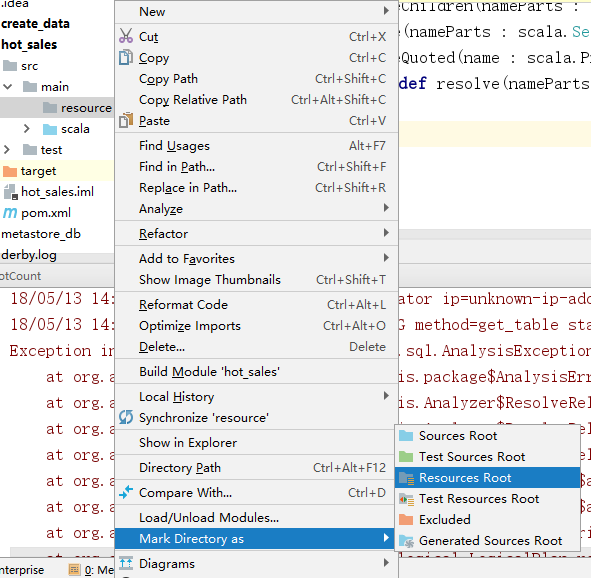
1. **让HiveContext与hive进行连接**

将hive-site.xml、hdfs-site.xml、core-site.xml文件加入resource目录

需要在cdh的版本下，去找到hive的配置文件

2.1：find / -name ‘hive-site.xml’发现有多个配置文件，建议使用/etc/目录下的文件

2.2：在Idea项目去创建一个resource目录



2.3将配置文件复制到resource目录下

注意：在sparksql去连接是通过主机名去访问的，要让本地的hosts的ip与主机名对应。

连接成功了。但是看不到相应的数据

在hive当中去创建了一个很简单的表，通过spark sql去测试，发现是可以正常读取出来的。

查询不到结果与表结构有关。

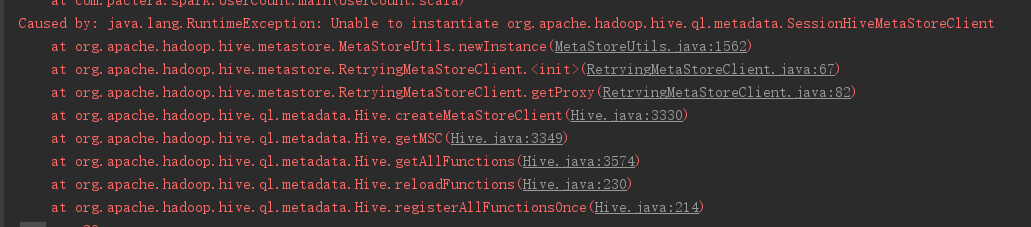
flume到hive的表需要要求：分桶与orc格式。

在spark 1.6以及之前不支持。

在spark 2.\*当中是可以通过spark sql来操作的。

将原始表中的数据，转换成我们计算的单元【根据需求来确定】

**2.4 提示一下错误**



问题原因：没有启动hive和metastore

解决办法：启动hive的metastore和hiveserve2

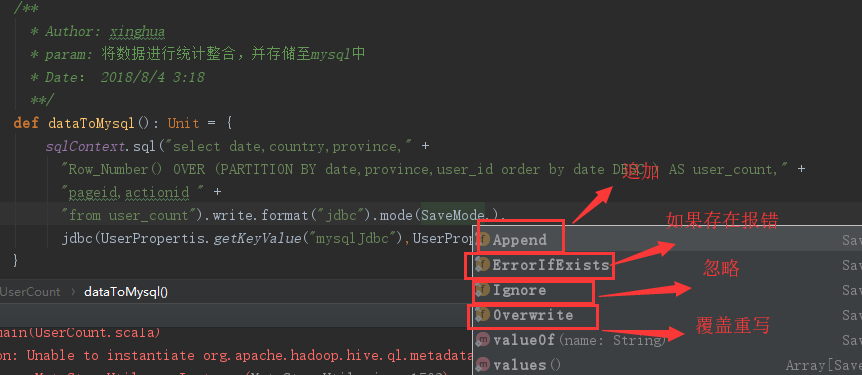
nohup hive --service metastore &

nohup hive --service hiveserver2 &

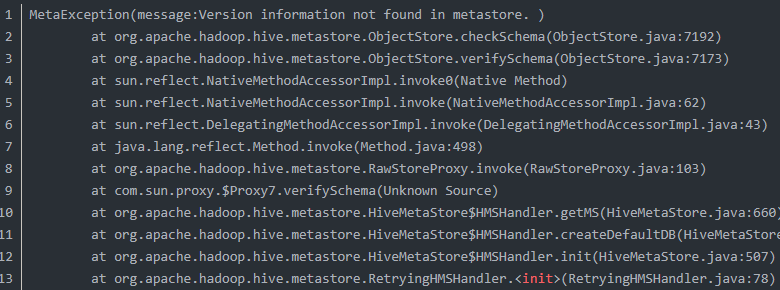
**2.5 spark通过jdbc将数据结果存储在mysql中**

|  |
| --- |
| sqlContext.sql("select date,country,province," +  "Row\_Number() OVER (PARTITION BY date,province,user\_id order by date DESC ) AS user\_count," +  "pageid,actionid " +  "from user\_count").write.format("jdbc").mode(SaveMode.Append).  jdbc(UserPropertis.getKeyValue("mysqlJdbc"),UserPropertis.getKeyValue("mysqlTable"),new Properties()) |

SaveMode：保存模式有以下几种方式



**2.6 hive 部署需要注意的几点以及Version information not found 错误解决办法**



安装HIVE 过程中要注意

1，mysql 是否正常运行

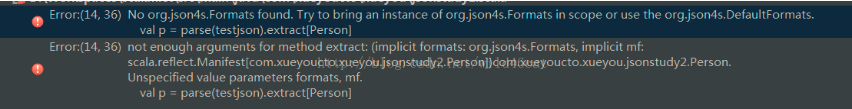
2. 创建好mysql 用户并分配好相应的访问权限以及数据库端口号等

3. mysql-connector-java-5.1.26-bin.jar 是否放到hive/lib 目录下 建议修改权限为777 （chmod 777 mysql-connector-java-5.1.26-bin.jar）

4. 修改conf/hive-site.xml 中的 “hive.metastore.schema.verification” 值为 false 即可解决 “Caused by: MetaException(message:Version information not found in metastore. )”

|  |
| --- |
| <property>  <name>hive.metastore.schema.verification</name>  <value>false</value>  </property> |

**2.7在进行json转换的时候可能会遇到这个错误**



这个时候需要一个转换：

**implicit val formats = Serialization.formats(ShortTypeHints(List()))**

还有依赖的包：

**import org.json4s.jackson.Serialization**