Description

You are provided with historical sales data for 45 stores located in different regions - each store contains a number of departments. The company also runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks.

[ **Promotional markdowns:** These are discounts that derive from any type of promotional sale such as a temporary price reduction, circular promotion, coupons, endcap promotions and more.]

Data files – Stores, Features and Sales

Stores

Information about the 45 stores, indicating the type and size of store

Features

Contains additional data related to the store, department, and regional activity for the given dates.

* Store - the store number
* Date - the week
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA
* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday week

Sales

Historical sales data, which covers to 2010-02-05 to 2012-11-01. Within this tab you will find the following fields:

* Store - the store number
* Dept - the department number
* Date - the week
* Weekly\_Sales -  sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week

The Task

1. **Data exploration**  
   Based on the dataset, decide on the datatype to be used to store the information.
2. **Data Copy from source to HDFS**  
   Create a **bash job (unix script)** to copy the datafiles to respective directories on hdfs. (No manual copy of data is allowed)
3. Create a **hive script** that will create three raw tables each for stores, features and sales.

While creating these raw tables, follow this nomenclature **raw\_<tablename>**

1. Analysis the data through hive tables for any missing information. Try to run hql queries directly on raw tables to understand it in depth.

(Hint : You can run simple queries to understand the data in a specific table, run joins to understand overall understanding of data )

1. Create a **hive script** that will create the final tables each for stores, features and sales.

The names of these final tables should be **Stores**, **Features** and **Sales**

1. Create another hive script that reads from the raw tables, supplies any missing data or add any new information (if needed) and inserts into final tables.
2. Create a hive script that returns the department-wide sales for each store in a given year
3. Create a hive script that returns the region-wide sales for each store in a given year
4. Write a sqoop to move these statistics to mysql tables (create required tables in mysql)
5. Prepare an oozie job to chain the processes here in a job (need to have a single workflow for this processing)
6. Schedule you oozie workflow to trigger once a week.