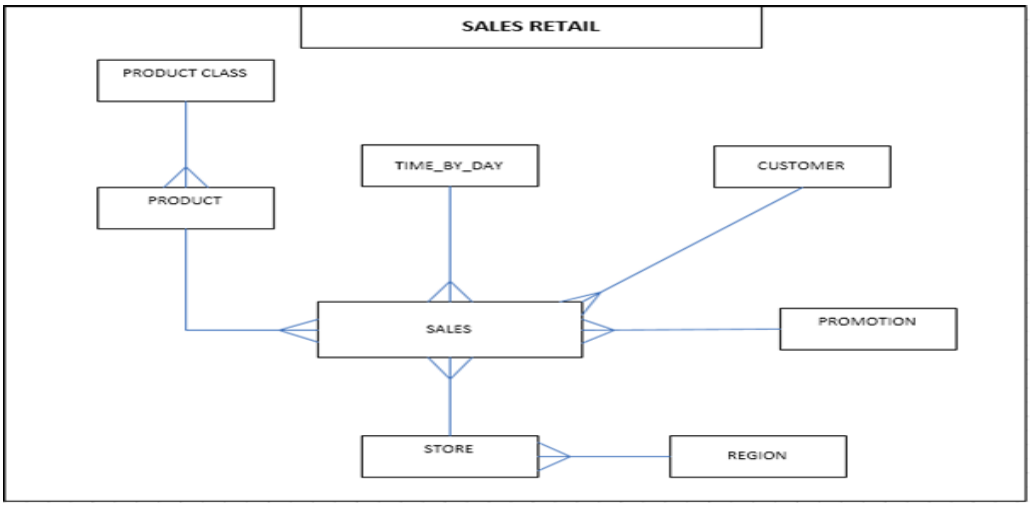
Case Study for Retail Sales

Objectives

The objective is to enable participants to get a feel of working with production datasets and analyzing data using tools in Hadoop ecosystem. This POC requires extracting required data from MySQL database, loading into Hadoop and analyzing data for the below listed retail schema.

Retail Schema:



Foodmart DB for MySQL can be downloaded from the below link:
Foodmart DB: http://pentaho.dlpage.phi-integration.com/mondrian/mysql-foodmart-database
Foodmart Schema: http://www2.dc.ufscar.br/~gbd/download/files/courses/DW&OLAP_2009/foodmart.jpg

Assignment:

a. Find total Promotion sales generated on weekdays and weekends for each region, year & month
b. Find the most popular promotion which generated highest sales in each region

Steps Involved:

Load data into HDFS using SQOOP

a. Load retail data from foodmart database into MySQL.

b. Extract following columns from **Promotion** table in Food mart DB using SQOOP.
PromotionID, Promotion Name, Promotion Cost

c. Extract following columns from **Sales_Fact_1997 & Sales_Fact_1998** tables in Food mart DB using SQOOP.
Region_id, ProductID, StoreID, PromotionID, Store Sales, the_day (day of week), the_month, the_year

Case Study for Retail Sales

Write a Map Reduce Program to find total Promotion sales generated on weekdays and weekends

a. Use the file loaded from Sqoop as input
b. In the Mapper Class, filter out sales records which are not part of any promotion.
c. In reducer class, perform the following transformations:
    a. Find total StoreSales for weekdays and weekends for given regionID, promotionID, sales_year, sales_month
    b. Load Promotion file into Distributed cache and lookup the file to get Promotion Name and Promotion Cost for given promotionID
    c. Save the following output as a **JSon** or **CSV** File
    sales_year, sales_month, region_id, PromotionID, Promotion Name, Promotion Cost, Week_day_sales, week_end_sales

Load data into Hive table and execute Queries in Spark (using scala)

a. Read the JSon / CSV file using a Hive External Table

b. Create and load a Hive Partitioned table partitioned by region_id, sales_year, sales_month

c. Write the following queries using Apache Spark and save the output in another hive table:
    i. Query1: List the total weekday sales & weekend sales for each promotions:
    Region ID, Promotion ID, Promotion Cost, total weekday sales, total weekend sales

    ii. Query 2: List the promotions, which generated highest total sales (weekday + weekend) in each region. Following columns are required in output:
    Region ID, Promotion ID, Promotion Cost, total sales

Publish Output to a Kafka Cluster (using scala)

a. Modify the spark code to write the result set of query 1 above to a Kafka topic called "**RetailSales**"
b. Write a Kafka Consumer which will read records in topic **"RetailSales"** and load data into a HBase Tables