**Objectives**

In this project, we will be migrating the existing Retail project to use the New Architecture using Spark, Airflow and Kafka.

**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:**

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

**Steps Involved:**

1. Create pySpark scripts for initial and incremental loads. The script will read sales and promotion tables based on last\_update\_date column from mysql and store them in AVRO format in S3 buckets. You might want to add a last\_update\_date in the tables.

* Source to Raw S3 bucket
  + AVRO Format

1. A second pySpark script will read the AVRO files, filter out all non-promotion records from input, join the promotion and sales tables and save the data in Parquet format in S3 buckets.

1. The Parquet file is aggregated by regionID, promotionID, sales\_year, sales\_month to generate total StoreSales for weekdays and weekends and the output is saved as a CSV file in S3 buckets.
2. The CSV file generated is loaded into a Snowflake database.
3. Following queries are executed on Snowflake table

Query1: List the total weekday sales & weekend sales for each promotions:

Region ID, Promotion ID, Promotion Cost, total weekday sales, total weekend sales

Query 2: List 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

1. Automate the workflow using Airflow scheduler

**Additional Assignment:**

1. a. Modify the spark code in step a. to write the result to Kafka topics called “**RetailSales**” and “**PromoData**”
2. b. Write a Spark Kafka Consumer which will subscribe to above topics and write the output in AVRO format in S3 buckets.