We did a case study on data ingestion process using pySpark, Python, Airflow, AWS (S3) and Snowflake. In this case study, objective is to load:

* source data to S3 Raw Bucket (Avro format);
* S3 Raw Bucket to S3 Cleansed Bucket;
* S3 Cleansed Bucket to Curator S3 bucket;
* Curator S3 bucket to Snowflake database.

Dumping data Filter Data

Raw

S3

Dumping Data

Cleansed S3

Filter data

MySql Database

Transformation(s)

Staging

S3

Semantic S3

Curator S3

Snowflake

Cloud DB

To achieve this, we downloaded a sample database from open source – **Foodmart DB for MySQL**

Steps followed:

Data Exploration & Data Modeling

**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>

* The foodmart database has data on sales promotions, stores, and regions. In this case study, our goal was to find the total sales generated on weekdays and weekends for each promotion, region, year, and month.
* After this, we found the most popular promotion which generated highest sales in each region.
* We

1. - Created a pyspark script for initial\_load where it will read sales and promotion tables based on last\_update\_date column from MySQL database and stored them in AVRO format in S3 bucket.
2. Then created a separate incremental load script to load the updated and newly added sales and promotions to the tables.

Data Transformations:

1. Another script created to read the AVRO files, to filter out all non-promotion records from input, joined the promotion and sales tables and saved the data in Parquet format in S3 bucket.
2. Then 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.

Data Quality:

1. Then the CSV file generated is loaded into a Snowflake database. Then queried to:
   * List the total weekday sales & weekend sales for each promotion:
   * List promotions, which generated highest total sales (weekday + weekend) in each region.

Airflow Automation - Scheduler

1. As a final step, created an Airflow DAG to automate the workflow using Airflow scheduler.