As part of the capstone project, you will assume the role of the Associate Data Warehouse Engineer who has recently joined an e-commerce organization. You will be presented with a business challenge that requires building a data platform for retail data analytics.

In this Capstone project, you will: Design a data platform that uses MySQL as an OLTP database and MongoDB as a NoSQL database, design and implement a data warehouse and generate reports from the data, design a reporting dashboard that reflects the key metrics of the business, extract data from OLTP and NoSQL databases, transform it and load it into the data warehouse, and then create an ETL pipeline, and finally, create a Spark connection to the data warehouse and then deploy a machine learning model.

In Module 1, you will design the OLTP database for an e-commerce website, populate the OLTP Database with the data provided, and automate the export of the daily incremental data into the data warehouse.

In Module 2, you will set up a NoSQL database to store the catalog data for an E-Commerce website, load the E-Commerce catalog data into the NoSQL database, and query the E-Commerce catalog

data in the NoSQL database.

In Module 3, you will design the schema for a data warehouse based on the schema of the OLTP and NoSQL databases. You’ll then create the schema and load the data into the fact and dimension tables, automate the daily incremental data insertion into the data warehouse, and create Cubes and Rollups to

make the reporting easier.

In Module 4, you will create a Business Intelligence dashboard. You will create a Cognos data source

that points to a data warehouse table, create a bar chart of quarterly sales of cell phones, create a pie chart of sales of electronic goods by category, and create a line chart of total sales per month for a given year.

In Module 5, you will extract data from an OLTP, NoSQL, and MongoDB databases into CSV format. You will then transform the OLTP data to suit the data warehouse schema, and then load the transformed data into the data warehouse. Finally, you will verify that the data is loaded properly.

You are a data engineer at an e-commerce company. You need to keep data synchronized between different databases/data warehouses as a part of your daily routine. One task that is routinely performed is the sync up of staging data warehouse and production data warehouse. Automating this sync up will save you a lot of time and standardize your process. You will be given a set of python scripts to start with. You will use/modify them to perform the incremental data load from MySQL server which acts as a staging warehouse to the IBM DB2 or PostgreSQL which is a production data warehouse. This script will be scheduled by the data engineers to sync up the data between the staging and production data warehouse.

**Scenario**

Write a pipeline that analyzes the web server log file, extracts the required lines(ending with html) and fields(time stamp, size ) and transforms (bytes to mb) and load (append to an existing file.)

In this assignment, you will author an Apache Airflow DAG using **Bash operators** that will:

* Extract data from a web server log file
* Transform the data
* Load the transformed data into a tar file

In the sixth and final module, you will use your skills in Big Data Analytics to create a Spark connection to the data warehouse and then deploy a machine learning model on SparkML for making sales projections.