Module 4: Deep Dive into Apache Spark Framework

Case Study 2

edureka!

Case Study: Instacart
Domain: E-commerce

Instacart is a grocery ordering and delivery app that aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. Instacart's data science team plays a big part in providing this delightful shopping experience. Currently, they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session.

The dataset is a relational set of files describing customers' orders over time. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.

Tasks:

As a Big Data consultant, you are helping the data science team to deal with a large amount of data. To solve this, you are pitching in to move the transactional data from RDBMS to HDFS by:

1. Verify the cluster health including HDFS and Spark
2. Test the spark environment by executing the spark's sort.py example.
3. Try to implement the same example in scala and perform spark-submit.
4. Analyze the behavior of spark application on Spark web UI
5. Add custom logs in your application and re-execute the application. Once executed check the Spark logs.
6. Transfer complete dataset from RDBMS to HDFS
7. Validate the loaded data by comparing the statistics of data both in source and HDFS
8. Create a new directory in HDFS named cheeses and load only rows where aisle is "specialty cheeses"
9. update "specialty cheeses" to "specialty cheese" and transfer only updated rows in the above-created directory.

Dataset: You can download the required dataset from LMS