Big Data Hadoop and Spark Developer

Course-End Project Problem Statement

Market Basket Analysis Using Instacart Online Grocery Dataset

Problem Statement:
Instacart is a grocery delivery and pick-up service that is available in the United States of America and Canada. The company's services can be accessed through a website and a mobile app. The data was collected anonymously and contained a sample of over 3 million grocery orders from over 200,000 instacart consumers.
Each user receives between 4 and 100 orders, with each order containing the order of products purchased. The company also provides the week, hour, and day of the order, as well as the time interval between orders. The tables are included in this dataset, and a description of each row is provided below:

Dataset Description:

1. **Orders.csv** : Consists of 3.4 million rows, 206 thousand users
2. **Products.csv** : 50 thousand rows
3. **Aisles.csv**: 134 rows
4. **Departments.csv**: 21 rows
5. **order_products__SET**: 30 million + rows where SET is defined as:
    a. **order_products _prior.csv**: 3.2 million previous orders
    b. **order_products_train.csv:** 3.2 million order information

Objective:

1. To evaluate online shopping company data in order to assist businesses in identifying the day when the most orders were placed in order to provide deals for that day
2. To determine which department is responsible for the most product launches

Steps Overview:
Step 1: Upload the **insta-cart** file to the HDFS

1. Download the relevant dataset from the **Reference Materials** section or the project description
2. Upload the dataset to the **FTP** lab from your local system

3. To move the dataset to **HDFS** from the **Terminal** use the put command

Syntax of put command for reference id given below:

hdfs dfs -put <FTP_folder _name_source> <hdfs_path_destination>

Example:

hdfs dfs -put insta-cart /user/abcsimplilearn/insta-cart-project

Step 2: Perform the below tasks on the uploaded dataset using pySpark:

1. Explore the orders dataset and create a dataframe
2. Replace all the null values
3. Identify the busiest day
4. Calculate the busiest hour
5. Identify the most popular item based on the order count by exploring order_products__prior and products datasets
6. Explore the department dataset and create a dataframe
7. Recognize the department which has published the maximum products

Tasks to be performed:

Task 1: Explore the orders dataset and create a dataframe

Aim: To read the **orders.csv** file stored in **HDFS** and create a dataframe

Steps to perform:

1. Read the orders data as a dataframe in pySpark

   Note: The column **days_since_prior_order** may contain some NULL values.
2. Display the data up to 10 rows

Task 2: Replace all the null values

Aim: To delete existing **Null** values

Steps to perform:

1. Replace all null values with a dummy **999** value in the dataframe that was created in task
2. Show top 10 records

Task 3: Identify the busiest day

Aim: To examine the order information and find the busiest day of the week by reading the data as a pySpark dataframe

Steps to perform:

1. Read the **orders.csv** data stored in **HDFS**

2. Store the table for a particular spark session
3. Compute the total number of orders as **Total_Orders** placed on each day of the week

   Hint: The column **order_dow** represents the day of the week

   Wherein:

   Day 0 is Sunday

   Day 6 is Saturday and so on
4. Display the result that contains the total orders placed on each day of the week (Monday to Sunday)

   Example:

Total_Orders	Day_of_the_week
600905	Sunday
587578	Monday

Task 4: Calculate the busiest hour
Aim: Give a breakdown of orders by the hour
Steps to perform:

1. Read the **orders.csv** data stored in **HDFS**
2. Store the table for a particular spark session
3. Select the number of order IDs as **Total_Orders** and the hour at which the order was placed using spark sql
4. Display the result that contains total orders and the hour

   Example:

Total_Orders	Hour
22758	0
345678	1

Task 5: Identify the most popular item
Aim: To identify the top 10 popular items
Steps to perform:

1. Read the **order_products__prior.csv** and **products.csv** data
2. Store the tables for a particular spark session
3. Calculate the top 10 popular items based on the count of orders

4. Display the result that contains the product name as **Popular_product_name** and the count of order id as **Order_Count**

Example:

Order_Count	Popular_Product_Name
22758	0
345678	1

Task 6: Explore the department dataset and create a dataframe
Aim: Read the department CSV as a pySpark dataframe
Steps to perform:

1. Read the data from the **departements.csv** file
2. Display the data stored

Task 7: Recognize the department that published the maximum number of products
Aim: Identify the department ID which has published the maximum products
Steps to perform:

1. Read the **department.csv** data
2. Store the tables for a particular spark session
3. Display the department id that has published maximum products

Example:

Department_ID	Max_Product
10	38
2	548