Big Data Hadoop and Spark Developer Market Basket Analysis Using Instacart







Objectives

- To analyze company data to assist businesses in identifying when the most orders were placed to provide deals for that day
- To determine which department is responsible for the most product launches







Prerequisites



- Python
- PySpark
- Hadoop
- Spark



Industry Relevance



- **Python:** It is used for web development, data science and data analysis, machine learning, startups, and the finance industry.
- **PySpark**: It writes Spark apps in Python.
- Hadoop: It is used in public sectors like intelligence, science, cyber security, and defense.
- **Spark:** It is used for machine learning and streaming data.



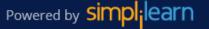


Problem Statement



- 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 application. The data was collected anonymously and contained a sample of over 3 million grocery orders from over 200,000 Instacart consumers.
- The company also provides the week, hour, and day of the order, as well as the time interval between orders to their customers.





Tasks to Perform



- Perform the following tasks on the dataset provided using PySpark:
- 1. Explore the orders CSV file and create a DataFrame
 - Read the orders data as a DataFrame in PySpark
- Note: The column "days_since_prior_order" may contain NULL values
 - Display the data up to 10 rows
- 2. Replace all null values with a dummy "999" value in the DataFrame that was created in task 1
- 3. Examine the orders CSV file and find the busiest day of the week by reading the data as a PySpark DataFrame
 - Display the result that contains the total orders placed on each day of the week (Monday to Sunday)



Tasks to Perform



- Perform the following tasks on the dataset provided using PySpark:
- 4. Give a breakdown of orders by the hour and identify the busiest hour
 - Select the number of order IDs as "Total_Orders" and the hour at which the order was placed
 - Display the result that contains total orders and the hour
- 5. Identify the most popular item based on the order count by exploring order_products__prior and products datasets
 - Calculate the top 10 popular items based on the count of orders
- 6. Explore the department dataset and create a DataFrame
- 7. Recognize the department which has published the maximum products
 - Display the department ID that has published the maximum products



Project Outcome



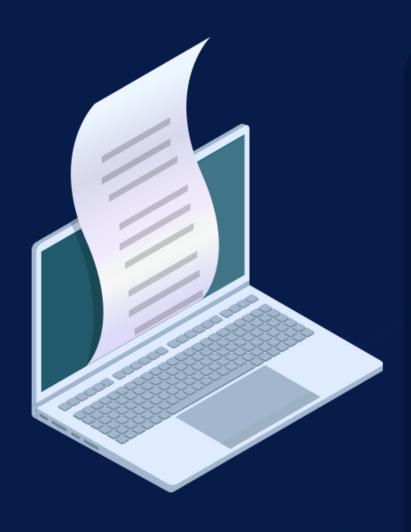
• This project is designed to help understand working with the dataset and performing analysis.

• You should be able to process the dataset for this project to produce reports.

You will be able to use PySpark, perform analysis, and obtain the desired results.



Submission Process



- 1. Complete the project in the Simplilearn lab
- 2. Complete each task listed in the problem statement
- 3. Take screenshots of the results for each question and the corresponding code
- 4. Save it as a document and submit it using the assessment tab
- 5. Tap the "Submit" button (this will present you with three choices)
- 6. Attach three files and then click "Submit"
- Note: Be sure to include screenshots of the output



