**Publishing Date : 30 August 23 Submission Date: September 06-23 - 23:59**

**BIG DATA SYSTEMS**

1. **Business Context**

In the context of taxi business analysis, the yellow taxi trip records are a valuable resource for businesses that operate in the transportation industry. By analyzing this dataset, businesses can gain insights into passenger demand, pricing strategies, service improvement, and new product development.

**2.** **Business Problem Understanding**

The Department of Transportation is seeking assistance in understanding passenger demand, businesses can determine where and when taxis are needed most. This information can be used to optimize taxi fleet size and deployment, which can help to reduce wait times and improve customer satisfaction. Our objective is to analyze this dataset and provide actionable insights to enhance businesses and gain insights that can help them to improve their operations and to better serve their customers.

Your team has been appointed to take a closer look at the data and analyze it to get insights for:

a) Identifying factors contributing to requesting the service

b) Recommending strategies for improved service.

1. **Data Understanding**

For this analysis, the department is expecting your team to explore the usage of Hadoop and Spark for storage and querying data. The data is downloaded from Kaggle.com and is attached. The attached data set is a scaled down version of the data available on Kaggle.com.

1. **Data preparation and Exploratory Data Analysis**

You are supposed to utilize appropriate data pre-processing techniques on the given data set. If required, make appropriate assumptions and make it explicitly known while using them in the query. Make appropriate selection of the attributes with sound justification for the same. The data set allows for several new combinations of attributes and attributes exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the analysis.

1. **Expected Outcomes**

You are expected to set up a Hadoop Spark based analytics platform and find out the answers to questions framed as part of the analytics.

**1. System setup:**

Configure a single / multi node Hadoop cluster on Linux / Windows machine.

Move the attached file *yellow\_tripdata\_2020-06.csv* to a folder in HDFS.

Install Apache Spark and configure it to make use of the YARN job scheduler of the Hadoop framework.

**2. Analytics program development:**

**2.1 Write Spark programs either in pySpark or Scala to do the following.**

**1.** Read the data in *yellow\_tripdata\_2020-06.csv* file into a dataframe created in spark.

– **1 Mark**

**2**. Count the number of taxi trips for each hour – **1 Mark**

**2.2 Create a table view of the data frame created in step 1 above and write**

**SparkSQL queries to find out the following:**

1. Average fare amount collected by hour of the day – **1 Mark**
2. Average fare amount compared to the average trip distance – **1 Mark**
3. Average fare amount and average trip distance by day of the week. – **1 Mark**
4. In the month of June 2020, find the zone which had maximum number of pick ups. – **1 Mark**
5. In the month of June 2020, find the zone which had maximum number of drops. – **1 Mark**
6. Average no of passengers by hour of the day. – **1 Mark**
7. Total number of payments made by different type for the month. – **1 mark**
8. Configuring Hadoop cluster and Spark installation on the cluster – **2 Marks**
9. Video uploading – **1 marks**

**Total – 12 Marks**

1. **Instructions for submitting the assignment**

**a) The assignment consists of 11 questions. (Total 12 Marks)**

**b) The submission should consist of 2 files:**

**A PDF file containing answers to all the questions based on the analysis that you have carried out earlier along with the supporting queries, pySpark, Scala programs that you have written to extract the answers. Also include the Hadoop, Spark configuration files along with the environment variables you have set.**

* + - Name the PDF file in format like "Grp\_<your\_group\_number>.doc" only. Don't add anything into the file names. Add the group member names in the PDF.
    - Make sure that you upload the file well ahead of the deadline. At last moments, we have seen several groups have faced issues while doing the submissions.
    - The group leader will be giving a contribution weightage of each group member and marks will be awarded to the group members accordingly.
    - Note - Since it is a group assignment, only one submission is expected from each group. Unnecessarily don’t upload the solution on individual basis. If it’s observed, then a penalty (25% reduction) will be applicable on it.

**d) Every group should record an mp4 video which should contain the executions with queries/answers.**

Name the video file in format like "Grp\_<your\_group\_number>.mp4" only. Don't add anything into the file names.

**e) Plagiarism will be strictly dealt with and if found will result in cancellation of the Assignment and 0 marks being awarded to all the group members**.

**f) Submissions done as mail attachments will be rejected**

**g) Images/photographs of text/PDF documents will not be evaluated**

**h) The last date of submission will not be extended in any case.**

**7. References**

* [Yellow Taxi Dataset](https://www.kaggle.com/datasets/microize/newyork-yellow-taxi-trip-data-2020-2019) (Drop Fields not used in the queries)

**About DataSet**

**Context**

The yellow taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP).

**Content**

Column Description

**tpep\_pickup\_datetime :** The date and time when the meter was engaged.

**tpep\_dropoff\_datetime :** The date and time when the meter was disengaged.

**Passenger\_count :** The number of passengers in the vehicle.( This is a driver-entered value )

**Trip\_distance :** The elapsed trip distance in miles reported by the taximeter.

**PULocationID :** TLC Taxi Zone in which the taximeter was engaged

**DOLocationID :**TLC Taxi Zone in which the taximeter was disengaged

**Payment\_type** A numeric code signifying how the passenger paid for the trip.

---- 1= Credit card

---- 2= Cash

---- 3= No charge

---- 4= Dispute

---- 5= Unknown

---- 6= Voided trip

**Fare\_amount :** The time-and-distance fare calculated by the meter.

**Extra :** Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.

**MTA\_tax :** $0.50 MTA tax that is automatically triggered based on the metered rate in use.

**Improvement\_surcharge :** $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.

**Tip\_amount :** Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.

**Tolls\_amount :** Total amount of all tolls paid in trip.

**Total\_amount :** The total amount charged to passengers. Does not include cash tips