**Final Project**

Hemalatha Subbiah

College of Science and Technology, Bellevue University

[DSC650- Big Data](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_512542_1)

Professor. Nasheb Ismaily

May 31, 2024

**Introduction:**

The WhatsgoodlyData dataset appears to be a collection of survey responses or poll data, likely gathered from users regarding their preferences or behaviors. Each row in the dataset represents a specific response to a particular question. The dataset is structured with multiple attributes that describe each response in detail.

**Columns**:

**1.Question**: contains the survey or poll question that was asked to the respondents.

**2**. **Segment Type**: categorizes the type of segment the respondents belong to.

**3**. **Segment Description**: column provides a more detailed description of the segment.

4. **Answer: column contains the possible answers or choices provided for the question**

**5.Count:** column shows the number of respondents who chose each particular answer.

**6. Percentage:** column represents the percentage of respondents who chose each answer, expressed as a decimal.

**Problem Statement:**

To analyze the preferences of mobile users in response to notifications from different social media platforms (Instagram, Facebook, Snapchat, and LinkedIn) and determine which platform is most frequently checked first by users. Additionally, to identify trends and patterns in user behavior based on the segment type and segment description.

**Key Questions**:

1. Which social media platform's notification is most likely to be checked first by mobile users?

2. What percentage of mobile users check each social media platform's notification first?

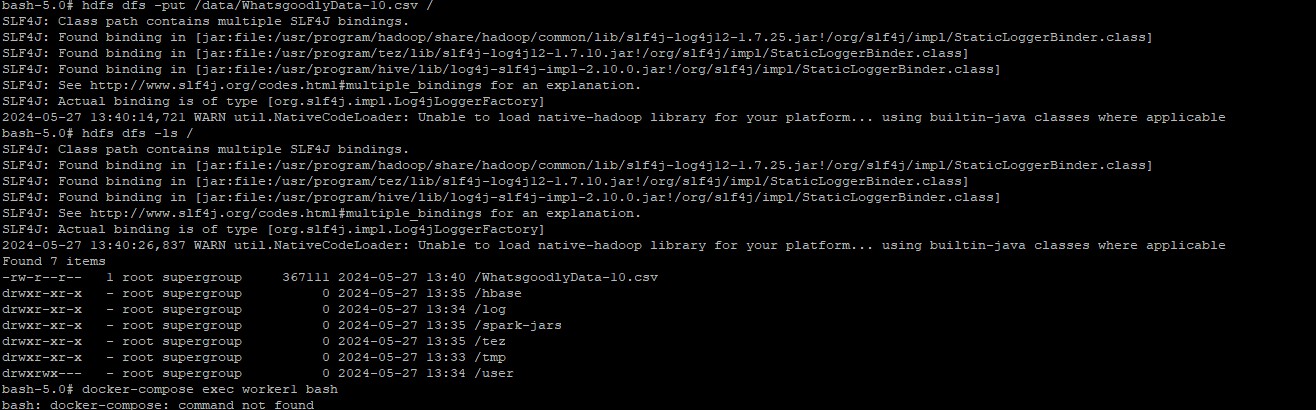
3. Are there any noticeable trends or patterns in user preferences based on segment descriptions?

4. How do the preferences vary across different segments of mobile users?

I present the basic steps to ingest data from this dataset to a Hadoop Spark cluster, store the data into HBase, a NoSQL columnar database, and consume this data using SQL language.

The data ingestion process will consist of two distinct phases, namely HDFS file transfer and subsequent HBase data storage.

**Loading Data to HDFS:**



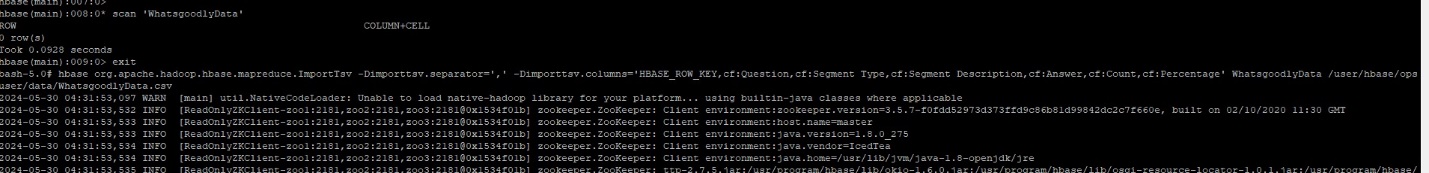
**Moving data to HBASE:**

Copied the file into the HDFS filesystem from where Hbase can read the file to load the data into the Hbase table.

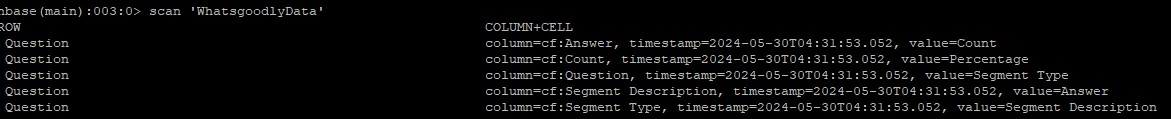
A screenshot of a computer program

Description automatically generated

For loading the CSV data into Hbase table I used the MR job which will load the data from CSV file into the Hbase table.



Scanned the table



Upon the successful completion of the file transfer, I leveraged a Spark dataframe to efficiently read the transferred data and automatically deduce its schema. I did this through PySpark.

I created a SparkSession and load data into a DataFrame using PySpark

A screenshot of a computer program

Description automatically generated

I performed various operations on the dataframe, such as filtering, aggregating, and transforming the data using PySpark’s DataFrame API.

In this example, we filter the dataframe to select rows where the value in the “column name” column is greater than 100.

A screen shot of a computer

Description automatically generated

The map transformation applies a function to each element in the RDD or DataFrame and returns a new RDD or DataFrame.

A black screen with white text

Description automatically generated

**Grouping and Aggregating Data:**

Grouping and aggregation operations enable summarizing data based on certain grouping criteria. I aggregated data by segment type, segment description, and answer to get total counts and percentages for each combination.

A black screen with white text

Description automatically generated

A screen shot of a computer

Description automatically generated

A computer screen with white text

Description automatically generated

**Conclusion:**

PySpark provides a Python API for interacting with Spark, enabling data manipulation, integration with machine learning algorithms, and real-time stream processing and the integration of Spark and HBase represents a powerful and efficient approach to processing social network data.

From my all my transformations like select, group by, aggregate and sort transformations able to conclude that **SnapChat** is most likely to be checked first by mobile users. Younger users may prefer Snapchat, while professionals may prefer LinkedIn. Marketers can target advertisements based on preferred platforms of different user segments. Based on the outcome social media platforms and marketers can make data-driven decisions to enhance user experience and engagement.