* **Introduction**

For this big data class project, I will develop a data pipeline that illustrates the integration of essential components in the big data ecosystem. I will utilize five technologies, including NiFi, HDFS, Hive, HBase, and Spark. This pipeline will show the complete data journey, from ingestion to storage, transformation, and analysis, helping me understand how data flows in a distributed environment. By constructing this pipeline, I will emphasize the advantages of each tool: NiFi for data collection and flow control, HDFS for scalable storage Hive for effective querying and structured data storage, and Spark for advanced data processing and transformation. By the end of this project, I will demonstrate my capability to build strong, scalable, and efficient big data solutions with interconnected components suited for managing high-volume data and real-time processing.

* **Data**

For this Project I did chose the mobile User dataset from Kagle. This dataset provides a thorough look at how people use mobile devices. It includes 700 samples with specific details like app usage time, screen-on time, battery usage, data consumption, and more. Each entry is sorted into one of five categories based on user behavior, ranging from light to heavy usage, which helps in effective modeling and analysis. Important features include user demographics, device information (like model and operating system), and daily mobile engagement metrics.

I will study mobile user behavior using this dataset by applying linear regression. This will help us find the main factors that affect usage patterns and predict user engagement. I will look at features like app usage time, screen-on time, battery drain, data usage, and demographics to create regression models. For instance, I can analyze how the number of apps installed, age, or device model affects app usage time or battery drain. This method will reveal which factors have the biggest impact on user behavior, giving us insights into mobile engagement trends and allowing us to make data-driven suggestions for improving device performance and user experience.

See below links for dataset:

GitHub

https://raw.githubusercontent.com/Mfrakso/DSC650/refs/heads/main/mobile\_user.csv

Kagle

<https://www.kaggle.com/datasets/valakhorasani/mobile-device-usage-and-user-behavior-dataset>

* **Pipeline Overview**

In this data pipeline, Apache NiFi is the first tool used to ingest a csv file which includes our dataset. It cleans and standardizes the data before sending it to HDFS for safe and scalable storage. HDFS serves as the main data hub, allowing access for further processing. Apache Spark then retrieves data from HDFS to carry out complex transformations, such as feature engineering and linear regression analysis, to find key factors influencing user behavior. The modified data is saved in HBase, which allows for quick querying and easy access for analytics and reporting. Finally, a visualization tool like Tableau or Power BI can be connected to Hive to present insights from the regression analysis, offering clear visuals of mobile user engagement trends and helping to identify actionable insights into usage patterns. This pipeline ensures a seamless transition from data collection to storage, processing, and analysis, showcasing the effectiveness of big data tools in understanding mobile user behavior.

* **Issues Encountered**

Frequent freezing and the need to restart everything, including the Google VM, can be very annoying, especially when using HDFS in a virtual setup. This problem happens because of not having enough resources, as HDFS, Spark, and similar services require a lot of them. When the VM does not have enough memory or CPU power, it freezes.

I had to shut down and restart the VM often which was very time consuming. At the end I had to stop NiFi after ingesting the data to HDFS which did help a bit in having some extra memory.

A computer screen shot of a program

Description automatically generated

Moreover, low disk space can make HDFS or the whole VM unresponsive, so it is important to regularly check disk usage and delete unnecessary files. Properly allocating memory based on workload needs can help avoid crashes and freezing, leading to a more stable environment for data-heavy tasks.

I forgot NiFIi running which loaded too much data into HDFs and I had to do some Logs cleaning before being able to start the system again.

You can see here the disc space used at 100% A screen shot of a computer program

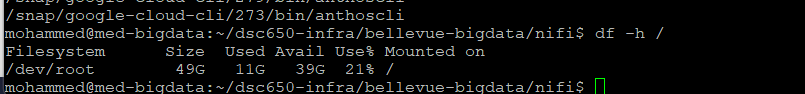
Description automatically generated

And this is how I solved it

docker system prune -a --volumes -f

sudo find / -type f -size +100M 2>/dev/null

after the cleaning only 21% used



I ran into datatype issues while applying machine learning algorithms in spark. I had this error: IllegalArgumentException: Data type string of column ... is not supported, this typically occurs when PySpark's machine learning (ML) transformations encounter columns with unsupported data types.

I had to change datatypes using the following code

from pyspark.sql import functions as F

mobile\_df = mobile\_df \ .withColumn("screen\_time", F.col("screen\_time").cast("double")) \ .withColumn("app\_usage", F.col("app\_usage").cast("double")) \ .withColumn("battery\_drain", F.col("battery\_drain").cast("double")) \ .withColumn("apps\_installed", F.col("apps\_installed").cast("double"))

* **Screenshots of System Deployment**

NiFi Flow including UpdateAttribute to conver data and put it into HDFS

A screenshot of a computer

Description automatically generated

Chowing our data in HDFS under project directory

A screen shot of a computer code

Description automatically generated

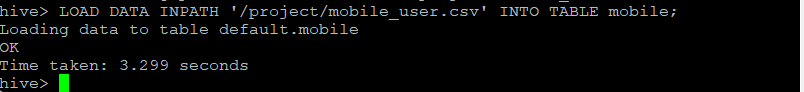
Creating the hive Table

A screen shot of a computer program

Description automatically generated

Loading the data to the table in hive

LOAD DATA INPATH '/project/mobile\_user.csv ' INTO TABLE mobile;



Showing table with 700 rows

A black screen with white text

Description automatically generated

Moving to Spark

A computer screen with white text

Description automatically generated

Changing names to correct format

A computer code on a black background

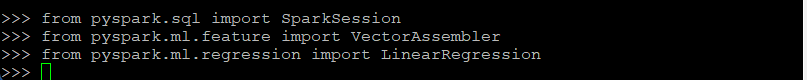
Description automatically generated

Showing the data with new header

A screenshot of a computer program

Description automatically generated

Importing libraries for linear regression



making sure we have the right data types

A computer screen with white text

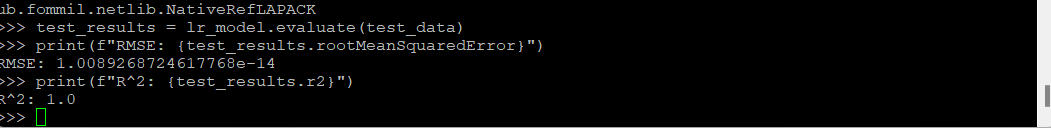
Description automatically generated

Starting ML by dividing data randomly

A screen shot of a computer program

Description automatically generated

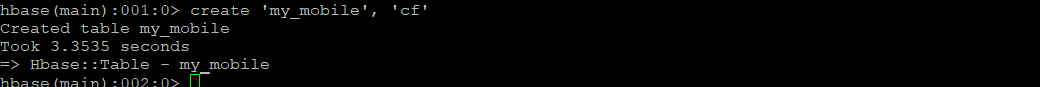
the model performance metrics

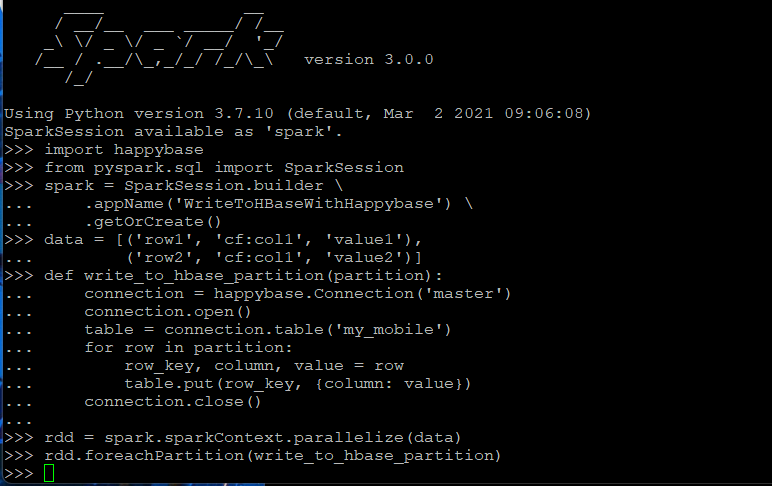


Results interpretation:

The very low RMSE (about 1.0089…e-14) and an R² of 1.0 show that the model's predictions closely match the actual values, indicating it accounts for 100% of the variance in the test data. While this suggests high accuracy, it may also raise concerns like overfitting, where the model might struggle with new data, or data leakage, where the test data may have unintentionally affected the training process, leading the model to memorize instead of generalizing. To ensure the model's reliability, it is important to test it on a different data set to verify its performance across various data.

Saving data in HBase:





A screen shot of a computer program

Description automatically generated

Showing table in Hbase

A screen shot of a computer

Description automatically generated

* **Code**

<https://raw.githubusercontent.com/Mfrakso/DSC650/refs/heads/main/Code.md>

<https://github.com/Mfrakso/DSC650/blob/main/Code.md>

* **Conclusion**

This project demonstrates the integration of key big data tools—NiFi, HDFS, Hive, HBase, and Spark—to build a comprehensive data pipeline that processes, stores, and analyzes mobile user data. Using a dataset with detailed metrics on mobile usage, this pipeline begins with NiFi for data ingestion, which then sends cleaned data to HDFS for centralized storage. Spark processes the data through transformations and linear regression analysis to identify factors influencing user behavior, and the refined data is stored in HBase for fast access. Finally, a visualization tool like Tableau or power BI can be connected to Hive, enabling clear, actionable insights.

Although challenges like memory constraints and datatype errors occurred, solutions such as optimizing disk usage and converting data types allowed smooth operation. This pipeline showcases efficient data flow from ingestion to analysis, highlighting the utility of big data tools in generating insights into mobile user behavior. The project's results, including low RMSE and high R² metrics, underline the model's accuracy, though further testing on new datasets is essential to validate its robustness.