|  |
| --- |
|  |

|  |
| --- |
|  |

|  |  |
| --- | --- |
|  |  |



Supervised by: Mr. Aziz NABIL

Intern: Ayoub ETOULLALI

Internship Title: Airbnb Database Analysis and ML Modeling on the Data Center Platform

Internship Period: 26/06/2023 to 26/09/2023

LinkedIn Profile

ACTIVITY REPORT

table of contents

1

[INTRODUCTION 2](#_Toc145419942)

[Methodology of Work 2](#_Toc145419943)

[Access to the Platform 3](#_Toc145419944)

[Database Integration with Hive Using HDFS 4](#_Toc145419945)

[Description about the Issue 4](#_Toc145419946)

[Data Preparation: ETL from HDFS to PostgreSQL 5](#_Toc145419947)

[Project Overview 5](#_Toc145419948)

[Database Description 6](#_Toc145419949)

[Key Database Concepts 6](#_Toc145419950)

[Data Schema 6](#_Toc145419951)

[Data Engineering and Data Science Ideas 6](#_Toc145419952)

[Business Problem 6](#_Toc145419953)

[Documentation of Airbnb New User Bookings 6](#_Toc145419954)

[Machine Learning Model with Spark MLlib 7](#_Toc145419955)

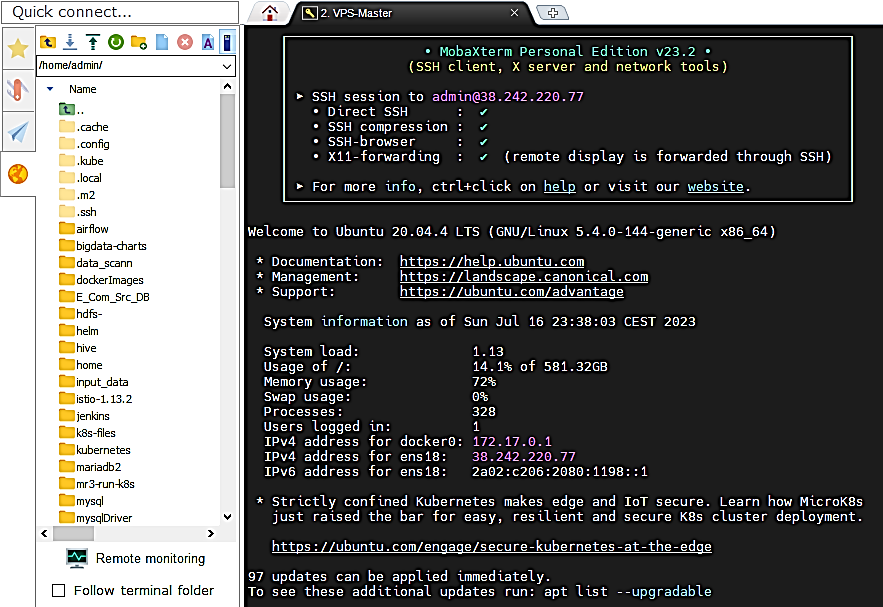
[CONCLUSION 8](#_Toc145419956)

|  |
| --- |
| INTRODUCTION |

|  |
| --- |
| During my internship, I would like to provide a brief overview of the internship's objectives and context. I had the privilege of completing my internship at **NTT DATA**, and during this period, I was entrusted with a range of tasks and projects. This report will offer insights into the purpose, the organization, and the various assignments I undertook during my internship. |
| Methodology of Work  During my internship, our work was conducted remotely, with a team of four members collaborating closely. Our typical workday consisted of daily meetings where we discussed our progress, challenges, and goals for the day. We followed a dynamic approach to our tasks, often working in parallel when tasks could be divided and distributed effectively. This allowed us to maximize efficiency and complete projects promptly.  Furthermore, our supervisor played a crucial role in guiding our work. As we completed tasks, our supervisor would provide valuable feedback and assign new tasks based on our collective progress and skill development. This iterative process ensured that we continually engaged with challenging and relevant projects, enhancing our learning experience throughout the internship. |

2

|  |
| --- |
| Access to the Platform |

I had the opportunity to access the company's data platform during my internship, which proved to be a pivotal aspect of my work. This platform, known as the "Centers Data Platform," is the core of the company's Big Data operations. My initial task was to gain access to this platform, which was facilitated through the use of the MobaXterm tool. MobaXterm enabled me to establish a secure connection to the platform, granting me access to the entire Big Data infrastructure of the organization.

3

Understanding the technologies integrated into the platform was a significant part of my role. Here are some of the key technologies that constitute the platform:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *NiFi* | *Spark* | *HDFS* | *PostgreSQL* |  | *Hive* | *Grafana* |  | *Airflow* | *Jenkins* |
| for Data Ingestion | for batch processing and data science pipelines | for storing large volumes of data | for the Airflow’s database backend |  | for the analytical data store | for data visualization |  | for the orchestration of the workflows | for execution spark jobs |

**Application

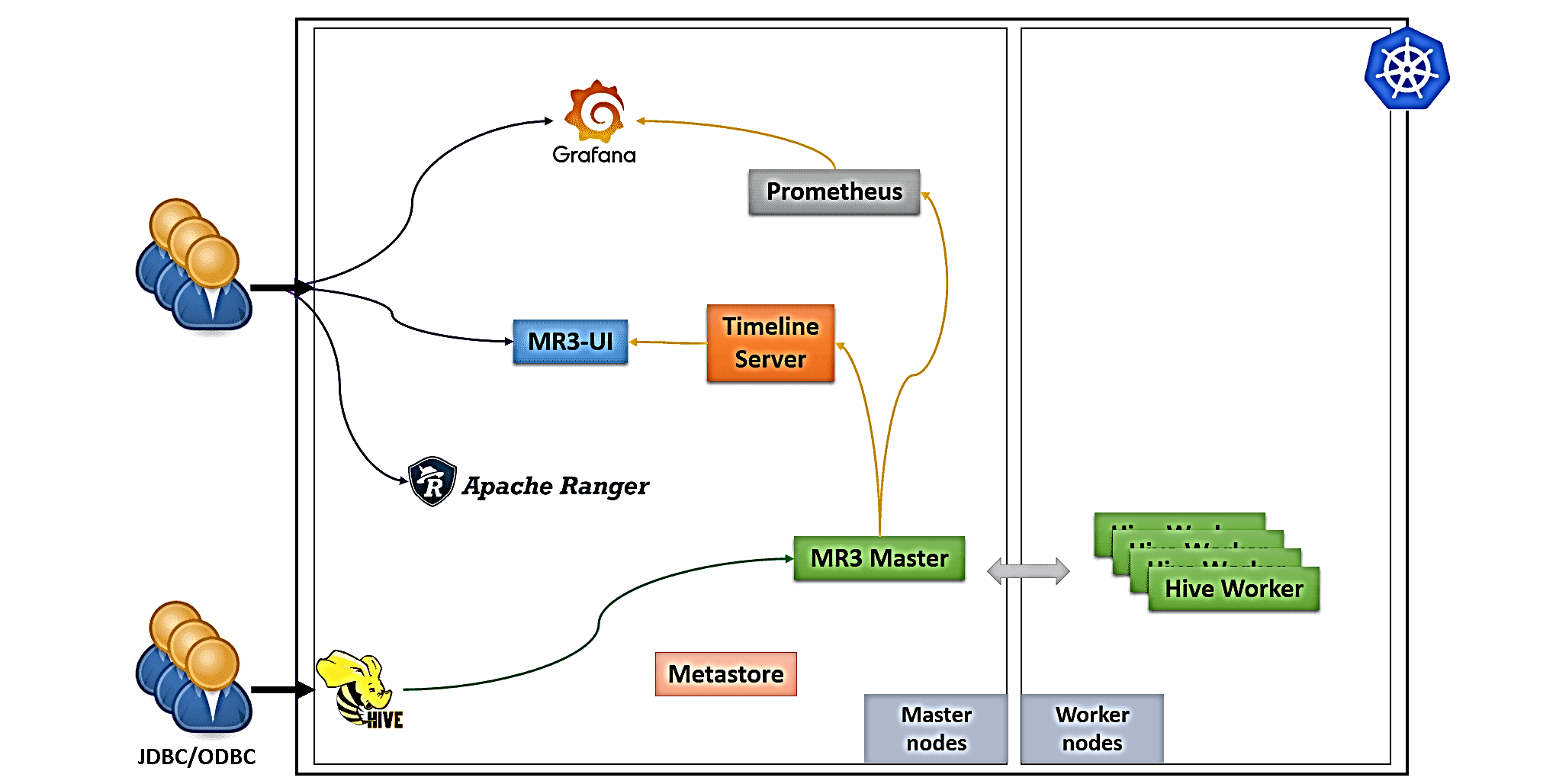
Description automatically generated with medium confidence**

These technologies, among others, were seamlessly integrated into the platform to create a robust environment for managing and analyzing massive datasets. My role included gaining proficiency in these tools, understanding how they interacted within the platform, and leveraging them effectively to accomplish various data-related tasks. This experience was instrumental in enhancing my understanding of Big Data technologies and their practical applications in a real-world setting.

|  |
| --- |
| Database Integration with Hive Using HDFS |

One of the primary objectives of my internship was to integrate the "Airbnb New User Bookings" database with Apache Hive using the Hadoop Distributed File System (HDFS). This integration aimed to enable the company to run SQL queries on the massive datasets stored within HDFS.

4

However, it's worth noting that the problem remained unresolved. I encountered a technical issue with Hive that made this task more challenging than anticipated. The problem was identified as an incorrect configuration within Hive, and it took me some time to troubleshoot and rectify this configuration to achieve a successful integration. This experience allowed me to develop my technical troubleshooting skills and gain a deeper understanding of Hive's internal workings. While the problem persisted, it served as a valuable learning experience and highlighted the complexities involved in working with Big Data technologies.

Description about the Issue

While this issue presented challenges during the internship, it also provided valuable learning experiences in Kubernetes cluster management and troubleshooting, which are essential skills in the field of DevOps and cloud-native application deployment.

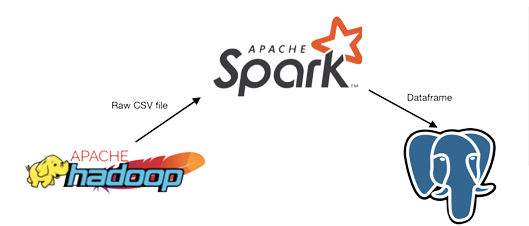
|  |
| --- |
| Data Preparation: ETL from HDFS to PostgreSQL |

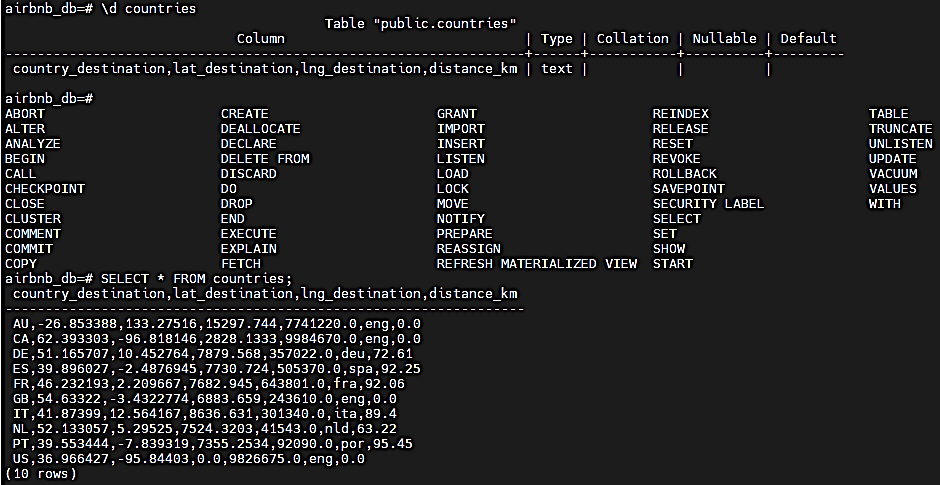
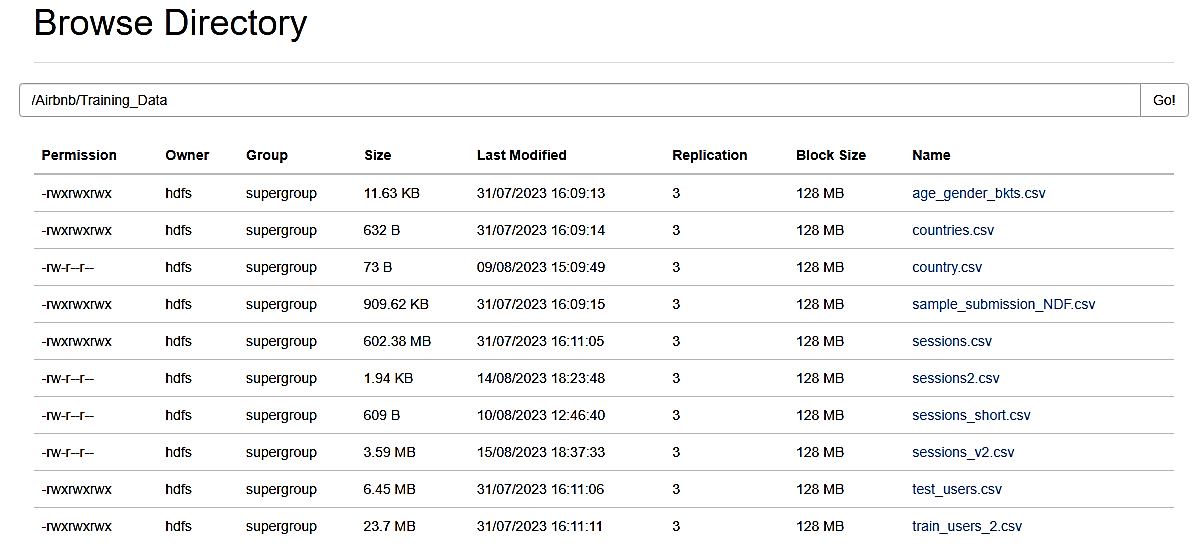
During my internship, I took on the responsibility of data preparation. This crucial phase involved the Extract, Transform, Load (ETL) process to transfer data from HDFS (Hadoop Distributed File System) to a PostgreSQL database. To streamline and ensure the quality of data transfers, I developed custom ETL scripts.

5

Project Overview

|  |  |  |
| --- | --- | --- |
| Objective The primary objective of this ETL job was to extract data from HDFS, perform necessary transformations, and load it into a PostgreSQL database for future analysis. | Loading method Scala Spark was utilized for this process. The ETL job involved a series of steps, summarized as follows:   1. Initialization 2. Data Reading 3. Data Transformation 4. Writing to PostgreSQL 5. Reporting | Conclusion This ETL job automated the data loading process from HDFS to PostgreSQL using Scala Spark. It provided flexibility to manage diverse data sources, perform customized transformations, and load data into PostgreSQL for future analysis. |





|  |  |
| --- | --- |
|  | Database Description  During my internship, I had the opportunity to work with the Airbnb New User Bookings database. This database is a rich source of information containing crucial data related to users, their demographics, web session records, and summary statistics.  The primary objective of this database is to predict the initial destination of a new Airbnb user traveling from the United States. There are twelve possible destination outcomes, including popular locations like the United States, France, Canada, and others. It's important to note that "NDF" (No Destination Found) differs from "other" because "other" signifies a booking in a country not on the list, whereas "NDF" means no booking was made.  Key Database Concepts  The database comprises various categorical variables, such as gender and registration method, along with timestamps for web sessions and session durations in seconds. The "user\_id" column in the "sessions.csv" file can be joined with the "id" column in "train\_users.csv" and "test\_users.csv" to integrate web session records with user information.  Data Schema  To provide clarity on how data is organized, a data schema is depicted below, illustrating the structure and relationships between different datasets used in the project. A clear data schema is vital for understanding data organization.  Data Engineering and Data Science Ideas  The report offers several data engineering and data science ideas. Data engineering ideas encompass data cleaning, preprocessing, feature engineering, data integration, and pipeline creation. Data science ideas encompass exploratory data analysis, predictive modeling, clustering analysis, and recommendation systems.  Business Problem  The section highlights the core business problem this project aimed to solve: predicting the initial destination for new Airbnb users. Emphasis is placed on the significance of this prediction for Airbnb's goal of providing personalized experiences and optimizing its operations. |

6

|  |
| --- |
| Documentation of Airbnb New User Bookings |

|  |
| --- |
| Machine Learning Model with Spark MLlib |

|  |  |
| --- | --- |
| During my internship, I had the exciting opportunity to work on constructing a Machine Learning (ML) model using Spark MLlib. The aim of this model was to predict user behaviors based on the collected data, and through this process, I explored various ML techniques and developed a model that showed promising results. | |
| |  | | --- | |  | | Exploratory Data Analysis (EDA) Exploratory Data Analysis (EDA) is a crucial step in the data analysis process. It involves a deep examination and understanding of the features and structures within a dataset. The main objective of EDA is to uncover hidden insights, trends, patterns, and anomalies in the data.   |  |  | | --- | --- | | **Univariate Analysis** | **Bivariate Analysis** | | In this phase, we examine each variable in isolation, visualizing distributions, descriptive statistics, and handling missing values. The dataset displayed a significant imbalance in user destinations, with the majority traveling solely within the United States ("US"). | This stage involves studying relationships between two variables. We observed relationships between country destinations and the types of devices used by users and between user age and session duration. |  Choosing Machine Learning Models In building the predictive model, I experimented with several machine learning algorithms available in Spark MLlib. These included Random Forest, Logistic Regression, and Gradient-Boosted Trees. I conducted hyperparameter tuning to optimize model performance, leveraging techniques like grid search and cross-validation. Results and Insights I began by selecting the most appropriate Machine Learning models for our prediction task. After experimenting with several algorithms, the Random Forest (RF) model yielded the best results, with an accuracy of 0.87613 on Kaggle. Addressing Performance Issues One challenge was the dataset's size, which exceeded my computer's capacity. To address this, I employed sampling techniques to work with data subsets, reducing computational load and speeding up the process. Experimentation and Evaluation Time Model experimentation and evaluation were time-consuming due to complex models and a large dataset. A systematic experimentation process and experiment tracking tools helped streamline this. Handling Imbalanced Dataset The dataset featured an imbalance in destinations, with a majority not booking any destination ("NDF"). |

Building a Machine Learning model with Spark MLlib was a crucial step in our project. Challenges included model selection, performance issues, experimentation time, and handling an imbalanced dataset. However, a systematic approach and appropriate tools led to a high-performing Random Forest model with an accuracy of 0.87613 on Kaggle. These skills in Machine Learning and project management will be invaluable for my future career as a data scientist.

7

|  |
| --- |
| CONCLUSION |

My internship at **NTT DATA** has been a transformative experience, providing valuable insights into the world of data science and technology. Here, I summarize the key takeaways from this enriching journey:

8

In conclusion, my internship at **NTT DATA** was a rewarding journey that equipped me with practical skills, instilled a problem-solving mindset, and nurtured effective teamwork. These experiences have not only enriched my academic knowledge but also provided a strong foundation for my future career. I am grateful for this opportunity and look forward to applying the skills and knowledge gained here to make meaningful contributions in the field of data science and technology.