Thadomal Shahani Engineering College, Bandra, Mumbai

Department of Computer Engineering Big Data Analytics ( Mini Project) SEM VII

**Subject: Big Data Analytics ( CSC702)**

**AY: 2024-25**

**Experiment 10 (Mini Project)**

**Aim:** Design the infrastructure of a Big Data Application.

**Tasks to be completed by the students:**

Task 1: Choose a problem definition which requires handling Big Data. Task 2: Design the data pipeline for your application.

Task 3: Deploy your project on suitable platform.

Task 4: Test your application with different volume, variety and velocity of data.

Thadomal Shahani Engineering College, Bandra, Mumbai

Department of Computer Engineering Big Data Analytics ( Mini Project) SEM VII

## Report on Mini Project

**Subject: Big Data Analytics ( CSC702) AY: 2024-25**

# Housing Data Analysis Using HIVE

## Parth Puranik : 2103142 Mohib Abbas Sayed : 2103158 Hamza Sayyed : 2103159 Om Shete : 2103163

**Guided By**

## (Dr. Arti Deshpande)

Thadomal Shahani Engineering College, Bandra, Mumbai

Department of Computer Engineering Big Data Analytics ( Mini Project) SEM VII

## CHAPTER 1: INTRODUCTION

**1.1 Aim**

The aim of this project is to perform a comprehensive analysis of the housing dataset using big data technologies, specifically Hadoop and Hive, to extract actionable insights into the real estate market. By analyzing trends and patterns within the dataset, the project aims to assist property investors, real estate professionals, and potential homebuyers in making informed decisions regarding housing investments, pricing strategies, and market behavior.

**1.2 Objective**

The primary objective of this project is to harness the power of distributed computing through Hadoop and data querying via Hive. By analyzing a large dataset of housing information, we aim to:

* Uncover geographic investment hotspots.
* Identify key property features that influence pricing.
* Provide a clear picture of market fluctuations over time. The insights generated from this analysis will empower stakeholders to make data-driven decisions regarding real estate investments and market strategies.

**1.3 Project Overview**

In today’s fast-evolving real estate market, having access to reliable and comprehensive data is crucial for making informed decisions. This project uses the Kaggle housing dataset, a rich resource with details on property prices, area, number of rooms, and other important features. Using Hadoop’s distributed processing capabilities and Hive’s SQL-like query functionality, we processed and analyzed this data to extract valuable insights. The key focus of this project is to explore trends that can help in understanding the factors driving housing prices and how these factors vary across different locations.

Thadomal Shahani Engineering College, Bandra, Mumbai

Department of Computer Engineering Big Data Analytics ( Mini Project) SEM VII

## CHAPTER 2: DATA DESCRIPTION AND ANALYSIS

* 1. **Dataset Overview**

The dataset used in this project is a housing dataset sourced from Kaggle. It contains various attributes related to properties such as:

* **Price**: The sale price of the property in rupees.
* **Area**: The total area of the property in square feet.
* **Bedrooms**: The number of bedrooms in each property.
* **Bathrooms**: The number of bathrooms in each property.
* **Stories**: Number of floors in the property.
* **Mainroad**: Whether the property is located on a main road (yes/no).
* **Guestroom**: Presence of a guestroom (yes/no).
* **Basement**: Whether the property has a basement (yes/no).
* **Airconditioning**: Whether the property is air-conditioned (yes/no).
* **Parking**: Number of available parking spaces.
* **Furnishingstatus**: The furnishing status of the property, such as ‘furnished’ or ‘semi-furnished.’
  1. **Data Preprocessing**

Before analyzing the data, preprocessing steps were conducted to ensure data integrity and usability. This included:

* **Data Cleaning**: Handling missing values, removing duplicates, and correcting inconsistent entries.

Thadomal Shahani Engineering College, Bandra, Mumbai

Department of Computer Engineering Big Data Analytics ( Mini Project) SEM VII

* **Feature Transformation**: Converting categorical features such as "Furnishingstatus" and "Mainroad" into binary or numerical values for analysis.
* **Data Normalization**: Standardizing features like "Area" and "Price" to ensure uniform scaling for better comparison.
  1. **Exploratory Data Analysis**

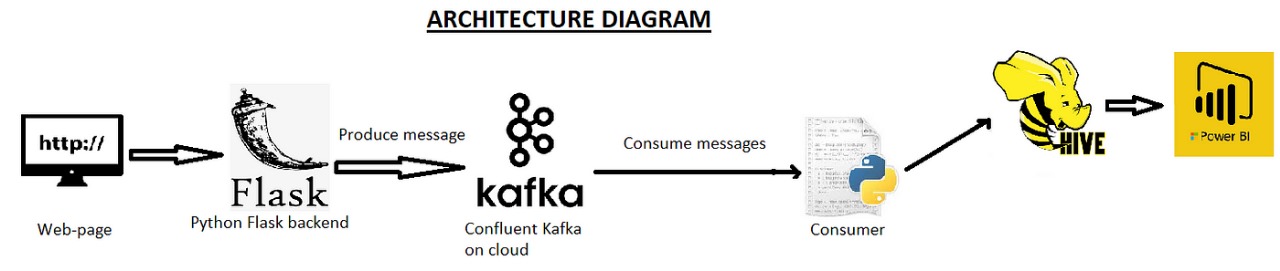
Exploratory Data Analysis (EDA) was conducted to identify trends and patterns in the data. Key insights included:

* **Price Distribution**: Analysis revealed that properties in preferred areas tend to have higher prices, with the average price being significantly influenced by the property’s area and number of stories.
* **Geographic Insights**: Properties located on the main road are more expensive on average, and regions with good amenities have a higher concentration of valuable properties.

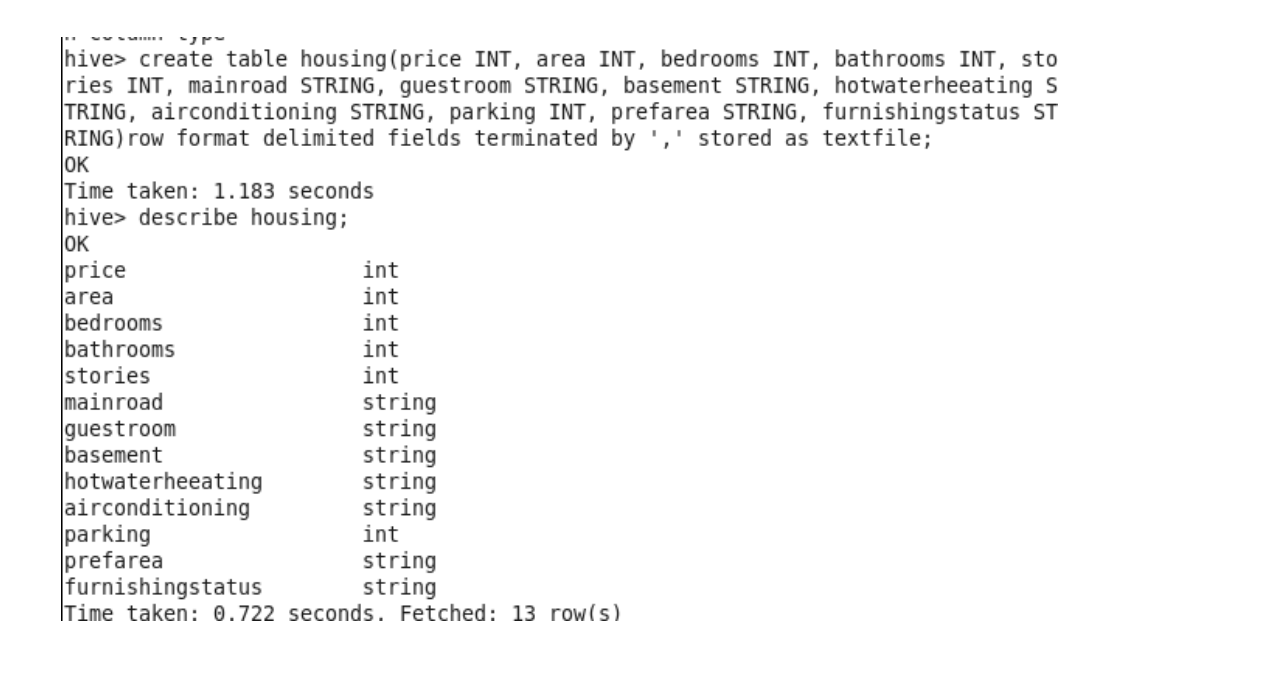
Thadomal Shahani Engineering College, Bandra, Mumbai

Department of Computer Engineering Big Data Analytics ( Mini Project) SEM VII

## CHAPTER 3: DESIGN OF DATA PIPELINE

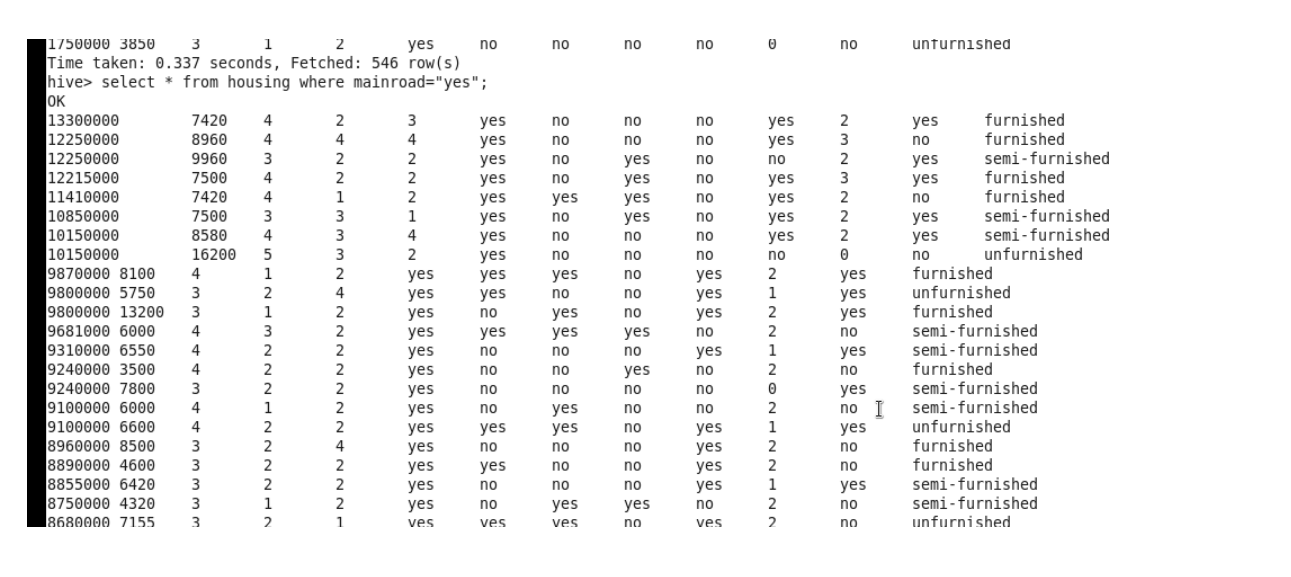
****

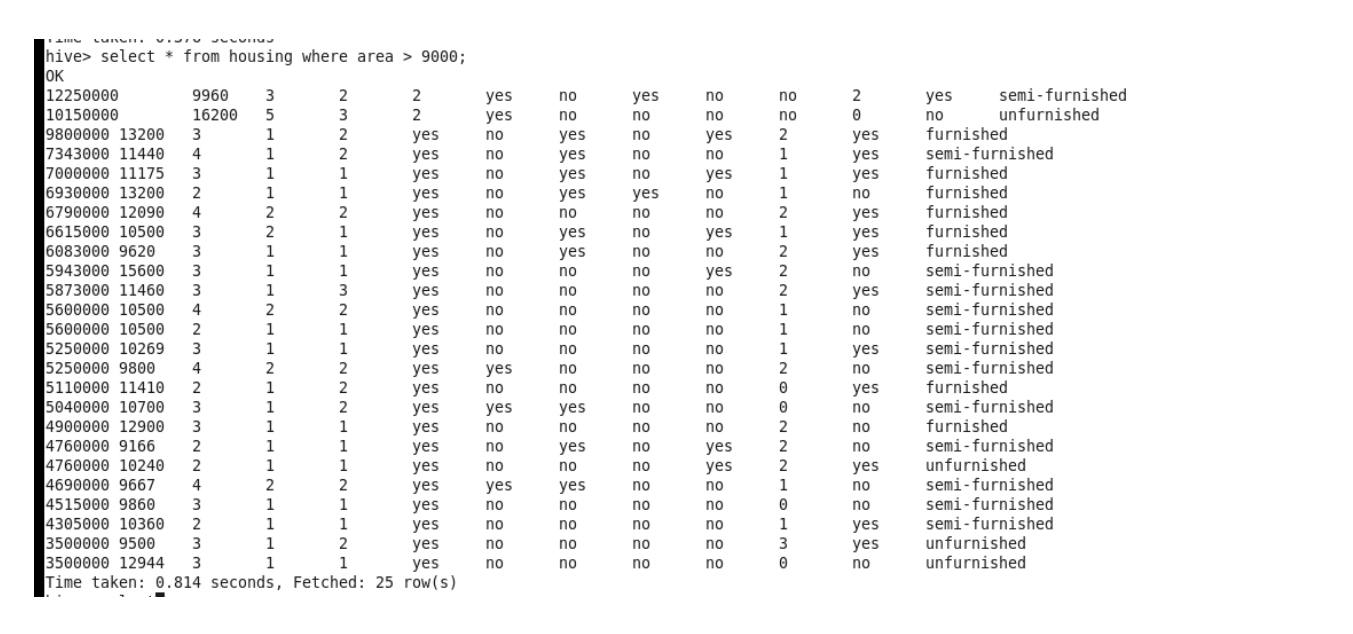
**CHAPTER 4: RESULT ANALYSIS**

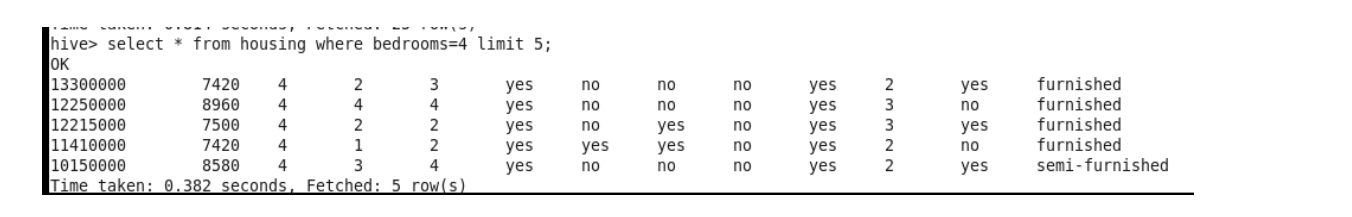


Thadomal Shahani Engineering College, Bandra, Mumbai

Department of Computer Engineering Big Data Analytics ( Mini Project) SEM VII







## CHAPTER 5: CONCLUSION AND FUTURE SCOPE

### Conclusion:

This project successfully demonstrated the value of using Hadoop and Hive for large-scale data analysis in the real estate market. The housing dataset was analyzed to extract insights that could assist various stakeholders, from property investors to homebuyers, in making data-driven decisions. The results highlighted key factors that influence property prices, as well as geographic regions with significant investment potential.

### Future Scope:

The project has laid the groundwork for further exploration into real estate analytics. Potential extensions include:

1. **Machine Learning Models**: Implementing predictive models to forecast property prices based on historical data and key features.
2. **Integration with Other Datasets**: Including datasets such as economic indicators, interest rates, and population growth to provide a more comprehensive analysis.
3. **Real-Time Analytics**: Enhancing the system to provide real-time insights using streaming data technologies like Apache Spark.