Real Estate Data Pipeline for Market Analysis

Problem Statement: A real estate company wants to analyze real estate data to understand market trends, price fluctuations, and other patterns. They need to build a data pipeline that collects, processes, and analyzes data from various sources to generate meaningful insights for decision-making. As a data engineer, your task is to design and implement the data pipeline to support this analysis.

Dataset:
The dataset for this project will consist of the Zillow Economics Data, which includes the following information:

- RegionID, RegionName, RegionType, StateName: Various details about the location of the property.
- SizeRank: Size rank of the property.
- Zhvi, MoM, QoQ, YoY, 5Year, 10Year: Various details about the price and price changes of the property.

Project Steps:

Data Collection:
Download the dataset from Kaggle.

Data Ingestion:
Create an ingestion process to receive and store the raw data from the CSV file.
Use tools like Apache Sqoop for batch data ingestion.
Store the data in a data lake or distributed file system (e.g., HDFS).

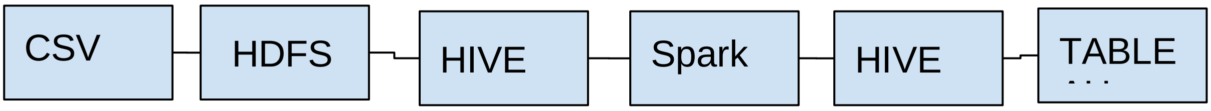
Data Processing:
Design ETL (Extract, Transform, Load) processes to cleanse and transform the raw data.
Implement data quality checks and filtering to ensure data integrity.
Utilize Apache Spark for distributed data processing and transformation.
Apply data modeling techniques (e.g., data normalization, denormalization) as per the analysis requirements.

Data Storage:
Choose a suitable database or data warehouse (e.g., Apache Hive, Apache HBase) for storing processed data.
Create optimized tables and partitions for efficient querying and analysis.
Ensure data security and privacy measures are in place.

Data Analysis and Visualization:
Use SQL queries or Spark SQL to extract relevant insights from the processed data.
Perform market analysis based on price fluctuations, market trends, and other patterns.

Generate reports, dashboards, and visualizations using tools like Apache Superset, Tableau, or Power BI.

Overall architecture flow:



CSV

HDFS

HIVE

Spark

HIVE

TABLE ...

Task 1: Data Ingestion and Storage

Outcome: Ingest the real estate data into Hadoop using Apache Sqoop, process it using Apache Spark, and store the results in HBase.

Deliverables:
Sqoop command to ingest data from the CSV file into Hadoop.
Spark code to process the ingested data.
HBase commands to store the processed data.

Task 2: Market Analysis

Outcome: Analyze the real estate data to identify market trends, price fluctuations, and other patterns using Apache Spark and SQL queries.

Deliverables:
Spark code and SQL queries to analyze the real estate data.
A report detailing the market trends, price fluctuations, and other patterns in the data.

Task 3: Data Visualization

Outcome: Create visualizations of the market trends, price fluctuations, and other patterns using a tool like Apache Superset, Tableau, or Power BI.

Deliverables:
Visual representations (e.g., line charts, bar charts, heatmaps) of the market trends, price fluctuations, and other patterns.
A dashboard that displays the visualizations and allows users to interact with the data.

- Average property price per region
- Trend of property prices over time
- Top 10 regions with the highest property prices
- Correlation between property size and price

For all these tasks, you can use the Zillow Economics Data mentioned above. Please note that you might need to preprocess the dataset to fit the exact requirements of the assignment. Also, remember to respect the terms of use for the dataset.

Tools required to achieve the end dashboard:

1. **Hadoop (hdfs , hive , spark)**
2. **Tableau/Power BI**