An Industry Oriented Mini Project Presentation

On

Handling Optimizations using View formats on Hive Data Warehouse

By

PARIMI JAHNAVI RANGASAI(20WJ1A6745)

PASNOOR SHREYA(20WJ1A6746)

DADAPURAM ANOOP REDDY (20WJ1A6712)

Under the Esteemed Guidance of

Mr. Ch. Srinivas

Asst. Professor, CSE-Data Science

Department of Computer Science Engineering – Data Science

GURU NANAK INSTITUTIONS TECHNICAL CAMPUS (AUTONOMOUS)

2023-2024

OVERVIEW

- Abstract
- Introduction
- Literature Survey
- Existing System
- Proposed System
- Module Description
- Design
- Implementation & Testing
- Results
- Conclusion
- Future Enhancement
- Thank you

ABSTRACT

- To enhance Hive data warehouse performance for real-time stock data analysis.
- Implementation of specialized views based on profit categories:

  -View 1: Targets customers with significant profits for investment plans.

  - View 2: Focuses on customers with minimal profits, guiding stock suggestions.
- Avoiding table duplication by using views, ensuring real-time updates.
- Emphasizing query optimization and rigorous performance testing for low-latency analysis.

INTRODUCTION

- Stock consultancy relies on effective data use.
- Abundant customer data holds stock insights.
- Hive aids in creating distinct views of profit and loss perspectives.
- These views serve as specialized lenses, providing a granular and comprehensive analysis of stock performances.
- Leveraging Hive's robustness for data warehousing.
- Crafting specialized analytical lenses with Hive.
- Enhancing stock consultancy insights via Hive's capabilities.
- Employing Hive to extract nuanced stock transaction details.

LITERATURE SURVEY

Title : Data Processing in Hive vs. SQL Server: A comparative analysis in the query performance
Author: Nadeem Ahmed; Shakil Ahamed; Jahir Ibna Rafiq; Sifatur Rahim

Description:

- Relational Database Management Systems (RDBMS) like MySQL, SQL Server, Oracle, and SQLite play a crucial role in data processing.
- Big data technology is popular for handling extremely large datasets in large organizations.
- Comparison study between traditional databases (SQLite, SQL Server) and Hive on Hadoop for Small Enterprises (SE).
- Suggests using traditional databases if the dataset fits on a single computer and there's no plan for handling vast amounts of data soon.

EXISTING SYSTEM

Existing System Overview:

- Data Sources :Utilizes MySQL databases and flat files for managing customer stock transaction data.
- Tooling: Relies on Excel for data manipulation and Tableau for visualization purposes.

Drawbacks of the Existing System:

- Data Fragmentation
- Query Response Times
- Scalability Issues

PROPOSED SYSTEM

Objective of the Proposed System:

- Centralized Data Handling :Leverage Hive's distributed data warehouse to centralize and optimize storage for diverse customer stock transaction data.
- Efficient Analysis: Utilize structured querying within Hive for streamlined and efficient analysis of stock transactions.

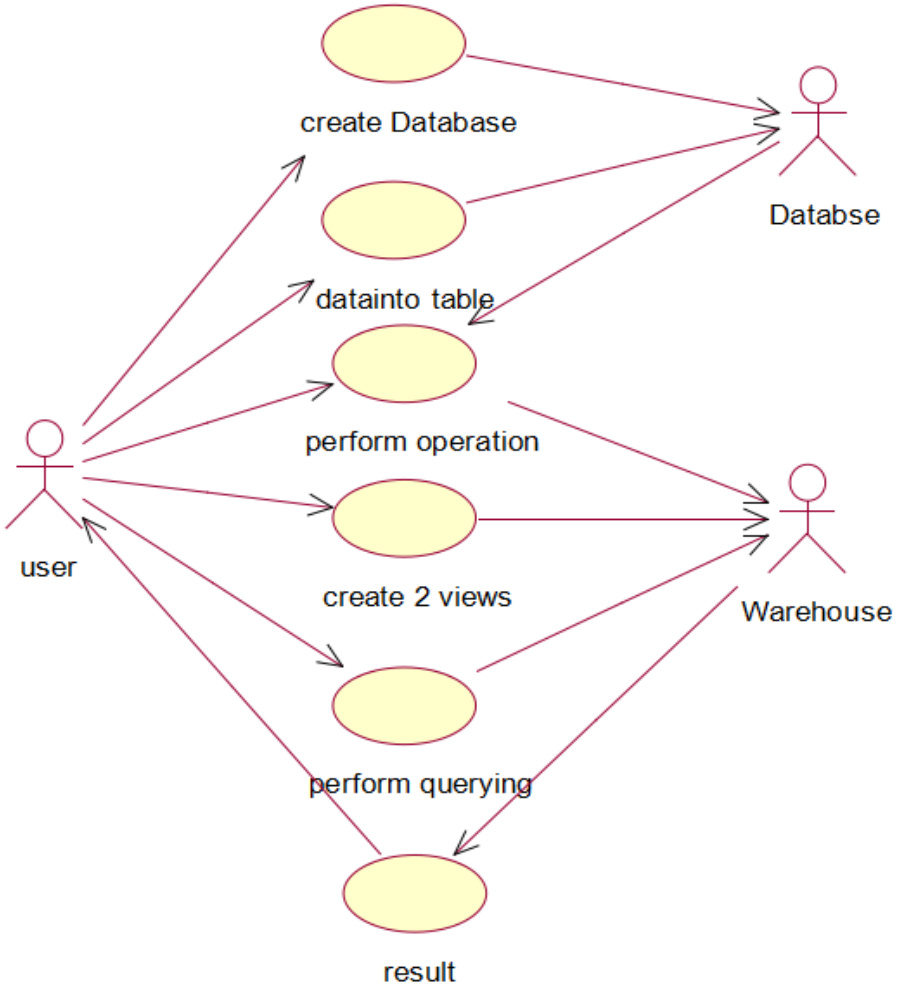
Advantages of the Proposed System:

- Centralized Data Storage
- Scalability
- Enhanced Query Performance

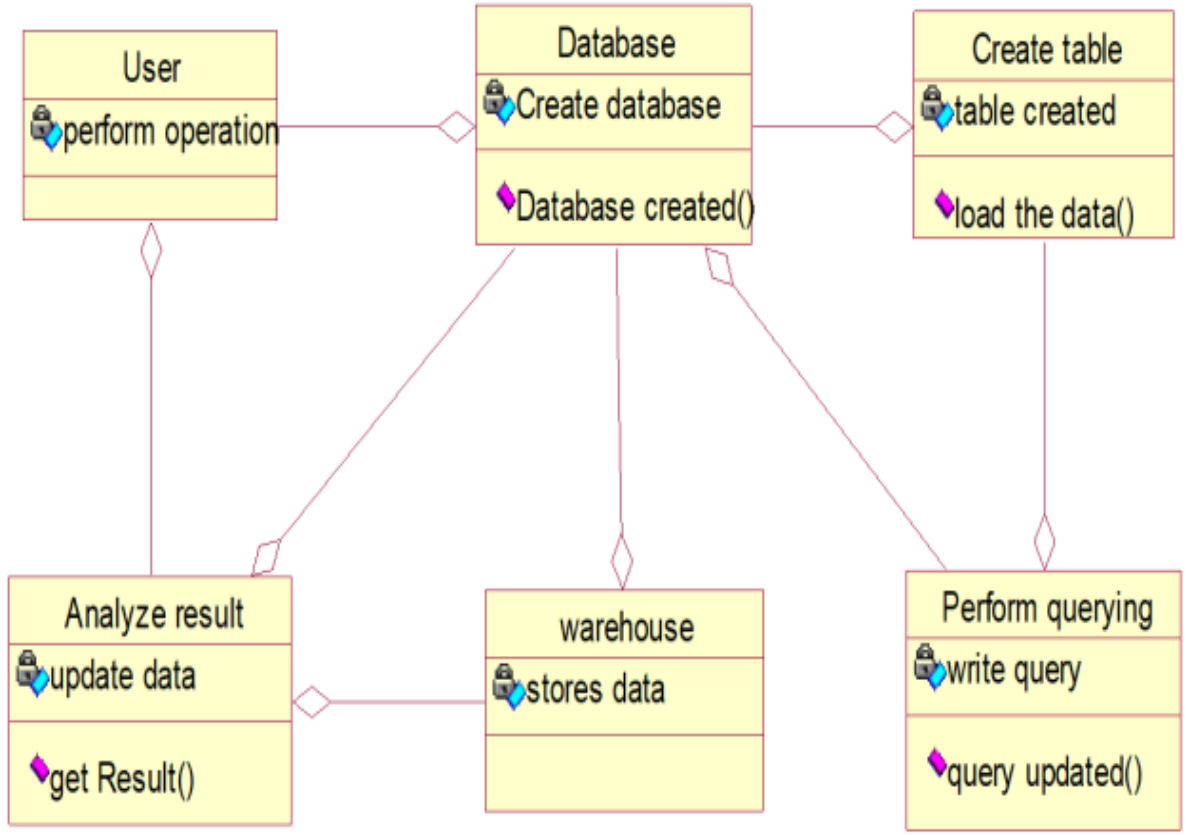
MODULE DESCRIPTION

- Data Ingestion
- Query Processing
- Custom View Creation
- Scalability and Performance
- Visualization and Reporting

DESIGN

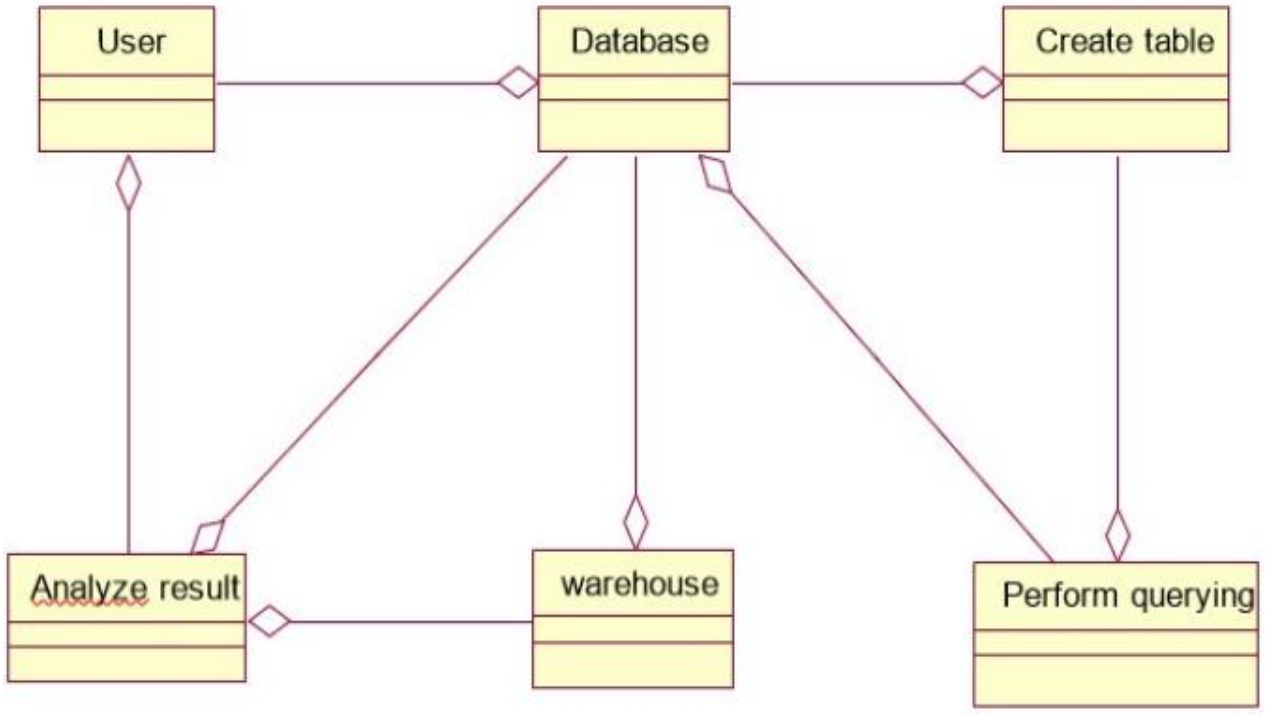


USE-CASE DIAGRAM

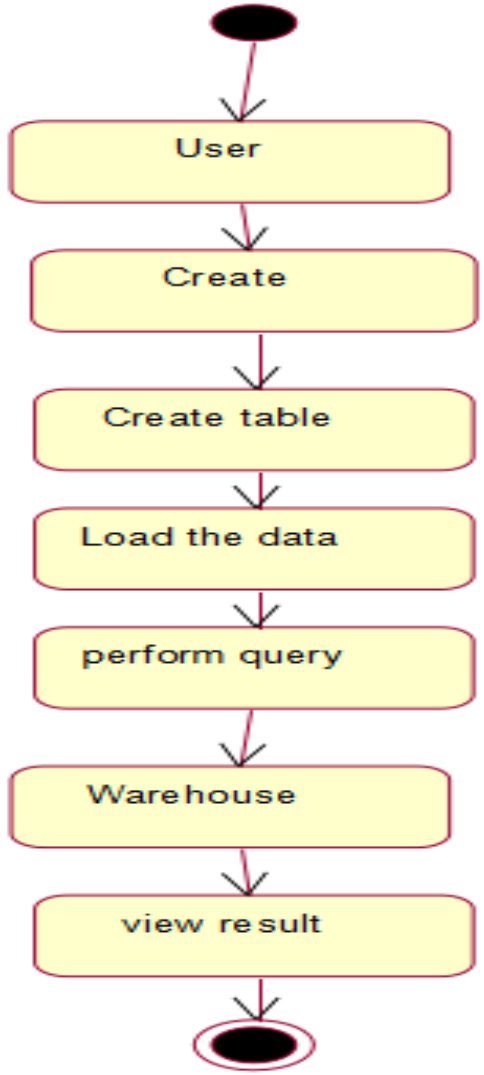


CLASS DIAGRAM

DESIGN

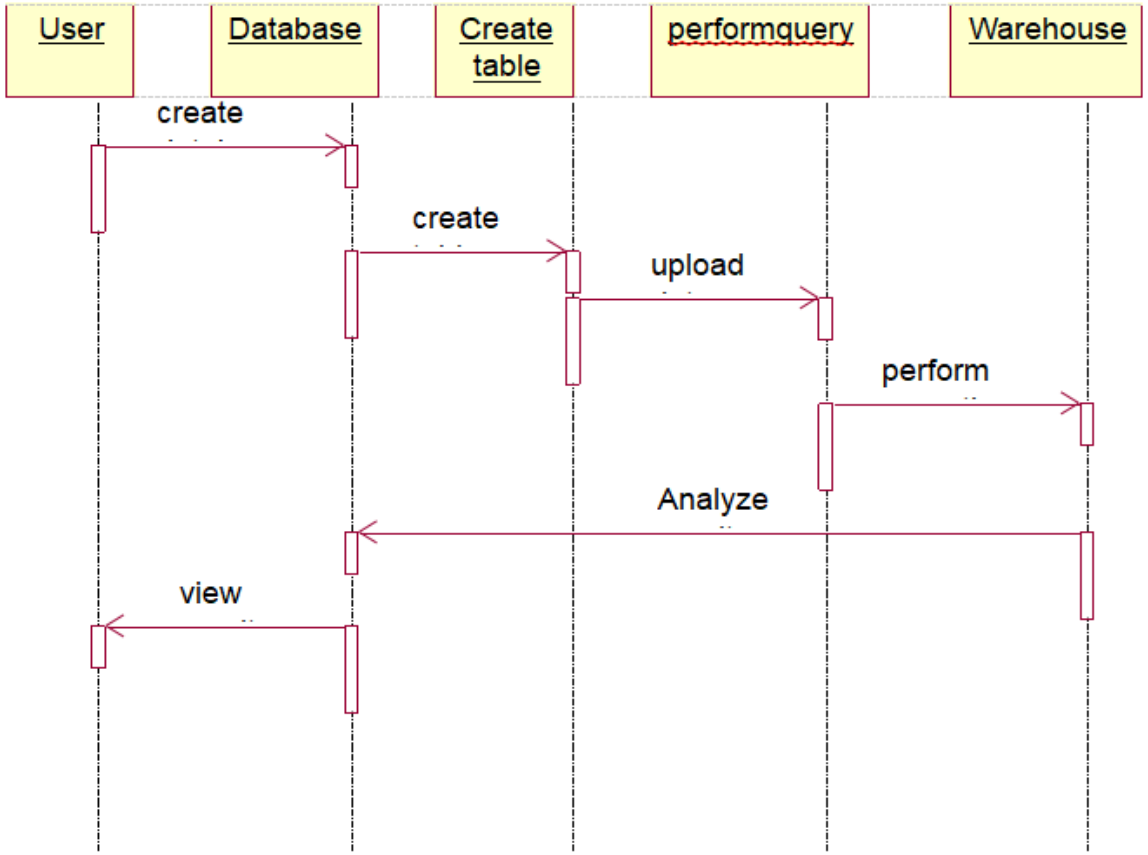


OBJECT DIAGRAM

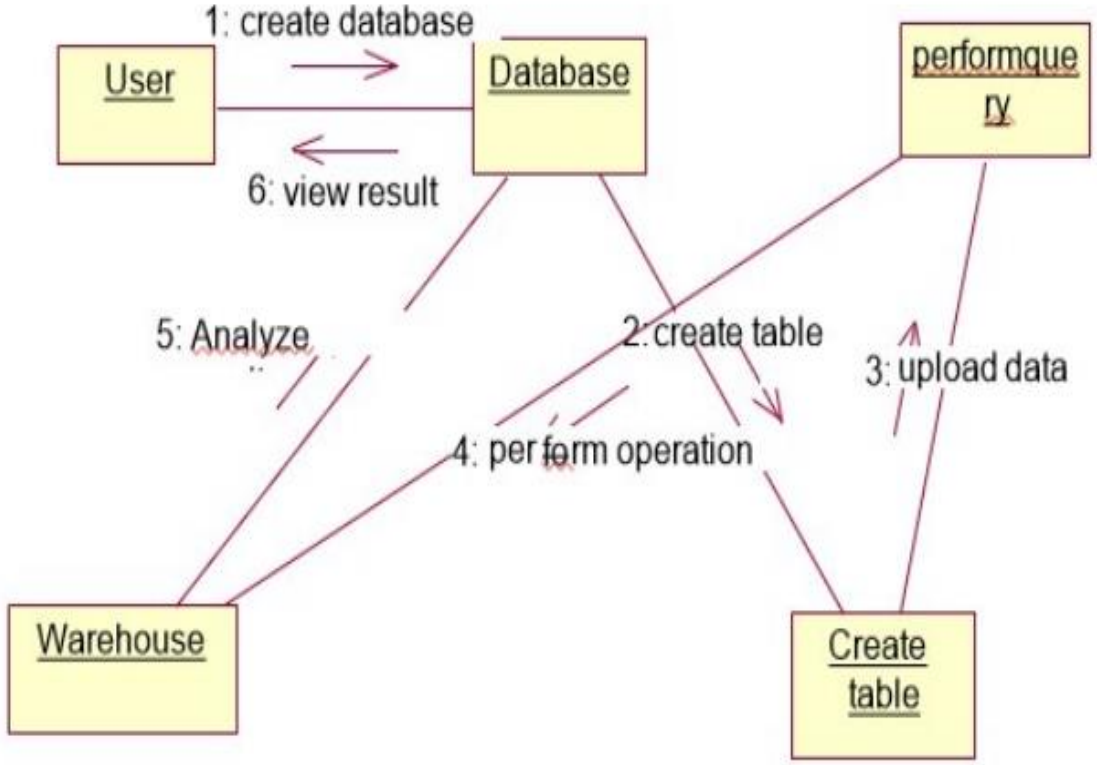


STATE-CHART DIAGRAM

DESIGN

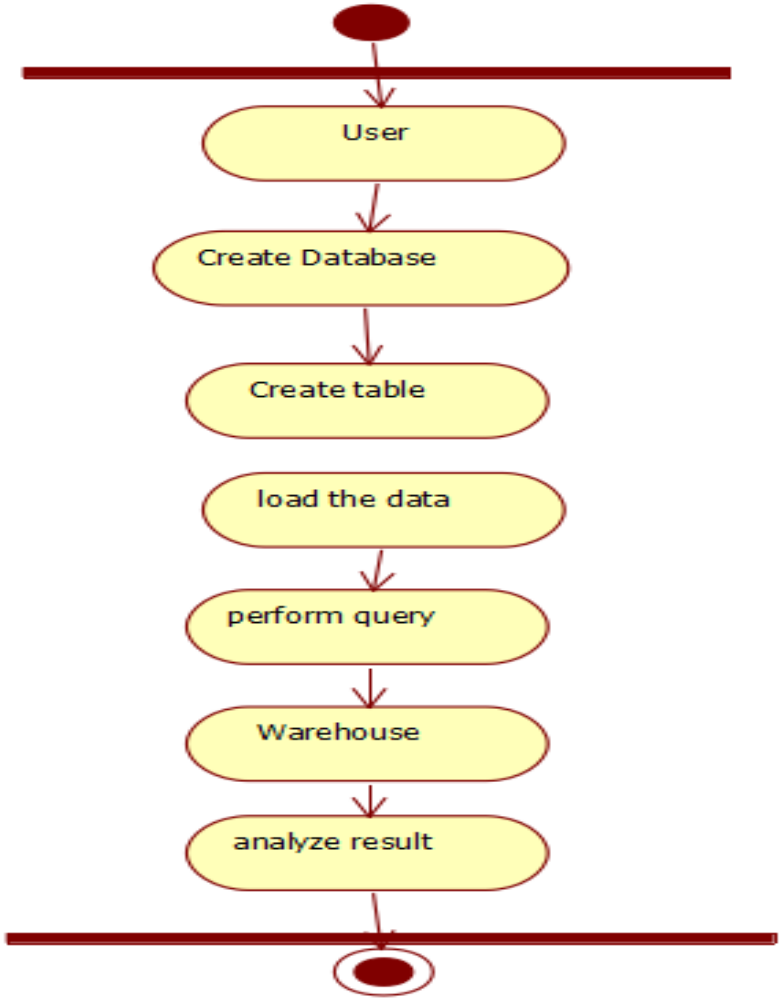


SEQUENCE DIAGRAM

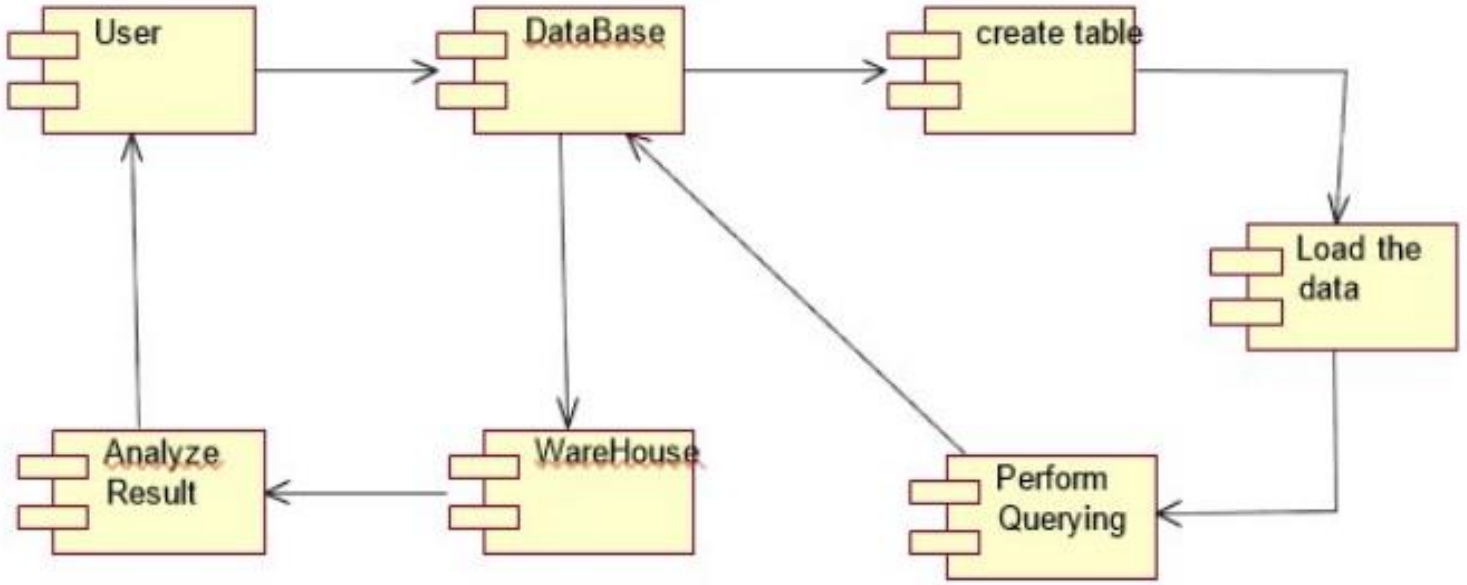


COLLABORATION DIAGRAM

DESIGN

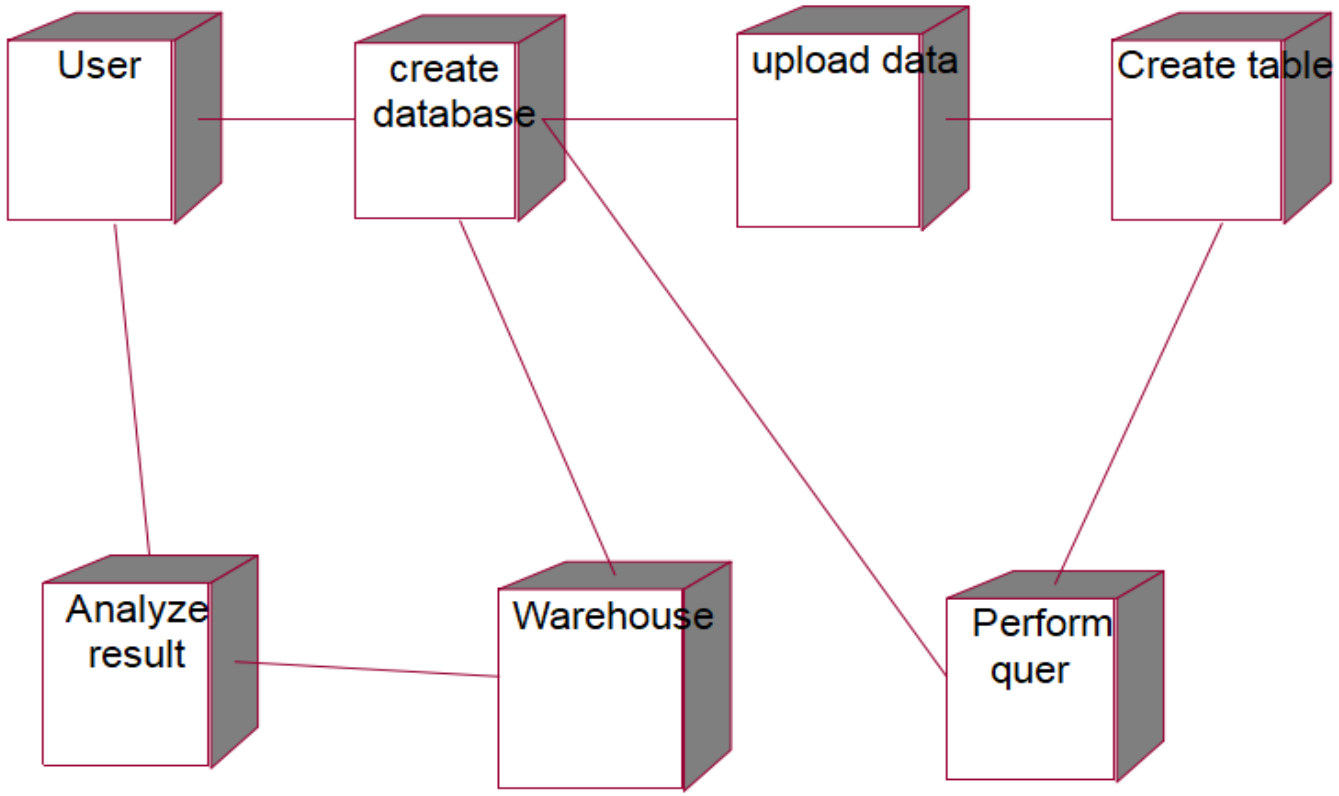


ACTIVITY DIAGRAM



COMPONENT DIAGRAM

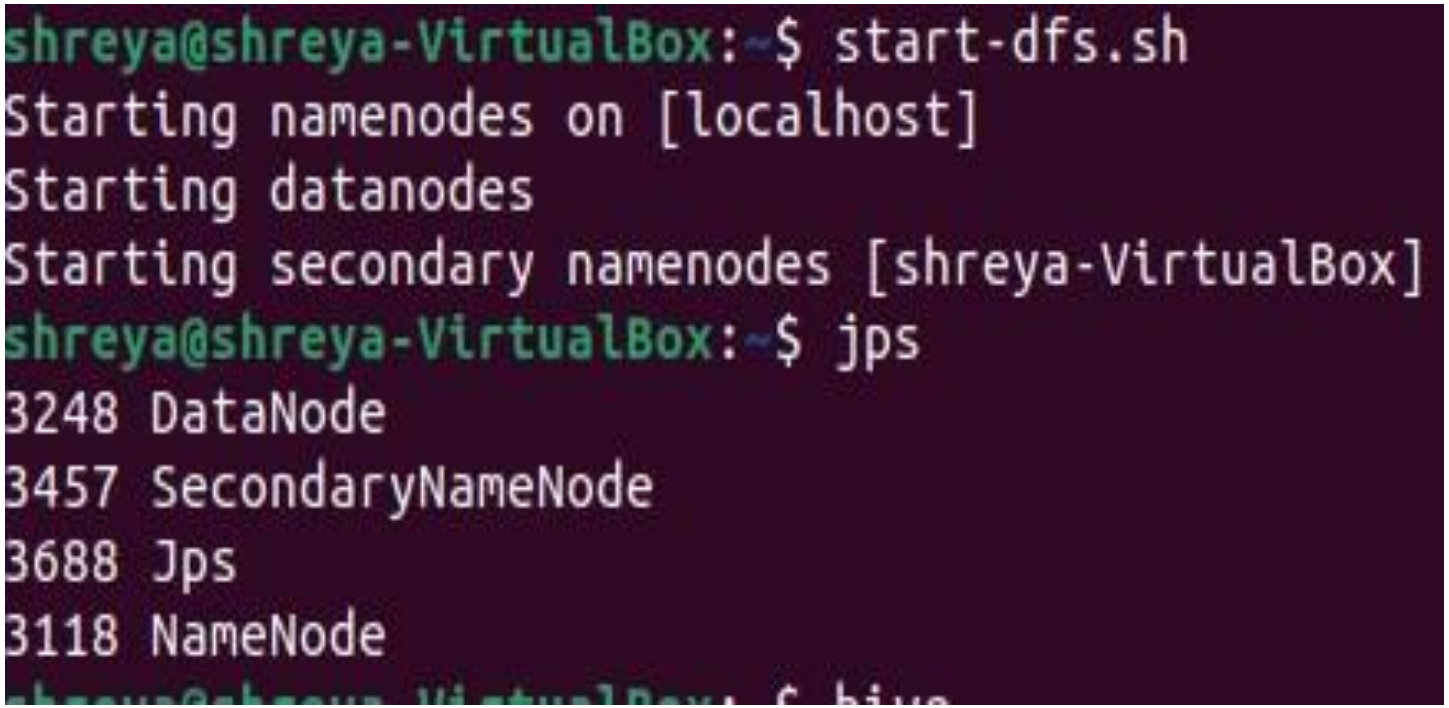
DESIGN



DEPLOYMENT DIAGRAM

IMPLEMENTATION& TESTING

1. Starting Hadoop environment to initialize hive



IMPLEMENTATION& TESTING

2. Starting Hive and creating database

```
shreya@shreya-VirtualBox:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/shreya/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/shreya/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 8393de39-af6b-4863-9090-d7b80b0c99d3

Logging initialized using configuration in jar:file:/home/shreya/hive/lib/hive-common-3.1.3.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Hive Session ID = 532ae79b-03f0-4bd0-8f6e-edbc43a3c695
hive> show databases;
OK
batch11
batch_11
default
Time taken: 0.779 seconds, Fetched: 3 row(s)
hive> create database batch__11;
OK
Time taken: 0.26 seconds
hive> use batch__11;
OK
Time taken: 0.067 seconds
hive>
```

IMPLEMENTATION& TESTING

3. Creating table in database created and loading data

```
hive> create table stocks(stockid int,first_name string,last_name string,stock_name string,location string,purchase_value int, current_value int,current_dt date,trade int,good_to_sale boolean) row format
delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.865 seconds
hive> load data local inpath "/home/shreya/Downloads/stock.unknown" into table stocks;
Loading data to table batch_11.stocks
OK
Time taken: 1.381 seconds
hive> desc stocks;
OK
stockid                 int
first_name              string
last_name               string
stock_name              string
location                string
purchase_value          int
current_value           int
current_dt              date
trade                   int
good_to_sale            boolean
Time taken: 0.126 seconds, Fetched: 10 row(s)
hive>
```

IMPLEMENTATION& TESTING

4. Creating views on the basis of stocks table, for detailed analysis

```
hive> create view good_to_sale as select * from stocks where good_to_sale=true;
OK
Time taken: 0.403 seconds
hive> create view not_good_to_sale as select * from stocks where good_to_sale=false;
OK
Time taken: 1.041 seconds
hive> show views;
OK
good_to_sale
not_good_to_sale
stock_view
Time taken: 0.104 seconds, Fetched: 3 row(s)
hive>
```

IMPLEMENTATION& TESTING

```
hive> select * from good_to_sale;
OK
101     Jonas   Jules   AAPL    Delhi   1500    1800    2023-01-10      1       true
102     John    Smith   MSFT    Mumbai  2000    2200    2023-01-10      1       true
104     Sarah   Davis   TSLA    Hyderabad       2500    3000    2023-05-10      1       true
105     jonas   jules   AAPL    Chennai 1500    1800    2023-01-10      1       true
107     shawn   mendes  CAPGEM  Delhi   3500    1500    2023-03-10      1       true
109     james   smith   AAPL    Delhi   1500    1800    2023-03-10      0       true
111     ronald  clark   TCS     Hyderabad       1500    1800    2023-01-10      1       true
114     michelle        johnson APPL    Delhi   1400    1800    2023-01-11      1       true
116     daniel  clark   AAPL    Mumbai  4500    2800    2023-02-10      1       true
118     nancy   jules   TSLA    Hyderabad       1500    1800    2023-01-10      1       true
119     laura   williams        AAPL    Chennai 1300    1800    2023-01-11      1       true
121     sarah   lee     MSFT    Delhi   1400    1800    2023-02-10      1       true
122     jonas   smith   AAPL    Mumbai  2000    1800    2023-01-10      1       true
124     kevin   hill    AAPL    Chennai 2000    1800    2023-01-10      1       true
126     karen   robert  AAPL    Pune    2000    2800    2023-02-10      1       true
127     linda   jones   MSFT    Chennai 1500    1800    2023-01-10      1       true
128     sarah   jules   AAPL    Delhi   1500    1800    2023-01-10      1       true
131     michael scott   MSFT    Chennai 6500    3800    2023-02-10      1       true
133     white   scott   TCS     Delhi   2000    1800    2023-01-10      1       true
134     anthos  jules   AAPL    Mumbai  3000    3500    2023-01-10      1       true
137     helen   adams   AAPL    Delhi   1500    1800    2023-01-10      1       true
139     george  jules   AAPL    Bangalore       1500    2000    2023-03-10      1       true
142     sandra  davis   AAPL    Delhi   2000    1800    2023-01-10      1       true
144     elsa    gilbert AAPL    Bangalore       2500    1800    2023-01-10      1       true
147     jonas   allen   TCS     Delhi   2500    3800    2023-02-11      1       true
148     sandra  parker  AAPL    Mumbai  2000    1800    2023-01-11      1       true
150     donna   young   TSLA    Hyderabad       2500    3200    2023-02-10      1       true
152     charles gonzalez        AAPL    Delhi   2500    3800    2023-02-11      1       true
154     edward  nelson  AAPL    Bangalore       2000    1800    2023-01-12      1       true
155     carol   jules   AAPL    Hyderabad       1500    1800    2023-01-10      1       true
156     jonas   jules   AAPL    Pune    1500    1800    2023-01-10      1       true
158     ruth    taylor  AAPL    Mumbai  1500    1800    2023-01-10      1       true
159     susan   moore   TCS     Bangalore       2500    2800    2023-01-12      1       true
160     marget  jules   AAPL    Hyderabad       1500    1800    2023-01-10      1       true
163     brian   jules   TCS     Hyderabad       2000    3000    2023-01-12      1       true
164     dorthy  carter  MSFT    Chennai 2000    3500    2023-01-10      1       true
166     jules   chen    AAPL    Mumbai  2000    3000    2023-01-11      1       true
169     sharon  jules   AAPL    Pune    1500    1800    2023-01-10      1       true
171     jenna   stenn   ACCEN   Bangalore       1500    1800    2023-01-10      1       true
173     jonas   jenny   ACCEN   Pune    1500    1800    2023-01-10      1       true
174     stefan  salvatore       AAPL    Delhi   6000    5000    2023-01-10      1       true
180     bonnie  bennete AAPL    Mumbai  1500    1800    2023-01-10      1       true
181     damon   salvtoer        MSFT    Bangalore       1500    1800    2023-01-10      1       true
182     jonas   jules   ACCEN   Hyderabad       2500    1800    2023-01-10      1       true
```

View showcasing profitable stocks data

```
hive> select * from not_good_to_sale;
OK
103     Micheal Williams        GOOGL   Bangalore       3000    2800    2023-02-10      0       false
106     wick    john    ACCEN   Pune    3000    2000    2023-02-07      0       false
108     justin  bieber  TCS     Mumbai  4000    2000    2023-01-08      0       false
110     christ  anderson        GOOGL   Bangalore       4000    2300    2023-01-11      0       false
112     mary    wright  MSFT    Chennai 4500    3000    2023-01-10      0       false
113     lisa    mitchell        TSLA    Pune    2500    1800    2023-03-10      0       false
117     jonas   smith   MSFT    Bangalore       1500    1800    2023-01-10      0       false
120     mark    jules   AAPL    Pune    3500    2800    2023-08-10      0       false
123     robert  jules   CAPGEM  Hyderabad       2000    1800    2023-02-01      0       false
129     jonas   jules   AAPL    Mumbai  1500    1800    2023-01-11      0       false
130     jason   hill    AAPL    Hyderabad       1500    1300    2023-01-10      0       false
132     laura   jules   ACCEN   Pune    5000    3000    2023-02-10      0       false
135     donald  davis   AAPL    Hyderabad       2000    2500    2023-02-11      0       false
136     betty   walker  CAPGEM  Pune    2000    1300    2023-02-11      0       false
138     jeff    marlin  TSLA    Mumbai  1500    1800    2023-01-10      0       false
140     smith   john    AAPL    Hyderabad       2000    2500    2023-01-12      0       false
141     jenny   allen   ACCEN   Pune    1000    1800    2023-01-11      0       false
143     jonas   baker   MSFT    Mumbai  3000    2800    2023-01-11      0       false
145     jonathan        smith   ACCEN   Hyderabad       2000    1800    2023-01-11      0       false
146     richard parker  AAPL    Pune    2500    3000    2023-03-10      0       false
149     maria   jules   AAPL    Bangalore       2000    1800    2023-01-10      0       false
151     jones   denn    AAPL    Pune    1000    2000    2023-01-13      0       false
153     richie  moore   MSFT    Mumbai  2500    2000    2023-01-10      0       false
157     brian   king    TCS     Delhi   1570    1800    2023-01-11      0       false
161     ava     volkov  AAPL    Pune    2000    1800    2023-02-10      0       false
162     bridget larsen  AAPL    Delhi   2000    3000    2023-02-11      0       false
165     stella  alonso  AAPL    Delhi   1800    2800    2023-01-10      0       false
167     thomas  robin   AAPL    Bangalore       1500    1800    2023-01-10      0       false
168     ruth    taylor  evans   Mumbai  3000    2000    2023-02-11      0       false
170     collen  hover   MSFT    Delhi   2500    2800    2023-02-11      0       false
172     marie   jules   AAPL    Hyderabad       2000    3000    2023-02-10      0       false
175     jenn    gilbert AAPL    Mumbai  2000    2500    2023-02-10      0       false
176     sandra  jules   MSFT    Chennai 2000    1800    2023-02-11      0       false
177     sentil  joone   AAPL    Hyderabad       2000    3000    2023-01-10      0       false
178     jenny   harry   ACCEN   Pune    3000    2800    2023-03-02      0       false
179     carl    eleza   AAPL    Delhi   3000    2500    2023-01-10      0       false
184     rhys    larsen  AAPL    Delhi   1500    1800    2023-01-10      0       false
187     christian       harper  MSFT    Chennai 2300    1800    2023-01-10      0       false
190     christ  harper  MSFT    Hyderabad       2500    2800    2023-01-11      0       false
191     jonas   jules   SAMSUNG Delhi   1500    1800    2023-01-10      0       false
194     jonas   sandy   AAPL    Hyderabad       1500    1800    2023-01-10      0       false
195     jonas   jules   AAPL    Pune    2500    3800    2023-01-10      0       false
197     jones   katru   AAPL    Mumbai  2500    2300    2023-01-10      0       false
198     sandra  jules   MSFT    Chennai 3500    4000    2023-03-11      0       false
199     jonas   jules   AAPL    Hyderabad       1500    1800    2023-01-10      0       false
```

View showcasing non-profitable stocks data

RESULTS

- Reduced query response time.
- Enabled clear and detailed analysis of stock transaction data.
- Improved overall efficiency in data processing tasks.
- Improved decision-making.
- Achieved a unified approach in data handling, eliminating fragmentation
- Overcame scalability limitations.

CONCLUSION

- Hive utilizes batch processing for massive data examination
- Offers fault-tolerant system with HDFS for Big Data analysis
- Uses HiveQL, akin to SQL, for communication with large databases
- Project emphasizes optimization via views in Hive for real-time scenarios
- Focuses on stocks data for showcasing low-latency query handling

FUTURE ENHANCEMENT

- Integration of streaming data technologies (Kafka, Flink).
- Exploration of cost-based optimizers.
- Implementation of partitioning and bucketing strategies for data organization/
- Investigation into alternative query engines such as Presto and Trino.
- Continuous pursuit of enhancements to improve the speed, accuracy, and efficiency of real-time profit and loss analysis, fostering data-driven investment decisions.

THANK YOU

ANY QUERIES?