**ANALYSING** **BIG DATA USING HDFC AND HIVE**

**A PROJECT REPORT**

*Submitted by*

**Gurmeet Singh Bhatia**

In

University Of Computing



**Chandigarh University**

10th November, 2022.



**BONAFIDE CERTIFICATE**

Certified that this project report **ANALYSING** **BIG DATA USING HDFC AND HIVE** is the bona fide work of **GURMEET SINGH BHATIA** who carried out the project work under my/our supervision.

SIGNATURE SIGNATURE

**Joe Arun Raja Ponnusamy**

(SUPERVISOR) (HEAD OF THE DEPARTMENT)

*2021-2023*

University of Computing

Masters of Computer Application

Submitted for the project viva-voce examination held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to several individuals and organization for supporting me throughout my Graduate study. First, I wish to express my sincere gratitude to my supervisor **Joe Arun Raja Ponnusamy**, for his enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped me tremendously at all times in my research and writing of this thesis.  His immense knowledge, profound experience and professional expertise in the project topic has enabled me to complete this research successfully. Without his support and guidance, this project would not have been possible. Thanks for all your encouragement! Finally, I also extend my heartiest thanks to my parents, friends, and well-wishers for being with me and extending encouragement throughout the project.

**TABLE OF CONTENTS**

* PROJECT DESCRIPTION
* INPUT FILES
* PROBLEM STATEMENT
* ENVIRONMENT SETUP
* PROJECT MODULES
* PLACING THE GIVEN DATASET IN HDFS
* CREATE DIRECTORY IN HDFS
* PLACING THE INPUT FILES IN THE HDFS DIRECTORY
* IMPLEMENTATION IN HIVE
* CREATING HIVE DB
* CREATING & LOADING THE HIVE TABLES WITH THE GIVEN DATASETS
* PROBLEM SCENARIO 1 - WHAT IS THE TOTAL NUMBER OF CONSUMERS FOR BRANCH1
* TYPE 1: CREATING SINGLE PHYSICAL TABLE WITH SUB QUERIES
* TYPE 2: CREATING MULTIPLE PHYSICAL TABLES
* SOLUTION
* PROBLEM SCENARIO 2 – WHAT IS THE NUMBER OF CONSUMERS FOR THE BRANCH2?
* TYPE 1: SUB QUERIES SELECTION WITHOUT TABLE CREATION
* TYPE 2: CREATING MULTIPLE PHYSICAL TABLES SOLUTION
* PROBLEM SCENARIO 3 - WHAT IS THE MOST CONSUMED BEVERAGE ON BRANCH1
* EXPLANATION
* SOLUTION
* PROBLEM SCENARIO 4 - WHAT ARE THE BEVERAGES AVAILABLE ON BRANCH10, BRANCH8, AND BRANCH1
* EXPLANATION

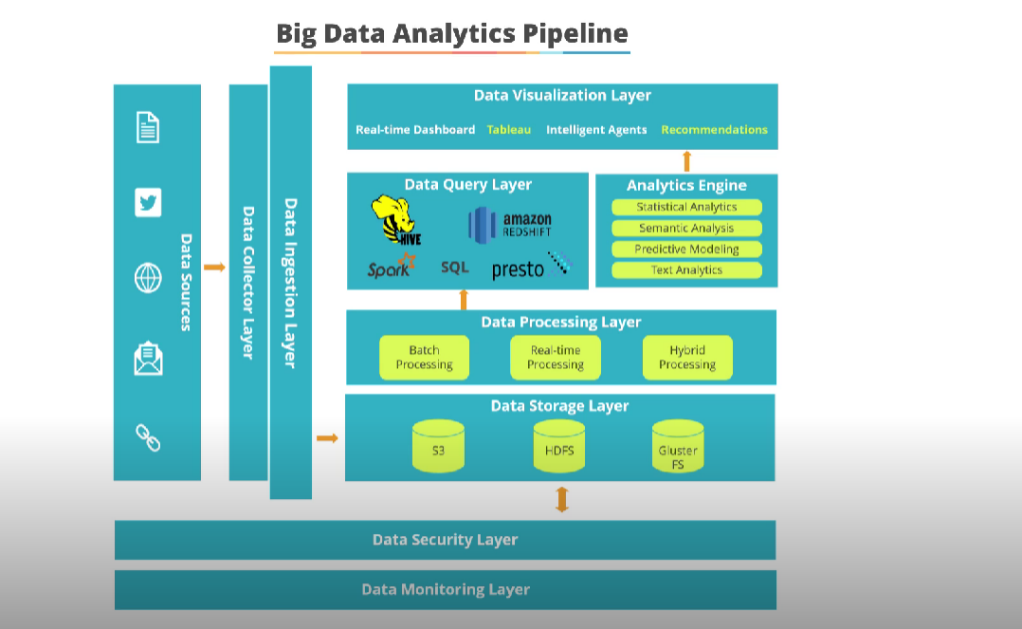
*CHAPTER 1*

**Project Identification**

The project hovers around how large set of data from anywhere around the globe that can be analysed by means of computational techniques to draw patterns and reveal the common or recurring points that would help to predict the next step—especially human behaviour, like future consumer actions based on an analysis of past purchase patterns.

**Task Identification**

The Task involves steps very similar to processing data in the transactional or [data warehouse](https://www.sciencedirect.com/topics/computer-science/data-warehouse) environments.  It shows the different stages involved in the processing of Big Data; the approach to processing Big Data is:



*Fig 1.1*

Heterogeneity

If data is not in natural language or in heterogeneity format

American Journal of Information Science and Computer Engineering Vol. 1, No. 2, 2015, pp. 38-43 39

then it may not provide valuable depth, because machine

algorithms are suited for homogeneous data. So for data

analysis data must be structured carefully [1]. Data can be

structured by using Map Reduce techniques, because all keys

generated by the Map Reduce must fit into main memory [10].

2.2. Inconsistency and Incompleteness

Since big data comprises of lot of information coming from

different sources (each source has different nature of

reliability), so error and missing values is a challenge for

managing them. These errors and incompleteness must be

managed in data analysis [1]. However, there is also

challenges include with data analysis such as efficient

representation, access, and analysis of unstructured or semi-

structured data in the further researches. To remove noise and

correct inconsistencies, different types of data preprocessing

techniques can be applied such as data cleaning, data

integration, data transformation and date reduction [6].

2.3. Scale

Size of big data is an important challenge, although there are

many researches to handle this issue such as handle big data

with processor speed but some time increasing volume of data

is faster than processor speed [1][6]. In Big Data applications,

the state-of-the-art techniques and technologies cannot ideally

solve the real problems, especially for real-time analysis. As

Big Data requires a more storage and medium, if Hard Disk

Drives (HDDs) are used for such purpose, then HDDs is

slower than data processing engines, this challenges can be

handled by using Solid State Drives (SSDs) and Pulse-code

modulation (PCM) technologies [6]. A sheer volume of data

requires very high speed [4] because data grows as the degree

of granularity increases. To store such data there is also

memory issues. Faster growth of data and memory issues can

be solved by using grid computing approach. Big Data is

comprised of large no of inputs, outputs and attributes, these

are lead to the complexity related to running time [9], and to

handle such types of issue distributed frameworks with

parallelized machines are preferred.

2.4. Velocity of Data

In Bigdata, data is generated with very high speed, and this

speedy data requires processing in timely manner. This

problem in learning algorithm can be solved by using online

learning approaches [9].

2.5. Timeliness

How data without outlier can be filtered at real time for

storage purpose [1]. This issue can be handled through Index

structure of traffic management system. If data is not

analyzed quickly and there is not proper framework for users

then data for decision making will not be fruitful. To address

data quality, there are some proactive methods that can

address data quality and timely [4].

2.6. Privacy and Data Ownership

For data analysis, data from all relevant side is requiring, but

in some situations, where, there are strict laws governing

what data can be revealed in different contexts. Another issue

is that many online services today require us to share private

information (think of Facebook applications), but beyond

record-level access control we do not understand what it

means to share data, how the shared data can be linked, and

how to give users fine-grained control over this sharing in an

intuitive [1], so proper preventive measures are taken to

protect such sensitive data, to ensure its safety [7]. Keeping

track of a particular individual’s data throughout big data

analytics contexts is merely an organizational requirement

that can e.g. be met by means of log files. Linkage of disjoint

datasets can often be performed without relying on linkage

via user’s identities, but based on other types of data. In the

same direction, many types of data can be preprocessed with

proper anonymization or pseudonymization prior to sharing,

such that linkage of datasets remains feasible, but linkage to

an individual’s identity becomes hard [8].

2.7. The Human Perspective

For data analysis, there should be considerable volume of

data that must be understandable for a human [1].

Understanding right data is an important issue that can be

handled by using visualization as a part of data analysis [4].

Companies have difficulty identifying the right data and

determining how to best use it. Building data-related business

cases can be used for identifying useful data [5].

2.8. Visualization

Due to large and high dimension of Big Data, visualization of

data is very difficult. There are some tools for visualization

but mostly have poor performance and response time. New

framework for visualization is highly necessary [6].

There is no communication path between data points, so

companies cannot aggregate and manage the data across the

enterprise. Recently [5] using smart grid such as real-time

grid management has provided such solution.

2.9. Displaying Meaningful Results

As in Big Data, there are extremely large amounts of

information or a variety of categories of information so

representation of such information is a big challenge. This

issue can be resolved by making data cluster into a higher-

level view where smaller groups of data become visible [4].

Heterogeneity

If data is not in natural language or in heterogeneity format

American Journal of Information Science and Computer Engineering Vol. 1, No. 2, 2015, pp. 38-43 39

then it may not provide valuable depth, because machine

algorithms are suited for homogeneous data. So for data

analysis data must be structured carefully [1]. Data can be

structured by using Map Reduce techniques, because all keys

generated by the Map Reduce must fit into main memory [10].

2.2. Inconsistency and Incompleteness

Since big data comprises of lot of information coming from

different sources (each source has different nature of

reliability), so error and missing values is a challenge for

managing them. These errors and incompleteness must be

managed in data analysis [1]. However, there is also

challenges include with data analysis such as efficient

representation, access, and analysis of unstructured or semi-

structured data in the further researches. To remove noise and

correct inconsistencies, different types of data preprocessing

techniques can be applied such as data cleaning, data

integration, data transformation and date reduction [6].

2.3. Scale

Size of big data is an important challenge, although there are

many researches to handle this issue such as handle big data

with processor speed but some time increasing volume of data

is faster than processor speed [1][6]. In Big Data applications,

the state-of-the-art techniques and technologies cannot ideally

solve the real problems, especially for real-time analysis. As

Big Data requires a more storage and medium, if Hard Disk

Drives (HDDs) are used for such purpose, then HDDs is

slower than data processing engines, this challenges can be

handled by using Solid State Drives (SSDs) and Pulse-code

modulation (PCM) technologies [6]. A sheer volume of data

requires very high speed [4] because data grows as the degree

of granularity increases. To store such data there is also

memory issues. Faster growth of data and memory issues can

be solved by using grid computing approach. Big Data is

comprised of large no of inputs, outputs and attributes, these

are lead to the complexity related to running time [9], and to

handle such types of issue distributed frameworks with

parallelized machines are preferred.

2.4. Velocity of Data

In Bigdata, data is generated with very high speed, and this

speedy data requires processing in timely manner. This

problem in learning algorithm can be solved by using online

learning approaches [9].

2.5. Timeliness

How data without outlier can be filtered at real time for

storage purpose [1]. This issue can be handled through Index

structure of traffic management system. If data is not

analyzed quickly and there is not proper framework for users

then data for decision making will not be fruitful. To address

data quality, there are some proactive methods that can

address data quality and timely [4].

2.6. Privacy and Data Ownership

For data analysis, data from all relevant side is requiring, but

in some situations, where, there are strict laws governing

what data can be revealed in different contexts. Another issue

is that many online services today require us to share private

information (think of Facebook applications), but beyond

record-level access control we do not understand what it

means to share data, how the shared data can be linked, and

how to give users fine-grained control over this sharing in an

intuitive [1], so proper preventive measures are taken to

protect such sensitive data, to ensure its safety [7]. Keeping

track of a particular individual’s data throughout big data

analytics contexts is merely an organizational requirement

that can e.g. be met by means of log files. Linkage of disjoint

datasets can often be performed without relying on linkage

via user’s identities, but based on other types of data. In the

same direction, many types of data can be preprocessed with

proper anonymization or pseudonymization prior to sharing,

such that linkage of datasets remains feasible, but linkage to

an individual’s identity becomes hard [8].

2.7. The Human Perspective

For data analysis, there should be considerable volume of

data that must be understandable for a human [1].

Understanding right data is an important issue that can be

handled by using visualization as a part of data analysis [4].

Companies have difficulty identifying the right data and

determining how to best use it. Building data-related business

cases can be used for identifying useful data [5].

2.8. Visualization

Due to large and high dimension of Big Data, visualization of

data is very difficult. There are some tools for visualization

but mostly have poor performance and response time. New

framework for visualization is highly necessary [6].

There is no communication path between data points, so

companies cannot aggregate and manage the data across the

enterprise. Recently [5] using smart grid such as real-time

grid management has provided such solution.

2.9. Displaying Meaningful Results

As in Big Data, there are extremely large amounts of

information or a variety of categories of information so

representation of such information is a big challenge. This

issue can be resolved by making data cluster into a higher-

level view where smaller groups of data become visible [4].

Heterogeneity

If data is not in natural language or in heterogeneity format

American Journal of Information Science and Computer Engineering Vol. 1, No. 2, 2015, pp. 38-43 39

then it may not provide valuable depth, because machine

algorithms are suited for homogeneous data. So for data

analysis data must be structured carefully [1]. Data can be

structured by using Map Reduce techniques, because all keys

generated by the Map Reduce must fit into main memory [10].

2.2. Inconsistency and Incompleteness

Since big data comprises of lot of information coming from

different sources (each source has different nature of

reliability), so error and missing values is a challenge for

managing them. These errors and incompleteness must be

managed in data analysis [1]. However, there is also

challenges include with data analysis such as efficient

representation, access, and analysis of unstructured or semi-

structured data in the further researches. To remove noise and

correct inconsistencies, different types of data preprocessing

techniques can be applied such as data cleaning, data

integration, data transformation and date reduction [6].

2.3. Scale

Size of big data is an important challenge, although there are

many researches to handle this issue such as handle big data

with processor speed but some time increasing volume of data

is faster than processor speed [1][6]. In Big Data applications,

the state-of-the-art techniques and technologies cannot ideally

solve the real problems, especially for real-time analysis. As

Big Data requires a more storage and medium, if Hard Disk

Drives (HDDs) are used for such purpose, then HDDs is

slower than data processing engines, this challenges can be

handled by using Solid State Drives (SSDs) and Pulse-code

modulation (PCM) technologies [6]. A sheer volume of data

requires very high speed [4] because data grows as the degree

of granularity increases. To store such data there is also

memory issues. Faster growth of data and memory issues can

be solved by using grid computing approach. Big Data is

comprised of large no of inputs, outputs and attributes, these

are lead to the complexity related to running time [9], and to

handle such types of issue distributed frameworks with

parallelized machines are preferred.

2.4. Velocity of Data

In Bigdata, data is generated with very high speed, and this

speedy data requires processing in timely manner. This

problem in learning algorithm can be solved by using online

learning approaches [9].

2.5. Timeliness

How data without outlier can be filtered at real time for

storage purpose [1]. This issue can be handled through Index

structure of traffic management system. If data is not

analyzed quickly and there is not proper framework for users

then data for decision making will not be fruitful. To address

data quality, there are some proactive methods that can

address data quality and timely [4].

2.6. Privacy and Data Ownership

For data analysis, data from all relevant side is requiring, but

in some situations, where, there are strict laws governing

what data can be revealed in different contexts. Another issue

is that many online services today require us to share private

information (think of Facebook applications), but beyond

record-level access control we do not understand what it

means to share data, how the shared data can be linked, and

how to give users fine-grained control over this sharing in an

intuitive [1], so proper preventive measures are taken to

protect such sensitive data, to ensure its safety [7]. Keeping

track of a particular individual’s data throughout big data

analytics contexts is merely an organizational requirement

that can e.g. be met by means of log files. Linkage of disjoint

datasets can often be performed without relying on linkage

via user’s identities, but based on other types of data. In the

same direction, many types of data can be preprocessed with

proper anonymization or pseudonymization prior to sharing,

such that linkage of datasets remains feasible, but linkage to

an individual’s identity becomes hard [8].

2.7. The Human Perspective

For data analysis, there should be considerable volume of

data that must be understandable for a human [1].

Understanding right data is an important issue that can be

handled by using visualization as a part of data analysis [4].

Companies have difficulty identifying the right data and

determining how to best use it. Building data-related business

cases can be used for identifying useful data [5].

2.8. Visualization

Due to large and high dimension of Big Data, visualization of

data is very difficult. There are some tools for visualization

but mostly have poor performance and response time. New

framework for visualization is highly necessary [6].

There is no communication path between data points, so

companies cannot aggregate and manage the data across the

enterprise. Recently [5] using smart grid such as real-time

grid management has provided such solution.

2.9. Displaying Meaningful Results

As in Big Data, there are extremely large amounts of

information or a variety of categories of information so

representation of such information is a big challenge. This

issue can be resolved by making data cluster into a higher-

level view where smaller groups of data become visible [4].

Heterogeneity

If data is not in natural language or in heterogeneity format

American Journal of Information Science and Computer Engineering Vol. 1, No. 2, 2015, pp. 38-43 39

then it may not provide valuable depth, because machine

algorithms are suited for homogeneous data. So for data

analysis data must be structured carefully [1]. Data can be

structured by using Map Reduce techniques, because all keys

generated by the Map Reduce must fit into main memory [10].

2.2. Inconsistency and Incompleteness

Since big data comprises of lot of information coming from

different sources (each source has different nature of

reliability), so error and missing values is a challenge for

managing them. These errors and incompleteness must be

managed in data analysis [1]. However, there is also

challenges include with data analysis such as efficient

representation, access, and analysis of unstructured or semi-

structured data in the further researches. To remove noise and

correct inconsistencies, different types of data preprocessing

techniques can be applied such as data cleaning, data

integration, data transformation and date reduction [6].

2.3. Scale

Size of big data is an important challenge, although there are

many researches to handle this issue such as handle big data

with processor speed but some time increasing volume of data

is faster than processor speed [1][6]. In Big Data applications,

the state-of-the-art techniques and technologies cannot ideally

solve the real problems, especially for real-time analysis. As

Big Data requires a more storage and medium, if Hard Disk

Drives (HDDs) are used for such purpose, then HDDs is

slower than data processing engines, this challenges can be

handled by using Solid State Drives (SSDs) and Pulse-code

modulation (PCM) technologies [6]. A sheer volume of data

requires very high speed [4] because data grows as the degree

of granularity increases. To store such data there is also

memory issues. Faster growth of data and memory issues can

be solved by using grid computing approach. Big Data is

comprised of large no of inputs, outputs and attributes, these

are lead to the complexity related to running time [9], and to

handle such types of issue distributed frameworks with

parallelized machines are preferred.

2.4. Velocity of Data

In Bigdata, data is generated with very high speed, and this

speedy data requires processing in timely manner. This

problem in learning algorithm can be solved by using online

learning approaches [9].

2.5. Timeliness

How data without outlier can be filtered at real time for

storage purpose [1]. This issue can be handled through Index

structure of traffic management system. If data is not

analyzed quickly and there is not proper framework for users

then data for decision making will not be fruitful. To address

data quality, there are some proactive methods that can

address data quality and timely [4].

2.6. Privacy and Data Ownership

For data analysis, data from all relevant side is requiring, but

in some situations, where, there are strict laws governing

what data can be revealed in different contexts. Another issue

is that many online services today require us to share private

information (think of Facebook applications), but beyond

record-level access control we do not understand what it

means to share data, how the shared data can be linked, and

how to give users fine-grained control over this sharing in an

intuitive [1], so proper preventive measures are taken to

protect such sensitive data, to ensure its safety [7]. Keeping

track of a particular individual’s data throughout big data

analytics contexts is merely an organizational requirement

that can e.g. be met by means of log files. Linkage of disjoint

datasets can often be performed without relying on linkage

via user’s identities, but based on other types of data. In the

same direction, many types of data can be preprocessed with

proper anonymization or pseudonymization prior to sharing,

such that linkage of datasets remains feasible, but linkage to

an individual’s identity becomes hard [8].

2.7. The Human Perspective

For data analysis, there should be considerable volume of

data that must be understandable for a human [1].

Understanding right data is an important issue that can be

handled by using visualization as a part of data analysis [4].

Companies have difficulty identifying the right data and

determining how to best use it. Building data-related business

cases can be used for identifying useful data [5].

2.8. Visualization

Due to large and high dimension of Big Data, visualization of

data is very difficult. There are some tools for visualization

but mostly have poor performance and response time. New

framework for visualization is highly necessary [6].

There is no communication path between data points, so

companies cannot aggregate and manage the data across the

enterprise. Recently [5] using smart grid such as real-time

grid management has provided such solution.

2.9. Displaying Meaningful Results

As in Big Data, there are extremely large amounts of

information or a variety of categories of information so

representation of such information is a big challenge. This

issue can be resolved by making data cluster into a higher-

level view where smaller groups of data become visible [4].

Heterogeneity

If data is not in natural language or in heterogeneity format

American Journal of Information Science and Computer Engineering Vol. 1, No. 2, 2015, pp. 38-43 39

then it may not provide valuable depth, because machine

algorithms are suited for homogeneous data. So for data

analysis data must be structured carefully [1]. Data can be

structured by using Map Reduce techniques, because all keys

generated by the Map Reduce must fit into main memory [10].

2.2. Inconsistency and Incompleteness

Since big data comprises of lot of information coming from

different sources (each source has different nature of

reliability), so error and missing values is a challenge for

managing them. These errors and incompleteness must be

managed in data analysis [1]. However, there is also

challenges include with data analysis such as efficient

representation, access, and analysis of unstructured or semi-

structured data in the further researches. To remove noise and

correct inconsistencies, different types of data preprocessing

techniques can be applied such as data cleaning, data

integration, data transformation and date reduction [6].

2.3. Scale

Size of big data is an important challenge, although there are

many researches to handle this issue such as handle big data

with processor speed but some time increasing volume of data

is faster than processor speed [1][6]. In Big Data applications,

the state-of-the-art techniques and technologies cannot ideally

solve the real problems, especially for real-time analysis. As

Big Data requires a more storage and medium, if Hard Disk

Drives (HDDs) are used for such purpose, then HDDs is

slower than data processing engines, this challenges can be

handled by using Solid State Drives (SSDs) and Pulse-code

modulation (PCM) technologies [6]. A sheer volume of data

requires very high speed [4] because data grows as the degree

of granularity increases. To store such data there is also

memory issues. Faster growth of data and memory issues can

be solved by using grid computing approach. Big Data is

comprised of large no of inputs, outputs and attributes, these

are lead to the complexity related to running time [9], and to

handle such types of issue distributed frameworks with

parallelized machines are preferred.

2.4. Velocity of Data

In Bigdata, data is generated with very high speed, and this

speedy data requires processing in timely manner. This

problem in learning algorithm can be solved by using online

learning approaches [9].

2.5. Timeliness

How data without outlier can be filtered at real time for

storage purpose [1]. This issue can be handled through Index

structure of traffic management system. If data is not

analyzed quickly and there is not proper framework for users

then data for decision making will not be fruitful. To address

data quality, there are some proactive methods that can

address data quality and timely [4].

2.6. Privacy and Data Ownership

For data analysis, data from all relevant side is requiring, but

in some situations, where, there are strict laws governing

what data can be revealed in different contexts. Another issue

is that many online services today require us to share private

information (think of Facebook applications), but beyond

record-level access control we do not understand what it

means to share data, how the shared data can be linked, and

how to give users fine-grained control over this sharing in an

intuitive [1], so proper preventive measures are taken to

protect such sensitive data, to ensure its safety [7]. Keeping

track of a particular individual’s data throughout big data

analytics contexts is merely an organizational requirement

that can e.g. be met by means of log files. Linkage of disjoint

datasets can often be performed without relying on linkage

via user’s identities, but based on other types of data. In the

same direction, many types of data can be preprocessed with

proper anonymization or pseudonymization prior to sharing,

such that linkage of datasets remains feasible, but linkage to

an individual’s identity becomes hard [8].

2.7. The Human Perspective

For data analysis, there should be considerable volume of

data that must be understandable for a human [1].

Understanding right data is an important issue that can be

handled by using visualization as a part of data analysis [4].

Companies have difficulty identifying the right data and

determining how to best use it. Building data-related business

cases can be used for identifying useful data [5].

2.8. Visualization

Due to large and high dimension of Big Data, visualization of

data is very difficult. There are some tools for visualization

but mostly have poor performance and response time. New

framework for visualization is highly necessary [6].

There is no communication path between data points, so

companies cannot aggregate and manage the data across the

enterprise. Recently [5] using smart grid such as real-time

grid management has provided such solution.

2.9. Displaying Meaningful Results

As in Big Data, there are extremely large amounts of

information or a variety of categories of information so

representation of such information is a big challenge. This

issue can be resolved by making data cluster into a higher-

level view where smaller groups of data become visible [4].

**Project Description**

The project aims to display Beverages-to-Multiple Branches of one Coffee Shop (many-tomany relation) using Hive. In other words, each Beverage might be available on many Branches, and each Branch of the Coffee shop might distribute many Beverages. Assuming each branch send their sales report as a csv file. The project aims to stage them to HDFS and further analysis to be performed using Hive for the given problem statement below.

The description of the data is as below:

• Beverages Name does not have spaces.

• Coffee shop Branches have mentioned as Branch1, Branch2 and etc.

• Beverages can be ordered many times, with different counts

• A Beverage and Branch combination might appear multiple times

• Beverages could be available on multiple Branches

• The output should have no commas or punctuation, only 1 space between the Beverages and Count of Consumed people.

**Goals and Objectives.**

One of the key objectives of this research work is to define how HDFS, Hive and MapReduce help address the challenges of big data. The HDFS component will describe how big data are stored and MapReduce will define how big data are processed. With this main objective, the research study will also have other objectives which are as follows:

1) To address the challenges of big data using Hadoop.

2) Using Hadoop can easily store and process big data.

3) HDFS and MapReduce help reduce the challenges of storing and processing of big data.

**INPUT FILES**



**PROBLEM STATEMENT**

• What is the total number of consumers for Branch1?

• What is the number of consumers for the Branch2?

• What is the most consumed beverage on Branch1?

• What are the beverages available on Branch10, Branch8, and Branch1?

**ENVIRONMENT SETUP**

Software Specification

VM Used – Oracle Virtual Machine.

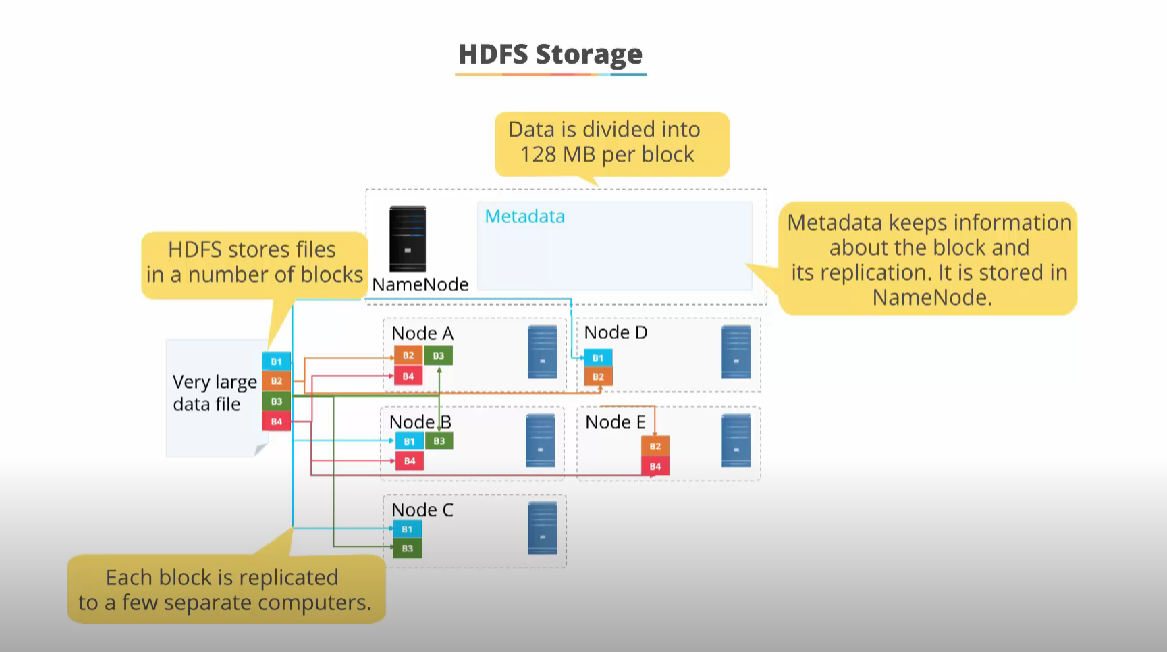
Hadoop version – Hadoop **2.10** line

Hive version – Hive version **3.1.3**

WinSCP – version 5.9.4

*CHAPTER 2 -*

**PLACING THE GIVEN DATASET IN HDFS CREATE DIRECTORY IN HDFS**



*Fig 1.2*

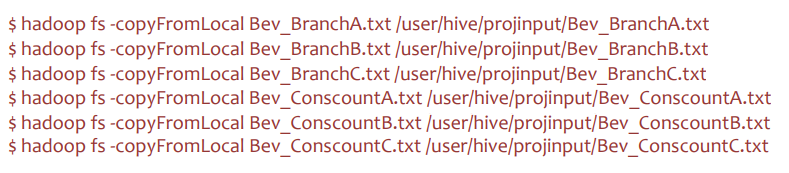
Step 1: Before creating directories in HDFS, ensure all the daemons in hadoop are started. The below code is for creating directory called “projinput” as follows,



**Placing the input files in the HDFS directory**

Step 1: Copying all the given dataset files from local to HDFS directory in a separate directory.

The code as follows:



Step 2: After the Step 1, check whether the files got placed in the hdfs in browser. **Implementation In Hive Creating Hive Db**

Hive> create database if not exits hadoophiveproj with

dbproperties(‘creator’=’Gurmeet Singh Bhatia’,’date’=’10/11/2022’);

**Creating & Loading the HIVE tables with the given datasets**

Step 1: Create separate raw tables for the Beverages-Counsumercount different datasets each in “hadoophiveproj” database.

The given file (Bev\_Conscount \*.txt) consist of (A Beverage and the number of consumers). Example Bev\_Conscount\*.txt:

Special\_Lite, 21

Triple\_Espresso,

38 Mild\_LATTE,

73 LARGE\_Coffee,

144 Cold\_cappuccino,

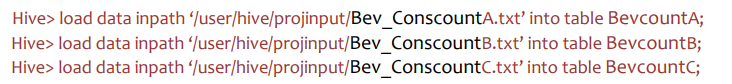
287 SMALL\_cappuccino,

574 …

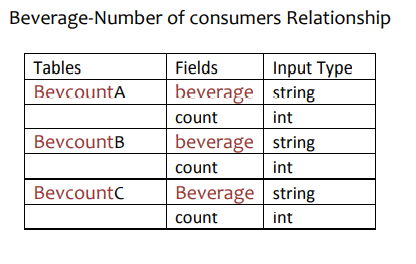
The codes for creating tables are as follows:



**Step 2**: Loading the Beverage -Number of consumers’ raw tables from the given text files individually. The code as follows:

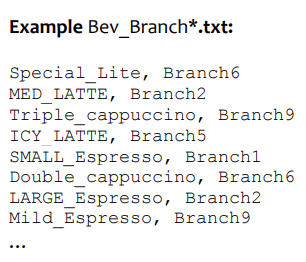


**Schema definition for the created tables as follows:**

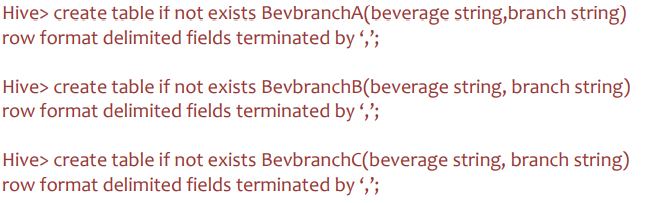


**Step 3:** Create separate raw tables for the Beverages-Branches different datasets each in “hadoophiveproj” database.

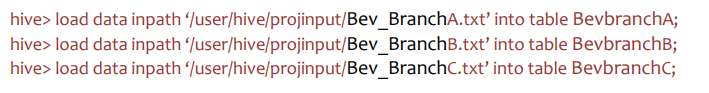
The given file (Bev\_Branch\*.txt) consist of (A Beverages and the Branches it was on)



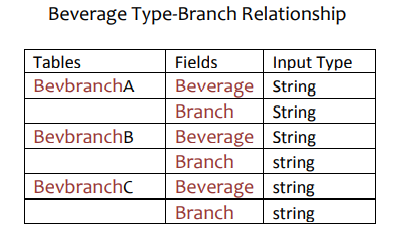
The codes for creating tables are as follows:



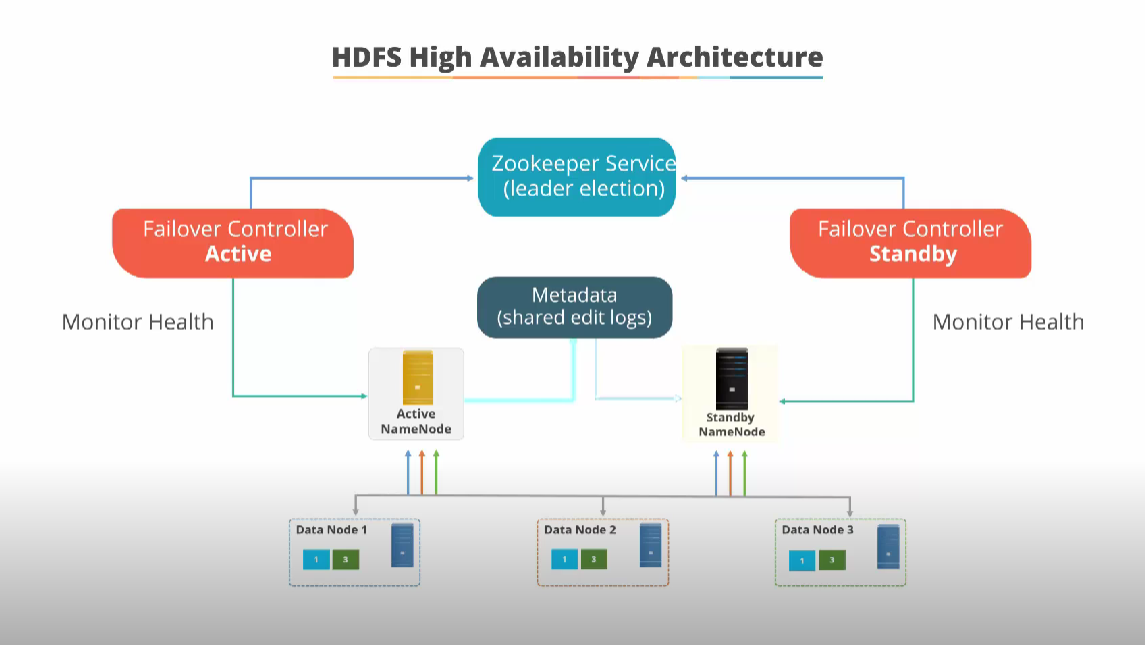
**Step 4**: Loading the Beverage type-Branch raw tables from the given text files individually. The code as follows:



Schema definition for the created tables as follows:



**Project Architecture:**



*Fig 1.3*

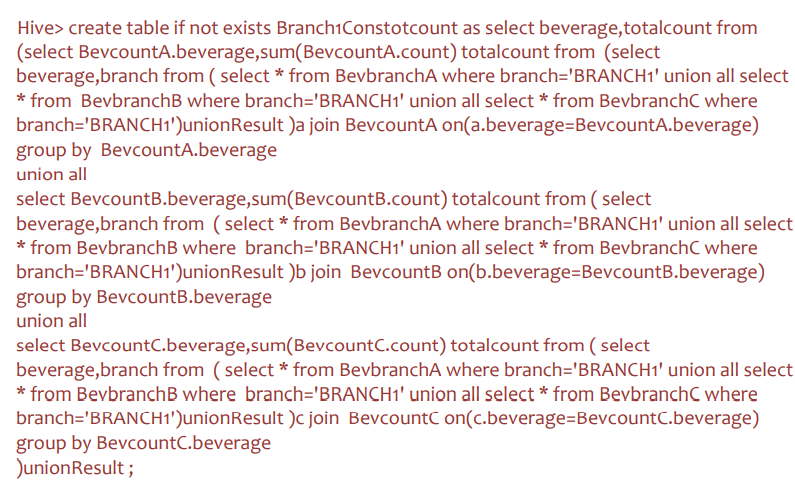
**Problem Scenario 1** - What is the total number of consumers for Branch1?

Type 1: Creating single physical table with sub queries

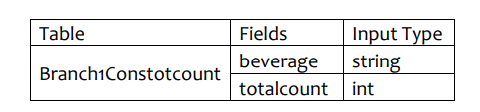
Explanation

Step 1: Creating single table “Branch1Constotcount” with sub queries in which the Branch1 consumers alone selected and counted.

The code as follows:



**Step 2:** The created table have the following table structure.



**Step 3**: Finally, the summation of the table gives the end result.

Code as follows:

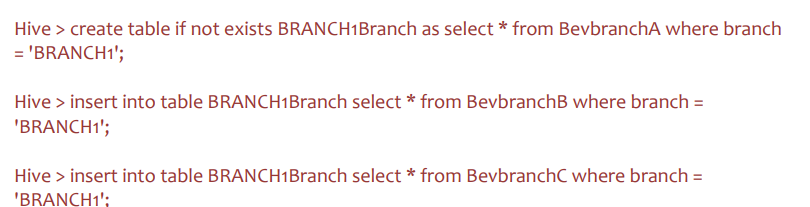


**Type 2:** Creating multiple physical tables

Explanation

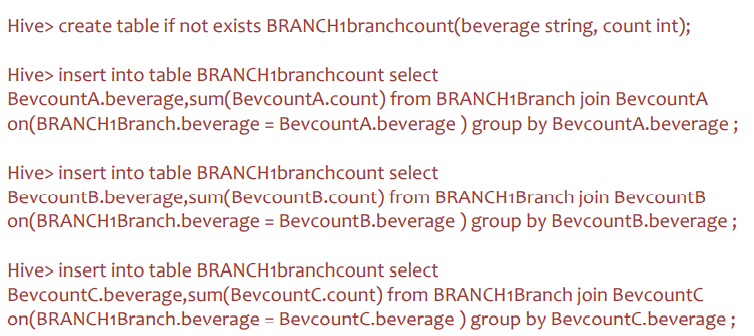
Step 1: Creating physical table “BRANCH1Branch” from the previous tables where the branch name is equal to BRANCH1.

The code as follows:



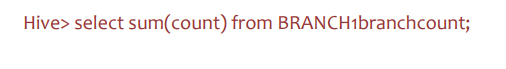
**Step 2**: Creating another table “BRANCH1branchcount” which merge tables which has the number of consumers with Beverages.

The code as follows:

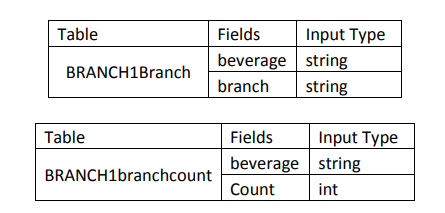


**Step 3**: Finally, summation of the fields from total number of consumers in the newly created table “BRANCH1branchcount”.

The code as follows:



**Step 4**: Created tables schema definition is as follows:



**Solution**

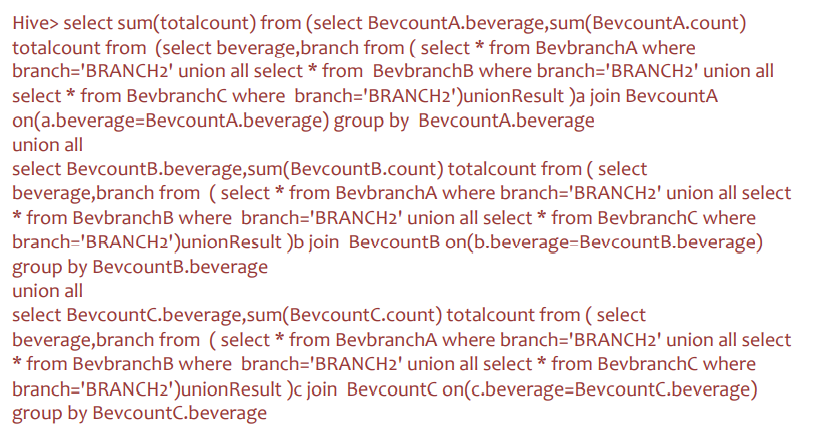
1115974

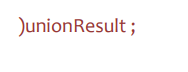
**Problem Scenario 2** – What is the number of consumers for the Branch2?

Type 1: Sub queries selection without table creation

Explanation

**Step 1**: Selecting data from previous tables with sub queries in which the branch BRANCH2 consumers alone selected and counted. The code as follows:

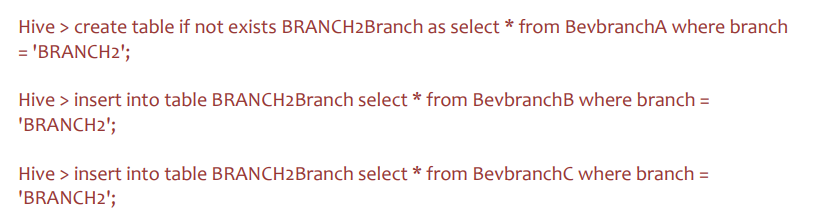




**Type 2:** Creating multiple physical tables

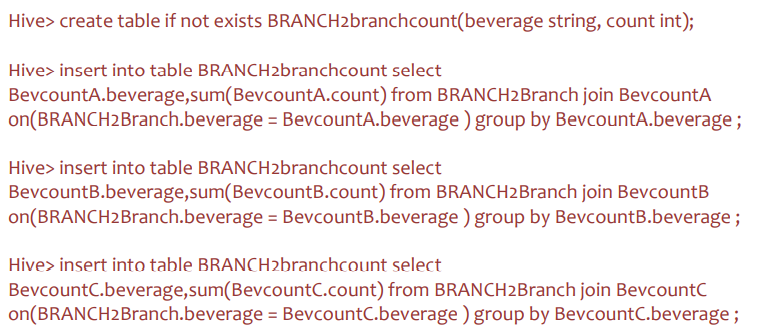
Explanation

**Step 1:** Creating physical table “BRANCH2Branch” from the previous tables where the branch name is equal to BRANCH1. The code as follows:



**Step 2**: Creating another table “BRANCH2branchcount” which merge tables which has the number of consumers with Beverages.

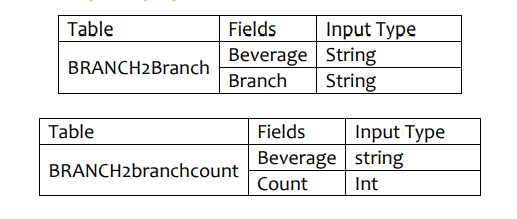
The code as follows:



**Step 3**: Finally, summation of the fields from total number of consumers in the newly created table “BRANCH2branchcount”. The code as follows:



**Step 4:** Created tables schema definition is as follows:



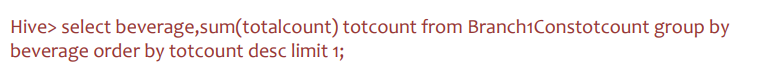
**Solution**

5099141

**Problem Scenario 3** - What is the most consumed beverage on Branch1?

Explanation

**Step 1:** Selecting the aggregate count from the previously created table Branch1Constotcount and ordering the data in descending. The code as follows**:**



**Solution**

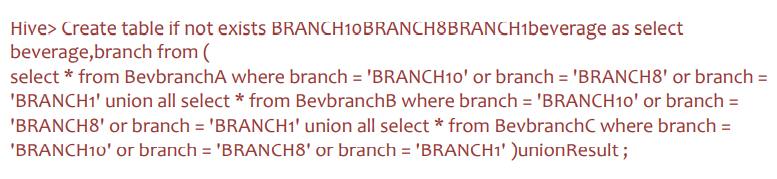
Special\_cappuccino 108163

**Problem Scenario 4** - What are the beverages available on Branch10, Branch8, and Branch1?

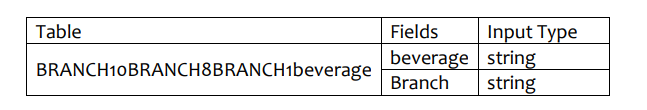
Explanation

**Step 1:** Creating a table “BRANCH10BRANCH8BRANCH1beverage” from the previous tables which has the records from the BRANCH10, BRANCH8, BRANCH1 branches.

The code as follows:

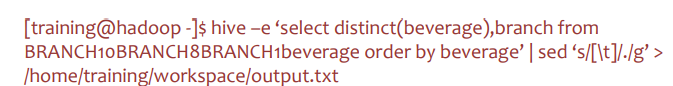


**Step 2:** “BRANCH10BRANCH8BRANCH1beverage” table schema definition



**Step 3**: Storing the output data table in a local comma separated file in hadoop.

The code as follows:



**OUTPUT FILE:**

****

**REFERENCES**

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Widom, J. (2020). Challenges and opportunities with big data.

Ozcan, S. (2013). Difference between Hadoop and RDBMS. Retrieved April 24, 2019.

Sharma, V., & Joshi, N. K. (2015). The evolution of big data security through Hadoop incremental security model.

**CONCLUSION**

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have produced a unique moment in the history of data analysis. The convergence of these trends means that we have the capabilities required to analyse astonishing data sets quickly and cost-effectively for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability.

The Age of Big Data is here, and these are truly revolutionary times if both business and technology professionals continue to work together and deliver on the promise.