Journal of Advanced Engineering Trends

ISSN : 2682 - 2091

Vol. 44, No. 2, July 2025

Research for Big Data Storage and Analysis Based on Artificial Intelligence

Rabab Hussien Mohamed[1,*], Kamel Hussien Rahouma[2]

[1] *Electrical Engineering Dep., Assuit University, Assuit, Egypt*
[1] *Electrical Engineering Dep., Minia University, Minia, Egypt*

* *Corresponding author(s) E-mail:* rababhussien21@yahoo.com

A R T I C L E I N F O

Article history:
Received: 23 July 2023
Accepted: 30 May 2025
Online: 1 June 2025

Keywords:
Big Data
Hadoop
HBase
Hive
Tableau
Artificial Intelligence

A B S T R A C T

In the age of big data, users generate a huge amount of data daily due to the rapid development of technology and the internet. These data are impossible to store or process by a single machine or in a traditional way. So, the need to use distrusted storage and processing systems was an emergency, such as the Apache Hadoop system, which provides a fault-tolerant, dependable, horizontally scalable, and effective service. It is based on the Hadoop distributed file system (HDFS) and MapReduce. Also, as experts and businessmen say, business is data. The need for analysis to understand business patterns and get significant insights from the available data is growing exponentially with the huge amount of data. Various organizations require an understanding analytical principle using machine learning, data prediction, and statistical techniques. Previously, only developers could perform these tasks; however, company workers can now immediately access these capabilities with cutting-edge tools. This research aims to integrate artificial intelligence with big data storage and analysis systems, using Hadoop, PySpark, Artificial Intelligence Algorithms, and Tableau to improve data processing efficiency and provide accurate analytical insights.

1. Introduction

Concrete has a significant impact effect on the environment, consuming materials from nature and generating one ton of carbon dioxide for every ton of (OPC) produced. Five percent of the greenhouse gases released into the atmosphere worldwide are produced during the cement making process [1, 6,16]. In 2050, it is expected that annual greenhouse gas emissions will exceed 2.34 billion tons if current conditions continue [3-4].

Currently, vast databases cannot be handled traditionally due to the quick growth of information technology and the expanding use of artificial intelligence. Additionally, advanced digitalization techniques combined with newly created modern technologies enable better, more value-added, and more economical manufacturing and service operations.

This research explores the effect of artificial intelligence (AI) in big data storage and analysis and uses business intelligence (BI) tools for data visualization. The main objectives:

- Improve big data management and storage using the Hadoop distributed system and Cloudera QuickStart VM for distributed data management. It leverages HDFS, Hive, and Spark for efficient storage and processing.
- Utilize AI techniques such as K-means Clustering, Random Forest, and Decision Tree by using Python and PySpark for effective data analysis. The integration of AI enhances pattern recognition, trend analysis, and predictive modeling, enabling informed decision-making.
- Evaluate AI model performance based on accuracy.
- Present results using Tableau to facilitate interpretation and data-driven decision-making.
- Tableau is a business intelligence tool based on AI for data analysis and visualization.

This study provides a practical framework for optimizing big data processing in real-world applications. For these objectives, we use the Hadoop echo system on Cloudera quick start VM to store data and retrieve it, apply AI algorithms for big data analysis, and use Tableau as a business intelligence tool based on AI for data analysis, visualization, and translating complex data into actionable insights. By combining distributed storage, AI-driven analysis, and interactive visualization,

This paper is structured into seven sections. Section one gives a brief introduction to the objectives and material used. Section two presents a background of big data and analysis, including the meaning of big data, the latest statistics for uses of big data, some basic definitions for the Hadoop ecosystem, and the importance of data analysis. Section three gives a literature review. Section four discusses the Methodology and processes of storing the data, retrieving it, and its analysis. Section five highlights the results of our experiments. Section six presents the discussion and conclusion. Section seven gives the future work.

2. Background

2.1. Big data meaning

Big data is defined as enormous, complicated datasets requiring novel computer techniques rather than typical ones. They might be unstructured, semi-structured, or structured. Gartner describes big data as "high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing to enable enhanced insight, decision making, and process automation [1]. It is characterized by 5 V's: Volume (Size), Velocity (Speed), Variety (Complexity), Veracity (data quality), and value of the data.

2.2. The latest Statistics of Big data

- Every day, 2.5 quintillion bytes of data are created.
- The global big data market, estimated to be worth $307.52 billion, is expected to grow to $745.15 billion by the end of 2030.
- Almost 97% of companies globally have invested in big data [2].

2.3. Basic definition used in this paper

Apache Hadoop: Apache Hadoop is a distributed data processing framework that uses easy algorithms based on the Google File System (GFS). It offers high throughput, fault tolerance, ample file storage, Scalability, reliability, and cost-effectiveness, making it suitable for large data sets [3]. It consists of:

Hadoop Distributed File System (HDFS): provides a highly reliable, fault-tolerant, and scalable distributed file system for storing big data across the clusters by dividing big files into blocks with replicated files and distributing them across nodes on clusters.

MapReduce: The programming model and processing engine enables the distributed processing of big data across clusters by breaking tasks into maps, reducing phases, and executing them across nodes on the cluster parallel distribution.

Spark: Spark is a fast and general computing engine for Hadoop data and a wide range of applications such as ETL, Machine Learning, Stream processing, and graph analysis. It allows the user to load data into memory and query it repeal.

HBase: It is a data warehouse tool that supports large amounts of sparse data.it is a No-SQL database that uses a key-value store run on top of Hadoop and provides real-time read and write access to data stored in HDFS.

Hive: Data warehouse infrastructure provides an SQL-like query language called HiveQL, which allows users to query and analyze data stored in HDFS.

Flume: it is a distributed, reliable service used for efficiently collecting, aggregating, and moving large amounts of log data.

Sqoop: The tool easily transfers structured data from an RDBMS to HDFS while preserving structure. That enables us to query the data; it stands for SQL to Hadoop. It works by spawning tasks on multiple data nodes to download various data portions in parallel. When you're finished, each piece of data is replicated to ensure reliability and spread out across the cluster to ensure you can process it in parallel on your cluster; we use sqoop in this project to automatically load data from MySQL to HDFS.

Pig: it is a high-level platform for creating MapReduce programs used for data processing and analysis

Oozie: the coordinator and workflow scheduler systems manage our Apache Hadoop jobs. It coordinates jobs that are triggered by frequency or data availability.

Cloudera Manager: a tool for monitoring and managing the cluster's configuration.

Parquet: is the file format, a columnar storage arrangement explicitly designed for large-scale queries typical in data warehouse scenarios.

Hue: is Impala's app to query our data and provide an interface for many tools on CDH on port (8888) on the Cloudera manager.

Impala: it is Cloudera's open source for massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop, also used for the sharing of databases and tables between the two components by integrating with the Apache Hive meta store database.

Tableau is the most potent growing data analytics data visualization platform. It uses several algorithms that help people to understand and view the data, such as:

Machine learning algorithms such as clustering, classification, and recommendation

Descriptive statistics such as bars, pie charts, and heat maps.

Predictive analytics is used for predictive analytics models, showcasing trends, and displaying forecasted values.

Spatial Analytics is used for geospatial visualizations, mapping, and location-based insights.

Text Analytics algorithms visualize sentiment analysis, word frequency, or other text-based insights.

2.4. Big Data Analysis and its Importance:

Big data analytics is essential in various fields. It analyzes vast, huge volume and unstructured datasets to identify correlation patterns and forecast consumer preferences. It aids in informed decision-making, product customization, and innovation, enhancing customer experiences, productivity, and cost reduction by streamlining processes and identifying development opportunities [4].

2.5. *Big Data Analysis using AI, BI tools based on Machine and deep learning*

Business intelligence systems (BI) combine operational and historical data with analytical tools to offer competitive and significant information. It aims to improve data quality and timeliness, enabling managers to understand company trends. This study uses Tableau as a BI tool that provides live visual analytics [5].

3. Literature Review

In [6], they combined several big data analytical methods to analyze integrated customer data. In this research, they used several techniques to optimize more effective and intelligent strategies for customer segmentation; they combined the regency, frequency, and monetary value (RFM) model, K-means clustering Algorithm, Naïve Bayes' Algorithm, and linked Bloom filters. They focus on big data mining algorithms and propose the analytics steps to determine customer segmentation strategies. They used the RFM model to divide the customers into favorite, general, and Inactive customers depending on their behavior and transactions, then applied the K-means Clustering Algorithm to divide data of sales to favorite and general, then used Naïve Bayes' Algorithm to predict new orders based on the previous analysis, lastly, the store data of inactive customers using LBF.

[7] presents a tensor-based big data management technique to reduce the dimensionality of data gathered from a smart city's Internet-of-Energy (IoE) environment. They extract the core data from collected data by using tensor operations such as vectorization, participation, and tenderization with the help of higher-order singular value decomposition, then store the core data on the cloud after reducing the dimensionality of data. They identify users who take part in the demand response (DR) mechanism and classify the end-users (residential and commercial) by using vector machine (SVM)--based classifiers into normal, overloaded, and under-loaded categories. Their result shows that the suggested tensor-based method for DR management is superior to the existing scheme.

In [8], this research proposed a review of recent trends in the big data life cycle on the floor and describing technologies, approaches, and strategies for every phase of the seven stages of the big data life cycle in manufacturing, with an emphasis on the user interface, maintenance, automation, quality control, decision-making, energy optimization, and flexibility. It also proposed the challenges and future research directions in Shopfloor's big data life cycle, data collection techniques, and data transmission protocols. It lays the foundation for possible future study avenues. The result shows that Regardless of the data's source, data privacy and any legal limitations that might be applicable have been emphasized as crucial factors, and The removal and filtering of unnecessary data, which could lower costs and processing power in applications with budgetary or storage limits, is not given enough attention. [8].

In [9], they examine big data analytics research using artificial intelligence methods. They choose relevant research publications using the Systematic Literature Review (SLR) method. These mechanisms are investigated by four groups: machine learning, search methods, optimization theory, knowledge-based and reasoning approaches, and decision-making algorithms. They also discuss the advantages and disadvantages of the chosen AI-driven big data analytics methods and analyze the relevant metrics, contrasting them regarding privacy, Scalability, efficiency, and precision. Their results show that most Machine learning-based systems have improved efficiency and accuracy as the key components. However, using inconsistent or insufficient data could lead to inaccurate results. Using search-based optimization techniques is very precise and efficient. However, these approaches are not sufficiently scalable. Using the knowledge base, reasoning, and knowledge-based methods enhances the quality of the analytics. Their development's relative simplicity. Even if there is less coverage for various scenarios, high precision will be provided by the scenarios these systems cover. A constraint programming problem simulates a decision-making problem, and a utility function maximizing is used to determine the desired solution. These techniques perform well in terms of accuracy, efficiency, and Scalability. They also present challenges, such as fog computing, processing vast quantities of data, security, qualitative parameters and metrics, and data quality [9].

From the literature reviews above, we can notice that [9] has pointed to the uses of artificial intelligence in big data analysis and mentioned the advantages and disadvantages of AI-driven big data analytics methods, such as privacy, Scalability, and efficiency. They all concurred that the K-means clustering Algorithm is the most effective and appreciated data mining method in the research community; therefore, we took advantage of them and applied the same technique to big data analysis. Most of the study only addresses one of the two topics: big data analysis or storage. Still, we merged the two and will discuss research procedures in future research.

4. Methodology

This paper provides an overview of Big data storage and analyses and the benefits of using AI on big data. The first objective is storing big data; we used Retail sample superstore databases on Cloudera quick start VM which is a virtual machine image provides a simple exploration and experiment on the Cloudera data platform (CDP) and Hadoop ecosystem and other big data technologies on the local Machine without the extensive and the complexities of setting up a full-scale cluster and adjust it's setting to be (8GB Ram – 2 core processor – 64GB Memory) as a single node cluster. The second objective is to analyze the big data; we Applied AI Algorithms such as Random forest, Discussion tree, and K-means clustering. Finally,We use the Tableau program for analysis and visualization, one of the Business Intelligence tools that depend on AI and machine learning.

4.1. *Storage of Big Data using Cloudera quick start VM*

We used Apache Hadoop to store the data, the most significant framework for storing and processing big data in parallel and distributed ways. It is also a method for resolving big data challenges [10]. The data source used was a Retailed dataset of 12.8 MB.it is preinstalled on the Cloudera quickstart VM and consists of 6 tables.

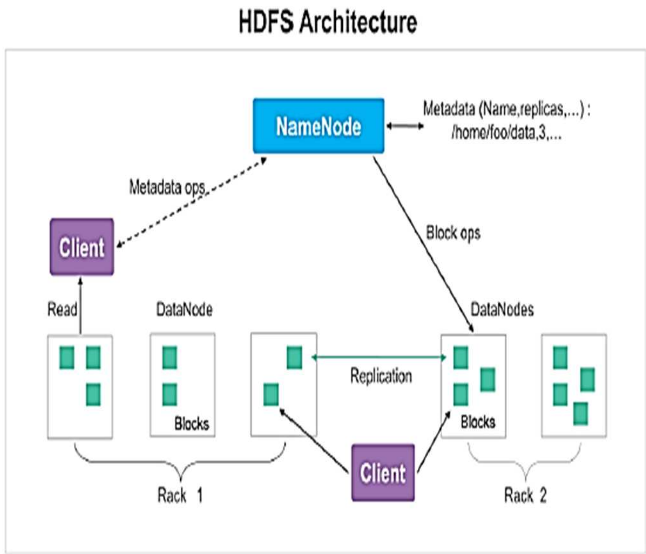


Figure 1: HDFS Architecture [11]

HDFS is one of the Hadoop ecosystems, as shown in Figure 1. HDFS architecture consists of four parts:

1. Name Node: it is the master node on the HDFS architecture, used to store and generate metadata of the file system and directories, such as file names, permissions, and block locations. There is only one active name node on the cluster.
2. Secondary Name Node (SNN): It is not a standby name node; it collects file system metadata from the active Name Node at regular intervals and combines it with the file system namespace's current state. In case of a failure, this procedure helps minimize the time needed for the Name Node to restart.
3. Data Node: used to store and manage actual data and report the data blocks; it manages to name nodes periodically with the list of blocks they store. HDFS runs multiple data node instances.
4. Client: allows services to access Hdfs and returns data obtained from name node and data node to services. HDFS runs multiple client instances [11].

The Data Writing process in detail

1. The Client sends a request to write to the Name node via API invoked by the service application.
2. Name Node oversees the management of the file system's metadata, which includes block locations, file permissions, and namespace structure. It creates a file node in the metadata after the Client connects to it and ensures that the file is new and the Client has the necessary permissions to create it.
3. The Client connects to data nodes and obtains the position and data block number from the name node, which connects to The Data node. The DFSOutputStream divides the data written by the Client into fixed-size blocks, typically 128 MB or 256 MB in size, which it then sends to the info queue, an indoor queue. The Data Streamer consumes the data queue and is responsible for selecting appropriate data nodes from the inventory to store the replicas and requesting the name node to allot new blocks. The set of data nodes creates a pipeline; in this case, we'll suppose that there are three nodes in the pipeline due to the replication factor of three (primary and two copies). The main data node in the pipeline receives the packets from the Data Streamer and stores them before forwarding them to the second data node in the pipeline. Then, the data node sends the Client a confirmation message after completing the data writing, and services invoke it to close the file.
4. The Client establishes a connection with the name node to ensure the data writing is finished and to make the metadata long-lasting, as shown in Figure 2.

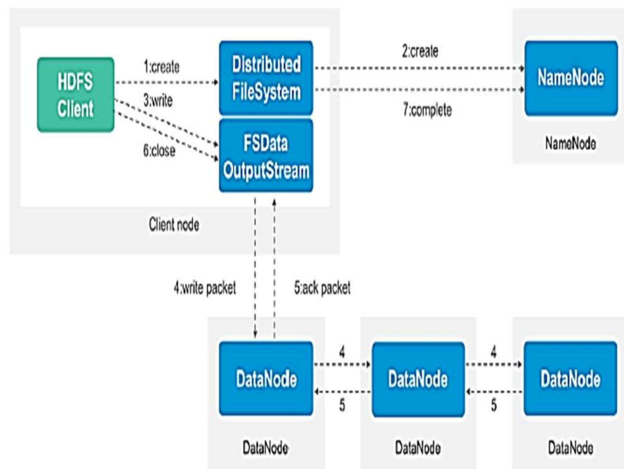


Figure 2: The Block Diagram of Data on HDFS [11]

The Data Read process in detail

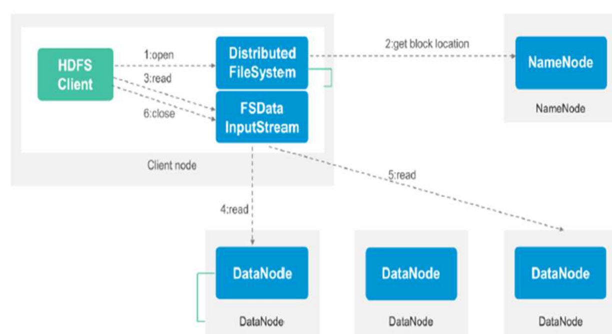


Figure 3: The Block Diagram of Data Read on HDFS [11]

A pseudo-code for the data-writing process:

Algorithm : Write the data into HDFS

Input : mysql database (Retail_db) consist of 6 Tables {Departments- Categories – Products – Order_Times –Orders – Customers }

output : Data base will be stored on HDFS on the cloudera quick start vm

Steps :

1- set up the cloudera quick start vm
  - Install it on my system ,start it

2- Access Hadoop Ecosystem services
  - Launch the cloudera manger and ensure that all the Hadoop services are running such as HDFS ,Hive and Hbase .

3- Write the code for storing the data into HDFS
  - Use Apache sqoop to transfer the data from RDMS to HDFS to enable us to query it easily .
  - Import the data into a form which prepared for Impala (the open source analytic query engine included with CDH)
  - Using the Apache Avro File formate for loading the data into Impala to be easily accessible since Avro is Hadoop optimized file formate

```
-- sqoop import-all-tables \
   -m 1 \
   --connect jdbc:mysql://quickstart:3306/retail_db \
   --username=***** \
   --password=*****\
   --compression-codec=snappy \
   --as-parquetfile \
   --warehouse-dir=/user/hive/warehouse \
   --hive-import
```

4- Verify that the data has been imported to HDFS

```
-- hadoop fs -ls /user/hive/warehouse/
-- hadoop fs -ls /user/hive/warehouse/categories/
```

A pseudo-code for the data Reading process

Algorithm : Read the data from HDFS

Input : distributed data base stored on HDFS

output : Data read from HDFS

Steps :

1- Verify that the data has been imported to HDFS
  -- Display the directory information by the command

  -- hdfs dfs -ls /path

  -- hadoop fs –ls // shows the number of items founded on HDFS and display it like the example below :

```
Found 3 items
drwx------  - cloudera cloudera    0 2023-03-01 02:49 .staging
drwxr-xr-x  - cloudera cloudera    0 2023-03-01 02:49 departments
drwxr-xr-x  - cloudera cloudera    0 2023-01-02 20:45 input
```

2- display the tables and it's content

```
--hadoop fs –ls /user/hive/warehouse / departments
Found 2 items
-rw-r--r--  1 cloudera cloudera    0 2023-03-01 02:49 departments/_SUCCESS
-rw-r--r--  1 cloudera cloudera   60 2023-03-01 02:49 departments/part-m-00000
-- hadoop fs -ls departments/part-m-00000
```

1. The Client sends a request for reading to the Name node via API invoked by the service application.
2. The Client establishes a connection with The name node To access file information (data block and data node information)
3. To read the file, the service application calls an API.
4. By using the information from the name node, the Client establishes a connection to The data node to locate
5. The Client connects to the data node depending on the information from the name node to locate the nearest corresponding data blocks.
6. The service application uses the close API to end the connection once the data reading is finished, as shown in Figure 3

4.2. Big Data Analysis

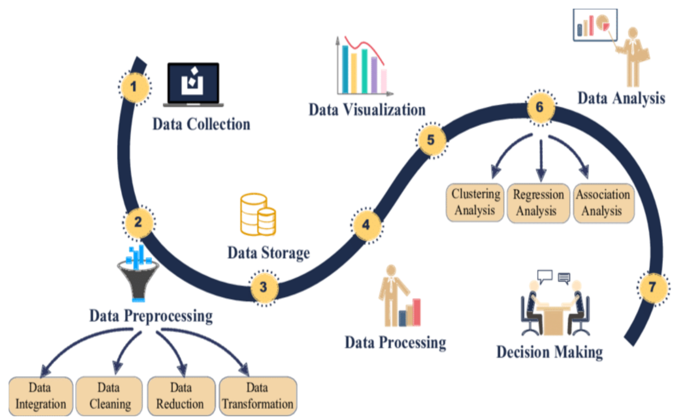


Figure 4: shows the stages of big data processes and analysis [12]

Integrating big data analytics and business intelligence has altered decision-making for many firms. These two tools are essential for corporate success and innovation to maintain a competitive edge, as shown in Figure 4. Businesses can increase performance, improve customer experiences, and achieve sustainable growth by implementing unique methodologies and platforms to address their big data concerns through the adoption of best practices and the exploration of creative solutions, as many Fortune 1000 organizations have done [13]

Big Data Analysis using Artificial Intelligence (AI).

The study presented an example of using AI algorithms such as classification to forecast jobs based on input features and clustering to divide jobs into groups and identify hidden patterns. Finally, it provides visualizations to help users understand the data and forecasts. Furthermore, the study employs PySpark in an ordered pipeline to efficiently process large amounts of data and apply scalable machine learning and artificial intelligence techniques. The data set used is a LinkedIn dataset with a size of 5 GB.

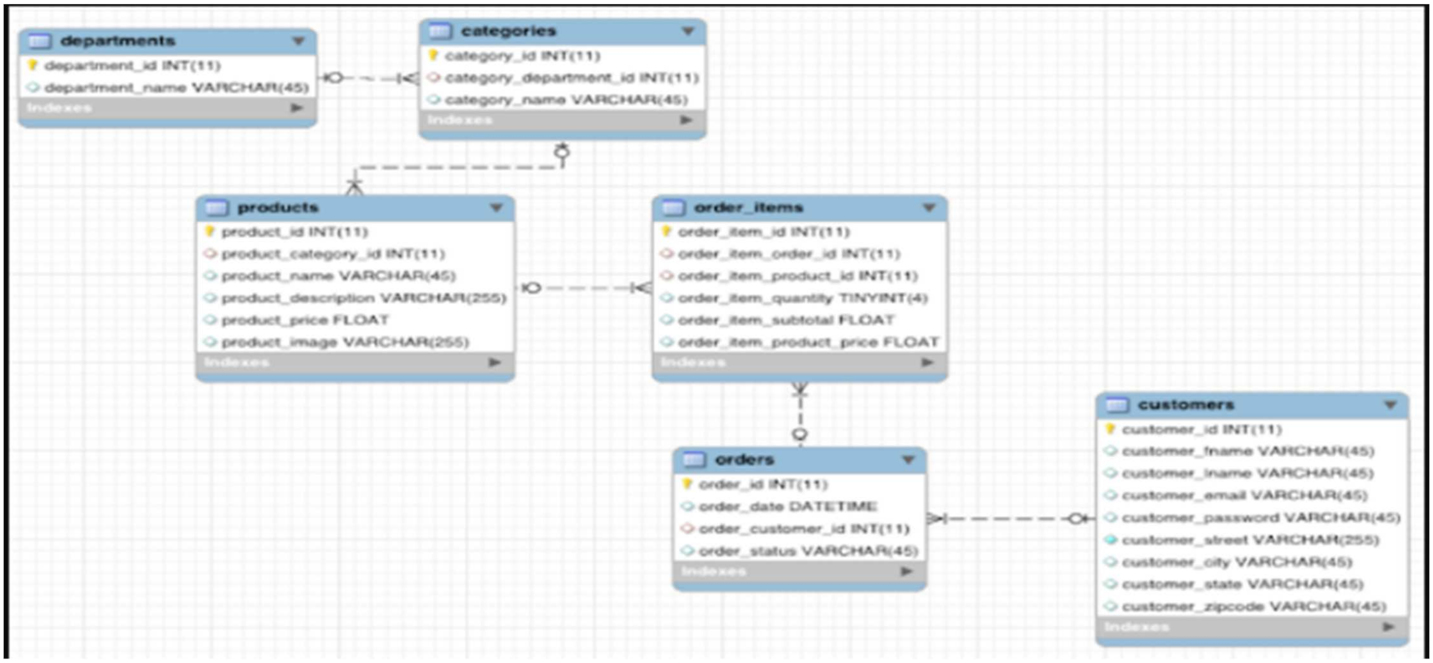


Figure 5: The ERD of the retail database

5. Results

5.1. Importing the data into HDFS

We used a retailed database base, a MySQL database already installed on the local Cloudera Quick Start VM host. It can also be downloaded from [14]. The data comprises 6 Tables [Departments- Categories – Products – Order Times –Orders – Customers] with a size of 12.8 MB. Figure 5 shows the database's ERD and some MySQL analysis before transferring the data to HDFS, as shown in Figures [6,7].

```
SELECT c.customer_fname, c.customer_lname,
o.order_date,o.order_status

FROM customers c JOIN orders o ON
(c.customer_id=o.order_customer_id)
 WHERE c.customer_state='TX' LIMIT 20;
```

```
mysql> SELECT c.customer_fname,c.customer_lname,o.order_date,o.order_status FROM
 customers c JOIN orders o ON(c.customer_id=o.order_customer_id) WHERE c.custome
r_state='TX' LIMIT 20;
+----------------+----------------+---------------------+------------------+
| customer_fname | customer_lname | order_date          | order_status     |
+----------------+----------------+---------------------+------------------+
| Brian          | Wilson         | 2013-07-25 00:00:00 | CLOSED           |
| Mary           | Marshall       | 2013-07-25 00:00:00 | COMPLETE         |
| Joan           | Smith          | 2013-07-25 00:00:00 | CLOSED           |
| Mary           | Smith          | 2013-07-26 00:00:00 | COMPLETE         |
| Amy            | Smith          | 2013-07-26 00:00:00 | COMPLETE         |
| Jack           | James          | 2013-07-26 00:00:00 | PROCESSING       |
| Mary           | Garner         | 2013-07-26 00:00:00 | CLOSED           |
| Donald         | Smith          | 2013-07-26 00:00:00 | PENDING_PAYMENT  |
| Mary           | Werner         | 2013-07-26 00:00:00 | PENDING          |
| Mary           | Gutierrez      | 2013-07-26 00:00:00 | PENDING_PAYMENT  |
| Kathy          | Little         | 2013-07-27 00:00:00 | PROCESSING       |
| Angela         | Peterson       | 2013-07-27 00:00:00 | COMPLETE         |
| Jane           | Day            | 2013-07-27 00:00:00 | PROCESSING       |
| Carolyn        | Price          | 2013-07-27 00:00:00 | PROCESSING       |
| Harold         | Johnson        | 2013-07-27 00:00:00 | COMPLETE         |
| Michael        | Smith          | 2013-07-27 00:00:00 | COMPLETE         |
| Mary           | Chapman        | 2013-07-27 00:00:00 | CLOSED           |
| Michelle       | Bennett        | 2013-07-27 00:00:00 | PENDING          |
| Sandra         | Bennett        | 2013-07-27 00:00:00 | CLOSED           |
| Mary           | Smith          | 2013-07-27 00:00:00 | COMPLETE         |
+----------------+----------------+---------------------+------------------+
20 rows in set (0.12 sec)
```

Figure 6: Shows the orders of customers who live in Texas

```
SELECT COUNT (*) as count FROM customers c JOIN orders o ON
(c.customer_id=o.order_customer_id) WHERE
o.order_status='PENDING_PAYMENT';
SELECT DISTINCT (Year (order_date)) FROM
orders WHERE
order_status='PENDING_PAYMENT';
```

```
Database changed
mysql> SELECT COUNT(*) as count FROM customers c JOIN orders o ON(c.customer_id=
o.order_customer_id) WHERE o.order_status='PENDING_PAYMENT';
+-------+
| count |
+-------+
| 15030 |
+-------+
1 row in set (0.19 sec)

mysql> SELECT DISTINCT(Year(order_date)) FROM orders WHERE order_status='PENDING
_PAYMENT';
+--------------------+
| (Year(order_date)) |
+--------------------+
|               2013 |
|               2014 |
+--------------------+
2 rows in set (0.12 sec)
```

Figure 7: Shows the count of pending payment orders and its date

We loaded the database into HDFS using sqoop. Then, we loaded it to an open-source analytics platform called Impala on CDH using Avro file format to analyze the data efficiently, as shown in Figure [8,9]. We used Cloudera Quick Start VM to set specifications (2 core, 8GB RAM, and 64GB hard disk). In our practical experience, We wanted to measure the speed of writing and reading data into HDFS, as shown in Figure 10. Our results show that writing six records with size 60 bytes in 35.5755 seconds (speed 1.6866 bytes /sec).

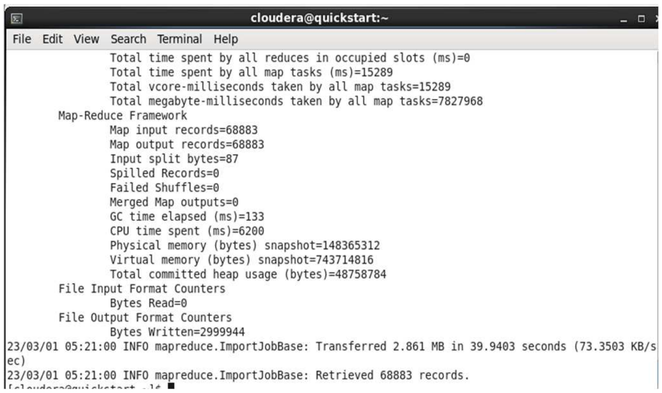


Figure 8: Importing data into HDFS

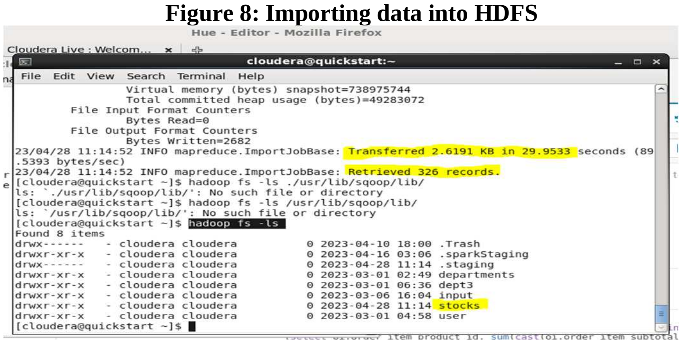


Figure 9: Data stored on HDFS

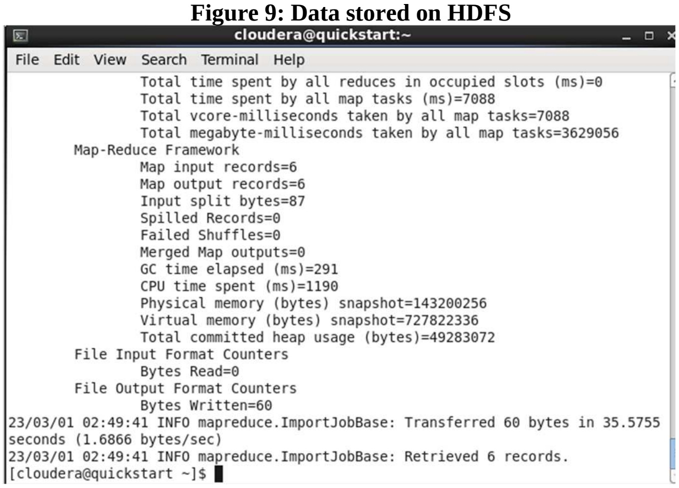


Figure 10: Shows the speed of writing small data.

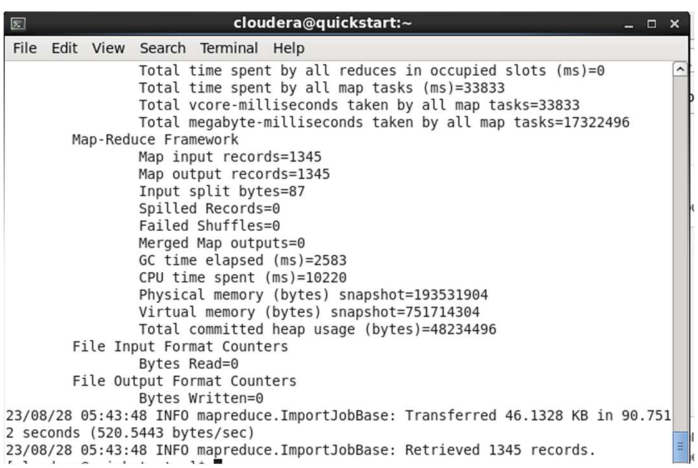


Figure 11: Shows the speed of writing data

Figure 11 shows that writing 1345 records with size 46.13 KB in 90.751 seconds (speed 520.5 bytes /sec). Figure 12 shows that writing 68883 records with size 2.861 MB in 39.94 seconds (speed 73.35 KB /sec). Figure 13 shows that reading 1024 records with size 0.23 MB in 12.6 seconds (speed 18 KB /sec).

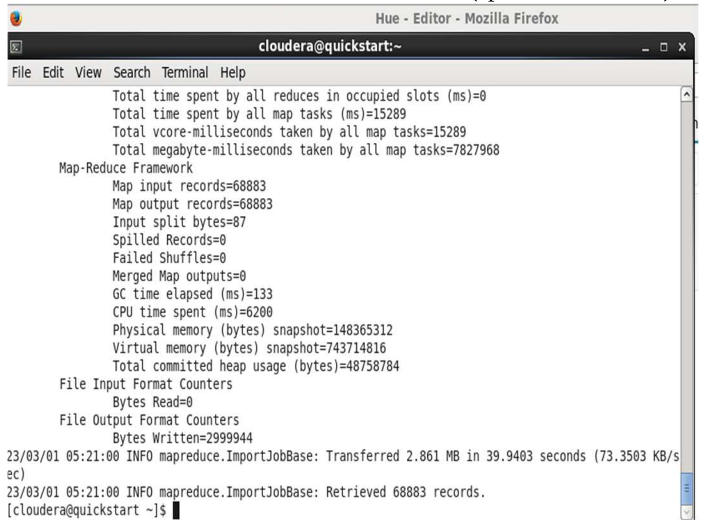


Figure 12: Shows the speed of the writing order table

Figure 13 shows that reading 1024 records with size 0.23 MB in 12.6 seconds (speed 18 KB /sec)

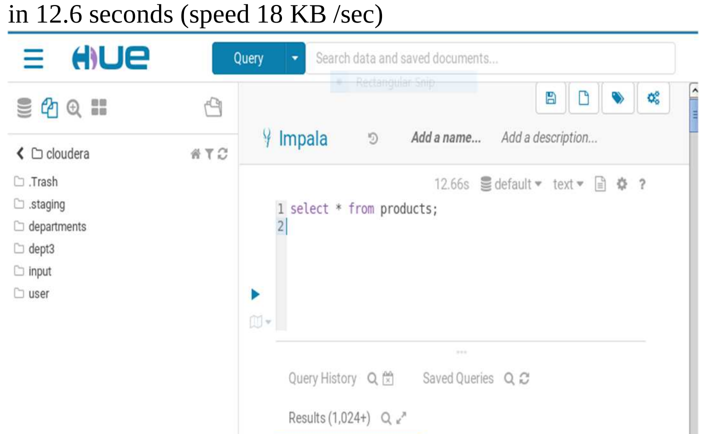


Figure 13 shows the speed of the reading process

5.2. Big Data Analysis using Hive and Impala

```
-sqoop import-all-tables \
  -m 1 \
  --connect jdbc: mysql: //quickstart:3306/retail_db \
  --username=retail_dba \
  --password=cloudera \
  --compression-codec=snappy \
  --As-parquet file \
  --warehouse-dir=/user/hive/warehouse \
  --hive-import
```

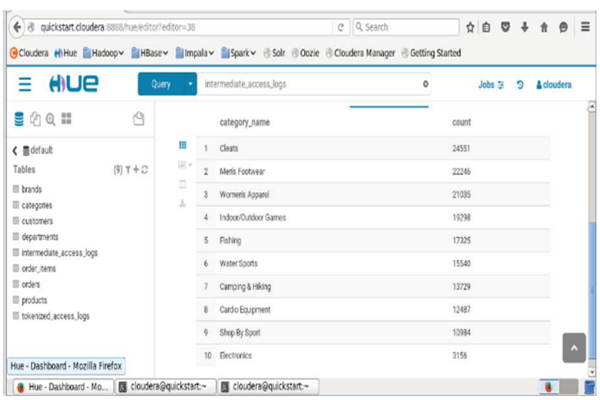
The data is loaded through Impala using Avro (Hadoop optimized file format) using the command above; this command launches MapReduce jobs and creates tables on Apache Hive to represent it via Impala. We used the Apache parquet format for transferring the data from MySQL to HDFS, which groups data into columns instead of rows by default and is intended for analytical applications on the Hadoop ecosystem; the numbers of .parquet equal the numbers of MapReduce jobs; we do some analysis on data using Hive and Impala as shown in Figures [14-16].

	product_id	product_name	revenue
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	6637668.2823181152
2	365	Perfect Fitness Perfect Rip Deck	4233794.3682899475
3	957	Diamondback Women's Serene Classic Comfort Bi	3946837.0045471191
4	191	Nike Men's Free 5.0+ Running Shoe	3507549.2067337036
5	502	Nike Men's Dri-FIT Victory Golf Polo	3011600
6	1073	Pelican Sunstream 100 Kayak	2967851.6815185547
7	1014	O'Brien Men's Neoprene Life Vest	2765543.314743042
8	403	Nike Men's CJ Elite 2 TD Football Cleat	2763977.4868011475
9	627	Under Armour Girls' Toddler Spine Surge Runni	1214896.220287323
10	565	adidas Youth Germany Black/Red Away Match Soc	63490

```
select p.product_id, p.product_name, r.revenue
from products p inner join
(select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as revenue
from order_items oi inner join orders o
on oi.order_item_order_id = o.order_id
where o.order_status <> 'CANCELED'
and o.order_status <> 'SUSPECTED_FRAUD'
group by order_item_product_id) r
on p.product_id = r.order_item_product_id
order by r.revenue desc
limit 10;
```

Figure 14: Top 10 revenue-generating products Result

```
select c.category_name, count(order_item_quantity) as count
from order_items oi
inner join products p
on oi.order_item_product_id = p.product_id
inner join categories c
on c.category_id = p.product_category_id
group by c.category_name
order by count desc
limit 10;
```



	category_name	count
1	Cleats	24551
2	Men's Footwear	22246
3	Women's Apparel	21035
4	Indoor/Outdoor Games	19298
5	Fishing	17325
6	Water Sports	15540
7	Camping & Hiking	13729
8	Cardio Equipment	12487
9	Shop By Sport	10984
10	Electronics	3156

Figure 15: Most popular product categories and their count

```
select max(order_item_quantity) as highest_quantity , p.product_name ,o.order_status
from order_items oi join products p on oi.order_item_product_id =p.product_id
join orders o on o.order_id =oi.order_item_order_id
group by p.product_name ,o.order_status
order by highest_quantity desc
limit 20;
```

	highest_quantity	product_name	order_status
1	5	Clicgear Rovic Cooler Bag	CLOSED
2	5	Glove It Women's Imperial Golf Glove	CLOSED
3	5	Glove It Women's Mod Oval 3-Zip Carry All Gol	CLOSED
4	5	Clicgear 8.0 Shoe Brush	ON_HOLD
5	5	Bridgestone e6 Straight Distance NFL Tennesse	PENDING_PAYMENT
6	5	Hirzl Women's Soffft Flex Golf Glove	CANCELED
7	5	Hirzl Women's Hybrid Golf Glove	PROCESSING
8	5	Titleist Pro V1x Golf Balls	COMPLETE
9	5	Team Golf Texas Longhorns Putter Grip	PROCESSING
10	5	Nike Women's Free 5.0 TR FIT PRT 4 Training S	CLOSED

Figure 15: Highest quantity for each order

5.3. *Big data analysis using AI algorithms*

We have applied various AI and machine learning (ML) models, including Random forest Classifier (RF), as shown in Figure 17, Decision Tree Classifier (DT) for supervised learning, as shown in Figure 18, and K- means clustering for unsupervised learning as shown on Figure 19. The LinkedIn dataset is used in this section with a size of 5 GB. The dataset has been prepared and cleaned to handle and remove missing values. The result shows that the prediction of the most jobs is chosen according to the level and type of jobs and the search country. The model was implemented using PySpark MLlib and Python libraries. It was trained using 80% of the data and tested with the remaining 20%.

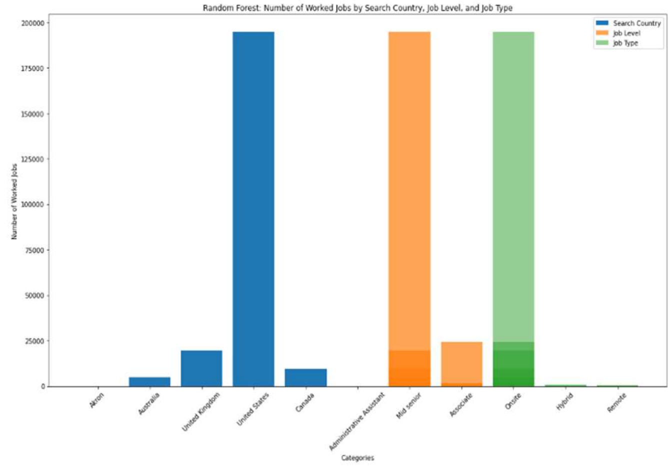


Figure 17: Random forest

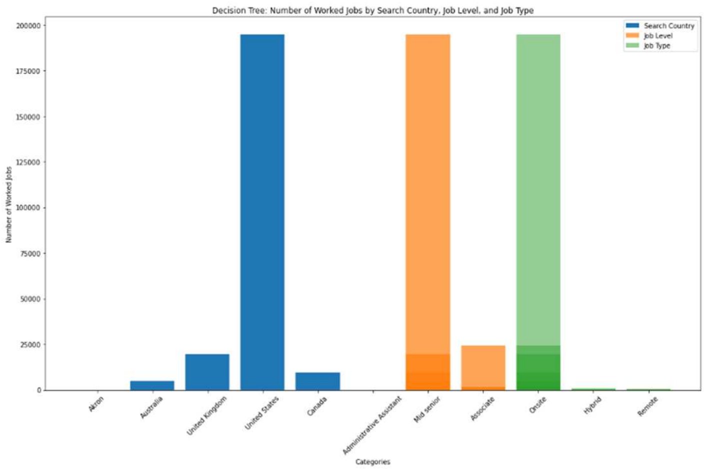


Figure 18: Discussion tree result

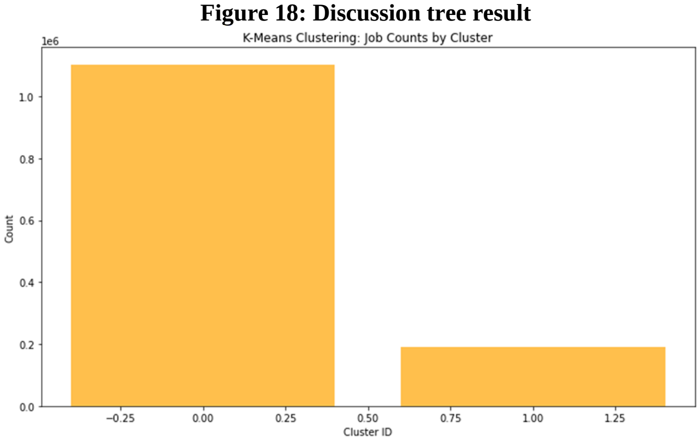


Figure 19: K- means clustering

5.4. Data Analysis using Tableau

- We used Tableau as a powerful BI tool for Data Analysis and Visualization using AI and ML techniques; it is a powerful data visualization solution that helps people view and understand data and does not require any technical or programming experience. Many companies use Tableau, such as Honey Well, Hello Fresh, Lenovo, Verzon, etc. It is user-friendly, applicable to any business, easy to access data reports, and has no necessary coding. Tableau facilitates connections to a broad spectrum of data stored in different locations.
- We used Tableau with a database sample super store downloaded from [15] with a size of 3.306 MB; we prepared the data and cleaned it for better results using Joins, Relationships, union, Data blending, Data interpreter, pivot, Aggregation, sort, group, and split.
- We applied filters to data on different categories to organize the data to achieve its demand; many types of filters can be used, such as Dimension Filter, Measure Filter, Data Filter, Interactive Filter, Data source Filter, and Context Filter, as shown in Figure [17-19].
- We analyze data and create charts, maps, and tables for a better representation of the information, a great way to make a decision, and a high-level overview of data collection and extraction, as shown in the results.

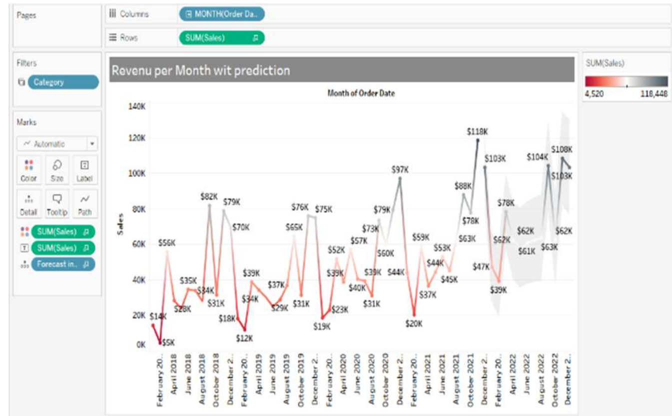


Figure 20: Shows the trend of a sum of Sales (actual & forecast) for Order Date Month. Details are shown for the Forecast indicator. The data is filtered into Categories, which include Furniture, Office Supplies, and Technology

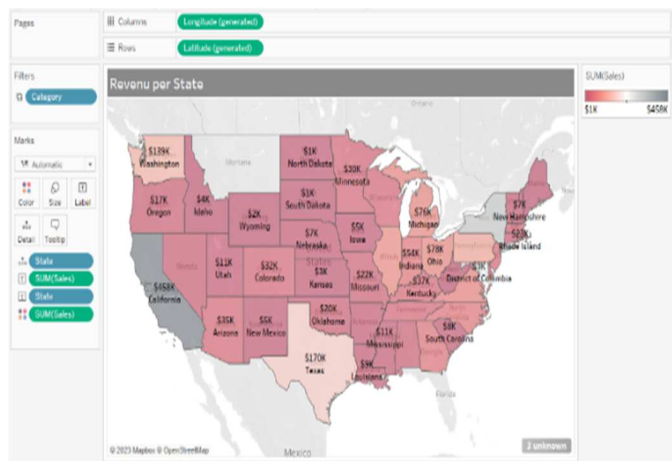


Figure 21 shows revenue per state as a map visualization.

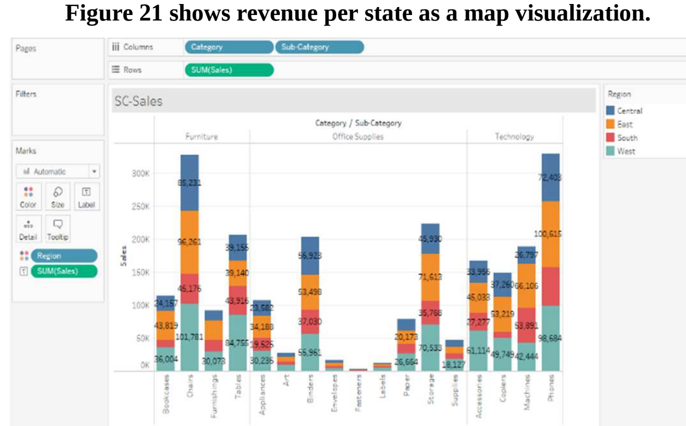


Figure 22 shows sales of each sub-category and region.

- We applied some calculations to make the data interpretation easy.

Tableau software has three types of Calculation (Basic expression, Level of Detail expression known as LOD, and Table Calculation) as shown in Figures [20- 23].

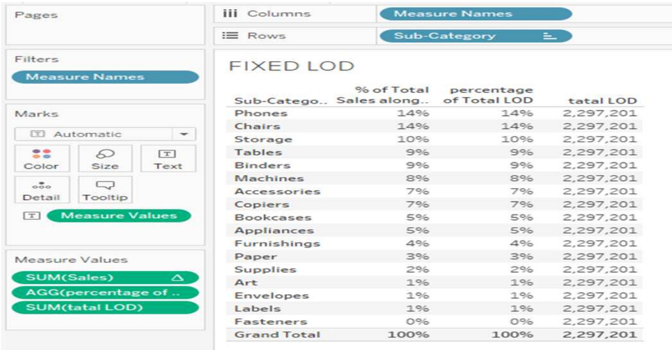


Figure 23: Shows Fixed Calculation applied to sales and subcategories

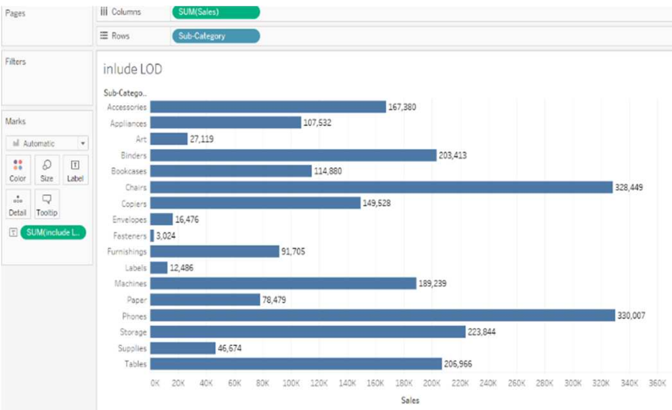


Figure 24: shows subcategories by sales using Included LOD Calculation

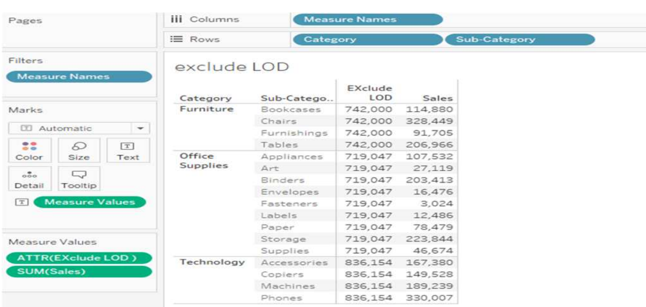


Figure 25: Subcategory and category by sales using Exclude LOD Calculation

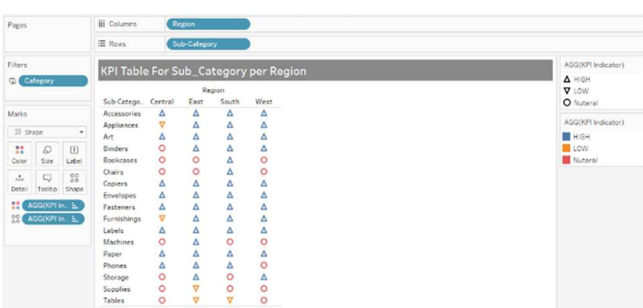


Figure 26 shows the performance of the Profit Ratio for each subcategory

Additionally, charts were created to show the contribution of positive and negative dimension members to the total value as a Waterfall Chart or charts containing both bar and line charts.

- Creating A Dashboard and Stories is a real-time and easy-to-read user interface that presents data graphically as a combination of various charts, Tables, Maps, and Calculations in one place to convey Business insights as shown in Figures [24-27].

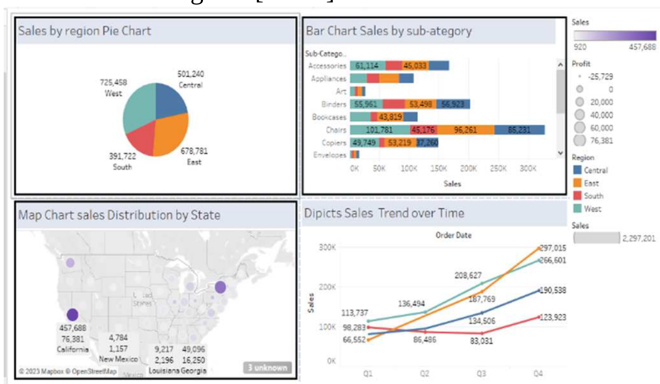


Figure 27: Example of Dashboard

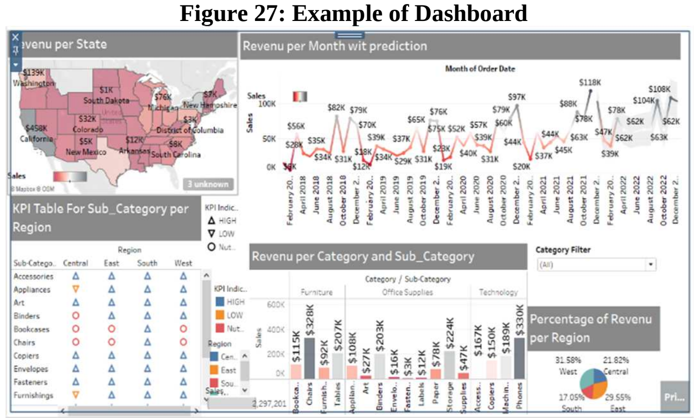


Figure 28 shows an entire dashboard containing all data analyses applied to the data set for better understanding

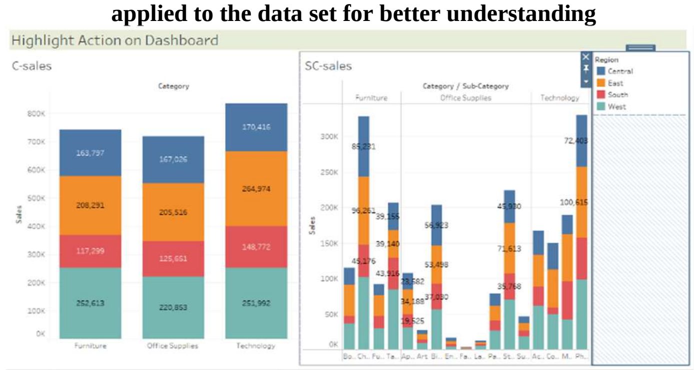


Figure 29: using Highlight on Dashboard

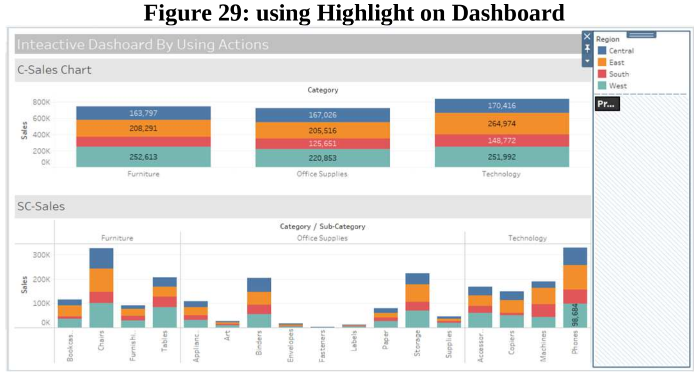


Figure 30: Use Action on Dashboard to make it more interactive.

Tableau has many products besides Tableau Desktop, such as Tableau Reader, Tableau Public Tableau Server, Tableau Online, Tableau Prep Builder, Tableau Mobile, Tableau Cloud, Tableau Prep, and Tableau CRM as shown in Figure 28.

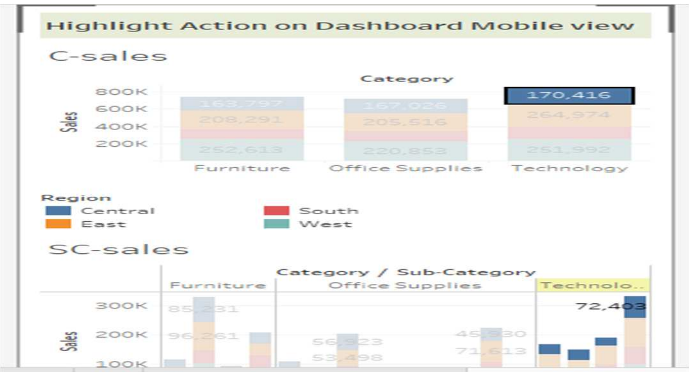


Figure 31: Mobile view for a dashboard

6. Discussion and Conclusion

Previous researchers have discussed various approaches for storing or analyzing big data. Still, we wanted to merge the two processes into one study and talk about the most effective ways to do so. We used the Hadoop ecosystem on the Cloudera quickstart VM to store and extract the data. AI algorithms were applied to study big data analysis, and we used Tableau for data visualization.

We noticed from practical work that the speed of reading and writing data in the Hadoop ecosystem might differ depending on several factors, including:

1. Network communication speed: HDFS performance depends heavily on it, but other bottlenecks exist. The data transmission speeds of hard drives are highly critical, particularly in high-speed local or metropolitan area networks.
2. Data size and distribution: HDFS is very efficient for large file processing, but it does not apply to large numbers of small files due to large numbers of map tasks created while storing files, which reduces the file system's performance. Larger files typically result in faster read and write operations than smaller files due to reduced overhead.
3. Data Formats and Compression: The data formats and compression methods can affect read and write speeds. For instance, read performance can be enhanced by employing columnar storage formats like Apache Parquet or Apache ORC, particularly for workloads involving analysis. Although compression can lower storage needs, it can also result in higher CPU overhead when reading and writing data.

So, The performance of Cloudera Quick Start VM may not be representative of a production-grade Hadoop cluster because it is primarily a virtual environment intended for learning and development rather than production use, so these results are not for general to store big data because of the speed varies depending on the configuration on the VM and the size and number of files being read or written.

The speed of data analysis using Tableau also depends on many factors, such as:

1. Data source: the type, size, and complex query affect the speed of data analysis on Tableau.
2. Hardware configuration: more RAM and a solid-state drive (SSD) can improve performance.
3. The complexity of the data model used, such as the number of calculations, filters, and visualizations, simplifies it and increases the speed.
4. Using Dashboard design: An efficient dashboard is better than many complex visualizations that limit the number of parameters and fitters to increase performance.

The Benefits of AI in Big Data Analysis:

- Scalability: AI algorithms like those implemented in Apache Spark can handle large-scale datasets without compromising performance.
- Automation: Machine learning models reduce manual effort by automatically detecting patterns and making predictions.
- Enhanced Decision-Making: AI models provide actionable insights, helping businesses optimize strategies based on data-driven conclusions.
- Improved Accuracy: Advanced classification models like Random Forest and Decision Tree enhance the accuracy of predictions compared to traditional methods.
- Unsupervised Learning for Discovering Patterns: Clustering methods such as K-Means help identify hidden structures in data that might not be evident through conventional analysis.

By leveraging AI, businesses can enhance workforce planning, recruitment strategies, and market insights, making data analysis more efficient and impactful.

7. Future Work

- We will focus more on prediction and data analysis.
- Studying the security of big data storage and analysis.
- Studying the search process on big data

The data is loaded through Impala using Avro (Hadoop optimized file format) using the command above; this command launches MapReduce jobs and creates tables on Apache Hive to represent it via Impala. We used the Apache parquet format for transferring the data from MySQL to HDFS, which groups data into columns instead of rows by default and is intended for analytical applications on the Hadoop ecosystem; the numbers of .parquet equal the numbers of MapReduce jobs; we do some analysis on data using Hive and Impala as shown in Figures [14-16].

Conflict of Interest

The authors declare no conflict of interest.

References

[1] K. Vassakis, E. Petrakis, and I. Kopanakis, "Big Data Analytics: Applications, Prospects and Challenges," Mobile Big Data, vol. 10, pp. 3–20, Nov. 2018, doi: https://doi.org/10.1007/978-3-319-67925-9_1.

[2] R. Shewale, "65 Big Data Statistics 2023 (Facts, Trends & More)," Demandsage, Sep. 02, 2023. https://www.demandsage.com/big-data-statistics/

[3] "Apache Hadoop 3.3.1 – HDFS Architecture," hadoop.apache.org. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html.

[4] Md. Toriqul Islam and Borhan Uddin Khan, "Big Data and Analytics," The gold Open Access encyclopedia, Encyclopedia of Information Science and Technology, no. Sixth Edition, pp. 1–30, doi: https://doi.org/10.4018/978-1-6684-7366-5.ch048.

[5] Jasmin Praful Bharadiya, "A Comparative Study of Business Intelligence and Artificial Intelligence with Big Data Analytics," American Journal of Artificial Intelligence, vol. 7, no. 1, Jun. 2023, doi: https://doi.org/10.11648/j.ajai.20230701.14.

[6] S.-C. Wang, Y.-T. Tsai, and Y.-S. Ciou, "A hybrid big data analytical approach for analyzing customer patterns through an integrated supply chain network," Journal of Industrial Information Integration, vol. 20, p. 100177, Dec. 2020, doi: https://doi.org/10.1016/j.jii.2020.100177.

[7] A. Jindal, N. Kumar, and M. Singh, "A unified framework for big data acquisition, storage, and analytics for demand response management in smart cities," Future Generation Computer Systems, Mar. 2018, doi: https://doi.org/10.1016/j.future.2018.02.039.

[8] T. Pulikottil et al., "Big Data Life Cycle in Shopfloor–Trends and Challenges," IEEE Access, vol. 11, pp. 30008–30026, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3253286.

[9] A. M. Rahmani et al., "Artificial intelligence approaches and mechanisms for big data analytics: a systematic study," PeerJ Computer Science, vol. 7, p. e488, Apr. 2021, doi: https://doi.org/10.7717/peerj-cs.488.

[10] "Anatomy of File Read and Write in HDFS," GeeksforGeeks, Jun. 12, 2020. https://www.geeksforgeeks.org/anatomy-of-file-read-and-write-in-hdfs/

[11] "HCIA-Big Data V2.0 Training Material," pdfcoffee.com. https://pdfcoffee.com/hcia-big-data-v20-training-material-3-pdf-free.html (accessed Jun. 08, 2024).

[12] S. Yu, Q. Qing, C. Zhang, A. Shehzad, G. Oatley, and F. Xia, "Data-Driven Decision-Making in COVID-19 Response: A Survey," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1016–1029, Aug. 2021, doi: https://doi.org/10.1109/tcss.2021.3075955.

[13] O. O. Olaniyi, A. I. Abalaka, and S. O. Olabanji, "Utilizing Big Data Analytics and Business Intelligence for Improved Decision-Making at Leading Fortune Company," Journal of Scientific Research and Reports, vol. 29, no. 9, pp. 64–72, Sep. 2023, doi: https://doi.org/10.9734/jsrr/2023/v29i91785.

[14] nitendratech, "Analyze retail DB using Structured Query Language (SQL)," Technology and Trends, Jan. 05, 2019. https://www.nitendratech.com/database/retail-data-sql/

[15] "Tableau Community Forums," community.tableau.com. https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls