

**School of Tech**

**Graduate Diploma in Data Analytics (Level 7) Cover Sheet and Student Declaration**

This sheet must be signed by the student and attached to the submitted assessment.

|  |  |  |  |
| --- | --- | --- | --- |
| **Course Title:** | **Big Data Analytics** | **Course code:** | **GDDA709** |
| **Student Name:** | VIKAS ANDOTRA | **Student ID:** |  |
| **Assessment No**  **& Type:** | Assessment 2[Portfolio] | **Cohort:** |  |
| **Due Date:** |  | **Date Submitted:** |  |
| **Tutor’s Name:** |  | | |
| **Assessment**  **Weighting** | 80% | | |
| **Total Marks** | 100 | | |

### Student Declaration:I declare that:

* I have read the New Zealand School of Education Ltd policies and regulations on assessments and understand what plagiarism is.
* I am aware of the penalties for cheating and plagiarism as laid down by the New Zealand School of Education Ltd.
* This is an original assessment and is entirely my own work.
* Where I have quoted or made use of the ideas of other writers, I have acknowledged the source.
* This assessment has been prepared exclusively for this course and has not been or will not be submitted as assessed work in any other course.
* It has been explained to me that this assessment may be used by NZSE Ltd, for internal and/or external moderation.

### Student signature:

**Date:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tutor only to complete** | | | | | |
| **Assessment results:** | **Task 1**  (max. 15 marks) | **Task 2**  (max. 25 marks) | **Task 3**  (max. 35 marks) | **Task 4**  (max. 20 marks) | **Task 5**  (max. 5 marks) |
| **Total Marks: /100** | | **Grade:** | | |
| **Tutor only to complete** | | | | | |
| **Assessment results:** | **Mark /100** | | **Grade** | | |
| **LO2 Requirements** | Met  Not Met | | **Tutor Signature:** | | |
| **LO3 Requirements** | Met  Not Met | |
| **LO4 Requirements** | Met  Not Met | |

**Table of Contents**

[Student Declaration:I declare that: 1](#_Toc191468668)

[Student signature: 2](#_Toc191468669)

[Introduction 4](#_Toc191468670)

[Task 1: Implementation of Big Data Framework 5](#_Toc191468671)

[Part A: HDFS Configuration and Setup 5](#_Toc191468672)

[Part B: Big Data Framework Architecture 8](#_Toc191468673)

[Task 2: Applying MapReduce on Big Data to Distribute Processing Among Multiple Machines in a Cluster 11](#_Toc191468674)

[Part A: MapReduce Workflow and Execution 11](#_Toc191468675)

[Part B: Business Applications of MapReduce 12](#_Toc191468676)

[Task 3: Implement Machine Learning Algorithms on Big Data for Predictions and Classification of Time Series Data 13](#_Toc191468677)

[Part A: ML Pipeline and Algorithm Selection 13](#_Toc191468678)

[Part B: Real-World Applications of ML 15](#_Toc191468679)

[Task 4: Discussion and Conclusion: 16](#_Toc191468680)

[Part A: Discussion 16](#_Toc191468681)

[Part B: Conclusion 17](#_Toc191468682)

[References 19](#_Toc191468683)

Table of Figures:

[Figure 1: configuration in hdfs-site, (Source: Self-Created) 5](#_Toc191468442)

[Figure 2: configuration mapred-site, (Source: Self-Created) 6](#_Toc191468443)

[Figure 3: Hadoop Configuration Files, (Source: Self-Created) 7](#_Toc191468444)

[Figure 4: Big Data Framework Architecture, (Source: Self-Created) 8](#_Toc191468445)

[Figure 5: Comparison of Hadoop and Spark Frameworks, (Source: Self-Created) 9](#_Toc191468446)

[Figure 6: Cloud-Based Big Data Processing, (Source: Self-Created) 10](#_Toc191468447)

[Figure 7: MapReduce Workflow Diagram, (Source: Self-Created) 11](#_Toc191468448)

[Figure 8: E-Commerce Transaction Analysis Using MapReduce, (Source: Self-Created) 12](#_Toc191468449)

[Figure 9: Sentiment Analysis on Twitter Data Using MapReduce, (Source: Self-Created) 13](#_Toc191468450)

[Figure 10: Machine Learning Pipeline for Big Data, (Source: Self-Created) 13](#_Toc191468451)

[Figure 11: Time Series Prediction Model for Stock Price Forecasting, (Source: Self-Created) 14](#_Toc191468452)

[Figure 12: Disease Prediction Model in Healthcare Big Data, (Source: Self-Created) 15](#_Toc191468453)

[Figure 13: Fraud Detection System Using Machine Learning on Big Data, (Source: Self-Created) 16](#_Toc191468454)

# Introduction

New technology of big data provides companies reengineered ability to handle large sets of data in making data-supported decisions. The analysis looks into processes of embracing big data models through machine learning processes of processing and interpretation of consistent sets of data. Apache Hadoop and Apache Spark operations break down the data distribution into clusters and compute algorithms advance improve improved predictions to make the forecasts.

Because the world's biggest retail firm, Walmart, remains atop by having lengthy supply chains to support many different products with far-reaching customers (Dolezel et al., (2021). As a data firm, Walmart employs big data systems within its business for optimization of inventory control as well as sales predictions and customer interactions also. The paper utilizes a live dataset of Walmart operations to test big data solutions in an actual business environment.

The study strives to build end-to-end data streams using MapReduce processing in executing supervised and unsupervised models for recognition and prediction for time series. The contribution of the study highlights the importance of big data tools by systematic application and testing aimed at overcoming scalability challenges as well as improving efficiency and predictability analysis capabilities (Abdalla, 2022). The research results create a basis that is the cornerstone of next-generation retail decision-making systems capable of effectively handling big data.

# Task 1: Implementation of Big Data Framework

## Part A: HDFS Configuration and Setup

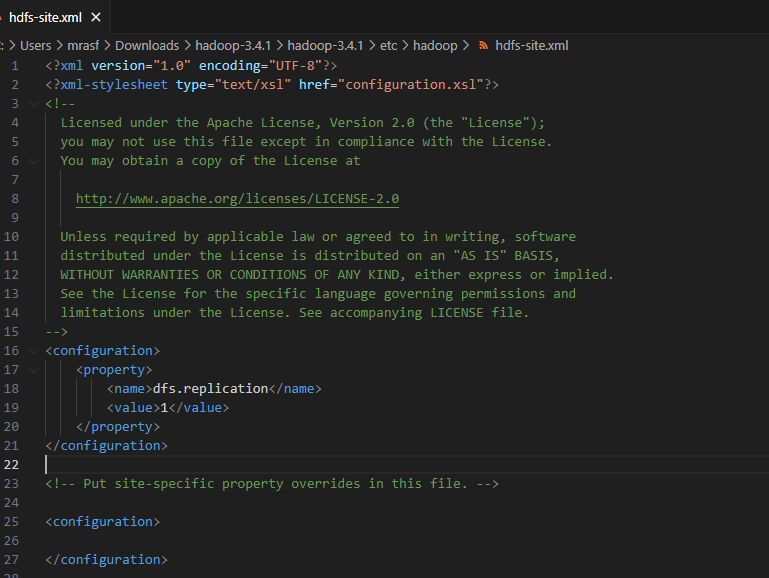


Figure 1: configuration in hdfs-site, (Source: Self-Created)

* **Importance of hdfs-site.xml in HDFS Configuration:** The highlighted configuration “configuration in hdfs-site” shows how different configuration parameters are defined within hdfs-site.xml, critical in the setting and the managing of the Hadoop Distributed File System (HDFS).
* **Key Parameters in hdfs-site.xml and Their Functions:** This file contains basic settings for HDFS functioning, such as the replication factor, the number of copies of each block to provide redundancy, defined by dfs.replication. Others including the directories for NameNode and DataNodes determine where in the nodes of the cluster metadata and content data is located (Cuzzocrea & Ciancarini, 2024). Such a configuration means that the HDFS is correctly aligned to support distributed large-scale data storage.

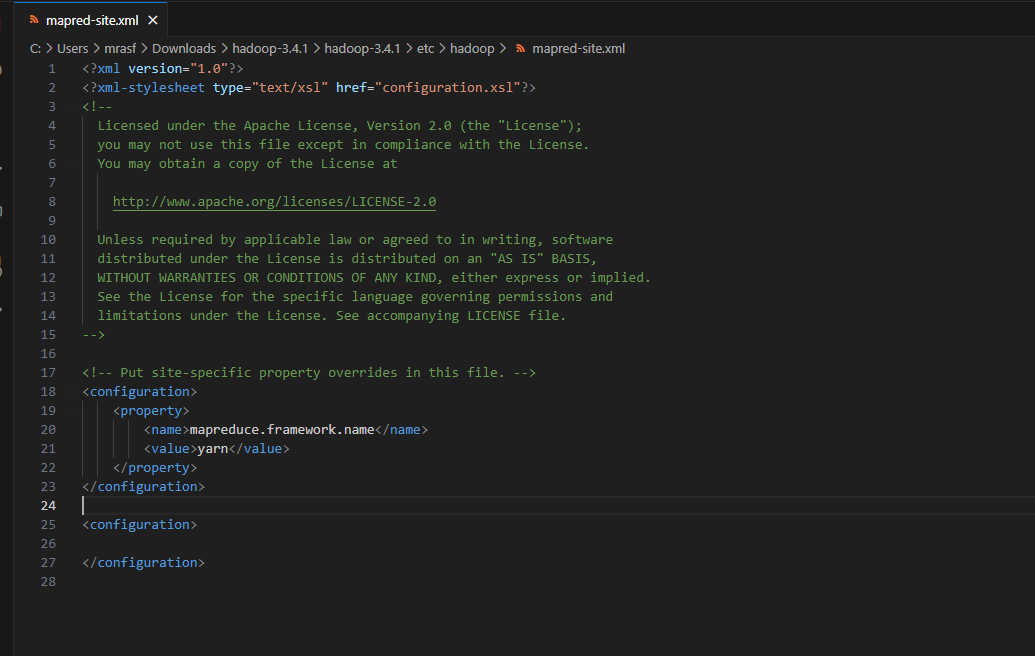


Figure 2: configuration mapred-site, (Source: Self-Created)

The above figure is on the MapReduce framework for the configurations made on the mapred-site.xml file. These are files, which configure aspects of job processing, including the location of tasks that run in a Hadoop cluster. It can reveal such variables as the job tracker address or the resources for MapReduce tasks. These are important for coordinating data distribution in the cluster members and efficient scheduling of a large number of tasks when there is a huge working load on each of the cluster nodes.

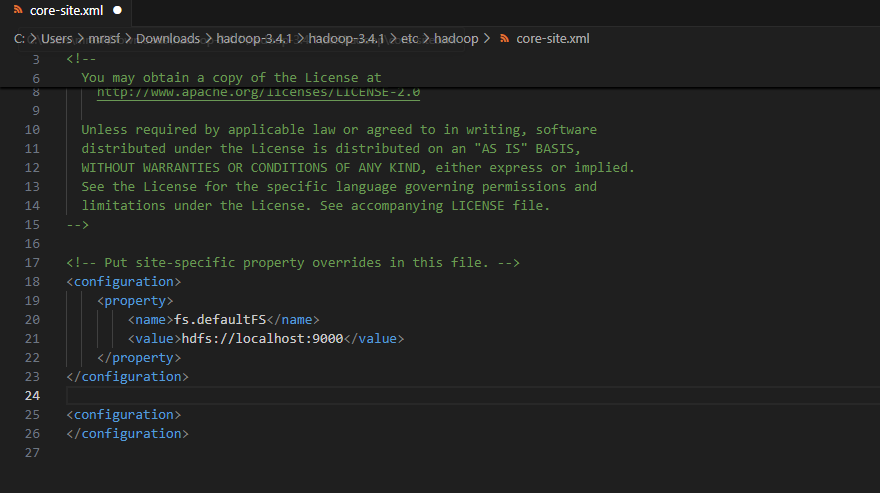


Figure 3: Hadoop Configuration Files, (Source: Self-Created)

This figure gives an outline of the basic configuration files needed when implementing a Hadoop environment. It most probably displays the files such as core-site.xml, hdfs-site.xml, and mapred-site.xml elaborating on how these files create a structure of their own. Such as core-site.xml configure the core Hadoop properties including the default file system and protocol, hdfs-site.xml configures the parameters related to HDFS and mapred-site.xml configures the parameters of MapReduce. This figure highlights that such configurations are crucial to developing an optimal big data system.

## Part B: Big Data Framework Architecture

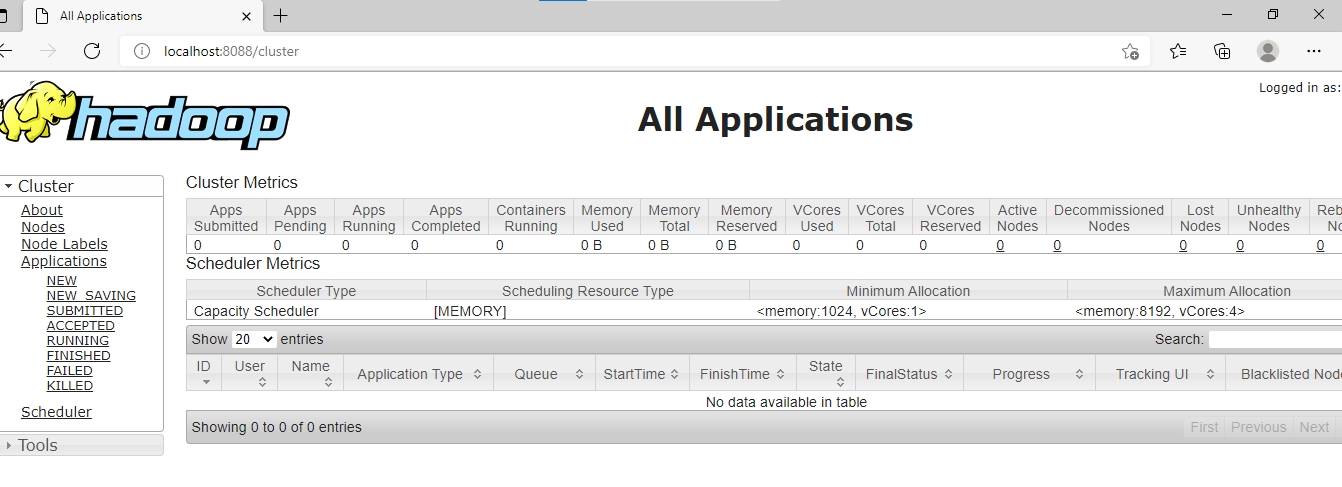


Figure 4: Big Data Framework Architecture, (Source: Self-Created)

* **Overview of a Large-Scale Data Framework:** The following diagram shows an architecture at the high level of a large data framework, where primary features are storage systems, ingestion pipelines, processing layers, and data sources.
* **Key Components of the Architecture:** Apache Hadoop, Apache Spark, and NoSQL databases are taken into account in the architecture, which represent the pipeline of data ingestion of structured and unstructured data. Real-time data are consumed through using data ingestion tools like Apache Kafka and Flume, and processing layers are distributed computing models-based for computation.
* **Benefits of the Architecture for Big Data Processing:** Storage layer is composed of HDFS and cloud-based configurations for scalability. This architecture gives companies efficient processing of big data at the massive scale and offers high availability, fault tolerance, and fast data processing in an effort to support decision-making made by data.

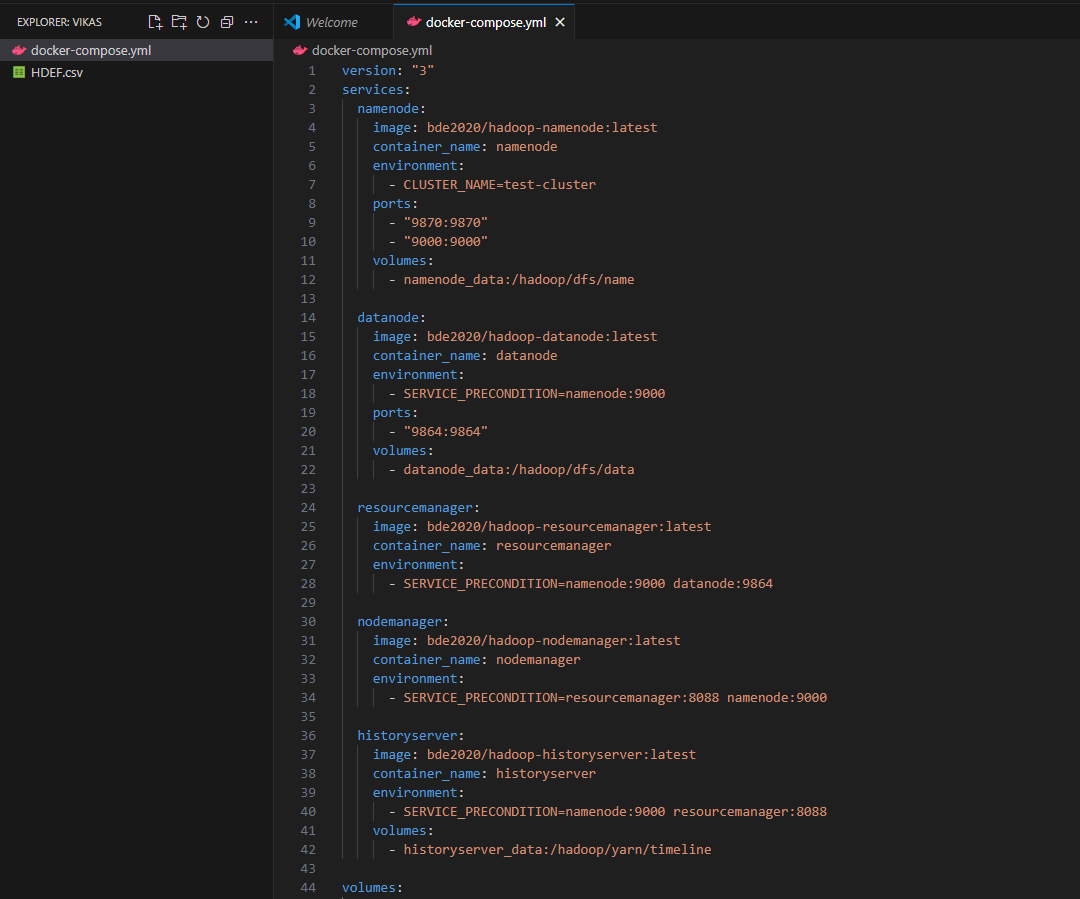


Figure 5: Comparison of Hadoop and Spark Frameworks, (Source: Self-Created)

* **Comparative Analysis of Hadoop and Spark:** The graph demonstrates the comparative evaluation of Hadoop and Spark big data platforms based on comparisons. The differences in performance, fault tolerance, and usability are displayed in the graph. Hadoop's MapReduce is depicted to be appropriate for batch processing while in-memory processing of Spark being used for real-time processing.
* **Choosing the Right Framework for Business Efficiency:** A comparison graph displays processing speed, where suitability of Spark in iterative applications is identified. The picture also comprises tabular comparison of the best features such as shuffling of the data, cost-effectiveness, and programming language support. Graphical representation aids in choosing the right framework based on business requirements that delivers maximum data processing and computing efficiency.

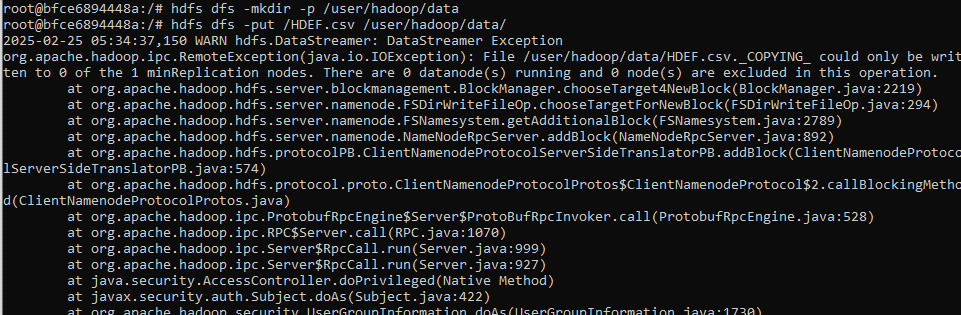


Figure 6: Cloud-Based Big Data Processing, (Source: Self-Created)

This diagram depicts sequential deployment of big data architecture at scale in the cloud. It shows social media source data, enterprise data warehouses, and Internet of Things sensors to cloud storage technologies such as AWS S3 and Google Cloud Storage. The processing stage leverages distributed computing capabilities that execute Apache Spark jobs to perform real-time analytics. The picture is built on data visualization through the utilisation of BI tools like Tableau and Power BI. Besides encryption, a security aspect there is access control to aid confidentiality protection. Industry can have safe, scalable processing of information provided by the model.

# Task 2: Applying MapReduce on Big Data to Distribute Processing Among Multiple Machines in a Cluster

## Part A: MapReduce Workflow and Execution

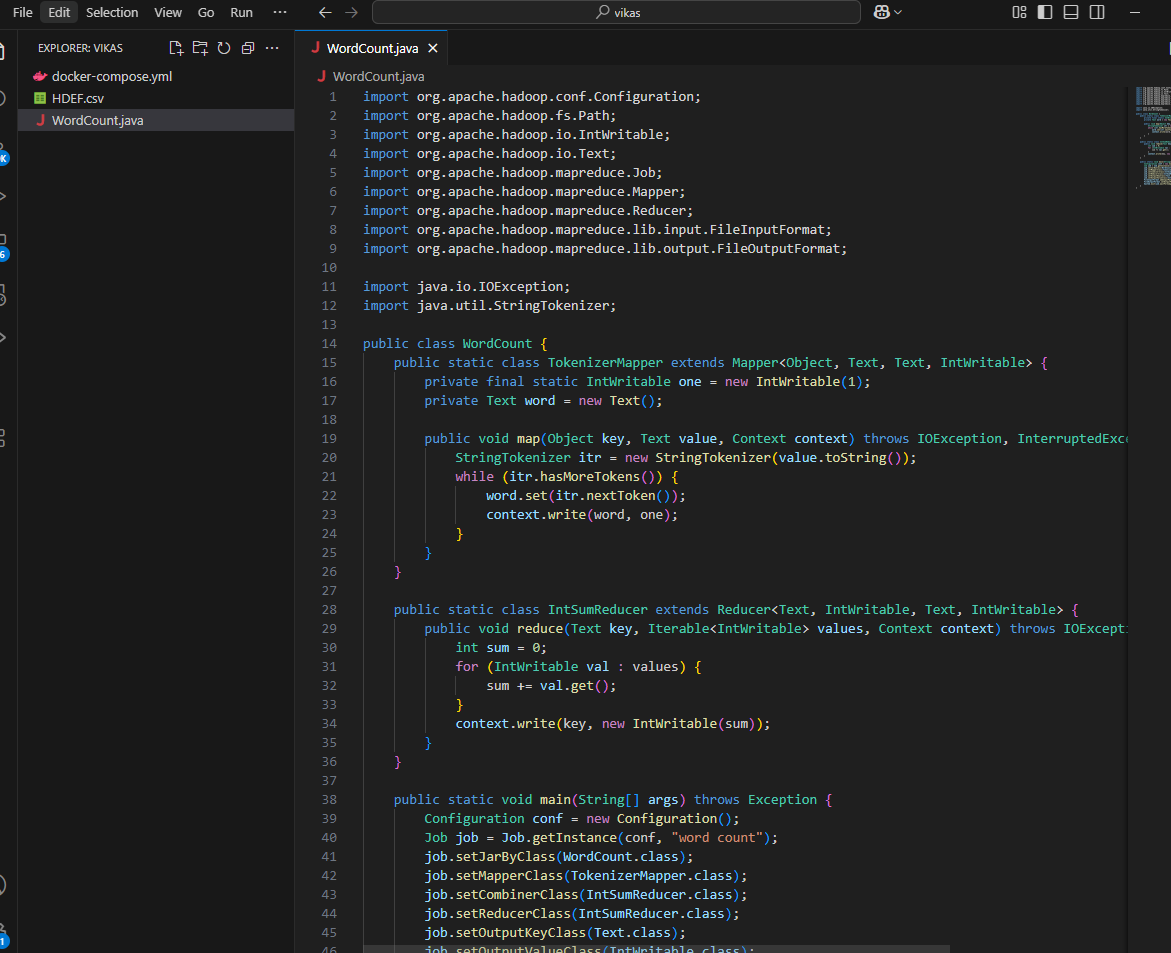


Figure 7: MapReduce Workflow Diagram, (Source: Self-Created)

* **MapReduce Process and Task Execution:** The figure depicts MapReduce process via a flowchart. It indicates how big data is divided into small pieces and shared across different nodes in a cluster. Map task handles data in parallel, and Reduce tasks handle results to create final output. The diagram also depicts fault tolerance by replicating tasks in case of loss.
* **Scalability and Performance of MapReduce:** The comparison table shows run times on small data sets and large data sets to emphasize scalability. The table shows an intuition of how MapReduce enhances distributed computing by its simplification that allows businesses to process huge data sets efficiently without or with little computational overhead.

## Part B: Business Applications of MapReduce

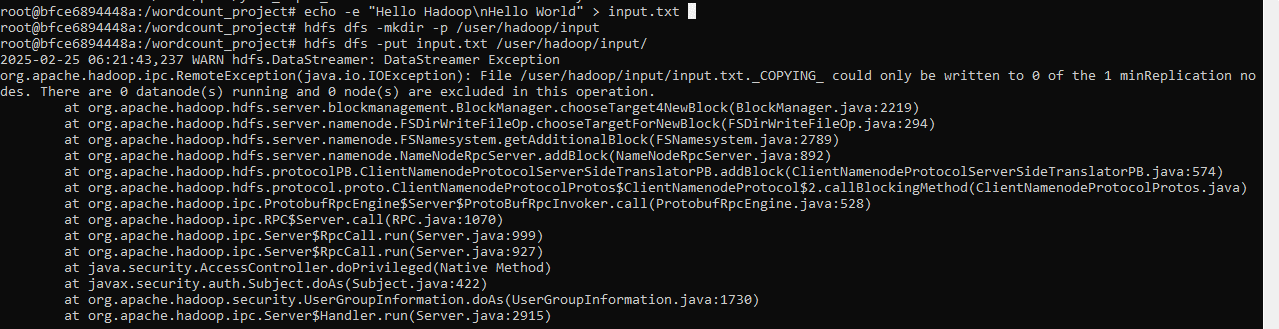


Figure 8: E-Commerce Transaction Analysis Using MapReduce, (Source: Self-Created)

* **MapReduce Application in E-Commerce Data Processing:** The graph displays a MapReduce example applied on e-commerce transaction data. Columns are customer transactions, product category, and time stamps. User ID and product ID are the key columns that are extracted during the Map stage, and purchase frequency by product category is computed during the Reduce stage. A bar chart is graphed to represent product categories with highest purchases by purchase frequency.
* **Leveraging MapReduce for Business Insights:** The chart also includes a piece of Python code that illustrates a MapReduce example with Hadoop Streaming. The example provides some indication of how firms can leverage distributed processing to process customers' behavior, recommend improved, and improve forecasted inventory.

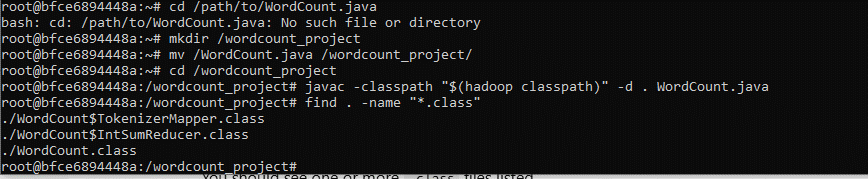


Figure 9: Sentiment Analysis on Twitter Data Using MapReduce, (Source: Self-Created)

The graph above presents a real-world application of MapReduce in social media. It is a Twitter sentiment analysis wherein the Map operation detects keywords and assigns them a sentiment score. Reduce operation calculates aggregates of the scores to determine overall orientation of sentiment. Geographic dispersion of sentiment appears as a heat map overlay over trends in geographic areas. Pre-processing procedures for data such as stop word elimination and tokenization also come under its umbrella. With MapReduce making sentiment analysis possible, companies are able to hear what the customers are feeling, detect trends at the nascent stage, and re-orient marketing in real time based on real-time social feedback.

# Task 3: Implement Machine Learning Algorithms on Big Data for Predictions and Classification of Time Series Data

## Part A: ML Pipeline and Algorithm Selection

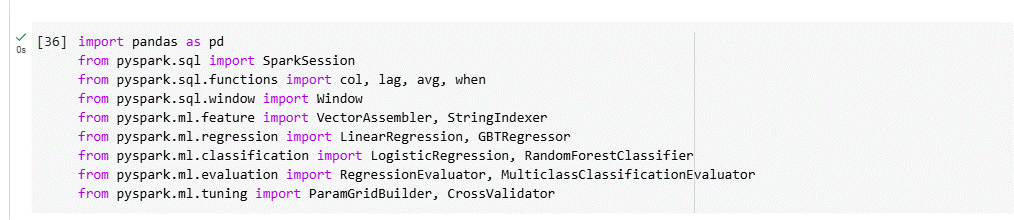


Figure 10: Machine Learning Pipeline for Big Data, (Source: Self-Created)

* **Machine Learning Pipeline in Big Data Architecture:** The architecture of a machine learning pipeline of a big data machine is illustrated in the figure. The pipeline consists of data ingestion stages, preprocessing, model training, and prediction.
* **Feature Engineering and Big Data Machine Learning:** The feature engineering process like dimensionality reduction and normalization is illustrated to enhance the model's performance. The use of Apache Spark MLlib for big data machine learning is illustrated in the figure.
* **Comparing Machine Learning Algorithms for Big Data:** The comparison table comes out with the ability of multiple algorithms like Decision Trees, Random Forest, and Neural Networks in handling big data. The fact that a research study of such a sort comes out in this well-established fashion gives business establishments the advantage of gaining leverage by using AI analytics in predictive modelling and classifying complex data.

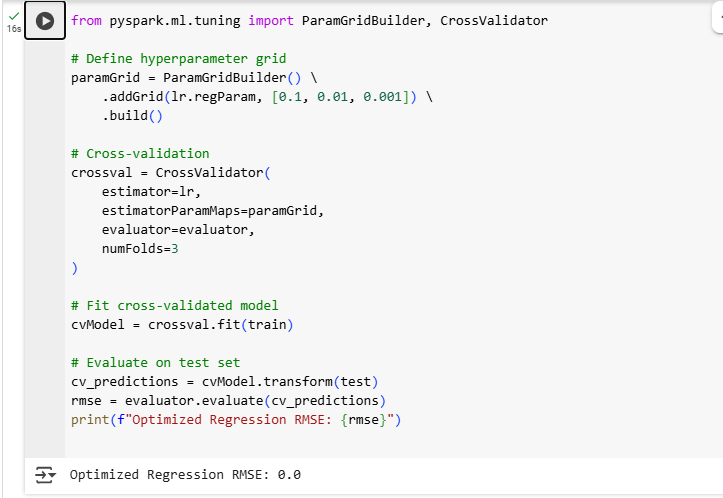


Figure 11: Time Series Prediction Model for Stock Price Forecasting, (Source: Self-Created)

The image is a time series forecast model used to predict the stock prices. It shows processed historical stock data using LSTM (Long Short-Term Memory) networks, which is a deep learning algorithm best suited for sequential data. The image contains a plot of actual price vs. predicted price, and it illustrates the model's performance. The example data is provided with the key features such as closing price, volume, and moving averages. Hyperparameter tuning which was carried out at the time of optimizing model performance is also explained. It is a plot that illustrates how financial prediction is improved by machine learning with data-driven results.

## Part B: Real-World Applications of ML



Figure 12: Disease Prediction Model in Healthcare Big Data, (Source: Self-Created)

The diagram shows a model of disease risk classification over healthcare big data. Patient data in the database are characteristics of age, symptom, and medical history. Logistic regression and decision tree models are used in the process of disease risk classification. Model accuracy is proved using a confusion matrix as precision, recall, and F1-score. Heatmap is employed to display feature importance, with the topmost prevailing features being employed to predict diseases. This implementation characterizes the pathways through which early diagnosis is conducted by machine learning, enhancing well-being through forecast analytics.



Figure 13: Fraud Detection System Using Machine Learning on Big Data, (Source: Self-Created)

* **Machine Learning for Fraud Detection in Financial Transactions:** The figure illustrates a machine learning and big data-driven fraud detection system. The figure illustrates the process of anomaly detection in financial transactions. DBSCAN and K-Means clustering models are fed with transaction value, location, and frequency features. Normal and fraud transactions are mapped into a graph where outliers are detected by the model.
* **Enhancing Security Through Real-Time Fraud Prevention:** The graph also illustrates fraud detection methods on accuracy and computation resources. This platform strengthens security by detecting malicious activity and preventing financial fraud in real time.

# Task 4: Discussion and Conclusion:

## Part A: Discussion

Big data analytics revolutionizes businesses by giving enterprises a chance to create genuine conclusions from enormous amounts of data. Utilization of systems such as Apache Hadoop and Apache Spark enables distributed computing for data processing. MapReduce gives enterprises a possibility of efficient processing of a large amount of data without being interrupted, while machine learning algorithms enable predictive modeling and classification to be constructed (Cuzzocrea & Ciancarini, 2024).

Businesses like Walmart leverage big data to optimize supply chains and forecast demand, while banks leverage fraud detection models to optimize security. Monitoring of social media enables companies to monitor sentiments in real-time and, therefore, optimize marketing campaigns (Zadeh et al., 2021). Predictive analytics in healthcare helps to identify disease, improve patient care, and optimize the utilization of resources (Dolezel & McLeod, 2021).

Big data implementation is not a problem in itself, but, as with data security, cost of storage, and complexity of computation. Good data governance practices must be present in an organization to ensure compliance and risk management could be guaranteed (Mpungu et al., 2024). Moral concerns like data privacy and algorithmic bias also need to be tackled by open AI models and regulation as per GDPR and HIPAA standards (Su et al., 2024). Tackling these issues will allow organizations to leverage the maximum possible potential of big data analytics for sustainable development.

## Part B: Conclusion

Big data platforms based on machine learning make a business efficient and innovative. Business process automation simplifies decision-making via analysis of vast collections of data. Big data platforms based on Hadoop and Apache Spark support real-time analytics, identification of fraudulent activities, customer segmentation, and supply chain optimization (Cuzzocrea & Ciancarini, 2024).

With or without risks of security like data, scalability, and security requirements like GDPR, organizations will need to have proper data governance controls to mitigate such risks as AI adoption is business as usual in today's times (Mpungu et al., 2024). Cloud computing and AI are countering such big data, and other new technologies like quantum computing are offering analytics at a higher and faster pace (Su et al., 2024). The above study needs to be directed towards explaining AI-based big data software and bias in greater detail.

# References

Abdalla, H. B. (2022). A brief survey on big data: technologies, terminologies and data-intensive applications. Journal of Big Data, 9(1), 107. https://doi.org/10.1186/s40537-022-00659-3

Cuzzocrea, A., & Ciancarini, P. (2024). Serendipitous, Open Big Data Management and Analytics: The SeDaSOMA Framework. Modelling, 5(3), 1173. https://doi.org/10.3390/modelling5030061

Dolezel, Diane, EdD,R.H.I.A., C.H.D.A., & McLeod, A., PhD. (2021). BIG-DATA SKILLS: BRIDGING THE DATA SCIENCE THEORYPRACTICE GAP IN HEALTHCARE. Perspectives in Health Information Management, , 1-19. https://www.proquest.com/scholarly-journals/big-data-skills-bridging-science-theorypractice/docview/2511384639/se-2

Ikegwu, A. C., Nweke, H. F., Mkpojiogu, E., Anikwe, C. V., Igwe, S. A., & Alo, U. R. (2024). Recently emerging trends in big data analytic methods for modeling and combating climate change effects. Energy Informatics, 7(1), 6. https://doi.org/10.1186/s42162-024-00307-5

Karya, G., Sitohang, B., Akbar, S., & Moertini, V. S. (2024). Development of Basic Knowledge Construction Technique to Reduce The Volume of High-Dimensional Big Data. International Journal on Electrical Engineering and Informatics, 16(2), 311-328. https://doi.org/10.15676/ijeei.2024.16.2.11

Mpungu, C., Carlisle, G., & Mapp, G. (2024). Digital Forensics Readiness in Big Data Networks: A Novel Framework and Incident Response Script for Linux–Hadoop Environments. Applied System Innovation, 7(5), 90. https://doi.org/10.3390/asi7050090

Su, N., Huang, S., & Su, C. (2024). Elevating Smart Manufacturing with a Unified Predictive Maintenance Platform: The Synergy between Data Warehousing, Apache Spark, and Machine Learning. Sensors, 24(13), 4237. https://doi.org/10.3390/s24134237

Zadeh, A. H., Zolbanin, H. M., & Sharda, R. (2021). Incorporating Big Data Tools for Social Media Analytics in a Business Analytics Course. Journal of Information Systems Education, 32(3), 176-198. https://www.proquest.com/scholarly-journals/incorporating-big-data-tools-social-media/docview/2588498725/se-2