# CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

|  |  |
| --- | --- |
| Module Title: | Advanced Data Analytics, Big Data Processing and Storage |
| Assessment Title: | Integrated CA 1 SEM 2 MSc Data Analytics |
| Lecturer Name: | David McQuaid  Muhammad Iqbal |
| Student Full Name: | Umar Iqbal |
| Student Number: | 2023463 |
| Assessment Due Date: | 5 April 2024 |
| Date of Submission: |  |

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Innovative Applications of Deep Learning and Big Data in Healthcare

***Abstract-* With advancement in technology and medicine, the healthcare facilities have improved a lot. The key player in the domain is data, which not only provide insights but also forecast the upcoming trends. The enormous amount of data which is generated each day in healthcare and other fields is so massive that it is impossible for traditional technology to evaluate and study. These type of data that falls in the category of Big Data is processed differently by the data scientists which create advanced algorithms to study it. These specialized algorithms are too complex that they behave like human brain, the scientists have created deep learning models which are known as neural networks which behave like neurons and make decisions appropriately. The intersection of both these fields provide a path for future developments that will not only revolutionize the healthcare but will change the core in all other domains. This paper provides an overview, applications, challenges and future aspects of Deep Learning and Big Data in Healthcare. The data storage management system, the architecture of distributed file system, big data stack and limitations of legacy approaches to big data are evaluated. The various different types of neural networks and set of requirements to determine the type of neural network are also discussed.**

***Keywords- Machine Learning, Deep Learning, Neural Networks, Big Data, Data Analytics, Medicine, Healthcare***

I. INTRODUCTION

In this decade, the healthcare industry has witnessed a drastic change driven by the advancements in technology especially in the areas of Artificial Intelligence and Machine Learning. Deep learning which is subset of Machine Learning and Artificial Intelligence has emerged as a powerful tool for extracting meaningful insights from large and complex databases. Similarly, Big Data, characterized by the massive volume, velocity and variety of healthcare data, presents both challenges and opportunities for transforming healthcare in a great way.

The collaboration of Deep Learning along with Big Data and healthcare with change the various aspects of healthcare including diagnosis of the disease,

treatment therapy, preventive measures, medicine and surgical intervention through bots and AI. The management sector will also be positively impacted by the data driven decisions. The medical sector generates many different categories of data ranging from structured, semi-structured and unstructured. The diagnosis facilities in the healthcare include studying images, graphs and visualizations. The mobile devices and wearable gadgets help keep record of vitals and generate daily logs incorporating huge amount of data which lays the foundation for studying the trends, recognizing patterns and make predictions via modelling which is the core of advanced data analytics.

II. BIG DATA

*A. Understanding the term*

The term Big Data is coined to demonstrate the enormous data which cannot be studied by the traditional data studying tools. To understand the complexity and magnitude of big data, the five core fundamentals volume, velocity, variety, veracity and variability plays an important role. The **Volume** represents the huge amount of data which is generated every single day. The media, sensors and devices creates terabytes of data which needs robust infrastructure to study. The **Velocity** at which the data flows also plays a crucial role in decision making. The transactions, streaming devices, online interactions generate a huge flow which needs to be studied for the insights.

The images, audio files, videos, spreadsheets and text files contribute to the **variety** of data that needs to be processed. This create a challenge for the data analysts to understand, manage and modify the data. The **Variability** refers to the fluctuations in the data generated over the time. The inconsistent nature of the data poses a problem for accurate predictions which needs to be addressed. In the vast sea of data, it is mandatory to check the quality of data being created which constitutes the **veracity** factor. It pertains to accuracy and reliability of data.

*B. Limitations of Legacy approaches to Big Data*

The traditional relational databases and on premises data warehouses have several limitations which include scalability, cost, performance, variety, security, processing speed and flexibility. The exponential growth of data needs horizontal scaling, but as the volume of the data increases, the capacity constraint becomes an issue. The upfront investment in hardware, maintenance cost of scaling the infrastructure to handle big data is expensive as compared to the cloud based services which have pay as you use price models. The traditional systems were optimized for structured data which pose a problem for different variety of data like images, texts and other semi structured or unstructured data. These data types are not supported by the relational databases. Processing these type of data require complex modeling and transformation. Another issue with legacy systems is it could not handle the velocity at which the big data is generated and processed. These systems were not able to handle real time analytics and decision making. The storage and modification of existing data also cause the lack in processing due to the time consuming, inflexible and rigid traditional systems. The system also lack the robust security features. Data stored in on premise centers may be susceptible to breaching and data loss.

*C. Big Data Storage and Management*

There are various processes and technologies involved in analyzing big data. The storage system has to handle enormous amount of data which traditional systems struggle to organize. Distributed storage systems like Hadoop, Apache Cassandra, Apache HBase, Google Cloud Storage and Azure Blob are designed to scale data across multiple servers and reduce fault tolerance. Technologies like Apache Spark, Flink and Kafka provide parallel processing capabilities to handle big data and process it. Machine learning and real time stream processing are some of the core functionalities. Cloud computing and containerization provides scalability by ensuring elasticity and dynamic resource allocation. Data management also includes managing the lifecycle of data by creating guidelines for data retention, deletion and archival. The utilization of storage resources is core of data lifecycle management. The metadata includes information about the structure, lineage and source. For better understanding, search, discovery and performance effective metadata management is necessary. The reason for data storage and management is to study the data to develop useful insights. Various dashboards like Tableau and Power BI are used to visualize, compare, analyze the data and generate reports. Data management and storage is a complex aspect of dealing with big data. Data Warehouses like Amazon Redshift, snowflake and Google Big Query are specialized systems designed to assist big data storage and management.

*D. Architecture*

Distributed file system (DFS) architecture is designed to store large amount of data across multiple nodes. This DFS is stored in a distributed computing environment. The typical architecture of a distributed file system includes Name node, Data node and Client node. The name node is responsible for storing the metadata about the file. It keeps track of which blocks constitute each file and their locations in the cluster. The actual data is stored in the data nodes which are known as the worker nodes. Each data node is responsible for serving read and write requests on its disk. The applications which interact with the distributed file system to read, write and manipulate data is referred as the client node. Client node send operational commands to the name node, which then coordinate with data node to perform the task. The client head may also store data in cache to improve performance and reduce traffic. The data is stored in blocks of fixed size typically 128 MB and there are multiple replicas of each block across the system to ensure there is no data loss and this feature is called replication. It is fundamentally important to make system redundant and fault tolerance.

*E. Big Data Stack*

There are various layers and each layer has many different technologies which form the stack of big data. The layers along with technologies are discussed below:

1. Ingestion and Collection:

* Apache Kafka: It is a distributed stream-processing platform used for building real time data pipelines.
* Apache Flume: Flume is a log collection service for collecting, aggregating and moving data chunks efficiently.
* Apache Nifi: It is a data flow management tool used in the process of automation of data ingestion.

1. Data Storage:

* Hadoop: Hadoop stores large datasets across multiple servers in a cluster. It is distributed file system for big data.
* Apache HBase: It is a scalable NoSQL database which allow real time modifications like read and write in large datasets.
* Amazon Simple Storage Service: S3 is developed by Amazon Web Services AWS for storing objects. It is scalable service for storing and retrieving data as objects.

1. Processing and Analytics:

* Apache Spark: It is a distributed computing engine used to process large scale data. It has various in memory capabilities which makes the computation fast.
* Apache Flink: Flink is a framework used to process live data streams with low latency.
* Apache Hive: Hive is made on top of Hadoop. It uses SQL like queries to analyze datasets. It is a data warehouse infrastructure.

1. Querying and Indexing:

* Apache Hadoop MapReduce: It is programming model which is used for parallel processing of large data sets.
* Apache Impala: Impala is open source SQL query engine used for parallel processing of data store in Hadoop clusters.
* Elasticsearch: It is a search engine that has analyzing capabilities. It is distributed system which offers real time searching.

1. Visualization:

* Tableau: A popular data visualization tool which help users to create interactive dashboards and reports.
* Power BI: It is a business analytics tool developed by Microsoft. It allows sharing insights across the organization. It is also used to visualize the data and generate reports.
* Apache Superset: It is an open source platform used for exploration and visualization. It supports various data sources and generates interactive charts.

1. Machine Learning and AI:

* Tensor Flow: It is developed by Google for building and training neural networks. An open source framework for machine learning and building models.
* Apache Mahout: It is an open source Hadoop library used for scalable algorithms.
* PyTorch: It is developed by Facebook for deep learning. This framework uses python programming language for computation.

1. Governance and Security:

* Apache Ranger: Ranger is a Hadoop framework for security and access control. It also offers centralized administration.
* Apache Atlas: It is used for managing and cataloging metadata. Atlas is management and governance platform.
* Apache Sentry: It is used with Hadoop for authorization and access to Meta data and data on cluster.

III. DEEP LEARNING

*A. Understanding neural networks*

The data analytics and machine learning focus on the detailed and comprehensive study of data which is possible using deep neural networks. The deep learning which is subset of machine learning, is inspired by the functioning of human brain, with layers of interconnected nodes which act as neurons to process the data. Deep neural networks have multiple hidden layers along with input layer and the output layer. Each node has its own weight, threshold and is connected to another node. The weight and threshold are continually adjusted during training to get the similar output as training data. If the output of any node is more than its threshold, the node send the data to another node thus activating the next layer. Neural networks learn by training the data and their accuracy improves over time. They allow us to classify and cluster data at high velocity. Neural networks are powerful tool in artificial intelligence domain.

*B. Applications*

The Computer Vision, speech technology, natural language processing, recommendation systems are some of the applications of the Deep Learning. Object orientation, image classification, facial recognition and segmentation are few examples where neural networks are used in the domain of image processing and computer vision. The security surveillance intelligence systems are also designed based on neural networks. Neural networks are the core of natural language preprocessing which include the tasks such as language translation, text summarization and AI chatbots. Virtual assistants like Apple Siri and Alexa recommends the content using neural networks. The speech to text feature, voice dictation systems, voice security systems all works on the speech recognition technology.

The recommendation systems used by Amazon, Netflix, YouTube and Social media apps like Instagram for suggesting movies, content, posts and music are powered by neural networks which work on real time data. Autonomous vehicles perceives, interprets the environment and collect data which is then analyzed by deep neural networks for object detection, path planning and various other decision making scenarios. Robotics use deep learning algorithms for service bots, industry bots, drones and humanoid bots. They navigate and learn by experiencing their surroundings. The healthcare sector rely on neural networks for disease prediction, medical image analysis like MRI and CT scans. Finance sector also use neural networks for predicting and identifying patterns, fraud detection, risk analysis, credit scoring, algorithmic trading and financial forecasting.