

BIG DATA (BDA) FOUNDATION

(by Futureskills prime-Nasscom & Digital Vidya)

An Industrial Internship Report

submitted by

SHARADINDU ADHIKARI

(19BCE2105)

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING



VIT[®]

Vellore Institute of Technology

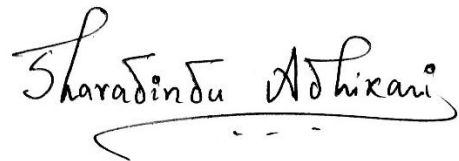
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

NOVEMBER 2021

DECLARATION BY THE CANDIDATE

I hereby declare that the Industrial Internship report entitled “**BIG DATA (BDA) FOUNDATION**” submitted by me to Vellore Institute of Technology, Vellore in partial fulfilment of the requirement for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide industrial training undertaken by me under the supervision of **Mr. Hitesh Gupta, Futureskills prime-Nasscom**. I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.



Name: Sharadindu Adhikari

Reg. Number: 19BCE2105



VIT[®]

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

BONAFIDE CERTIFICATE

This is to certify that the Industrial Internship report entitled “**BIG DATA (BDA) FOUNDATION**” submitted by **SHARADINDU ADHIKARI (19BCE2105)** to Vellore Institute of Technology, Vellore in partial fulfilment of the requirement for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide Industrial Internship undertaken by him under my supervision. The training fulfils the requirements as per the regulations of this Institute and in my opinion, meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Signature of the Supervisor

SUPERVISOR

Date:

Date:

<Signature>

Internal Examiner(s)

<Signature>

External Examiner(s)

ACKNOWLEDGEMENT

I hereby acknowledge that I have reviewed the recorded presentation, reviewed information on the informational website and received information on the course “Big Data (BDA) Foundation” in its entirety. I agree to abide by the principles that were explained in this training. I understand that if I have any questions about the training, materials presented or information not addressed in the training, or if I encounter any problems, it is my responsibility to seek clarification from the designated Human Resources Liaison and/or Human Resources.

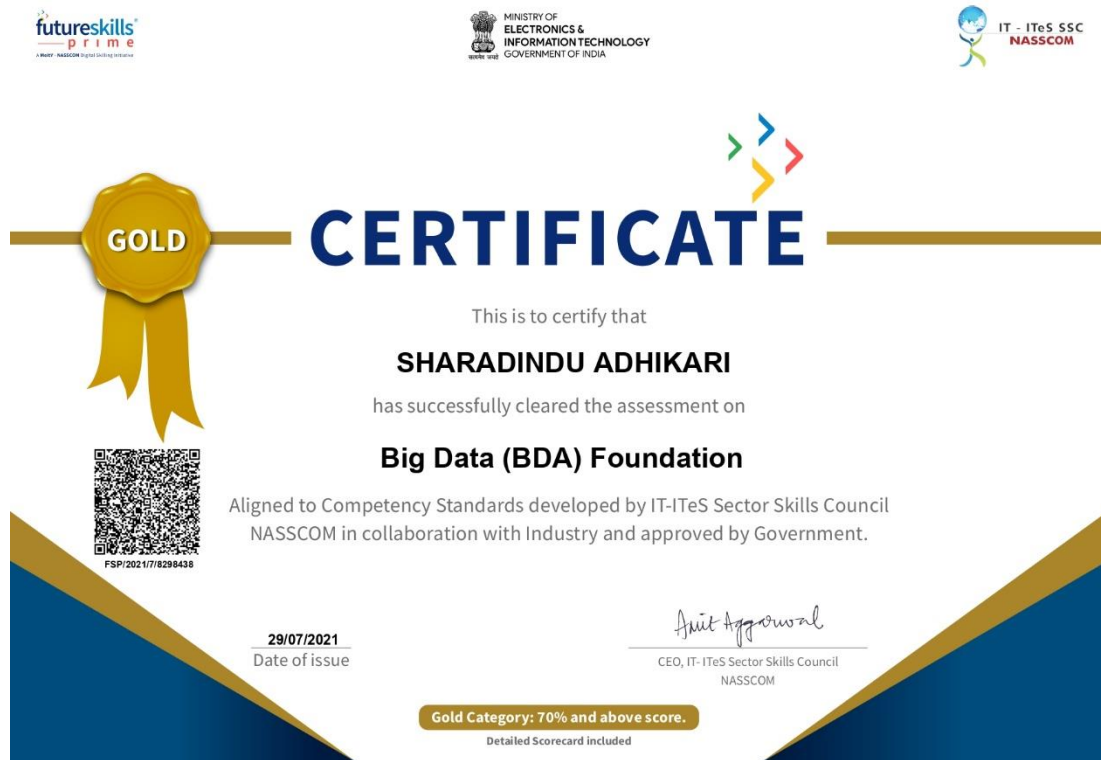
Place: Vellore

Sharadindu Adhikari

Date: 21/11/2021

CERTIFICATE

Link: <https://fsp-assessment-certificates.s3-ap-southeast-1.amazonaws.com/SharadinduAdhikari-64971411.pdf>



Certificate Details

Candidate Name	Sharadindu Adhikari
Assessment/Course Name	Big Data (BDA) Foundation
Date of Issue	29/07/2021
Certification ID	FSP/2021/7/8298438
Category Gold >=70% / Silver 60%-69% / Bronze 50%-59%	Gold



FSP/2021/7/8298438

Assessment Score

Module Name/NOS ID	NSQF Level	Maximum Marks	Marks Obtained	Percentage
M001	NA	12.00	12.00	100.00
M002	NA	35.00	31.00	88.57
M003	NA	53.00	37.00	69.81
Total		100.00	80.00	80.00



TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	LIST OF TABLES	ix
	LIST OF FIGURES	ix
1.	INTRODUCTION	1
	1.1. Synopsis	1
2.	ABOUT THE COMPANY	2
	2.1. Introduction	2
	2.2. Nasscom	2
	2.3. Digital Vidya	2
3.	SKILLSET	3
4.	KNOWLEDGE ACQUIRED FROM IN-PLANT TRAINING	5
	4.1. Module1	5
	4.1.1. Big Data Introduction	5
	4.1.2. History of Big Data	5
	4.1.3. Infrastructure of Big Data	6
	4.1.4. Marketing of Big Data	6
	4.1.5. Big Data Opportunities	6
	4.1.6. Project Plan of Big Data	7
	4.1.7. Intelligence Analyst and Data Scientist	7
	4.2. Module 2	8
	4.2.1. Data Storage	8
	4.2.2. Hadoop Introduction	9
	4.2.3. Hadoop Map Reduce	10

	4.3. Module 3	13
	4.3.1. Big Data Preprocessing	13
	4.3.2. Mongo Database	14
	4.3.3. Big Data Pipelines	18
5.	APPLICATION OF THE GAINED KNOWLEDGE IN THE TRAINING	21
6.	SELF - EVALUATION	23
7.	CONCLUSION	24

LIST OF TABLES

S.NO.	TABLE	PAGE NO.
1.	Difference between Data Warehouse and Data Lake	20

LIST OF FIGURES

S.NO.	FIGURES	PAGE NO.
1.	Indexing	16
2.	MongoDB Management Service	16
3.	Load Balancing	17
4.	Sharding	17
5.	Evolution of Data Pipelines	19
6.	Making a Hadoop File Directory	21
7.	Using the map reduce function available in hadoop	21
8.	Execution of map reduce function	21
9.	Map Reduce Framework	22
10.	Output after map reduce operation	22

1. INTRODUCTION

1.1 SYNOPSIS

This internship report consists of a thorough description of the 1 month (33 days) internship program I did during my summer vacation. It consists of the skills I gained from my course curriculum and training and also discusses the workflow, project architecture and all the knowledge that was imparted to me throughout the course. The program helped me gain experience in the field of Big Data Analytics. It improved my thinking and reasoning abilities in big data processing and management, provided me with ways which will improve my big data analytic skills. It helped me in understanding different big data tools such as Apache Hadoop, MongoDB and other tools. There were 3 modules in the course and each module covered a specific topic that will be used in Big Data Analytics. Big Data has increasing demand in this era and many big MNCs need Data Scientists. Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions. With the growth in the Internet of Things, data streams into businesses at an unprecedented speed and must be handled in a timely manner. This report also contains the assignment problems that were solved by me.

2. ABOUT THE COMPANY

2.1. INTRODUCTION

This section will go into detail about the certification course provider. It will also highlight the key technologies and directions of each of the specific organizations.

2.2. NASSCOM

NASSCOM, a not-for-profit industry association, is the apex body for the 194 billion dollar IT BPM industry in India, an industry that has made a phenomenal contribution to India's GDP, exports, employment, infrastructure and global visibility. In India, this industry provides the highest employment in the private sector. Established in 1988 and ever since, NASSCOM's relentless pursuit has been to constantly support the IT BPM industry, in the latter's continued journey towards seeking trust and respect from varied stakeholders, even as it reorients itself time and again to remain innovative, without ever losing its humane and friendly touch. NASSCOM is focused on building the architecture integral to the development of the IT BPM sector through policy advocacy, and help in setting up the strategic direction for the sector to unleash its potential and dominate newer frontiers. NASSCOM's members, 3000+, constitute 90% of the industry's revenue and have enabled the association to spearhead initiatives at local, national and global levels. In turn, the IT BPM industry has gained recognition as a global powerhouse.

2.3. DIGITAL VIDYA

Digital Vidya started delivering Social Media training workshops across India in 2009. After a successful stint, we decided to launch a full fledged certified digital marketing course in 2013, which achieved tremendous acceptance, thereby making Digital Vidya the leaders of online training in Asia. In 2017, Digital Vidya started Data Science & Analytics training programs holding on to our legacy of being forward thinkers & qualified educators. The excellent quality of training & education we have provided has led us to get partnered with Google, Microsoft, LinkedIn, Facebook, Vskills, & NASSCOM. Continuing the tradition of providing training on cutting-edge technologies, we have now ventured into Cyber-Security & Full-Stack Development training. As early stage professionals, we have a knack of training Corporates & professionals including CXOs.

3. SKILLSET

As the program was completed in between my 4th and 5th semester, I had become well-versed in a variety of core computer science fundamentals and concepts through the subjects offered by the University. I had also gained experience and skills through my personal projects and participation in various competitions organized by the university. Some of the major skill-sets possessed prior to training include:

- **Programming** - The core university subjects allowed me to become well-versed in multiple programming languages and concepts. I was adept enough in C, C++, Java, and Python prior to the program and had also been aware of the pillars of object-oriented programming and data structures. It made me well-versed in different programming concepts such as conditional statements, iteratives (for, while, do-while), function creation etc.
- **Data Structures and Algorithms** - This course helped me gain knowledge about the various data structures that are used in the current industry (stacks, linked lists, queues, heaps) and the process to formulate and manipulate them. It also helped me learn graph theory and some efficient algorithms for sorting, searching etc.
- **Database Management System** - This course helped me to understand the basics of databases and how to write SQL queries. Many concepts of the database were also taught such as database rollbacking, paging, different types of databases such as sql as well as nosql databases. The basics of making an Entity Relationship Diagram was one of the most important parts as it is necessary for making a database.
- **Internet of Things** - In this course different types of communications such as Zigbee, WiFi, GPS as well as different IOT protocols such as SCADA, MQTT, CoAP etc. I also learnt how to program microcontrollers so that we can take data from the sensors. There are various and large types of data that can be retrieved from the sensors which need to be managed. We learnt how

to do that and at the end we were also taught how to store the data in a cloud database.

- **Software Engineering** - In software engineering, many different process models of making projects were taught. This is one of the main courses that is necessary and needs to be documented before starting a project so that all the requirements and other things would be known beforehand. There should also be verification and validation of the projects.
- **Discrete Mathematics** - This course helped me understand how the basic language formulation for machines take place. It helped me understand discrete logic, graph theory, hypothesis formulation, conclusion derivation, lattice graphs, coloured graphs etc.

The courses offered to me above allowed me to develop projects and showcase my learned knowledge through application. I was able to develop many personal projects in these courses.

4. KNOWLEDGE ACQUIRED FROM IN-PLANT TRAINING

4.1. Module 1

This module included a thorough introduction about Big Data, Big Data systems, the advantages of Big Data, the problems with traditional data and how big data seeks to solve them.

4.1.1. Big Data Introduction

In this topic, first we studied about what is Big Data and the five Vs of Big Data, these are volume, velocity, variety, veracity and value of data. Volume describes the quantity of the data, variety describes the amount of diversity, veracity describes the authenticity of the data, velocity describes the rate at which the data is collected and value describes the new information that can be obtained from the data. Following this, the instructor describes the problems with traditional large-scale systems that were used in the industry before Big Data. We learned about the architecture of traditional grid computing and its bottlenecks; the architecture involved a brief detour on the field of distributed computing and its characteristics. Next, we were introduced to the Big Data grid computing architecture and how it differs from traditional grid computing.

4.1.2. History of Big Data

This topic starts with a brief introduction about the various software solutions available in the Big Data market since 2007 till 2016. Some of the important solutions, such as HDFS, Map Reduce, Spark, mentioned here are discussed in detail in upcoming modules. After this introduction, the course shifts to a detailed analysis of compliance, opportunity and need of Big Data in various industries. The aggressive adoption of Big Data technologies in the the financial sector (especially banks) were highlighted along with the various financial services, such as fraud detection, EDW optimisation and risk management, in which big data can be used. Healthcare is the next sector that was analysed, the main focus was on the emergence of IoT smart devices in healthcare and the data generated from them with less focus

on the role of big data in electronic health record and reducing frauds. Different use cases and their values were discussed at the end of this topic.

4.1.3. Infrastructure of Big Data

This section is about different considerations that need to be kept in mind when implementing big data solutions to different industries and the Big Data infrastructure. First, we were introduced with a six-step general guideline that must be kept in mind when implementing Big Data solutions.

Next, security and physical infrastructures of Big Data were explained. In physical infrastructure, performance, availability, scalability, flexibility and cost of different Big Data solutions were examined and explained. In security infrastructure, data access, application access, data encryption and threat detection features of Big Data solutions were examined.

4.1.4. Marketing of Big Data

This section first provides a list of Big Data analytic providers in the current. Next it focuses on the market analysis performed by Forrester Wave. It placed IBM, MapR technologies, Cloudera and Hortonworks as the market leaders and Pivotal Software as strong market performers. Finally different career paths and jobs in the Big Data field were examined and presented to us.

4.1.5. Big Data Opportunities

This section is about the business opportunities in the field of Big Data for both existing companies and upcoming startups, that were examined. First, we learned about the Big Data maturity models (BDMM), these provide a capability assessment tool and help guide development through milestones. The core components of BDMM are business monitoring, business insights, business optimisation, data monetisation and business metamorphosis. Business monitoring includes examining business trends and benchmarking business systems and shares to produce a report on the adoption of big data in the business. Business insights involves identifying problems and opportunities for Big Data in the business. Business optimization evaluates business insights from the previous step to better adopt the Big Data technologies. Data monetisation includes ways to utilize the information collected

from the Big Data into profit. Finally, business metamorphosis is the evolution of the business in time to better adopt Big Data technologies.

4.1.6. Project Plan of Big Data

This section involves understanding how organisations make money from Big Data, identifying key business initiatives for Big data, barnstorming Big Data business imports, fleshing out the use cases, proving the use cases and finally designing and implementing Big Data solutions. In identifying key business initiatives, we learn about roles and responsibilities of business stakeholders, key performance indicators, timeframe for delivery, critical success factors and desired outcomes and key tasks. Barnstorming Big data business impact involves detailed mining of data at the lowest level, integrating new unstructured data sources, providing low-latency data access and integrating predictive analytics. In designing and implementing Big Data solutions, data source and access requirements, data management and modelling capabilities, real-time data access capabilities, analysis requirements and user experience requirements are analysed.

4.1.7. Intelligence Analyst and Data Scientist

This section deals with different roles of BI analyst and data scientist in the field of Big Data. First, we differentiate between the task of BI analyst and data scientist, some of those differences are BI analyst focuses on hindsight analysis while data scientist focuses on foresight analysis, BI analyst reports on what happened while data scientist focuses on what will happen. BI analyst works with a determined single truth while data scientist works with different scenarios with different probabilities. In the next part of the section, the role of data scientist is further examined. These are, in the given order, data discovery, data preparation, model planning, model building and finally communicating the results to the management.

4.2. MODULE 2

Module 2 of the course is Big Data Fundamentals and Platforms. This module delves into the details of data warehousing, terminologies used in data warehouse and cap theorem.

4.2.1. Data Storage

Here, we learn that a data warehouse contains data from operational systems, such as Excel. It also consists of a data staging area and a presentation area.

Data in this context can be structured, semi-structured or unstructured. There are 3 types of data warehouses, enterprise data warehouse, operational data store and data mart.

We learn that the steps in data warehousing are a) to gather business data b) the gathered data is then cleaned and pre-processed through extraction transformation and loading process which is called ETL c) the data is then stored in a data warehouse and d) various business analysis tools are used to derive hidden pattern and customer behaviour insights from the data.

A modern data warehouse usually has heterogeneous sources of data such as server log files, no sql application data, social media alongside traditional relational data. The technologies used to store and process this data include Apache Hadoop that uses the power of parallel and cloud computing to handle large data sets.

Parallel computing and distributed computing are two terms that I learnt help in the processing of large data. Parallel computing refers to the execution of an application on several processors simultaneously to achieve faster results. Distributed computing refers to the execution of a process on several nodes that are located on different networks simultaneously.

We also learn the various components of Hadoop ecosystem are:

a. Hadoop distributed file system that stores data and acts as interface to hadoop

- b. HBase which is a no sql database management system to query the stored data
- c. Yarn is the operating system in hadoop environment that helps in job scheduling and cluster allocation
- d. Sqoop that is a connector between the file system and the database management system
- e. Apache Spark cloud computing framework for analyzing data
- f. Apache Flume is a distributed service that collects web log data and sends it to the hadoop file system
- g. Hadoop MapReduce is a program that takes a big data set, breaks it into small datasets and maps a function onto the small data parts. The reduce function then agglomerates these results
- h. Apache Pig helps in data manipulation and reuse of code
- i. Hive is used to query large datasets in the hadoop system
- j. Apache Drill is used to process no SQL databases
- k. Apache Zookeeper helps in synchronization and coordination
- l. Oozie is used for maintaining workflows in a hadoop cluster.

4.2.2. Hadoop Introduction

This section discussed the concepts of Apache Hadoop in detail such as its installation and administration. The course instructor first guides us on how to install apache hadoop on our systems and configure it to safe mode. I also learnt about other commands such as cat, chmod, copyFromLocal, copyToLocal, du, get.

Among these, some commands I learnt for the first time and their use include:

put: this command moves copies files from the local file system to the destination file system

get: copies files to the local file system

I also learnt that HiveQL translates user entered queries into MapReduce code and thus allows users to use MapReduce more easily. Hive can be used via the command line, web interface, hiver server and JDBC/ODBC

An example of a query in HiveQL is

create database if not exists Database_Aritri

The Pig Latin is a script designed for exploring large datasets. It has queries such as load, for each, filter, dump

Sqoop is a command line tool that allows us to import data from MySQL or MariaDB to HDFS easily.

4.2.3 Hadoop Map-Reduce

This section primarily focuses on understanding the concept of Map reduce.

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster

A MapReduce program is composed of a map procedure, which performs filtering and sorting (such as sorting students by first name into queues, one queue for each name), and a reduce method, which performs a summary operation (such as counting the number of students in each queue, yielding name frequencies)

“Map”, in simple words, is taking a big data set and dividing it into smaller and more concise datasets. After this a common function (map) is applied on each dataset repeatedly.

“Reduce” combines the output from all smaller sub datasets, which are the final outputs of the mapper function used in the “Map” stage.

The concept was further clarified by taking the “Word count” example. In the MapReduce word count example, the frequency of each word is found. Here, the job of Mapper function is to map the keys to the existing values and the role of Reducer is to aggregate the keys of common values. As a result, everything is arranged in the form of a key value pair.

The functioning of map reduce in the context of Big data was then explained. The Hadoop namenode has two primary components called the Job Tracker (present in the map reduce layer) and the NameNode part (present in the HDFS layer). The Job tracker keeps a record of all the incoming and submitted jobs and then divides each job into tasks and comes up with a plan to run each task by continuously listening and monitoring task trackers. The Task Trackers execute tasks by assigning multiple tasks per processing node. It also monitors the execution of each task and relays feedback to the Job Tracker in a continuous fashion.

The data flow of map reduce is a very efficient concept altogether. It primarily encompasses 6 phases:

- **Input Reader:** The incoming data is read by the input reader which is responsible for splitting it into data blocks of desirable size (64MB to 128MB). Each data block is packaged with a mapping function. Once the reader completely ingests the data, it generates the subsequent key-value pairs. The input files reside in HDFS.
- **Map function:** The map function then processes the upcoming key-value pairs and generates the corresponding output key-value pairs. The map input and output type may differ from each other.
- **Partition function:** It is the job of the partition function to assign the output generated from each map function and passes it to the appropriate reducer. The current key values are provided to the functions. It then returns the index of the reducers.
- **Shuffling and Sorting:** The data is then shuffled between/within nodes to

make it ready to process for reduce function. The shuffling of data seldom takes a lot of time in regards to computation.

- **Reduce Function:** The reduce function is then applied over each unique key. The keys that are provided as input to this function are already sorted. The corresponding values of the keys can loop through the reduce and generate the required output.
- **Output Writer:** After all the above phases are completed, the output writer starts its execution. The responsibility of Output Writer is to write the output generated by the Reduce function to a safe and stable storage.

To decide the total no. of mapped tasks, a number of factors have to be taken into consideration. It usually depends on the total size of inputs which is nothing but the no. of blocks of the input files. The right level of parallelism for maps seems to be around 10-100 maps per-node.

Hadoop is a fault tolerant system. Intermediate data between mappers and reducers are materialized to simple & straightforward fault tolerance.

In case a task fails, map reduce detects the failure, intimates a message to the job tracker which in turn re-schedules the task.

In the case of data node failure, both the namenode and job tracker detect the failure and all tasks assigned to the failed node are rescheduled. Finally the namenode replicates the user's data to another node. If a Namenode is down, the entire cluster stops functioning.

4.3. MODULE 3

This module tells us about the different types of big data processing, management and analytics.

4.3.1. Big Data Preprocessing

This section basically describes the process of data ingestion and how data is prepared for analysis. It is all about data storage and further its analysis, which is possible by using various Tools, Design Patterns, and few Challenges. In the era of the Internet of Things and Mobility, a huge amount of data is becoming available quickly. There is also the need for an efficient Analytics System. It includes the below mentioned steps:

- Extract which means taking the data from its current location
- Transforming the data by cleaning and normalizing the data
- The final step is the loading of the data in the database where it can be analysed

Data scientists must be prepared to expect difficulties and plan accordingly. Usually, they spend a lot of time on running algorithms, examining results and then refining them further for the next sequence of input. In reality, we need to wrangle the data into shape. As the size of data surges, this phase of the job becomes more and more cumbersome and time consuming.

The need of the hour is to automate the process of data ingestion. It is a known fact that data in recent times has grown too large, both in variety and size to be analysed manually. A user should be able to define information like metadata, schema and rules in a spreadsheet. This spreadsheet is then read by a tool that implements the specified metadata. This kind of automation can greatly reduce cumbersome tasks of data ingestion. However, it is often seen that it fails to address the problem of ingestion bottleneck considering the vast amount of tables involved. Logically, it is more optimal to fill out a lot of spreadsheets instead of writing the same amount of ingestion scripts. We can use artificial intelligence for the same.

Some examples of automation were also discussed. One option is to infer the global schema from the local tables mapped to it. The algorithm must be able to classify each local table into a global table in which it can be ingested. This is done by inferring synonyms during the data normalization phase. It also includes the detection of duplicate records.

One more solution is to make data ingestion a self-service task. This can be done by giving users self-service to cleanse and detect missing values, outlier values and duplicate records before they are aggregated into the global database.

It is also necessary to govern the data to ensure it is always clean. The task includes defining schema, cleansing rules and decisions to ingest particular data into data source, treatment of dirty data.

We also have to ensure data security standards are in place which in turn should be compliant with regulatory standards such as GDPR and master data protection and management.

Once the data source is cleaned, it is necessary to ensure that other users can find the data easily. It is equally important to ensure that the data is generalized so that it is relevant to as many people as possible rather than a single group. Organizations can also set up a publish-subscribe model, along with a registry of previously cleansed data available for lookup by all other users.

4.3.2. Mongo Database

There are many reasons for using MongoDB such as it contains aggregation framework, BSON format, sharding, ad-hoc query, schema-less, capped collection, indexing, file storage, replication and at last MongoDB Management Services (MMS).

Aggregation Framework - It consists of 2 different varieties of Sets and they are:

Map(): It performs operations like filtering the data and then performing sorting on that dataset.

Reduce(): It performs the operation of summarizing all the data after the map() operation.

BSON Format - BSON stands for Binary JSON. It is a JSON-like storage format. We can add data types like date and binary. BSON format makes use of `_id` as a primary key over here. The `_id` is being used as a primary key so it is having a unique value associated with itself called as ObjectId, which is either generated by application driver or MongoDB service.

Sharding - Scaling is a major problem in any of the web or mobile applications. To solve this issue MongoDB has added the sharding features. In this particular method the data is being distributed across various and multiple platforms. The sharding provides MongoDB with sharding features.

Ad hoc queries - There is a major difference between SQL and NoSQL.

SQL Statement – `SELECT * FROM Students WHERE stud_name LIKE '%ABC%';`

MongoDB Query – `db.Students.find({stud_name:/ABC/ });`

Schema-Less - The structure of the data to be stored is described by the schema. In the case of relational databases, the tables, its fields are defined in the schema. The relationship between each table and each of the fields is also kept in the schema. The data needs to be according to the schema, this is it has to comply with defined structure (tables, columns, data types and relations). Such that every register in a table has the same number of columns and format.

Capped Collections - MongoDB supports capped collections. This means that it has a fixed size of collections in it. An insertion order is being maintained. When the limit is reached it starts behaving like a circular queue. Example – Limiting our capped collection to 4MB.

```
db.createCollection('logs', {capped: true, size: 4194304})
```

*Here, $4 * 1024 * 1024 = 4194304$ bytes i.e. 4MB

Indexing - There is an indexing feature in MongoDB which helps in improving the performance of the searches, any field can be indexed either primary or secondary. This is particularly the reason why MongoDB's database engine can efficiently resolve the queries.

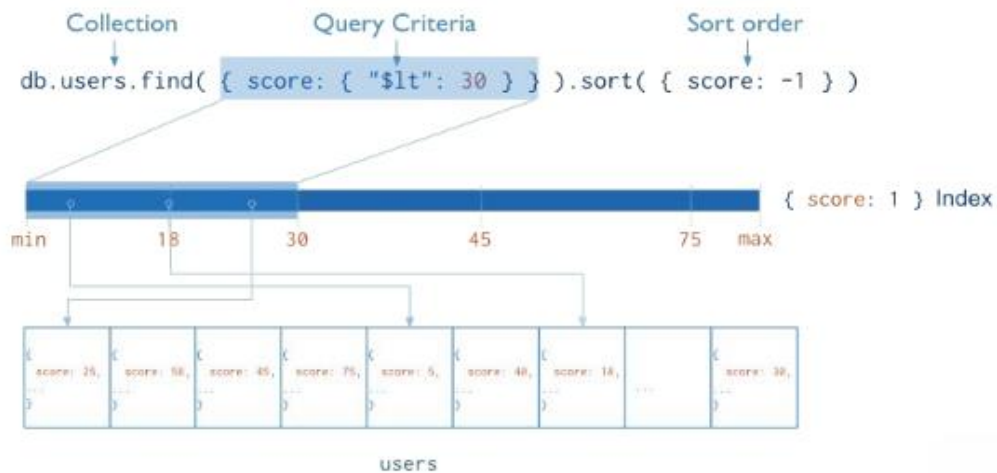


Fig.1. Indexing

File Storage - MongoDB has many specifications for storing and retrieving files that exceed the BSON-document size limit of 16 MB such as GridFS. It divides the file into chunks and then stores it accordingly.

Replication - This feature is being provided by distributing the data across different machines. It only has one primary node but it can have more than one secondary node in it. This basically acts like a master-slave. The master can perform the read and write and all the slaves copy the data from the master as a backup only for the read operation.

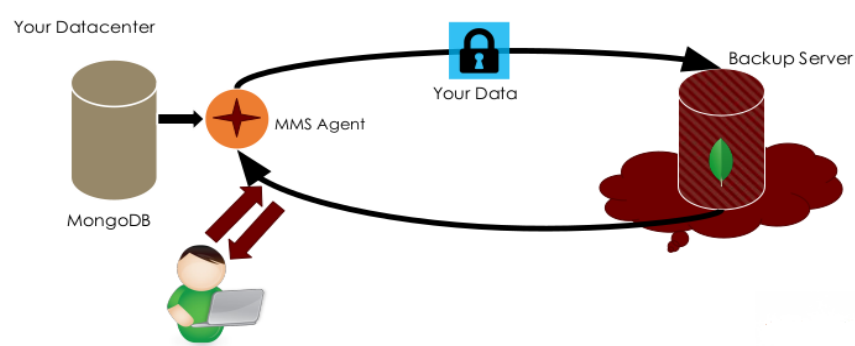


Fig.2. MongoDB Management Service

The benefits of MongoDB are load balancing, sharding, flexibility and speed.

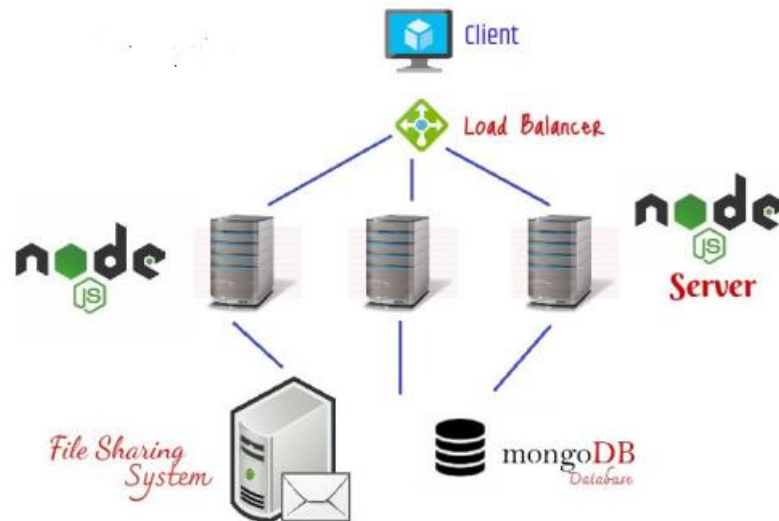


Fig.3. Load Balancing

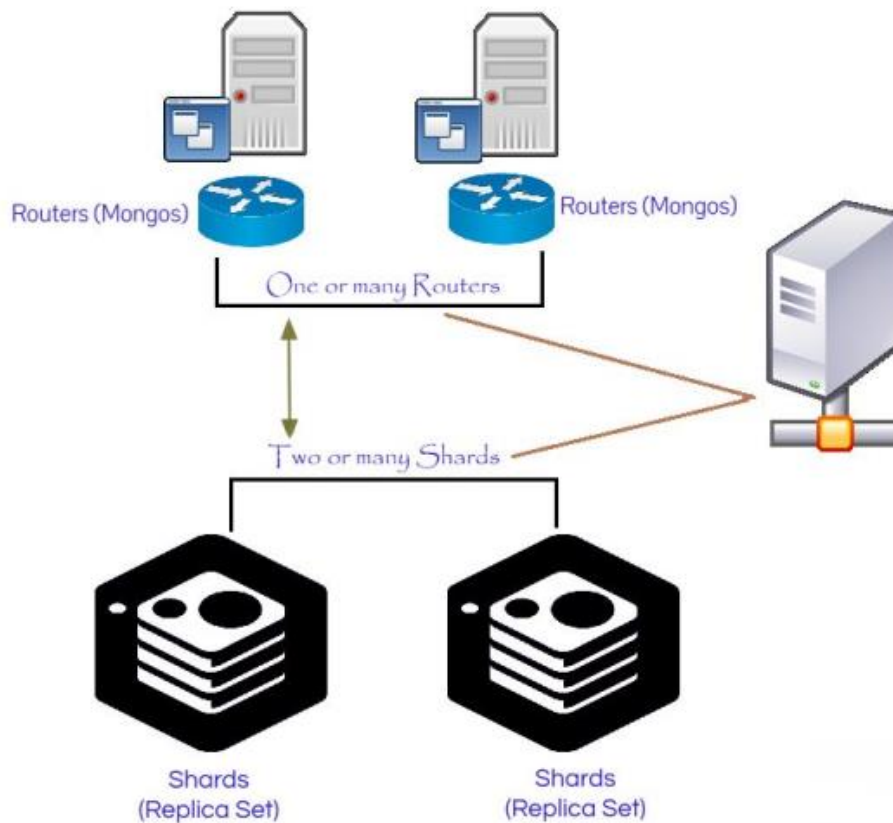


Fig.4. Sharding

Flexibility -

- Document Model
- Dynamic Schema
- Powerful Query Language
- Secondary Indexes

Speed - MongoDB can quickly and easily process the data, but is only valid until your data is in document format. The speed of the processing automatically increases as it is handling a large amount of unstructured data within seconds which feels like magic.

4.3.3. Big Data Pipelines

Data Pipelines - The core component of data science is building data pipelines at the startup. First we need to collect data and process. Data Lake is the destination of the data pipeline such as the Hadoop or the parquet files on S3, or a relational database such as Redshift. The data pipeline views all data as the streaming data and it allows the schema to be flexible. There is no need for an ultimate destination for a data pipeline to be a data warehouse. Basically the pipeline is a commonplace for everything related to data whether they need to ingest data, store data for analyzing the data.

The components of Big Data Pipelines are Compute, Storage and Messaging. There are many tools to get the data processed such as Hadoop MapReduce, Apache Storm, Apache Flink, Apache Heron and Apache Shark.

The storage components that are used by Big Data Pipelines are HDFS, S3 for other cloud filesystems, Local Storage and NoSQL Database.

The messaging components available are Apache Kafka, RabbitMQ and Apache Pulsar.

There are 3 types of data:

Raw Data - This includes the tracking Data and JSON.

Processed Data - This includes the decoded data and schema data.

Cooked Data - This includes the aggregated data and number of sessions.

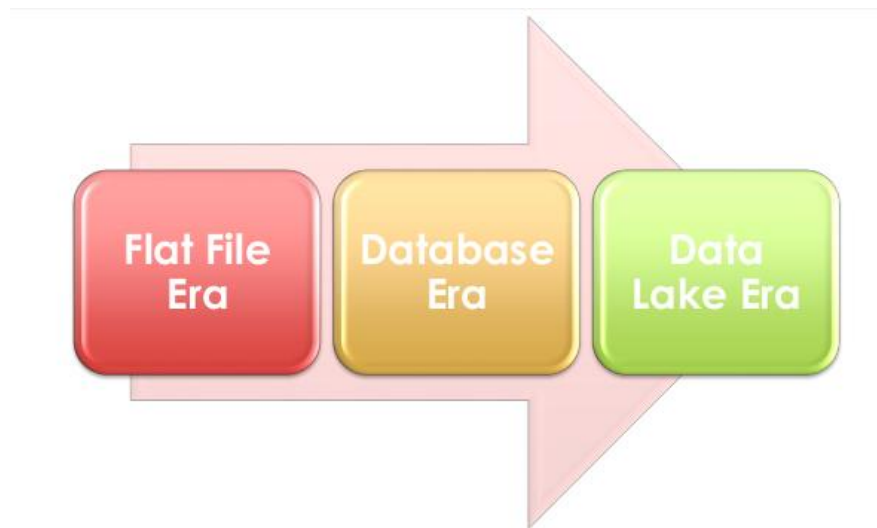


Fig.5. Evolution of Data Pipeline

Flat File Era - It basically stores the data in the plain text format. In the relational database the flat file includes one record per line. These are used widely in data warehousing projects to import the data. These are text documents which contain data separated by commas or tabs.

Database Era - As we know that in the relational database the data can be stored in the form of tables.

Data Lake Era - It is one of the concepts that has appeared in the big data era. The idea basically originated from business fields instead of the academic field. It is a newly conceived idea with a revolutionized concept and it brings many challenges for its adoption.

The properties of data pipelines are Low Event Latency, Scalability, Interactive Querying, Versioning, Monitoring, Testing.

Comparison	Data Warehouse	Data Lake
Data	Structured, processed data	Structured/semi-structured, unstructured data, raw data, unprocessed data
Processing	Schema-on-write	Schema-on-read
Storage	Expensive, reliable	Low cost storage
Agility	Less agile, fixed configuration	High agility, flexible configuration
Security	Matured	Maturing
Users	Business professional	Data Scientists

Table.1. Difference between Data Warehouse and Data Lake

The different layers in the Data Pipelines are:

- Data Ingestion Layer
- Data Collection Layer
- Data Processing Layer
- Data Storage Layer
- Data Query Layer
- Data Visualization Layer

The big data pipelines help in better event framework designing and the persistence of data. There is an ease of scalability at the coding end. There is workflow management because the pipeline is automated and it has many scalability factors. It provides the serialization framework. There are some disadvantages of data pipelines but they have many ways to manage that using other softwares. The economic resources may affect the performance because mainly the data pipeline is suited for large datasets only. The maintenance of job processing and cloud management is quite easy. There is no need for privacy on the cloud for critical data.

5. APPLICATION OF THE GAINED KNOWLEDGE IN THE TRAINING

```
bitnami@debian:~$ hadoop fs -ls /inputdir/*
-rw-r--r-- 1 hadoop supergroup 167 2021-11-21 10:29 /inputdir/datafile
bitnami@debian:~$ cat datafile
This practical application is done by Kaustubh.
This practical has been done according to the big data course.
This big data course from digital vidya was quite good.
bitnami@debian:~$ cat datafileddew_
```

Fig.6. Making a Hadoop File Directory

```
bitnami@debian:/opt/bitnami/hadoop/share/hadoop/mapreduce$ hadoop jar ./hadoop-mapreduce-examples-3.
B.1.jar wordcount /inputdir/datafile /output
2021-11-21 10:34:51,149 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManage
r at /0.0.0.0:8032
2021-11-21 10:34:51,913 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/
hadoop-yarn/staging/hadoop/.staging/job_1637490031909_0001
2021-11-21 10:34:52,330 INFO input.FileInputFormat: Total input files to process : 1
2021-11-21 10:34:52,439 INFO mapreduce.JobSubmitter: number of splits:1
2021-11-21 10:34:52,704 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637490031909_00
01
2021-11-21 10:34:52,711 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-21 10:34:53,153 INFO conf.Configuration: resource-types.xml not found
2021-11-21 10:34:53,156 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-21 10:34:53,659 INFO impl.YarnClientImpl: Submitted application application_1637490031909_00
01
2021-11-21 10:34:53,856 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/ap
plication_1637490031909_0001/
2021-11-21 10:34:53,873 INFO mapreduce.Job: Running job: job_1637490031909_0001
```

Fig.7. Using the map reduce function available in hadoop

```
2021-11-21 10:34:53,856 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/ap
plication_1637490031909_0001/
2021-11-21 10:34:53,873 INFO mapreduce.Job: Running job: job_1637490031909_0001
2021-11-21 10:35:07,497 INFO mapreduce.Job: Job job_1637490031909_0001 running in uber mode : false
2021-11-21 10:35:07,512 INFO mapreduce.Job: map 0% reduce 0%
2021-11-21 10:35:14,782 INFO mapreduce.Job: map 100% reduce 0%
```

Fig.8. Execution of map reduce function

```

Map-Reduce Framework
  Map input records=3
  Map output records=28
  Map output bytes=279
  Map output materialized bytes=271
  Input split bytes=104
  Combine input records=28
  Combine output records=22
  Reduce input groups=22
  Reduce shuffle bytes=271
  Reduce input records=22
  Reduce output records=22
  Spilled Records=44
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=107
  CPU time spent (ms)=1050
  Physical memory (bytes) snapshot=393715712
  Virtual memory (bytes) snapshot=6735962112
  Total committed heap usage (bytes)=328208384
  Peak Map Physical memory (bytes)=248717312
  Peak Map Virtual memory (bytes)=3364433920
  Peak Reduce Physical memory (bytes)=144998400
  Peak Reduce Virtual memory (bytes)=3371528192
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=167
File Output Format Counters
  Bytes Written=177
bitnami@debian:/opt/bitnami/hadoop/share/hadoop/mapreduce$ P_

```

Fig.9. Map Reduce Framework

```

bitnami@debian:/opt/bitnami/hadoop/share/hadoop/mapreduce$ hadoop fs -cat /output/part-r-00000
Kaustubh.      1
This           3
according      1
application    1
been           1
big            2
by             1
course         1
course.        1
data           2
digital        1
done           2
from           1
good.          1
has            1
is             1
practical      2
quite          1
the            1
to             1
vidya          1
was            1

```

Fig.10. Output after map reduce operation

6. SELF - EVALUATION

Throughout the course I was able to obtain quite a lot of knowledge about big data and its applications in big data management. I learned many things from this certification course such as follows:

- Learnt about the 4 Vs in Big Data i.e. Volume, Velocity, Variety and Veracity.
- Learned about Big Data Market, Infrastructure, how it is used in Business Intelligence.
- How all the data is stored in the data warehouse and how it is managed.
- Through practical sessions I got to know how to use Apache Hadoop.
- Also learnt how the map reduce is used to manage the data.
- I learnt how data ingestion is done i.e. before the processing of the data.
- Got to know about MongoDB, how to run commands in it and use it.
- In the last, I learnt about data pipelines and how it is used to manage the data.

7. CONCLUSION

Through the completion of the certification course, I was able to achieve an intermediate level of knowledge about the different concepts of big data platforms. The processing, management and analytics using the Big Data Management tools. It helped me improve my big data knowledge and took me to a new level such that I know how to manage big data practically. The knowledge about how to use MongoDB to manage a large amount of data is also very insightful. This helped me to get ahead of my peers in many ways as the knowledge I gained from doing this course is quite a lot. My future goals would be to do more certified courses related to big data so that I can become a data scientist at one of the reputed firms. This was one of the best beginner level courses to start my big data journey.