Solution to Data Management Challenges for Big Data Analytics through GM³'s Framework

Gendlal M Vaidya
Research Scholar, Department of Information Technology, Yeshvantrao Chawan College of Engineering. Nagpur, India, gendlalvaidya@gmail.com

Dr. Manali M. Kshirsagar
Professor, Yeshvantrao Chawan College of Engineering , Nagpur, India,manali_kshirsagar@yahoo.com

Abstract

Many researchers are contributing to the field and lot of results are carried out through their research. the data generation process through various sectors will be continually growing and in spite of too many hurdles the process will not stop here. Thus the challenges also rising in the domain. Data collected at a platform and need to give more accurate and useful information. The in between process from generation of data to results, have to face many challenges in data management and data processing. This study mainly focuses on the several challenges that might occur at every step of data management. Further we are trying to give the solutions to each challenge through the GM³'s framework. After preprocessing of data through ETL tools, some challenges like centralization, decentralization and indexing of data proposed within this study. In further studies, we will make some additions of approaches into the framework to carry out more efficient and reliable results.

Keywords—Data management, Hadoop, MapReduce

DOI Number: 10.14704/nq.2022.20.9.NQ44433 **Neuro Quantology 2022; 20(9):3777-3783**

I. Introduction

The term big data is obviously huge in quantity of data. It may contain various formats within it. In today's life, every software firms, IT companies, public and private sectors are collecting lots of data from their available sources. Once the data is stored into storage devices then they may face several challenges like data management, data security and many other data processing related issues. Once they started with pure and clean data, then further complications can be avoided. But, think if the proper neat data is not processed then, the generating results can not conclude to proper further decisions and therefore the ETL (extraction, transform and load) method plays a very important role in the data processing technique. Data warehousing is a platform where the data is integrated. The starting ETL process is not as easy and low priced. It all depends on the design of the data warehouse. Firstly the data extracted from various sources and after this, cleaned and standardized data stored in data warehouse. Business intelligence tool are used for the decision making purposes[1]. In previous work was related to preprocessing of data through some open source ETL tools, which is one of the basic requrement of data analytics.

While considering various challenges in data management, mainly issues related to storage, centralization, decentralization and indexing are may arise. So, to manage all these issues, here we implemented one partially completed framework which solves the raised issues. The experimental results will show how it overcomes all the issues. Apache foundation's contribution to the field of data science is very appreciable. The provided tools like Hive, Spark, Hadoop and so on solves the evolved issues. According to size, shape and spped of data, it is modeled into 5V's model which are known as Volume, Variety, Velocity, Veracity and Value. The heavy volume or quantity of data is tremendously generating with high speed and various form of data can be observed within it as shown in the following figure.

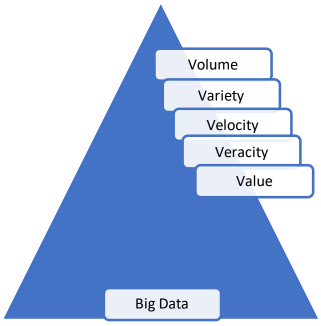


Figure 1. 5V's Model of Big Data

II. RELATED WORK

Hafiz A. Alaka and et al[2], This study implemented a fbig data analytics framework CB-FBM where data of thousands of construction firms was tested. The work given a comparison of iterations of MapReduce and their framework. Many researchers used Haloop, Spark, Daytona and Twister and more tools for the better experimentations. Instead using traditional tools, they used Spark and ANN to reduce the iterations in the datasets of construction firms. From the available metadata, it took nine seconds to complete the required iterations which shows the good accuracy in all manners. The proposed work leads to increase in the performance by framework. Table 1 below shows the some of initiatives that can be applied for processing of big data. It also shows whether the existing systems are supporting machine learning tools and fault tolerance or not.

Big data analytics initiative	Type/processing systems	Implementation platform	Distributed file system (Data Access)	Single or cluster/cloud Support	ML tools	Fault tolerance (FT)
Old Apache Mahout	MapReduce	Hadoop platform	HDFS	Both	Yes	Yes
Daytona	MapReduce runtime	Microsoft Azure	Not requiredb	Cloud based only	Yes	Yes
Twister	MapReduce runtime	Twister platform	Twister toold	Both	Yes	No
Haloop	Modified Hadoop MapReduce	Hadoop Platform	HDFS	Cluster/cloud only	Yes	Yes
New Apache Mahout	Spark	Any platforme	Any system	Both	Yes	Yes

Table 1: Features of Big Data [2]

Chenyang Xu and et al[3], this study deals with a computational engine name Octopus which integrates machine learning algorithms and standard queries. Using the adaptive scheduling it optimizes the running time of subtasks which was separated through main process. The novel idea of Octopus engine was tested on real time data and achieved the accuracy. Apache Spark contributed more in the running time required for processing of big data. This study mainly focused and supported the problems arises within traditional backend systems. Entire real time data proven the performance of their ideas and hence future study was relating to make engine more efficient for data analytics related challenges.

Manoj Muniswamaiah and et al[4], to obtain decision making through big data analytics heavy computation strategies are applied and exact results generations are not a easy tasks because heterogeneous data processing requires heavy setup. The proposed work of Appropriate Query

Processing deals with a single database. The work provides solution by using AQP as middleware solution for heterogeneous database. The experimental setup tested with 5GB and 20 GB databases where less response time is observed. This study provides one to one computation of different database. If data from various resources are taken into considerations then response time my get increased beyond the threshold values.

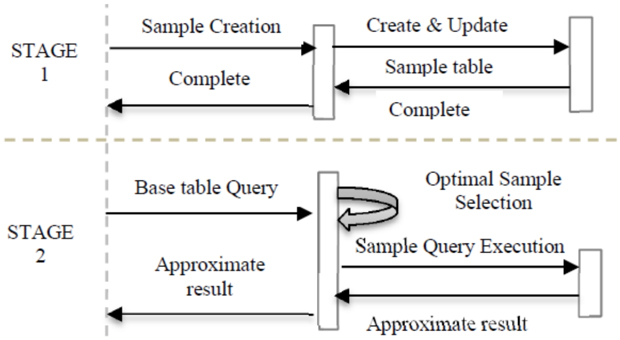


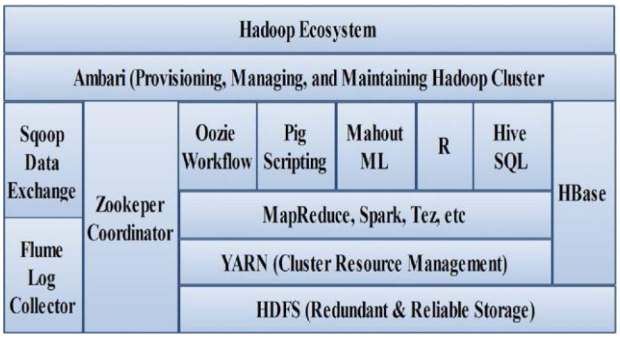
Figure 2: Query execution stages in AQP

Maksim Podkorytov and et al[5], focused some challenges in unified data model, design of the query optimizer, design of storage engine and design of the query compiler. They proposed a hybrid poly architecture with distribute in-memory multi-model storage engine. The distributed storage allowed large volume of data for experimental setup. The distributed storage hence overcomes some challenges arise in data storage and data management.

Yali Zaho [6] adn et al, contributed the area of big data analytics through profit optimization with scheduling algorithm. Tehy also proposed PO resource scheduling algorithmbased on the problems of scheduling. Minimal response time with cost saving resources was driven through the novel idea. Most parameters of BDA prooved by comparing with some previos work. Such concepts are more helpful for the processing of multiple data sets which may sac=ve the resouce costs and can achieve less response time.

III. Big Data Processing Platforms

3.1 HDFS - Hadoop distributed files system is most commonly used environment for BD processing. Though it is not fully supportive for all types of data but most researchers proven their results with the help of HDFS. The amin role of HDFS is to create different data nodes of multiple times which handles the one end of hadoop environment.



Fifure 3: Hadoop Network

3.2 MapReduce - This is another end of hadoop wherer Mapper and Reducer functions works individually for the mapping of data and provides reduced results. Basically both the strategies can be re implemented with advances in map() and reduce() methods.

3.3 Presto - It is an open source platform used for distributed SQL query engine. It supports heterogeneous data sources and process of GB's of data is possible on it. It contained with three types of servers i.e resource manager, coordinator and workers. Based on the experiment's requirement we can use either or all types together.

3.4 Spark - It is again open source platform used for workloads. It is having in memory caching concept and fast exaction of queries with variable size can be processed. Most of the big data analytics frameworks uses this platform.

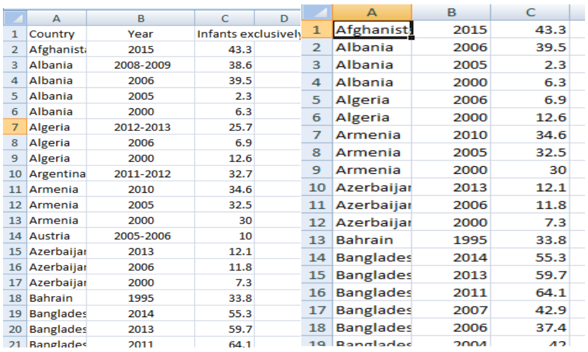
3.5 Hive - Its ued for the development of SQL scripts and widely used for MapReduce operations. Initial development of Hive completed by Facebook but later Apache foundation developed and enhanced its features. Generally it is assumed as relational database but actually it loads schema into the database.

IV. Experimental Setup

We installed hadoop environment on Ubuntu platform. Initially we tested functionality of Map and Reducer functions over it which results into all the available entries are mapped with '1' and finally Reducer function summarized it into its count. Experiment is carried out in following steps;

1) Data Collection : For the experimentation we chosen medical and healthcare datasets. The dataset contains country wise diseases related information. Firt we analyzed the entire information and and only useful entities are taken for further experiment. Meanwhile we took another datasets related to accounts information where several expenses carried out from advertisement agencies.

2) Preprocessing of datasets: In this section, we first cleaned up entire information with round figure values and used open source preprocessing tool Talend. for the processing of information, there are may open source preprocessing tools available but the Talend gives the filtered information with its design and pseudo code.



	A	B	C
1	Country	Year	Infants exclusively
2	Afghanist	2015	43.3
3	Albania	2008-2009	38.6
4	Albania	2006	39.5
5	Albania	2005	2.3
6	Albania	2000	6.3
7	Algeria	2012-2013	25.7
8	Algeria	2006	6.9
9	Algeria	2000	12.6
10	Argentina	2011-2012	32.7
11	Armenia	2010	34.6
12	Armenia	2005	32.5
13	Armenia	2000	30
14	Austria	2005-2006	10
15	Azerbaijar	2013	12.1
16	Azerbaijar	2006	11.8
17	Azerbaijar	2000	7.3
18	Bahrain	1995	33.8
19	Banglades	2014	55.3
20	Banglades	2013	59.7
21	Banglades	2011	64.1

	A	B	C
1	Afghanist	2015	43.3
2	Albania	2006	39.5
3	Albania	2005	2.3
4	Albania	2000	6.3
5	Algeria	2006	6.9
6	Algeria	2000	12.6
7	Armenia	2010	34.6
8	Armenia	2005	32.5
9	Armenia	2000	30
10	Azerbaijar	2013	12.1
11	Azerbaijar	2006	11.8
12	Azerbaijar	2000	7.3
13	Bahrain	1995	33.8
14	Banglades	2014	55.3
15	Banglades	2013	59.7
16	Banglades	2011	64.1
17	Banglades	2007	42.9
18	Banglades	2006	37.4

Figure 4: Before and After use of tool

By comparing both the datasets it is observed that the the preprocessing tool filtered out the years mentioned like 2008-2009, because such kind of information while considering any analytics work may not produce meaningful information and hence decision making goes in wrong ways therefore such parts to be filtered out.

3) Decentralization of data: Datasets when imported to the platform, was unstructured and decentralized. But to generate some meaningful results we need to centralize it and avoid unwanted fields to be removed from. This activity carried out using Talend tool and using filters we got the finalized design of the dataset.

4) Centralization of data: This is the next step where Map and Reduce to be carried out. It outcomes with sequential summarized data.

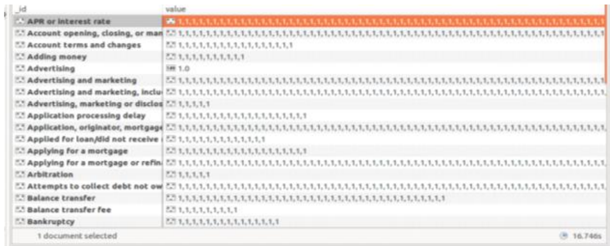
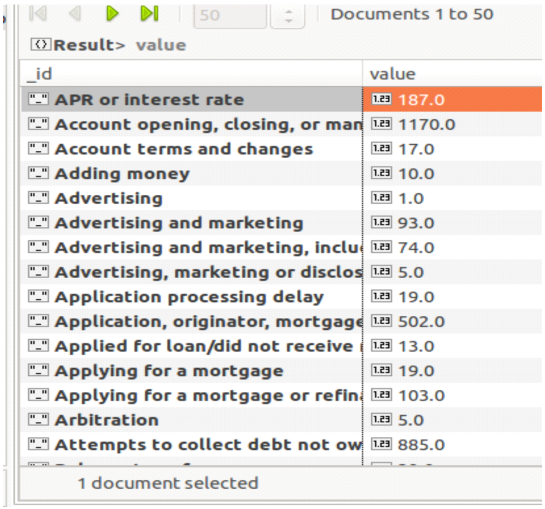


Figure 6: Results of Map() function



Documents 1 to 50
Result> value

_id	value
APR or interest rate	187.0
Account opening, closing, or man	1170.0
Account terms and changes	17.0
Adding money	10.0
Advertising	1.0
Advertising and marketing	93.0
Advertising and marketing, inclu	74.0
Advertising, marketing or disclos	5.0
Application processing delay	19.0
Application, originator, mortgage	502.0
Applied for loan/did not receive	13.0
Applying for a mortgage	19.0
Applying for a mortgage or refin	103.0
Arbitration	5.0
Attempts to collect debt not ow	885.0

1 document selected

Figure 7: Result of Reduce() function

5) Indexing of the information: The indexing of sample dataset

or any information is required for sequential output genetation. we performed some queries on entities like;

db.abc.createIndex({"Product":-1})

db.abc2.createIndex({"Company":1})

db.abc2.ensureIndex({"State":-1})

.

6) Merging of datasets: This step provides concatenation of two or more datasets where we can do the analysis of several entities targeting to multiple values extraction. The following queries are executed for getting the results.

db.abc.find(("Product": "Mortgage")).forEach(function(d){db.getSibling DB('com017')['mycollection'].insert(d);});

db.abc2.find({"cust_id":""}).forEach(function(d){ db.getSiblingDB('com017')['myCollection'].insert(d);});

By adding more queries we can combine multiple datasets and hence comparision between all the datasets can be done ata a glance. Such strategies gives the proper analytics results.

In this section, we tested various datasets and observed meaning results. This section provides the best way for big data management and solves may challenges involved within it.

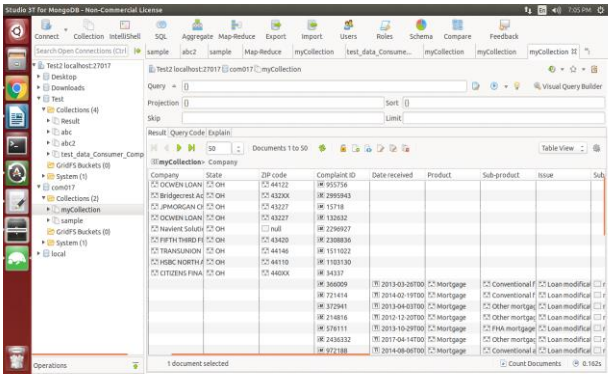


Figure 8: Merged datasets

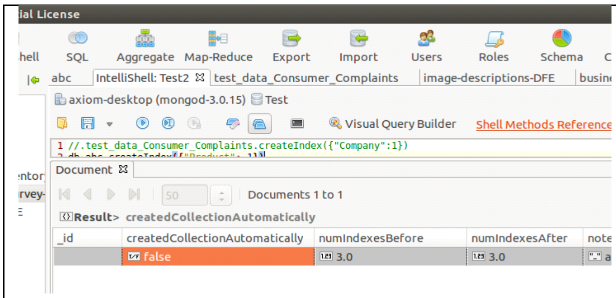
V. RESULTS AND DISCUSSION

A. *Solution to Preprocessin()ETL) of data:*

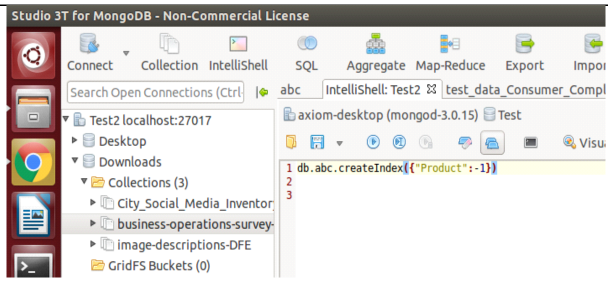
Extraction-Transformation and Loading of impure data always remains a challenging task for any researchers or big data organizations. Online platform provides many tools for preprocessing of data. If we uses paid versions of tools then we can integrate all of its functionality into proposed work. Many tools cannot fulfill all the basic requirements of researchers. Our proposed framework is capable to solve this issue through data integration tool Talend, which can filter out raw information and provides the design and pseudo code for the same datasets. There is a provision in framework to make the changes in dataset design and can produce more effective design of available datasets.

B. *Solution to Challenging issues in Big data management:*

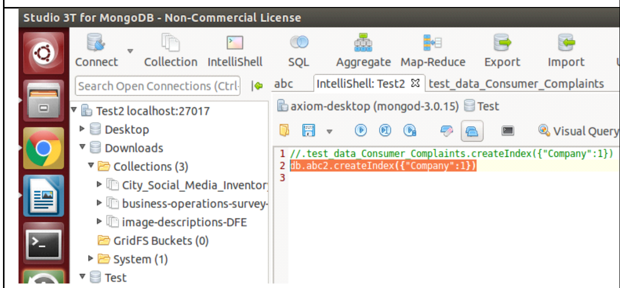
Once we received finalized pure datasets after ETL process, then data managemnt are again having challenges like centralization, decentralization and indexing of information. The proposed framework comes with solution to all the available challenges. By using HDFS and MongoDB within the framework we resolved the issues.



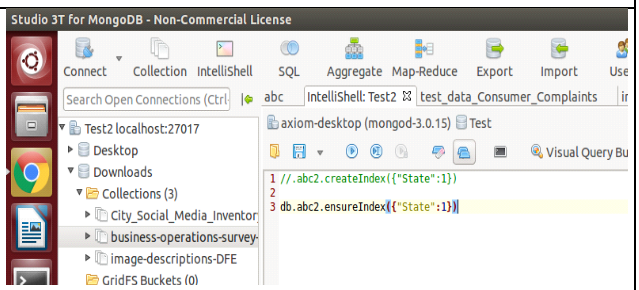
Decentalized data indexing



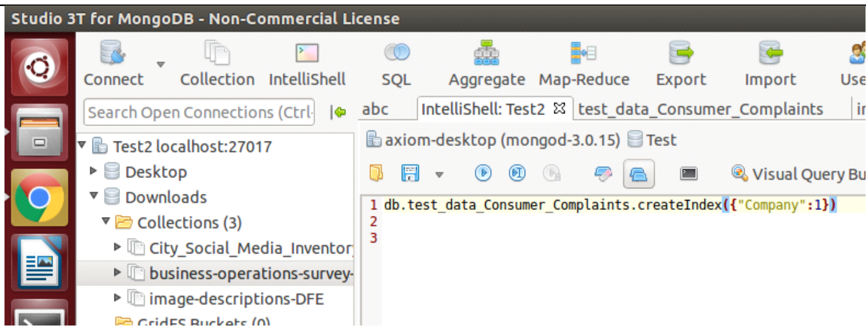
db.abc.createIndex({"Product":-1})



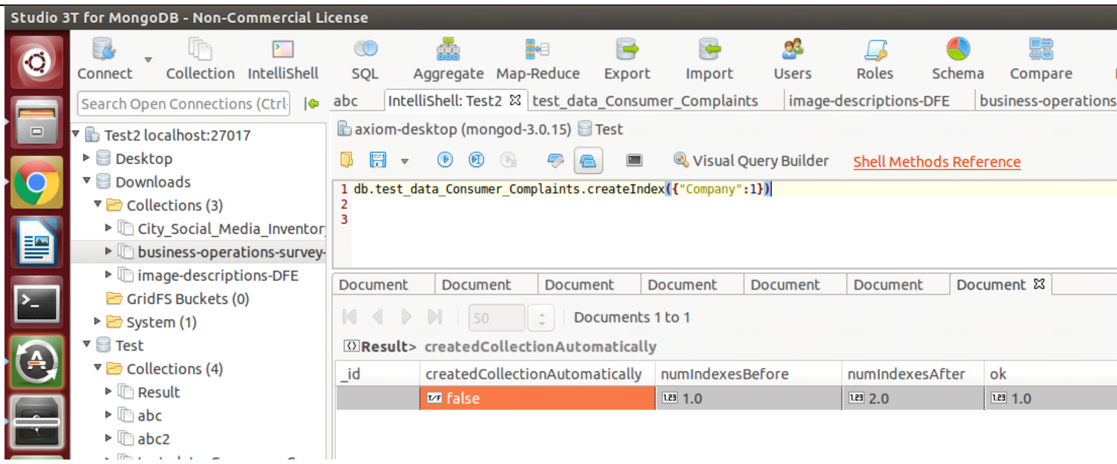
db.abc2.createIndex({"Company":1})



db.abc2.ensureIndex({"State":-1})



Centralized Data Indexing



db.test_Consumer_Complaints.createIndex({“Company”:1})

Table 2: Output Screens of Tested Queries

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented the results of solutions provided for challenges in data management. In future, we will work on query optimizations with respect to the developed framework and MongoDB.

VII. ACKNOWLEGEMENT

We are thankful to all the anonymous reviewers. Due to their suggestions, we achieved purity in the article. I dedicate this GM³'s framework to my parents and supervisor.

REEFERENCES

[1] Papa Senghane Diouf and et al, “Variety of data in the ETL processes in the cloud:state of the art”, published in IEEE International Conference on Innovative Research and Development (ICIRD), dated 11-12 May 2018,Bangkok Thailand, page no 1 to 5

[2] Hafiz A. Alaka and et al, "A framework for big data analytics approach to failure prediction of construction firms ", published in Applied Computing and Informatics (2018),

[3] Chenyang Xu, Yanjie Chen, Qin Liu, Weixiong Rao, Hong Min, and Gong Su, titled "A Unified Computation Engine for Big Data Analytics" published in 2015 IEEE/ACM 2nd International Symposium on Big Data Computing, page 73-77

[4] Manoj Muniswamaiah, Tilak Agerwala and Charles C. Tappert, titled " Approximate Query Processing for Big Data in Heterogeneous Databases", published in 2020 IEEE International Conference on Big Data (Big Data), page 6765-5767

[5] Maksim Podkorytov, Dylan Soderman, Michael Gubanov, titled "Hybrid.poly: An Interactive Large-scale In-memory Analytical Polystore" published in 2017 IEEE International Conference on Data Mining Workshops, pages 43-50.

[6] Yali Zhao, Rodrigo N. Calheiros, James Bailey, Richard Sinnott, "SLA-Based Profit

Optimization for Resource Management of Big DataAnalytics-as-a-Service Platforms in Cloud Computing Environments", published in 2016 IEEE International Conference on Big Data (Big Data), page 432 - 441

[7] Evdokia Kassela, Ioannis Konstantinou, Nectarios Koziris, "Towards a Multi-engine Query Optimizer for Complex SQL Queries on Big Data", published in 2019 IEEE International Conference on Big Data (Big Data), Page No 6095-6097

[8] Yali Zhao, Rodrigo N. Calheiros, James Bailey, Richard Sinnott, "SLA-Based Profit Optimization for Resource Management of Big Data Analytics-as-a-Service Platforms in Cloud Computing Environments", 2016 IEEE International Conference on Big Data (Big Data), Page No 432 to 441

[9] Congkai Bao, Meiyang Cao, "1Query Optimization of Massive Social Network Data Based on HBase" published in 2019 the 4th IEEE International Conference on Big Data Analytics, Page No 94 to 97

Gendlal M Vaidya, woring as Assistant Professor in Computer Technology Department at Yeshwantrao Chavan College of Engineering, Nagpur (India). His total teachning experience is of 14 years. He completed UG degree in Computer Engineering and Masters degree in Embedded System and Computing from Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur in 2004 and 2013. Currently, he is pursuing PhD at Yeshwantrao Chawan College Egineering under Nagpur university, Nagpur.His area of interest is Big Data, Networking and different programming platforms

Dr.Manali M. Kshirsagar, currently working as a Professor at Computer Technology Department at Yeshwantrao Chavan College of Engineering, Nagpur. She completed her Masters and Ph.D in Computer Science and Engineering from Amravati University, Amravati. She is having total 28 Years of academics experience. Her area of interest are Data Science, Database Management System, Operating System. .