Big Data Management Using Hadoop

Majida yaseen khalil¹, Murtadha M. Hamad²
Department of Computer Science, College of Computer Science and IT, University of Anbar, Ramadi, Iraq
majdhsyasyns@gmail.com
dr.mortadha61@gmail.com

Abstract. Today, one of the key issues is the design of systems and software to deal with the storage, management and processing of large amounts of data as a result of the exponential rise in data. In unstructured forms, these data are found. Due to the large and complex data sizes, data management with traditional approaches is unacceptable. Hadoop is an appropriate solution for the continuous growth of data sizes. We have suggested in this paper techniques and algorithms dealing with big data including data collection, preprocessing of data. The Fragmentation algorithm will take the function of a distributed implementation of the traditional file system time-sharing model, where various users share files and storage resources. Also, in this research we used a framework to improve the performance of a query and reduce the response time called the HADOOP. The Apache Hadoop project for safe, scalable and distributed computing. The results showed that Hadoop is the best way to deal with big data during calculating the rate of response time of a complex query for example at (00:00:01) per second and comparing it with the response time of the same queries on the fragmentation algorithm at (00:01:11) per second and the standard database at (00:05:13) per second. We concluded that Total time Access for complex queries in distributed processing is faster than in non-distributed processing.

Keywords. Big Data, Hadoop, Distributed data file system, Fragmentation Algorithm, OLAP and Structured & Unstructured data.

1. Introduction
At present, large volumes of data in heterogeneous sources (e.g. commercial and academic, finance) seem unimaginable. The abundance of smart computers and things on the Internet would make them a very technological character. Clear behind-the-scenario systems and distributed applications support multiple overlapping systems (smart grid systems for example) [1,2]. Big Data is used to describe huge amount of structured and unstructured data which is large. It was very difficult to process the data in the Big Data using the traditional databases and software technologies. [3].

Using most relational database management systems, desktop stats and simulation packages, it is hard to operate with enormous information. Instead, it needs "big parallel programs operating on dozens, hundreds or even thousands of servers". What is regarded to be "huge data" varies depending on the group management organization's capacities and the application capacities traditionally used to process and evaluate the data set in its own domain. "For some organizations, addressing hundreds of gigabytes of data for the first time may lead to a reconsideration of data management options, and for others it may take tens or hundreds of terabytes of data before the data volume becomes an important issue."
A possible solution is to found a technique for distributed data file system (Fragmentation algorithm), using Hadoop that decrease above limitations and the summation of total query maintenance cost and response time of the selected views which is regarded the view selection problem. The researchers have found algorithms and frameworks that solve the problem to reach the best solution of distributed System. The main contribution of this work is to achieve:
1- high query performance,
2- improving scalability and availability;
3- integrating data from multiple sites; integrating heterogeneous data;
4- offloading batch processing.

Therefore, the term Big Data refers to vast amounts of data that come from different sources. Hence Big Data refers not only to this huge data volume but also to the diversity of data types, delivered at various speeds and frequencies [4]. See figure 1.

![Figure 1. Data Growth Over the Years](image)

2. Related Works:
In 2015, their work is called Jie Songa, Chaopeng Guoa, Zhi Wanga, YichanZhanga, Ge Yub and Jean-Marc Piersone (HaoLap: A Hadoop-based Big Data OLAP System) in this paper HaoLap (Hadoop-based OLap), Big data application OLAP (OnLine Analytical Processing). Designed a hadoop-based OLAP and applied multiple algorithms to each specific work to roll-up dimensional hierarchy operations using dimension coding and traverse algorithms Dimensions and measurements were then stored using the partition and linearization algorithm. Effective results in OLAP and complex query. [5].

In 2016, their research is named Ikbal Taleb, Hadeel T. El Kassabi, Mohamed Adel Serhani, Rachida Dssouli, and Chafik Bouhaddioui (Big Data Quality: Evaluation of Quality Dimensions). Data is proud of being the most precious asset firms. The effects are unpredictable when their quality degrades and can lead to full misunderstandings. By implementing sampling strategies on big data sets, they suggested an effective data quality assessment system. The sampling will decrease the information size to a representative population sample for quick quality assessment. The findings showed that the mean sample quality score was representative of the initial information and illustrated the significance of sampling in order to decrease the cost of computing when evaluating big data quality [6].

In 2017, their work (DATA WAREHOUSE AND BIG DATA INTEGRATION) was called Sonia Ordonez Salinas and Alba Consuelo Nieto Lemus. Different opinions on warehouse data and large data Many believed that database data had vanished with the advent of large data, while others have achieved the fusion of the two by finding the intersection points and the gap between them and the combined tasks. They proposed models, for example the integration of the target, the technology used, the architecture and other common characteristics. [7].

In 2017, their work was called Xiaolei Li, Zhenyu Tu, Quanchao JiaXinjiang Man, Hui Wang and Xiuli Zhang (Deep-Level Big Data Quality Management) By using large data analysis to improve performance and raise rates, new business opportunities can be gained. The data analysis was carried out using industrial companies and the off-line data reference model library was developed. By using Spark to introduce the web application, we conclude that unlabeled data can be processed in real time in the manufacturing sector [8].
In 2018, their work was called Konstantinos Vassakis, Emmanuel Petrakis and Ioannis Kopanakis (Big Data Analytics: Applications, Perspectives and Challenges) The enormous increase in data varies from generation to generation. The increase of industrial enterprises in the previous generation, People and advanced technology has led to companies competing with each other, but the growth is now the result of rapidly growing Web and social networking sites. The ability to achieve, evaluate and action on statistics (“data-driven decision systems”) It is an effective system that can benefit governments and is distinguished as an important resident [9].

2.1. Characteristics of Big Data:

   Big data can be described by the following characteristics:

   1- Volume
   The volume is one distinctive which needs to be considered while dealing with Big Data. The amount of created and stored data. It is expected that the volume of data around the world will reach 40 ZB by 2020.

   2- Velocity
   In this context, Emerging information will be faster than before, Data is produced and processed to satisfy growth and development requirements and difficulties. Big data are frequently accessible in real time.

   3- Variety
   It refers to the data's type and nature. This enables individuals analyzing it to use the resulting understanding efficiently. Big data derives from text, pictures, audio, video, and through data fusion complete missing pieces.

   4- Veracity
   The data quality of captured data can vary greatly, affecting the accurate analysis [10].

2.2. Big Data and Hadoop

Approximately more that 44 times of data will be available over the next generation than today. It is also apparent that there is some restricted capacity in every data storage even in grid [11]. In order to manage this big quantity of data, it is necessary to divide the issue into components and to distribute these components to various computers in parallel. Whenever in collaboration various machines are used, the possibility of failure increases. Faults are expected and common in a distributed setting. In case of switches and routers break down, networks may face partial or complete failure. Data may not arrive at a specific stage in time due to the network congestion issue. Individual compute nodes also suffer from mistakes such as overheat, crash, hard drive failures, or memory or storage space failures. Data may be corrupted, transferred maliciously or incorrectly. Multiple implementations of client software can talk to slightly distinct protocols. The distributed system should be able to recover and proceed to create progress from the component failure or temporary error condition. Hadoop is the Big data issue solution. It's an open source framework that apache develops. Hadoop [ 12, 13] is one of those architectures that takes advantage of the characteristics listed above. Hadoop operates on commodity server clusters and can scale up to support thousands of hardware nodes and huge data volumes. It utilizes a distributed file system namesake that is intended to provide quick access to data across nodes in a cluster, plus fault-tolerant capacities to allow apps to continue running when individual nodes fail. [14].

2.2.1. Hadoop

Hadoop has published in Java and was began in 2006 as a top-level Apache project. It emphasizes discovery in order to realize near-impossible feats from the viewpoint of scalability and evaluation. Doug Cutting established Hadoop as an open-source project collection for which Google. Map Reduce programming environment in a distributed system could be implemented. It is currently being used on big quantities of information. Businesses can harness information with Hadoop that was earlier hard to handle and evaluate. About 63 percent of organizations use Hadoop to handle a massive amount of
unstructured records and occurrences. In specific, with variable structures (or no structure available), Hadoop can process highly big quantities of data [15]. See figure 2.

Figure 2. Hadoop

2.2.2. Key characteristics of Hadoop

1- Economic: Hadoop is open source, with no cost of licensing. Even if you buy Hadoop's suppliers from Cloudera or MAPR, then distribution is also free, but they only charge for assistance [16].

2- Flexible: Without any sort of system down, we can edit or delete a node. Hadoop is schema-free and can absorb any kind of data from any amount of sources, organized or not [16].

3- Scalable: A cluster can be extended without having to relocate, reformat, or alter dependent analytical workflows or apps by adding fresh servers or resources. It can handle any amount of data where only 10 TB of data could be handled by the ancient system [17].

4- Reliable: The system is constructed in such a manner that you still have data accessibility when there are failures [17].

3. The Proposed System

The proposed system illustrates the main steps from data collection to results obtained using the following algorithms and techniques.

3.1. Distributed Processing.

A major challenge in dealing with large data is the distributed file system as it uses several computers connected to each other using any available networks and in the case of a particular request, these computers will be sent and the immediate reply will be responded to, thereby saving time in data recovery.

3.1.1. Data Fragmentation

To handle large data, the data is distributed to multiple computers either horizontally or vertically in compliance with the Fragmentation algorithm and then handled with Client-Server architecture in need of a comprehensive OLAP complex. As shown in algorithm 1.

Algorithm 1 Fragmentation algorithm

Input: Hierarchy tree centralized.
Input: Integer X, Y\f The min and max number of fragmentation by Data.
Input: Integer E\f the end level of hierarchy.
Output: N\f The number of resulting fragments.
Boolean Find solution

Start.
Step 1 \{s=1; m=1;\}
Step 2. S=Select (Horizontal table (s, m));
Step3. A=Apply Primary Horizontal fragmentation (S (s, m));
Step4. Apply Derived Horizontal fragmentation (fact table) based on (A);
Step5. Calculate (N)
If(x<=n) OR(N<=Y) then
{Find solution =True;
Print ("Candidate Level is ", s, m)
Go to (6);
Else
If (m=E) then {
Find solution =False;
Go to (2);
Step6.}
End.

3.1.2. Replication of data
Replication is one of the technologies used to copy the data to more than one site to maintain in the case of loss of data from the designated place because it is located in the other and used with the process of fragmentation as integrated work in the architecture of Client-Server therefore, the data is stored more accurately and provide more data and give a detailed report of anything whether homogeneous or not.

3.1.3. Network Regulation
Distributed data operation within the network environment, where possible, should be within the area of building (LAN) or city (MAN). implementation of the system was based on an internal network (LAN) within organization building. The work will be in the architecture Client-Server.

3.2. Dealing with Hadoop
Hadoop became a fundamental data management platform for big data analytics uses. The stack of Hadoop technology has become the de facto standard for analyzing big data. There are many tools in the Hadoop ecosystem HBase, Storm, Pig, Hive, Oozie, and Ambari, to name just a few. We can definitely use Azure VMs to create your own custom Hadoop solution. Or we can use the Azure platform to provide and handle one for us via HDInsight service. We can even use Windows or Linux to deploy HDInsight clusters. It can be a significant time saver to provide Hadoop clusters with HDInsight (as opposed to doing the same manually).

1- System Requirements
We need Internet access and a browser (Internet Explorer 10 or greater to access Microsoft Azure and use Visual Studio to display ideas used in Azure applications development. So we're going to need Visual Studio 2019. The requirements for the system are:
• Windows 10
• Computer that has intel (R)core(TM)i7-8550u cpu @1.80GHz 2.00GHz or faster processor (2GHz recommended).
• 2 GB (64 Bit) RAM (Add 512 MB if running in a virtual machine).
• 8 GB of available hard disk space
• 5400 RPM hard disk drive
• DirectX 9 capable video card running at 1024 x 768 or higher-resolution display
• DVD-ROM drive (if installing Visual Studio from DVD)
• Internet connection
After installing Visual Studio, you must also install the Azure Tools and SDK for the language.

2-Using SQL Server Management Studio to connect
Open SQL Server Management Studio once we have added your IP address to the list of permitted IP addresses. We need to know the SQL Database server's complete name. The complete server name can be found in the SQL Database Blade's Essentials section along with other important information.

When the SQL Server Management Studio Connect to Server dialog box opens, enter the complete server name, Select SQL Server Authentication and provide the administrative login and password that we set when the database was created.

So Microsoft will upload the SQL database to azure. HD Insight Hadoop will deal with a SQL database; Hadoop will respond very rapidly when we ask for any complex query.
4. Implementation of the System
Interfaces display the execution of the designed system and the results of the execution are discussed. The results are also presented in tables and displayed through interfaces, and graphs that are a kind of visualization tool represent the final results.

4.1. Distributed processing
The machines in a distributed system interact with each other through different media, such as high-speed networks or local network. They have no main memory or disks to share.

● Data Fragmentation

If the LSalesDB - relationship is fragmented, LSalesDB will be divided into a number of db1, db2, db3, db4. These fragments have enough information to reconstruct the original relationship. This reconstruction can take place either through the Union procedure or through a unique type procedure on the different pieces. There are two distinct schemes (Horizontal & Vertical) for fragmenting a relationship. So, we have divided the original data into four sections (db1, db2, db3.db4), each section into client (client 1, client2, client3, client4) horizontally. As shown in figure 3.

![Figure 3. Fragment the Original Data (LSales DB)](image)

In Network Regulation Distributed data process within the network environment. The four computers are referred to as client1, client2, client3, client4 connected by the network (LAN or MAN), the one is working as a server, the rest four computers as a client as explained in chapter 3. As shown in figure 4.

![Figure 4. Explains Sample Client –Server Architecture of Proposed System](image)

In order to connect to four Client, we will set for each Client local host and serial port starting from 5000, 6000,7000 and 8000 sequentially and then press start button.

Now go to the original server and press the Execute button will show the next main interface then we click on any complex query found within the appendix (B) The results will be shown as in the following figure 5. Note that each Client (Storage Node) will take 860.160 kb of data.
Figure 5. The Main Interface to Implement Complex Queries from Server to Clients

- Data Replication

Replication is a collection of techniques to copy and distribute information and database items from one database to another and then synchronize databases in order to preserve consistency. As shown in figure 6 (a, b).

4.2. Interfaces of the Implementation Hadoop on SALES System

Faults are expected and common in a distributed setting. If switches and routers or local host break down, networks may face partial or complete failure. Data may not arrive at a specific stage in time due to the network congestion issue. Individual compute nodes also suffer from mistakes such as overheat, crash, hard drive failures, or memory or storage space failures. Data may be corrupted, transferred
maliciously or incorrectly. Thus, such a platform needs to be designed that offers high levels of reliability, effectiveness, accessibility, and scalability. Hadoop is the Big data issue solution. When we are configuring Cluster, we choose the type of cluster is Hadoop and operating system (OS) is Linux and the Version is Hadoop 2.7.3(HDI 3.6)

The SQL data base (SALES DATA) will be uploaded to the cloud computing. And the connection to the database is by the following connection:

```
(Server=majida.database.windows.net;Database=LSalesDB;UserId=majida;Password=majida123456;Connection Timeout=0;)
```

Now we will click on the start button in above window will show the main window of the Hadoop and then click on any button of complex query will appear results below as shown in the following figures 7,8.

![Fig.7 The Main Interface for Executing Complex Queries Using the Hadoop.](image)

![Fig.8 Implementation of Query 8 in Hadoop](image)

5. Results and Discussion

By applying the proposed system algorithms on the sales system data we found:

- Response Time of Query

The query response time in the OLAP and decision support systems is critical and very important. By applying distributed processing algorithms to the sales system We concluded that when processing large data time saving (i.e. the system requires a few minutes), high quality and data retrieval speed. Therefore, the implementation of the query on the distributed processing(Fragmentation) provides us with fast response time and speeds up decision making. For example, if we implemented queries fifteen that were implemented by each Client separately in distributed processing would be faster compared to
if they implemented themselves without distributed processing. But if we use the Hadoop in the execution of these queries will be much faster because the Hadoop already deal with large data and analysis. Hadoop depends on the speed of the Internet and the complexity of query. As shown in the table 1 and figure 9.

**Table 1.** Execution time of OLAP query in Standard and Fragmentation and Hadoop

| Queries | Standard | Fragmentation | Hadoop |
|---------|----------|---------------|--------|
| Query1  | 0:12:15  | 0:10:05       | 0:00:23|
| Query2  | 0:18:14  | 0:10:01       | 0:00:41|
| Query3  | 1:00:20  | 0:56:09       | 0:02:18|
| Query4  | 0:05:13  | 0:01:11       | 0:00:01|
| Query5  | 0:15:55  | 0:14:45       | 0:00:36|
| Query6  | 0:45:10  | 0:40:10       | 0:01:00|
| Query7  | 1:01:02  | 0:58:00       | 0:03:08|
| Query8  | 0:10:44  | 0:04:10       | 0:00:05|
| Query9  | 0:46:22  | 0:43:20       | 0:01:08|
| Query10 | 0:07:29  | 0:03:08       | 0:00:02|
| Query11 | 0:15:21  | 0:10:10       | 0:00:24|
| Query12 | 0:20:25  | 0:17:20       | 0:01:27|
| Query13 | 0:41:38  | 0:39:33       | 0:01:01|
| Query14 | 0:09:14  | 0:05:09       | 0:00:04|
| Query15 | 0:21:50  | 0:20:55       | 0:01:34|

Fig.9 Comparison of Queries Implementation Between Standard and Fragmentation and Hadoop

6. Conclusions
By studying distributed data file system (Fragmentation algorithm) and using Hadoop theoretically, besides the practical part, the implementation of proposed algorithms and results discussion. We obtain the following conclusions and recommendations for the future work.

1. The main idea behind designing this system and implementing it is to get the optimizing queries performance and minimum response time to answer complex, ad-hoc queries.
2. The use of Hadoop is the best solution for dealing with big data which is preferred over all other methods.
3. Big data storage is largely dependent on Hadoop.
4. Hadoop is an open source software project that enables big datasets to be handled distributed across service server bunches. It is regarded to scale up to thousands of techniques from a single server with a very great mark of tolerance of fault.
5. The assessment focused on some aspects of data quality such as completeness and reliability.
Out of the results of the present work and depending on our expectations and viewpoints about the research topics and to make the subject more important, reliable, transparent for future development, we propose the following recommendations:

1. Potential of using machine learning with the Hadoop instead of using the alone Hadoop to answer complex queries as in our current system.
2. The possibility of using social media database such as (Facebook or twitter) instead of using videos, digital and texts data and implement the current system on it.
3. Implementation of the proposed system using spark apache.
4. Implementation of the proposed system using java environment.
5. using Classification algorithms like Support Vector Machines (SVMs), neural network for answering new entered queries.

References

[1] Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences, 30(4), 431-448.
[2] Kune, R., Konugurthi, P. K., Agarwal, A., Chillarige, R. R., & Buyya, R. (2016). The anatomy of big data computing. Software: Practice and Experience, 46(1), 79-105.
[3] Dhaenens, C., & Jourdan, L. (2016). Metaheuristics for big data. John Wiley & Sons.
[4] Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 2053951716631130.
[5] Song, J., Guo, C., Wang, Z., Zhang, Y., Yu, G., & Pierson, J. M. (2015). HaoLap: a Hadoop based OLAP system for big data. Journal of Systems and Software, 102, 167-181.
[6] Taleb, I., El Kassabi, H. T., Serhani, M. A., Dssouli, R., & Bouhaddioui, C. (2016, July). Big data quality: Aquality dimensions evaluation. In 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld) (pp. 759-765). IEEE.
[7] Salinas, S. O., & Lemus, A. C. (2017). Data warehouse and big data integration. Int. Journal of Comp. Sci. and Inf. Tech, 9(2), 1-17.
[8] Li, X., Tu, Z., Jia, Q., Man, X., Wang, H., & Zhang, X. (2017, October). Deep-level quality management based on big data analytics with case study. In 2017 Chinese Automation Congress (CAC) (pp. 4921-4926). IEEE.
[9] Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: applications, prospects and challenges. In Mobile Big Data (pp. 3-20). Springer, Cham.
[10] Owais, S. S., & Hussein, N. S. (2016). Extract five categories CPIVW from the 9V’s characteristics of the big data. International Journal of Advanced Computer Science and Applications, 7(3), 254-258.
[11] Rathi, D. R., & Lohiya, S. (2014). Big Data and Hadoop. International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), 2(2).
[12] Jablonski, J. (2011). Introduction to Hadoop. A Dell Technical White Paper.
[13] Apache Hadoop. http://hadoop.apache.org/
[14] https://searchdatamanagement.techtarget.com/definition/Hadoop
[15] Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, M., Kamaleldin, W., ... & Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. The Scientific World Journal, 2014.
[16] Beakta, R. (2015). Big Data and Hadoop: A Review Paper. Baddi University of Emerging Sciences & Technology, Baddi, India Volume 2, Spl. Issue 2 ISSN: 1694-2329 2015.
[17] Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, 2(1), 24.