



BIG DATA MANAGEMENT FOR INFORMATION TECHNOLOGY COMPANIES IN MALAYSIA

A project report submitted in partial fulfillment of the requirement for the award of
the degree of Bachelor (Hons.) of Technology Management (Innovation)



ARVIND A/L NARASAMMAN

B061910405

001025-10-1225

FACULTY OF TECHNOLOGY MANAGEMENT AND TECHNOPRENEURSHIP

2022

SUPERVISORS DECLARATION

I certify that this thesis entitled “ BIG DATA MANAGEMENT FOR IT COMPANIES IN MALAYSIA ” was prepared by ARVIND A/L NARASAMMAN(B061910405) has meet the required standard for submission in partial fulfillment of requirement for the award of Bachelor (Hons.) of Technology Management (Innovation) at Universiti Teknikal Malaysia Melaka.

Approved by.

Signature:

Supervisor's Name:

PROFESOR MADYA DR NORAIN BINTI ISMAIL

Date:

Signature:

Panel's Name : ADILAH BINTI MOHD DIN

Date :

9/2/2023



STUDENTS DECLARATION

'With the exception of citations and quotations that have been properly acknowledged, I hereby declare that this thesis is based on my original work.' I further declare that it has not been submitted for any degree or award at Universiti Teknikal Malaysia Melaka or any other institution earlier or concurrently.'

Signature : *arvind*

Name : ARVIND A/L NARASAMMAN

No Matric :B061910405

Date : 29/1/2023



**BIG DATA MANAGEMENT FOR INFORMATION TECHNOLOGY
COMPANIES IN MALAYSIA**

ARVIND A/L NARASAMMAN

A project report produced as part of the requirements for the award of a bachelor's
degree.

Bachelor (Hons.) of Technology Management (Innovation)



JUNE 2022

DEDICATION

My dissertation effort is a gift to my family . My families words of encouragement and admonition to persevere ring in my ears, and I'm grateful for them. They've always been there for me, and they're very wonderful.

In addition, I would like to thank my supervisors and members of my final year project community for their encouragement and support during the writing process. I'll always be grateful for their aid , particularly my father , my mother and my sister who taught me how to use the leader dots and helped me improve my technical abilities.



ACKNOWLEDGEMENT

The success of this initiative would not have been possible without the help of numerous individuals. My Family and Dr Norain , my Supervisor, was instrumental in helping me sort through my jumbled thoughts. Please accept my sincere appreciation for the help and support that I received from the members of my final year project committee

اونيورسيتي تيكنيكل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Abstract

Big data is a term used to describe enormous data sets that have a high Velocity, high Volume, high Variety, and complicated structure, as well as challenges with management, analysis, storage, and processing . Big data's characteristics make it extremely challenging to manage, analyse, store, transport, and analyse the data using current conventional methods. In this work, big data management and analysis are introduced. Big data's initial introduction, its challenges and solution to the challenges of big data management, including its use with cloud computing, Hadoop , HDFS , MapReduce, and its conclusion and future research are all covered.

KEYWORD : BIG DATA , BIG DATA MANAGEMENT , CHALLENGES , SOLUTION

TABLE OF CONTENT

SUPERVISOR DECLARATION	2
STUDENT DECLARATION	3
TITLE PAGE	4
DEDICATION	5
ACKNOWLEDGEMENT	6
ABSTRACT	7

	 	
	CHAPTER 1	
INTRODUCTION		14
BACKGROUND		14
	UNIVERSITI TEKNIKAL MALAYSIA MELAKA	
PROBLEM STATEMENT		15-16
RESEARCH QUESTION		17
RESEARCH OBJECTIVE		17
SCOPE OF RESEARCH		17
RESEARCH SIGNIFICANCE		17-18



CHAPTER 2 LITERATURE REVIEW

INTRODUCTION	19
BIG DATA OVERVIEW	19-20
BIG DATA HANDLING METHODS	21-22
ADOPTION OF BIG DATA	23
RELATIONSHIP BETWEEN IT EMPLOYEES AND BIG DATA MANAGEMENT	24


CHARACTERISTICS OF BIG DATA	25
TYPES OF BIG DATA	26
CHALLENGES OF BIG DATA	27-29
SOLUTION TO OVERCOME THE CHALLENGES OF BIG DATA	30-31
THEORETICAL FRAMEWORK	32-33
SUMMARY	33



CHAPTER 3 RESEARCH METHODOLOGY

INTRODUCTION	34
RESEARCH DESIGN	34
METHODOLOGICAL CHOICES	35
SOURCES OF DATA	35-36
RESEARCH STRATEGIES	36

INTERVIEW DESIGN	36-37
SAMPLING DESIGN	37
SAMPLING TECHNIQUE	37
SAMPLING SIZE	38
SAMPLING LOCATION	38
TIME HORIZON	39
DATA ANALYSIS METHOD	40-42
SUMMARY	43




UNIVERSITI TEKNIKAL MALAYSIA MELAKA

اونيورسيتي تیکنیکل ملیسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

CHAPTER 4 DATA ANALYSIS AND FINDINGS	
4.1 INTRODUCTION	44
4.2 DEMOGRAPHICS OF INTRVIEWEES	44
4.2.1 CATEGORIES OF INTERVIEWEES	44
4.3 RESEARCH FINDINGS	45
4.3.1 INTRODUCTION	45

4.3.2 TYPES OF BIG DATA	45-46
4.3.3 CHALLENGES OF BIG DATA MANAGEMENT IN IT COMPANIES	47-49
4.3.4 WAYS TO OVERCOME THE CHALLENGES OF BIG DATA -52 MANAGEMENT IN IT COMPANIES	50
4.4 RESEARCH FINDINGS (THEMES)	53
4.4.1 THEMES FOR TYPES OF BIG DATA	53
4.4.2 THEMES FOR CHALLENGES OF BIG DATA MANAGEMENT FOR IT COMPANIES	54
4.4.3 THEMES FOR SOLUTIONS OF CHALLENGES OF BIG DATA 55 MANAGEMENT IN IT COMPANIES	55
	
CHAPTER 5 CONCLUSION	
5.1 INTRODUCTION	
56	
5.2 OVERALL SUMMARY OF THE STUDY	56
5.3 CONTRIBUTION OF RESEARCH	57
5.4 LIMITATIONS OF RESEARCH	58

5.5 RECOMMENDATIONS OF FUTURE RESEARCH	59
5.6 GANTT CHART	60
5.7 REFERENCE	61-66
5.8 APPENDIXS	67-72
5.9 SIMILARITY REPORT	
73	



CHAPTER 1

INTRODUCTION

A large amount of data referred as Big Data, as the term implies. Big Data refers to a data collection that is both huge in volume and complicated. Big Data cannot be handled by traditional data processing software due of its enormous size and growing

complexity . Big Data is a term used to describe datasets that contain a large amount of unstructured and structured data .(Ishwarappa, J. Anuradha 2015)

BACKGROUND

Since the early 1990s, the term "Big Data" has been in usage. Although the term's originator is unknown, John R. Mashey (a Silicon Graphics employee at the time) is generally credited with popularising it . (Enterprise Big Data Framework , 2022)

Despite the fact that Big Data is a relatively recent phenomenon, it is not a wholly new one. Throughout history, individuals have tried to employ data analysis and analytics approaches to help them make decisions. In the library of Alexandria, Egypt's ancient Egyptians tried to gather all of the available "data" circa 300 BC. Furthermore, the Roman Empire utilised to evaluate military statistics in order to identify the best way to distribute their troops. (Enterprise Big Data Framework , 2022)

As a result of this, the amount and speed of data generation has increased exponentially during the past two decades. In 2013, the world's total data storage capacity was 4.4 zettabytes. By the year 2020, that number is expected to soar to 44 zettabytes. As an example, 44 zettabytes is equal to 44 trillion gigabytes in terms of capacity. Data analysis is impossible, even with the most cutting-edge technology available today. Traditional data analysis was transformed into "Big Data" in the previous decade because of the requirement to analyse these ever-larger (and unstructured) data sets. (Enterprise Big Data Framework , 2022)

There are three distinct stages in the evolution of Big Data that can be summarised. In each phase, there is a unique set of features and powers. A thorough understanding of Big Data's modern environment requires a thorough understanding of each phase's contribution. (Enterprise Big Data Framework , 2022)

BIG DATA PHASE 1	BIG DATA PHASE 2	BIG DATA PHASE 3
Period: 1970-2000	Period: 2000-2010	Period: 2010-present
DBMS-based, structured content: <ul style="list-style-type: none"> • RDBMS & data warehousing • Extract Transfer Load • Online Analytical Processing • Dashboards & scorecards • Data mining & statistical analysis 	Web-based, unstructured content <ul style="list-style-type: none"> • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis 	Mobile and sensor-based content <ul style="list-style-type: none"> • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization • Human-Computer-Interaction

PROBLEM STATEMENT

Expertise in cutting-edge technology and big data solutions is sought after by companies. Data scientists, data analysts, and data engineers are among the experts who will use the technology to analyse and interpret huge amounts of data. One of the most challenging Big Data challenges that any company has to deal with is the scarcity of big data expertise. In many cases, this is due to the fact that data management technologies have advanced rapidly, although most experts have not close the gap with concrete actions . Because of a lack of expertise, companies efforts to harness the power of Big Data have been a resounding failure. It's possible that employees don't know what data is, how it's stored and processed, how important it is, or even where it came from. Others, such as data specialists, may be in the dark about what's happening. The failure to back up critical data may be due to employees who do not comprehend the need of knowledge storage, for example. They couldn't appropriately store data in databases. This means it's difficult to get this critical information when it's needed . (XenonStack ,2020)

The proper storage of these enormous volumes of data is one of the most difficult aspects of Big Data . When it comes to commercial data centres and database storage, the volume of data is always increasing. As the size of a data collection grows, managing it becomes more complex. The majority of the data is unstructured and

comes from a wide range of media, including books, films, audio files, text documents, and more. Because they have not been entered into the database, this implies that they are not available. To cope with these huge data sets, companies adopt cutting-edge methods like compression, tiering, and deduplication. An important aspect of compression is that it reduces the amount of bits in a piece of data. Data dedupe is the process of eliminating redundant and unnecessary information from a collection of knowledge. Many storage tiers may be used for data tiering in the world of companies. It makes certain that the data is kept in the most suitable location possible. Depending on the volume and importance of the data, several tiers of storage are used, such as public cloud, private cloud, and flash storage. Big Data technologies like Hadoop, NoSQL, and others are also being used by companies . (Xenon , 2020)

High cost of data solution is also considered to become a headache when the topic big data comes to light . After understanding how your companies will benefit most from implementing data solutions, you're likely to find that buying and maintaining the necessary components can be expensive. Along with hardware like servers and storage to software, there also comes the cost of human resources and time. (Solvexia ,2019)



RESEARCH QUESTION

- i. What are the types of Big Data for Information Technology companies in Malaysia ?
- ii. What are the challenges of Big Data Management for Information Technology companies in Malaysia ?
- iii. How to overcome the challenges of Big Data Management for Information Technology companies in Malaysia ?

RESEARCH OBJECTIVE

- I. To study types of Big Data for Information Technology companies in Malaysia
- II. To study challenges of Big Data Management for Information Technology companies in Malaysia .
- III. To analyse ways to overcome challenges of Big Data Management for Information Technology companies in Malaysia .

SCOPE OF RESEARCH

The general purpose of this research is to study about Big Data . The population or sample that this research will focus is basically on the employees who specialise in IT . The participant for this research is chosen based on few criteria such as the experience in IT , specialist in handling big data , capable of answering the interview questions , skills of the respondent in the IT and also knowledge on IT . Other than that ,the duration this research will take is approximately around 6 months . The topics and theories that will be studied are definitions , types , where is it used , challenges and solution on how to overcome . The geographical location that this research will take place is in Selangor , Malaysia .

RESEARCH SIGNIFICANCE

First of all , it is significant from the aspect of national development . At this point, the globe has fully entered the information age. The widespread use of the Internet, Internet of Things, Cloud Computing, and other emerging IT technologies has resulted in an unprecedented increase in the number of data sources, while also increasing the complexity of data formats and kinds. Deep analysis and the use of big data will play an essential role in fostering sustainable economic growth and

increasing company competitiveness.

Second of all , it is significant from the aspect of industrial upgrades . Big data is presently a prevalent concern for many businesses, posing significant hurdles to their digitalization and informationization. Research on common big data challenges, particularly advancements in fundamental technologies, will help industries to harness the complexity generated by data connectivity and manage uncertainties caused by redundancy and/or data scarcity .

Last but not least , it is significant from the aspect of scientific research . Big data has compelled the scientific world to reconsider its research technique , resulting in a revolution in scientific thought and methodologies. Experiments were used to conduct the first scientific inquiry in human history. Later, theoretical science evolved, which was distinguished by the investigation of many laws and theorems. However, because theoretical analysis is too complicated and impractical for tackling actual issues, people began to look for simulation-based methodologies, which led to the development of computational science .

Finally , it is significant for emerging interdisciplinary research . Big data technology and the accompanying basic research have emerged as academic research priorities. Data science is a new multidisciplinary subject that is progressively taking shape. This study uses big data as its research object and seeks to generalise knowledge extraction from data. It includes information science, mathematics, social science, network science, system science, psychology, and economics.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides an overall literature review of the previous study relating to the title of this research. Besides, within the section, the study distance identified of earlier studies was also clarified. In this chapter, the researcher will determine how the use of previous research will generate the information and detail for this study. The researcher will explain about what is Big Data , Types ,Challenges in Big Data Management and also ways to overcome the Challenges in the Big Data by referring to previous research.

2.2 Big Data Overview

User-provided content and social media are only a few of the many sources of big data. Other sources include data collected via sensor networks or through commercial operations like sales enquiries and purchases . Furthermore, genomics, health care, engineering, operations management, the Internet of Things, and finance all contribute to the pervasiveness of big data . Powerful computational tools are required to explore the vast socioeconomic statistics to find patterns and trends . It is possible that new data-value extraction insights may have a considerable impact on mostly stagnant government statistics, surveys, and archive data sources by providing depth and knowledge from community experiences in live time and bridge both information and temporal gaps. Perhaps the misunderstanding stems from the "bigness" of massive data, which continuously draws researchers' attention to the dataset's size. More and more practitioners agree that the defining criterion these days is the "smartness" of data, that is the insights that the volume of data can genuinely bring. "Big data" is defined by its fine-grained structure, which shifts emphasis from a large number of participants toward specific information about each individual. A Formula 1 racing vehicle, for example, has 150 sensors that produce 20 terabytes of data that

can be used to analyse not just the technical performance of the car's components, but also driver responses, pit stop delays, and crew-to-driver communication, all of which contribute to overall performance (Munford, 2014). As a consequence, the focus moves from race results (win/loss) to each proximal, contributing component discovered for each second of the race. Similarly, by using remote sensors to map mobility patterns onto physical workspace layouts, individuals' social networks and engagement behaviours might be examined together with meeting room use patterns to provide insight into the communication and coordination requirements associated with more complicated projects with rapidly approaching deadlines. Micro data provides a wealth of data on individual actions and behaviours that management research has yet to fully use. The use of "big" or "smart" data to predict human behaviour is gaining traction in business and government policy, as well as in scientific domains where physical and social sciences collide (dubbed "social physics" lately) (Pentland, 2014).



2.3 BIG DATA HANDLING METHODS

There is a constant stream of new data in a variety of forms that most firms have to deal with. Insights generated by big data have the potential to transform any organisation. MapReduce, a framework for processing big data, has generated a whole new sector. Developed by Google, Data processing framework MapReduce uses the divide and conquer method to break down huge data problems into smaller chunks that may be processed in parallel . There are two phases in the MapReduce process: the Map Step, which divides up the data from the master node into smaller subproblems, and the Reduce Step, which divides it up even more. Reducers may access the results stored in the local file system of a worker node, which is controlled by a JobTracker node. The map stages' input data is examined and aggregated in the Reduce Step. Several reduction tasks may be run on worker nodes under the control of the JobTracker to parallelize the aggregate . (Seref Sagiroglu , Duygu Sinanc)(2014)

BigTable, the Google File System, and MapReduce all served as inspiration for Hadoop. There are several uses for Hadoop, which is a Java-based framework and open source platform. An ETL (Extraction, Transformation, Load) technique is not meant to be used in lieu of this tool. There are three main components to Hadoop: a distributed file system, an analytics platform, and a configuration management layer. If you need to handle streams of complex events in real time, this is not the platform for you. HDFS (Hadoop Distributed File System) connects the file systems on separate input and output data nodes in a Hadoop cluster to form a single large file system. (Seref Sagiroglu , Duygu Sinanc)(2014)

HPCC (High Performance Computing Cluster) Systems are open source computation systems for big data workflow management and distributed data-intensive computing. Unlike Hadoop, HPCC's data model is user-defined. A high level ECL foundation may readily explain the key to difficult issues. HPCC guarantees that ECL is completed as quickly as feasible while simultaneously treating nodes. Furthermore, third-party technologies like as GreenPlum, Cassandra, RDBMS, Oozie, and others are not required for the HPCC Platform . HPCC Data Refinery (Thor) is a massively parallel ETL engine for large-scale data integration and batch-oriented data manipulation. HPCC Data Delivery Engine (Roxie) is a massively parallel, high throughput, ultra fast, low latency engine for efficient multi-user data retrieval and structured query response . (Seref Sagiroglu , Duygu Sinanc)(2014)

2.4 ADOPTION OF BIG DATA

Big data adoption is defined by Günther, Rezazade Mehrizi, Huysman, and Feldberg (2017) as a process that permits an innovation to transform an organization's infrastructure. Adoption of big data necessitates enhanced data processing methods and technology that help decision-making (Raguseo, 2018). It provides new possibilities for firms to leverage data to achieve a competitive edge (Ur Rehman et al., 2019). The utilisation of big data increases productivity, predicts risk, and improves customer service (Al-Qirim, Tarhini & Rouibah, 2017) . Choosing which technology to utilise is the process of adoption for a business or organisation (Agrawal, 2015) . Adoption of big data allows organisations and industries to outperform their competition. Adoption of big data may be time-consuming and expensive, although long-term advantages might lead to success (Al-Qirim et al., 2017). Almost every data is a byproduct of business (Moat et al., 2014) . According to a Zoom data study, 41% of firms are now using big data, and 46% want to do so in the near future.(Zoomdata, 2017). Organizations may anticipate greater issues as the amount of data grows on a daily basis (Zhao, Yu, Li, Han & Du, 2019). These difficulties might be connected to concerns about data security, privacy, and ownership (Osman, 2019).

The archaic technology couldn't handle data overload (Sivarajah, Kamal, Irani, & Weerakkody, 2017). Digital developments have been absorbed more swiftly in developed countries, but they must also grasp, accept, and utilise digital