ISBN:978-93-95341-11-0

INTRODUCTION TO BIG DATA

Dr. T. Arumuga Maria Devi
Dr. G. Heren Chellam
Dr. T. J. Benedict Jose
Dr. D. Sharmila
Mrs. A. Premalatha

INTRODUCTION TO BIG DATA

Dr. T. Arumuga Maria Devi
Assistant Professor,
Centre for Information Technology and Engineering,
Manonmaniam Sundaranar University,
Tirunelveli, Tamil Nadu, India.

Dr. G. Heren Chellam
Assistant Professor,
Department of Computer Science,
Rani Anna Government College for Women,
Tirunelveli, Tamil Nadu, India.

Dr. T. J. Benedict Jose
Assistant Professor,
Department of Computer Applications,
Government Arts and Science College,
Palkulam, Kanyakumari, Tamil Nadu, India.

Dr. D. Sharmila
Assistant Professor,
Department of Computer Applications,
Government Arts and Science College,
Palkulam, Kanyakumari, Tamil Nadu, India.

Mrs. A. Premalatha
Assistant Professor,
Department of Computer Science,
Rani Anna Government College for Women,
Tirunelveli, Tamil Nadu, India.

Published by

SK Research Group of Companies
The International Journal & Book - SKRGC Publication
142, Periyar Nagar, Madakulam
Madurai - 625003, Tamil Nadu, India
skrgc.publisher@gmail.com/+91 9790120237
https://skrgcpublication.org/isbn/

SKRGC Publication
Read | Write | Teach

Title: INTRODUCTION TO BIG DATA

Author's: Dr. T. Arumuga Maria Devi
Dr. G. Heren Chellam
Dr. T. J. Benedict Jose
Dr. D. Sharmila
Mrs. A. Premalatha

Published by: SK Research Group of Companies,
Madurai 625003,
Tamil Nadu, India

Publisher's Address: 142, Periyar Nagar,
Madakulam, Madurai 625003,
Tamil Nadu, India

Edition Details: I

ISBN: 978-93-95341-11-0

Pages: 334

Price: ₹450/-

PREFACE

Dear Readers!

This book Introduction to Big Data comprising the current topics is an outcome of our perceptible effort to fulfill your intellectual requirements.

Hope it helps!

Explore all the topics!

ACKNOWLEDGMENTS

First and foremost, I want to express my gratitude and appreciation to my parents, my family, the Almighty, and all of my well-wishers for their blessings throughout the writing of my book and for helping it to be published successfully.

My heartfelt thanks to Senior Professor Dr. N. Krishnan, Professor & Head, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli for his continued academic assistance and encouragement to contribute towards intellectual achievements.

My sincere and heartfelt thanks to my husband Dr. P. Kumar, Associate Professor, and Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli for his support and encouragement in successful completion of this book.

I owe a great deal to the co-authors of this book Dr. G. Heren Chellam, Dr. T. J. Benedict Jose, Dr. D. Sharmila, Mrs. A. Premalatha for their hard work, helpful suggestions and commitment to complete this book.

- Dr. T. Arumuga Maria Devi

Our special thanks to all the faculty members of Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli for their support.

Our heartfelt thanks to Dr. M. Sathish Kumar, Publisher, SK Research Group of Companies (SKRGC Publication), Madurai for his effective co-ordination and the publication team for taking this book for publication.

Without the participation and support of several individuals, many of whose names may not be listed here, this undertaking would not have been able to be completed. We truly appreciate and gladly acknowledge their contributions.

Finally, thanks to all our family and friends for their moral support.

Dr. T. Arumuga Maria Devi,

Dr. G. Heren Chellam,

Dr. T. J. Benedict Jose,

Dr. D. Sharmila,

Mrs. A. Premalatha

CONTENTS

Chapter 1

BIG DATA

Introduction

Big data is a term that describes large, hard-to-manage volumes of data – both structured and unstructured – that inundate businesses on a day-to-day basis. But it's not just the type or amount of data that's important, it's what organizations do with the data that matters. Big data can be analyzed for insights that improve decisions and give confidence for making strategic business moves

According to Gartner, the definition of Big Data

"Big data, which is defined as high-volume, high-velocity, and diverse information assets, necessitates creative, cost-effective methods of information processing for improved understanding and decision-making.

What is Big Data? is answered succinctly by this definition. Big Data is the term used to describe complicated and substantial data collections that need to be processed and examined in order to produce useful information for businesses and organisations.

But there are several fundamental Big Data principles that will make it much easier to answer the question, "What is Big Data?"

It refers to a massive amount of data that keeps on growing exponentially with time.

It is so voluminous that it cannot be processed or analysed using conventional data processing techniques.

ISBN: 978-93-95341-11-0

Data mining, data archiving, data analysis, data sharing, and data visualisation are all included. The phrase refers to everything related to processing and analysing data, including data, data frameworks, tools, and processes.

The Development of Big Data

Large data sets have their roots in the 1960s and 1970s, when the first data centres and the relational database were being developed, and although the idea of big data is still a relatively recent one.

People started to understand how much data users were producing through Facebook, YouTube, and other online services around 2005. That same year, Hadoop (an open-source framework designed primarily to store and analyse massive data sets) was launched. At this time, NoSQL also started to gain prominence.

The emergence of big data was dependent on the creation of open-source frameworks like Hadoop (and more recently, Spark), which made massive data more manageable and less expensive to keep. Since then, the amount of big data has exponentially increased. Although not just people are producing vast volumes of data, users are nonetheless doing so.

More products and devices are now online thanks to the Internet of Things (IoT), which is gathering information on consumer usage trends and product performance. The development of machine learning has led to the creation of even more data.

Big data has gone a long way, but its utility is still in its infancy. The potential uses of big data have been further

ISBN: 978-93-95341-11-0

Chapter 2

HADOOP

HDFS, HADOOP AND HADOOP INFRASTRUCTURE

- Big Data Technology
- Distributed processing of large data sets
- Open source
- Map Reduce- Simple Programming model

Why Hadoop?

- Handles any data type
- Structured/unstructured
- Schema/no Schema
- High volume/Low volume
- All kinds of analytic applications
- Grows with business
- Proven with Petabyte scale
- Capacity & Performance grows
- Leverages commodity hardware to mitigate costs

Hadoop Features

- 100% Apache open source
- No Vendor locking
- Rich Eco system & community development
- To Derive compute value of all data
- More affordable cost-effective platform

Hadoop and Databases

RDMS (Relational Database Management System): RDBMS is a data model-based information management

system. Tables are used in RDBMS to store information. The table's columns stand in for data attributes, and each row represents a record. In contrast to other databases, RDBMS has a different structure for data and different methods for manipulating it. The ACID (atomicity, consistency, integrity, and durability) properties needed for database architecture are ensured by RDBMS. To store, manage, and retrieve data as rapidly and accurately as possible is the goal of RDBMS.

Hadoop is an open-source software framework used to run programmes and store data on a collection of inexpensive devices. It features a lot of storage space and a powerful processor. It has the ability to control several concurrent processes simultaneously. It is utilised in machine learning, data mining, and predictive analysis. Both structured and unstructured forms of data can be handled by it. Compared to standard RDBMS, it is more adaptable at storing, processing, and managing data. Hadoop, in contrast to conventional systems, enables several analytical operations on the same data at once. It is quite scalable and flexibly. Below in Table 2.1 of differences between Data Science and Data Visualization:

Table 2.1 RDBMS Vs Hadoop

S.NO.	RDBMS	HADOOP
1.	Structured database approach	Structured and Unstructured database approach
2.	Traditional row-column based databases, basically used for data storage,	An open-source software used for storing data and running applications or processes concurrently.

Chapter 3

MAP-REDUCE

Hadoop Mapreduce Framework

- Hadoop's MapReduce programming framework is used to process data in parallel. Hadoop's processing engine, MapReduce, handles and computes enormous amounts of data.
- It consists of the following two phases: Map and Reduce, as indicated. in Fig. 3.1.

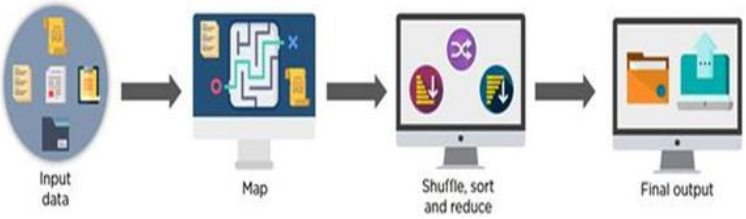

Fig. 3.1 Components of Hadoop

The generalized workflow of MapReduce is shown in Fig. 3.2.

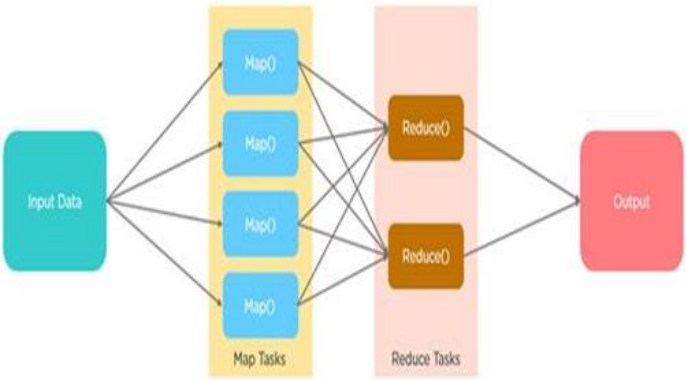

Fig. 3.2 Generalized Workflow of MapReduce

In order to accomplish parallel processing, MapReduce shown in Fig. 3.3.

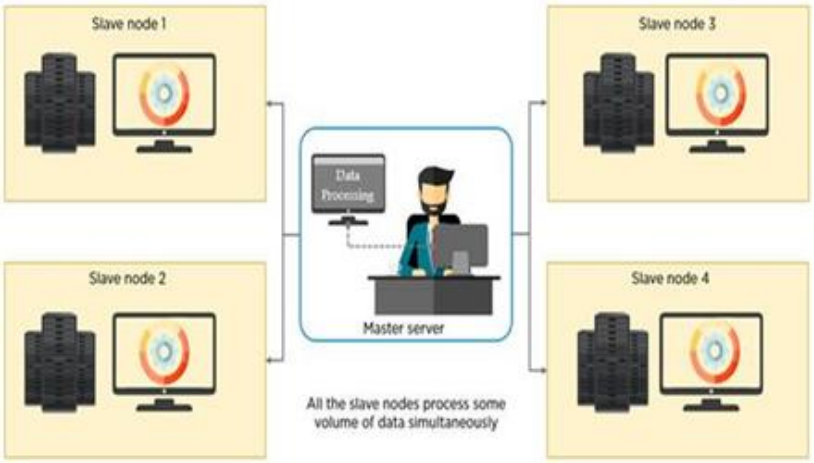

Fig. 3.3 Parallel Processing of MapReduce

The entire MapReduce workflow is displayed. in the Fig. 3.4:

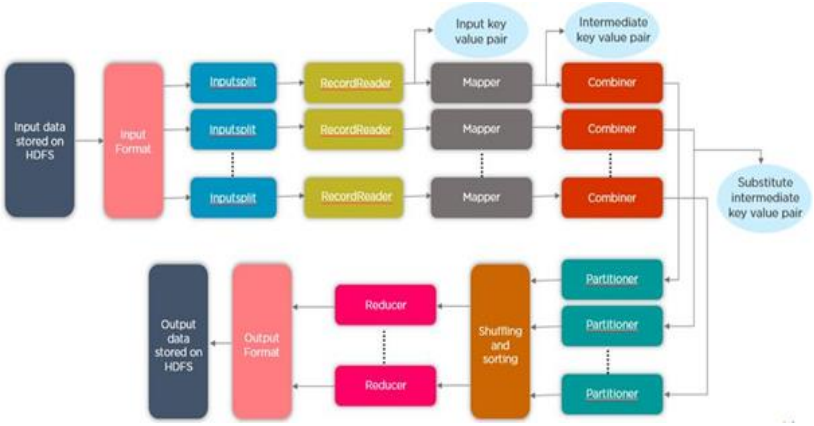

Fig. 3.4 Detailed Workflow of MapReduce

Chapter 4

HADOOP PROJECT ENVIRONMENT

4.1 Introduction to HBase

On top of the Hadoop file system, the distributed column-oriented database HBase was created. Similar to Google's large table, Base is a data model created to offer speedy random access to enormous amounts of structured data. It makes use of the Hadoop File System's fault tolerance (HDFS). It offers random real-time read/write access to data in the Hadoop File System and is a component of the Hadoop ecosystem. Data can be directly stored in HDFS or indirectly using HBase. Utilizing HBase, a data consumer randomly reads and accesses the data in HDFS.

Need for Hbase is as follows:

- Hadoop can perform only batch processing
- data will be accessed only in a sequential manner.
- That means one has to search the entire dataset even for the simplest of jobs.
- a new solution is needed to access any point of data in a single unit of time (random access)

HBase sits on top of the Hadoop File System and provides read and write access. Figure 4.1 shows the Hbase read and write

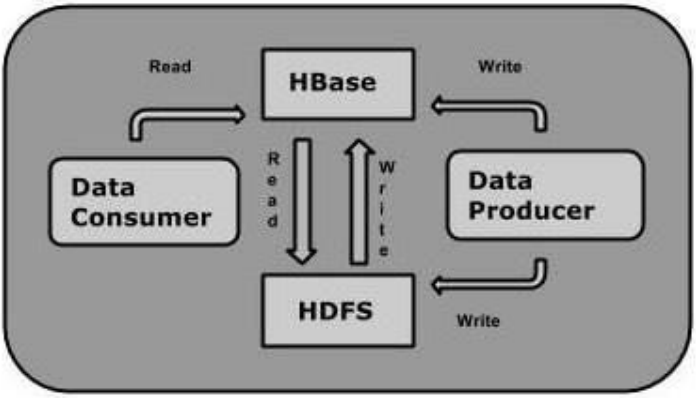

Fig. 4.1 Hbase Read/Write

Relational DBMS Vs Column oriented database

In a row-oriented indexed system, the primary key is the rowid that is mapped from indexed data. In Hbase, HBase is built on top of HDFS where the HBase data is stored in HFiles and HFiles are stored on HDFS. The purpose for introducing HBase was to enable random access to data for interactive querying which was not supported by pure HDFS implementation.

Column-oriented vs Row-oriented storages

Column-oriented Database	Row oriented Database
When the situation comes to process and analytics we use this approach. Such as **Online Analytical Processing** and it's applications.	**Online Transactional process** such as banking and finance domains use this approach. Online banking, Online airline ticket booking Sending a text message, Order entry Add a book to shopping cart

Chapter 5
INTRODUCTION TO HIVE AND PIG

Large datasets with a lot of volume, speed, and variety that are constantly growing are referred to as "big data" collections. Processing Big Data using conventional data management solutions is challenging. To address the issues of managing and analysing Big Data, the Apache Software Foundation developed the Hadoop framework.

Hadoop

A distributed system may store and handle Big Data using the open-source Hadoop platform. It consists of two modules: Hadoop Distributed File System and MapReduce (HDFS).

- MapReduce is a parallel programming model for handling massive volumes of organised, semi-structured, and unstructured data on vast clusters of affordable hardware.
- The Hadoop framework includes HDFS: Hadoop Distributed File System, which is used to store and process datasets. It offers a file system that can run on common hardware while being fault-tolerant.

The Hadoop ecosystem includes various side projects (tools), such Sqoop, Pig, and Hive, which support the Hadoop modules.

- Sqoop: This tool is utilised for data import and export between HDFS and RDBMS.
- Pig: This platform for procedural language develop ent is used to create scripts for MapReduce operations.
- Hive is a framework for creating SQL-style scripts that perform MapReduce processes.

ISBN: 978-93-95341-11-0

Note: There are other methods for carrying out MapReduce operations:

- The conventional method for organised, semi-structured, and unstructured data utilising the Java MapReduce software.
- Using Pig to script MapReduce, which can handle both structured and semi-structured data.
- The Hive Query Language for MapReduce, which uses Hive to process structured data.

Describe Hive

A Hadoop infrastructure utility for processing structured data is called Hive. To summarise Big Data, it sits on top of Hadoop and simplifies querying and analysis.

Initially created by Facebook, Hive was later taken up and further developed as an open source project under the name Apache Hive by the Apache Software Foundation. It is utilised by various businesses. Amazon utilises it, as an illustration, in Amazon Elastic MapReduce.

Bee is not a correct

An online transaction processing (OLTP) design, a relational database, and a language for real-time queries and row-level updates

Qualities of a Hive

It is intended for OLAP; it saves schema in a database and processed data in HDFS.

- It offers a familiar, quick, scalable, and extensible SQL-like querying language called HiveQL or HQL.

For Full Book Contact

Dr.T.Arumuga Maria Devi

Assistant Professor,

Centre for Information Technology and Engineering,

Manonmaniam Sundaranar University,

Tirunelveli, Tamil Nadu, India

Email: deviececit@gmail.com

Email: arumugamariadevi@msuniv.ac.in

Mobile: +91 8667899606

About the Authors

Dr. T. Arumuga Maria Devi B.E., M.Tech., Ph.D., Received B.E degree in Electronics & Communication Engineering from Manonmaniam Sundaranar University, Trinelveli, Tamil Nadu, India in 2003. Completed M.Tech degree in computer & Information Technology from Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India in 2005. Received Ph.D Degree in Information Technology Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India in 2012. She is the Assistant Professor of Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli since 2005. Her Research Includes Signal Processing, Remote Communication, Multimedia and Mobile Computing.

Dr. G. Heren Chellam is currently working as an Assistant Professor in the Department of Computer Science, Rani Anna Government College for Women, Tirunelveli-627 008. She received her MCA in the Dept. of Computer Science and Engineering, Annamalai University, Chidambaram and M.Phil Degree in Mother Teresa Women's University, Kodaikanal. She received her Doctorate in Computer Applications from Manonmanium Sundaranar University. She has published papers in many National and InternationalJournalsand presented papers inInternational and National Conferences. She has attended many Workshops, Seminars and Faculty Development Programs. Her research interest are in the field of Neural Networks, Digital Image Processing, Pattern Recognition. She has 29 years of teaching experience.

Dr. T. J. Benedict Jose received his B.Sc degree in Computer Science from St. Xavier's College (Autonomous), Palayamkottai, Tamilnadu, India in the year 2004. He obtained his M.Sc degree in the field of Computer Science from St.Xavier's College (Autonomous), Palayamkottai, Tamilnadu, India in the year 2006. He received his Master of Philosophy (M.Phil) in Computer Science from Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India in the year 2008. He obtained his Doctor of Philosophy (Ph.D) in Computer Science in the area of Digital Image Processing from Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India in the year 2017. Previously he worked as Assistant Professor in the Department of Computer Science, St. Xavier's College, Palayamkottai, Tamilnadu, India. At present he is working as Assistant Professor in the Department of Computer Applications, Government Arts and Science College, Palkulam, Kanyakumari, Tamilnadu, India. He is also the Red Ribbon Club Programme Officer in his college. He has been approved as a Research Supervisor by Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India.

Dr. D. Sharmila received her B.Sc degree in Computer Science from Womens Christian College Nagercoil. She obtained her MCA & MPhil degrees from Madurai Kamaraj University. She received her Doctor of Philosophy (Ph.D) in Computer Science in the area of Digital Image Processing from Mother Teresa Womens University, Kodaikanal. Previously she worked as Professor/HOD in the Department of Master of Computer Applications, Lord Jegananth College of Engg & Technology. At present she is working as an Assistant Professor in the Department of Computer Applications, Government Arts and Science College, Palkulam, Kanyakumari. She has published many research papers in International Journals and Conferences. She has presented many papers in National and International Conferences. She has attended many Faculty Development Programmes, Seminars, Workshops and Training Courses.

Mrs. A. Premalatha received her B.Sc degree in Computer Science from Rani Anna Government College for Women, Tirunelveli. She obtained her MCA degree from Manonmananiam Sundaranar University and M.Phil degree from Madurai Kamaraj University. Previously She worked as an Assistant Professor in the Department of Master of Computer Applications, St.Xavier's College Palayamkottai. At present She is working as an Assistant Professor in the Department of Computer Science, Rani Anna Government College for Women, Tirunelveli. She has presented many papers in National and International Conferences. She has attended many Faculty Development Programmes, Seminars, Webinar, Workshops and Training Courses.

Published by

SK Research Group of Companies

The International Journal & Book - SKRGC Publication

142, Periyar Nagar, Madakulam

Madurai - 625003, Tamil Nadu, India

skrgc.publisher@gmail.com/+91 9790120237

https://skrgcpublication.org/isbn/

SKRGC Publication

Read | Write | Teach

ISBN 939534111-4

9 789395 341110