Big Data Analytics with Hadoop

Dr. G. UMA DEVI

Professor,

Computer Science & Engineering

University of Engineering &

Management, Jaipur

Jupiter Publications Consortium

www.jpc.in.net

TEXTBOOK ON

Big Data Analytics with Hadoop

Authors:
Dr. G. UMA DEVI

First Published: 10th February 2022

ISBN: 978-93-91303-31-0

Digital Object Identifier (DOI):
https://doi.org/10.47715/JPC.B.79.2022.
9789391303310

Pages: 232
Price: 375/-

Publisher & Imprint:
Jupiter Publications Consortium
22/102, Second Street, Virugambakkam
Chennai, Tamil Nadu, India.
Email: director@jpc.in.net

COPYRIGHT DISCLAIMER

Publisher's Address:
Jupiter Publications Consortium
22/102, Second Street, Virugambakkam
Chennai, Tamil Nadu, India.
Email: director@jpc.in.net

TITLE VERSO

Title of the Book:

Big Data Analytics with Hadoop

Author's Name:

Dr. G. UMA DEVI

Published By:

Jupiter Publications Consortium

Publisher's Address:

22/102, Second Street, Venkatesa Nagar,

Virugambakkam

Chennai 600 092. Tamil Nadu, India.

Printer's Details:

Jupiter Publications Consortium

Edition Details:

FIRST EDITION

ISBN: 978-93-91303-31-0

Dedicated to my beloved family

This Page Intentionally Left Blank

PREFACE

The phrase “big data” has gained popularity to refer to an exciting new collection of tools and approaches for developing contemporary, data-driven applications that revolutionise the way the world computes. To the dismay of statisticians, this ubiquitous word seems to be widely utilised to include applying well-known statistical methods to big datasets for predictive purposes. Although the term “big data” has become a catchphrase, the reality is that current, distributed processing methods enable studies of datasets far more significant than those previously analysed, with astonishing results. However, distributed computing alone does not imply data science. The combination of constantly growing datasets created by the Internet and the insight that these data sets may power prediction models has resulted in a new economic paradigm known as data products. Stunning data modelling accomplishments across vast, diverse datasets.

Hadoop has developed from a cluster computing abstraction to a big data operating system by offering a framework for distributed data storage and parallel processing. Spark expanded on these concepts and simplified cluster computing for data scientists. However, data scientists and analysts unfamiliar with distributed computing may believe that these technologies are designed for programmers rather than analysts. This is because paradigm changes in handling and computing data in a parallel rather than sequential approach are required. This textbook is designed to help Engineering, and Technological Undergraduates comprehend the principles, methods, and procedures involved in big data analytics with Hadoop and serve as a springboard for further exploring subject areas.

Dr. G. UMA DEVI

This Page Intentionally Left Blank

SYLLABUS

Contents:

Module-1: Introduction to Big Data and Hadoop: Types of Digital Data, Introduction to Big Data, Big Data Analytics, History of Hadoop, Apache Hadoop, Analysing Data with Unix tools, Analysing Data with Hadoop, Hadoop Streaming, Hadoop Echo System, IBM Big Data Strategy, Introduction to Infosphere Big Insights and Big Sheets.

Module-2: HDFS (Hadoop Distributed File System) The Design of HDFS, HDFS Concepts, Command Line Interface, Hadoop file system interfaces, Data flow, Data Ingest with Flume and Scoop and Hadoop archives, Hadoop I/O: Compression, Serialization, Avro, and File-Based Data structures.

Module-3: Map Reduce Anatomy of a Map Reduce Job Run, Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce Types and Formats, Map Reduce Features.

Module-4: Hadoop Eco System Pig: Introduction to PIG, Execution Modes of Pig, Comparison of Pig with Databases, Grunt, Pig Latin, User Defined Functions, Data Processing operators. Hive: Hive Shell, Hive Services, Hive Metastore, Comparison with Traditional Databases, HiveQL, Tables, Querying Data and User Defined Functions. HBase: HBasics, Concepts, Clients, Example, Hbase Versus RDBMS. Big SQL: Introduction

Module-5: Machine Learning: Introduction, Supervised Learning, Unsupervised Learning, Collaborative Filtering. Big Data Analytics with Big R.

Textbooks:

1.Tom White “Hadoop: The Definitive Guide” Third Edit on, O’reily Media, 2012.

2.Seema Acharya, Subhasini Chellappan, “Big Data Analytics” Wiley 2015.

Reference Books:

1.Michael Berthold, David J. Hand, “Intelligent Data Analysis”, Springer, 2007.

2.Jay Liebowitz, “Big Data and Business Analytics” Auerbach Publications, CRC press (2013)

3.Tom Plunkett, Mark Hornick, “Using R to Unlock the Value of Big Data: Big Data Analytics with Oracle Enterprise and Oracle R Connector for Hadoop”, McGraw-Hill/Osborne Media (2013), Oracle press

TABLE OF CONTENTS

Module-1: Introduction to Big Data and Hadoop

Module-2: HDFS (Hadoop Distributed File System)

Module-3: Map Reduce Anatomy of a Map Reduce

1 Module

1.0 Introduction to Big Data and Hadoop

1.1 Types of Digital Data

Three types of big data exist:

- Structured Data
- Unstructured Data
- Semi-Structured Data

While these three words are theoretically relevant to all levels of analytics, they are critical in the context of big data. Understanding where raw data originates and how it must be processed prior to analysis becomes even more critical when dealing with large amounts of big data. Because there is so much of it, information extraction must be efficient to justify the effort.

The data's structure dictates how to work with it and what insights it may provide. Before data can be evaluated, it must undergo an extract, transform, and load (ETL) process. It is a literal term: data is gathered, structured so that an application can

access it, and then saved for later use. Each data structure requires a unique ETL method.

Let us define what they signify and how they relate to big data analytics.

1.1.1 Structured Data

The most straightforward kind of data to deal with is structured data. It is highly ordered, with parameters defining its size.

Consider spreadsheets; each item of data is organised into rows and columns. Certain variables specify specific components that are readily discoverable.

It is comprised of all our quantitative data:

- Age
- Billing
- Contact
- Address
- Expenses
- Numbers of debit/credit cards

Because structured data is already composed of concrete numbers, it makes data collection and sorting considerably more uncomplicated for software.

Structured data is organised according to schemas; road maps to data points effectively. These schemas define the location and meaning of each item.

A payroll database will include employee identifying information, pay rates, hours worked, and the way money is distributed, among other things. Each of these dimensions will be defined by the schema for the application that uses it. The software will not

have to sift through data to ascertain its meaning; it can immediately gather and analyse it.

1.1.2 Unstructured Data

Not all data is as neatly packaged and organised with use instructions as structured data. The general view is that no more than 20% of all data is organised.

Thus, what constitutes the remaining four-fifths of all available information? Due to the lack of structure, naturally, this is unstructured data.

We may deduce why it comprises such a large portion of the current data library. Almost all activities performed on a computer create unstructured data. Nobody is transcribing their phone conversations or categorising each tet they send.

While organised data saves time throughout an analytical process, the time and effort required to make unstructured data readable are inconvenient.

The ETL method is straightforward for structured data. It is cleaned and checked before the information is loaded into a database during the converting step. Hover, with unstructured data, this second step becomes far more challenging.

To get anything approaching valuable data, the dataset must be interpretable. Hover, the work might be much more beneficial than the effortless alternative to unstructured data processing. In athletics, as they say, we get what we put in.

1.1.3 Semi-structured data:

Semi-structured data straddles the structured and unstructured worlds. Typically, this converts to unstructured data with associated metadata. This may be intrinsic data acquired during

the collection process, such as time, location, device ID stamp, email address, or a semantic tag added to the data subsequently.

Assume we capture a photograph of our pet using our phone. It routinely accounts for the date and time the photograph was shot. The GPS coordinates now of capture and our device's ID. Our account information is associated with the file using a b-based storage service, such as iCloud.

When we send an email, the time it was sent, the email addresses to and from, the internet protocol address of the device from which the email was sent, and other bits of information are associated with the email's content.

In both cases, the actual content (i.e., the pixels that comprise the picture and the text that comprise the email) is unstructured. However, some components enable the data to be categorised according to qualities.

1.2 Big Data

1.2.1 What is Big Data?

Big Data refers to large amounts of massive data yet increases exponentially in size over time. Data is so extensive and complicated that no usual data management methods can effectively store or handle it. Big data is likewise data, but it is enormous.

Examples of Big Data.

- Discovering consumer shopping habits
- Finding new customer leads
- Fuel optimisation tools for the transportation industry
- Live road mapping for autonomous vehicles
- Monitoring health conditions through data from wearables

- Personalised health plans for cancer patients
- Personalised marketing
- Predictive inventory ordering
- Real-time data monitoring and cybersecurity protocols
- Streamlined media streaming
- User demand prediction for ridesharing companies

Big Data gives us unprecedented insights and opportunities, but it also raises concerns and questions that must be addressed:

Data privacy: The Big Data we now generate contains a lot of information about our personal lives, much of which we have a right to keep private

Data security: Even if we decide we are happy for someone to have our data for a purpose, can we trust them to keep it safe?

Data discrimination: When everything is known, will it become acceptable to discriminate against people based on data we have on their lives? We already use credit scoring to decide who can borrow money, and insurance is heavily data-driven.

Data quality: Not enough emphasis on quality and contextual relevance. The trend with technology is collecting more raw data closer to the end user. The danger is data in raw format has quality issues. Reducing the gap between the end user and raw data increases issues in data quality.

Facing up to these challenges is an important part of Big Data, and they must be addressed by organisations who want to take advantage of data. Failure to do so can leave businesses vulnerable, not just in terms of their reputation, but also legally and financially.

1.2.2 Big Data Characteristics

Three attributes stand out as defining Big Data characteristics:

1. **Volume**: Huge volume of data: Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.

2. **Variety:** Complexity of data types and structures: Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.

3. **Velocity:** Speed of new data creation and growth: Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

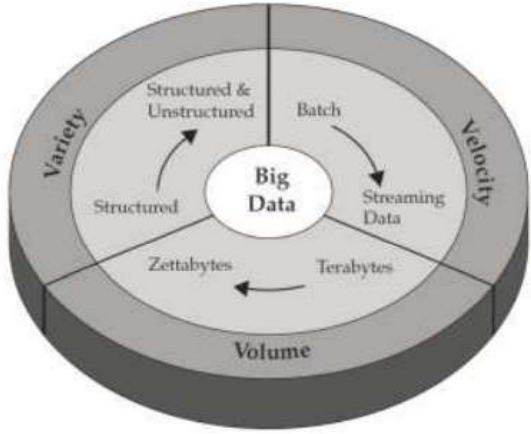

Fig. 1: Big data V's

This can be extended to add a fourth and fifth V's

4. Veracity: It is equivalent to quality. We have all the data, but could we be missing something? Are the data "clean" and accurate? Do they really have something to offer?

5. Value: There is another V to take into account when looking at big data: Value having access to big data is no good unless we can turn it into value. Companies are starting to generate amazing value from their big data.

1.3 Big Data Analytics

Big data analytics studies massive volumes of data in order to unearth previously unknown patterns, correlations, and other insights. With today's technology, it can analyse our data and get answers nearly instantly - a process that would take much longer and be less efficient with more conventional business intelligence solutions.

1.3.1 Big data analytics' history and development

The concept of big data has emerged for more than two decades; most firms now recognise that by capturing all data that enters their company, they can apply analytics and derive tremendous value. However, even in the 1950s, decades before the phrase "big data" were coined, companies were uncovering insights and patterns using rudimentary analytics.

However, the new advantages that big data analytics delivers include speed and efficiency. Whereas a few years ago, a firm would have collected data, performed analytics, and discovered insights for future choices, today's business can find insights for current decisions. The capacity to work more quickly — and remain nimble – provides firms with a competitive advantage they previously lacked.

1.3.2 What is the significance of big data analytics?

Big data analytics enables businesses to harness their data and use it to discover new possibilities. This results in better-informed company decisions, more effective operations, more earnings, and happier consumers. Tom Davenport, IIA Director of Research, interviewed more than 50 firms for his paper Big Data in Big Companies to better understand how they employed big data. He discovered that they added value in the following ways:

Cost savings: When it comes to storing massive volumes of data, big data technologies such as "Hadoop and cloud-based analytics provide considerable cost savings". They may also help uncover more effective methods of conducting business.

Decision-making that is more rapid and accurate: Businesses can evaluate information instantaneously – and make choices based on what they have learned – thanks to the speed of Hadoop and in-memory analytics and the capacity to study new sources of data.

Introducing new goods and services: With the capacity to assess consumer demands and satisfaction through analytics comes the ability to deliver on client desires. Davenport notes that more businesses are developing new goods to fulfil client requirements due to big data analytics.

1.4 History of Hadoop

Hadoop is an Apache Software Foundation-managed open-source framework developed in Java for storing and analysing massive information on commodity hardware clusters. There are primarily two issues with big data. The first is to store such a massive quantity of data, and the second is to process it. Thus, Hadoop serves as a solution to the issue of big data, namely the storage and processing of large amounts of data with specific additional capabilities. Hadoop is composed chiefly of Hadoop Distributed File System (HDFS) and Yet Another Resource Negotiator (YARN).

1.4.1 Hadoop's Historical Background

Hadoop was originated in 2002 and founded by Doug Cutting and Mike Cafarella as part of their work on the Apache Nutch project. The Apache Nutch project was tasked with developing a search engine system capable of indexing one billion documents. After doing extensive study on Nutch, they determined that such a system would cost roughly half a million dollars in hardware and a monthly operating cost of approximately $30 000, which is rather costly. As a result, they discovered that their project design would not cope with the billions of online pages. As a result, they sought a practical solution to minimise the implementation cost and store and process massive datasets.

In 2003, they discovered a document describing the design of Google's distributed file system, GFS (Google File System), which Google released to store massive data collections. They now see that this research can resolve their storage of huge files created by web crawling and indexing operations. However, this research provided just a partial answer to their difficulty. In 2004, Google produced another article on MapReduce technology used to handle such massive datasets. For Doug Cutting and Mike Cafarella, this report was another half-solution to their Nutch project. Both approaches (GFS and MapReduce) were previously only available as white papers at Google. Google did not use any of these approaches. Doug Cutting recognised through his work on Apache Lucene (a free and open-source information retrieval software library that Doug Cutting first wrote in Java in 1999) that open-source is an excellent approach to sharing technology with a broader audience. As a result, he began working with Mike Cafarella on open-source implementations of Google's algorithms (GFS & MapReduce) in the Apache Nutch project.

Cutting discovered in 2005 that Nutch is confined to clusters of between 20 and 40 nodes. He quickly saw two issues:

Nutch would not reach its full potential until it could run stably on more prominent clusters (b), which seemed unachievable with just two workers (Doug Cutting & Mike Cafarella). The engineering work in the Nutch project was far more than he anticipated. As a result, he began looking for work with a firm willing to invest in their efforts. Moreover, he discovered that Yahoo! has a sizable engineering staff ready to work on this project.

Thus, Doug Cutting joined Yahoo in 2006 and the Nutch project. With the assistance of Yahoo, he wanted to present the world with an open-source, dependable, and scalable computing architecture. Thus, the first separate the distributed computing components of Nutch and establishes a new project at Yahoo. Hadoop (He provided the name Hadoop since it was the name of a yellow toy elephant that Doug Cutting's kid had. because it was simple to say and was a one-of-a-kind term.) Now he desired to optimise Hadoop's performance on hundreds of nodes. As a result, he began working on Hadoop using GFS and MapReduce.

Yahoo began utilising Hadoop in 2007 after successfully testing it on a 1000-node cluster. In January 2008, Yahoo donated Hadoop to the Apache Software Foundation as an open-source project (Apache Software Foundation). Additionally, in July 2008, the Apache Software Foundation successfully tested Hadoop on a 4000-node cluster. Hadoop was successfully tested in 2009 for sorting a PB (PetaByte) of data in less than 17 hours for processing billions of queries and indexing millions of web pages. Moreover,

Doug Cutting Departed Yahoo to join Cloudera to take on the task of bringing Hadoop to new sectors.

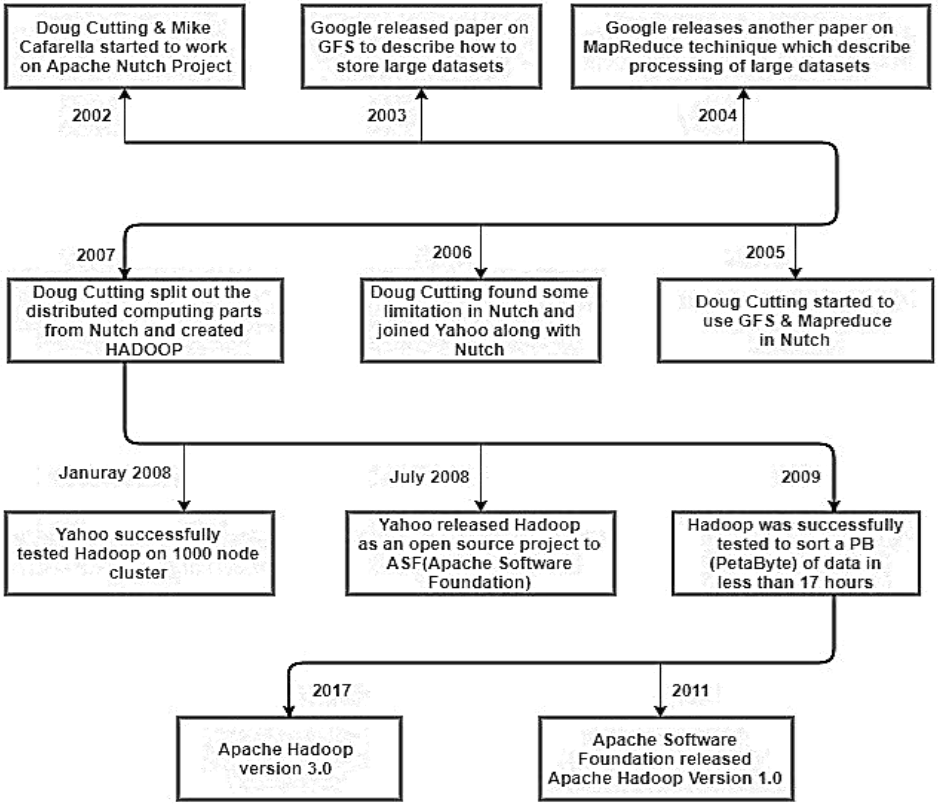

Fig. 2: Hadoop History

- Apache Hadoop version 1.0 was published by the Apache Software Foundation in December 2011.
- Additionally, Version 2.0.6 was released in August 2013.
- Furthermore, as of December 2017, we have Apache Hadoop version 3.0.

1.5 Apache Hadoop

Apache Hadoop is a free and open-source platform for storing and processing massive information ranging in size from gigabytes to petabytes. Rather than storing and processing data on a single colossal computer, Hadoop enables clustering many computers to analyse enormous datasets in parallel.

1.5.1 Four significant modules of Hadoop:

HDFS — A distributed file system that works on commodity or low-end hardware. HDFS outperforms conventional file systems in data performance, fault tolerance, and native support for massive datasets.

YARN — manages and monitors cluster nodes and resource utilisation. It automates the scheduling of jobs and tasks.

MapReduce — A framework that enables parallel computing on data by programmes. The map job turns the input data into a dataset calculated in key-value pairs. It is reducing tasks that consume the output of the map task in order to aggregate it and produce the desired result.

Hadoop Common — Provides a set of shared Java libraries utilised by all modules.

1.5.2 How Hadoop Operates

Hadoop simplifies the process of using all the data storage capacity available in cluster computers and executing distributed algorithms against massive volumes of data. Hadoop offers the

foundation for the development of additional services and applications.

Applications that gather data in various forms may upload data to the Hadoop cluster by connecting to the NameNode through an API function. The NameNode maintains the directory structure of each file and the location of "chunks" for each file, which is duplicated among DataNodes. To launch a job that queries the data, supply a MapReduce job consisting of several maps and reduce jobs executing on the data stored in HDFS across the DataNodes. Each node executes map tasks against the specified input files, while reducers execute to aggregate and arrange the final output.

Due to Hadoop's flexibility, the ecosystem has evolved tremendously over the years. Today, the Hadoop ecosystem comprises a variety of tools and applications that aid in the collection, storage, processing, analysis, and management of large amounts of data. Several of the most prominent uses include the following:

Spark — An open-source distributed processing technology often used to handle large amounts of data. Apache Spark provides general batch processing, streaming analytics, machine learning, graph databases, and ad hoc queries through in-memory caching and efficient execution.

Presto — A distributed SQL query engine geared for low-latency, ad-hoc data processing. It adheres to the ANSI SQL standard, which includes sophisticated searches, aggregations, joins, and window functions. Presto can handle data from various sources,

6 Bibliography

6.0 Bibliography

1. Agneeswaran, V. S. (2014). *Big data analytics beyond hadoop: real-time applications with storm, spark, and more hadoop alternatives*. FT Press.
2. Anuradha, J. (2015). A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia computer science*, *48*, 319-324.
3. Augustine, D. P. (2014). Leveraging big data analytics and Hadoop in developing India's healthcare services. *International Journal of Computer Applications*, *89*(16), 44-50.
4. Azeroual, O., & Fabre, R. (2021). Processing big data with apache hadoop in the current challenging era of COVID-19. *Big Data and Cognitive Computing*, *5*(1), 12.
5. Dhyani, B., & Barthwal, A. (2014). Big data analytics using Hadoop. *International Journal of Computer Applications*, *108*(12).
6. Gupta, B., & Jyoti, K. (2014). Big data analytics with hadoop to analyze targeted attacks on enterprise data. *(IJCSIT) International Journal of Computer Science and Information Technologies*, *5*(3), 3867-3870.
7. Kousalya, D. R., & Sindhupriya, T. (2017). Review on big data analytics and Hadoop framework. *International Journal of Innovations in*

Scientific and Engineering Research (IJISER), ISSN: 2347-9728 (print), *4*(3MAR), 101.

8. Kumar, Y., Sood, K., Kaul, S., & Vasuja, R. (2020). Big data analytics and its benefits in healthcare. In *Big data analytics in healthcare* (pp. 3-21). Springer, Cham.
9. Malhotra, J., Sethi, J. K., & Mittal, M. (2021). Analysis of big data using two mapper files in hadoop. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, *13*(1), 69-77.
10. Niu, Y., Ying, L., Yang, J., Bao, M., & Sivaparthipan, C. B. (2021). Organizational business intelligence and decision making using big data analytics. *Information Processing & Management*, *58*(6), 102725.
11. Priyanka, E. B., Thangavel, S., Meenakshipriya, B., Prabu, D. V., & Sivakumar, N. S. (2021). Big data technologies with computational model computing using hadoop with scheduling challenges. In *Deep Learning and Big Data for Intelligent Transportation* (pp. 3-19). Springer, Cham.
12. Rehman, A., Naz, S., & Razzak, I. (2021). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems*, 1-33.
13. Wu, W., Lin, W., Hsu, C. H., & He, L. (2018). Energy-efficient hadoop for big data analytics and computing: A systematic review and research insights. *Future Generation Computer Systems*, *86*, 1351-1367.
14. Zakir, J., Seymour, T., & Berg, K. (2015). Big Data Analytics. *Issues in Information Systems*, *16*(2).
15. Zhang, X., & Wang, Y. (2021). Research on intelligent medical big data system based on Hadoop and blockchain. *EURASIP Journal on Wireless Communications and Networking*, *2021*(1), 1-21.