Learning Guide Unit 2

Site:	University of the People	Printed by:	Ryohei Hayashi
Course:	CS 3440-01 Big Data - AY2025-T3	Date:	Thursday, 30 January 2025, 5:26 PM
Book:	Learning Guide Unit 2		

Description

Learning Guide Unit 2

Table of contents

Overview

Introduction

Reading Assignment

Discussion Assignment

Written Assignment

Learning Journal

Self-Quiz

Checklist

Overview

UNIT 2: Big Data Tools, Techniques, and Systems

Topic

- Introduction to tools and techniques used in Big data: Hadoop, MongoDB, Apache Spark

Learning Objectives

By the end of this Unit, you will be able to:

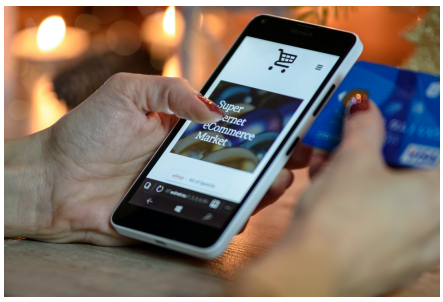
1. Describe the three main components of HADOOP.
2. Compare and Contrast HADOOP and Apache Spark.
3. Discuss the usage of MongoDB in Big Data sphere.

Tasks

- Peer assess Unit 1 Written Assignment
- Read the Learning Guide and Reading Assignments
- Participate in the Discussion Assignment (post, comment, and rate in the Discussion Forum)
- Complete and submit the Written Assignment
- Make entries to the Learning Journal
- Take and submit the Self-Quiz

Introduction

In this world of social media and advancements in data collection practices companies need a way to collect, organize, analyze and report on highly enormous data sets being collected daily. HADOOP has come into the picture rescue and has been a dominant player in the Big Data arena. It's a very stable platform and processing power outshines standard computing technology.



Also, with the popularity of online shopping companies like Amazon, and eBay is collecting millions of rows of transaction data daily, which using HADOOP and other Big Data processing platforms can give them insights into customers buying habits and provide them with valuable marketing information for business decision functions.

By analyzing this data, organizations can learn trends about the information they are measuring and the people generating it. This extensive data analysis aims to provide more customized service and increased efficiencies in whatever industry the data is collected from. Though there are many Big data tools we are going to focus on: HADOOP, Apache Spark, and MongoDB. In this unit, we will focus on their history and background, as well as the workings of HADOOP, Adobe Spark, and MongoDB for processing and storing big data. You will gain a better understanding of the structure and use of Adobe Spark and how it is closely related to and structured to Hadoop. Furthermore, you will learn about MongoDB and how it uses NoSQL for data processing and storing data for use in big data analysis.

HADOOP

HADOOP is the open-source software framework Apache Hadoop was developed for the storage and large-scale processing of data sets on clusters. Hadoop is an Apache top-level project built and used by a community of contributors and users from around the globe. Apache Hadoop is licensed under the Apache License 2.0, meaning users have not to worry about infringing on any software patents and are free to use the software as needed (GeeksforGeeks, 2021). Apache HADOOP is used extensively with big data to help store, organize and analyze data sets that are normally too large for other tools to manage.

Apache Spark

Apache Spark is a cluster computing technology that was designed for providing fast processing time for computations. It is loosely based on Hadoop MapReduce and improves on the MapReduce model by using it for different types of computations and interactive queries (Tutorials Point, 2022).

Like Apache HADOOP, Spark was developed to help with the management, organization, and analysis of big data datasets. It functions in the same manner as Hadoop given its development is based on the MapReduce model which is central to the Hadoop architecture.

MongoDB

“MongoDB is an open-source document database that is built on a horizontal scale-out architecture that uses a flexible schema for storing data” (Mongo DB, 2022). MongoDB was built on a scale-out architecture from its founding, which is a structure that allows many small machines to work together to create fast systems and handle enormous amounts of data (Day, 2022). MongoDB is built in such a manner that allows it to handle large amounts of data, with access to the data done quickly. It is used by many big data analysts and developers given the normalized structure in which it stores data.

Throughout this unit, you will learn about Hadoop, Apache, and MongoDB and the roles they each plan in the collection and analysis of big data. You will get a good understanding of the background of each of these technologies, and apply what you have learned in the written assignment and learning journal.

References

Apache spark - introduction. (n.d.). Tutorials Point. from https://www.tutorialspoint.com/apache_spark/apache_spark_introduction.htm

Day, F. (2022, June 17). Top 10 NOSQL databases for data science. *Classes near me blog*. https://www.nobledesktop.com/classes-near-me/blog/top-nosql-databases-for-data-science

History of hadoop - the complete evolution of hadoop ecosytem. (2021, August 25). DataFlair. https://data-flair.training/blogs/hadoop-history/

Vivek5252. (2021, June 29). *Hadoop - introduction.* GeeksforGeeks. Retrieved: August, 7, 2022, from https://www.geeksforgeeks.org/hadoop-introduction/?ref=lbp

Why Use MongoDB and when to use It? (2022, February 8). MongoDB.com. https://www.mongodb.com/why-use-mongodb

Reading Assignment

Read through the following to better understand the background and structure of Hadoop, Apache Spark and MongoDB. These readings will provide you with information about each of the three items for you to gain a better understanding of how and why they are used with big data and some of the elementary components of each.

Aggarwal, A. (2019, January 18). *Hadoop – history or evolution.* GeeksforGeeks. Retrieved: August, 7, 2022, from https://www.geeksforgeeks.org/hadoop-history-or-evolution/

- This article provides a timeline of Hadoop's history.

Introduction to apache spark. (n.d.). AWS. https://aws.amazon.com/big-data/what-is-spark/

This webpage speaks about Apache Spark's definition, history, how it works, benefits and use cases.

Jones, N. (2022, March 8). Big data and the 5 major advantages of Hadoop. ITpro Portal. https://www.itproportal.com/2013/12/20/big-data-5-major-advantages-of-hadoop/

- This website explains the advantages of HADOOP when used with Big data.

Kerzner, M., &Maniyam, S. (2016). *Hadoop Illuminated.* Elephant Scale LLC. https://elephantscale.com/wp-content/uploads/2020/05/hadoop-illuminated.pdf licensed by CC BY-NC-SA 3.0

- This text explains what big data is, and how Hadoop is used with big data in different functional areas such as finance, health care, travel and retail settings. Read the following:

1. **Chapter 4, from pages 9 - 11 -** This chapter goes into how Hadoop is built, as well as business cases for using Hadoop.
2. **Chapter 7, from pages 16 - 23 -** This chapter goes into detail on why you should use Hadoop, as well as some of the component parts that make up this system.
3. **Chapter 15, from pages 60 -61 -** This chapter presents in brief some of the challenges with Hadoop.

MongoDB history. (n.d.). Quick programming tips. https://www.quickprogrammingtips.com/mongodb/mongodb-history.html

- This article provides a detailed history of MongoDB from its origin to present release.

O'Reilly Radar Team. (2012). *Big data now.* O'Reilly Media, Inc. http://cdn.oreillystatic.com/oreilly/radarreport/0636920028307/Big_Data_Now_2012_Edition.pdf

- Read Chapter 3: Big data tools, techniques and strategies. This chapter presents the reader with a big data predictive modeling technique called the drivetrain approach for use in analyzing and making predictions using big data as its base.

Pedamkar, P. (2022). *Advantages of Hadoop*. Educba. https://www.educba.com/advantages-of-hadoop/

- This website explains the advantages of HADOOP when used with Big data.

Salloum, S., Dautov, R., Chen, X., Peng, P. X., & Huang, J. Z. (2016). Big data analytics on apache spark. *International Journal of Data Science and Analytics, 1*(3), 145-164. DOI: https://doi.org/10.1007/s41060-016-0027-9

- This article covers the key components and features of Apace Spark. It specifically shows what Apache Spark has for the design and implementation of algorithms for big data.

Vivek5252 (2021, July 29). Hadoop - Introduction. GeeksforGeeks. Retrieved: August, 7, 2022, from https://www.geeksforgeeks.org/hadoop-introduction/?ref=lbp

- This article gives an introduction of Hadoop.

Optional Reading

learntek. (2017, April 11). *Why MongoDB is so important for big data.* https://www.learntek.org/blog/mongodb-important-big-data/

- This article explains why MongoDB is important for Big data.

Video Resources

Simplilearn. (2021, January 21). *Hadoop in 5 minutes | what is hadoop? | introduction to hadoop | hadoop explained* [Video]. YouTube. https://www.youtube.com/watch?v=aReuLtY0YMI

- This video presents an explanation of the term Big Data to help you under the importance of Hadoop and it use in the big data space.

Simplilearn. (2021, March 16). *What is mongoDB? | what is mongoDB and how It works | mongoDB tutorial for beginners |* [Video]. YouTube. https://www.youtube.com/watch?v=SnqPyqRh4r4

- This video will give you a good understanding of the working and applications of the DBMS. The video explains what MongoDB is, its salient features, and its applications.

Discussion Assignment

MongoDB is used by many companies that deal with big data due to its unique structure and processing speed. Discuss why MongoDB is used within the Big Data sphere and how it can help companies collect and report on captured data.

Your Discussion should be a minimum of **200 words** in length and not more than **300 words**. Please include a word count. Following the APA standard, use references and in-text citations for the textbook and any other sources.

Use APA citations and references for the textbook and any other sources used; you should use at least 1 APA citation and reference, but you can use more if needed. Refer to the UoPeople APA Tutorials in the LRC for help with APA citations. You are required to post an initial response to the question/issue presented in the Forum and then respond to at least 3 of your classmates' initial posts. You should also respond to anyone who has responded to you. Don't forget to rate the postings of your classmates according to the Rating Guidelines. Review the Discussion Forum rating guidelines to see how your classmates will be rating your post.

After posting an appropriate, meaningful, and helpful response to your three classmates, you must rate their posts on a scale of 0 (unsatisfactory) to 10 (excellent).

10 (A) - Excellent, substantial, relevant, insightful, enriching, and stimulating contribution to the discussion. Also, uses external resources to support positions where required and/or applicable.
8 - 9 (B) - Good, quite substantial and insightful, but missing minor details which would have otherwise characterized it as an excellent response.
6 - 7 (C) - Satisfactory insight and relevance, but required some more information and effort to have warranted a better rating.
4 - 5 (D) - Limited insight and relevance of the material; more effort and reflection needed to have warranted a satisfactory grading.
0 - 3 (F) - Unsatisfactory insight/relevance or failure to answer the question, reflecting a poor or limited understanding of the subject matter and/or the guidelines of the question.

The rating scores are anonymous; therefore, do NOT mention in your remarks the separate rating score you will give the peer. The instructor is the only person who knows which score matches the comment given to a peer. Some classmates may worry that some peers will not provide a fair rating, or be unable to provide accurate corrections for grammar or other errors. It is the instructor's responsibility to ensure fairness and accuracy.

Written Assignment

For this week's written assignment, answer the following questions:

- Describe three main components of Hadoop, and summarize at least three main principles behind each of these three components
- Lastly, discuss the importance of each of the components mentioned for processing big data.

You will be assessed based on:

- Description of the three main components of HADOOP and summarization of the main principle behind each of these three components.
- Discussion on the importance of each of the components mentioned for processing big data.
- Organization and style (including APA formatting)

Submit a paper that is at least 2 pages in length exclusive of the reference page, double-spaced using 12-point Times New Roman font. The paper must cite a minimum of two sources in APA format and be well-written. Check all content for grammar, spelling and be sure that you have properly cited all resources (in APA format) used. Refer to the UoPeople APA Tutorials in the LRC for help with APA citations.

Learning Journal

Reflect on the learning from this week around Apache Spark and discuss the following:

- Distinguish the differences between the Apache Spark and Apache Hadoop frameworks. You need to write at least 4 points each.
- Which framework do you feel provides faster processing and analysis features of big data? Justify your response with supporting references.

The Learning Journal entry should be a minimum of 500 words and not more than 750 words. Use APA citations and references if you use ideas from the readings or other sources

The rubric detailing how you will be graded for this assignment can be found within the unit's assignment on the main course page.

Self-Quiz

The Self-Quiz gives you an opportunity to self-assess your knowledge of what you have learned so far.

The results of the Self-Quiz do not count towards your final grade. However,the quiz is an important part of the University's learning process and it is expected that you will take it to ensure understanding of the materials presented. Reviewing and analyzing your results will help you perform better on future Graded Quizzes and the Final Exam.

Please access the Self-Quiz on the main course homepage; it is listed inside the Unit.

Checklist

- Peer assess Unit 1 Written Assignment
- Read the Learning Guide and Reading Assignments
- Participate in the Discussion Assignment (post, comment, and rate in the Discussion Forum)
- Complete and submit the Written Assignment
- Make entries to the Learning Journal
- Take the Self-Quiz