Learning Guide Unit 4

Site: University of the People
Course: CS 3440-01 Big Data - AY2025-T1
Book: Learning Guide Unit 4

Printed by: Mejbaul Mubin
Date: Thursday, 5 September 2024, 2:30 PM

Description

Learning Guide Unit 4

Table of contents

Overview

UNIT 4: Querying Big Data

Topics

- Big data query- definition and techniques
- Apache Spark and Hive SQL Querying Process
- Challenges and advantages of querying big data

Learning Objectives

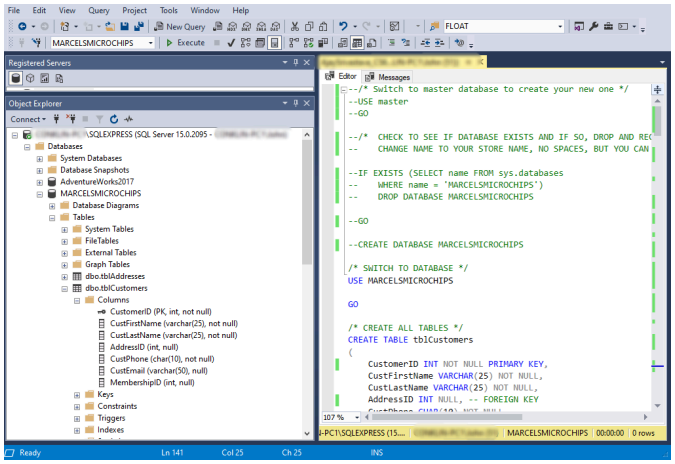
By the end of this Unit, you will be able to:

1. Implement query techniques for reporting on big data.
2. Compare Apache Hive and Hive SQL for querying big data.
3. Discuss the challenges of querying big data.

Tasks

- 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



Note: This is a screenshot of a SQL Server Management Studio query window.

Big data query is an architecture developed by Google to help developers and organizations manage their big data enterprises. The Big data query platform offers its users many different tools to aid in the management, querying, and processing of big data. The platform also provides users with spreadsheet interface capabilities, real-time analytics, and real-time data change capture capabilities. Users can also run Spark and MapReduce against this platform which supports standard SQL structure for querying.

The ability to query big data is an extremely important area for organizations that capture big data, and there are different techniques that can be used to help accomplish this. One approach that has begun to become widely used is data visualization. Data visualization gives users to ability to perform a series of analysis tasks that are usually not always possible with more common data analysis techniques. Apache Hive SQL is one product available to organizations for use in querying big data, and like any technology, it has its advantages and disadvantages. Apache Hive querying is completed with Apache Spark. Spark SQL is a fast cluster computing framework designed for fast computations. It provides users with the ability to access various data sources. It makes the process possible to weave SQL queries with code transformations, resulting in it being a very powerful tool and aiding in the querying of big data (Apache.com, 2022).

As part of an effort to speed up the processing of big data analytical queries, both Hadoop and MapReduce, discussed in an earlier unit, offer a processing engine that is well-tested and best suited for handling large datasets by utilizing a batch processing model. This batch-processing model provides a significant advantage for processing big data and has been improving the data analysis space since its inception.

Reference

Apache.com (2022). Built-in SQL Commands. Retrieved August 25, 2022 from https://spark.apache.org/docs/latest/api/sql/index.html

Reading Assignment

Read through the following to better understand about big data query, the process of using Apache Hive and Spark to query big data databases and some of the different data techniques used in data analysis. You will also learn some of the challenges around querying big data and how the industry has responded to the challenge.

Quick start. (2022). Apache spark,3.3.0. https://spark.apache.org/docs/latest/quick-start.html

- This is the Apache home site for the Apache Spark framework and provides the users with a quick start on how to Spark for querying big data.

Damji, J.S., Wenig, B., Das,T., & Lee, D. (2020). *Learning spark* (2nd Ed). O'Reilly Media, Inc. https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/ch04.html

- Chapter 4 provides the reader with an introduction to some of the built-in data sources available in Spark SQL. It goes into SQL with Spark and explains what data frames are how they are used in this querying.

Dayananda, S. (2022, March 28). *Spark SQL tutorial – Understanding spark SQL with examples.* Edureka. https://www.edureka.co/blog/spark-sql-tutorial/

- This article provides the reader with a simple tutorial on Spark SQL and provides examples and outlines the apache Spark querying process

Gurusamy, V., Kannan, S., & Nandhini, K. (2017). The real-time big data processing framework: Advantages and limitations. *International Journal of Computer Sciences and Engineering*, *5*(12), 305-312.

https://www.researchgate.net/profile/Vairaprakash-Gurusamy/publication/322550872_The_Real_Time_Big_Data_Processing_Framework_Advantages_and_Limitations/links/5a5f4658458515c03ee1d07c/The-Real-Time-Big-Data-Processing-Framework-Advantages-and-Limitations.pdf?_sg%5B0%5D=uwVJhtrrr3M7NIWrhvls-UstxE9TJpysCbcZmRheshs4C7wYaf2wXw60_nBVeVXN4pimX1iEGMemSq2AnFnU2A.lhaUyDPsKIHsCSjF-je_WWD83to8fD-njoCBH-kRVpHA5r1iVi2HxQKrqHsTW8DUZXlxPAysqzM_MPkw0yepUQ&_sg%5B1%5D=02APsySDXJ2zV5Py7JWaQ3Mknh4SgUZDwNfLZx2rnRROkOIs4K5ZjZQJtTG02oW0je_WWD83to8fD-njoCBH-kRVpHA5r1iVi2HxQKrqHsTW8DUZXlxPAysqzM_MPkw0yepUQ&_iepl=

- This paper discusses some of the advantages and limitations of the big data processing framework used in querying big data.

Małysiak-Mrozek, B., Wieszok, J., Pedrycz, W., Ding, W., & Mrozek, D. (2022). High-efficient fuzzy querying with HiveQL for big data warehousing. IEEE transactions on fuzzy Systems, *30*(6), 1823-1837. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9388934

- This article covers a method for querying big data warehouses with HiveQL

Video Resources:

BigDataELearning. (2017, December 17). *What is apache hive? : Understanding hive* [Video]. YouTube. https://www.youtube.com/watch?v=cMziv1iYt28

- This video explains the use of Apache Hive and how it interacts with Hadoop for data querying.

Simplilearn. (2017, November 2). *Big Data Tools and Technologies | Big Data Tools Tutorial | Big Data Training | Simplilearn* [Video]. YouTube. https://youtu.be/Pyo4RWtxsQM?t=103

- Watch the video from 1:43 to 2:38. This section of the video discusses big data challenges, which affect the way big data is queried.

Computerphile. (2018, December 12). *Apache spark- computerphile* [Video]. YouTube. https://www.youtube.com/watch?v=tDVPcqGpEnM

- This video discusses the use of Apache Spark in querying big data and the advantages of MapReduce.

Discussion Assignment

- Discuss three challenges organizations face when querying big data. Provide your justification for the three challenges you selected.

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 other sources.

Use APA citations and references for the textbook and any other sources; 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 your classmates' postings according to the Rating Guidelines. Review the Discussion Forum rating guidelines to see how your classmates will rate 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, it 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:

- Identify the three different querying techniques used to query big data that can benefit organizations.
- Discuss how organizations are implementing the three techniques you mentioned for querying big data.

You will be assessed based on:

- Identification of the three different querying techniques used to query big data that can benefit organizations.
- Discussion on how organizations are implementing the three techniques you mentioned for querying 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, and spelling and be sure you have correctly cited all resources (in APA format) used. Refer to the UoPeople APA Tutorials in the LRC for help with APA citations.

Learning Journal

Compare processing data using Apache Spark and Hive SQL. Provide a detailed explanation of how these systems are used with big data processing.

The Learning Journal entry should be a minimum of 400 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

- 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