Web server log processing using Hadoop in Azure

Agenda:

- Data processing is a crucial step in understanding any data. As a part of this video series, we understand

●Various features and techniques available in Hadoop to store data in a distributed manner on HDFS.

●Hive to create a projection of data stored in HDFS.

●Flume to ingest data from external systems to HDFS.

●Using Spark and Scala to process the log and transform data in gaining insights on NASA log Dataset.

Two months' worth of HTTP requests to the NASA Kennedy Space Center WWW server in Florida are contained in these two traces. The first log was kept for 31 days, from July 1, 1995, to July 31, 1995. The second log was kept for seven days, from August 1, 1995, to August 31, 1995. There were 3,461,612 requests in this two-week period.

Aim:

In this project, you will use Hadoop, Flume, Spark and Hive to process the Web Server logs dataset to get more insights on the log data. As part of this, you will create Azure Virtual Machine and install Hadoop, Flume, Spark, Scala and Hive to perform Flume agent execution, Build Scala code, submit Spark jobs and Hive Queries using the dataset.

Data Format:

The logs are an ASCII file with one line per request, with the following columns:

1.host making the request. A hostname when possible. Otherwise the Internet address.
2.Timestamp in the format "DAY MON DD HH:MM:SS YYYY", where DAY is the day of the week, MON is the name of the month, DD is the day of the month, HH:MM:SS is the time of day using a 24-hour clock, and YYYY is the year. The timezone is -0400.

3.Request given in quotes.

4.HTTP reply code.

5.Bytes in the reply.

Tech Stack:

➔ Language: Scala

➔ Services: Azure VM, Hadoop, Hive, Flume, Spark

Scala

Scala is a multi-paradigm, general-purpose, high-level programming language. It's an object-oriented programming language that also supports functional programming. Scala applications can be converted to bytecodes and run on the Java Virtual Machine (JVM). Scala is a scalable programming language, and Javascript runtimes are also available.

Hive

Apache Hive is a fault-tolerant distributed data warehouse that allows for massive-scale analytics. Using SQL, Hive allows users to read, write, and manage petabytes of data. Hive is built on top of Apache Hadoop, an open-source platform for storing and processing large amounts of data. As a result, Hive is inextricably linked to Hadoop and is designed to process petabytes of data quickly. Hive is distinguished by its ability to query large datasets with a SQL-like interface utilizing Apache Tez or MapReduce.

Flume

Flume is a service for rapidly gathering, aggregating, and transporting massive amounts of log data that is distributed, reliable, and available. Its architecture is simple and adaptable, based on streaming data flows. It has configurable reliability techniques as well as several failover and recovery mechanisms, making it resilient and fault tolerant. It employs a straightforward extensible data model that enables online analytic applications.

Approach

1. Sign-in to the Microsoft Azure account

2. Create a virtual machine

    - Select the tab to create a new VM
    - Add the basic configuration details to create a VM instance.

3. Connect to the Virtual machine

- Download and install the putty application
- Add the configuration details.

4. Install hadoop, hive, flume, scala, spark

5. Execute the Scripts in code.Zip step by step.

Project Takeaways

- Understanding the problem statement
- Overview of Microsoft Azure
- Creating a Virtual machine (VM) instance in Microsoft Azure.
- Connect to the Virtual machine using Putty application
- Install various big data technologies on Virtual machine
- What are log files and different types of log file
- How to process log files and importance of processing them
- What are a referrer and user agent
- What are the contents of a log file and uses of a log file
- Why flume and how does flume work, flume agent and its role
- Processing and ingestion of log data using Flume
- Process the log data using Spark
- Performing analysis on log data using Hive

Step 1 – Setting up Virtual Machine on Azure

We created resource groups, furthermore, creating Virtual Machine using PUTTY.

Step 2 – Installation of Hadoop, Hive and Flume.

We set up the environment in order to kickstart the ETL process.

Step 3 – We used Flume after understanding how it works and how does it flume agent work after understanding what its role is.

Step 4 – We Processed and ingested log data using Flume.

Step 5 – We performed analysis on log data using Hive.

Step 6 – We processed the log data using Spark.