Naan Mudhalvan

Big Data Analytics

Module 1 Homework

Linux and Introduction to Big Data Hadoop

1. Open the file /var/log/messages in the vi editor and delete line number 150.

$vi /var/log/messages This is used to open a message file.

: set number This is used to set a specific line of line inside the file. :150 d This is a representation of a particular line to be deleted

:wq This command is used to exit & save from a file.

2. Write a shell script to add two numbers?

Script :

!/bin/bash

Calculate the sum of two integers with pre initialize values in a shell script a=10 b=20

sum=$(($a + $b))

echo "Sum is: $sum"

Answer: *Sum is:30*

3. User root wants to copy /etc, including all subdirectories and files to /tmp. How will you achieve this task?

Using the below command can we achieve copy /etc, including all subdirectories and files to /tmp.

sudo cd /etc

cp -r /etc tmp

4. Create a file that contains only the username and the user id of all the users present on the server.

$cat /etc/passwd

By this Unix command, we can create a file that contains only the username and user id of all the users present on the server.

5. How will you provide a count of all users on the system except for adm user?

$ who Using the above command we can count all users on the system except for adm users.

6. How will you list all files in /tmp in increasing order of their size?

"*$ ls -laShr*" Using this command we can list the files in increasing order.

To list all files and sort them by size, use the -S.

By default, it displays output in descending order

-l flag means long listing and -a tells ls to list all files including (.) or hidden files. human-readable format by adding the -h

And to sort in reverse order, add the -r flag.

7. What command is used to clear history on the Linux server?

To clear history

If you want to delete a particular command, enter history

-d <line number> .

To clear the entire contents of the history file, execute history *-c .*

8. Explain “Big Data” and what are five V’s of Big Data?

BigData :

The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.

Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can’t manage them. But these massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before.

5V’s

- **Volume:** the size and amounts of big data that companies manage and analyze
- **Value:** the most important “V” from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits
- **Variety:** the diversity and range of different data types, including unstructured data, semi-structured data and raw data
- **Velocity:** the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time
- **Veracity:** the “truth” or accuracy of data and information assets, which often determines executive-level confidence

9. What is Hadoop and its components?

Hadoop can be defined as a collection of Software Utilities that operate over a network of computers with Software Frameworks on a distributed storage environment in order to process the Big Data applications in the Hadoop cluster.

The Core Components of Hadoop are as follows:

- MapReduce
- HDFS
- YARN