Big Data Analytics

Linux and introduction to Hadoop

Homework -1

Q1. 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 specific line of the line inside the file.

:150 d This is representation of particular line to be delete

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

Q2. Write a shell script to add two numbers?

!/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

Q3. 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

Q4. 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 can we create a file that contains only the username and the user id of all the user present on the server.

Q5. 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 user.

Q6. 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 (biggest to smallest in size).

-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.

Q7. What command is used to clear history on the Linux server?

Removing 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 .

Q8. Explain "Big Data" and what are five V's of Big Data?

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.

five characteristics: volume, value, variety, velocity, and veracity.

Q9. What is Hadoop and its components?

Hadoop is a framework that uses distributed storage and parallel processing to store and manage big data. It is the software most used by data analysts to handle big data, and its market size continues to grow. There are three components of Hadoop:

Hadoop HDFS - Hadoop Distributed File System (HDFS) is the storage unit.

Hadoop MapReduce - Hadoop MapReduce is the processing unit.

Hadoop YARN - Yet Another Resource Negotiator (YARN) is a resource management unit.