**CP2285 Big Data Programming with Java**

**Adam Barry (20154153)**

**January 18th, 2023**

1) What is Big Data? Give a brief definition.

Big Data is generally a set of large data that displays characteristics of volume, velocity, and variety to extent the makes data unsuitable for management by a database system. Big data comes from a variety of sources, including web logs, social media, mobile devices, and more. It can be used to identify trends, customer preferences, and other patterns that can help businesses/companies make better decisions.

2) What are the traditional 3 Vs of Big Data? Briefly define each

The three traditional Vs of Big Data are the following….

1. Volume – The quantity of data stored.
2. Velocity – The speed that data is entering a system and the speed the data is getting processed.
3. Variety – The variations in the structure of data being stored in a system.

3) Explain why companies like Google and Amazon were among the first to address the big data problem?

Companies like Google and Amazon were among the first to address the big data problem because a combination of text graphics video and I resources combine into complex structures created challenges for data management that involves all three characteristics. Web Startup companies failed as a result to this but companies like Google and Amazon experienced significant growth and were among the first feel the pressure imagine big data. Popular web platforms like Facebook quickly followed and these companies became pioneers or create new technologies to address big data problems. Google create a Big Table data store Amazon created Dynamo add Facebook created Cassandra technologies which would help with the big data problem that arise.

4) Explain the difference between scaling up and scaling out.

The difference between scaling up and scaling down is scaling up is keeping the same number of systems the same but the process for dealing with large amounts of data growth that involves migrating the same structure to a bigger and more powerful system whereas scaling out means the workload exceeded the capacity of the server known as clustering. This is the process for dealing with data growth that involves the distributing of data storage structures across a cluster of commodity servers.

5) What is stream processing, and why is it sometimes necessary?

Stream Processing is the process of data inputs in order to make decisions about which data to keep in which data to discard before storage. This is sometimes necessary because stream processing focuses on the input processing. In some situations, large volumes of data can enter the system at a rapid pace that is not feasible to store all the data at once. The data must be processed and filtered as it enters the system determine which they had to keep and which data to discard.

6) How is stream processing different from feedback loop processing?

Stream processing is different from feedback loop processing because Stream processing focuses more on the input processing, and it requires analysis of the data stream as it enters a system. Whereas Feedback loop processing refers to the analysis of the data to produce actionable results. While string processing could be thought of a focused-on inputs feedback loop processing can be thought of a focus on outputs.

Choose 6 HDFS commands

1) hadoop fs -ls (Used to list files and directories. The contents of your current working directory)

![Graphical user interface

Description automatically generated]()

2) Hadoop fs -cat (Used to print the content of a file onto the standard output stream)

![Text

Description automatically generated with medium confidence]()

3) hadoop fs -mkdir (allows users to create or make new directories)

![Graphical user interface

Description automatically generated]()

4) hadoop fs -rmdir (allows users to remove directories)

![Graphical user interface, text

Description automatically generated]()

5) hadoop fs -help (A command that is used to provide for information about a given command)

![Text

Description automatically generated]()

6) Hadoop fs -rm (A command to remove a give file)

![Text

Description automatically generated]()