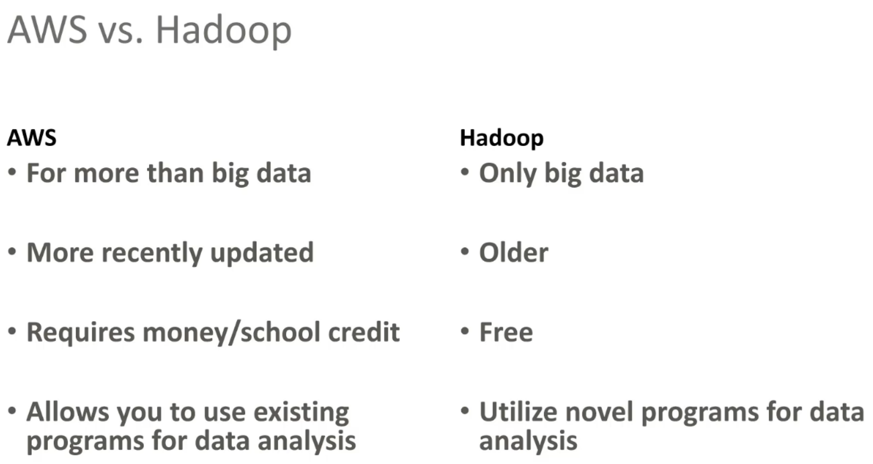
This exam serves as the assessment for those students who cannot utilize the Hadoop system and/or Ambari GUI.

1. Compare and contrast the Hadoop and AWS big data experiences.



AWS vs. Hadoop

AWS	Hadoop
• For more than big data	• Only big data
• More recently updated	• Older
• Requires money/school credit	• Free
• Allows you to use existing programs for data analysis	• Utilize novel programs for data analysis

Overall, AWS is used for more things and Hadoop is free and niche-like.

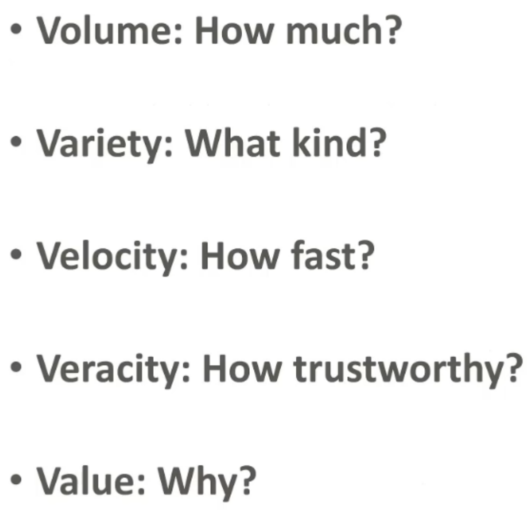
2. Which Hadoop tool do you think is most useful, and why?

We use several Hadoop tools in this program like SQL in Hive, HDFS, and Ambari.

Ambari is a web-based tool that supports many of the other tools and components so that's why I think it is most useful.

[What is Hadoop Good For? (Best Uses, Alternatives, & Tools) - HostingAdvice.com]()

3. What are the Vs of big data?



- **Volume: How much?**
- **Variety: What kind?**
- **Velocity: How fast?**
- **Veracity: How trustworthy?**
- **Value: Why?**

4. Why is Hadoop important to big data? What functions does it provide?

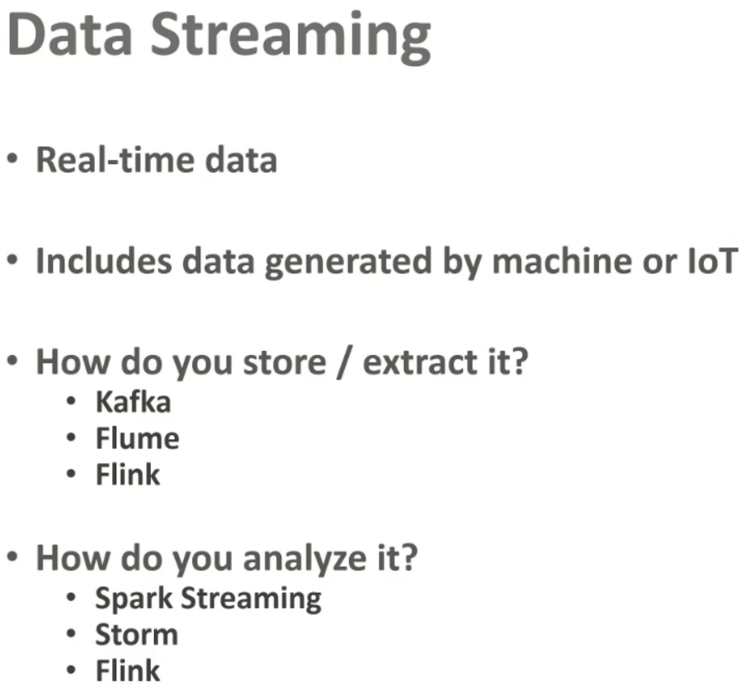
Hadoop is a framework to store and process big data. Hadoop is specifically designed to provide distributed storage and parallel data processing that big data requires.

Big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful business information.

Why Hadoop used for Big Data Analytics ? - Interview Questions (wikitechy.com)

Hadoop is important to big data as the data is huge and Hadoop can store and process data.

5. Explain the concept and overall process of data streaming.

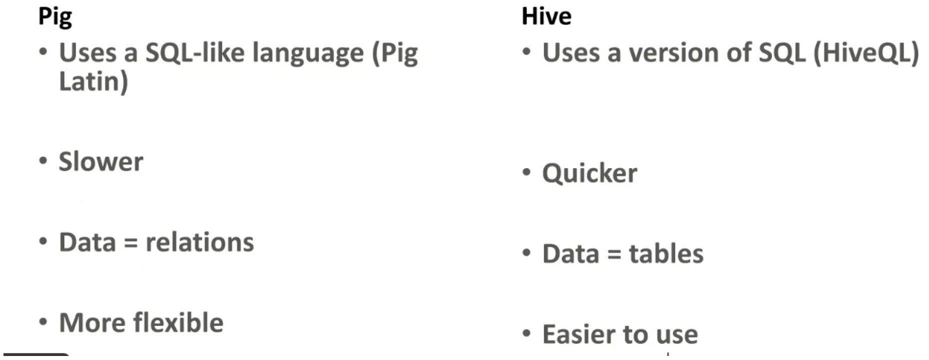


Data Streaming

- Real-time data
- Includes data generated by machine or IoT
- How do you store / extract it?
  - Kafka
  - Flume
  - Flink
- How do you analyze it?
  - Spark Streaming
  - Storm
  - Flink

Data streaming is real-time data and it could use micro-batches like in Spark Streaming. Kafka is the first program mentioned in our textbook for storing but it doesn't analyze it. Flink is the most versatile program of these mentioned.

6. Compare and contrast Pig and Hive.



Pig	Hive
• Uses a SQL-like language (Pig Latin)	• Uses a version of SQL (HiveQL)
• Slower	• Quicker
• Data = relations	• Data = tables
• More flexible	• Easier to use

Hive is quicker, uses tables, and SQL than Pig which is slower, uses relations, and uses SQL-like language.

7. Why is MapReduce important?

This is important because it is the original big data system. Since MapReduce is a two-step process, the work will be split into two parts. Map is when we find the data that we need and transform it. Reduce is when we simply the data.

8. What area of big data would you like to learn more about and why?

Big data is making an impact on the financial, IT. and marketing industries. 5 Key Areas Where Big Data Is Making a Major Impact (techopedia.com)

Big data is helping businesses keep customers, lower costs, provide insights, know more about competitors, risk management, innovation of products, and supply chain.

7 Areas of Your Business That Can Be Streamlined By Big Data | Socialnomics

In big data, data analysis can make a difference in the business. I would like to continue learning how I can help them.