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

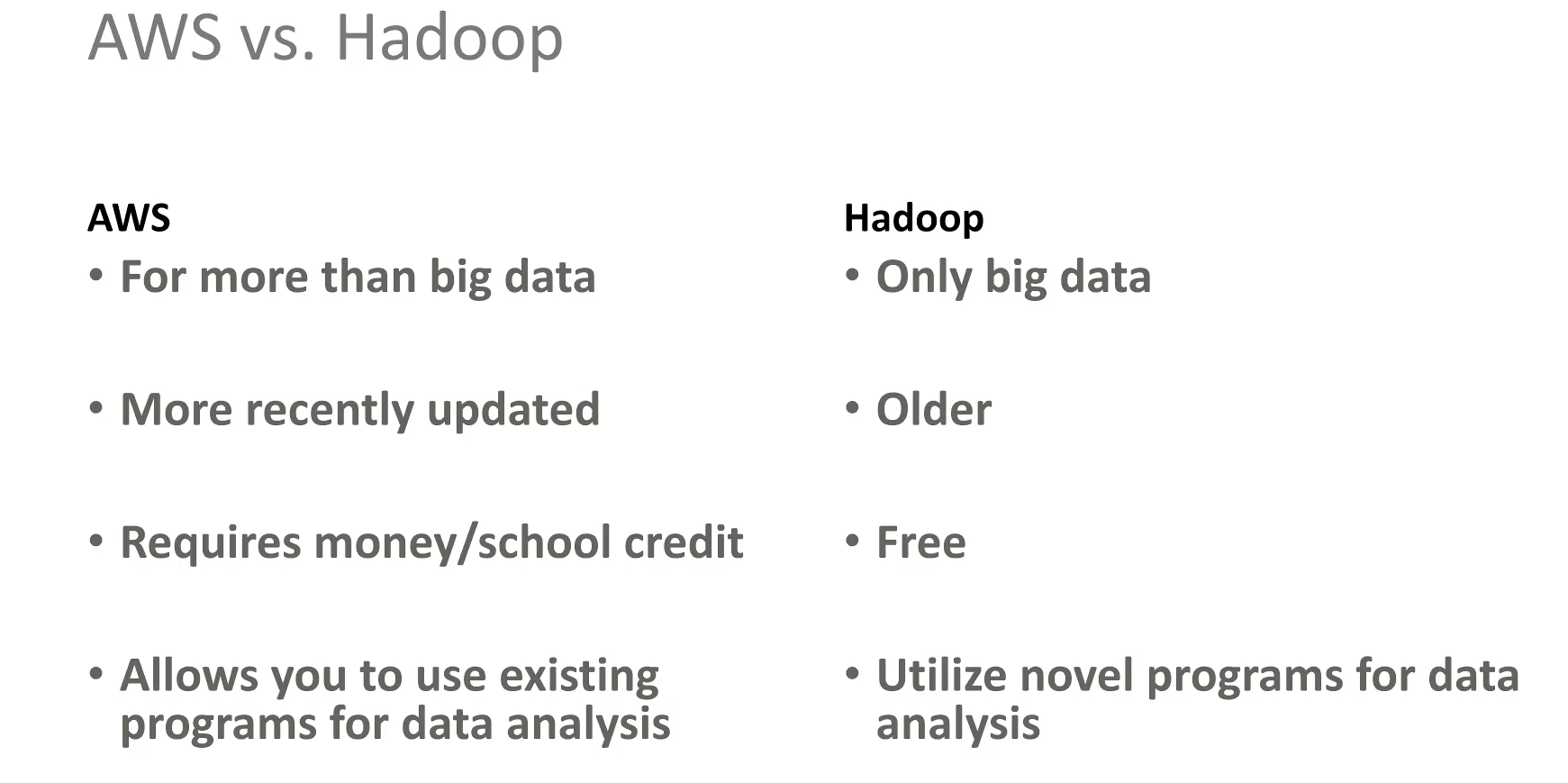
1. **Compare and contrast the Hadoop and AWS big data experiences.**

AWS stands for Amazon Web Services. Essentially, it is Amazon’s cloud platform that allows for processing big data and [data analytics](https://blog.amitechsolutions.com/how-to-win-big-with-healthcare-data-analytics). AWS can offer businesses across industries a platform to host their data warehousing systems.

Some benefits of AWS EMR include:

* Auto-scaling clusters
* Relatively easy to use
* On-demand processing power

As opposed to AWS , which is a cloud platform, Hadoop is a data storage and analytics program developed by Apache. You can think of it this way: if AWS is an entire car, then Hadoop is akin to the engine.



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

Although all tools have their own specific pros and cons, Spark looks the most promising because :

Apache Spark is a unified analytics engine for processing big data and for machine learning applications. It is the biggest open-source data processing project and has seen a very wide-spread adoption.

While Hadoop is a great tool to process large data, it relies on disk storage making it slow. This makes interactive data analysis a difficult task. Spark, on the other hand, processes in-memory making it many, times faster.

Spark’s RDD (Resilient Distributed Dataset) data structure makes it possible to distribute data across the memory of many machines. Spark also supports several tools including Spark SQL, MLib (for machine learning), and GraphX (for graph processing).

1. **What are the Vs of big data?**

 5 Vs, these are Volume, Velocity, Variety, Variety, and Veracity.

* Volume – It refers to an enormous size of data and this size determines the volume of the data. Volume determines whether data is big or not.
* Velocity – It refers to the speed of generation of data. It shows how fast data is generated and processed for analysis.
* Variety – It means the heterogeneous nature of the data in question. Today, the data is of different types like photos, videos, emails, audio, etc.
* Variability – This is the inconsistency of data that can affect how we manage or process it effectively.
* Veracity – It refers to the trustworthiness and messiness of data. Due to the different forms of large data, it becomes important to control accuracy and quality of the data.

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

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

Hadoop makes it easier to use all the storage and processing capacity in cluster servers, and to execute distributed processes against huge amounts of data. Hadoop provides the building blocks on which other services and applications can be built.

Why is Hadoop important?

* Ability to store and process huge amounts of any kind of data, quickly. With data volumes and varieties constantly increasing, especially from social media and the Internet of Things (IoT), that's a key consideration.
* Computing power. Hadoop's distributed computing model processes big data fast. The more computing nodes you use, the more processing power you have.
* Fault tolerance. Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.
* Flexibility. Unlike traditional relational databases, you don’t have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.
* Low cost. The open-source framework is free and uses commodity hardware to store large quantities of data.
* Scalability. You can easily grow your system to handle more data simply by adding nodes. Little administration is required.

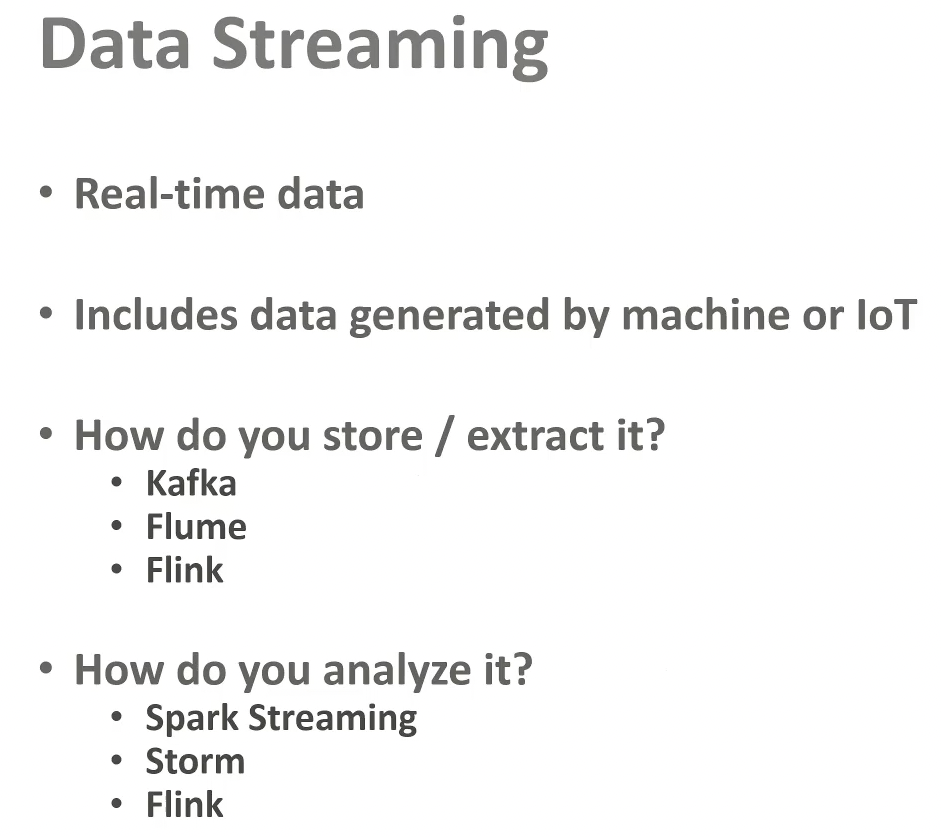
## Two major functions of Hadoop

Firstly providing a distributed file system to big data sets.

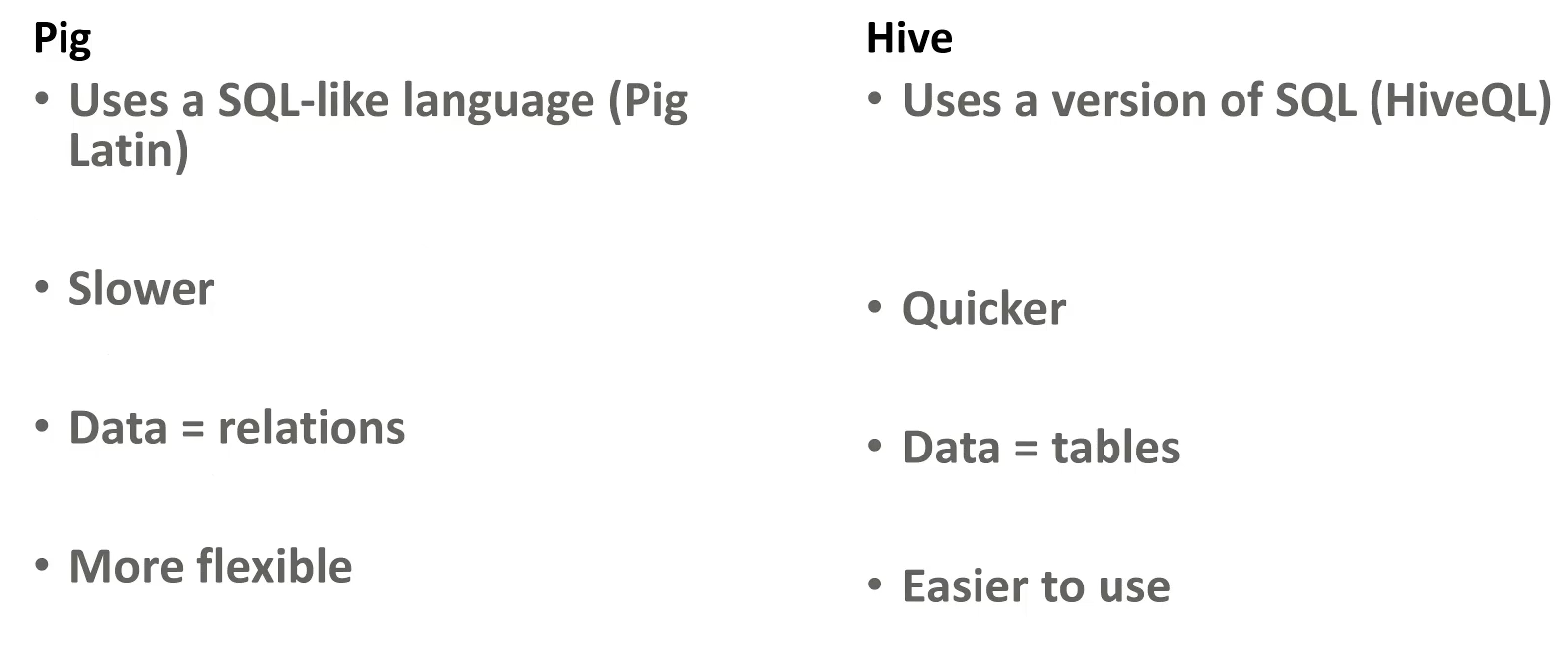
Secondly, transforming the data set into useful information using the MapReduce programming model. Big data sets are generally in size of hundreds of gigabytes of data. For such a huge data set, it provides a distributed file system (HDFS). This allows to store them in clusters of different commodity machines and then access them parallelly.

The HDFS replicates the data sets on all the commodity machines making the process more reliable and robust. If there is a failure on one node, Hadoop can detect it and can restart the task on other healthy nodes. The framework is also highly scalable and can be easily configured anytime according to the growing needs of the user. Setting up the Hadoop framework on a machine doesn’t require any major hardware change. The machine just needs to meet some basic minimum hardware requirements such as RAM, disk space and operating system.

1. **Explain the concept and overall process of data streaming.**



1. **Compare and contrast Pig and Hive.**



1. **Why is MapReduce important?**

The MapReduce framework is scalable, flexible, cost-effective, and fast processing system.

It offers security, fault-tolerance, and authentication. MapReduce is a simple model of programming and offers parallel programming.

Features of MapReduce

1. Scalability

Apache Hadoop is a highly scalable framework. This is because of its ability to store and distribute huge data across plenty of servers. All these servers were inexpensive and can operate in parallel. We can easily scale the storage and computation power by adding servers to the cluster.

Hadoop MapReduce programming enables organizations to run applications from large sets of nodes which could involve the use of thousands of terabytes of data.

Hadoop MapReduce programming enables business organizations to run applications from large sets of nodes. This can use thousands of terabytes of data.

#### 2. Flexibility

MapReduce programming enables companies to access new sources of data. It enables companies to operate on different types of data. It allows enterprises to access structured as well as unstructured data, and derive significant value by gaining insights from the multiple sources of data.

Additionally, the MapReduce framework also provides support for the multiple languages and data from sources ranging from email, social media, to clickstream.

The MapReduce processes data in simple key-value pairs thus supports data type including meta-data, images, and large files. Hence, MapReduce is flexible to deal with data rather than traditional DBMS.

#### 3. Security and Authentication

The MapReduce programming model uses HBase and HDFS security platform that allows access only to the authenticated users to operate on the data. Thus, it protects unauthorized access to system data and enhances system security.

#### 4. Cost-effective solution

Hadoop’s scalable architecture with the MapReduce programming framework allows the storage and processing of large data sets in a very affordable manner.

#### 5. Fast

Hadoop uses a distributed storage method called as a Hadoop Distributed File System that basically implements a mapping system for locating data in a cluster.

The tools that are used for data processing, such as MapReduce programming, are generally located on the very same servers that allow for the faster processing of data.

So, Even if we are dealing with large volumes of unstructured data, Hadoop MapReduce just takes minutes to process terabytes of data. It can process petabytes of data in just an hour.

#### 6. Simple model of programming

Amongst the various features of Hadoop MapReduce, one of the most important features is that it is based on a simple programming model. Basically, this allows programmers to develop the MapReduce programs which can handle tasks easily and efficiently.

The MapReduce programs can be written in Java, which is not very hard to pick up and is also used widely. So, anyone can easily learn and write MapReduce programs and meet their data processing needs.

#### 7. Parallel Programming

One of the major aspects of the working of MapReduce programming is its parallel processing. It divides the tasks in a manner that allows their execution in parallel.  
The parallel processing allows multiple processors to execute these divided tasks. So the entire program is run in less time.

#### 8. Availability and resilient nature

Whenever the data is sent to an individual node, the same set of data is forwarded to some other nodes in a cluster. So, if any particular node suffers from a failure, then there are always other copies present on other nodes that can still be accessed whenever needed. This assures high availability of data.

One of the major features offered by Apache Hadoop is its fault tolerance. The Hadoop MapReduce framework has the ability to quickly recognizing faults that occur.

It then applies a quick and automatic recovery solution. This feature makes it a game-changer in the world of big data processing.

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

I would like to learn more about Machine Learning, MapReduce, Hadoop and also Data Visualization

**Data Visualisation –**It is an upcoming big data field where business context for data is in question. Data visualization allows the stakeholders and non-analysts to understand the data model and take decisions according to it.