Introducing Hadoop

**Question 1 of 6**

Which statement is true about Hadoop?

* 

It adds a wide column store capability.

* 

It provides a fast NoSQL database.

* 

It replaces your existing RDBMS.

* 

**It will commonly be used as an alternative file system.**

Which type of example data would be best suited for Hadoop?

* incorrect

records of ACH banking transfers

* 

flight schedules for a major airline

* 

**customer purchasing data for sales analytics**

**customer purchasing data for sales analytics**

**Correct**

**Non-critical data about a customer's interactions with a business is an excellent fit for Hadoop. HDFS supports storing unstructured data in a manner that is scalable and accessible.**

Relational databases lack the scalability and speed to accommodate most big data projects.

Correct

While relational databases such as MongoDB and SQL variants will always have their uses, requirements for big data projects have generally grown beyond the ability of relational databases to deliver.

* 

databases of patient prescriptions

records of ACH banking transfers

Incorrect

Structured data, such as bank transaction reports, is best suited for conventional relational databases, partially because of the integrity and durability of relational databases.

Structured data, such as flight schedules, is better suited to standard relational databases, partly due to the consistency and reliability of relational databases.

Which type of data is a good fit for Hadoop?

* 

money-related activities

* 

medical diagnostics

* 

transactions

* 

**behavioral**

In the context of CAP, consistency supports \_\_\_\_\_ while partitioning supports \_\_\_\_\_.

* 

up-time; availability

* 

**transactions; scalability**

* 

availability; transactions

* 

complexity; flexibility

Which aspect of the CAP theory does Hadoop gain the most benefits in comparison to traditional relational databases?

* 

availability

**Scalability**

**scalability**

**Correct**

**The infinite scalability of Hadoop allows storage of any potential size to be partitioned any kind of hardware, even commodity servers**.

**Which aspect of the CAP theory does Hadoop gain the most benefits in comparison to traditional relational databases?**

availability

scalability

flexibility

consistency

flexibility

consistency

Which statement about the use of relational databases for big data projects is true?

Relational databases lack the scalability and speed to accommodate most big data projects

**Relational databases lack the scalability and speed to accommodate most big data projects.**

**Correct**

**While relational databases such as MongoDB and SQL variants will always have their uses, requirements for big data projects have generally grown beyond the ability of relational databases to deliver.** .

Relational databases have no use anymore.

Relational databases have superior scalability and speed for big data applications.

Relational databases support sophisticated processing techniques for machine learning tasks but lack scalability for big data projects.

Which statement is true regarding the Hadoop market?

Hadoop growth is relatively slow.

Hadoop is a relatively narrow ecosystem.

**Hadoop is considered a top job trend.**

Hadoop implementation only requires developer skills.

What is HBase used for

tuning the underlying Hadoop file layout****

**making Hadoop data accessible to less technical teams**

natively accessing the Hadoop file system

writing the MapReduce processing logic

As a beginner, which distribution should you consider for quickly setting up a core Hadoop instance?

Apache Hadoop's core package

a Hortonworks platform

a Cloudera platform

**a PaaS solution, such as Google Dataproc**

Which Hadoop component functions as the processing API?**MapReduce MapReduce**

**Correct**

**MapReduce is a processing method and a distributed computing application architecture written in java.** 

HDFS

Hive

Hbase

Which Hadoop distribution is open-source?

Hortonworks

**Apache Hadoop Apache Hadoop**

**Correct**

**Apache Hadoop is a series of open-source software utilities that make it easy to solve problems utilizing massive amounts of data and computing using a network of many servers**

MapR

Cloudera

Which Hadoop function does the MapReduce component handle?

library consolidation

**processing API**

data storage

cloud distribution

Which level of abstraction would be most accessible for analysts?

**Hbase Hbase**

**Correct**

**HBase is an open-source non-relational distributed database that is similar to a wide column store, which should be familiar to SQL veterans.**

Java

MapReduce

HDFS

In which type of Hadoop distribution does the vendor provide and manage the image as a service categorized as?

Infrastructure as a Service (IaaS)

commercial

open source

**Platform as a Service (PaaS) Platform as a Service (PaaS)**

**Correct**

**Software as a service (PaaS) or application platform as a service (aPaaS) is a cloud-based service delivering a platform that allows customers to develop, run and manage apps without the complexities of creating or managing the typical app-development infrastructure.**

Unlike the pseudo-distributed mode, the standalone mode \_\_\_\_\_.

cannot be used in production

replicates each file three times

**uses the regular file system**

cannot be used for testing

If a Hadoop implementation uses multiple JVM daemons spread across several nodes, uses the HDFS filesystem, and stores every file in triplicate, which Hadoop mode is the implementation running?

pseudo-distributed

single node

cloud managed

**fully distributed fully distributed**

**Correct**

**The code running on a real Hadoop cluster is conducted in this mode. It is the mode in which you see Hadoop's true power when you run your code on thousands of servers and against a massive input.**

Because Hadoop processes run in JVMs, they \_\_\_\_\_.

are identical between all Hadoop versions

**do not share state**

only work in Hadoop version 2.0

share a state like other databases

You would like to quickly deploy a Hadoop testing cluster that uses real data, and that you will scale to production later. Which option is likely best for this task?

**using a GCP Dataproc cluster**

creating your own cluster using Docker

building your own cluster from scratch

using Cloudera Hue Community Edition

When setting up a Hadoop cluster, which configuration is simplest to scale and can use a vendor-provided image for minimal setup with technical support?

**virtual machines on public cloud When setting up a Hadoop cluster, which configuration is simplest to scale and can use a vendor-provided image for minimal setup with technical support?**

**virtual machines on public cloud**

docker containers on bare metal

virtual machines on bare metal

docker containers on cloud

Which Cloudera product allows you to try some Hadoop functions in a browser?

Cloudera Express

Cloudera Enterprise

**Cloudera Live**

Cloudera Director

Which is true about Apache Spark?

It supports HBase but not HDFS. It support HDFS, Hbase Hive any Hadoop InputFormat

**It supports both Java and Python**.

It uses the disk for processing. uses in-memory processin

It is only available as a commercial solution it is an open source library.

What kind of task is Apache Spark well suited to?

**distributed processing of very large dataset distributed processing of very large datasets**

**Correct**

**Apache Spark is a distributed, open-source data workload processing system that uses in-memory caching and streamlined query execution to easily query any size data and is simply a fast, general engine for the processing of large-scale data.**



version management of Hadoop modules

consistent querying of transactional databases

long term storage of large databases

Which component in the Hadoop ecosystem is used for predictive analytics?

**Mahout** for Machine learning/predictive analysis

Sqoop

Flume

YARN

What language library is used to query Hbase?

mahout

pig

**Apache Pig is a large data set analysis framework composed of a high-level language for data analysis programs to be represented, combined with an architecture for analyzing these programs**.

oozie

**HQL Hive offers a method for projecting a structure onto the Hadoop data and for querying the data using the HiveQL (HQL) language using HBase.**

What is the most important difference between Java Virtual Machines (JVMs) and traditional database processing systems?

Different Hadoop processes can run in the same JVM.

**JVMs do not share state**

**What is the most important difference between Java Virtual Machines (JVMs) and traditional database processing systems?**

JVMs cannot perform batch processing.

JVM processes are exactly the same between Hadoop 1.0 and 2.0.

What has to be installed so you can develop MapReduce jobs in Java?

**a JDK**

a Sublime text editor

an IDE

the R bridge

Which **Hadoop library adds SQL-like queries that run as interactive processes**?

**Impalainteractive process use memory in the machine**

Storm

Hive

Pig

Which console command checks the version of Hadoop included in the Cloudera VM?

version 

version --hadoop 

hadoop release 

**hadoop version**

What is the first consideration when setting up your development environment for Hadoop?

which vendor distribution to work with

which data storage model to use

which libraries to use as part of your installation

**whether to work with a local install using open source binaries or work with a vendor distribution Correct**

**Before anything else, you must decide if you have the ability to manage a cluster manually, or whether utilizing a vendor will be more helpful. This will also depend on available hardware.**

Assuming you have a master node with 4 vCPUs, what is the maximum number of worker nodes with 2 vCPUs each that can be configured to not exceed the number of CPUs allowed per region?

three

one

**two**

zero

**When running a Hadoop cluster, which part of the cluster performs the main job of processing and storing data?**

Master nodes Incorrect

Master nodes in distributed Hadoop clusters host the different storage and processing management resources listed in this list for the entire Hadoop cluster.



Google Cloud Platform Incorrect

Google Cloud Platform is a digital infrastructure provider for the deployment and operation of web applications. Its focus is to provide a platform for individuals and companies to create and run apps, and to use the web to communicate to the users of the said apps. 

**Worker nodes In a Hadoop cluster, the worker nodes comprise most of the virtual machines and do the job of storing the data and performing computations.**

HDFS

Which IDE often used with the Java programming language can be installed with Hadoop plugins?

Sublime

**Eclipse**

**Correct**

**Apache Hadoop-Eclipse provides useful Apache Hadoop integrations from the Eclipse IDE**.

RStudio

Vim

Which core library will give you the ability to have a SQL-like query for batch MapReduce jobs?

Spark

Pig

**Hive Correct**

**Apache Hive is a project focused on the Apache Hadoop data warehouse platform for querying and analyzing data. Hive offers an SQL-like interface for querying data contained in a variety of databases and file systems that interact with Hadoop.**



Impala

When installing the Cloudera developer distribution, which component are you responsible for installing

Cloudera tools

the OS inside the VM

**virtualization software**

Cloudera samples

What is the most important consideration when downloading a Hadoop virtual machine?

**downloading the VM that matches your virtualization software**

**Correct**

**If the image we download isn't in the correct format for our virtualization software, we will not be able to run it.**

OS of the virtual machine

your download speed

download size of the VM

What is the disadvantage of setting up Hadoop as a local install?

It is expensive.

**It takes a long time**.

It requires installing virtualization software.

It requires using a full HDFS.

What is the main advantage of using the Google Cloud Platform as opposed to local clusters?

HDFS availability

**direct job monitoring direct job monitoring**

**Correct**

**Using Google Cloud Platform, you are provided a variety of job monitor and visualization tools.**

enhanced processing speed Incorrect

As with any cloud service, your processing speed likely depends on your configuration.

accessibility through SQL-like DBMs

When will you elect to use the streaming method for developing a MapReduce job?

when you want to send the job results directly to the console

when you want to connect your logic to abstraction libraries

when you need to develop the mapping and reducing classes using C++

**when you want to create the mapping or reducing logic in a language different from Java**

Which Hadoop file system shell command input from the command line will run a MapReduce() job from a JAR file?

hadoop jar jobname.jar /home/output /home/input

**hadoop jar jobname.jar /home/input /home/output Correct**

**This hadoop file system shell command input from the command line will run a MapReduce() job from a specified JAR file.**

run jar jobname.jar /home/input /home/output

hadoop jobname.jar /home/input /home/output

Which MapReduce phase is based on code that you have to write?

**Mapping**

Shuffling

Sorting

Splitting

When setting up a MapReduce job on GCP, where would you specify the output destination for the results?

the Cluster configuration

the Main class field

**the Arguments field**

the Jar files field

What must be considered before running a new JAR file in Google Cloud Platform?

**whether the Hadoop version matches Correct**

**If the image we download isn't in the correct format for our virtualization software, we will not be able to run it.**

your OS

your download speed

file size of the JAR

What should you plan for when running Hadoop jobs?

* 

**Jobs should be monitored to confirm successful completion**.

Job errors are typically fixed by the Hadoop developer.

Verbose logging should always be left at the default setting.

Newer job outputs will replace old outputs on the HDFS.

Which Hadoop file system shell command input from the Linux terminal can be used to read the contents of a file?

hadoop fs pwd /home/file.txt

**hadoop fs cat /home/file.txt Which Hadoop file system shell command input from the Linux terminal can be used to read the contents of a file?**

cat /home/file.txt

hadoop fs read /home/file.txt

When implementing the MapReduce functionality, which important aspect of the Map function should you consider?

The same Map function instance executes on all nodes.

**The Map function is implemented as a static class**.

The Map function on one node can share state with other nodes.

The Map function aggregates key-value pairs from some nodes.

Which function of MapReduce 1.0 outputs a single combined list of key-value pairs?

Comparison()

Partition()

**Reduce()Correct The role of The Reducer is to process data that comes from the mapper. After processing, it generates a new set of output, which will be stored in the HDFS.**

Map()

Why does the functional programming aspect of Hadoop cause the requirement for task trackers on each node to have to run in separate JVMs?

Hadoop does not support object-oriented programming.

**JVMs do not share states. Correct**

**Unlike traditional relational database management systems, JVMs do not share state.**

Hadoop nodes have a maximum per-node memory limitation.

It is false that task trackers must run on separate nodes for each JVM.

What is it called when MapReduce() kills a job based on parameters you have set?

Pig Latin

configurable parameters

Pipes

**speculative execution Correct**

**Speculative execution is an optimization strategy where a computer system executes certain tasks that may not be required.**

What is the main form of MapReduce() output, regardless of format?

long integers

32-bit floats

**key and value pairs Correct**

**In MapReduce, a certain key-value pair is processed by the map function which outputs a certain number of key-value pairs, while the Reduce function processes values grouped by the same key and outputs another set of key-value pairs.**

64-bit floats

Which command creates the directory /user/hadoop/mydir in the Hadoop file system?

hadoop -mkdir /hadoop/mydir 

hadoop fs -FromHadoop -mkdir /mydir 

**hadoop fs -mkdir /user/hadoop/mydir** 

mkdir -hadoop /user/hadoop/mydir