**1.Before you create a Jupyter notebook in Watson Studio, which two items are necessary? (Two answers)**

1. Project
2. file
3. scala
4. url
5. spark instance

**2.What is the architecture of Watson Studio centered on?**

1. Collaborator
2. Projects
3. Analytic assets
4. data assets

**3.Which Watson Studio offering used to be available through something known as IBM Bluemix?**

1. Watson Studio Cloud
2. // // Local
3. // // Desktop
4. // // Business

**4.You need to add a collaborator to your project. What do you need?**

1. The email of the collaborator
2. The list of deployments
3. list of your saved bookmarks
4. your project ID

**5.Where does the unstructured data of a project reside in Watson Studio?**

1. Wrapper
2. Tables
3. Database
4. Object Storage

**6.What does the user interface for Jupyter look like to a user?**

1. Common desktop app
2. Database Interface
3. App in web browser
4. Linux SSH session

**7.What is a “magic” command used for in Jupyter?**

1. Running common statistical analyses.
2. Autoconfig
3. extending the core language with shortcuts
4. Parsing and loading data into a notebook.

**9.Why might a data scientist need a particular kind of GPU (graphics processing unit)?**

1. To display a simple bar chart of data on the screen
2. to collect video for use in streaming data applications
3. to perform certain data transformations quickly
4. to input commands to a data science notebook

**10.What is a markdown cell used for in a data science notebook?**

1. Documenting the computational process
2. writing code to transform data
3. holding the output of a computation
4. configuring data connections

**11.Which is an advantage that Zeppelin holds over Jupyter?**

1. Users must auth
2. Notebooks can be used by multiple ppl at the same time
3. notebooks can be connected to big data engines such as spark
4. Zeppelin is able to use the R language.

**12.Which statement about Apache Spark is true?**

1. it supports HDFS, MS-SQL, and Oracle
2. it is much faster than MapReduce for complex applications on disk
3. It runs on hadoop clusters with ram drives configured on each DataNode
4. It features API’S for C++ and .NET

**13.Which two are valid watches for ZNodes in ZooKeeper?**

1. NodeExpired.
2. NodeDeleted.
3. NodeRefreshed
4. NodeChildrenChanged

**14.Which component of an Hadoop system is the primary cause of poor performance?**

1. disk latency
2. network
3. CPU
4. Ram

**15.What is an example of a NoSQL datastore of the “Document Store” type?**

1. Cassandra(column stores)
2. REDIS(key value)
3. HBase(column store)
4. MongoDB(document store)

**16.What is the preferred replacement for Flume?**

1. Hortonworks Data Flow
2. Storm
3. NiFi
4. Druid

**17.Which statement describes an example of an application using streaming data?**

1. An application evaluating sensor data in real-time
2. one time export and import of a db
3. a web application that supports 10,000 users
4. a system that stores many records in a database

**18.Which Spark Core function provides the main element of Spark API?**

1. MLlib
2. RDD
3. Mesos
4. YARN

**19.Which statement is true about the combiner phase of the MapReduce architecture?**

1. it aggregates all input data before it goes through the Map Phase
2. it reduces the amount of data that is sent to the Reducer task nodes
3. It determines the size and distribution of data split in the map phase
4. It is performed after the reducer phase to produce the ...

**20.Apache Spark provides a single, unifying platform for which three of the following types of operations? (Three answers)**

1. transaction processing
2. ACID transaction
3. machine learning
4. record locking
5. batch processing
6. graph operations

**21.Which is the java class prefix for the MapReduce v1 APIs?**

1. org.apache.hadoop.mapred(v1)
2. org.apache.mr
3. org.apache.hadoop.mr
4. org.apache.mapreduce(v2)

**22.Which two factors in a Hadoop cluster increase performance most significantly?**

1. solid state disks
2. immediate failover of failed disk
3. parallel reading of large data
4. data redundancy on management nodes
5. high-speed networking between nodes
6. large number of small data

**23.Which statement accurately describes how ZooKeeper works?**

1. There can be more than one leader server at a time.
2. .all servers keep a copy of shared data memory
3. clients connect to multiple servers at the same time
4. Writes to a leader server will always succeed

**24.Hadoop 2 consists of which three open-source sub-projects maintained by the Apache Software Foundation?**

1. HDFS
2. Hive
3. BigSQL
4. MapReduce
5. YARN
6. CloudBreak

**25.Hadoop uses which two Google technologies as its foundation?**

1. Google File System
2. HBase
3. YARN
4. Ambari
5. MapReduce

## **26.Which component of the Apache Ambari architecture provides statistical data to the dashboard about the performance of a Hadoop cluster?**

1. Ambari Server
2. Ambari metric system

**27.What are two security features Apache Ranger provides?**

Authorization, auditing

**28.Under the YARN/MRv2 framework, which daemon arbitrates the execution of tasks among all the applications in the system?**

ResourceManager

**29.Which data encoding format supports exact storage of all data in binary representations such as VARBINARY columns?**

1. Sequence files
2. Parquet
3. RCFile
4. Flat

**30.Which feature makes Apache Spark much easier to use than MapReduce?**

1. Suitable for transaction processing.
2. Applications run in-memory
3. Libraries that support SQL queries
4. APIs for Scala, Python, C++, and .NET

**31.Which Hadoop ecosystem tool can import data into a Hadoop cluster from a DB2, MySQL, or other databases?**

1. Oozie
2. Sqoop
3. Accumulo
4. HBase

**32.Which two are attributes of streaming data?**

1. Simple, numeric data.
2. Data is processed in batch.
3. Requires extremely rapid processing.
4. Sent in high volume.

**33.Which type of cell can be used to document and comment on a process in a Jupyter notebook?**

1. output
2. markdown
3. code
4. kernel

**34.Which component of the Spark Unified Stack allows developers to intermix structured database queries with Spark's programming language?**

1. MLlib
2. Mesos
3. Java
4. Spark SQL

**35.What is the default directory in HDFS where tables are stored?**

1. /app/hive/warehouse/data
2. /app/hive/warehouse/bigsql
3. /app/hive/warehouse/schema
4. /app/hive/warehouse/

**36.What is an advantage of the ORC file format?**

1. Big SQL can exploit advanced features
2. Efficient compression
3. Supported by multiple I/O engines
4. Data interchange outside Hadoop

**37.What two security functions does Apache Knox provide?**

1. Management of Kerberos in the cluster.
2. Database field access auditing.
3. Proxying services.
4. API and perimeter security.

**38.Which Apache Hadoop application provides an SQL-like interface to allow abstraction of data on semi-structured data in a Hadoop datastore?**

1. YARN
2. Spark
3. Hive
4. Pig

**39.Which two commands would you use to give or remove certain privileges to/from a user?**

1. REVOKE
2. GRANT
3. INSERT
4. LOAD
5. SELECT

**40.When connecting to an external database in a federation, you need to use the correct database driver and protocol. What is this federation component called in Big SQL?**

wrapper

**41.Using the Java SQL Shell, which command will connect to a database called mybigdata?**

./jsqsh mybigdata

**42.Which Big SQL feature allows users to join a Hadoop data set to data in external databases?**

A. Integration

B. Impersonation

C. Grant/Revoke privileges

D. Fluid query

**43.Which definition best describes RCAC?**

A. It limits the rows or columns returned based on certain criteria.

B. It grants or revokes certain user privileges.

C. It grants or revokes certain directory privileges.

D. It limits access by using views and stored procedures.

**44.You need to enable impersonation. Which two properties in the bigsql-conf.xml file need to be marked true?**

bigsql.alltables.io.doAs

bigsql.impersonation.create.table.grant.public

**45.Which command creates a user-defined schema function?**

CREATE FUNCTION

**46.You are creating a new table and need to format it with parquet. Which partial SQL statement would create the table in parquet format?**

STORED AS parquetfile

**47.Which tool would you use to create a connection to your Big SQL database?**

DSM

**48.Which directory permissions need to be set to allow all users to create their own schema?**

777

**49.How many Big SQL management nodes do you need at minimum?**

A. 1

B. 4

C. 3

D. 2

**50.You need to determine the permission setting for a new schema directory. Which tool would you use?**

umask

**51.What are Big SQL database tables organized into?**

Schemas

**52.What is the first step in a data science pipeline?**

acquisition

**53.What command is used to list the "magic" commands in Jupyter?**

%lsmagic

**54.Under the MapReduce v1 programming model, which optional phase is executed simultaneously with the Shuffle phase?**

combiner

**55.Which hardware feature on an Hadoop datanode is recommended for cost efficient performance?**

JBOD

**56.Which statement is true about Hortonworks Data Platform (HDP)?**

It is a Hadoop distribution based on a centralized architecture with YARN at its core.

**57.What is the name of the Hadoop-related Apache project that utilizes an in-memory architecture to run applications faster than MapReduce?**

Spark

**58.What are two ways the command-line parameters for a Sqoop invocation can be simplified?**

Include the --options-file command line argument.

Place the commands in a file.

**59.If a Hadoop node goes down, which Ambari component will notify the Administrator?**

Ambari Alert Framework

**60.Which three programming languages are directly supported by Apache Spark? (Three answers)**

Scala

Python

Java

**61.Which Apache Hadoop application provides a high-level programming language for data transformation on unstructured data?**

Pig

**62.Which component of an Hadoop system is the primary cause of poor performance?**

disk latency

**63.Under the YARN/MRv2 framework, the Scheduler and ApplicationsManager are components of which daemon?**

ResourceManager

**65.What is the final agent in a Flume chain named?**

Collector

**66..Which component of the Hortonworks Data Platform (HDP) is the architectural center of Hadoop and provides resource management and a central platform for Hadoop applications?**

YARN

**68.What are two primary limitations of MapReduce v1?**

Scalability

Resource utilization

## **69.Which description characterizes a function provided by Apache Ambari?**

## A wizard for installing Hadoop services on host servers.

## **70.How can a Sqoop invocation be constrained to only run one mapper?**

## Use the -m 1 parameter.

## **71.What does the split-by parameter tell Sqoop?**

## The column to use as the primary key

## **72.Which NoSQL datastore type began as an implementation of Google's BigTable that can store any type of data and scale to many petabytes?**

## HBase

## **73.Under the MapReduce v1 programming model, which shows the proper order of the full set of MapReduce phases?**

## Map -> Combine -> Shuffle -> Reduce

## **74.Which computing technology provides Hadoop's high performance?**

## Parallel Processing

## **75.Which statement is true about MapReduce v1 APIs?**

## MapReduce v1 APIs are implemented by applications which are largely independent of the execution environment.

## **76.Under the YARN/MRv2 framework, the JobTracker functions are split into which two daemons?**

## ResourceManager

## ApplicationMaster

## **77.Which component of the Apache Ambari architecture integrates with an organization's LDAP or Active Directory service?**

## Authorization Provider

## **78.Which Apache Hadoop component can potentially replace an RDBMS as a large Hadoop datastore and is particularly good for "sparse data"?**

## HBase

## **79.What is the name of the Hadoop-related Apache project that utilizes an in-memory architecture to run applications faster than MapReduce?**

## Spark

## 

## **80.What are three IBM value-add components to the Hortonworks Data Platform (HDP)?**

## Big Replicate

## Big Match

## Big SQL

## **81.Which three are a part of the Five Pillars of Security?**

## Audit

## Data Protection

## Administration

## **82.What are two services provided by ZooKeeper?**

## Providing distributed synchronization.

## Maintaining configuration information.

**83.Under the YARN/MRv2 framework, which daemon is tasked with negotiating with the NodeManager(s) to execute and monitor tasks?**

1. TaskManager
2. ResourceManager
3. JobMaster
4. ApplicationMaster

## 

## **84.Which statement best describes a Big SQL database table?**

## A directory with zero or more data files.

## 

## **85.You have a distributed file system (DFS) and need to set permissions on the the /hive/warehouse directory to allow access to ONLY the bigsql user. Which command would you run?**

A. hdfs dfs -chmod 755 /hive/warehouse

B. hdfs dfs -chmod 666 /hive/warehouse

C. hdfs dfs -chmod 770 /hive/warehouse

D. hdfs dfs -chmod 700 /hive/warehouse

**86.What is an example of a Key-value type of NoSQL datastore?**

REDIS

## 

## **87.Under the MapReduce v1 programming model, what happens in a "Reduce" step?**

Data is aggregated by worker nodes.

**88.Apache Spark can run on which two of the following cluster managers? (Two answers)**

Hadoop YARN

Apache Mesos

**89.What are two common issues in distributed systems? (Two answers)**

Finding a particular node within the cluster.

Partial failure of the nodes during execution.

**90.Apache Spark can run on which two of the following cluster managers?**

A. Apache Mesos

B. oneSIS

C. Nomad

D. Linux Cluster Manager

E. Hadoop YARN