Which of the following options better describe the Hadoop framework

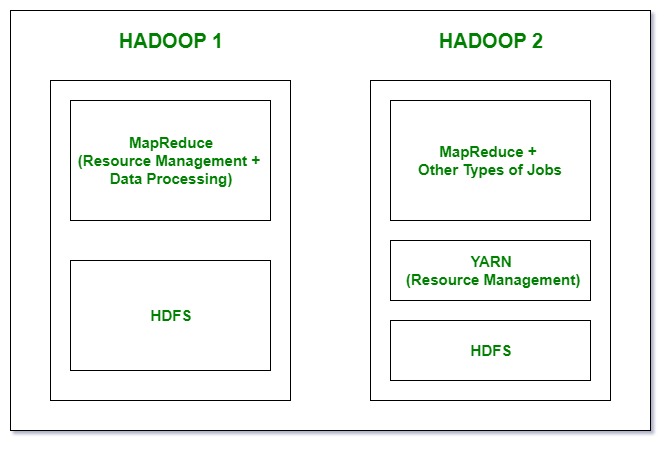
Ans ))

Default block size in Hadoop

Ans))128MB

What is the improvements in2’x Hadoop

Ans))



Which of the following is component of Hadoop

Ans))) the above(Yarn ,hdfs,mapreduce)

Which of the following file is used to specify the scheduling file

Ans)))23, a scheduling policy can be used to specify the predicates and priorities process. ... For example, you can set a scheduling policy by running kube-scheduler --policy-config-file <filename> or kube-scheduler --policy-configmap <ConfigMap> .

Types of tables in hive

Ans))External and internal tables

Which of the following function is used to read data in PIG?

PigStorage is the default load function.

Each version of data within a cell adds a versioning information through a

Ans)))) value

Deepest level in spark execution hierarchy

Ans))))

Application (instance of spark context can be used for single batch job).

Which of the following is not indicated for data streaming

Ans)))Data analytics

\_\_\_\_\_\_\_\_\_\_ is the subset of the replicas list that is currently alive and caught up to the leader. Ans))))“isr” is the set of “in-sync” replicas.

What is Apache kafka

Ans))))

Kafka is primarily used to build real-time streaming data pipelines and applications that adapt to the data streams. It combines messaging, storage, and stream processing to allow storage and analysis of both historical and real-time data.

What is true about Kafka?

Kafka is a distributed streaming platform that is used publish and subscribe to streams of records. Kafka gets used for fault tolerant storage. Kafka replicates topic log partitions to multiple servers. Kafka is designed to allow your apps to process records as they occur.

Which of the following technology drive big data

Ans))) Apache Hadoop

10key technologies which drive the big data and enable the Big Data analytics for business

Ans))) predictive, no SQL database, kdd, stream analytics, in-memory data, data visualisation, distributed storage, data integration, data quality, data preprocessing

Big challenges of big data

Ans))))

* Lack of knowledge Professionals. To run these modern technologies and large Data tools, companies need skilled data professionals. ...
* Lack of proper understanding of Massive Data. ...
* Data Growth Issues. ...
* Confusion while Big Data Tool selection. ...
* Integrating Data from a Spread of Sources. ...
* Securing Data.

How do hdfs client create a file

Ans))))

create() on DistributedFileSystem(DFS).

How data or file is returned into hdfs

Ans))))

Map reduce frameworKs

Ans)))))

Which of the following is not a characteristic of data lake

1. Data is not searchable easily (B)data is in a single centralised repository (c)data is not governed by audit and compliance for resources(d) data is easily accessible using query and visualisation tools

Ans))))

When data is saved in Azure data Lake how many copies of data are saved by default

Ans)))

3

What will be the output of the below code

Import pandas

Df=pandas.DataFrame(33,2, 55, 66)

Print(type (df))

A …class pandas.com dot frame dot data frame

B…. Class dataframe

C….. class pandas

D…. Class list

Ans))))

b

What is the return type of fetchall() function

Ans))))

fetchall() fetches all the rows of a query result. It returns all the rows as a list of tuples. An empty list is returned if there is no record to fetch.

Which of the following is not a valid numpy function

1…numpy.Max

2…numpy.min

3…numpy.sort

4….numpy.pop

Anss))))

Numpy.pop

How can you change Thomas into Michael in the last name column in the users table

Ans))))

Which of the following statement correct to display all the records from the weather table from which humidity is in the range of 60 to 75

A…. Select star from weather where humidity in 62 to 75

B….. select star from weather where humidity between 60 and 75

C….. select star from weather humidity not in 60 and 65

D…. Select star from weather weather humidity and not between 60 and

Ans))))

Which of the following SQL statement can be used for adding a row in a table

Ans))))

You have a very well run data management program in your organisation which is very secure you use analytics for making big data driven decisions in recent months you realise that whatever decisions you are making based on the analytics are proving to be wrong and harmful for the organisation after consulting with data analyst and data stewards you conclude that this is due to poor data quality what should be the next step of action

A…. Install install a firewall

B…change the analytical strategies

C…. Minimise Data loss by changing data ownership

D…. Find the breach in the data

Ans)))

Other than gaining real time insights off data what is another major advantage of streaming analytics

A .. enhancement of

B…real-time predictive modelling

C…effective analysis

D…real time dashboards

Ans))))

Identify essential features of strategic information

Elastic

preserves

Data integrity time

variant highly scalable

(Multiple choices)

Ans))))

Select the component used to combine multiple database records to a single records in talend studio

Ans))))

Select the component used to perform lookup data in talend studio

Ans))))

What is an essential factor to ensure good data quality

A …focus on enterprise-wide data quality improvement

B…. Fix data quality issues

C…. Aggregate poor quality data

D…Manually correct bad data

Ans))))

Reducing the number of fields in the output is an example of what type of partitioning

Ans))))

Where does the library of job components reside in talend open studio UI

Pallet ….preferences ….job tools …business models

Ans))))

Which statement is true about unstructured data

A…. Common examples of unstructured data include audio video text

B…structured data is information that does have a predefined values

C…. Unstructured information is typically not text-heavy and almost audio and video files

D…. Unstructured data is information that is well organised

Ans))))

Specify some of the outcomes of a data warehouse realisation

Database management

Transactional management

Intuitive dashboards

Predictive analytics reports

It's a multiple choice

Ans))))

Data in ….bytes size is called big data

Ans))))

Which correctly describes the difference between static and dynamic data

A…static data is also known as historical data and dynamic data is known as latest

B…. Static data is rarely used for analytical purpose but dynamic is used for

C…. Static data can be both transactional and non transactional

D…. Static data does not require updating whereas dynamic require constant

Ans))))

d

Complex data types in hive

Ans)))

arrays, maps, and structs

Single point of failure in Hadoop is which version

Ans))) Hadoop 1.0

What are the Hadoop features

Ans)))

1. Open Source: Hadoop is open-source, which means it is free to use. ...
2. Highly Scalable Cluster: Hadoop is a highly scalable model. ...
3. Fault Tolerance is Available: ...
4. High Availability is Provided: ...
5. Cost-Effective: ...
6. Hadoop Provide Flexibility: ...
7. Easy to Use: ...
8. Hadoop uses Data Locality:

Hive row format is delimited with

Ans))))

comma (“,”)

Namenode only responds to

Ans)))

 RPC requests issued by DataNodes or client.

What are the three v's of big data

Ans))))

 Volume, Velocity, and Variety

Which tool is used to collect from the multiple sources

Ans)))

Which is the command to check the status of demons in Hadoop

Ans)))

jps

What is the default replication factor to achieve fault tolerance

Ans)))

3

Which of the following process schedules the task across the nodes in Hadoop system

Ans)))

Hadoop FIFO scheduler

Which file contains configuration of resource manager

Ans)))

xml: This file contains the configuration settings related to YARN . For example, it contains settings for Node Manager, Resource Manager, Containers, and Application Master.

The default block size in hdfs with 2.6 release is

Ans))))

128

From the below steps to choose right sequences of steps required when an hdfs client wants to create a file on the data note

1. Hdfs client write the packet on data node
2. Hdfs client gets the block location from the namenode
3. Data node will send acknowledgement back to hdfs client
4. Hdfs client will send a create file request to the namenode

choose the most appropriate sequence from the below options

Ans)))

Mapper and reducer implementations can use the ……… to report programs or just indicate that they are alive choose the appropriate option

Path over

Output collector

Reporter

None of the above

Ans))))

Which file contains configuration for resource manager

Ans)))

xml: This file contains the configuration settings related to YARN . For example, it contains settings for Node Manager, Resource Manager, Containers, and Application Master.

Which is not an advantage of views

1. They reduce query complexity
2. Can be used to reach actions
3. They slow logo creation of thread
4. They can be updated manually

Ans)))

Which tool can be used for analysing unstructured data without writing the Hadoop Java code

Ans)))

hive

Embedded metal stores in the Hive DW make use of which built-in database

Ans))

In hbase a table can be

1. dropped directly
2. Drop after
3. Only disabled not dropped
4. only compressed not dropped

Ans)))

Replication factor for a topic in kafka guarantees

Which of the following code blocks reads in paraquet file “/home/vagrant/imports.paraquet” as a dataframe

Ans)))

Your organisation want to develop onestream application which gives the latest status vaccination dose completed in your organisation since covid-19 vaccination started which method wilfull fill recruitment

Ans)))

Why pointers used in streaming application select all that apply

A)))

The only metadata retained on a per consumer basis is the position of the consumer in the log called

Ans))))

offset

What is true about yarn

Ans))) one of the Apache Hadoop score components and is responsible for allocating system resources to various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster

Hadoop framework is written in which language

Ans)))

JAVA

Hdfs clients write the data packet in

Ans))))

a client needs to interact with master i.e. namenode (master).

Which of the following is not true about pig

Ans)))) pig cannot perform all the data manipulation operations in Hadoop

Pig uses grunt shell

Ans))))

Yes, Grunt shell is a shell command. The Grunts shell of Apace pig is mainly used to write pig Latin scripts. Pig script can be executed with grunt shell which is native shell provided by Apache pig to execute pig queries. We can invoke shell commands using sh and fs.

Alter table command is used for

Ans)) alter table is used to add delete or drop or modify columns in the existing table and also used to drop or add various constraints on existing table

What is Spark streaming checkpoint

Ans))) a process of writing received records a checkpoint intervals to hdfs

What is the method used while developing windows-based streaming application

Ans)))

Which storage does Hive use

hive does not provide any storage, it just uses a database (derby as default, MySQL would be a good choice if you want to use a different db) as a metastore just to store the metadata related to the tables, partitions, views, buckets etc..