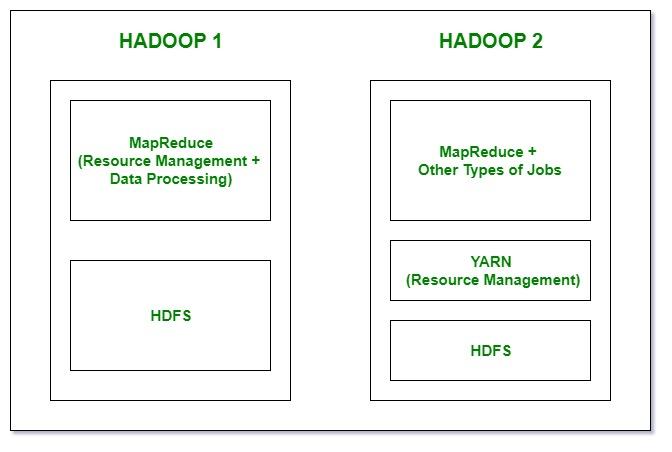
Default block size in Hadoop

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))))When a client wants to write a file to HDFS. Client communicates to namenode for metadata. The Namenode responds with details of datanodes where client will start writing the data. Then, on basis of information from NameNode, client directly writes data on the datanode.
* Map reduce frameworKs
* Ans)))))The MapReduce framework is the core of Apache Hadoop. This programming paradigm provides for massive scalability across hundreds or thousands of servers in a Hadoop cluster. ... MapReduce applications can process large data sets in parallel by using a large number of computers, known as clusters.
* 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
* 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))))
* 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)))
* Open Source: Hadoop is open-source, which means it is free to use. ...
* Highly Scalable Cluster: Hadoop is a highly scalable model. ...
* Fault Tolerance is Available: ...
* High Availability is Provided: ...
* Cost-Effective: ...
* Hadoop Provide Flexibility: ...
* Easy to Use: ...
* 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
* Hdfs client write the packet on data node
* Hdfs client gets the block location from the namenode
* Data node will send acknowledgement back to hdfs client
* 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
  + They reduce query complexity
  + Can be used to reach actions
  + They slow logo creation of thread
  + 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
  + dropped directly
  + Drop after
  + Only disabled not dropped
  + 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..
  + ….. is the orchestration system
  + Ans)))kubernets
  + Which is not true about pigs
  + Ans))) it runs MR jobs under the hood
  + What is the command to check Hadoop version
  + Ans)) Hadoop version
  + Optimisation can be performed by using MR jobs by using a combiner to work on the mapper output before it is sent to the reducer and by increasing number of reducers
  + True
  + What is the default replication factor for hdfs
  + Ans))))3
  + Which files are used for storing metadata in namenode
  + Ans)))both a and b
  + Which of the following phases occur simultaneously in mapreduce framework choose the most appropriate option
  + Ans))) Shuffle and sort
  + Which file contains configuration for resource manager
  + Yarn-site.xml
  + Which of the following process schedules the task across nodes in Hadoop system
  + Ans))Resource manager
  + Which data type can be used to implement a bounded variable length character string
* Ans))) String
* What does the Hive metastore do
* Ans))) none
* Pig operates in many how many nodes
* Ans))2
* All mapreduce jobs reading from nhb stable accept their (k1,V1) pair in the form of
* Rowkey:scan result
* Assume that sparkcontext difference is available as SC which of the following statement will help us create a spark SQL context SSC successful
* Ssc=SQLContext(sc)
* Which attribute is not indicated for data streaming
* Ans))limited data
* Which coordination service is used in kafka cluster
* Zookeeper
* Which of the following statement is true about Apache
* Both a and b
* .Which of the following are the daemon processes of Hadoop?
  + Ans: All of the above
* 2.In which language is Hadoop framework written?
  + Ans: Java
* 3.What is the command to check Hadoop version?
  + Ans: hadoop version
* 4.What is the command to check the status of daemons of Hadoop?
  + Ans: jps
* 5.Which of the following options better describe Hadoop Framework?
  + Ans: write-once-read-many
* 6.Single Point of Failure of Name-node was the major issue in which Hadoop version?
  + Ans: Hadoop 1.0
* 7.Which is not a scheduler option available with YARN?
  + Ans: Optimal Scheduler
* 8.What is/are true about YARN?
  + Ans: All the above
* 9.What is the default port number for YARN Web UI?
  + Ans: 8088
* 10.Which file contains configuration for resource manager?
  + Ans: yarn-site.xml
* 11.Which file(s) is/are used for storing meta data in Name node?
* Ans: both A & B ( editlogs & fsimage)
* 12. What is the default block size in hadoop 2.0 ?
  + Ans: 128 MB
* 13.What is the default replication factor to achieve fault tolerance?
  + Ans: 3
* 14. Which of the following is not a HDFS command?
  + Ans: cd
* 15. hadoop fs -put Customer.csv /dataset
* What will happen when you execute above command
* Ans: Customer.csv file copied from Local to HDFS /dataset directory
* 16. What is map reduce?
  + Ans: both Distributed Framework based on HDFS & Programming Model
* 17. If your file size is 2 GB then how many blocks created in hadoop 2.0?
  + Ans: 16
* 18. If you have used 4 partitions in your map reduce job then ideally how many reducer required?
  + Ans: 4
* 19. What will happen if P1 directory does not exist?
  + Ans: Hadoop MapReduce job will successfully executed
* 20. which directory contains mapred-site.xml file?
  + Ans: /etc/Hadoop
* 21. pig -x mapreduce is the command to run Pig grunt shell in cluster environment
  + Ans: True
* 22. Pig enables us to write MR jobs without knowing Java
  + Ans: True
* 23. Hive is schema on read
  + Ans: True
* 24. Metastore exposes the data in HDFS files as stored in tables in Hive
  + Ans: True
* 25. HIVE Provides more built-in functions than SQL
  + Ans: True
* 26. Result of Hive query can be stored on HDFS and Local FS
  + Ans: True
* 27. Hive is Data Warehouse for Hadoop system.
  + Ans:True
* 28. The command ""create database test;"" will create test database in hive
  + Ans: True
* 29. the two different tables supported by Hive are External Table & Global Table
  + Ans: False
* 30. "LOAD DATA INPATH '/dataset/emp.csv' INTO TABLE emp;
* As per above statement your input file storage is on Local FS?
  + Ans: False
* 31. the property ""set hive.cli.print.header=true"" is used to set to display the header fields for table in Hive
  + Ans: True
* 32. Sqoop is used for Data cleansing
  + Ans: False
* 33. In Sqoop operations, num-mappers argument is used to achieve Parallelism while Import
  + Ans: true
* 34. Partitioning and bucketing help to improve query performance
  + Ans: True
* 35. Partitioning column is a real column
  + Ans: False
* 36. Hive commands can only be run from CLI
  + Ans: False
* 37. Bucketing creates separate folders inside table directories.
  + Ans: False
* 38. Hive does not support user defined functions
  + Ans: False
* 39. While dynamically loading partitioned tables we specify the partition name
  + Ans: False
* 40. Hive can be used without Hadoop
  + Ans: False
* 41. The storage levels options available in scala, java and python applications are
* Memory\_only, Memory\_and\_disk, Disk\_only Memory\_and\_disk\_ser,
  + Ans: False
* 42. If you are intending to use input Dstream based on the receiver for a 3-node cluster environment, which of the following configuration definition is true
  + Ans: "val conf = new SparkConf().setMaster(""local[2]"").setAppName(""MyStreamingApp"")"
* "val conf = new SparkStreamingConf().setMaster(""local[2]"").setAppName(""MyStreamingApp"")"
* 43. Which one of the below code is correct to start kafka server?
  + Ans: bin/kafka-server-start.sh config/server.properties
* 44. Which among the following is a fault -tolerant group of elements that is worked on in parallel?
  + Ans: RDD
* 45. Which of the following statement is false?
  + Ans: Maven dependency on Spark is optional
* 46. Which of the below is not the data format spark uses:
  + Ans: distFile
* 47. An Mechanism used to safely execute the updating of variable in a split in processing environment of a cluster?
  + Ans: Accumulators
* 48. what is default retention policy for messages in kafka?
  + Ans: 7 Days
* 49. If in multi-node kafka cluster replication factor is 3 then how many leader & followers possible?
  + Ans: 1 Leader, 2 Followers
* 50. A complex and costly redistributing data mechanism to group differently across multiple partition is,
  + Ans: Shuffle
* 51. Which of the following operation does not bring the values together across all partitions to compute the final result?
  + Ans: Parallelize
* 52. Which of the following is the default data source of SparkSQL
  + Ans: Parquet
* 53. The below is a method which generates code dynamically and uses a specific format to perform various operations without deserializing the serialized data
  + Ans: Encoder
* 54. "Which of the following are true regarding Spark Streaming?
* a. There can be only one streamingContext will be active at one pint of time.
* b. When the streaming Context is stopped the spark context also stops.
* c. A single sparkContext cannot be reused for multiple streamingContext.",
  + Ans: A&B
* 55. What is/are tasks of zookeeper in kafka cluster?
  + Ans: All the above(Configuration management, Leader detection, Synchronization)
* 56. Which of the following are the component of kafka cluster:
  + Ans: a,b,c (Producer, Consumer, Broker)
* 57. Which is the default scheduler used by SparkContext?
  + Ans: DAG
* 58. As per the below code:
* val Me=spark.read.json(“/emp.csv”) which component is created"
  + Ans: Dataframe
* 59. Which of the statements is/are not correct?
  + Ans: Datasets can be created using toDF()
* 60. Which of the list are considered as Pair RDD's
  + Ans: Both A & B (reduceByKey() , groupByKey())

BIG DATA: The Big Picture

* Which V out of all V's of big data led to inspiring of trust?
* Veracity
* Choose all the best reasons that led the technology move from Hadoop to Spark.
* More Memory Oriented
* Stream Processing
* Which of the following is not a SQL/Data Lake Query Engine?
* Kafka
* Apache Paraquet stores data in \_\_\_\_\_\_\_\_\_\_\_ fashion.
* Columnar
* Which act as a checkpoint node in HDFS?
* Secondary Name Node
* Which among the following holds the location of data?
* Name Node
* Which of the following is True about Map Reduce?
* MapReduce Works with Large quantities of Datasets
* The Mapper Output <Key,Value> obtained is combined based on Key and then fed to Reducer.
* Series of Values for Each Distinct Key is given as Input to Reducer
* The keys will not be unique in Mapper.
* Which are the two Reasons(Advantages) to prefer ELT(Extract Load and Transform) over the other methods?
* Raw data can be more efficiently processed on Destination Platform than ETL platform.
* Resilient Processing
* When Cloudera and Hortonworks merged, which of the following were de-emphasized?
* Apache Orc
* Sentry
* Trino (formerly called PrestoSQL) is basically a fork of an original project. Name that original project.
* PrestoDB
* In the BI world, Google acquired \_\_\_\_\_\_\_\_ .
* Looker
* Which of the following are focused on both analytics and AI?
* Alteryx, Databricks, Cloudera
* Which of the following are Analytical Tool Approaching?
* Visuals
* Code
* Arrange the Maturity Model in sequence.
* Queries and Spreadsheet
* Business Intelligence
* Data Lakes or Warehouse
* Some combo of batch, streaming, ML
* Analysis of all data, everywhere

Getting started with HDFS

* If the schema of the table does not match with the data types present in the file containing the table then hive
* Reports Null values for mismatched data
* The default load function in pig is
* PigStorage
* Resources in hdfs 1.0 are managed by
* map reduce
* What does ls-u do
* it uses access time for displaying and sorting files
* Which command is used to format HDFS
* bin/hadoop namenode -format
* Hadoop runs the jobs by dividing them into
* Tasks
* Identify the actions that can be performed in HDFS**.**
* Create
* Delete
* append
* Which HDFS command is used to create a file of zero length?
* Touchz
* Which among these is correct sqoop command to import data from MySQL to HDFS?
* sqoop import --connect jdbc:mysql://localhost/test --table tableName -m 1
* Sqoop command to display tables in MySQL database?
* sqoop list-tables --connect jdbc:mysql//localhost/test1
* What is the correct syntax to fetch specific number of rows from Hbase table
* scan 'table\_name', {'LIMIT' => 10}
* How to insert values in a table in Hbase?
* put 'tableName', 'rowNumber', 'columnName', 'value'
* Unix Scripting language that processes POSIX commands?
* Bash
* Benefits of Bash Scripting in HDFS?
* Runs on cluster
* Easy to learn
* MapReduce Jobs

Hive

* Match the following hadoop components with the most appropriate description
* HDFS:File system
* MapReduce: Parallel processing framework
* YARN:Resource Negotiator
* The technology used by hadoop to index and look up the search result is similar to that of which search engine?
* Google
* Location where the hive stores all its data on HDFS is:
* /user/hive/warehouse
* One shot execution of a hive script is stored with which of the following extension?
* HQL
* The key constraints Hive support are?

Options:

Foreign keys

Primary keys

Unique keys

* Correct Answer: None
* The join that hive supports is :
* Equi-join
* Function works on one row as input and returns multiple rows as output.
* User Defined Tabular
* Which of the following is/are not the characteristics of Map data type?
* Ordered collection of pairs
* Fixed size
* Collection data type
* What are the privilegeS Given To ThE default directory '/user/hive/warehouse'?
* Group Right(Read & Write)
* An exception is thrown by the hive while inserting the data due to something called CBO, what's the acronym of CBO?
* Cost Based Optimized
* with \_\_\_\_\_\_\_ The explode function can be used on both ARRAYs and MAPs.
* Lateral View
* Which Statements are True according to Views ?
* Views are Virtual Table's
* Views have no representation in underlying HDFS store
* Normalized Storage in tradition databases helps IN
* Minimum Redundancy
* Optimized storage
* To Updates Data fast
* Denormalized Storage in Hive:
* Have Minimize Joins
* Will optimize number of disk seeks
* Which are the set operations that we can use on hive?
* Union
* Union All

PySpark

* Arrange the following steps in the correct order in RDD.
* Read, Split, Flatmap, Map, ReduceByKey
* -------is an action that returns all the elements of the dataset, RDD of a PySpark
* Collect
* Apache Spark is developed in which language
* Scala
* Which cluster manager do spark support?
* Apache MESOS
* Hadoop Yarn
* Standalone Cluster Manager
* Which of the following is not true for Hadoop and Spark?
* Both have their own file system
* In how many ways RDD can we created?
* 3
* What are the features of Spark RDD?
* Lazy evaluation
* Fault tolerance
* In-memory computation
* Fault Tolerance in RDD is achieved using
* DAG
* Which of the following is the entry point of Spark Application?
* SparkContext
* Can we modify RDD?
* No
* Which of the following is the entry point of Spark SQL?
* SparkSession
* which of the following is not a transformation?
* Reduce

Spark2

* In which of the following Action the result is not returned to the driver.
* foreach()
* Which of the following is action?
* CountByValue()
* catalyst optimizer works well with which type of data?
* Semi-structured data
* what is the function to create table that is per-session?
* Create or Replace Temp View
* GroupBy can only be used on pair RDDs?
* True
* What is the heart of any spark application?
* spark context
* The map() operation is not supported by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
* Dataframes
* The columns in DataFrame can contain only primitive datatype values, which makes it easier to access.
* False
* Which of the following number refers to current row in a frame?
* 0
* Temporary tables can be accessed from multiple spark sessions.
* False
* An “add-only” shared variable that tasks can only add values to is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
* Accumulator
* Once a SparkConf object is passed to Spark, it is cloned and can no longer be modified by the user
* True
* what are the services run inside the driver ?
* SparkEnv
* DAGScheduler
* SparkIU
* Spark 2.0 uses \_\_\_\_\_\_,an engine that speeds up execution \_\_\_\_\_\_\_\_
* Tungsten,10-20x

Stream Processing with Spark Streaming

* What are the parameters defined to specify window operation
* Window length, sliding interval
* Which of the following is true for stateful transformation?
* Uses data or intermediate results from previous batches and computes the result of the current batch.
* Which of the following is not output operation on Dstream
* . ReduceByKeyAndWindow
* How many tasks does Spark run on each partition
* one Task
* How many messages are included in one RDD if the log messages are received at the rate of four messages per second and the batch interval is five seconds.in the DStream?
* 20
* What are the basic components of any Spark application?
* Driver, Executor and Receiver
* Which of the following options represent the objectives of Clustering?
* Maximize intra-cluster similarity
* Maximize inter-cluster differences
* Minimize inter-cluster similarity
* Explain in brief. What is Forgetfulness Metric? Why is it Important in ML? What are the 2 ways it can be applied in K-Means Clustering Algorithm?
* Measure to balance the importance of new data vs old data. Makes algorithm adaptive to changing dataset. By specifying Decay factor or Half life
* What should be the decay factor if we want to use only the most recent data?
* 0
* What are the ways to specify the half-life?
* Number of batches and number of data points
* Advantages of checkpoint are:
* limiting state re-computation in case of failure
* fault tolerance for driver jobs
* \_\_\_\_\_\_\_\_\_is the open standard for token based authorization which allows a user's account to be used by a 3rd party without exposing password.
* Oauth
* Spark streaming provides exactly how many number of semantics for processes?
* 1
* The coordinator process which executes the user program?
* Driver

Kafka

* Which maven package contains the Kafka-Spark streaming connector?

spark-streaming-kafka-0-8\_2.11

spark-core\_2.11

kafka-streaming

kafka-clients

Ans: a

* What is needed to create a Kafka stream using the KafkaUtils.createStream function?

Spark streaming context

GroupId

Zookeeper address

Topics

Kafka parameters

Broker list

Ans:a,b,c,d

* Which Spark streaming function is used to combine streams that are running in parallel?

map()

unify()

union()

combine()

reduce()

Ans:c

* Which stream function should be used to output data to Kafka?

foreachRDD()

print()

output()

send()

foreach()

Ans:a

* Which data structure is used to optimize the Kafka output process in parallel streams?

pool

map

constant

list

Ans:a

* What is needed to create a Kafka stream using the KafkaUtils.createDirectStream function?

Spark streaming context

GroupId

Topic list

Zookeeper address

ConsumerStrategy

LocationStrategy

Ans:a,e,f

* Which location strategy should be implemented when both Spark executors and Kafka brokers are on the same host?

Subscribe

Assign

PreferBrokers

PreferFixed

PreferConsitent

Ans:c

* Which KafkaUtils function is used to create an RDD based on offset ranges?

map()

foreachRDD()

RDD()

createRDD() (a)

createDirectStream()

Ans:d

* Which function is used to get current Kafka offsets for an rdd?

KafkaUtils.getOffsets(rdd)

rdd.getOffsets()

((HasOffsetRanges) rdd.rdd()).offsetRanges() (a)

((HasOffsetRanges) rdd. offsetRanges())

Ans:c

* Which Spark stream function should be used to split a string into an array of words?

map()

foreachRDD()

reduce()

split()

Ans:

* What are the five core APIs of Kafka

Ans:Producer,consumer,streams,connect,admin

* Specify the levels of the kafka protocol

Ans: Application

presentation

session

transport

network

datalink

physical

* which of the one is SSL-encrypted communication

Presentation

session

transport

network

Ans:a

* Which one of the following describes data in streaming data?

bounded and continuous

bounded and discrete

unbounded and continuous

unbounded and discrete

Ans:c

* Which is the acknowledgment level that provides the best delivery guarantee?

total

zero

all

one

Ans: c

* Given a high load topic with small message sizes and no requirements from the latency perspective, which property would help to increase the throughput the most?

max.in.flight.request.per.second

batch.size

acks

linger.ms

Ans:d

* What kind of mechanism do Kafka Consumers use for retrieving records?

command

push

pull (a)

event-based

Ans:c

* Given a topic with five partitions, what is the maximum number of consumers from a consumer group that can be active at the same time?

infinite

one

five

ten

Ans: c

* Given a topic that should contain only filtered data from another topic, which technology would you use to achieve this?

kafka broker

kafka connect

kafka producer

kafka streams

Ans:d

* Deleting records before certain offsets can be done by which of the following?

partition

cluster

broker

topic

Ans:a

* Given an external system that your consumer needs to send data to, when is the best time to commit the offsets?

Immediately after the request has been sent

When the consumer chooses to commit

After a response from the external system

About one minute after consuming the message

Ans:c

* Which tool should a new custom application be integrated with to produce data to Kafka?

kafka producer

rest proxy

kafka connect

admin client

Ans:a

* In which format the kafka broker stores records?

Serialize

Deserialize

Binary

Both a and b

Ans:c

Quiz

* \_\_\_\_\_\_\_\_\_\_ helps to speed up the data retrieval in traditional relational databases.

Ans:Indexes

* Default directory for storing Hive table data is\_\_\_\_\_\_\_\_\_\_

Ans:/user/lkmuser/warehouse

* Command used for viewing the partitions?

Ans:

* For optimizing join of three tables, the largest sized tables should be placed as

The first table in the join clause

Second table in the join clause

Does not matter

Third table in the join clause

Ans:d

* \_\_\_\_\_ mode allows all partitions to be dynamic

Ans:Non-strict

* To see the partitions key present in the hive table the command used is?

Describe

show

show extended

Describe extended

Ans:d