Which commonly available Hadoop distribution is offered as a provision-able service through both Amazon Web Services Elastic MapReduce (EMR) and Google Compute Engine?

- [x] MapR
- [ ] Cloudera
- [ ] Hortonworks
- [ ] Hadapt

Which of the following vendors offers an MPP Data Warehouse appliance?

- ✗ Hortonworks
- ○ Tableau
- ✓ Teradata
- ○ Informatica

Pig was developed by Facebook as a data warehouse application for Hadoop.

- ✓ False
- ✗ True

What new resource manager was released in HDFS 2.0?

- ✗ HBase
- ○ Pig
- ✓ YARN
- ○ Amabari

Massively Parallel Processing (MPP) data warehouses are always queried using imperative code.

- ✗ True
- ✓ False

Which of the following is a use case for Sqoop?

- [x] Bulk transfers of data of structured data into HDFS
- [ ] Moving large amounts of log files into HDFS
- [ ] Analysis unstructured or semi-structured data
- [ ] Allows random read-and-write in MapReduce

Which command shares the same functionality as HDFS dfs commands?

- [x] Hadoop fs
- [ ] HDFS dfsadmin
- [ ] Hadoop dfs
- [ ] Hadoop ds

HDFS allows for files to be updated.

- [x] False
- [ ] True

Hadoop hybrid products allow MapReduce querying of relational data and/or co-locate Hadoop and a relational database engine on the same cluster.

- False
- True

Which of the following is not a typical Hadoop stack member/distribution component?

- ✗ Mahout
- ○ Pig
- ✓ Thunderbird
- ○ Hive

```
id: integer
name: string
account_idlist: array<string>
```

What is the query you would run to get the list of all accounts along with the customer ids for all customers?

- ○ select id, explode(account_idlist) as account_ids from customers;
- ✗ select id, account_ids from customers
  lateral view explode(account_idlist) as account_ids;
- ✓ select id, account_ids from customers
  lateral view explode(account_idlist) accountsTable as account_ids;

Next question

Which of these scheduling policies have the longest wait time?

- ○ Last In First Out Scheduler
- ✗ Capacity Scheduler
- ○ Fair Scheduler
- ✓ First In First Out

```
Orders: 100GB
Products: 500MB
Reviews: 1GB
```

If the most common join operation performed was on the Products and Reviews tables and you wanted to make the join operation between these tables as fast as you possibly could, what design decisions would you take?

- ✗ Bucket the table on the product id column, have the number of buckets in one table be a multiple of another and join on the product id column.
- ✓ Bucket the table on the product id column, have an equal number of buckets on each table and sort each bucket on product id, join on the product id column.

Next question

Question 6 of 10

Which of these is the difference between a managed and an external table?

- ✗ Managed table metadata is stored in HDFS, and external table metadata is stored in the metastore.
- ✓ Managed table records are stored in the warehouse directory, and external table records are stored elsewhere in HDFS.
- ○ Managed table records are stored in the metastore, and external table records are stored in HDFS.

Next question

Question 10 of 10

What is the container that YARN uses when scheduling tasks on the cluster?

- ✗ The memory and CPU required across all nodes in the cluster to complete the task
- ✓ The logical unit of all the resources that a process needs, memory and CPU
- ○ The memory and CPU required on the name node to process a task
- ○ The storage capacity of a node which is required to process a task

See results

What is the default location for Hive data?

- ○ `/user/hive/warehouse` on the local file system
- ✓ `/user/hive/warehouse` on HDFS
- ○ `/warehouse` on the local file system
- ✗ `/warehouse` on HDFS

Next question

Why is the left semi-join faster than the IN/EXISTS subqueries?

- [x] The left semi-join scans the table on the right side of the join only until the matching record is found, it does not scan the entire table.
- [ ] The left semi-join uses only one column in the join operation which is more efficient than using multiple columns like in the IN/EXISTS subqueries.
- [ ] The left semi-join runs entirely in memory which makes it more performant.

Next question

Question 7 of 10

What are the advantages of using the Hive Beeline interface over the older Hive command line?

- Faster and more efficient queries because of additional parallelism
- ✓ Support for multiple concurrent clients and more robust security
- Better formatting and color coding on the terminal
- Greater fault tolerance due to more replicas of data

Next question

Question 1 of 10

Which of the following is NOT a reason that map-only joins are faster than joins which run both map and reduce phases?

- ✗ Complete MapReduce operations have additional steps such as shuffle and sort between the map and reduce phases which gets eliminated in map-only joins.
- ○ Map-only joins reduce data transfer within a cluster.
- ○ Improved processing time because one phase is entirely eliminated.
- ✓ The memory overhead of a MapReduce job is much greater than a Map-only job.

Next question

Which are possible ways in which join performance can be optimized?

- [x] Structuring joins such that the reduce phase of the underlying MapReduce is eliminated
- [ ] Matching of join records is made really fast by caching data
- [ ] Ensuring that multiple processes work on data in large tables
- [ ] Structuring joins in such that the reduce phase is run on more than one machine

Next question

How would you specify a window between the first row in a result set and the current row?

- ✗ `rows between 6 preceding and current row`
- ✓ `rows between unbounded preceding and current row`
- ○ nothing, this is the default specification
- ○ `rows between unbounded preceding and and unbounded following`

Next question

Analytical processing jobs include all of these characteristics EXCEPT:

- ✗ Works on huge datasets from multiple sources
- ○ Reads data, does not usually edit or update it
- ✓ Deals mostly with recent data, collected in the last few hours or days
- ○ Tends to involve long running jobs

Next question

Question 8 of 10

Which of the following is an important reason that Hive stores data in a denormalized form?

- Joins are performant and can be performed to get all data related to an entity in one go.
- Redundancy is not acceptable even though disk space is cheap.
- Read operations are more common in Hive, so disk seeks should be minimized for better performance.

Next question

When the table on the right side of a join operation is much smaller than the table on the left side of a join operation, which of the following are map-only joins?

- ✗ Full-outer join
- ○ Right-outer join
- ✓ Left-outer join

Next question

Question 7 of 10

What is the name of the tool used to automate deployments in the cloud?

- ○ Cloudera Distribution including Hadoop
- ✗ Cloudera CDH
- ✓ Cloudera Director
- ○ Cloudera Manager

Next question

Which of these is NOT a characteristic of views in Hive?

- ✗ Views are only stored as a query when created and are not executed till used.
- ○ Views can include one or more tables in the query.
- ✓ Data in the view is frozen in time and does not change along with the underlying data in the table.
- ○ The structure is frozen in time and not affected by changes in the underlying table.

Next question

Which of these is NOT a supported scalar type in Pig?

- ✗ • double**
- ○ **float**
- ○ **bytearray**
- ✓ **string**
- ○ **int**

Next question

Why are joins with one join column faster than joins with 2 join columns?

- [x] The number of MapReduce jobs run to perform the join is equal to the number of join columns. Fewer columns, fewer jobs; so faster.
- [ ] The amount of data processed is increases as the number of join columns increase. Less data; faster joins.
- [ ] The amount of data loaded into memory is directly proportional to the number of join columns. Fewer columns, less memory used; so faster.

Question 2 of 10

Which are the names of the HBase processes that run when HBase is started up in pseudo-distributed mode?

- RegionServer, Master, Zookeeper
- NameNode, DataNode and SecondaryNameNode
- NodeManager, ResourceManager
- HRegionServer, HMaster, HQuorumPeer

What do you call the main program in a Spark application?

- ✗ Cluster manager
- ✓ Driver
- ○ Session
- ○ Context

Which command would you use to see all tables that are present in HBase?

- ✗ status
- ○ dir
- ✓ list
- ○ ls
- ○ describe

Question 6 of 10

Which of the following are the only 2 operations possible on an RDD?

- [x] Transformation and Action
- [ ] Aggregation and Update
- [ ] Transformation and Retrieval
- [x] Update and Retrieval
- [ ] Aggregation and Action

Which of the following functions is NOT a valid load/store function Pig?

- ✗ **JsonLoader()**
- ○ **HBaseStorate()**
- ○ **AvroStorage()**
- ○ **CSVExcelStorage()**
- ✓ **RDBMSStorage()**

How would you perform grouping and aggregation operations on HBase?

- ✗ Using intermediate tables to hold results
- ○ Using multiple shell commands piped together
- ○ Using the Structured Query Language
- ✓ Using the MapReduce programming model

families: personal, academic
personal:name = name of each student academic:year = grade in which the student currently studies
row key = student id
You want to find the total number of students in each grade. What would be the operation in the reduce phase?

- ✗ Find the minimum number of students for each grade key.
- ○ Find the maximum number of students for each grade key.
- ○ Average the number of students for each grade key.
- ✓ Sum the number of students for each grade key.

Next question

Question 4 of 10

What does CSD stand for?

- [x] Custom Service Descriptor
- [ ] Cloudera Service Description
- [x] Custom Service Definition
- [ ] Cloudera Standalone Data

What data structure represents the actions and transformations in a Spark application?

- ✗ Binary Tree
- ○ Maximum Heap
- ○ Queue
- ✓ Directed-acyclic Graph
- ○ Undirected Graph

Which is the correct Spark method to use to convert a Python list to an RDD?

- ✗ sc.fromDF()
- ○ sc.make()
- ○ sc.toRDD()
- ✓ sc.parallelize()

What is the name of the original core abstraction in Spark?

- ○ Dataset[Row]
- ✗ Dataset
- ✓ RDD
- ○ DataFrame

Which is the default sort order in Spark?

- [x] Ascending
- [ ] It is required to specify via a function
- [ ] Descending

Next question

Question 9 of 10

Which format do you use to read from external databases?

- ○ format('database')
- ○ format('mysql')
- ✓ format('jdbc')
- ✗ format('external')

Which of these functions is used to return data back to the driver?

- ✗ Foreach
- ○ For
- ✓ Collect
- ○ Print

Question 10 of 10

Which is the correct Spark method to use to convert a Python list to an RDD?

- sc.make()
- sc.toRDD()
- sc.fromDF()
- sc.parallelize()

You're analyzing the data loaded into a dataframe and you find that some records have some fields missing. How will you get rid of these records so you don't process them further?

- [x] df.dropna()
- [ ] df.filter(df['some_col'] != NaN)
- [ ] df.dropmissing()
- [ ] df['col1', 'col2', 'col3'], keep only those columns with no missing information

Which of these are examples of stream processing?

- ✗ Analyzing new users to see whether the rate of new signups is higher on mobile or the website
- ○ Analyzing reviews to see if the overall sentiment for a product is good or bad
- ○ Analyzing orders on an e-commerce site to see what zip codes they're concentrated at
- ✓ Analyzing log messages to see whether all pages of the e-commerce site are up and running

What is the major difference in applying k-means clustering on batch vs streaming data?

- [x] Determining how relevant older data is compared with newer data in the stream
- [ ] The algorithm needs to scale to handle more data as compared with batch operations.
- [ ] More iterations of the algorithm as newer batches of data arrive
- [ ] The algorithm changes based on how fast the newer data arrives.

What does the explode() function in Spark do?

- Creates a new batch for each group of elements
- Creates a new window for each group of elements
- ✓ Creates a new row in the dataframe for every element
- Creates a new column for each element

Given a NoSQL database, which component is most properly suited to transfer and adapt the data to stream processing?

- ✗ Kafka Streams
- ○ Kafka Producer
- ✓ Kafka Connect
- ○ Kafka Broker

You're tracking trending news stories over the last 2 days with your streaming application and you receive a new batch of updates every hour. **News received a day ago is only half as important as the news received in the last hour. **How would you specify this forgetfulness in your streaming k-means algorithm?

- ✗ Half-life of 12 batches
- ✓ Half-life of 24 batches
- ○ Half-life of 1000 news articles
- ○ Half-life of 24 articles

Question 7 of 10

Given a topic that has to store user-related information (name, address, etc.), which type of clean-up policy would be the most appropriate?

- ✗ Delete + Compact
- ○ None
- ✓ Compact
- ○ Delete

Next question

What is the relationship between a DStream and an RDD?

- [x] A DStream is made up of a sequence of RDDs, where every RDD contains entities which are received within a batch interval.
- [ ] An RDD is a group of DStreams where each DStream contains entities received within a batch interval.
- [ ] A DStream is just one RDD that contains all the entities available in a stream.

Which of the following are characteristics of RDDs in Spark?
A: They are distributed across multiple machines in the cluster
B: They can be reconstructed when a node crashes
C: They are stored on the hard disks of cluster machines
D: They are lazily evaluated

- ✗ A, B and C only
- ○ B and D only
- ○ B, C and D only
- ✓ A, B and D only
- ○ A and D only

Next question

How does a subscriber in Kafka receive messages from a publisher?

- [x] A subscriber subscribes to a topic and receives messages from all publishers to that topic
- [ ] A subscriber subscribes to a message and receives all messages which are of the same type
- [ ] A subscriber subscribes to an individual publisher and receives all messages that the publisher sends out

Question 6 of 10

Which of these operations require real-time processing?

- [x] Requests on sale days to an e-commerce site
- [x] Daily indexing of web pages to serve search operations
- [ ] Setting up a heat map of the location of the most active Twitter users in the last month
- [ ] Running operations to see weekly trends in sales on an e-commerce site

Your e-commerce sites sends every transaction with product id as a stream to a Spark application. The data sent can be assumed to always arrive on time.
You want to keep a global count of how many items of each product were sold and save this to an RDBMS.
What kind of output mode will you use when you write out this information if you want to minimize updates to the RDBMS?

- ✗ Append mode
- ○ Complete mode
- ✓ Update mode

Next question

You're tracking trending news stories over the last 24 hours with your streaming application and you receive news updates every hour. What kind of decay factor will you use to perform streaming k-means clustering?

- [x] 0 < decayFactor < 1
- [ ] decayFactor = 0
- [ ] decayFactor = 1

Question 2 of 10

What is a broker in Kafka?

- ○ A single topic in a Kafka cluster
- ✓ A single node in a Kafka cluster
- ○ A single subscriber in a Kafka cluster
- ✗ A single publisher in a Kafka cluster

multiplicative operation represented by this pseudocode:

```
multiple(a, b) {
  return a * b
}
```

**What is the pseudocode for the inverse function **of this in the sliding window operation?

✗
```
subtract(a, b) {
    return a - b
}
```

○
```
sum(a, b) {
    return a + b
}
```

✓
```
divide(a, b) {
    return a/b
}
```

Next question

Question 4 of 10

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

- [x] Partition
- [ ] Topic
- [ ] Broker
- [ ] Cluster

Given a Kafka cluster composed of five brokers, what is the maximum number of replicas that a topic can have?

- [x] Five Replicas
- [ ] Two Replicas
- [ ] One Replica
- [ ] Ten Replicas

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

- ✗ Kafka Connect
- ○ Admin Client
- ○ REST Proxy
- ✓ Kafka Producer

How does Kafka allow to implement transactions across microservices?

- [x] It can serve as a WAL for distibuted systems
- [ ] It implements a two-phase commit protocol across multiple databases
- [ ] It should be used as the only database in a system that should serve all user queries directly
- [ ] Different databases can coordinate update operations by communicating via Kafka

What is "stateless stream processing" refer to?

- ✗ ".map()" function in Kafka Streams
- ○ Storing an intermediate state to process each record
- ✓ Processing records in a stream when each record can be processed independently
- ○ All processing implemented using Kafka Streams

How are records ordered in Kafka?

- ✗ Records are ordered per partition in lexicographical order
- ○ Records are ordered per topic in the order they were written to Kafka
- ✓ Records are ordered per partition in the order they were written to Kafka
- ○ Records are not ordered in Kafka

Question 9 of 10

What approach to analytics allows us to perform ad-hoc queries?

- ○ Either analytics-on-read or analytics-on-write
- ✓ Analytics on-read
- ○ Analytics on-write
- ○ Neither of these approaches

What are the components of a Kafka record?

- ○ Topic name, key, value, headers, timestamp
- ✓ Key, value, headers, timestamp
- ○ Topic name, headers, timestamp
- ○ Key, headers, timestamp

How can we track removal operations with immutable events?

- ○ We should write removal events to a separate datastore
- ○ We can remove arbitrary records from a Kafka topic
- ✓ We need to create a separate event that marks a removal operation
- ○ We should copy all existing events from a Kafka topic, except the removed event

How does Kafka allow to implement transactions across microservices?

- [x] It can serve as a WAL for distibuted systems
- [ ] It implements a two-phase commit protocol across multiple databases
- [ ] It should be used as the only database in a system that should serve all user queries directly
- [ ] Different databases can coordinate update operations by communicating via Kafka

What is the main benefit of using Kappa architecture over Lambda architecture?

- ✗ Lambda architecture forces only to use mutable data instead of building on top of immutable events.
- ○ Kappa architecture allows implementing real-time applications, while Lambda architecture is only about batch processing applications
- ○ Lambda architecture does not allow to use Kafka
- ✓ With Kappa architecture, each algorithm should only be implemented once

What is the name of a window type in Kafka Streams that divides records into overlapping groups?

- ✗ Session window
- ○ Tumbling window
- ○ Sliding window
- ✓ Hopping window

What is the benefit of reading WAL to reflect updates from a database in Kafka?

- ○ It allows preserving timestamps of records
- ✗ It is easier to implement than alternatives
- ○ WAL is supported by more datastores
- ✓ It allows to keep track of updates and deletions in a database

What approach would you use to implement a complex distributed transaction if the goal is to make the result as maintainable as possible?

- ✗ Choreography
- ○ Neither of these options
- ○ Administration
- ✓ Orchestration

Which of the following are input arguments to the window function in structured streaming?
A: The event time
B: The window interval
C: The sliding interval
D: The batch interval

- ✗ B, C and D only
- ○ A and D only
- ○ A and B only
- ✓ A, B and C only
- ○ A, C and D only

Next question

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?

- ✗ acks
- ○ batch.size
- ✓ linger.ms
- ○ max.in.flight.request.per.second

Which of the following are important characteristics of any stream processing system?
A: Ability to work with time-sensitive data
B: Manage out of order events
C: Replay data for recalculations

- ✗ B and C only
- ○ B only
- ○ A only
- ○ C only
- ✓ A, B and C

Question 3 of 10

Your e-commerce sites sends every transaction with product id as a stream to a Spark application. The data sent can be assumed to always arrive on time.
You want to keep a global count of how many items of each product were sold and save this to an RDBMS.
What kind of aggregation would you perform on the incoming stream?

- [x] Group by product id and count the number of products
- [ ] Group by day and count the number of products
- [ ] Group by sale price and sum the number of products

Next question

Given a topic called "orders," how many producers can you use to produce data on this topic?

- ✗ Exactly two
- ✓ More than five
- ○ Exactly one
- ○ Less than five

Question 8 of 10

Which of the following is an improvement present in Spark 2 but not in Spark 1?

- ○ Ability to work with dataframes and datasets
- ✓ Unified APIs for batch as well as stream processing
- ✗ Lazy evaluation of datasets when an action is requested
- ○ Fault-tolerant distributed datasets

Question 5 of 10

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

- ○ Event-based
- ✓ Pull
- ○ Push
- ○ Command

Which of the following is NOT a valid difference between batch and streaming data?

- ✗ The order of data received is generally more important in streams
- ○ Streams are constantly updated, batches may not be
- ○ Batches are bounded datasets, streams are unbounded datasets
- ✓ Streams have a single global state that does not change over time

Why are RDDs, the basic Spark programming abstraction, considered resilient?

- [x] RDDs keep track of their lineage; they can be reconstructed from the source even when cluster nodes crash.
- [ ] After every transformation, an RDD has its data replicated to the memory of multiple nodes and can be retrieved from other machines if nodes crash.
- [ ] After every transformation, an RDD has its data replicated to multiple nodes in the file system and it's permanently stored on more than one machine.

Next question

Question 9 of 10

You receive log messages at the rate of two messages per second and your batch interval is five seconds. If the sliding window size is 20 seconds, how many RDDs are included in the window at any point in time?

- [x] 4
- [ ] 8
- [ ] 20
- [ ] 40
- [ ] 10

Which option best describes data in streaming?

- [x] Unbounded and Continuous
- [ ] Unbounded and Discrete
- [ ] Bounded and Continuous
- [ ] Bounded and Discrete

Question 2 of 10

What is checkpointing in streaming Spark applications?

- Backing up all input data sources on multiple machines in the cluster
- Periodically saving data so state can be reconstructed from this intermediate context
- Backing up all RDDs on stable storage after every transformation

Question 8 of 10

What do you need to do in order to be able to run SQL queries on your dataframe?

- [x] Create a temporary view of your data that registers it as a SQL table
- [ ] Set the enable SQL flag to true to use SQL queries
- [x] Format your data in rows and columns and write it out to disk

Which of these statements are true about HBase?

- ✗ Only primitives can be row keys.
- ✓ The column family is specified when you create the table.
- ○ The column families can be dynamically changed at any time.
- ○ The column values uniquely identify a row.
- ○ All rows should have the same column names.

What do you call a logical division of data?

- ✗ Block
- ○ File
- ○ Division
- ✓ Partition

Question 5 of 10

If the output data type of the mapper is ImmutableBytesWritable, LongWritable the input arguments to the reduce() method is:

- ✗ `ImmutableBytesWritable, LongWritable`
- ○ `ImmutableBytesWritable, Collection<LongWritable>`
- ✓ `ImmutableBytesWritable, Iterable<LongWritable>`
- ○ `LongWritable, ImmutableBytesWritable`

Question 7 of 10

Which of these best defines an RDD?

- A table of rows and columns which allow array manipulations
- A collection of records which are held in memory across multiple machines in a cluster
- A data type which allows operations that are idempotent
- A map of key values pairs which are held in memory and allow fast lookup

Question 4 of 10

What are the main features of the Hadoop distributed computing framework?

- [x] Storage and processing of data across multiple machines with fault tolerance and recovery
- [ ] Distribution of processes but not storage across multiple machines while handling fault tolerance and recovery
- [ ] Using single powerful machine which integrates with other dumb terminals all of which send their processes to the single machine
- [ ] Multiple machines which are connected to each other but each job runs only on one machine at a

Next question

When you sample data, what occurs if you specify withreplacement = True?

- ✗ The probability of each item from being selected increases as other items are taken
- ✓ The probability of each item from being selected remains constant as other items are taken
- ○ The probability of each item from being selected decreases as other items are taken

Question 10 of 10

What is the command to retrieve a single column value from a Result object representing a row in the HBase table?

- [x] byte[] grade = result.getValue(Bytes.toBytes("academic"), Bytes.toBytes("grade"))
- [ ] string grade = (String) result.getValue(Bytes.toBytes("academic"), Bytes.toBytes("grade"))
- [ ] byte[] grade = result.getValue("academic", "grade")
- [ ] byte[] grade = result.getValue(Bytes.toBytes("academic"))

See results

Question 6 of 10

Hadoop has some limitations such that it cannot be used as a database. Which of the following is NOT a limitation of Hadoop?

- [x] Data in Hadoop has a well-defined structure and that structure is well-enforced.
- [ ] Hadoop alone is more suited for long-running batch processes.
- [ ] Hadoop alone does not allow random access to individual records.
- [ ] Hadoop can store files of any kind, text, binary, media files, etc.

Question 2 of 10

How does lazy evaluation function in a Spark RDD?

- ○ All transformations write out intermediate results to disk and so can be lazily combined later.
- ✓ Transformations are only applied when a result is requested, they are evaluated on demand.
- ○ All operations on an RDD are performed in memory not on disk.

What is the command to see what partitions exist in a Hive table?

- ✓ show partitions <tablename>
- ✗ display partitions <tablename>
- ○ describe table <tablename>
- ○ describe formatted <tablename>

Which of the following set operations are allowed in Hive?

- ✗ minus
- ○ intersect
- ✓ union
- ○ not

Which of these is NOT a feature of a data warehouse?

- ✓ Loss of data is acceptable and even expected, and lost data is ignored.
- ○ Data may be available a couple of days after the updates were made.
- ○ They are optimized to serve reads, and certain write operations may not even be allowed.
- ○ Records may be stored for many years.

Question 6 of 10

What is the name of Cloudera's Platform-as-a-Service offering?

- ○ Cloudera Director
- ○ CDH
- ✗ Cloudera Manager
- ✓ Cloudera Altus

Question 3 of 10

What are the default data types for the input to the map phase when reading from an HBase table using a TableMapper class?

- ✗ `Byte, Result`
- ○ `Integer, String`
- ○ `ImmutableBytesWritable, Row`
- ✓ `ImmutableBytesWritable, Result`

Which of these is NOT a characteristic of a relation in Pig?

- ✓ A relation is analogous to a table in traditional databases and has strict schema and data type constraints.
- ○ A relation is immutable, contents of a relation cannot be edited in-place.
- ○ A relation is ephemeral, it survives only for the duration of a Pig session.
- ○ A relation is referenced by a name, similar to a variable in a programming language.

How does the MapReduce programming model take advantage of multiple machines in a cluster?

- The Map operation is performed on multiple machines on different subsets of data, and the Reduce operation is performed in parallel with the Map process

- The Map operation is performed on a single machine, and the Reduce operation is parallelized across multiple machines

- The Map operation is performed on multiple machines on different subsets of data, and the Reduce operation combines the output of various mappers

Which of the following is NOT a way to sample data from Hive tables?

- ✗ Bucket Sampling
- ✓ Random Sampling
- ○ The LIMIT Keyword
- ○ Block Sampling
- ○ Row Count Sampling

While implementing the MapReduce classes using Hadoop libraries, the output key and value type of the mapper have to match:

- ✗ The input key and value types of the mapper
- ○ The output key and value types of the reducer
- ✓ The input key and value types of the reducer

When would you choose to use broadcast variables in Spark?

- ○ When you want to announce cluster updates to all nodes
- ✓ When you want to efficiently copy over data to each node to reduce network data transfer
- ○ When you want to perform join operations on 2 dataframes

Question 4 of 10

Which of the following statements about Project Tungsten in Spark 2 is false?

- It optimizes Spark queries like a compiler would rather than like a DBMS would.
- ✓ It contains many virtual function calls to help model very complex operations.
- It is faster than the volcano iterator model of Spark 1.
- It heavily utilizes loop unrolling to make operations faster.

Which of these is NOT true of unions in Pig?

- ✗ Pig preserves duplicates in the final union
- ○ Pig does not preserve the order of tuples in the final union
- ○ Pig requires that the schema of the relations be compatible
- ✓ Pig will resolve schema incompatibilities by padding the tuples with nulls

Which of the following is true of the flatten function in Pig?

- ✗ It applies to the tuple complex data type and extracts individual fields as separate records.
- ○ It applies to the map complex data type and extracts key-value pairs as separate records.
- ✓ It applies to the bag complex data type and extracts each tuple as separate records.

```
employee_id: chararray
name: chararray
title: chararray
department: chararray
salary: double
address: tuple (apartment: chararray,
 street: chararray, zipcode: int)
training: map[chararray]
```

**

**

The `training` field is a map of all the training sessions the employee has attended. It maps the name of the training session to the date on which it was held.
Access the employee name and the dates on which they attended the 'Policy101' training and store it in a relation

```
**foreach employees generate
name,
training#'Policy101';**
```

Next question

Which of the following attributes is not true of MapReduce processing?

- ✗ Has a Map step for preprocessing and a Reduce step for aggregating
- ○ Is used by Hadoop, with Java in its native coding environment
- ✓ Processes relational or dimensional data using SQL or MDX exclusively
- ○ Pioneered at Google

What function would you use to apply a transformation to every element of an RDD?

- ✗ rdd.mapReduce()
- ○ rdd.withColumn()
- ○ rdd.filter()
- ○ rdd.transform()
- ✓ rdd.map()

What role does ETL, Extract, Transform and Load play when you work with data?

- [x] Allows you to pull data from original sources and pre-process them in useful ways before it is stored in a data warehouse for analysis
- [ ] Allows you to transform data when moving it from one database to another
- [ ] Has plugins to multiple databases to allow transforming data and loading into a variety of destination databases
- [ ] Allows you to work with multiple sources, even if they are remotely located to bring them all together in the same format

Next question

What is the Pig function that is used to deserialize and load data from the file system?

- ✗ **PigDeserializer()**
- ○ **PigLoader()**
- ✓ **PigStorage()**
- ○ **PigLoadStore()**
- ○ **PigStore()**

Question 1 of 10

If no data type is specified for a certain field when loading a relation, what is the default data type that Pig assumes for this field?

- ✗ **chars**
- ○ **bytes**
- ○ **chararray**
- ✓ bytearray
- ○ **int**

Question 2 of 10

Which is these is NOT a primitive data type in Hive?

- Decimal
- String
- Month
- Integer
- Boolean

Which of these is NOT a standard scheduling algorithm in YARN?

- ✗ Capacity Scheduler
- ○ First In First Out Scheduler
- ○ Fair Scheduler
- ✓ Shortest Job First Scheduler

Which is the correct syntax for a join operation on Hive tables?

- select students.id, students.name, courses.title from students, courses where students.id = courses.student_id

- select students.id, students.name, courses.title from students join courses where students.id < courses.student_id

- ✓ select students.id, students.name, courses.title from students join courses where students.id = courses.student_id

Which of the following vendors does not offer a Hadoop distribution of its own?

- [x] Cloudera
- [ ] IBM
- [ ] SAP/Sybase
- [ ] MapR

When writing a Bash script for HDFS, you must specify 'HDFS' in script.

- [x] False
- [ ] True

What is one major drawback of the Capacity Scheduler?

- [x] Underutilization of cluster resources
- [ ] Excessive overhead in queue management
- [ ] Starvation of jobs with short processing times
- [ ] Long wait times

HiveQL is best described as a:

- ✗ Expressive language
- ○ Scripting language
- ✓ Declarative language
- ○ Procedural language

Why does the pseudo-distributed mode require password-less Secure Shell (SSH) login to be set up?

- ✗ SSH is required to encrypt client interactions with the cluster so it is secure, and passwords ensure that this communication is smooth and not interrupted by password requests.
- ○ SSH is required for communication between master and slave nodes in the cluster and manual password logins do not work with machine to machine communication.
- ✓ SSH is required for communication between master and slave nodes in the cluster, and since this communication is very frequent, you reduce the overhead by configuring

Hadoop can be run on cloud computing infrastructure.

- [x] True
- [ ] False

What is the command to submit a MapReduce job to a Hadoop cluster using the command line?

- ○ `hadoop submit`
- ✓ `hadoop jar`
- ○ `hadoop run`
- ○ `hadoop fs`

Bash scripts can automate what HDFS task?

- ✗ Optimizing YARN applications
- ○ Checking HBase job status
- ✓ Ingesting, exporting, and moving data
- ○ Updating Masternode key pairs

Which of the following is a valid definition of big data?

- ○ Into an MPP data warehouse appliance
- ✓ All of these
- ○ Loaded onto Hadoop or NoSQL clusters
- ○ Analysis of petabyte-scale collections of data

When removing a file using the rm command, files are deleted permanently.

- ○ True
- ✓ False

What are the main characteristics of the Hadoop Distributed File System?

- ✗ Built on specialized hardware which has a very low failure rate, suited to batch operations and supports large datasets
- ○ Built on commodity hardware, fault tolerant, suited to low latency transactional operations and supports large datasets
- ✓ Built on commodity hardware, fault tolerant, suited to batch operations and supports large datasets
- ○ Built on commodity hardware, fault tolerant, suited to low latency transactional operations and supports large structured datasets with well defined fields

Why do reads from the HDFS have to access both the name node and the data nodes?

- ✗ Reads to the name node check the file permissions and reads to the data nodes read where the file is located and the file contents.
- ○ Reads to the name node update edit logs and reads to the data nodes retrieve file contents.
- ○ Reads to the name node retrieve the contents as the name node in turn reads from the data nodes.
- ✓ Reads to the name node look up where a file and the corresponding blocks are located on the cluster, and reads to the data nodes read the actual contents of the file.

What is the role of MapReduce and YARN in Hadoop versions after 2.0?

- ✗ MapReduce helps the various parallel processes communicate with each other and YARN is responsible for management of resources in the cluster to help run processes.
- ○ YARN defines the parallel processing tasks and MapReduce is responsible for running the processes on nodes in the cluster.
- ○ MapReduce contains the parallel processing logic and YARN partitions the data so the processes can run efficiently.
- ✓ MapReduce defines the parallel processing tasks and YARN is responsible for running the

Which of the following is a valid justification for use of NoSQL technology for Big Data applications?

- ✗ The Hadoop Distributed File System is explicitly tailored to unstructured data and RESTful APIs
- ○ Many NoSQL databases are based on open source technology
- ○ NoSQL's use of scale-out architectures work well for big data
- ✓ All NoSQL products use MapReduce in their query interfaces

How do you empty the trash in HDFS dfs?

- [x] expunge
- [x] empty
- [ ] trash_out
- [ ] refresh