Big Data: The Big Picture

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

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

- Which of the following is a valid justification for use of NoSQL technology for Big Data applications?
    - All NoSQL products use MapReduce in their query interfaces

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

- NoSQL and NewSQL are interchangeable terms.
    - False

- Hadoop can be run on cloud computing infrastructure.
    - True

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

- Which of the following attributes is not true of MapReduce processing?
    - Processes relational or dimensional data using SQL or MDX exclusively

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

- Which of the following is a valid definition of big data?
    - all of the above

- Which of the following is a typical motivation for replacement of the Hadoop Distributed File System (HDFS)?
    - All of the above

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

- ## Getting Started with HDFS

- HDFS allows for files to be updated.
  - False
- Bash scripts can automate what HDFS task?
  - Ingesting, exporting, and moving data
- Before dropping a table in HBase, what command must you use first?
  - `Disable`
- Which HDFS dfs command deletes the local copy once the file is put into HDFS?
  - `moveFromLocal`
- In Sqoop commands what does the "m" argument represent?
  - Number of MapReduce jobs
- Best practices for Bash scripts are to start with?
  - #!/bin/bash
- Pig was developed by Facebook as a data warehouse application for Hadoop.
  - False
- In what language is HDFS written?
  - Java
- How many columns can HBase Scale to?
  - Billions
- Sqoop ships with Hortonworks Data Platform.
  - True
- What is Pig's scripting language called?
  - Pig Latin
- What kind of database is HBase?
  - NoSQL
- When writing a Bash script for HDFS, you must specify 'HDFS' in script.
  - False
- NameNode is responsible for remembering where data is stored in the cluster.
  - True
- How do you empty the trash in HDFS dfs?

    - `expunge`

- What new resource manager was released in HDFS 2.0?
    - YARN

- Hive was designed for ____ data, while Pig excels with _____ data.
    - structured, unstructured

- Bash scripts are written in what language?
    - POSIX commands

- Sqoop is an application that allows you to import and export data from:
    - Relational and NoSQL Databases

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

- What allows for HDFS to work during hardware failures?
    - Fault tolerance

- Which of the following commands creates a zero-length file in HDFS?
    - `HDFS dfs -touchz`

- Hive and Pig both abstract away Java to allow developers to write MapReduce jobs in SQL-Like syntax.
    - True

- HiveQL is best described as a:
    - Declarative language

- Which of the following is a use case for Sqoop?
    - Bulk transfers of data of structured data into HDFS

- What permission model does HDFS default on files?
    - POSIX

- Best practices for Bash scripts are to start with?
    - #!/bin/bash

- HBase uses SQL.
    - False

The Building Blocks of Hadoop - HDFS, MapReduce, and YARN

- Why do reads from the HDFS have to access both the name node and 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 command to submit a MapReduce job to a Hadoop cluster using the command line?
  - `hadoop jar`

- What is the container that YARN uses when scheduling tasks on the cluster?
  - The logical unit of all the resources that a process needs, memory and CPU

- What is one major drawback of the Capacity Scheduler?
  - Underutilization of cluster resources

- In a distributed system, if you double the number of machines in the cluster, in the best case, your overall computing capacity and speed would be:
  - A little less than twice the original capacity and speed

- What are the characteristics of Hadoop running in pseudo-distributed mode?
  - It runs two JVMs, uses a single node and runs both HDFS and YARN.

- Why does the pseudo-distributed mode require password-less Secure Shell (SSH) login to be set up?
  - 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 authentication keys rather than using passwords.

- What is the trade-off on the block size specified in HDFS?
  - Balancing larger block sizes which results in less parallel processing and smaller block sizes which results in greater overhead of managing processes

- The website Glassdoor wants to analyze salaries based on profession, experience, and other factors across a cross section of its users. It sends a form to its users asking them for information and collects this in a massive text file in this format:
  - `Occupation, Company, Location, Years Of Experience, Gender, Salary`
  - Glassdoor wants use this data to show average salaries based on occupation. What is the key value output of the map phase?
  - `Occupation, Salary`

- What is the role of MapReduce and YARN in Hadoop versions after 2.0?
  - MapReduce defines the parallel processing tasks and YARN is responsible for running the processes on nodes in the cluster.

- What is checkpointing of the name node?

    - Merging of the `fsimage` and `edits` files on the secondary name node which are then copied over to the name node
- Which of these is NOT a standard scheduling algorithm in YARN?
    - Shortest Job First Scheduler

- Which of these scheduling policies have the longest wait time?
    - First In First Out

- What are the characteristics of Hadoop running in standalone mode?
    - It runs a single JVM, uses the local file system, and does not run HDFS and YARN processes.

- The default replication factor for pseudo-distributed mode is:
    - 1

- What is `jps` and why do you use it when you set up Hadoop on your local machine?
    - `jps` lists all the Java processes running on your machine and is used to check whether the Hadoop component processes are running.

- What are the main characteristics of the Hadoop Distributed File System?
    - Built on commodity hardware, fault tolerant, suited to batch operations and supports large datasets

- 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 reducer

- By default where does Hadoop run the web interface to monitor jobs on the cluster?
    - http://localhost:8088

- What does the basic software running on a distributed system do?
    - Distribute data across machines, manage processes, and handle multiple processes

- How do the metadata files `fsimage` and `edits` help with recovery of the name node?
    - The `fsimage` file contains a snapshot of the file system at start up, and the `edits` file tracks all changes made in the file system; together they help reconstruct the name node.

- What are the factors which govern how many replica locations are chosen for fault tolerance in HDFS?
    - More replicas maximize redundancy but increase the intra-cluster bandwidth for write operations; fewer replicas result in less fault tolerance.

• GETTING STARTED WITH HIVE

- 1.Consider a customer's table in a bank where a customer can have more than one account with
- these fields:
- `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, account_ids from customers
    lateral view explode(account_idlist) accountsTable as account_ids;
```

    - 2.Which of the following set operations are allowed in Hive?
        - Union
    - 3.Which of these is NOT a characteristic of views in Hive?
        - Data in the view is frozen in time and does not change along with the underlying data in the table.
- 4.Which of the following is an important reason that Hive stores data in a denormalized form?
    - Read operations are more common in Hive, so disk seeks should be minimized for better performance.
- 5.Which is the default database that Hive uses for the metastore?
    - The built-in Derby database
- 6.Which is these is NOT a primitive data type in Hive?
    - Month
- 7. What is the default location for Hive data?
    - `/user/hive/warehouse` on HDFS
- 8.What are the implications of Hive being built on top of a distributed computing framework?
    - Hive queries run as parallel processes on the underlying cluster to query data.
- 9.What is the role of the metastore in the Hive data warehouse?
    - Stores information such that files stored on HDFS can be represented as tables to the user.
- 10.Which of these is the difference between a managed and an external table?
    - Managed table records are stored in the warehouse directory, and external table records are stored elsewhere in HDFS.
- 11.Which of the following jobs would you classify as transaction processing rather than analytical
- processing?

- Editing bank account details when you deposit money into an account

- 12.If you have a table `"students"` in Hive in the default database, and you have one other database called `"ecommerce"`, what are the directories in the Hive warehouse?

  - A directory called `"students"` and a directory called `"ecommerce.db"`

- 13.Consider a customer's table in a bank where a customer can have more than one account with these fields:
- `id: integer`
- `name: string`
- `account_idlist: array<string>`
  - What is the query you would run to get the list of all accounts for all customers?
    - `select explode(account_idlist) as account_ids from customers;`
- 14. Consider a customer's table in a bank where a customer can have more than one account with these fields:
- `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, account_ids from customers`
- `lateral view explode(account_idlist) accountsTable as account_ids;`

- 15. Which of these is NOT a feature of a data warehouse?
  - Loss of data is acceptable and even expected, and lost data is ignored.

- 16. Which is the correct syntax for a join operation on Hive tables?
  - `select students.id, students.name, courses.title from students join courses where students.id = courses.student_id`

- 17. What are the advantages of using the Hive Beeline interface over the older Hive command line?
  - Support for multiple concurrent clients and more robust security

- 18.If you have a text file and where a `#` separates each field and collection items are separated by a `:`, what additional information will you specify as a part of the create table command?

  - `row format delimited fields terminated by '#'`
  - `collection items terminated by ":"`

- 19. Are Hive tables ACID-compliant?
  - No, they are not ACID-compliant by default, but they can be if certain properties are set in hive-site.xml.

- 20. How does Hive impose table schema on its underlying data?
- Hive uses schema-on-read where it checks to see if the data match the schema during read operations.

• Data Transformations with Apache Pig:

- 1.What is the Pig function that is used to deserialize and load data from the file system?
- Ans:PigStorage()

- 2.Which of the following are case-insensitive in Pig?
- Ans: Commands such as LOAD, ORDER BY, GROUP BY, STORE

- 3.What is lazy evaluation of a relation?
- Ans:The relation contents are associated with a schema only when they are needed i.e. when they are stored to a file or displayed on screen.

- 4.Let's say the field names and data types of your relation is as follows:

- Relation name: employees

- employee_id: chararray
- name: chararray
- title: chararray
- department: chararray
- salary: double
- address: tuple (apartment: chararray, street: chararray, zipcode: int)

- How would you write a foreach-generate statement to extract the name and zipcode for every employee in the organization into a resultant tuple?

- Field names are as specified.

- Ans:foreach employees generate name, address.zipcode;

- 5.Which of the following conditional operators can you use with maps and tuples?

- A: ==

- B: >=

- C: <=

- D: !=

- E: > and <
- Ans:A and D only

- 6.Which of these is NOT true of unions in Pig?
- Ans: Pig will resolve schema incompatibilities by padding the tuples with nulls

- 7.Which of the following is true of the flatten function in Pig?
- Ans: It applies to the bag complex data type and extracts each tuple as separate records.

- 8.When would we choose to use Pig to work with data?
- Ans:When the data has no fixed schema and may have missing fields

- 9.What role does ETL, Extract, Transform and Load play when you work with data?
- Ans: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

- 10.What is the language that you would use with Apache Pig for commands?

- Ans:Pig Latin

- 11.Let's say the field names and data types of your relation is as follows:

- Relation name: employees

- 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

- Ans:foreach employees generate name, training#'Policy101';

- 12.Which of the following functions is NOT a valid load/store function Pig?
- Ans:RDBMSStorage()

- 13.Which of these is NOT true of unions in Pig?
- Ans:Pig will resolve schema incompatibilities by padding the tuples with nulls

- 14.How does the nested foreach help when parsing large datasets
- Ans:Allows you to specify multiple operations as you are iterating over every record in a data set

- 15.How does the MapReduce programming model take advantage of multiple machines in a cluster?
- Ans:The Map operation is performed on multiple machines on different subsets of data, and the Reduce operation combines the output of various mappers

- 16.Which of these is NOT a supported scalar type in Pig?
- Ans:string

- 17.What is Pig's command line interface called?
- Ans: Grunt Shell

- 18.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?
- Ans:bytearray

- 19. Which of the following are supported complex data types in Pig?

- A: Tuple

- B: Relation

- C: Bag

- D: Set

- E: Map

- Ans:A, C and E only

- 20.Let's say the field names and data types of your relation is as follows:

- Relation name: employees

- employee_id: chararray
- name: chararray
- title: chararray
- department: chararray
- salary: double
- address: tuple (apartment: chararray, street: chararray, zipcode: int)

- How would you write a foreach-generate statement to extract the name, title and department of every employee in this organization into a resultant tuple?

- Use only field indexes, assume field names are not available.

- Ans:foreach employees generate $1, $2, $3;

- 21.Consider the follow relation to represent students in a school

- Relation name: students

- name: chararray
- grade: int
- contact: chararray

- If you group by the grade field what is the structure of the resultant relation?

- Ans:The "group" field is grade and the "students" field is a bag of tuples with "name", "grade" and "contact" as its fields

- 22.Which of these is a join that Pig does NOT support?
- Ans:Non-equi join

- 23.What are the reasons one would run Pig in local mode?
- Ans:To run on a single machine environment which makes it easy to develop scripts and debug them

- 24.Which of the following statements is NOT true in Pig?
- Ans:The DISTINCT keyword can act on a single field across all tuples

- 25.How does the nested foreach help when parsing large datasets?
- Ans:Allows you to specify multiple operations as you are iterating over every record in a data set

- 26.Which of these is NOT a characteristic of a relation in Pig?
- Ans:A relation is analogous to a table in traditional databases and has strict schema and data type constraints.

- 27.Consider the follow relation to represent students in a school

- Relation name: students

- name: chararray
- grade: int
- contact: chararray

- If you group by the grade field and store it in the grades_group relation how would you count the number of students per grade?
- Ans:foreach grades_group generate group, COUNT(students);

- 28.Which Apache distributed computing framework does Pig support?

- A: Flink

- B: MapReduce

- C: Tez

- D: Spark

- Ans:B, C and D

Getting Started with HBase: The Hadoop Database

- What are the main advantages of a columnar store over a traditional rows and columns layout?
    - Ans Correct choice:
    - Missing fields do not result in wasted space and attributes of each record can be specified dynamically.

- Which are the pieces of information needed to access **one** value in an HBase table?
- Correct choice:
    - Row key, Column Family, Column Name, Timestamp

- Which of these statements are true about HBase?
- Correct choice:
    - The column family is specified when you create the table.

- How would you check for the existence of a particular table?

```
connection -> the connection object to the HBase table
admin -> instance of administration object for the HBase table
conf -> HBase table configuration
```

- Correct choice:
    - admin.tableExists()

- Instantiate a new Put instance to add data to a table:
- Correct choice:

```
Put put = new Put(Bytes.toBytes("234"))
```

- What are the arguments that a RowFilter accepts?
    - Correct choice:

    - A compare operation and a comparator

- What are the default data types for the input to the map phase when reading from an HBase table using a TableMapper class?
- Correct choice:
    - `ImmutableBytesWritable, Result`

- What kind of filter would you use to filter records based on column values in HBase?
- Correct choice:
    - SingleValueColumnFilter

- How would you perform grouping and aggregation operations on HBase?
- You chose: correct choice:
    - Using the MapReduce programming model

- What do ACID properties of traditional databases enforce?
    - You chose: correct choice:
    - While booking a ticket, the seat reservation and the charge on your credit card will both occur or neither will occur.

- What are the main advantages of de-normalized storage in a columnar store?
- Correct choice:
    - Requires just one disk seek and read operation to read all details of a single entity.

- All of these are differences between HBase and a traditional database EXCEPT:
- Correct choice:
    - HBase displays ACID properties only at the column level, and relational databases enforce ACID properties in their entirety.

- What command would you use to access **only the names of students having student ids in the range 5 to 7, both inclusive**?

```
Table name: students
Column families: personal, academic
```

- Correct choice:

```
scan 'students', {COLUMNS => ['personal:name'], STARTROW => "5", STOPROW => "8"}
```

- How do you delete a table in HBase?

```
Table name: students
Column families: personal, academic
```

- Correct choice:
- `disable 'students' and drop 'students'`

- `Table name: students`
- `Column 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 use MapReduce to find the total number of students in each grade. What would be the output of the map phase?
- Correct choice:
    - `<academic:year, 1>`

- The input key, value pair of the reduce phase should match which of the following?
- Correct choice:
    - The output key, value pair of the map phas

- If the output data type of the mapper is ImmutableBytesWritable, LongWritable the input arguments to the reduce() method is:
- Correct choice:
    - `ImmutableBytesWritable, Iterable<LongWritable>`

- What are the main features of the Hadoop distributed computing framework?
- Correct choice:
    - Storage and processing of data across multiple machines with fault tolerance and recovery

- Which is the correct format of the command to add data to an HBase table?
- `Table name: students`
- `Column families: personal, academic`
- correct choice:
- `put 'students', 'Student123', 'personal:name', 'John'`

- What is the command to retrieve a single column value from a Result object representing a row in the HBase table?
- correct choice:
- `byte[] grade = result.getValue(Bytes.toBytes("academic"), Bytes.toBytes("grade"))`

- What are the main advantages of a columnar store over a traditional rows and columns layout?
- Correct choice:
    - Missing fields do not result in wasted space and attributes of each record can be specified dynamically.

- Which is the correct format of the command to add data to an HBase table?
- `Table name: students`
- `Column families: personal, academic`
- Correct choice:
- `put 'students', 'Student123', 'personal:name', 'John'`

- What kind of filter would you use to filter records based on column values in HBase?
- Correct choice:
    - SingleValueColumnFilter

- `Table name: students`
- `Column 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?
- Correct choice:
- Sum the number of students for each grade key.
- Instantiate a new Put instance to add data to a table:
- correct choice:
    - `Put put = new Put(Bytes.toBytes("234"))`

- What are the arguments that a RowFilter accepts?
- You chose: correct choice:
    - A compare operation and a comparator
- What are the main features of the Hadoop distributed computing framework?
- You chose: correct choice:
    - Storage and processing of data across multiple machines with fault tolerance and recovery

- Which command would you use to see all tables that are present in HBase?
- Correct choice:
    - list

- How do you delete a table in HBase?
- `Table name: students`
- `Column families: personal, academic`
- Correct choice:
    - `disable 'students' and drop 'students'`

- `Table name: students`
- `Column 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 use MapReduce to find the total number of students in each grade. What would be the output of the map phase?
- Correct choice:
    - `<academic:year, 1>`

- Hadoop has some limitations such that it cannot be used as a database. Which of the following is NOT a limitation of Hadoop?
- Correct choice:
    - Data in Hadoop has a well-defined structure and that structure is well-enforced.

- What command would you use to access **only the names of students having student ids in the range 5 to 7, both inclusive**?

- `Table name: students`
- `Column families: personal, academic`

- correct choice:
    - `scan 'students', {COLUMNS => ['personal:name'], STARTROW => "5", STOPROW => "8"}`

- Which are the names of the HBase processes that run when HBase is started up in pseudo-distributed mode?
- Correct choice:
    - HRegionServer, HMaster, HQuorumPeer

- Hadoop has some limitations such that it cannot be used as a database. Which of the following is NOT a limitation of Hadoop?
- Correct choice:
    - Data in Hadoop has a well-defined structure and that structure is well-enforced.

Developing Spark Applications with Python & Cloudera

- 1 What do you use to create a Dataset from a DataFrame?
  - .as[]

- 2. What is used in Python to indicate the continuation of a logical line?
  - Correct choice:
    - \

- 3. Which port is used by default for the Spark Web UI?
  - Correct choice:
    - 4040

- 4. What do you call the graph of transformation operations required to execute when an action is called?
  - Correct choice:
    - RDD lineage

- 5.What function do you use to provide configuration parameters?
  - Correct choice:
    - option()

- 6.What is the name of Cloudera's Platform-as-a-Service offering?
  - Correct choice:
    - Cloudera Altus

- 7.Which version of Scala do you need to run Spark 2 in a CDH cluster?
  - You chose: correct choice:
    - 2.11

- 8.Which method can be used to create a DataFrame?
  - You chose: correct choice:
    - createDataFrame

- 9.What Spark package should you use if you want to use TensorFlow?
  - You chose: correct choice:
    - TensorFrames

- 10.What type of function is collect?
  - You chose: correct choice:
    - An action

- 11.What is the name of the Cloudera Hadoop administration tool?
  - Correct choice:
    - Cloudera Manager

- 12.What type of class do you use to specify types in a Dataset?
    - Case class

- 13.Which type do you use to represent a schema?
    - Correct choice:
        - StructType
- 14.What is the name of the file used to configure application properties, which is used by Spark during execution?
    - Correct choice:
        - spark-defaults.conf
- 15.What command do you use to launch Spark 2 in a Python shell in a Cloudera distribution?
    - Correct choice:
        - pyspark2
- 16.Which function can be used to output the lineage of a particular RDD?
    - Correct choice:
        - toDebugString()
- 17.What is the name of Cloudera's Platform-as-a-Service offering?
    - You chose: correct choice:
        - Cloudera Altus
- 18.What does DSL stand for?
    - You chose: correct choice:
        - Domain Specific Language
- 19.What is the name of the original core abstraction in Spark?
    - You chose: correct choice:
        - RDD
- 20.How can you tell if a Spark function is a transformation?
    - You chose: correct choice:
        - It returns a new RDD
- 21.What does CSD stand for?
    - Correct choice:
        - Custom Service Descriptor
- 22.What is the name of the component that serves as an entry point to a Spark application when using the higher level API?
    - Correct choice:
        - SparkSession
- 23.Which function do you use to replace null values?
    - Correct choice:
        - fillna
- 24.How much faster is Spark than MapReduce?
    - Correct choice:
        - 10x to 100x

- 25.How can you define Map?
    - Correct choice:
        - Apply function to each element

- 26.Which of these lists is ordered from slowest to fastest method of working with data?
    - You chose: correct choice:
        - One at a time, parallel, and then pipeline
- 27.Which component of Spark optimizes for CPU and memory efficiency?
    - You chose: correct choice:
        - Project Tungsten
- 28.Which function can be used to register user defined functions?
    - You chose: correct choice:
        - udf
- 29.What is the recommended indentation according to PEP 8?
    - You chose: correct choice:
        - Use four spaces
- 30.When you save a Picklefile, what are you saving?
    - You chose: correct choice:
        - A SequenceFile with multiple Pickle objects
- 31.What version of IPython do you need to install to use with Python 2.x?
    - Correct choice:
        - IPython 5.x LTS
- 32.What do you need when you want to concatenate a string a column?
    - Correct choice:
        - lit('yourstring')
- 33.Which component of Spark optimizes queries, thus improving performance?
    - Correct choice:
        - Catalyst Optimizer
- 34.What are the three modes for handling corrupt records?
    - Correct choice:
        - PERMISSIVE, DROPMALFORMED, or FAILFAST
- 35.Which year was the RDD paper published?
    - Correct choice:
        - 2012
- 36.What does Spark use to determine how to shuffle data in a shuffle boundary?
    - You chose: correct choice:
        - Partitioner

- 37.Which component is in charge of getting resources for execution and allocating them to a particular job?
    - You chose: correct choice:
        - Cluster manager
- 38.What is another name for Python's interactive shell?
    - You chose: correct choice:
        - REPL
- 39.How many SparkContexts can exist per application?
    - You chose: correct choice:
        - 1
- 40.What is the name of the type of transformation applied to all items in a dataset?
    - You chose: correct choice:
        - Coarse grained transformations
- 41.When you sample data, what occurs if you specify withreplacement = True?
    - Correct choice:
        - The probability of each item from being selected remains constant as other items are taken
- 42.Which class do you use to work with missing or corrupt data?
    - Correct choice:
        - DataFrameNAFunctions
- 43.Which format is used for the StackOverflow/StackExchange dumps?
    - Correct choice:
        - XML
- 44.Which function do you use to see the schema of a DataFrame?
    - Correct choice:
        - printSchema
- 45. In which deployment mode does the Driver run in a node within the cluster even if launched from outside?
    - Correct choice:
        - Cluster deployment mode
- 46.Which function do you use to apply an operation to all items in a collection?
    - Correct choice:
        - Map
- 47.What is the name of the community index of third-party packages for Apache Spark?
    - Correct choice:
        - spark-packages.org
- 48.Which function can be used to clone a column in a DataFrame?
    - You chose: correct choice:

        - withColumn

- 49.When running Spark Standalone, what is the default assumed location?
    - You chose: correct choice:
        - file:///

- 50.What are the 3 cluster managers supported by Spark?
    - You chose: correct choice:
        - Standalone, Mesos, and Yarn

- 51.Which action is available in Spark to coalesce all rows in a partition into an array?
    - Correct choice:
        - Glom

- Related Clip: Partition Operations: MapPartitions and PartitionBy
- 52.Where can you discover the current Java version in Cloudera Manager?
    - Correct choice:
        - Support > About

- 53.What do you call the main program in a Spark application?
    - Correct choice:
        - Driver

- 54. What was the most important change in the release of Spark 2.0?
    - Correct choice:
        - The unification of the Dataset and DataFrame API

- 55.What is it called when there is data shuffling involved?
    - Correct choice:
        - Wide transformation

- 56.What are the available SaveModes?
    - Correct choice:
        - append, overwrite, error, and ignore

- 57.What year was Python 2 released?
    - You chose: correct choice:
        - 2000

- 58.What do you have in a cell instead of a value?
    - You chose: correct choice:
        - A Catalyst Expression

- 59. Which is the language used to create Spark?
    - You chose: correct choice:
        - Scala

- 60.What do you call a logical division of data?

- You chose: correct choice:
  - Partition

- 61.Which set operation explodes the size of the two datasets being joined?
  - Correct choice:
    - Cartesian
- 62.What do you use to install Spark 2 Standalone on a Mac?
  - Correct choice:
    - Homebrew
- 63.How do you save or persist a DataFrame to a table?
  - Correct choice:
    - saveAsTable
- 64.What do you use to install IPython in a Cloudera cluster?
  - Correct choice:
    - Anaconda
  - 65.What is the name of the component that serves as an entry point to a Spark application when using the higher level API?
  - Correct choice:
    - SparkContext
- 66.What happens if you assign an integer to a string variable in Python?
  - Correct choice:
    - The variable becomes of type integer and holds the value.
- 67.Which of these functions is used to return data back to the driver?
  - Correct choice:
    - Collect
- 68.Which of the following are the four languages that you can use in Spark?
  - Correct choice:
    - Scala, Python, Java, and R
- 69.What do you call to trigger execution?
  - Correct choice:
    - An action
- 70. What is another way of reading a parquet file instead of:
- `spark.read.format('parquet').load('/myfile')`
- You chose: correct choice:
  - spark.read.parquet('/myfile')

Getting Started with Stream Processing with Spark Streaming

- 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?
    - Ans: **Half-life of 24 batches**
- What is checkpointing in streaming Spark applications?
    - Ans: **Periodically saving data so state can be reconstructed from this intermediate context**
- What are the basic components of any Spark application?
    - Ans: **Driver, Executor and Receiver**
- Which of these operations require real-time processing?
    - Ans: **Requests on sale days to an e-commerce site**
- How is stream processing in Spark an extension of its batch processing framework?
    - Ans: **Stream processing operates on DStreams, which are made up of sequences of RDDs. Spark processes these individual RDDs in a stream just as it would during batch processing.**
- When are transformations considered stateful?
    - Ans**: When they operate across multiple RDDs rather than a single RDD**

- 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?
        - Ans: **4**

- In a sliding window operation, let's say the summary function you want to apply across a window is a 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?
- Ans:

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

- What attributes can represent a student in school?
    - Ans: **Age, grade, test scores**
- What is the major difference in applying k-means clustering on batch vs streaming data?
        - Ans**: Determining how relevant older data is compared with newer data in the stream**

- What is the relationship between a DStream and an RDD?

- Ans: **A DStream is made up of a sequence of RDDs, where every RDD contains entities which are received within a batch interval**.

- Under what conditions would you use the updateStateByKey() function? Ans: **Summarizing across an entire DStream where each RDD in the stream is a pair in RDD.**

- 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?
  - Ans**: 0 < decayFactor < 1**
- Which of these are examples of stream processing?
  - Ans: **Analyzing log messages to see whether all pages of the e-commerce site are up and running**
- Why are RDDs, the basic Spark programming abstraction, considered resilient?
  - Ans: **RDDs keep track of their lineage; they can be reconstructed from the source even when cluster nodes crash.**
- What are the two operations that can be performed on an RDD in Spark?
  - Ans**: A transformation to create a new RDD and an action to retrieve results from an RDD**
- Why does MapReduce NOT work well for streaming data?
  - Ans: **It operates on a huge number files stored in reliable storage, typically running jobs which can take hours or even days.**
- You receive log messages at the rate of two messages per second and your batch interval is five seconds.How many messages are included in one RDD in the DStream?
  - Ans**:10**

• Structured Streaming in Apache Spark 2 (Py Spark)

- 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?
- **…….UPDATE MODE**

- 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

- Correct choice:
  A, B and C only
- ---
- Which of the following is an improvement present in Spark 2 but not in Spark 1?
- Correct choice:
  Unified APIs for batch as well as stream processing
- ---
- Which of the following is a timestamp that is assigned at the source of the streaming data?
- Correct choice:
  Event time
- ---
- How does a subscriber in Kafka receive messages from a publisher?
- Correct choice:
  A subscriber subscribes to a topic and receives messages from all publishers to that topic
- ---
- What does the explode() function in Spark do?
- Correct choice:
  Creates a new row in the dataframe for every element
- ---
- 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?
  - *Ans))Update mode*

- ---
- What kind of aggregation would you perform on the incoming stream?
- correct choice:
  Group by product id and count the number of products
- ---
- Your e-commerce sites sends every transaction with product id and sale price as a stream to a Spark application. The data sent can be assumed to always arrive on time.
- You want to know how much revenue each product brought in on a per day basis.
- What kind of operations would you perform on the incoming stream?
- correct choice:
  Group by product id, apply a window for every 24 hours and sum the sale price
- I---
- Which of the following is an advantage of structured streaming in Spark 2?
- correct choice:
  A stream is simply batch data that is constantly appended to, both can be treated the same way
  ---
- What is a broker in Kafka?
- correct choice:
  A single node in a Kafka cluster
- ---
- A single column data type in the schema for structured streaming is specified using what data type?
- Ans)Struct field
- ---
- Which of the following relationships is correct regarding timestamps in streaming data?
- Choice)EventTime>InjestionTime>processingtime

- ----------
- 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
- Choice)A,B,C
- ----------
- Which of the following is true of structured streaming
- A: Can only work with streams using micro-batches
- B: Interactions with output sinks are managed by the system
- C: Ad hoc queries on streams are possible
- Choice)B and C
- ----------
- 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
- Ans)))ABC
- ----------
- A single column data type in the schema for structured streaming is specified using what data type
- Ans))struct field
- ----------
- In structured streaming, what is a trigger?
- You chose: correct choice:
  Ans)))Events which determine when transformations on input data need to be re-performed
- ----------
- 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

- *Ans))A,B,C*----------
- In structured streaming, what is a trigger?
- Ans))Events which determine when transformations on input data need to be re-performed
- ----------
- What is a watermark as it applies to streaming data?
- Ans))
- A watermark is the allowable lateness threshold for entities in a stream
- ----------
- Which of the following is an advantage of structured streaming in Spark 2?
  A stream is simply batch data that is constantly appended to, both can be treated the same way
- ----------
- 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?
- Ans)))Group by product id and count the number of products
- ----------
- Which of the following is correct about the how the complete output mode works in structured streaming?
  Ans))))))))))The entire result table is recalculated
- ----------
- Which of the following is NOT a valid difference between batch and streaming data?
  - Ans))))))))))Streams have a single global state that does not change over time

- # Handling Streaming Data with a Kafka Cluster:

- 1.Given a topic called “orders,” how many producers can you use to produce data on this topic?
- →More than five.
- 2.Given a topic that has to store user-related information (name, address, etc.), which type of clean-up policy would be the most appropriate?
  → compact.
- 3. Which is the acknowledgment level that provides the best delivery guarantee?
- →All.
- 4. 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?
  → linger.ms
- 5.What kind of mechanism do Kafka Consumers use for retrieving records?
- →Pull.
- 6.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?
- →Five.
- 7. Given a topic that should contain only filtered data from another topic, which technology would you use to achieve this?
- →Kafka Streams.
- 8. Deleting records before certain offsets can be done by which of the following?
- →Partition.
- 9. Given a NoSQL database, which component is most properly suited to transfer and adapt the data to stream processing?
- →Kafka Connect.
- 10. Given an external system that your consumer needs to send data to, when is the best time to commit the offsets?
- →After a response from the external system.
- 11. Given a high load topic with big message sizes and no requirements from the latency perspective, which property would help to increase the throughput the most?
- →batch.size
- 12. Given two consumers with different group IDs, how many times will the messages be consumed?
- →Each message can be consumed more than once by each consumer.
- 13. Which tool should a new custom application be integrated with to produce data to Kafka?
- →Kafka Producer
- 14. Which option best describes data in streaming?
- →Unbounded and Continuous
- 15. Given a Kafka cluster composed of five brokers, what is the maximum number of replicas that a topic can have?

- →Five Replicas