Fundamentals of Big Data Analytics

What does the "4Vs" refer to in describing big data?

Veracity, Volume, Variety, Velocity

What are some of the challenges faced by big data? (Choose all that apply)

Storing big data in an inherently parallel architecture while maintaining a reasonably fast access to data

Higher chances of failures due to more hardware used

Scalability of the algorithms for processing big data

Which statement is NOT true about the MapReduce?

All other options are true

MapReduce hides system-level details from the application developer.

MapReduce is a programming model for processing vast amounts of data on large clusters

All developers have to do is to write two functions: Map and Reduce and let the system manage the parallel execution, coordinate the tasks, and deal with task failures)

MapReduce sits on top of a Linux file system

Which statement is NOT true about MapReduce's functions?

All other answers are true

Reduce function returns the final output key-value pairs

Map function returns the extracted information as a new list of intermediate key-value pairs

Reduce function takes the intermediate key-value pairs produced by several map functions.

Map function takes a key-value pair as input

Which statement is NOT true about the Hadoop Distributed File System (HDFS)?

All other options are true

Performs much better with small number of large files rather than large number of small files

Designed more for batch processing of data

Used for storage of very large files

Consists of a name node and a set of data nodes

Divides a large data file into chunks or blocks and replicates them on multiple nodes in the cluster

There is only one ResourceManager (RM) that runs on a worker (or slave) node in a Hadoop cluster.

True

False

Yarn is a new resource management incorporated in Hadoop 2.x. Yarn allows multiple data processing engines to run on a single Hadoop cluster. In Yarn, the Resource Manager is responsible for allocating resources to competing applications.

True

False

Which statement is NOT true about the Hadoop Rack Awareness?

All other options are true

Never loose all data if entire rack fails

Keep bulky flows in-rack when possible

Assumption that in-rack is higher latency

Assumption that in-rack is higher bandwidth

Which statement is NOT true about the Namenode in the Hadoop Distributed File System (HDFS)?

The client reads and writes data directly into the namenode

All other options are true

Namenode stores metadata such as a name directory that keeps track of what node stores which block of data

The entire metadata of Namenode is stored in the main memory of Master node

Formatting name node causes loss of data that is stored in HDFS.

The secondary Namenode is a stand-by Namenode

True

False

Which of the following is NOT true about Pig Latin or Pig?

Pig Latin is a high-level language for expressing data analysis programs

All choices are true

Pig Latin is a parallel data flow engine

Piggybank is a collection of user contributed User Defined Functions (UDFs)

Pig is a platform for analyzing large data sets

Which of the following is NOT true about Pig Operations?

COUNT - requires a preceding GROUP ALL statement for global counts or GROUP BY statement for group counts

TOKENIZE - eliminates bag nesting

DUMP - display output results, will always trigger execution

LOAD - PigStorage() loads/stores relations using field-delimited text format

FOREACH ... GENERATE - iterates over the members of a bag, then results in another bag

Which of the following is NOT correct why we need Pig, according to the video clip-Apache Pig at Twitter?

All choices are true

Java MapReduce is extremly verbose, so 400 lines of Java becomes less than 30 lines of Pig

Java MapReduce's joins are very difficult

Java MapReduce is painful

Java MapReduce is difficult to make abstractions

Pig scripts are shown below. What is a name/meaning of $1?

students = LOAD 'student.txt' USING PigStorage() AS (name:chararray, age:int, gpa:float);
DUMP A;
(John,18,4.0F)
(Mary,19,3.8F)
(Bill,20,3.9F)

studentname = Foreach students Generate $1 as studentname;

Position notation
Field value
One dollar
Data Type
Name (variable)

We are going to find the sum of hours and miles logged by each driver. The Pig Latin scripts are shown below. Which command should you use in the blank?

```
drivers = LOAD 'Pig/drivers.csv' USING PigStorage(',');
raw_drivers = FILTER drivers BY $0>1;
drivers_details = FOREACH raw_drivers GENERATE $0 AS driverId,
                $1 AS name;
timesheet = LOAD 'Pig/timesheet.csv' USING PigStorage(',');
raw_timesheet = FILTER timesheet by $0>1;
timesheet_logged = FOREACH raw_timesheet GENERATE $0 AS driverId,
                 $2 AS hours_logged, $3 AS miles_logged;
grp_logged = GROUP timesheet_logged by driverId;
sum_logged = FOREACH grp_logged (  ?  ) GROUP as driverId,
           SUM(timesheet_logged.hours_logged) as sum_hourslogged,
           SUM(timesheet_logged.miles_logged) as sum_mileslogged;
```

FLATTEN
JOIN
GROUP
GENERATE
COUNT

Which of the following is NOT true about Apache HBase?

All is true
Support for updating records
Good for batch processing (scans over big files)
Provides Fast record lookup
Support for record-level insertion

In Google BigTable and HBase, all rows are always sorted lexicographically by their column family

True

False

Which of the following HBase Shell commands and descriptions is not correct?

List_namespace_tables: list or display the tables available in given namespace

Scan: view the data in HTable

List: list all the tables in HBase

Create: create a table

Disable: delete a table like 'drop'

HBase commands to create/insert/retrieve a table or data from a table are shown below. Which types of names/elements should be used in the blanks [A], [B], [C]?

```
// Create table
hbase(main):001:0> create 'test_tbl', [A]
0 row(s) in 2.4250 seconds
=> Hbase::Table - test_tbl

// Insert data
hbase(main):017:0> put 'test_tbl', [B], 'test_cf:test_column1','test_data1'
0 row(s) in 0.0590 seconds

// Retrieve
hbase(main):022:0> scan [C]
ROW COLUMN+CELL
rowkey1 column=test_cf:test_column1, timestamp=1..., value=test_data1
```

[A]: Column Family [B]: Rowkey [C]: Table

[A]: Rowkey [B]: Table [C]: Column Family

All choices are NOT correct

[A]: Column [B]: Rowkey [C]: Table

[A]: Rowkey [B]: Column Family [C]: Rowkey

The below shows the output of an HBase shell command. Which command is used to show the output?

```
hbase(main):001:0> <command> 'sslee777:truck_event'
Table sslee777:truck_event is ENABLED
sslee777:truck_event
COLUMN FAMILIES DESCRIPTION
{NAME => 'events', VERSIONS => '1', EVICT_BLOCKS_ON_CLOSE => 'false', NEW_VERSIO
N_BEHAVIOR => 'false', KEEP_DELETED_CELLS => 'FALSE', CACHE_DATA_ON_WRITE =>
'fa
lse', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0',
REPL
ICATION_SCOPE => '0', BLOOMFILTER => 'ROW', CACHE_INDEX_ON_WRITE => 'false',
IN_
```

MEMORY => 'false', CACHE_BLOOMS_ON_WRITE => 'false',
PREFETCH_BLOCKS_ON_OPEN =>
'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536'}
1 row(s)
Took 0.5548 seconds
Create
Show
Status
Echo
Describe

What are some of the challenges faced by big data? (Choose all that apply)
Storing big data in an inherently parallel architecture while maintaining a reasonably fast access to data
Scalability of the algorithms for processing big data

The MapReduce framework is a good solution for batch processing of very large data sets when the dataset is usually processed as a whole and is rarely updated in place.
Correct!
True
False

What is the input/output data structure in a MapReduce framework?
Files
Vectors
Trees
Hash Tables
Key-Value pairs

What is the role of the reduce function in MapReduce?
Filtering out unwanted records
Preparing data for processing
Sorting the data before mapping
Performing computation on mapped data

The output of all the reduce tasks in the MapReduce framework is written into the same file on HDFS.
True
False

"Under replication" in HDFS means which of the following?
No replication is happening in the data nodes
The frequency of replication in data nodes is very low
None of all choices

The number of replicated copies is less than as specified by the replication factor
Replication process is very slow in the data nodes

Which services does YARN provide via long-running daemons? (Choose all that apply)
Node Manager
Application Manager
Resource Manager
Container Manager
YARN Manager

Which of the following properties gets configured on hdfs-site.xml when you set up the Hadoop cluster? (Choose all that apply)
Host and port where MapReduce job runs
Java Environment variables
Directory names to store HDFS files
None of all choices
Replication factor

Which files contain the configuration setting for the min/max memory and the number of virtual cores to allocate to a single container when you setup the Hadoop cluster?
mapred-site.xml
yarn-site.xml
hdfs-site.xml
None of all choices
node-resource-mgr.xml

Which of the following is NOT true about Pig Latin or Pig?
Pig is a platform for analyzing large data sets
Pig Latin is a high-level language for expressing data analysis programs
Piggybank is a collection of user contributed User Defined Functions (UDFs)
All choices are true
Pig Latin is a parallel data flow engine

Which of the following is NOT true about Pig Operations?
FOREACH ... GENERATE - iterates over the members of a bag, then results in another bag
DUMP - display output results, will always trigger execution
TOKENIZE - eliminates bag nesting
COUNT - requires a preceding GROUP ALL statement for global counts or GROUP BY statement for group counts
LOAD - PigStorage() loads/stores relations using field-delimited text format

We are going to find the sum of hours and miles logged by each driver. The Pig Latin scripts are shown below. Which command should you use in the blank?

drivers = LOAD 'Pig/drivers.csv' USING PigStorage(',');
raw_drivers = FILTER drivers BY $0>1;
drivers_details = FOREACH raw_drivers GENERATE $0 AS driverId,
$1 AS name;
timesheet = LOAD 'Pig/timesheet.csv' USING PigStorage(',');
raw_timesheet = FILTER timesheet by $0>1;
timesheet_logged = FOREACH raw_timesheet GENERATE $0 AS driverId,
$2 AS hours_logged, $3 AS miles_logged;
grp_logged = GROUP timesheet_logged by driverId;
sum_logged = FOREACH grp_logged (?) GROUP as driverId,
SUM(timesheet_logged.hours_logged) as sum_hourslogged,
SUM(timesheet_logged.miles_logged) as sum_mileslogged;

COUNT
JOIN
FLATTEN
GROUP
GENERATE

Which of the following is NOT true about Apache HBase?

Support for record-level insertion
Provides Fast record lookup
Support for updating records
Good for batch processing (scans over big files)
All is true

How are column family members physically stored on the filesystem in HBase?

According to their data type
Based on alphabetical order
None of the others
Together in the same location
Separately for each member

What denotes a region in HBase?

The table it belongs to and its last row
Not any of the choices listed
Its first row and its last row
Its size threshold and its row boundary
The table it belongs to and its first row

HBase commands to create/insert/retrieve a table or data from a table are shown below. Which types of names/elements should be used in the blanks [A], [B], [C]?
// Create table
hbase(main):001:0> create 'test_tbl', [A]
0 row(s) in 2.4250 seconds

=> Hbase::Table - test_tbl

// Insert data
hbase(main):017:0> put 'test_tbl', [B], 'test_cf:test_column1','test_data1'
0 row(s) in 0.0590 seconds

// Retrieve
hbase(main):022:0> scan [C]
ROW COLUMN+CELL
rowkey1 column=test_cf:test_column1, timestamp=1..., value=test_data1
[A]: Column [B]: Rowkey [C]: Table
[A]: Rowkey [B]: Table [C]: Column Family
[A]: Rowkey [B]: Column Family [C]: Rowkey
All choices are NOT correct
[A]: Column Family [B]: Rowkey [C]: Table

Which of the following is NOT true about the Hive (data warehousing tool)?
Provides Java API and SQL-like interface (Hive QL)
All true
Can access HBase tables using Hive QL (Hive-HBase Integration)
Doesn't support 'group by' query
Provides data summarization, query and analysis

What is the primary advantage of using a standalone database configuration for the Hive metastore?
It allows for multiple concurrent sessions and users
None of the others
It provides better performance for SQL queries
It simplifies the setup process.
It reduces disk space usage

What is a trade-off of schema on read compared to schema on write?
Faster query time performance
Limited flexibility in schema design
Correct Answer
Slower initial load time
Data rejection if schema is not met
None of the others

Which of the following is NOT fully supported in Hive (HiveQL)?
transactions
Joins
Views

All others are fully supported
Indexes

Which of the following is NOT true about the Apache Hive?
Runs on the client machine
Generates MapReduce jobs that run on the Hadoop (or Spark) cluster
All true
High-level abstraction on top of MapReduce
Provides SQL-like interface (Hive QL)

Why is using an embedded metastore configuration in Hive not suitable for concurrent usage?
It lacks support for SQL queries
It is not compatible with Hadoop ecosystem
It requires additional authentication steps
None of the others
It can only support one Hive session accessing the same metastore at a time

In normal use, Hive runs on your workstation and converts your SQL query into a series of jobs for execution on a Hadoop cluster. Hive organizes data into tables, which provide a means for attaching structure to data stored in HDFS.
True
False

In schema on read approach, when is the data verified against the schema?
None of the others
Before data is copied or moved
When a query is issued
After data serialization
During data load

What is a major advantage of creating table partitions in Hive?
Simpler query syntax
Less RAM required by namenode
Effective storage memory utilization
Faster query performance
Isolation and security

The tables created in Hive are stored as
a hdfs block containing the database directory
a .java file present in the database directory
a block in main memory
a subdirectory under the database directory
a file under the database directory

Which of the following is NOT true about the SerDe in Hive?

When performing an INSERT, table's SerDe will serialize Hive's internal representation of a row of data into the bytes that are written to the output file.

When querying a table, SerDe will deserialize a row of data from the bytes in the file to objects used internally by Hive to operate on that row of data.

It stands for Serializer and Deserializer

All other choices are true

The file format dictates how rows, and the fields in a particular row, are stored. The file format is defined by a SerDe in Hive.

Which of the following is NOT true about the Hive Shell?

All other options are true

You can execute HiveQL statements in the Hive Shell

You can run Hive on a local mode (for testing purposes on a small sample of your dataset)

Run the 'hive' command to start the Hive shell

Each statement must be terminated with a colon

HiveQL query statement is shown below. The command is to

hive> DESCRIBE FORMATTED trucks;

Show the DDL to recreate the table

Show results after creating the table

None of all choices

Show a list of columns of the table only

Show a list of columns and additional metadata of the table

A HiveQL query statement for loading data from the Linux file system (not from HDFS) is shown below. Which of the following options should be used at {command}?

hive> LOAD DATA LOCAL {command} '/home/Data/trucks.csv'

INTO TABLE trucks;

Loading data to table trucks

OK

Time taken: 0.8 seconds

HDFS

LINUX

INPATH

FROM

LOCAL

Which RDD action combines the elements of an RDD together based on a given function?

combine

reduce

union
collect
take

Which of the following is true about running spark in the client mode on YARN cluster

Spark driver runs on the client machine which submitted the application
It does not require a two-way communication with the client machine
Spark driver runs as part of the application master
The client process can go away after submitting the application
All other choices are false

Running Spark on YARN provides the tightest integration with other Hadoop components and is the most convenient way to use Spark when you have an existing Hadoop cluster. Spark offers two deploy modes for running on YARN: YARN client mode, where the driver runs in the client, and YARN cluster mode, where the driver runs on the cluster in the YARN application master. Which of the following(s) is the command(s) for using YARN client mode? (Check all that apply)

pyspark --mode yarn-client
pyspark --master yarn-client
pyspark --master local
pyspark
spark-submit --master yarn

(Multiple Answers) Which statements are true that are related to the concept "lazy evaluation/operation" in Spark:

When running a transformation in spark-shell, the transformation is evaluated and applied right away and the resulting RDD is computed.
Lazy Evaluation of RDD helps spark recover from failure
RDD lineage is executed only when an action is called on an RDD.
You can call toDebugString method on an RDD to view its lineage and stages

(*Multiple Answer) Which statements are true about Resilient Distributed Datasets (RDD) in Spark?*

They can be created programmatically or from a file
RDD partitions are stored in the distributed memory and spilled to disk if not enough memory is available
An RDDs is mutable and its value can change during their life cycle.
If an RDD partition is lost, it can be reconstructed from the RDD lineage

Which of the following is not true for DataFrame?

We can build DataFrame from different data sources. structured data file, tables in Hive
The Application Programming Interface (APIs) of DataFrame is available in various languages
Both in Scala and Java, we represent DataFrame as Dataset of rows
DataFrame in Apache Spark is behind RDD

Which of the following is NOT true for Datasets and DataFrames?

Dataframe is a dataset organized in to named columns

Dataset is a distributed collection of typed objects, which similar to rdds, are partitioned across multiple nodes in a cluster and can be operated on in parallel

All other answers are true

Dataframes and Datasets are higher level spark data abstractions, allowing query language (SQL and hive) for data manipulation

Although rdds give low level control to spark distributed data, they are not optimized by spark and it is easy to build an inefficient rdd transformation chain

Which of the following is NOT true for SparkSession and SparkContext?

All other answers are true

SparkSession provides a unified channel to access all spark functionality

SparkSession allows working with spark dataframes and datasets

SparkContext encapsulates SparkSession

SparkSession Provides built-in support for Hive features(Hive tables, and HiveQL language)

Which of the following are the common features of RDD and DataFrame?

In-memory

Immutability

Partitioned

Named columns

Resilient

(Multiple answers) Which of the following techniques can be used to compute the distance between two word vectors in NLP?

N-grams

Lemmatization

Direct Distance

Cosine Similarity

Euclidean Distance

Which statement is not true about the bag of word (BOW) representation of a document in a corpus of documents?

All of the choices are true

BOW vector has an index for every word that appears in the corpus

The count of each word in BOW is normalized by the number of documents in which the word appears

BOW is stored as a sparse vector for each document

BOW does not capture ngrams

In Natural Language Processing (NLP), the process of identifying persons, organizations, locations from a given sentence or paragraph is called

Named Entity Recognition

Stemming
Lemmatization
Sentiment Analysis
Part of Speech (PoS) tagging

In text mining, converting text into tokens and then converting them into an integer or floating-point vectors can be done using

CountVectorizer
Probabilistic Topic Models
Vector Space Models
TF-IDF
Bag of Words

Part-of-Speech (PoS) tagging is marking up a word in a text (corpus) as corresponding to a particular part of speech

True
False

Which of the following is not an (input) feature for sequence labeling for NER?

Part-of-speech tags
Current word
Previous and next word
Label context
Target label (class)

Which of the following is not true about the Vector Space (VS) Model?

Documents are projected into a concept space
Distance between the vectors in a concept space is a relationship among documents
The model defines the distance metric
Each concept defines one dimension
Documents are represented by concept vectors

What are the three attributes associated with each concept in Latent Semantic Analysis (LSA)?

Importance score, frequency count, and document length
Affinity for each document, affinity for each term, and an importance score
None of the others (No answer)
Affinity for each word, frequency count, and relevance score
Term frequency-inverse document frequency, cosine similarity, and document frequency

How does LSA contribute to simplifying the representation of data within a corpus?

None of the others (No answer)
By distilling the corpus into relevant concepts
By increasing irrelevant noise
By merging unrelated strands

By adding more complexity to the dataset

Which of the following is NOT true about probabilistic topic models and related concepts in NLP?

Topic is a multinomial distribution over words

Fitting the probabilistic model to text

A document is “generated” by first sampling topics from some prior distribution

All others are true

Answer topic-related questions by computing various kinds of posterior distributions in topic modeling

—----------THE END—----------