Hadoop 2.0 Certification exam for Pig and Hive Developer

Hortonworks Apache-Hadoop-Developer

Total Questions: 10
Version Demo

https://dumpsarena.com
sales@dumpsarena.com

QUESTION NO: 1

You use the hadoop fs –put command to write a 300 MB file using and HDFS block size of 64 MB. Just after this command has finished writing 200 MB of this file, what would another user see when trying to access this life?

A. They would see Hadoop throw an ConcurrentFileAccessException when they try to access this file.
B. They would see the current state of the file, up to the last bit written by the command.
C. They would see the current of the file through the last completed block.
D. They would see no content until the whole file written and closed.

Answer: C

QUESTION NO: 2

Which TWO of the following statements are true regarding Hive? Choose 2 answers

A. Useful for data analysts familiar with SQL who need to do ad-hoc queries
B. Offers real-time queries and row level updates
C. Allows you to define a structure for your unstructured Big Data
D. Is a relational database

Answer: A, C

QUESTION NO: 3

Given the following Hive command:

```
CREATE EXTERNAL TABLE mytable (name string, age int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/home/user/mydata/';
```

Which one of the following statements is true?

A. The files in the mydata folder are copied to a subfolder of /apps/hlve/warehouse
B. The files in the mydata folder are moved to a subfolder of /apps/hive/wa re house
C. The files in the mydata folder are copied into Hive's underlying relational database
D. The files in the mydata folder do not move from their current location In HDFS

Answer: D

QUESTION NO: 4

Which project gives you a distributed, Scalable, data store that allows you random, realtime read/write access to hundreds of terabytes of data?

A. HBase
B. Hue
C. Pig
D. Hive
E. Oozie
F. Flume
G. Sqoop

Answer: A
Use Apache HBase when you need random, realtime read/write access to your Big Data.

Note: This project's goal is the hosting of very large tables -- billions of rows X millions of columns -- atop clusters of commodity hardware. Apache HBase is an open-source, distributed, versioned, column-oriented store modeled after Google's Bigtable: A Distributed Storage System for Structured Data by Chang et al. Just as Bigtable leverages the distributed data storage provided by the Google File System, Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.

Features

Linear and modular scalability.

Strictly consistent reads and writes.

Automatic and configurable sharding of tables

Automatic failover support between RegionServers.

Convenient base classes for backing Hadoop MapReduce jobs with Apache HBase tables.

Easy to use Java API for client access.

Block cache and Bloom Filters for real-time queries.

Query predicate push down via server side Filters

Thrift gateway and a REST-ful Web service that supports XML, Protobuf, and binary data encoding options

Extensible jruby-based (JIRB) shell

Support for exporting metrics via the Hadoop metrics subsystem to files or Ganglia; or via JMX

Reference: http://hbase.apache.org/ (when would I use HBase? First sentence)

QUESTION NO: 5

Consider the following two relations, A and B.

```
A = LOAD 'data1' AS (a1:int,a2:chararray);
DUMP A;
(1,apple)
(3,orange)
(4,peach)
(2,cherry)

B = LOAD 'data2' AS (b1:chararray,b2:int);
DUMP B;
(Jim,2)
(Brian,4)
(Kim,0)
(Terry,3)
(Chris,2)
```

A Pig JOIN statement that combined relations A by its first field and B by its second field would produce what output?

A. 2 Jim Chris 2
3 Terry 3
4 Brian 4

B. 2 cherry
2 cherry
3 orange
4 peach

C. 2 cherry Jim, Chris
3 orange Terry
4 peach Brian

D. 2 cherry Jim 2
2 cherry Chris 2
3 orange Terry 3
4 peach Brian 4

Answer: D

QUESTION NO: 6

Which one of the following statements is true about a Hive-managed table?

A. Records can only be added to the table using the Hive INSERT command.
B. When the table is dropped, the underlying folder in HDFS is deleted.
C. Hive dynamically defines the schema of the table based on the FROM clause of a SELECT query.
D. Hive dynamically defines the schema of the table based on the format of the underlying data.

Answer: B

QUESTION NO: 7

Which process describes the lifecycle of a Mapper?

A. The JobTracker calls the TaskTracker's configure () method, then its map () method and finally its close () method.
B. The TaskTracker spawns a new Mapper to process all records in a single input split.
C. The TaskTracker spawns a new Mapper to process each key-value pair.
D. The JobTracker spawns a new Mapper to process all records in a single file.

Answer: B
For each map instance that runs, the TaskTracker creates a new instance of your mapper.

Note:

* The Mapper is responsible for processing Key/Value pairs obtained from the InputFormat. The mapper may perform a number of Extraction and Transformation functions on the Key/Value pair before ultimately outputting none, one or many Key/Value pairs of the same, or different Key/Value type.

* With the new Hadoop API, mappers extend the org.apache.hadoop.mapreduce.Mapper class. This class defines an 'Identity' map function by default - every input Key/Value pair obtained from the InputFormat is written out.

Examining the run() method, we can see the lifecycle of the mapper:

```
/**
 * Expert users can override this method for more complete control over the
 * execution of the Mapper.
 * @param context
 * @throws IOException
 */
public void run(Context context) throws IOException, InterruptedException {
  setup(context);
  while (context.nextKeyValue()) {
    map(context.getCurrentKey(), context.getCurrentValue(), context);
  }
  cleanup(context);
}
```

setup(Context) - Perform any setup for the mapper. The default implementation is a no-op method.

map(Key, Value, Context) - Perform a map operation in the given Key / Value pair. The default implementation calls Context.write(Key, Value)

cleanup(Context) - Perform any cleanup for the mapper. The default implementation is a no-op method.

Reference: Hadoop/MapReduce/Mapper

QUESTION NO: 8

Which one of the following is NOT a valid Oozie action?

A. mapreduce
B. pig
C. hive
D. mrunit

Answer: D

QUESTION NO: 9

Which one of the following statements describes a Hive user-defined aggregate function?

A. Operates on multiple input rows and creates a single row as output
B. Operates on a single input row and produces a single row as output
C. Operates on a single input row and produces a table as output
D. Operates on multiple input rows and produces a table as output

Answer: A

QUESTION NO: 10

You want to perform analysis on a large collection of images. You want to store this data in HDFS and process it with MapReduce but you also want to give your data analysts and data scientists the ability to process the data directly from HDFS with an interpreted high-level programming language like Python. Which format should you use to store this data in HDFS?

A. SequenceFiles
B. Avro
C. JSON
D. HTML
E. XML
F. CSV

Answer: B

Reference: Hadoop binary files processing introduced by image duplicates finder