HDPCD Q&As

Hortonworks Data Platform Certified Developer

Pass Hortonworks HDPCD Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.leads4pass.com/hdpcd.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Hortonworks Official Exam Center

QUESTION 1

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

Correct Answer: B

Reference: Hadoop binary files processing introduced by image duplicates finder

QUESTION 2

Which best describes what the map method accepts and emits?

A. It accepts a single key-value pair as input and emits a single key and list of corresponding values as output.

B. It accepts a single key-value pairs as input and can emit only one key-value pair as output.

C. It accepts a list key-value pairs as input and can emit only one key-value pair as output.

D. It accepts a single key-value pairs as input and can emit any number of key-value pair as output, including zero.

Correct Answer: D

Explanation: public class Mapper extends Object Maps input key/value pairs to a set of intermediate key/value pairs.

Maps are the individual tasks which transform input records into a intermediate records. The transformed intermediate records need not be of the same type as the input records. A given input pair may map to zero or many output pairs.

Reference: org.apache.hadoop.mapreduce

Class Mapper

QUESTION 3

Which one of the following Hive commands uses an HCatalog table named x?

A. SELECT * FROM x;

B. SELECT x.-FROM org.apache.hcatalog.hive.HCatLoader(\\'x\\');

C. SELECT * FROM org.apache.hcatalog.hive.HCatLoader(\\'x\\');

D. Hive commands cannot reference an HCatalog table

Correct Answer: C

QUESTION 4

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

Correct Answer: D

QUESTION 5

Which HDFS command copies an HDFS file named foo to the local filesystem as localFoo?

A. hadoop fs -get foo LocalFoo

B. hadoop -cp foo LocalFoo

C. hadoop fs -Is foo

D. hadoop fs -put foo LocalFoo

Correct Answer: A

QUESTION 6

Workflows expressed in Oozie can contain:

A. Sequences of MapReduce and Pig. These sequences can be combined with other actions including forks, decision points, and path joins.

B. Sequences of MapReduce job only; on Pig on Hive tasks or jobs. These MapReduce sequences can be combined with forks and path joins.

C. Sequences of MapReduce and Pig jobs. These are limited to linear sequences of actions with exception handlers but no forks.

D. Iterntive repetition of MapReduce jobs until a desired answer or state is reached.

Correct Answer: A

Explanation: Oozie workflow is a collection of actions (i.e. Hadoop Map/Reduce jobs, Pig jobs) arranged in a control dependency DAG (Direct Acyclic Graph), specifying a sequence of actions execution. This graph is specified in hPDL (a XML Process Definition Language).

hPDL is a fairly compact language, using a limited amount of flow control and action nodes. Control nodes define the flow of execution and include beginning and end of a workflow (start, end and fail nodes) and mechanisms to control the workflow execution path (decision, fork and join nodes).

Workflow definitions Currently running workflow instances, including instance states and variables

Reference: Introduction to Oozie

Note: Oozie is a Java Web-Application that runs in a Java servlet-container - Tomcat and uses a database to store:

QUESTION 7

You want to run Hadoop jobs on your development workstation for testing before you submit them to your production cluster. Which mode of operation in Hadoop allows you to most closely simulate a production cluster while using a single machine?

A. Run all the nodes in your production cluster as virtual machines on your development workstation.

B. Run the hadoop command with the -jt local and the -fs file:///options.

C. Run the DataNode, TaskTracker, NameNode and JobTracker daemons on a single machine.

D. Run simldooop, the Apache open-source software for simulating Hadoop clusters.

Correct Answer: C

QUESTION 8

Assuming default settings, which best describes the order of data provided to a reducer\\'s reduce method:

A. The keys given to a reducer aren\\'t in a predictable order, but the values associated with those keys always are.

B. Both the keys and values passed to a reducer always appear in sorted order.

C. Neither keys nor values are in any predictable order.

D. The keys given to a reducer are in sorted order but the values associated with each key are in no predictable order

Correct Answer: D

Explanation: Reducer has 3 primary phases:

1.

Shuffle

The Reducer copies the sorted output from each Mapper using HTTP across the network.

2.

Sort

The framework merge sorts Reducer inputs by keys (since different Mappers may have output the same key).

The shuffle and sort phases occur simultaneously i.e. while outputs are being fetched they are merged.

SecondarySort

To achieve a secondary sort on the values returned by the value iterator, the application should extend the key with the secondary key and define a grouping comparator. The keys will be sorted using the entire key, but will be grouped using the grouping comparator to decide which keys and values are sent in the same call to reduce.

3. Reduce

In this phase the reduce(Object, Iterable, Context) method is called for each in

the sorted inputs.

The output of the reduce task is typically written to a RecordWriter via TaskInputOutputContext.write

(Object, Object).

The output of the Reducer is not re-sorted.

Reference: org.apache.hadoop.mapreduce, Class Reducer

QUESTION 9

Assuming the following Hive query executes successfully:

```
from inputdata select context_ngrams(sentences(lines),
array("you", "are", null), 80);
```

Which one of the following statements describes the result set?

A. A bigram of the top 80 sentences that contain the substring "you are" in the lines column of the input data A1 table.

B. An 80-value ngram of sentences that contain the words "you" or "are" in the lines column of the inputdata table.

C. A trigram of the top 80 sentences that contain "you are" followed by a null space in the lines column of the inputdata table.

D. A frequency distribution of the top 80 words that follow the subsequence "you are" in the lines column of the inputdata table.

Correct Answer: D

QUESTION 10

Table metadata in Hive is:

A. Stored as metadata on the NameNode.

B. Stored along with the data in HDFS.

C. Stored in the Metastore.

D. Stored in ZooKeeper.

Correct Answer: C

Explanation: By default, hive use an embedded Derby database to store metadata information. The metastore is the "glue" between Hive and HDFS. It tells Hive where your data files live in HDFS, what type of data they contain, what tables they belong to, etc.

The Metastore is an application that runs on an RDBMS and uses an open source ORM layer called DataNucleus, to convert object representations into a relational schema and vice versa. They chose this approach as opposed to storing this information in hdfs as they need the Metastore to be very low latency. The DataNucleus layer allows them to plugin many different RDBMS technologies.

Note:

*

By default, Hive stores metadata in an embedded Apache Derby database, and other client/server

databases like MySQL can optionally be used.

*

features of Hive include:

Metadata storage in an RDBMS, significantly reducing the time to perform semantic checks during query

execution.

Reference: Store Hive Metadata into RDBMS

QUESTION 11

Which two of the following are true about this trivial Pig program\\' (choose Two)

```
$ pig
grunt> ABC = LOAD 'myfile';
grunt> DUMP ABC;
```

A. The contents of myfile appear on stdout

B. Pig assumes the contents of myfile are comma delimited

C. ABC has a schema associated with it

D. myfile is read from the user\\'s home directory in HDFS

Correct Answer: AD

QUESTION 12

Given the following Hive command:

INSERT OVERWRITE TABLE mytable SELECT * FROM myothertable;

Which one of the following statements is true?

A. The contents of myothertable are appended to mytable

B. Any existing data in mytable will be overwritten

C. A new table named mytable is created, and the contents of myothertable are copied into mytable

D. The statement is not a valid Hive command

Correct Answer: B

QUESTION 13

You are developing a MapReduce job for sales reporting. The mapper will process input keys representing the year (IntWritable) and input values representing product indentifies (Text).

Indentify what determines the data types used by the Mapper for a given job.

A. The key and value types specified in the JobConf.setMapInputKeyClass and JobConf.setMapInputValuesClass methods

B. The data types specified in HADOOP_MAP_DATATYPES environment variable

C. The mapper-specification.xml file submitted with the job determine the mapper\\'s input key and value types.

D. The InputFormat used by the job determines the mapper\\'s input key and value types.

Correct Answer: D

Explanation: The input types fed to the mapper are controlled by the InputFormat used. The default input

format, "TextInputFormat," will load data in as (LongWritable, Text) pairs. The long value is the byte offset

of the line in the file. The Text object holds the string contents of the line of the file.

Note: The data types emitted by the reducer are identified by setOutputKeyClass()

andsetOutputValueClass(). The data types emitted by the reducer are identified by setOutputKeyClass()

and setOutputValueClass().

By default, it is assumed that these are the output types of the mapper as well. If this is not the case, the

methods setMapOutputKeyClass() and setMapOutputValueClass() methods of the JobConf class will

override these.

Reference: Yahoo! Hadoop Tutorial, THE DRIVER METHOD

QUESTION 14

Given the following Hive commands:

```
CREATE TABLE mytable (name chararray, age int) ROW FORMAT
DELIMITED FIELDS TERMINATED BY ',' STORED AS
TEXTFILE;
LOAD DATA INPATH '/home/user/mydata.txt' INTO TABLE mytable;
```

Which one of the following statements Is true?

A. The file mydata.txt is copied to a subfolder of /apps/hive/warehouse

B. The file mydata.txt is moved to a subfolder of /apps/hive/warehouse

C. The file mydata.txt is copied into Hive\\'s underlying relational database 0.

D. The file mydata.txt does not move from Its current location in HDFS

Correct Answer: A

QUESTION 15

You need to perform statistical analysis in your MapReduce job and would like to call methods in the Apache Commons Math library, which is distributed as a 1.3 megabyte Java archive (JAR) file. Which is the best way to make this library available to your MapReducer job at runtime?

A. Have your system administrator copy the JAR to all nodes in the cluster and set its location in the HADOOP_CLASSPATH environment variable before you submit your job.

B. Have your system administrator place the JAR file on a Web server accessible to all cluster nodes and then set the HTTP_JAR_URL environment variable to its location.

C. When submitting the job on the command line, specify the ?ibjars option followed by the JAR file path.

D. Package your code and the Apache Commands Math library into a zip file named JobJar.zip

Correct Answer: C

Explanation: The usage of the jar command is like this,

Usage: hadoop jar [mainClass] args...

If you want the commons-math3.jar to be available for all the tasks you can do any one of these

1. Copy the jar file in $HADOOP_HOME/lib dir

2.

or

Use the generic option -libjars.