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.passapply.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

In Hadoop 2.2, which one of the following statements is true about a standby NameNode?

The Standby NameNode:

A. Communicates directly with the active NameNode to maintain the state of the active NameNode.

B. Receives the same block reports as the active NameNode.

C. Runs on the same machine and shares the memory of the active NameNode.

D. Processes all client requests and block reports from the appropriate DataNodes.

Correct Answer: B

QUESTION 2

What are the TWO main components of the YARN ResourceManager process? Choose 2 answers

A. Job Tracker

B. Task Tracker

C. Scheduler

D. Applications Manager

Correct Answer: CD

QUESTION 3

Indentify which best defines a SequenceFile?

A. A SequenceFile contains a binary encoding of an arbitrary number of homogeneous Writable objects

B. A SequenceFile contains a binary encoding of an arbitrary number of heterogeneous Writable objects

C. A SequenceFile contains a binary encoding of an arbitrary number of WritableComparable objects, in sorted order.

D. A SequenceFile contains a binary encoding of an arbitrary number key-value pairs. Each key must be the same type. Each value must be the same type.

Correct Answer: D

Explanation: SequenceFile is a flat file consisting of binary key/value pairs.

There are 3 different SequenceFile formats:

Uncompressed key/value records.

Record compressed key/value records - only \\'values\\' are compressed here. Block compressed key/value

records - both keys and values are collected in \\'blocks\\' separately and compressed. The size of the \\'block\\'

is configurable.

Reference: http://wiki.apache.org/hadoop/SequenceFile

QUESTION 4

The Hadoop framework provides a mechanism for coping with machine issues such as faulty configuration or impending hardware failure. MapReduce detects that one or a number of machines are performing poorly and starts more copies of a map or reduce task. All the tasks run simultaneously and the task finish first are used. This is called:

A. Combine

B. IdentityMapper

C. IdentityReducer

D. Default Partitioner

E. Speculative Execution

Correct Answer: E

Explanation: Speculative execution: One problem with the Hadoop system is that by dividing the tasks across many nodes, it is possible for a few slow nodes to rate-limit the rest of the program. For example if one node has a slow disk controller, then it may be reading its input at only 10% the speed of all the other nodes. So when 99 map tasks are already complete, the system is still waiting for the final map task to check in, which takes much longer than all the other nodes. By forcing tasks to run in isolation from one another, individual tasks do not know where their inputs come from. Tasks trust the Hadoop platform to just deliver the appropriate input. Therefore, the same input can be processed multiple times in parallel, to exploit differences in machine capabilities. As most of the tasks in a job are coming to a close, the Hadoop platform will schedule redundant copies of the remaining tasks across several nodes which do not have other work to perform. This process is known as speculative execution. When tasks complete, they announce this fact to the JobTracker. Whichever copy of a task finishes first becomes the definitive copy. If other copies were executing speculatively, Hadoop tells the TaskTrackers to abandon the tasks and discard their outputs. The Reducers then receive their inputs from whichever Mapper completed successfully, first.

Reference: Apache Hadoop, Module 4: MapReduce

Note:

*

Hadoop uses "speculative execution." The same task may be started on multiple boxes. The first one to finish wins, and the other copies are killed.

Failed tasks are tasks that error out.

*

There are a few reasons Hadoop can kill tasks by his own decisions:

a) Task does not report progress during timeout (default is 10 minutes)

b) FairScheduler or CapacityScheduler needs the slot for some other pool (FairScheduler) or queue (CapacityScheduler).

c) Speculative execution causes results of task not to be needed since it has completed on other place.

Reference: Difference failed tasks vs killed tasks

QUESTION 5

You have a directory named jobdata in HDFS that contains four files: _first.txt, second.txt, .third.txt and #data.txt. How many files will be processed by the FileInputFormat.setInputPaths () command when it\\'s given a path object representing this directory?

A. Four, all files will be processed

B. Three, the pound sign is an invalid character for HDFS file names

C. Two, file names with a leading period or underscore are ignored

D. None, the directory cannot be named jobdata

E. One, no special characters can prefix the name of an input file

Correct Answer: C

Explanation: Files starting with \\'_\\' are considered \\'hidden\\' like unix files starting with \\'.\\'. # characters are allowed in HDFS file names.

QUESTION 6

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 7

You have just executed a MapReduce job. Where is intermediate data written to after being emitted from the Mapper\\'s

map method?

A. Intermediate data in streamed across the network from Mapper to the Reduce and is never written to disk.

B. Into in-memory buffers on the TaskTracker node running the Mapper that spill over and are written into HDFS.

C. Into in-memory buffers that spill over to the local file system of the TaskTracker node running the Mapper.

D. Into in-memory buffers that spill over to the local file system (outside HDFS) of the TaskTracker node running the Reducer

E. Into in-memory buffers on the TaskTracker node running the Reducer that spill over and are written into HDFS.

Correct Answer: C

Explanation: The mapper output (intermediate data) is stored on the Local file system (NOT HDFS) of each individual mapper nodes. This is typically a temporary directory location which can be setup in config by the hadoop administrator. The intermediate data is cleaned up after the Hadoop Job completes.

Reference: 24 Interview Questions and Answers for Hadoop MapReduce developers, Where is the Mapper Output (intermediate kay-value data) stored ?

QUESTION 8

You need to run the same job many times with minor variations. Rather than hardcoding all job configuration options in your drive code, you\\'ve decided to have your Driver subclass org.apache.hadoop.conf.Configured and implement the org.apache.hadoop.util.Tool interface.

Indentify which invocation correctly passes.mapred.job.name with a value of Example to Hadoop?

A. hadoop "mapred.job.name=Example" MyDriver input output

B. hadoop MyDriver mapred.job.name=Example input output

C. hadoop MyDrive -D mapred.job.name=Example input output

D. hadoop setproperty mapred.job.name=Example MyDriver input output

E. hadoop setproperty ("mapred.job.name=Example") MyDriver input output

Correct Answer: C

Explanation: Configure the property using the -D key=value notation:

-D mapred.job.name=\\'My Job\\' You can list a whole bunch of options by calling the streaming jar with just the -info argument Reference: Python hadoop streaming : Setting a job name

---

QUESTION 9

In a large MapReduce job with m mappers and n reducers, how many distinct copy operations will there be in the sort/shuffle phase?

A. mXn (i.e., m multiplied by n)

B. n

C. m

D. m+n (i.e., m plus n) E. mn (i.e., m to the power of n)

Correct Answer: A

Explanation: A MapReduce job with m mappers and r reducers involves up to m * r distinct copy operations, since each mapper may have intermediate output going to every reducer.

QUESTION 10

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 11

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.

Correct Answer: B

Explanation: 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 12

A NameNode in Hadoop 2.2 manages _____.

A. Two namespaces: an active namespace and a backup namespace

B. A single namespace

C. An arbitrary number of namespaces

D. No namespaces

Correct Answer: B

QUESTION 13

Can you use MapReduce to perform a relational join on two large tables sharing a key? Assume that the two tables are formatted as comma-separated files in HDFS.

A. Yes.

B. Yes, but only if one of the tables fits into memory

C. Yes, so long as both tables fit into memory.

D. No, MapReduce cannot perform relational operations.

E. No, but it can be done with either Pig or Hive.

Correct Answer: A

Explanation: Note:

*

Join Algorithms in MapReduce A) Reduce-side join B) Map-side join C) In-memory join / Striped Striped variant variant / Memcached variant

*

Which join to use? / In-memory join > map-side join > reduce-side join / Limitations of each? In-memory join: memory Map-side join: sort order and partitioning Reduce-side join: general purpose

QUESTION 14

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 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.