APACHE-HADOOP-DEVELOPER^{Q&As}

Hadoop 2.0 Certification exam for Pig and Hive Developer

Pass Hortonworks APACHE-HADOOP-DEVELOPER Exam with 100% Guarantee

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

https://www.passapply.com/apache-hadoop-developer.html

100% Passing Guarantee
100% Money Back Assurance

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

- ✪ **Instant Download** After Purchase
- ✪ **100% Money Back** Guarantee
- ✪ **365 Days** Free Update
- ✪ **800,000+** Satisfied Customers

QUESTION 1

Which Hadoop component is responsible for managing the distributed file system metadata?

A. NameNode

B. Metanode

C. DataNode

D. NameSpaceManager

Correct Answer: A

QUESTION 2

You want to understand more about how users browse your public website, such as which pages they visit prior to placing an order. You have a farm of 200 web servers hosting your website. How will you gather this data for your analysis?

A. Ingest the server web logs into HDFS using Flume.

B. Write a MapReduce job, with the web servers for mappers, and the Hadoop cluster nodes for reduces.

C. Import all users\\' clicks from your OLTP databases into Hadoop, using Sqoop.

D. Channel these clickstreams inot Hadoop using Hadoop Streaming.

E. Sample the weblogs from the web servers, copying them into Hadoop using curl.

Correct Answer: A

QUESTION 3

You wrote a map function that throws a runtime exception when it encounters a control character in input data. The input supplied to your mapper contains twelve such characters totals, spread across five file splits. The first four file splits each have two control characters and the last split has four control characters.

Indentify the number of failed task attempts you can expect when you run the job with mapred.max.map.attempts set to 4:

A. You will have forty-eight failed task attempts

B. You will have seventeen failed task attempts

C. You will have five failed task attempts

D. You will have twelve failed task attempts

E. You will have twenty failed task attempts

Correct Answer: E

Explanation: There will be four failed task attempts for each of the five file splits.

Note:

When the jobtracker is notified of a task attempt that has failed (by the tasktracker's heartbeat call), it will reschedule execution of the task. The jobtracker will try to avoid rescheduling the task on a tasktracker where it has previously failed. Furthermore, if a task fails four times (or more), it will not be retried further. This value is configurable: the maximum number of attempts to run a task is controlled by the `mapred.map.max.attempts` property for map tasks and `mapred.reduce.max.attempts` for reduce tasks. By default, if any task fails four times (or whatever the maximum number of attempts is configured to), the whole job fails.

QUESTION 4

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 5

Review the following andapos;dataandapos; file and Pig code.

```
M,38,95111
M,62,95102


A = LOAD 'data' USING PigStorage(',')
AS (gender:chararray, age:int, zip:chararray);
D = GROUP A BY gender;
DUMP D;
```

Which one of the following statements is true?

A. The Output Of the DUMP D command IS (M,{(M,62.95102),(M,38,95111)})

B. The output of the dump d command is (M, {(38,95in),(62,95i02)})

C. The code executes successfully but there is not output because the D relation is empty

D. The code does not execute successfully because D is not a valid relation

Correct Answer: A

QUESTION 6

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 7

In Hadoop 2.0, which TWO of the following processes work together to provide automatic failover of the NameNode?
Choose 2 answers

A. ZKFailoverController

B. ZooKeeper

C. QuorumManager

D. JournalNode

Correct Answer: AD

QUESTION 8

You are developing a combiner that takes as input Text keys, IntWritable values, and emits Text keys, IntWritable
values. Which interface should your class implement?

A. Combiner

B. Mapper

C. Reducer

D. Reducer

E. Combiner

Correct Answer: D

QUESTION 9

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 ? 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 10

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 11

MapReduce v2 (MRv2/YARN) splits which major functions of the JobTracker into separate daemons? Select two.

A. Heath states checks (heartbeats)

B. Resource management

C. Job scheduling/monitoring

D. Job coordination between the ResourceManager and NodeManager

E. Launching tasks

F. Managing file system metadata

G. MapReduce metric reporting H. Managing tasks

Correct Answer: BC

Explanation: The fundamental idea of MRv2 is to split up the two major functionalities of the JobTracker,

resource management and job scheduling/monitoring, into separate daemons. The idea is to have a global

ResourceManager (RM) and per-application ApplicationMaster (AM). An application is either a single job in

the classical sense of Map- Reduce jobs or a DAG of jobs.

Note:

The central goal of YARN is to clearly separate two things that are unfortunately smushed together in

current Hadoop, specifically in (mainly) JobTracker:

/ Monitoring the status of the cluster with respect to which nodes have which resources available. Under

YARN, this will be global.

/ Managing the parallelization execution of any specific job. Under YARN, this will be done separately for

each job.

Reference: Apache Hadoop YARN ?Conceptsand; Applications

QUESTION 12

You write MapReduce job to process 100 files in HDFS. Your MapReduce algorithm uses TextInputFormat: the mapper applies a regular expression over input values and emits key- values pairs with the key consisting of the matching text, and the value containing the filename and byte offset. Determine the difference between setting the number of reduces to one and settings the number of reducers to zero.

A. There is no difference in output between the two settings.

B. With zero reducers, no reducer runs and the job throws an exception. With one reducer, instances of matching patterns are stored in a single file on HDFS.

C. With zero reducers, all instances of matching patterns are gathered together in one file on HDFS. With one reducer, instances of matching patterns are stored in multiple files on HDFS.

D. With zero reducers, instances of matching patterns are stored in multiple files on HDFS. With one reducer, all instances of matching patterns are gathered together in one file on HDFS.

Correct Answer: D

Explanation: * It is legal to set the number of reduce-tasks to zero if no reduction is desired.

In this case the outputs of the map-tasks go directly to the FileSystem, into the output path set by setOutputPath(Path). The framework does not sort the map-outputs before writing them out to the FileSystem.

* Often, you may want to process input data using a map function only. To do this, simply set mapreduce.job.reduces to zero. The MapReduce framework will not create any reducer tasks. Rather, the outputs of the mapper tasks will be the final output of the job.

Note:

Reduce

In this phase the reduce(WritableComparable, Iterator, OutputCollector, Reporter) method is called for

each pair in the grouped inputs.

The output of the reduce task is typically written to the FileSystem via OutputCollector.collect

(WritableComparable, Writable).

Applications can use the Reporter to report progress, set application-level status messages and update

Counters, or just indicate that they are alive.

The output of the Reducer is not sorted.

QUESTION 13

You\'ve written a MapReduce job that will process 500 million input records and generated 500 million key-value pairs. The data is not uniformly distributed. Your MapReduce job will create a significant amount of intermediate data that it needs to transfer between mappers and reduces which is a potential bottleneck. A custom implementation of which interface is most likely to reduce the amount of intermediate data transferred across the network?

A. Partitioner

B. OutputFormat

C. WritableComparable

D. Writable

E. InputFormat

F. Combiner

Correct Answer: F

Explanation: Combiners are used to increase the efficiency of a MapReduce program. They are used to aggregate intermediate map output locally on individual mapper outputs. Combiners can help you reduce the amount of data that needs to be transferred across to the reducers. You can use your reducer code as a combiner if the operation performed is commutative and associative.

Reference: 24 Interview Questions and Answers for Hadoop MapReduce developers, What are combiners? When should I use a combiner in my MapReduce Job?

QUESTION 14

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 15

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.

*

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

Failed tasks are tasks that error out.

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

APACHE-HADOOP-
DEVELOPER Study Guide

APACHE-HADOOP-
DEVELOPER Exam
Questions

APACHE-HADOOP-
DEVELOPER Braindumps