HADOOP-PR000007 Q&As

Hortonworks Certified Apache Hadoop 2.0 Developer (Pig and Hive Developer)

Pass Hortonworks HADOOP-PR000007 Exam with 100% Guarantee

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

https://www.pass2lead.com/hadoop-pr000007.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 Ingest log files Into HDFS, which tool would you use?

A. HCatalog

B. Flume

C. Sqoop

D. Ambari

Correct Answer: B

QUESTION 2

Your cluster\\'s HDFS block size in 64MB. You have directory containing 100 plain text files, each of which is 100MB in size. The InputFormat for your job is TextInputFormat.

Determine how many Mappers will run?

A. 64

B. 100

C. 200

D. 640

Correct Answer: C

Explanation: Each file would be split into two as the block size (64 MB) is less than the file size (100 MB), so 200 mappers would be running.

Note:

If you\\'re not compressing the files then hadoop will process your large files (say 10G), with a number of mappers related to the block size of the file.

Say your block size is 64M, then you will have ~160 mappers processing this 10G file (160*64 ~= 10G).

Depending on how CPU intensive your mapper logic is, this might be an acceptable blocks size, but if you find that your mappers are executing in sub minute times, then you might want to increase the work done by each mapper (by increasing the block size to 128, 256, 512m - the actual size depends on how you intend to process the data). Reference: http://stackoverflow.com/

QUESTION 3

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 4

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 5

Which of the following tool was designed to import data from a relational database into HDFS?

A. HCatalog B. Sqoop

C. Flume

D. Ambari

Correct Answer: B

QUESTION 6

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 7

Identify the tool best suited to import a portion of a relational database every day as files into HDFS, and generate Java classes to interact with that imported data?

A. Oozie

B. Flume

C. Pig

D. Hue

E. Hive

F. Sqoop

G. fuse-dfs

Correct Answer: F

Sqoop ("SQL-to-Hadoop") is a straightforward command-line tool with the following capabilities:

Imports individual tables or entire databases to files in HDFS Generates Java classes to allow you to

interact with your imported data Provides the ability to import from SQL databases straight into your Hive

data warehouse

Note:

Data Movement Between Hadoop and Relational Databases Data can be moved between Hadoop and a

relational database as a bulk data transfer, or relational tables can be accessed from within a MapReduce

map function.

Note:

* Cloudera\\'s Distribution for Hadoop provides a bulk data transfer tool (i.e., Sqoop) that imports individual tables or entire databases into HDFS files. The tool also generates Java classes that support interaction with the imported data. Sqoop supports all relational databases over JDBC, and Quest Software provides a connector (i.e., OraOop) that has been optimized for access to data residing in Oracle databases.

Reference: http://log.medcl.net/item/2011/08/hadoop-and-mapreduce-big-data-analytics- gartner/ (Data Movement between hadoop and relational databases, second paragraph)

QUESTION 8

What data does a Reducer reduce method process?

A. All the data in a single input file.

B. All data produced by a single mapper.

C. All data for a given key, regardless of which mapper(s) produced it.

D. All data for a given value, regardless of which mapper(s) produced it.

Correct Answer: C

Explanation: Reducing lets you aggregate values together. A reducer function receives an iterator of input values from an input list. It then combines these values together, returning a single output value.

All values with the same key are presented to a single reduce task.

Reference: Yahoo! Hadoop Tutorial, Module 4: MapReduce

QUESTION 9

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

Correct Answer: A

Explanation: 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 10

Given the following Pig command:

logevents = LOAD andapos;input/my.logandapos; AS (date:chararray, levehstring, code:int, message:string);

Which one of the following statements is true?

A. The logevents relation represents the data from the my.log file, using a comma as the parsing delimiter

B. The logevents relation represents the data from the my.log file, using a tab as the parsing delimiter

C. The first field of logevents must be a properly-formatted date string or table return an error

D. The statement is not a valid Pig command

Correct Answer: B

QUESTION 11

Review the following data and Pig code:

```
M,38,95111
F,29,95060
F,45,95192
M,62,95102
F,56,95102

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

What command to define B would produce the output (M,62,95l02) when invoking the DUMP operator on B?

A. B = FILTER A BY (zip = = \\'95102\\' AND gender = = M");

B. B= FOREACH A BY (gender = = \\'M\\' AND zip = = \\'95102\\');

C. B = JOIN A BY (gender = = \\'M\\' AND zip = = \\'95102\\');

D. B= GROUP A BY (zip = = \\'95102\\' AND gender = = \\'M\\');

Correct Answer: A

QUESTION 12

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 13

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 14

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 15

Given the following Pig commands: Which one of the following statements is true?

```
logevents = LOAD 'input/my.log';
severe = FILTER logevents BY ($1 =='severe' AND $2 >= 500);
grouped = GROUP severe BY $2;
DUMP grouped;
```

A. The $1 variable represents the first column of data in \\'my.log\\'

B. The $1 variable represents the second column of data in \\'my.log\\'

C. The severe relation is not valid

D. The grouped relation is not valid

Correct Answer: B