Vendor: Hortonworks

Exam Code:APACHE-HADOOP-DEVELOPER

Exam Name: Hadoop 2.0 Certification exam for Pig and Hive Developer

Version: Demo

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 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 4

You want to count the number of occurrences for each unique word in the supplied input data. You\\'ve decided to implement this by having your mapper tokenize each word and emit a literal value 1, and then have your reducer increment a counter for each literal 1 it receives. After successful implementing this, it occurs to you that you could optimize this by specifying a combiner. Will you be able to reuse your existing Reduces as your combiner in this case and why or why not?

A. Yes, because the sum operation is both associative and commutative and the input and output types to the reduce method match.

B. No, because the sum operation in the reducer is incompatible with the operation of a Combiner.

C. No, because the Reducer and Combiner are separate interfaces.

D. No, because the Combiner is incompatible with a mapper which doesn\\'t use the same data type for both the key and value.

E. Yes, because Java is a polymorphic object-oriented language and thus reducer code can be reused as a combiner.

Correct Answer: A

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. The execution of combiner is not guaranteed, Hadoop may or may not execute a combiner. Also, if required it may execute it more then 1 times. Therefore your MapReduce jobs should not depend on the combiners execution.

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 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 6

In Hadoop 2.0, 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 7

What does the following command do? register andapos;/piggyban):/pig-files.jarandapos;;

A. Invokes the user-defined functions contained in the jar file

B. Assigns a name to a user-defined function or streaming command

C. Transforms Pig user-defined functions into a format that Hive can accept

D. Specifies the location of the JAR file containing the user-defined functions

Correct Answer: D

QUESTION 8

Review the following data and Pig code.

M,38,95111

F,29,95060

F,45,95192

M,62,95102

F,56,95102

A = LOAD andapos;dataandapos; USING PigStorage(andapos;.andapos;) as (gender:Chararray, age:int,

zlp:chararray);

B = FOREACH A GENERATE age;

Which one of the following commands would save the results of B to a folder in hdfs named myoutput?

A. STORE A INTO andapos;myoutputandapos; USING PigStorage(andapos;,andapos;);

B. DUMP B using PigStorage(andapos;myoutputandapos;);

C. STORE B INTO andapos;myoutputandapos;;

D. DUMP B INTO andapos;myoutputandapos;;

Correct Answer: C

QUESTION 9

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 10

How are keys and values presented and passed to the reducers during a standard sort and shuffle phase of MapReduce?

A. Keys are presented to reducer in sorted order; values for a given key are not sorted.

B. Keys are presented to reducer in sorted order; values for a given key are sorted in ascending order.

C. Keys are presented to a reducer in random order; values for a given key are not sorted.

D. Keys are presented to a reducer in random order; values for a given key are sorted in ascending order.

Correct Answer: A

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 11

You need to create a job that does frequency analysis on input data. You will do this by writing a Mapper that uses TextInputFormat and splits each value (a line of text from an input file) into individual characters. For each one of these characters, you will emit the character as a key and an InputWritable as the value. As this will produce proportionally more intermediate data than input data, which two resources should you expect to be bottlenecks?

A. Processor and network I/O

B. Disk I/O and network I/O

C. Processor and RAM

D. Processor and disk I/O

Correct Answer: B

QUESTION 12

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)
```

What is the output of the following Pig commands?

X = GROUP A BY S1;

DUMP X;

A. (group,{(apple,peach,cherry,orange)})

B. {apple,peach,cherry,orange}

C. {1,4,2,3}

D. (apple,{(1,apple)})
(peach,{(4,peach)})
(cherry,{(2,cherry)})
(orange,{(3,orange)})

A. Option A

B. Option B

C. Option C

D. Option D

Correct Answer: D