CCD-410 Q&As

Cloudera Certified Developer for Apache Hadoop (CCDH)

Pass Cloudera CCD-410 Exam with 100% Guarantee

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

https://www.certbus.com/ccd-410.html

100% Passing Guarantee
100% Money Back Assurance

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

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

QUESTION 1

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

Files starting with \\'_\\' are considered \\'hidden\\' like unix files starting with \\'.\\'.

characters are allowed in HDFS file names.

QUESTION 2

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

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 3

To process input key-value pairs, your mapper needs to lead a 512 MB data file in memory. What is the best way to accomplish this?

A. Serialize the data file, insert in it the JobConf object, and read the data into memory in the configure method of the mapper.

B. Place the data file in the DistributedCache and read the data into memory in the map method of the mapper.

C. Place the data file in the DataCache and read the data into memory in the configure method of the mapper.

D. Place the data file in the DistributedCache and read the data into memory in the configure method of the mapper.

Correct Answer: D

QUESTION 4

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 5

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

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 6

In the reducer, the MapReduce API provides you with an iterator over Writable values. What does calling the next () method return?

A. It returns a reference to a different Writable object time.

B. It returns a reference to a Writable object from an object pool.

C. It returns a reference to the same Writable object each time, but populated with different data.

D. It returns a reference to a Writable object. The API leaves unspecified whether this is a reused object or a new object.

E. It returns a reference to the same Writable object if the next value is the same as the previous value, or a new Writable object otherwise.

Correct Answer: C

Calling Iterator.next() will always return the SAME EXACT instance of IntWritable, with the contents of that instance replaced with the next value.

Reference: manupulating iterator in mapreduce

QUESTION 7

In a MapReduce job with 500 map tasks, how many map task attempts will there be?

A. It depends on the number of reduces in the job.

B. Between 500 and 1000.

C. At most 500.

D. At least 500.

E. Exactly 500.

Correct Answer: D

Explanation: From Cloudera Training Course: Task attempt is a particular instance of an attempt to execute a task There will be at least as many task attempts as there are tasks If a task attempt fails, another will be started by the

JobTracker Speculative execution can also result in more task attempts than completed tasks

QUESTION 8

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

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 9

A client application creates an HDFS file named foo.txt with a replication factor of 3. Identify which best describes the file access rules in HDFS if the file has a single block that is stored on data nodes A, B and C?

A. The file will be marked as corrupted if data node B fails during the creation of the file.

B. Each data node locks the local file to prohibit concurrent readers and writers of the file.

C. Each data node stores a copy of the file in the local file system with the same name as the HDFS file.

D. The file can be accessed if at least one of the data nodes storing the file is available.

Correct Answer: D

HDFS keeps three copies of a block on three different datanodes to protect against true data corruption.

HDFS also tries to distribute these three replicas on more than one rack to protect against data availability

issues. The fact that HDFS actively monitors any failed datanode(s) and upon failure detection immediately

schedules re-replication of blocks (if needed) implies that three copies of data on three different nodes is

sufficient to avoid corrupted files.

Note:

HDFS is designed to reliably store very large files across machines in a large cluster. It stores each file as a sequence of blocks; all blocks in a file except the last block are the same size. The blocks of a file are replicated for fault tolerance. The block size and replication factor are configurable per file. An application can specify the number of replicas of a file. The replication factor can be specified at file creation time and can be changed later. Files in HDFS are write-once and have strictly one writer at any time. The NameNode makes all decisions regarding replication of blocks. HDFS uses rack-aware replica placement policy. In default configuration there are total 3 copies of a datablock on HDFS, 2 copies are stored on datanodes on same rack and 3rd copy on a different rack.

Reference: 24 Interview Questions and Answers for Hadoop MapReduce developers , How the HDFS Blocks are replicated?

QUESTION 10

Given a directory of files with the following structure: line number, tab character, string:

Example: 1 abialkjfjkaoasdfjksdlkjhqweroij 2 kadfjhuwqounahagtnbvaswslmnbfgy 3 kjfteiomndscxeqalkzhtopedkfsikj

You want to send each line as one record to your Mapper. Which InputFormat should you use to complete the line: conf.setInputFormat (____.class) ; ?

A. SequenceFileAsTextInputFormat

B. SequenceFileInputFormat

C. KeyValueFileInputFormat

D. BDBInputFormat

Correct Answer: C

http://stackoverflow.com/questions/9721754/how-to-parse-customwritable-from-text-in-hadoop