HADOOP INTERVIEW QUESTIONS

Hadoop Interview Questions

(This article is part of our Hadoop Guide. Use the right-hand menu to navigate.)

Q: Is Hadoop a database?
A: No. Hadoop is a write-only file system. But there are other products like Hive and HBase that provide a SQL-like interface to Hadoop for storing data in RDMB-like database structures.

Q: What commands do you use to start Hadoop?
A:start-dfs.sh and start-yarn.dfs

Q: What does Apache Pig do?
A: It is a way to write MapReduce jobs using a far simpler, SQL-like syntax than using Java, which is very wordy.

Q: How do you copy a local file to the HDFS
A: hadoop fs -put filename /(hadoop directory)

Q: What is the Hadoop machine learning library called?
A: Apache Mahout.

Q: How is Spark different than Hadoop?
A: Spark stores data in memory, thus running MapReduce operations much faster than Hadoop, which stores that on disk. Also it has command line interfaces in Scala, Python, and R. And it includes a machine learning library, Spark ML, that is developed by the Spark project and not separately, like Mahout.

Q: Map is MapReduce?
A: Map takes an input data file and reduces it to (key->value) pairs or tuples (a,b,c,d) or other iterable structure. Reduce then takes adjacent items and iterates over them to provide one final result.

Q: What does safemode in Hadoop mean?
A: It means the datanodes are not yet ready to receive data. This usually occurs on startup.

Q: How do you take Hadoop out of safemode?
A: hdfs dfsadmin -safemode leave

Q: What is the difference between the a namenode and datanode?
A: Hadoop is a master-slave model. The namenode is the master. The slaves are the datanodes. The namenode partitions MapReduce jobs and hands off each piece to different datanodes. Datanodes are responsible for writing data to disk.

Q: What role does Yarn play in Hadoop?
A: It is a resource manager. What it does is keep track of available resources (memory, CPU, storage) across the cluster (meaning the machines where Hadoop is running). Then each application (e.g. Hadoop) asks the resource manager what resources are available and doles those out accordingly. It runs two daemons to do this: Scheduler and ApplicationsManager.

Q: How do you add a datanode?
A: You copy the whole Hadoop $HADOOP_HOME folder to a server. Then you set up ssh keys so that the Hadoop user can ssh to that server without having to enter a password. Then you add the name of that server to$HADOOP_HOME/etc/hadoop/slaves. That you run hadoop-daemon.sh --config $HADOOP_CONF_DIR --script hdfs start datanode on the new data node.

Q: How do you see what Hadoop services are running? Name them.
A: Run jps. You should see: DataNode on the datanodes and NameNode, SecondaryNameNode, and ResourceManager on the NameNodes and (optionally) the JobHistoryServer.

Q: How do you start the Hadoop Job History Server?
A:$HADOOP_HOME/sbin/mr-jobhistory-daemon.sh --config $HADOOP_HOME/etc/hadoop start historyserver

Q: What is linear regression?
A: This is a technique used to find a function that most nearly matches a set of data points. For example if you have one independent value x and one dependant variable y then linear regression will calculate the y = mx + b where m is the slope and b is the x intercept. This is used in predictive data models. This is used to find a correlation between variables, for example whether studying more (x) increases student grades (y).

Q: What do you think "Bring the computing to the data instead of bringing the data to the computing" means?
A: This is the whole idea behind Hadoop. It means to use the computing power of commodity virtual machines to process pieces of a dataset instead of having one central powerful computer process that in one place. This model also lets Hadoop operate over datasets of almost unlimited size.

Q: What is Apache Cassandra?
A: This is a noSQL column-oriented database that works in a ring topology. That means it has no central controlling server. It looks like a regular row-and-column RDBMS database, like MySQL, in that it supports a SQL-like syntax. But it groups data by columns for fast retrieval and writes and not

rows. And when it writes an item it writes one row-column combination at a time. That means unlike RDBMS there can be rows that omit certain columns.

Q: What are the main Hadoop config files?
A: hdfs-site and core-site.xml

Q: How does Hadoop replication work? What does rack aware mean?
A: The datanodes write data in data blocks, just like a regular disk drive would. It writes a copy of each block to other datanodes depending on the Replication Factor. To say that datanodes are rack aware means it does not write replicas all on the same rack in the data center where a power or other outage would result in the loss of the data and the back-up too.

Q: What are some of the Hadoop CLI options?
A: The full list is: appendToFile cat checksum chgrp chmod chown copyFromLocal copyToLocal count cp createSnapshot deleteSnapshot df du dus expunge find get getfacl getfattr getmerge help ls lsr mkdir moveFromLocal moveToLocal mv put renameSnapshot rm rmdir rmr setfacl setfattr setrep stat tail test text touchz truncate usage

Q: What file types can Hadoop use to store its data?
A: Avro, Parquet, Sequence Files, and Plain Text

Q: What is a NameSpace?
A: It is an abstraction of a directory and file across the cluster. In other words /directory/file is a namespace that represents some file in some directory. But it is not local. It is on the Hadoop cluster, meaning it is stored across the data nodes.

Q: What does Apache Hive us a SQL server like Derby or mysql for?
A: It stores the schema there while it stores the data in Hadoop.

Q: How can you call an external program from Hive, like a Python one:
A: Use TRANSFORM like SELECT TRANSFORM (fields) USING 'python programName.py' as (fields) FROM table;

Q: What is Apache Flume?
A: It is a way to write streaming data to Hadoop.

Q: What is Hadoop High Availability?
A: That means configuring a second namenode to work as a hot standby incase the primary namenode crashes.

Q: What does the fsck command do?
A: It checks for bad blocks (i.e., corrupt files) and problems with replication.

Q: What kind of security does Hadoop have? How can you add authentication?
A: Hadoop by default only has file permissions security like a regular UNIX file system. You change permissions on that using chown and chmod and the regular Linx account or LDAP account. But if you want to have a higher level of authentication you would enable Kerberos, which is the authentication system used by Windows and optionally used by Linux. Then the datanodes would need to authenticate to connect to other nodes.