edureka!

- Blog Home
- Webinars
- Courses »
- Interview Questions

Hadoop Interview Questions -Setting Up Hadoop Cluster

April 23, 2013 | Big Data and Hadoop, Interview Questions

ıt for Hadoop Interview Questions that are frequently asked by employers?
second list of Hadoop Interview Questions which covers setting up Hadoop Cluster…

the three modes in which Hadoop can be run?

odes in which Hadoop can be run are:

e (local) mode
listributed mode
ributed mode

he features of Stand alone (local) mode?

In stand-alone mode there are no daemons, everything runs on a single JVM. It has no DFS and utilizes the local file system. Stand-alone mode is suitable only for running MapReduce programs during development. It is one of the most least used environments.

What are the features of Pseudo mode?

Pseudo mode is used both for development and in the QA environment. In the Pseudo mode all the daemons run on the same machine.

Can we call VMs as pseudos?

No, VMs are not pseudos because VM is something different and pesudo is very specific to Hadoop.

What are the features of Fully Distributed mode?

Fully Distributed mode is used in the production environment, where we have 'n' number of machines forming a Hadoop cluster. Hadoop daemons run on a cluster of machines. There is one host onto which Namenode is running and another host on which datanode is running and then there are machines on which task tracker is running. We have separate masters and separate slaves in this distribution.

Does Hadoop follows the UNIX pattern?

Yes, Hadoop closely follows the UNIX pattern. Hadoop also has the '*conf*' directory as in the case of UNIX.

In which directory Hadoop is installed?

Cloudera and Apache has the same directory structure. Hadoop is installed in **cd /usr/lib/hadoop-0.20/.**

What are the port numbers of Namenode, job tracker and task tracker?

The port number for Namenode is '70′, for job tracker is '30′ and for task tracker is '60′.

What is the Hadoop-core configuration?

Hadoop core is configured by two xml files:
1. hadoop-default.xml which was renamed to **2. hadoop-site.xml**.
These files are written in xml format. We have certain properties in these xml files, which consist of name and value. But these files do not exist now.

What are the Hadoop configuration files at present?

There are 3 configuration files in Hadoop:

1. core-site.xml

2. hdfs-site.xml

3. mapred-site.xml

These files are located in the *conf/ subdirectory*.

How to exit the Vi editor?

To exit the Vi Editor, press ESC and type :q and then press enter.

What is a spill factor with respect to the RAM?

Spill factor is the size after which your files move to the temp file. Hadoop-temp directory is used for this.

Is fs.mapr.working.dir a single directory?

Yes, ***fs.mapr.working.dir*** it is just one directory.

Which are the three main hdfs-site.xml properties?

The three main hdfs-site.xml properties are:

1. **dfs.name.dir** which gives you the location on which metadata will be stored and where DFS is located – on disk or onto the remote.

2. **dfs.data.dir** which gives you the location where the data is going to be stored.

3. **fs.checkpoint.dir** which is for secondary Namenode.

How to come out of the insert mode?

To come out of the insert mode, press ESC, type :q (if you have not written anything) OR type :wq (if you have written anything in the file) and then press ENTER.

What is Cloudera and why it is used?

Cloudera is the distribution of Hadoop. It is a user created on VM by default. Cloudera belongs to Apache and is used for data processing.

What happens if you get a 'connection refused java exception' when you type hadoop fsck /?

It could mean that the Namenode is not working on your VM.

We are using Ubuntu operating system with Cloudera, but from where we can download Hadoop or does it come by default with Ubuntu?

This is a default configuration of Hadoop that you have to download from Cloudera or from Edureka's dropbox and the run it on your systems. You can also proceed with your own configuration but you need a Linux box, be it Ubuntu or Red hat. There are installation steps present at the Cloudera location or in Edureka's Drop box. You can go either ways.

What does ‘jps’ command do?

This command checks whether your Namenode, datanode, task tracker, job tracker, etc are working or not.

How can I restart Namenode?

1. Click on **stop-all.sh** and then click on **start-all.sh** *OR*
2. Write **sudo hdfs** (press enter), **su-hdfs** (press enter), **/etc/init.d/ha** (press enter) and then **/etc/init.d/hadoop-0.20-namenode start** (press enter).

What is the full form of fsck?

Full form of fsck is *File System Check.*

How can we check whether Namenode is working or not?

To check whether Namenode is working or not, use the command **/etc/init.d/hadoop-0.20-namenode status** or as simple as **jps.**

What does the command mapred.job.tracker do?

The command **mapred.job.tracker** lists out which of your nodes is acting as a job tracker.

What does /etc /init.d do?

/etc /init.d specifies where daemons (services) are placed or to see the status of these daemons. It is very LINUX specific, and nothing to do with Hadoop.

How can we look for the Namenode in the browser?

If you have to look for Namenode in the browser, you don’t have to give localhost:8021, the port number to look for Namenode in the brower is *50070.*

How to change from SU to Cloudera?

To change from SU to Cloudera just type exit.

Which files are used by the startup and shutdown commands?

Slaves and *Masters* are used by the startup and the shutdown commands.

What do slaves consist of?

Slaves consist of a list of hosts, one per line, that host datanode and task tracker servers.

What do masters consist of?

Masters contain a list of hosts, one per line, that are to host secondary namenode servers.

What does hadoop-env.sh do?

hadoop-env.sh provides the environment for Hadoop to run. JAVA_HOME is set over here.

Can we have multiple entries in the master files?

Yes, we can have multiple entries in the Master files.

Where is hadoop-env.sh file present?

hadoop-env.sh file is present in the *conf* location.

In Hadoop_PID_DIR, what does PID stands for?

PID stands for 'Process ID'.

What does /var/hadoop/pids do?

It stores the PID.

What does hadoop-metrics.properties file do?

hadoop-metrics.properties is used for '*Reporting*' purposes. It controls the reporting for Hadoop. The default status is '*not to report*'.

What are the network requirements for Hadoop?

The Hadoop core uses Shell (SSH) to launch the server processes on the slave nodes. It requires *password-less* SSH connection between the master and all the slaves and the secondary machines.

Why do we need a password-less SSH in Fully Distributed environment?

We need a *password-less* SSH in a Fully-Distributed environment because when the cluster is LIVE and running in Fully
Distributed environment, the communication is too frequent. The job tracker should be able to send a task to task tracker quickly.

Does this lead to security issues?

No, not at all. Hadoop cluster is an isolated cluster. And generally it has nothing to do with an internet. It has a different kind of a configuration. We needn't worry about that kind of a security breach, for instance, someone hacking through the internet, and so on. Hadoop has a very secured way to connect to other machines to fetch and to process data.

On which port does SSH work?

SSH works on Port No. **22,** though it can be configured. **22** is the default Port number.

Can you tell us more about SSH?

SSH is nothing but a secure shell communication, it is a kind of a protocol that works on a Port No. 22, and when you do an SSH, what you really require is a password.

Why password is needed in SSH localhost?

Password is required in SSH for security and in a situation where *password-less* communication is not set.

Do we need to give a password, even if the key is added in SSH?

Yes, password is still required even if the key is added in SSH.

What if a Namenode has no data?

If a Namenode has no data it is not a Namenode. Practically, Namenode will have some data.

What happens to job tracker when Namenode is down?

When Namenode is down, your cluster is OFF, this is because Namenode is the single point of failure in HDFS.

What happens to a Namenode, when job tracker is down?

When a job tracker is down, it will not be functional but Namenode will be present. So, cluster is accessible if Namenode is working, even if the job tracker is not working.

Can you give us some more details about SSH communication between Masters and the Slaves?

SSH is a password-less secure communication where data packets are sent across the slave. It has some format into which data is sent across. SSH is not only between masters and slaves but also between two hosts.

What is formatting of the DFS?

Just like we do for Windows, DFS is formatted for proper structuring. It is not usually done as it formats the Namenode too.

Does the HDFS client decide the input split or Namenode?

No, the Client does not decide. It is already specified in one of the configurations through which input split is already configured.

In Cloudera there is already a cluster, but if I want to form a cluster on Ubuntu can we do it?

Yes, you can go ahead with this! There are installation steps for creating a new cluster. You can uninstall your present cluster and install the new cluster.

Can we create a Hadoop cluster from scratch?

Yes we can do that also once we are familiar with the Hadoop environment.

Can we use Windows for Hadoop?

Actually, *Red Hat Linux* or *Ubuntu* are the best Operating Systems for Hadoop. Windows is not used frequently for installing Hadoop as there are many support problems attached with Windows. Thus, Windows is not a preferred environment for Hadoop.

To refer to the first list, click Hadoop Interview Questions – HDFS!

To refer to the third list, click Hadoop Interview Questions – MapReduce!

To refer to the fourth list, click Hadoop Interview Questions - PIG!

0 Comments http://www.edureka.in/blog/ Login

Sort by Best Share Favorite ★

Start the discussion…

Be the first to comment.

ALSO ON **HTTP://WWW.EDUREKA.IN/BLOG/** WHAT'S THIS?

Hadoop Admin Responsibilities

1 comment • 20 days ago

Souvik — Thank you :)

Free Webinar on 'Introduction to MongoDB'

12 comments • 5 days ago

EdurekaSupport — Hi, Sorry for the inconvenience. We have fixed it and is working fine now. You can also …

Sample HBase POC

1 comment • 4 months ago

Sagar Morakhia — nice sample example. :)

Is Big Data the Right Move for You?

13 comments • 2 months ago

Newb — ios dev

Subscribe Add Disqus to your site

- **Search**

-

- Recent Posts

  - Free Webinar on 'Introduction to MongoDB'
  - Importance of Data Science With Cassandra
  - How to become a Hadoop Administrator?
  - Why Big Data Professionals Need to Learn MongoDB?
  - Why Should you go for Hadoop Administration Course?

- Categories

  - Android
  - Apache Cassandra
  - Apache Storm
  - Big Data and Hadoop
  - Business Analytics With R
  - Cloud Computing
  - Data Science
  - Hadoop Administration
  - Interview Questions
  - Java
  - MongoDB
  - PMI-ACP
  - PMP Exam Preparation
  - Python for Big Data Analytics
  - Resources & Misc.
  - Uncategorized
  - Webinars

Enter your Email Address...

Subscribe

RSS Feed