**Hadoop面试45个题目和参考答案**

**1.**[Hadoop](http://www.superwu.cn/2014/09/06/1233/)集群可以运行的3个模式？

* 单机（本地）模式
* 伪分布式模式
* 全分布式模式

**2.**单机（本地）模式中的注意点？

在单机模式（standalone）中不会存在守护进程，所有东西都运行在一个JVM上。这里同样没有DFS，使用的是本地文件系统。单机模式适用于开发过程中运行MapReduce程序，这也是最少使用的一个模式。

**3.**伪分布模式中的注意点？

伪分布式（Pseudo）适用于开发和测试环境，在这个模式中，所有守护进程都在同一台机器上运行。

**4.**VM是否可以称为Pseudo？

不是，两个事物，同时Pseudo只针对Hadoop。

**5.**全分布模式又有什么注意点？

全分布模式通常被用于生产环境，这里我们使用N台主机组成一个Hadoop集群，Hadoop守护进程运行在每台主机之上。这里会存在Namenode运行的主机，Datanode运行的主机，以及task tracker运行的主机。在分布式环境下，主节点和从节点会分开。

**6.** Hadoop是否遵循UNIX模式？

是的，在UNIX用例下，Hadoop还拥有"conf"目录。

**7.** Hadoop安装在什么目录下？

Cloudera和Apache使用相同的目录结构，Hadoop被安装在cd/usr/lib/hadoop-0.20/。

**8.** [Namenode](http://www.superwu.cn/2014/11/04/1283/)、Job tracker和task tracker的端口号是？

Namenode，70；Job tracker，30；Task tracker，60。

**9.**Hadoop的核心配置是什么？

Hadoop的核心配置通过两个xml文件来完成：1，hadoop-default.xml；2，hadoop-site.xml。这些文件都使用xml格式，因此每个xml中都有一些属性，包括名称和值，但是当下这些文件都已不复存在。

**10.**那当下又该如何配置？

Hadoop现在拥有3个配置文件：1，core-site.xml；2，hdfs-site.xml；3，mapred-site.xml。这些文件都保存在conf/子目录下。

**11.** RAM的溢出因子是？

溢出因子（Spill factor）是临时文件中储存文件的大小，也就是Hadoop-temp目录。

**12.**fs.mapr.working.dir只是单一的目录？

fs.mapr.working.dir只是一个目录。

**13.** hdfs-site.xml的3个主要属性？

* dfs.name.dir决定的是元数据存储的路径以及DFS的存储方式（磁盘或是远端）
* dfs.data.dir决定的是数据存储的路径
* fs.checkpoint.dir用于第二Namenode

**14.**如何退出输入模式？

退出输入的方式有：1，按ESC；2，键入:q（如果你没有输入任何当下）或者键入:wq（如果你已经输入当下），并且按下Enter。

**15.**当你输入hadoopfsck /造成"connection refused java exception'"时，系统究竟发生了什么？

这意味着Namenode没有运行在你的VM之上。

**16.**我们使用Ubuntu及Cloudera，那么我们该去哪里下载Hadoop，或者是默认就与Ubuntu一起安装？

这个属于Hadoop的默认配置，你必须从Cloudera或者Edureka的dropbox下载，然后在你的系统上运行。当然，你也可以自己配置，但是你需要一个Linux box，Ubuntu或者是Red Hat。在Cloudera网站或者是Edureka的Dropbox中有安装步骤。

**17.** "jps"命令的用处？

这个命令可以检查Namenode、Datanode、Task Tracker、 Job Tracker是否正常工作。

**18.** 如何重启Namenode？

* 点击stop-all.sh，再点击start-all.sh。
* 键入sudo hdfs（Enter），su-hdfs （Enter），/etc/init.d/ha（Enter），及/etc/init.d/hadoop-0.20-namenode start（Enter）。

**19.**Fsck的全名？

全名是：File System Check。

**20.** 如何检查Namenode是否正常运行？

如果要检查Namenode是否正常工作，使用命令/etc/init.d/hadoop-0.20-namenode status或者就是简单的jps。

**21.** mapred.job.tracker命令的作用？

可以让你知道哪个节点是Job Tracker。

**22.** /etc /init.d命令的作用是？

/etc /init.d说明了守护进程（服务）的位置或状态，其实是LINUX特性，和Hadoop关系不大。

**23.**如何在浏览器中查找Namenode？

如果你确实需要在浏览器中查找Namenode，你不再需要localhost:8021，Namenode的端口号是50070。

**24.**如何从SU转到Cloudera？

从SU转到Cloudera只需要键入exit。

**25.** 启动和关闭命令会用到哪些文件？

Slaves及Masters。

**26.** Slaves由什么组成？

Slaves由主机的列表组成，每台1行，用于说明数据节点。

**27.** Masters由什么组成？

Masters同样是主机的列表组成，每台一行，用于说明第二Namenode服务器。

**28.** hadoop-env.sh是用于做什么的？

hadoop-env.sh提供了Hadoop中. JAVA\_HOME的运行环境。

**29.**Master文件是否提供了多个入口？

是的你可以拥有多个Master文件接口。

**30.**hadoop-env.sh文件当下的位置？

hadoop-env.sh现在位于conf。

**31.**在Hadoop\_PID\_DIR中，PID代表了什么？

PID代表了"Process ID"。

**32.**  /var/hadoop/pids用于做什么？

/var/hadoop/pids用来存储PID。

**33.** hadoop-metrics.properties文件的作用是？

hadoop-metrics.properties被用做"Reporting"，控制Hadoop报告，初始状态是"not to report"。

**34.**  Hadoop需求什么样的网络？

Hadoop核心使用Shell（SSH）来驱动从节点上的服务器进程，并在主节点和从节点之间使用password-less SSH连接。

**35.** 全分布式环境下为什么需求password-less SSH？

这主要因为集群中通信过于频繁，Job Tracker需要尽可能快的给Task Tracker发布任务。

**36.** 这会导致安全问题吗？

完全不用担心。Hadoop集群是完全隔离的，通常情况下无法从互联网进行操作。与众不同的配置，因此我们完全不需要在意这种级别的安全漏洞，比如说通过互联网侵入等等。Hadoop为机器之间的连接提供了一个相对安全的方式。

**37.** SSH工作的端口号是？

SSH工作的端口号是NO.22，当然可以通过它来配置，22是默认的端口号。

**38.**SSH中的注意点还包括？

SSH只是个安全的shell通信，可以把它当做NO.22上的一种协议，只需要配置一个密码就可以安全的访问。

**39.**  为什么SSH本地主机需要密码？

在SSH中使用密码主要是增加安全性，在某些情况下也根本不会设置密码通信。

**40.**  如果在SSH中添加key，是否还需要设置密码？

是的，即使在SSH中添加了key，还是需要设置密码。

**41.**假如Namenode中没有数据会怎么样？

没有数据的Namenode就不能称之为Namenode，通常情况下，Namenode肯定会有数据。

**42.**当Job Tracker宕掉时，Namenode会发生什么？

当Job Tracker失败时，集群仍然可以正常工作，只要Namenode没问题。

**43.** 是客户端还是Namenode决定输入的分片？

这并不是客户端决定的，在配置文件中以及决定分片细则。

**44.** 是否可以自行搭建Hadoop集群？

是的，只要对Hadoop环境足够熟悉，你完全可以这么做。

**45.** 是否可以在Windows上运行Hadoop？

你最好不要这么做，Red Hat Linux或者是Ubuntu才是Hadoop的最佳操作系统。在Hadoop安装中，Windows通常不会被使用，因为会出现各种各样的问题。因此，Windows绝对不是Hadoop的推荐系统。

想了解更多hadoop 相关资讯请加qq群204069925~