Turing AI 万维图灵 | 大数据系列课程

大数据技术-第二章：Hadoop运行及开发环境搭建

Hadoop集群环境搭建概述

CONTENTS

01. 集群概念

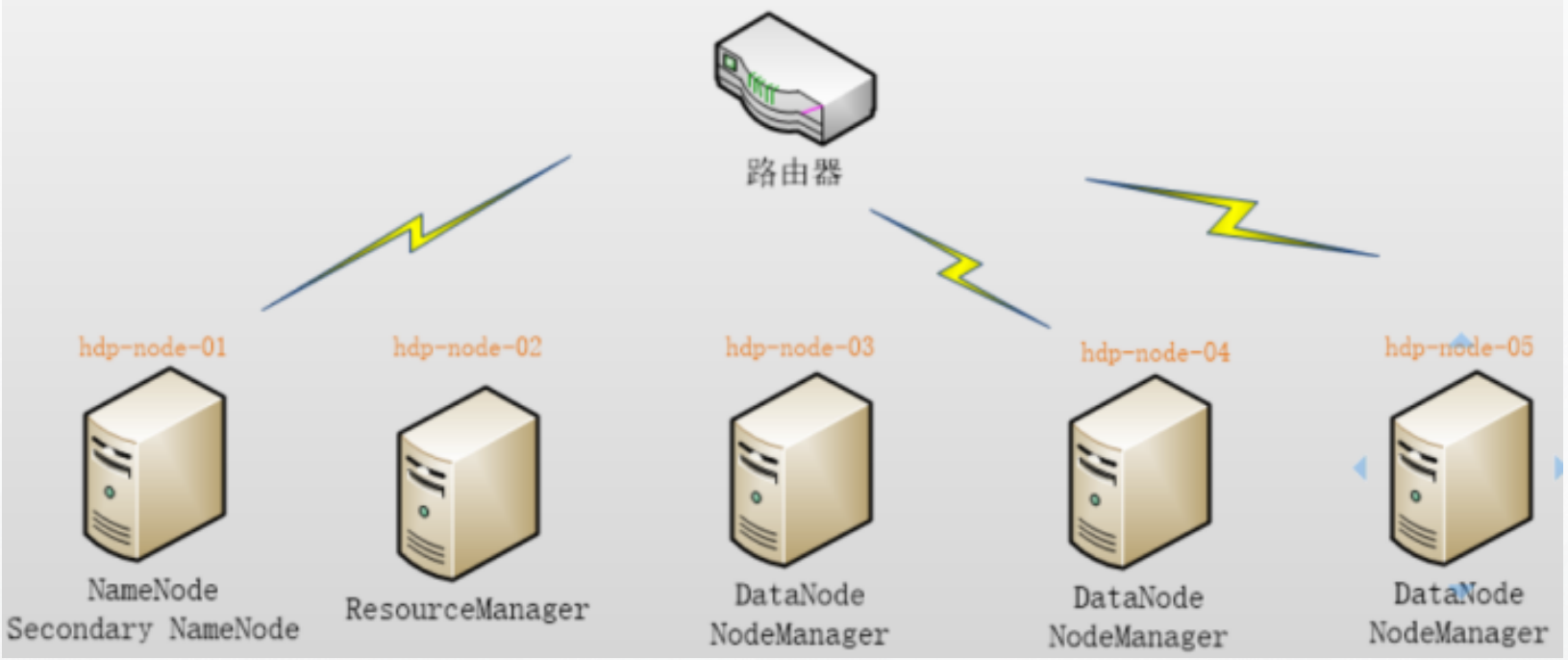
02 Hadoop集群部署

03. Hadoop三种运行模式

01

集群概念

集群概念



所谓集群，就是一组通过网络互联的计算机，集群中的每一台计算机称作一个节点，Hadoop集群搭建就是在这个物理集群之上安装部署Hadoop相关的软件，然后对外提供大数据存储和分析等相关服务。

02

Hadoop集群部署

Hadoop集群部署

Turing AI 万图 维灵
www.tringai.com

一个前提：Hadoop是为了在Linux平台上使用而开发的
一个现实：我的电脑不是Linux系统
如何解决：搭建虚拟机，在虚拟机上安装Linux操作系统

虚拟机是什么?
虚拟的计算机，功能和真实计算机几乎完全一样

如何搭建虚拟机?
在真实电脑上安装虚拟化软件来实现虚拟机的搭建

虚拟化软件有哪些?

VMware workstation和Virtualbox

版本选择及注意事项

9,10,11,12都可以，但是要注意输入对应版本的序列号

Linux运行环境的部署

搭建一个虚拟机，然后再在这个虚拟机上直接安转部署Linux操作系统来实现Linux运行环境。

03

Hadoop三种运行模式

Hadoop三种运行模式

单机模式（独立模式）（Local或Standalone Mode）

- 默认情况下，Hadoop即处于该模式，用于开发和调式。
- 不对配置文件进行修改。
- 使用本地文件系统，而不是分布式文件系统。
- Hadoop不会启动NameNode、DataNode、JobTracker、TaskTracker等守护进程，Map()和Reduce()任务作为同一个进程的不同部分来执行的。
- 用于对MapReduce程序的逻辑进行调试，确保程序的正确。

Hadoop三种运行模式

伪分布式模式（Pseudo-Distrubuted Mode）

- Hadoop的守护进程运行在本机机器，模拟一个小规模的集群
- 在一台主机模拟多主机。
- Hadoop启动NameNode、DataNode、JobTracker、TaskTracker这些守护进程都在同一台机器上运行，是相互独立的Java进程。
- 在这种模式下，Hadoop使用的是分布式文件系统，各个作业也是由JobTraker服务，来管理的独立进程。这种模式常用来开发测试Hadoop程序的执行是否正确。
- 修改3个配置文件：core-site.xml（Hadoop集群的特性，作用于全部进程及客户端）、hdfs-site.xml（配置HDFS集群的工作属性）、mapred-site.xml（配置MapReduce集群的属性）

Hadoop三种运行模式

全分布式集群模式（Full-Distributed Mode）

- Hadoop的守护进程运行在一个集群上
- Hadoop的守护进程运行在由多台主机搭建的集群上，是真正的生产环境。
- 在所有的主机上安装JDK和Hadoop，组成相互连通的网络。
- 在主机间设置SSH免密码登录，把各从节点生成的公钥添加到主节点的信任列表。
- 修改3个配置文件：core-site.xml、hdfs-site.xml、mapred-site.xml，指定NameNode和JobTraker的位置和端口，设置文件的副本等参数。

Turing AI 万维图灵 | 大数据系列课程

大数据

BIG DATA

智 / 能 / 科 / 技　　放 / 眼 / 未 / 来