Turing AI 万维图灵 | 大数据系列课程

大数据技术-第三章：HDFS分布式文件系统

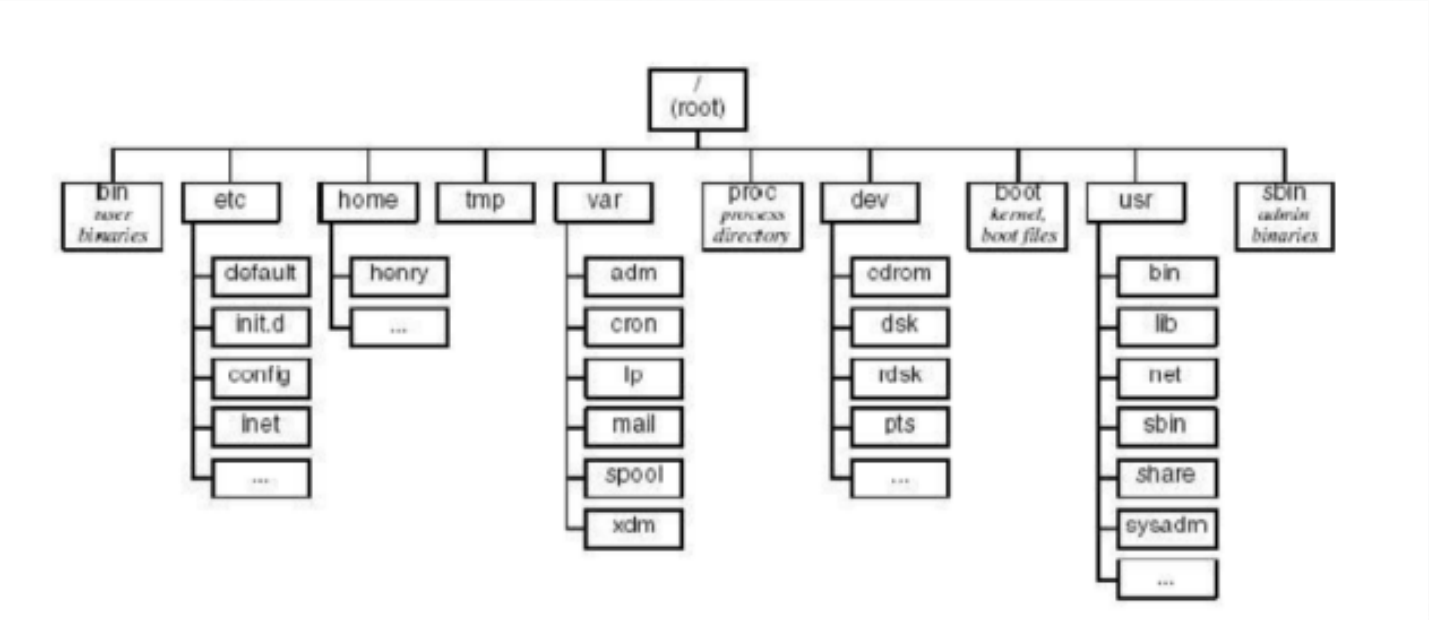
HDFS体系结构概述

CONTENTS

01

文件系统概述

文件系统概述



文件系统是操作系统提供的用于解决“如何在磁盘上组织文件”的一系列方法和数据结构。目前大家不用关心文件具体在磁盘上是如何存放的，只需要能够熟练掌握类似于指定文件的存储路径，往哪个路径下的文件写数据，从哪个路径下读取文件数据等基本的文件系统操作就可以了。

02

分布式文件系统概述

分布式文件系统概述

分布式文件系统是指利用多台计算机协同作用解决单台计算机所不能解决的存储问题的文件系统。比如：

- 单机负载可能极高
- 数据不安全
- 文件整理困难

03

HDFS分布式文件系统概述

HDFS分布式文件系统概述

HDFS是什么?

HDFS（Hadoop Distributed File System）是Hadoop项目的核心子项目，是分布式计算中数据存储管理的基础，是基于流式数据访问和处理超大文件的需求而开发的分布式文件系统。整个系统可以运行在由廉价的商用服务器组成的集群之上，它所具有的高容错性、高可靠性、高可扩展性、高获得性、高吞吐率等特征为海量数据提供了不怕故障的存储，给超大数据集的应用处理带来了很多便利。

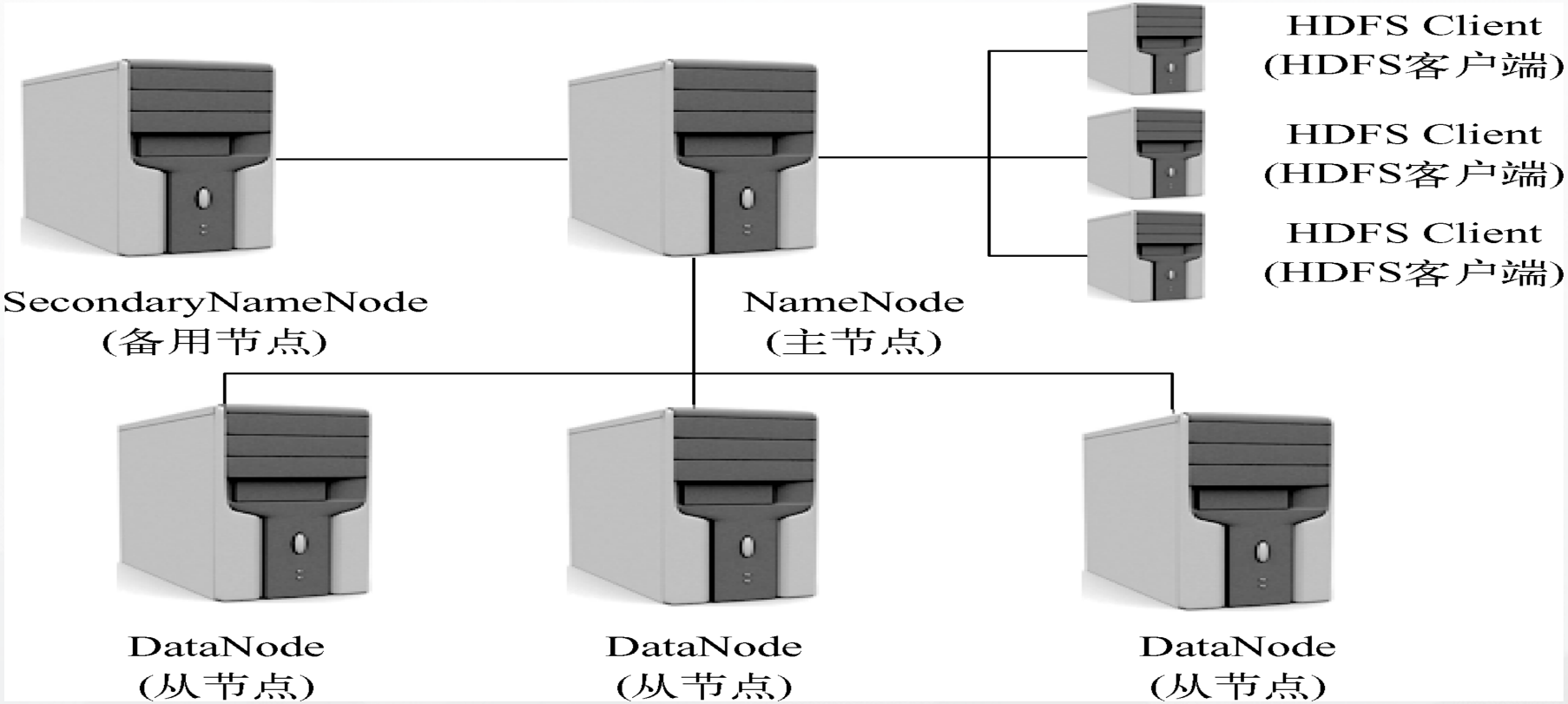
HDFS分布式文件系统概述

HDFS的设计理念

- 支持超大文件存储
- 流式数据访问
- 简单的一致性模型
- 硬件故障的检测和快速应对

HDFS分布式文件系统概述

HDFS体系结构



HDFS分布式文件系统概述

HDFS的优缺点

(1) HDFS的优点

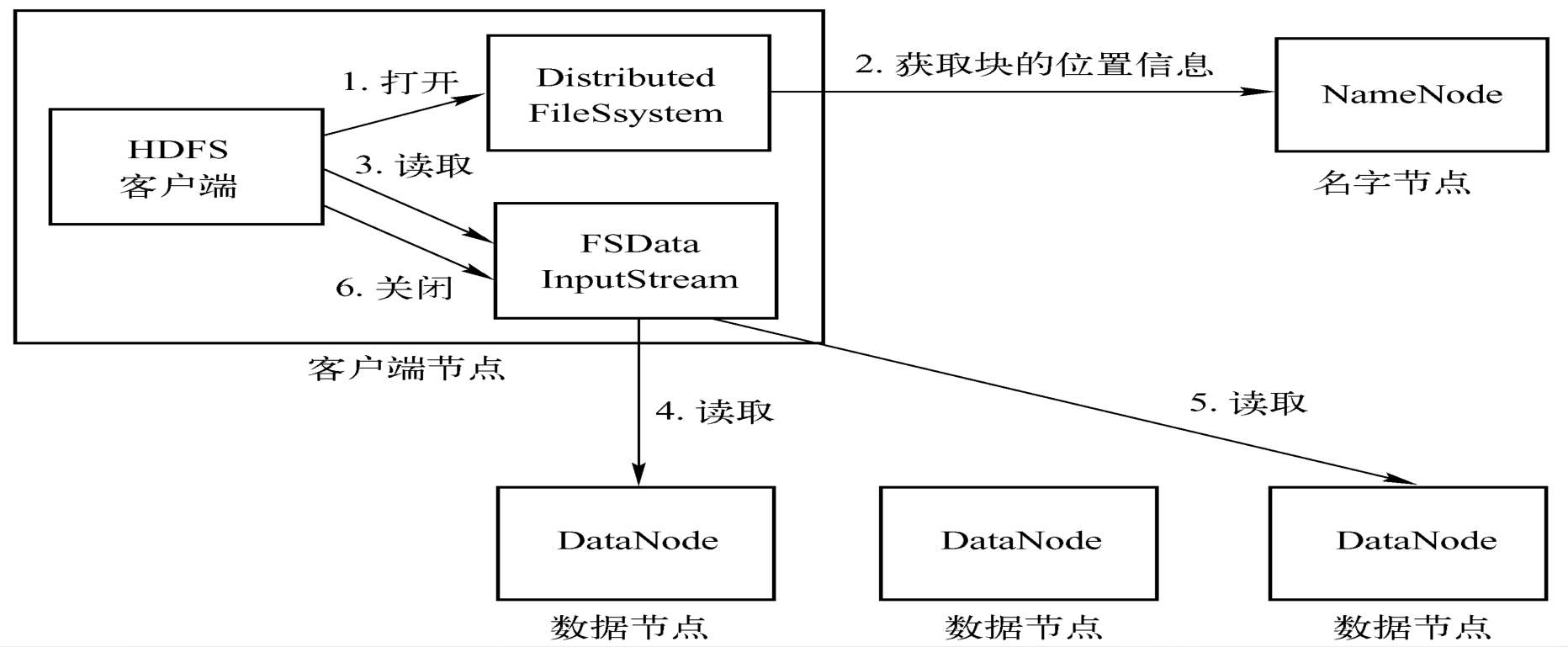
1) 高容错性
2) 适合大数据处理
3) 流式文件访问
4) 可构建在廉价的机器上

(2) HDFS的缺点

1) 不适合低延时数据访问
2) 不适合大量小文件的存储
3) 不适合并发写入、文件随机修改

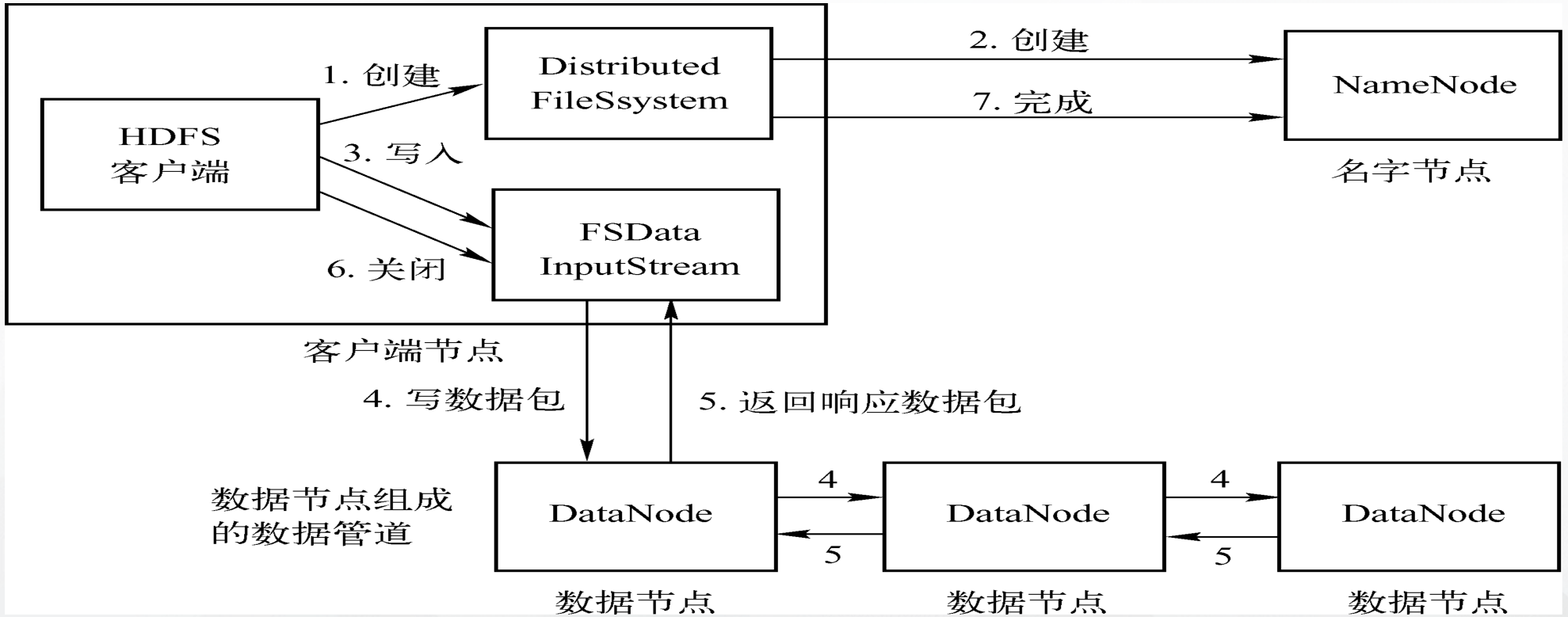
HDFS分布式文件系统概述

HDFS读数据流程



HDFS分布式文件系统概述

HDFS写数据流程

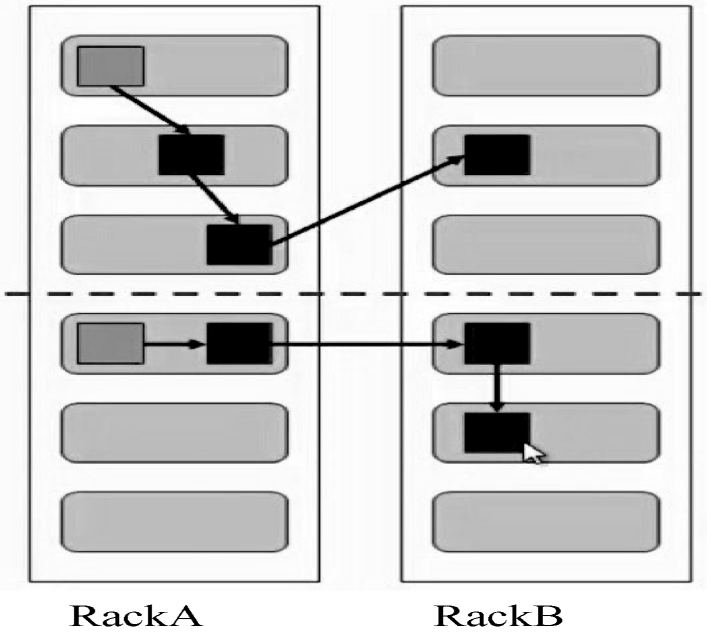


HDFS副本存放策略

HDFS副本放置策略

- Hadoop 0.17之前~
  - 副本1：同机架的不同节点
  - 副本2：同机架的另一个节点
  - 副本3：不同机架另一个节点
  - 其他副本：随机挑选

- Hadoop 0.17之后~
  - 副本1：同Client的节点上
  - 副本2：不同机架中的节点上
  - 副本3：同第二个副本的机架中的另一个节点上
  - 其他副本：随机挑选



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