Turing AI 万维图灵 | 大数据系列课程

大数据技术-第一章：Hadoop大数据概述

Hadoop的核心组件

CONTENTS

01

Hadoop组件概述

Hadoop组件概述

Hadoop Common是一个公共基础设施，用于支撑其他项目，包括RPC、序列化包等

可扩展、容错、高性能的分布式文件系统，异步复制，一次写入多次读取

分布式计算框架；
主要包含map（映射）和reduce（规约）过程

02

HDFS分布式文件系统

HDFS分布式文件系统

分布式文件系统(HDFS，Hadoop Distributed File System)

- 高度容错性的系统

  上传的数据自动保存多个副本，适合部署在廉价的机器上。

- 适合大数据的处理

  高吞吐量的数据访问，非常适合大规模数据集上的应用。

- 流式文件写入

  一次写入，多次读取。文件一旦写入，不能修改，只能增加。这样可以保证数据的一致性。

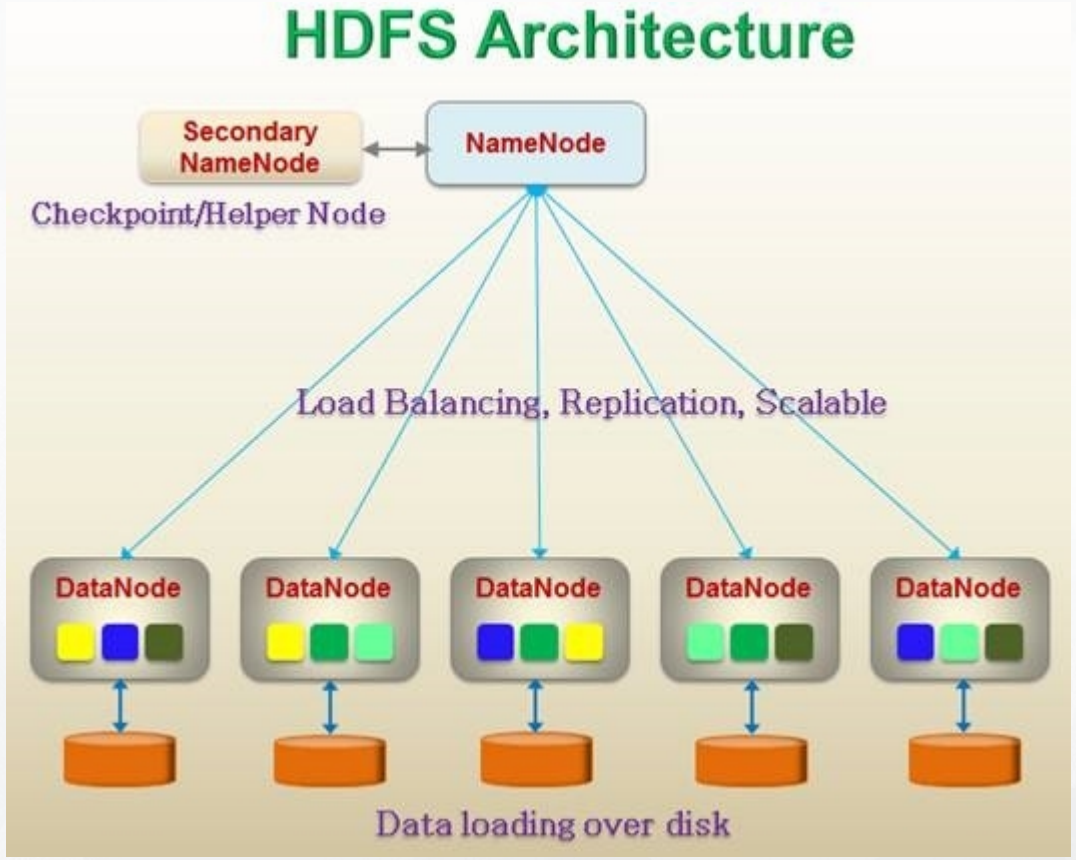
HDFS分布式文件系统

➢HDFS并不是一个单机文件系统，它是分布在多个集群节点上的文件系统。节点之间通过网络通信进行协作，提供个节点文件信息，让每个用户都可以看到文件系统的文件，让多机器上的多用户分享文件和存储空间。

➢文件存储时被分布在多个节点上。这里涉及到一个数据块的概念，数据存储不是按一个文件存储，而是把一个文件分成一个或多个数据块存储，数据块的概念在上一节已经描述过。数据块在存储时并不是都存在一个节点上，而是被分布存储在各个节点中，并且数据块会在其他节点存储副本。

➢数据读取从多个节点读取。读取一个文件时，从多个节点中找到该文件的数据块，分布读取所有数据块直到最后一个数据块读取完毕。

HDFS分布式文件系统



- NameNode：用于存储元数据以及处理客户端发出的请求；
- SN：一个Checkpoint来帮助NameNode更好的工作；
- DataNode：它为 HDFS 提供存储位置。

对外部客户机而言，HDFS就像一个传统的分级文件系统。可以创建、删除、移动或重命名文件，等等。

03

MapReduce框架

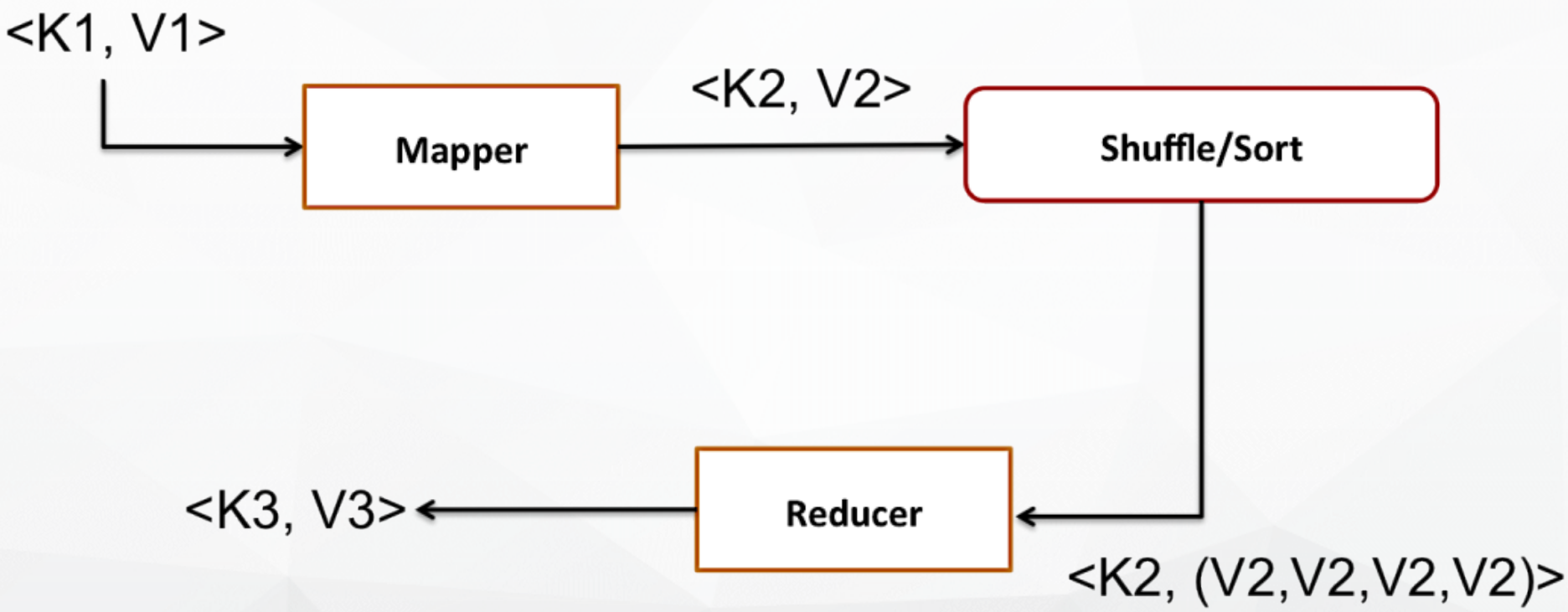
MapReduce框架

并行计算框架

➢MapReduce是Google提出的一个软件架构，用于大规模数据集（大于1TB）的并行运算。概念“Map（映射）”和“Reduce（归纳）”，及他们的主要思想，都是从函数式编程语言借来的，还有从矢量编程语言借来的特性。

➢当前的软件实现是指定一个Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的Reduce（归纳）函数，用来保证所有映射的键值对中的每一个共享相同的键组。

MapReduce框架



04

YARN资源管理系统

YARN资源管理系统

Turing AI 万图 维灵
www.tringai.com

YARN是Hadoop 2.0中的资源管理系统，它的基本设计思想是将MRv1（Hadoop1.0中的MapReduce）中的JobTracker拆分成了两个独立的服务：

- 全局的资源管理器ResourceManager
- 每个应用程序特有的ApplicationMaster。

其中ResourceManager负责整个系统的资源管理和分配，而ApplicationMaster负责单个应用程序的管理。

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