IB00109 云计算技术

授课教师：姜婧妍

jiangjingyan@sztu.edu.cn

2023年

课程介绍

序号	主要教学内容	课时量
1	云计算概论与云计算基础	2
2	云计算系统架构	2
3	虚拟化技术	2
4	数据中心与并行计算	2
5	Docker容器技术	4
6	介绍大数据处理架构 Hadoop 框架	4
7	分布式文件系统 HDFS 的基本原理	4
8	分布式数据库 HBase 的使用方法	4
9	分布式并行编程模型 MapReduce 原理	2
10	数据仓库 Hive 的基本原理和使用方法	2
11	NoSQL 数据库的概念和基本原理	2
12	云数据库	2
13	Spark原理	2
14	云计算在互联网等领域的典型应用	2

第六章

大数据与分布式

授课教师：姜婧妍

jiangjingyan@sztu.edu.cn

2023年

目录

Contents

学习目标

Learning Objectives

1. 了解什么是数据
2. 了解数据对现实生活而言有什么意义

数据是什么

进入21世纪，我们的生活就迈入了"数据时代"

作为21世纪的新青年，"数据"一词经常出现。

数据无时无刻的在影响着我们的现实生活

- 什么是数据?
- 数据又如何影响现实生活?

数据是什么

数据：一种可以被鉴别的对客观事件进行记录的符号。

简单来说就是：对人类的行为及产生的事件的一种记录。

我们无时无刻都在产生数据：

数据是什么

这些我们的日常活动所产生的信息记录

都是数据

数据的价值

有什么用?

数据的价值

性别　下单时间

地区　下单商品

手机型号　数量

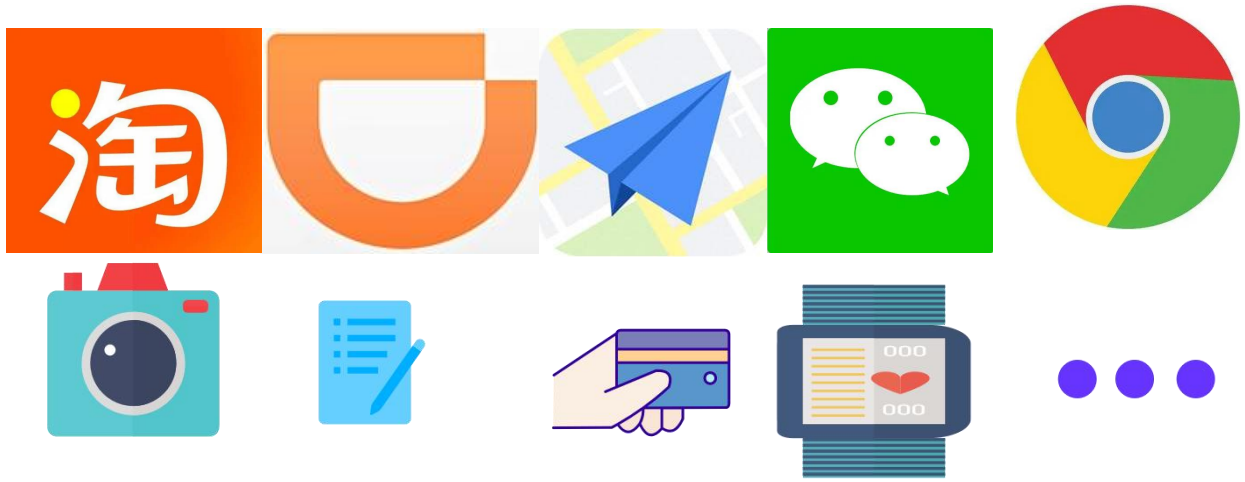
付款方式　评价

是否使用优惠券

.....

对淘宝来说：数据可以更好的了解用户

数据的价值



数据的背后都会隐藏着巨大的价值

丰富的数据支撑可以让我们更好的了解

事和物在现实世界的运行规律

创造价值、提高效率

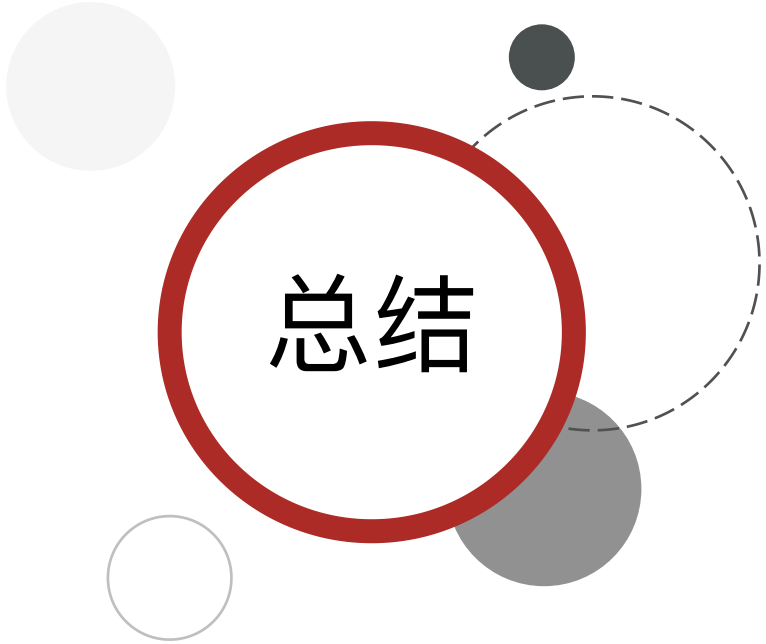
大数据时代

当下时代已经是数据的时代，数据非常重要并且蕴含巨大的价值。

大数据技术栈

对超大规模的数据进行处理并挖掘出数据背后的价值的技术体系

小结



1. 什么是数据?

人类的行为及产生的事件的一种记录称之为数据

2. 数据有什么价值?

- 对数据的内容进行深入分析，可以更好的帮助了解事和物在现实世界的运行规律
- 比如，购物的订单记录（数据）可以帮助平台更好的了解消费者，从而促进交易，提升销售额度。

目录

Contents

学习目标

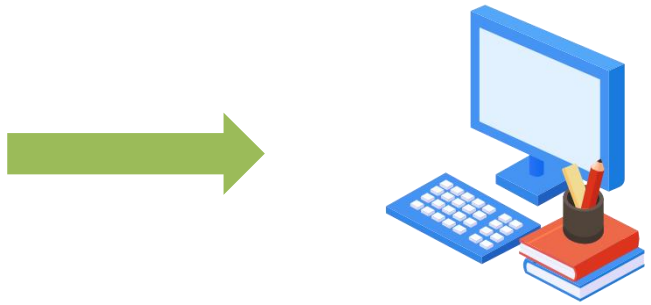
Learning Objectives

1. 了解大数据技术体系是如何诞生的
2. 了解Apache Hadoop对大数据体系的意义

大数据的诞生

大数据的诞生和信息化以及互联网的发展是密切相关的。

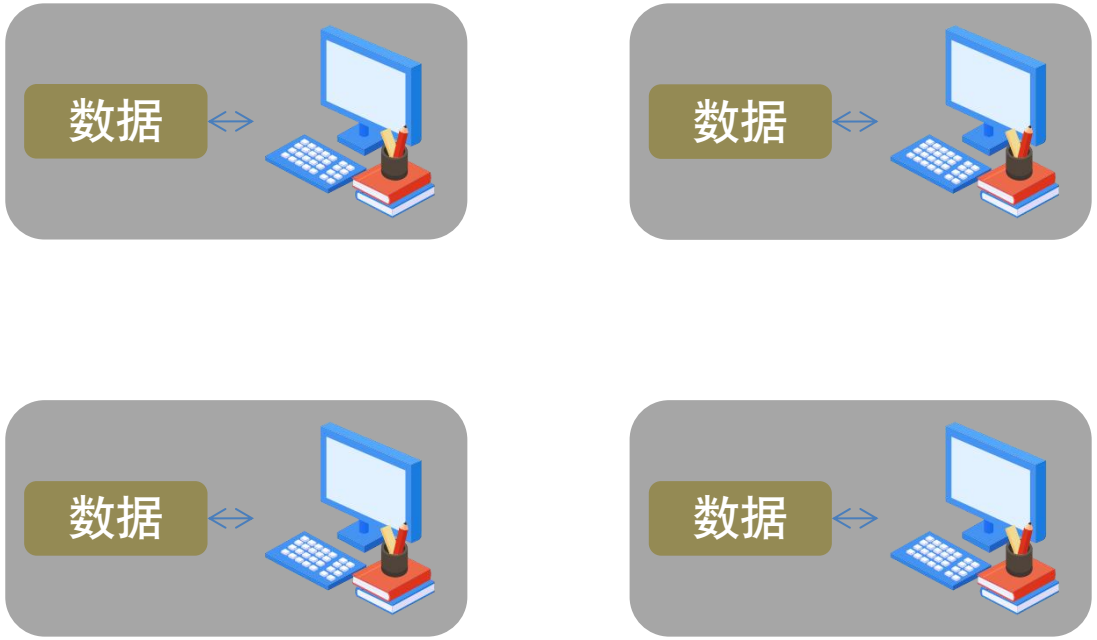
计算机发明前



计算机发明之后（上世纪50年代）

大数据的诞生

大数据的诞生和信息化以及互联网的发展是密切相关的。

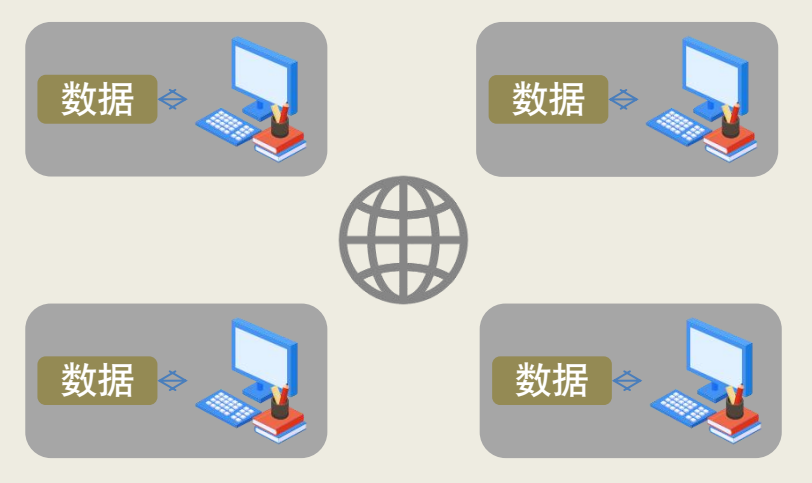


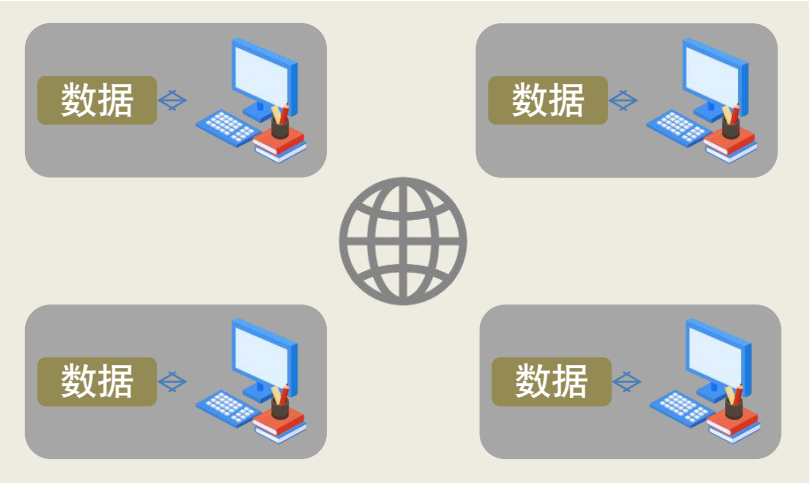
早期的计算机（上世纪70年代之前）

大多数是相互独立的，各自处理各自的数据

大数据的诞生

大数据的诞生和信息化以及互联网的发展是密切相关的。



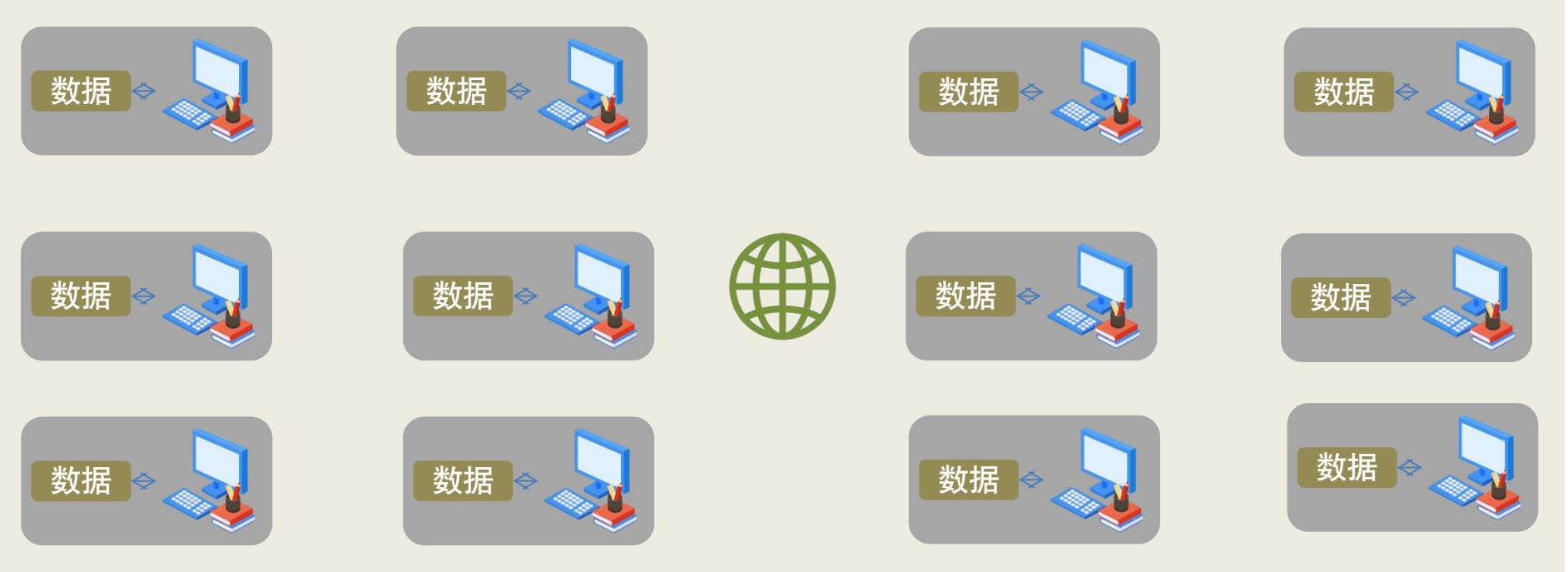


上世纪70年代后，逐步出现了基于TCP/IP协议的小规模的计算机互联互通。

但多数是军事、科研等用途。

大数据的诞生

大数据的诞生和信息化以及互联网的发展是密切相关的。

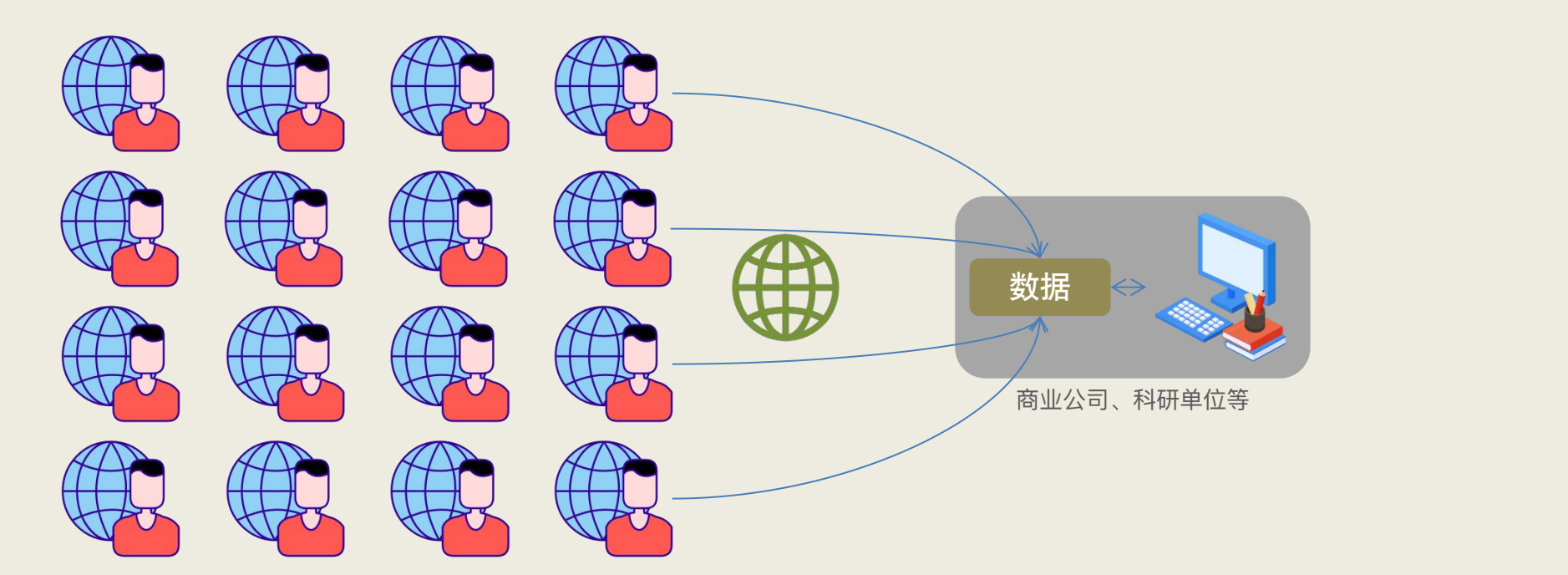


上世纪90年代左后，全球互联的互联网出现。

个人、企业均可参与其中，真正逐步的实现了全球互联。

大数据的诞生

大数据的诞生和信息化以及互联网的发展是密切相关的。



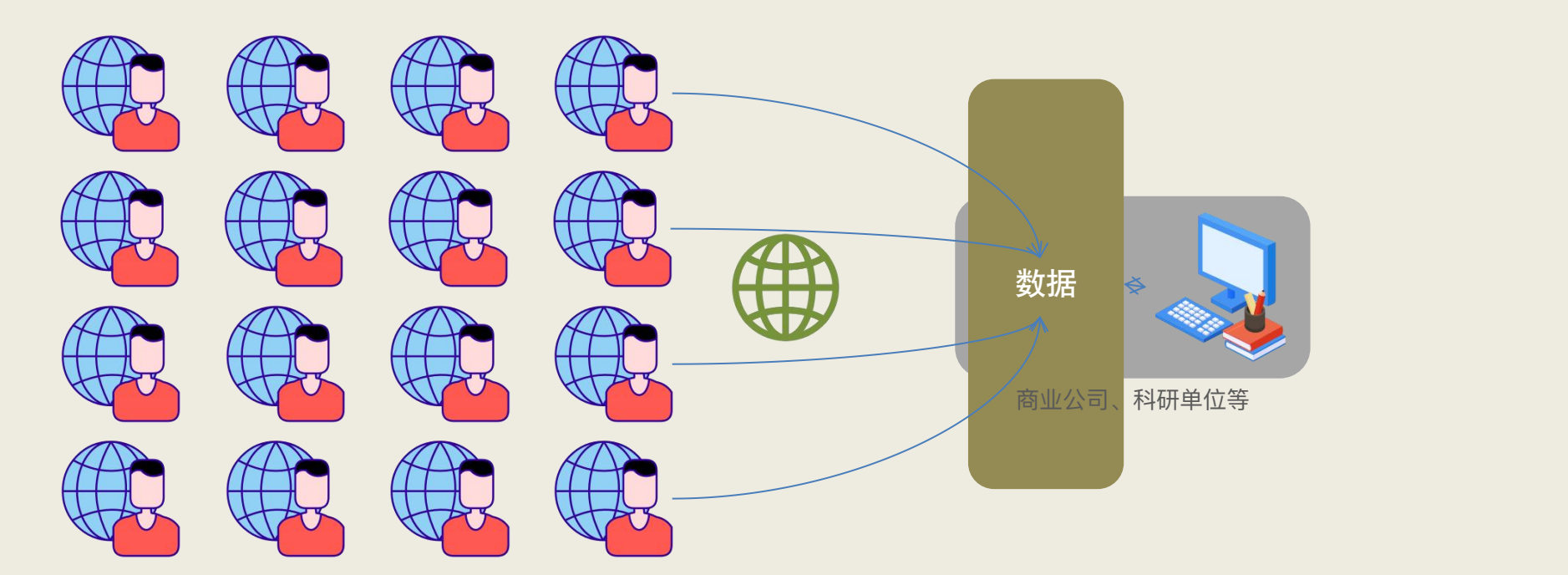
在2000年后，互联网上的商业行为剧增

现在知名的互联网公司（谷歌、AWS、腾讯、阿里等）也是在这个年代开始起步。

在互联网参与者众多的前提下，商业公司、科研单位等，所能获得的数据量也是剧增。

大数据的诞生

大数据的诞生和信息化以及互联网的发展是密切相关的。

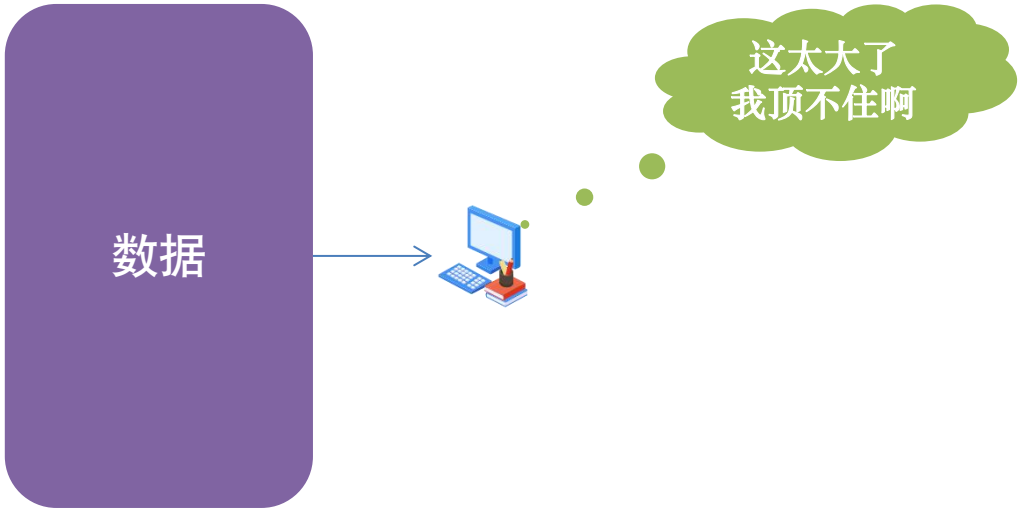


在2000年后，互联网上的商业行为剧增

现在知名的互联网公司（谷歌、AWS、腾讯、阿里等）也是在这个年代开始起步。

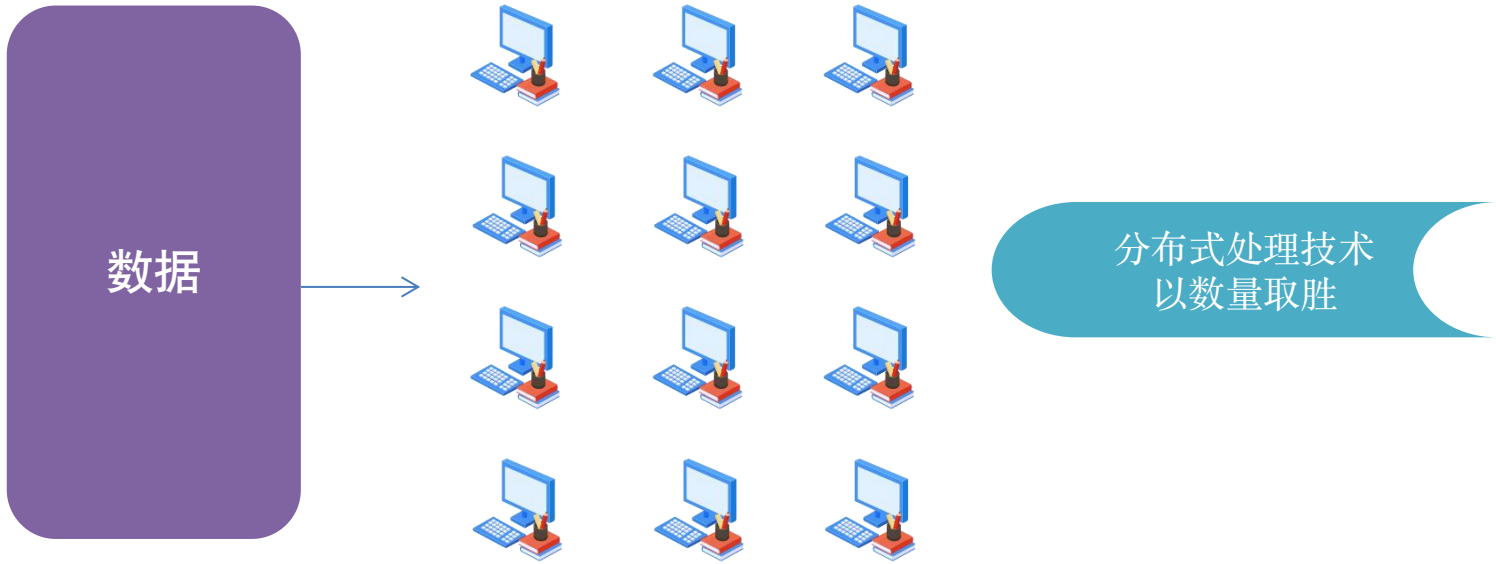
在互联网参与者众多的前提下，商业公司、科研单位等，所能获得的数据量也是剧增。

大数据的诞生



剧增的数据量，和羸弱的单机性能，让许多科技公司开始尝试以**数量**来解决问题。

大数据的诞生



剧增的数据量，和羸弱的单机性能，让许多科技公司开始尝试以**数量**来解决问题。

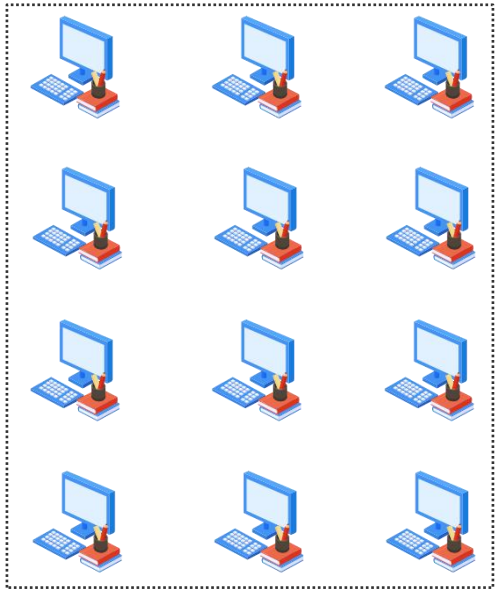
在这个过程中，分布式处理技术诞生了。

大数据的诞生

分布式处理技术

在**数据量巨大**的基础下

以服务器的**数量**来解决大规模数据处理问题



分布式服务器集群

用大量的服务器
解决大量的数据

大数据的诞生

分布式处理技术

在**数据量巨大**的基础下

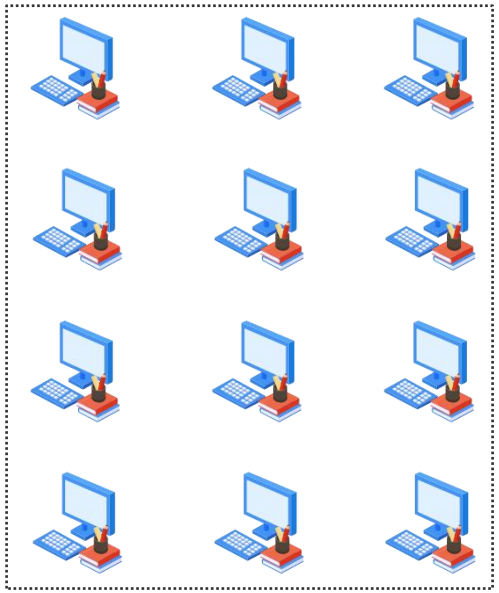
以服务器的**数量**来解决大规模数据处理问题

逐步演化

- 大规模服务器集群下的大规模数据存储
- 大规模服务器集群下的大规模数据计算
- 大规模服务器集群下的大规模数据传输技术

存 **用** **传**

大规模数据



分布式服务器集群

用大量的服务器解决大量的数据

大数据的诞生

分布式处理技术

在**数据量巨大**的基础下

以服务器的**数量**来解决大规模数据处理问题

逐步演化

- 大规模服务器集群下的大规模数据存储
- 大规模服务器集群下的大规模数据计算
- 大规模服务器集群下的大规模数据传输技术

大规模数据：存、用、传

2008年之前

这些在当时较为”高端”的分布式技术基本上还处于大企业内部专用且不够成熟

大数据的诞生

分布式处理技术

在**数据量巨大**的基础下

以服务器的**数量**来解决大规模数据处理问题

逐步演化

- 大规模服务器集群下的大规模数据存储 —— 存
- 大规模服务器集群下的大规模数据计算 —— 用
- 大规模服务器集群下的大规模数据传输技术 —— 传

大规模数据

2008年 Apache Hadoop开源
广大企业拥有了成熟的、开源的、分布式数据处理解决方案

大数据的诞生

Apache Hadoop 是一款开源的分布式处理技术栈
为业界提供了

- 基于Hadoop HDFS的：分布式数据存储技术
- 基于Hadoop MapReduce的：分布式数据计算技术
- 基于Hadoop YARN的：分布式资源调度技术

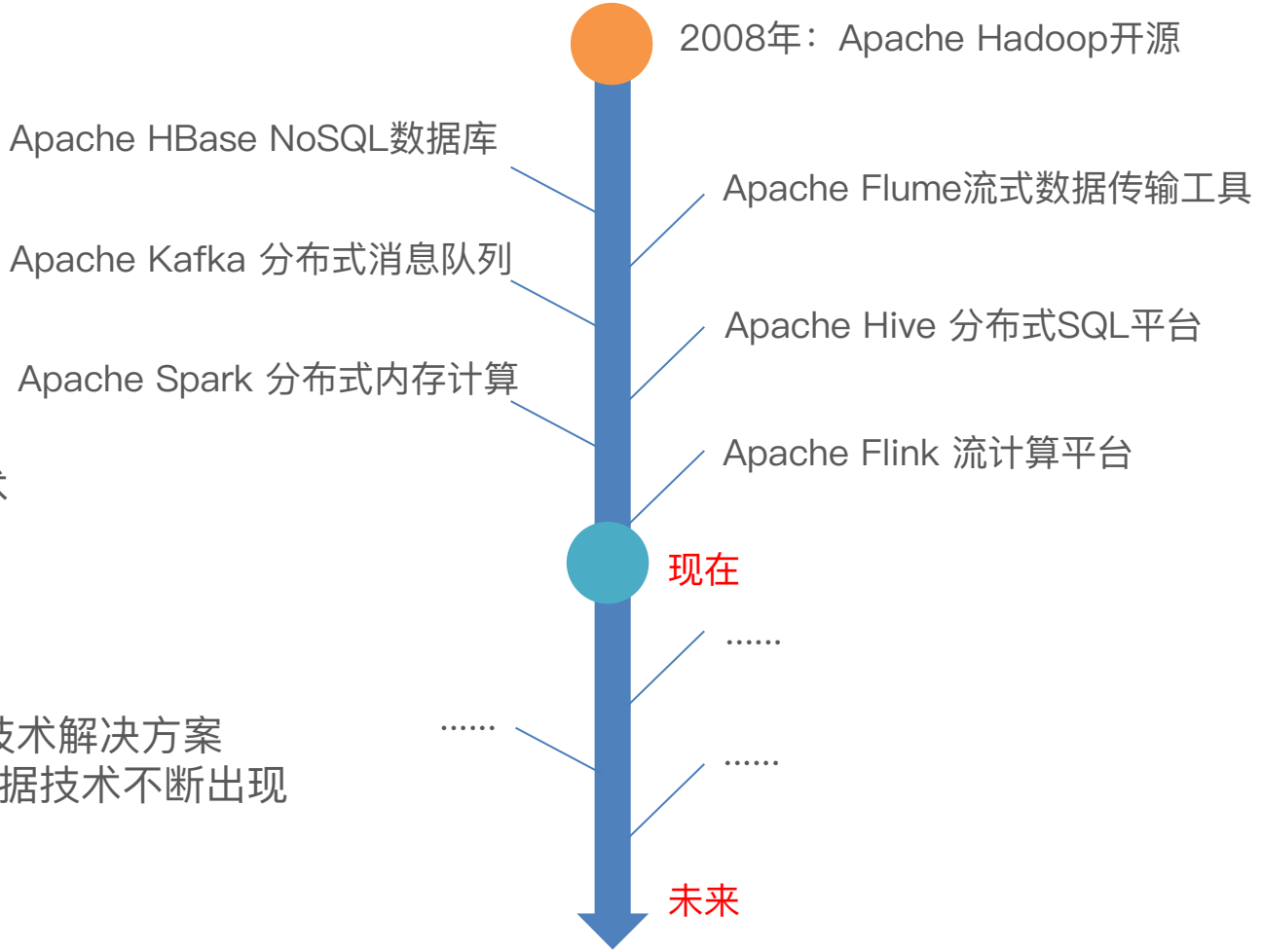
大数据的诞生

Apache Hadoop 是一款开源的分布式处理技术栈
为业界提供了

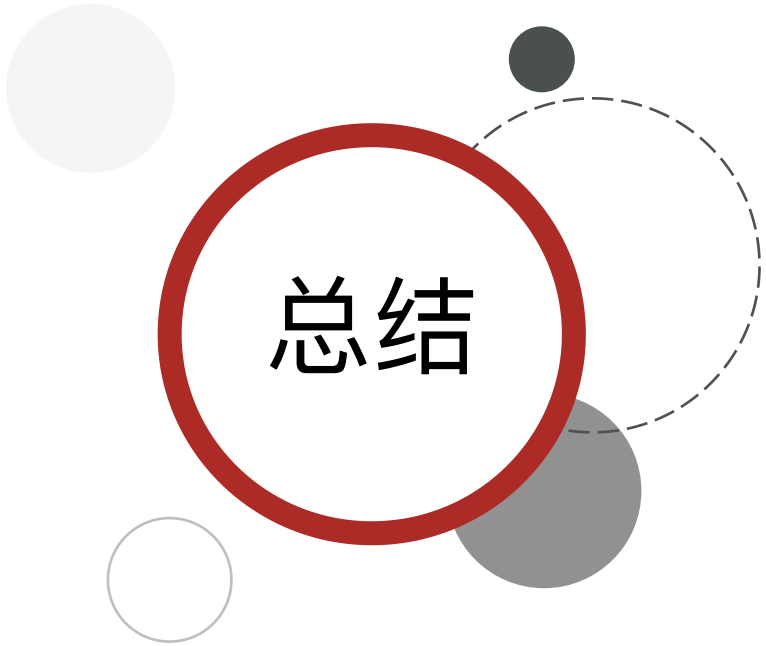
- 基于Hadoop HDFS的：分布式数据存储技术
- 基于Hadoop MapReduce的：分布式数据计算技术
- 基于Hadoop YARN的：分布式资源调度技术

Apache Hadoop的出现具有非常重大的意义：

- 为业界提供了”第一款”企业级开源大数据分布式技术解决方案
- 从Hadoop开始，大数据体系逐步建成，各类大数据技术不断出现



小结



1. 大数据的诞生是跟随着互联网的发展的

- 当全球互联网逐步建成（2000年左右），各大企业或政府单位拥有了海量的数据亟待处理。
- 基于这个前提逐步诞生了以分布式的形式（即多台服务器集群）完成海量数据处理的处理方式，并逐步发展成现代大数据体系。

2. Apache Hadoop对大数据体系的意义

- 第一款获得业界普遍认可的开源分布式解决方案
- 让各类企业都有可用的企业级开源分布式解决方案
- 一定程度上催生出了众多的大数据体系技术栈
- 从Hadoop开始（2008年左右）大数据开始蓬勃发展

目录

Contents

学习目标

Learning Objectives

1. 了解什么是大数据及其特征和核心工作内容

大数据概述

什么是大数据

通过大数据的诞生我们可以发现：大数据的出现，本质上是为了解决海量数据的处理难题。

大数据就是：使用分布式技术完成海量数据的处理，得到数据背后蕴含的价值。

狭义的（技术思维的）

广义的

什么是大数据

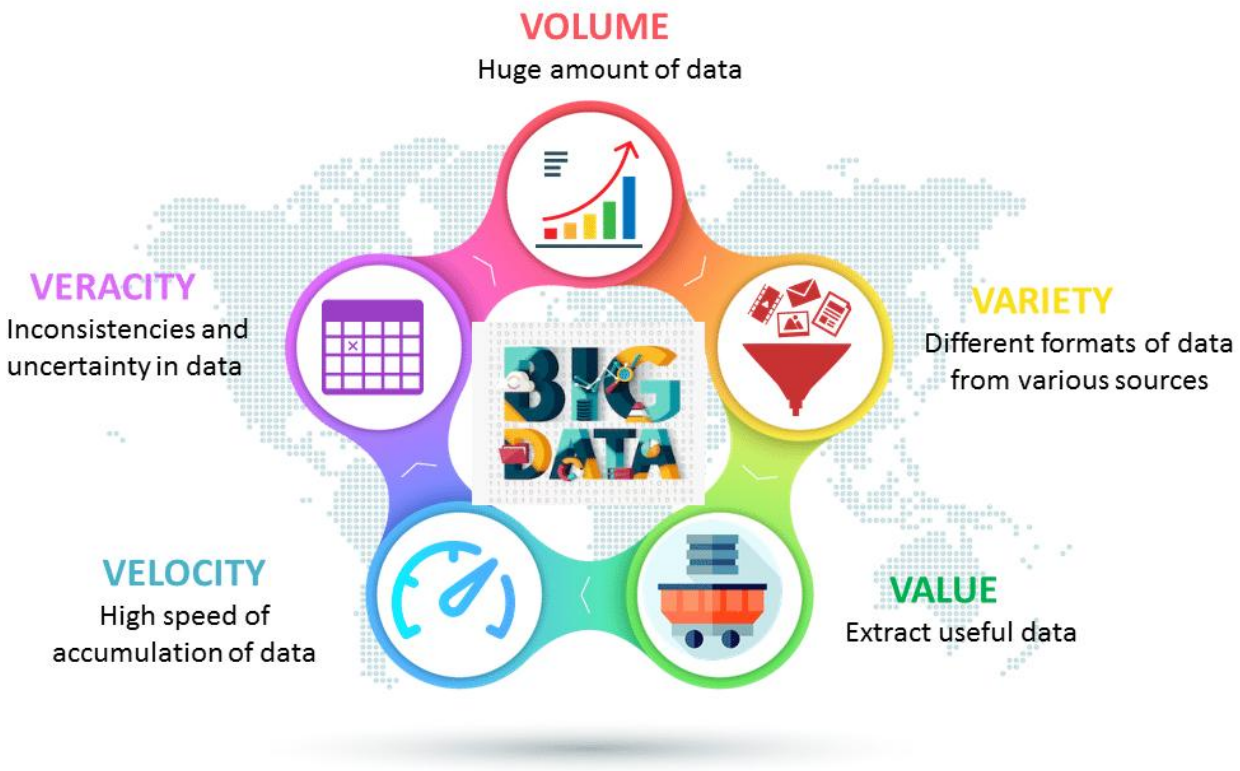
狭义上：大数据是一类技术栈，是一种用来处理海量数据的软件技术体系。

广义上：大数据是数字化时代、信息化时代的基础（技术）支撑，以数据为生活赋能。

大数据概述

大数据的特征

大数据有5个主要特征，称之为：5V特性



大数据概述

大数据的特征

Volume 体积

数据体量大

- 采集数据量大
- 存储数据量大
- 计算数据量大
- TB、PB级别起步

Variety 种类

种类、来源多样化

- 种类：结构化、半结构化、非结构化
- 来源：日志文本、图片、音频、视频

Value 价值

低价值密度

- 信息海量但是价值密度低
- 深度复杂的挖掘分析需要机器学习参与

Velocity 速度

速度快

- 数据增长速度快
- 获取数据速度快
- 数据处理速度快

Veracity 质量

数据的质量

- 数据的准确性
- 数据的可信赖度

大数据概述

大数据的特征

从海量的高增长、多类别、低信息密度的数据中挖掘出高质量的结果

大数据的核心工作

大数据的核心工作其实就是：从海量的高增长、多类别、低信息密度的数据中挖掘出高质量的结果

数据计算

数据存储

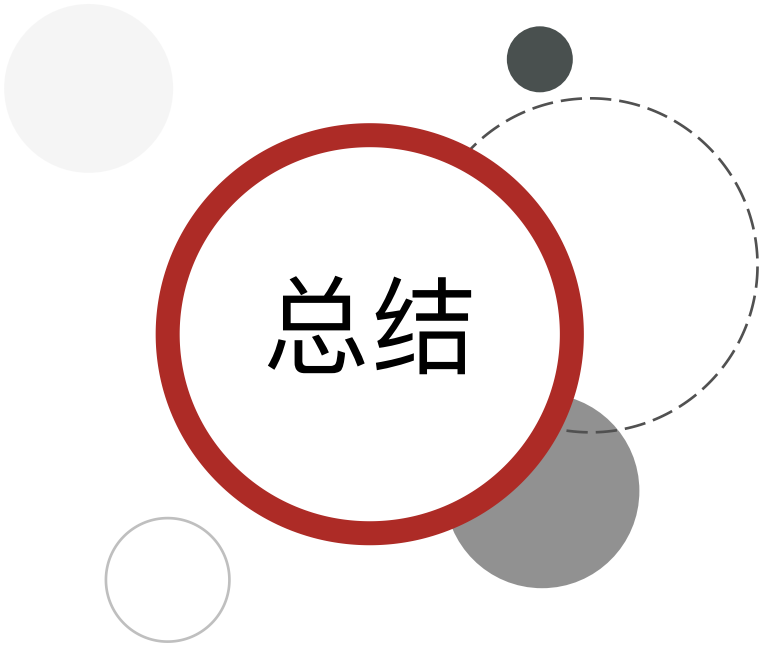
数据传输

大数据的核心工作

后续将学习的技术也是围绕着这三点来进行的，即：

- **分布式存储相关技术栈**
- **分布式计算相关技术栈**
- **海量数据传输相关技术栈**

小结



1. 什么是大数据

- 狭义上：对海量数据进行处理的软件技术体系
- 广义上：数字化、信息化时代的基础支撑，以数据为生活赋能

2. 大数据的5个主要特征

3. 大数据的核心工作：

- 存储：妥善保存海量待处理数据
- 计算：完成海量数据的价值挖掘
- 传输：协助各个环节的数据传输

目录

Contents

学习目标

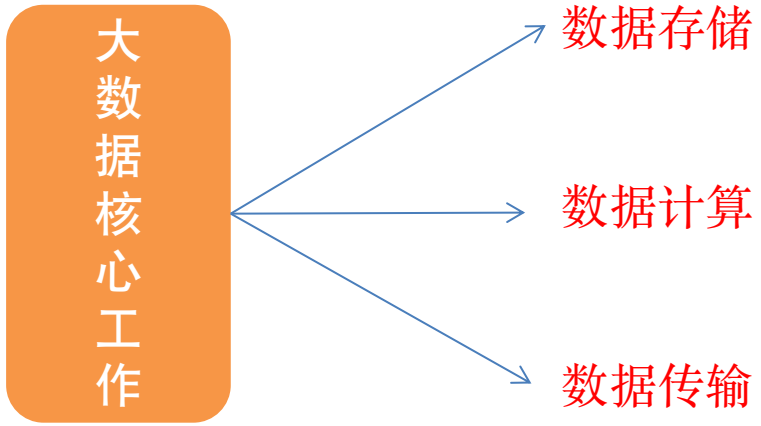
Learning Objectives

1. 了解常见的大数据技术框架及其负责的场景

大数据技术框架

大数据软件生态

大数据主体上分成如下三大核心工作体系。



大数据软件生态，也基本上围绕着三大工作体系

大数据技术框架

大数据软件生态

数据存储

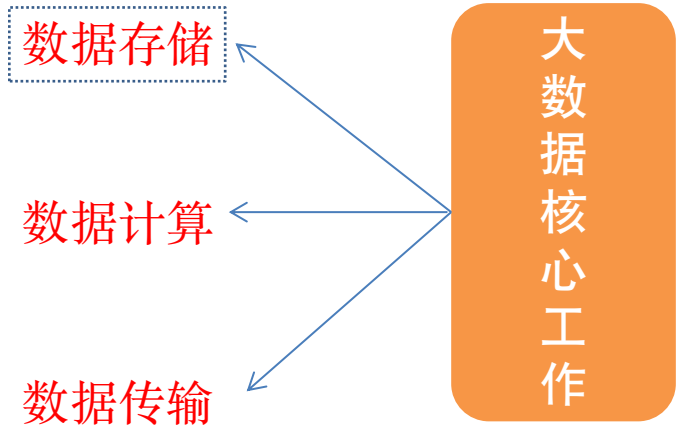
- Apache Hadoop – HDFS

Apache Hadoop框架内的组件HDFS是大数据体系中使用最为广泛的分布式存储技术

- Apache HBase

Apache HBase是大数据体系内使用非常广泛的NoSQL KV型数据库技术

HBase是基于HDFS之上构建的。



大数据技术框架

大数据软件生态

数据存储

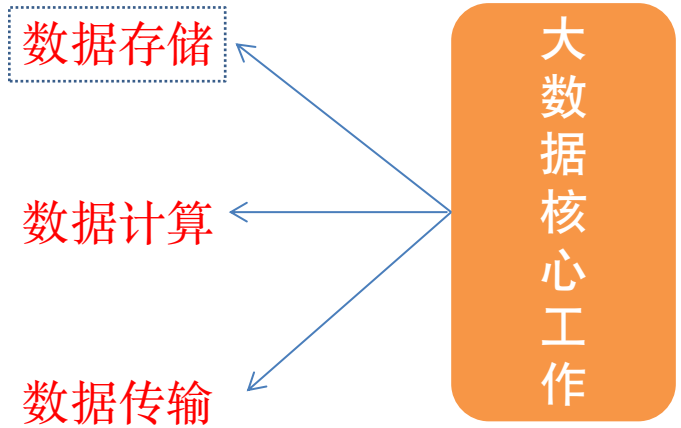
- Apache KUDU

Apache Kudu同样为大数据体系中使用较多的分布式存储引擎

- 云平台存储组件

除此以外，各大云平台厂商也有相应的大数据存储组件，如

阿里云的OSS、UCloud的US3、AWS的S3、金山云的KS3等等



大数据技术框架

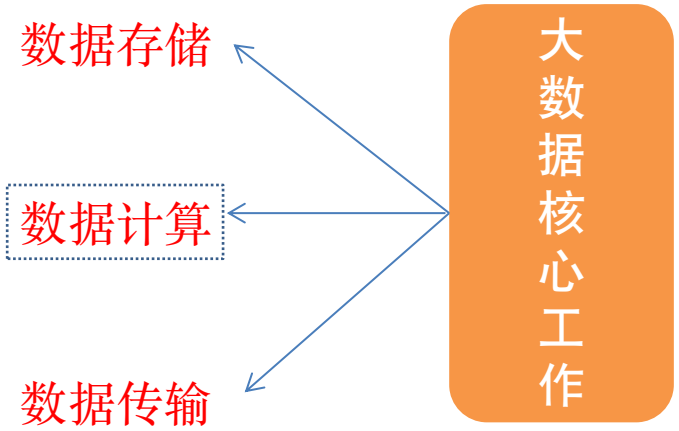
大数据软件生态

数据计算

- Apache Hadoop – MapReduce

Apache Hadoop的MapReduce组件是最早一代的大数据分布式计算引擎

对大数据的发展做出了卓越的贡献



- Apache Hive

Apache Hive是一款以SQL为要开发语言的分布式计算框架。其底层使用了Hadoop的MapReduce技术

Apache Hive至今仍活跃在大数据一线，被许多公司使用。

大数据技术框架

大数据软件生态

数据计算

- Apache Spark

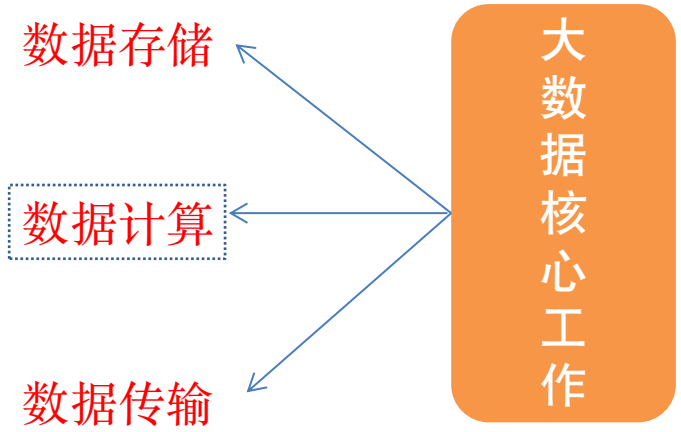
Apache Spark是目前全球范围内最火热的分布式内存计算引擎。

是大数据体系中的明星计算产品

- Apache Flink

Apache Flink同样也是一款明星级的大数据分布式内存计算引擎。

特别是在实时计算（流计算）领域，Flink占据了大多数的国内市场。



大数据技术框架

大数据软件生态

数据传输

- Apache Kafka

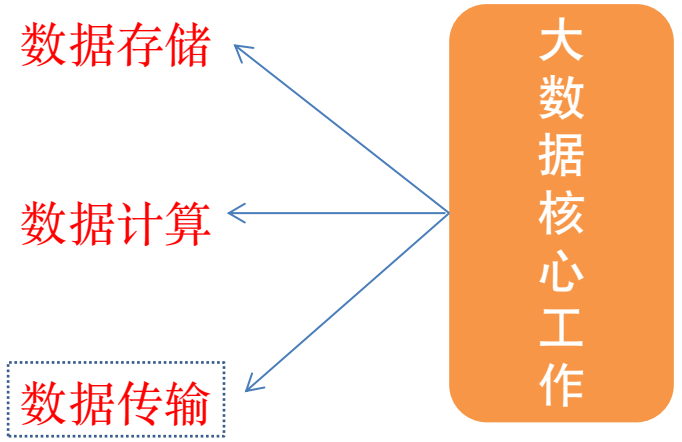
Apache Kafka是一款分布式的消息系统，可以完成海量规模的数据传输工作。

Apache Kafka在大数据领域也是明星产品

- Apache Pulsar

Apache Pulsar同样是一款分布式的消息系统。

在大数据领域同样有非常多的使用者。



大数据技术框架

大数据软件生态

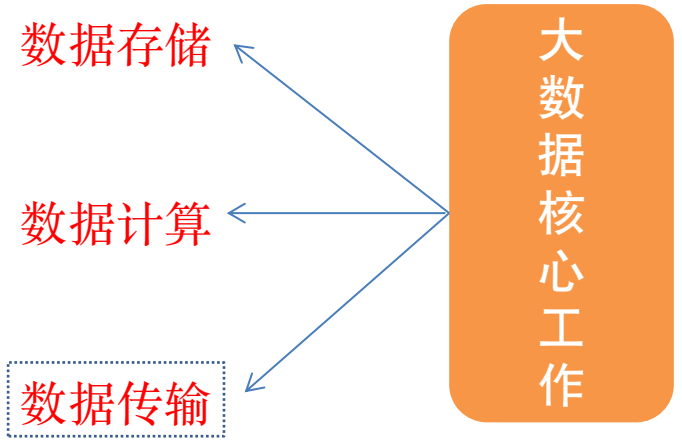
数据传输

- Apache Flume

Apache Flume是一款流式数据采集工具，可以从非常多的数据源中完成数据采集传输的任务。

- Apache Sqoop

Apache Sqoop是一款ETL工具，可以协助大数据体系和关系型数据库之间进行数据传输。

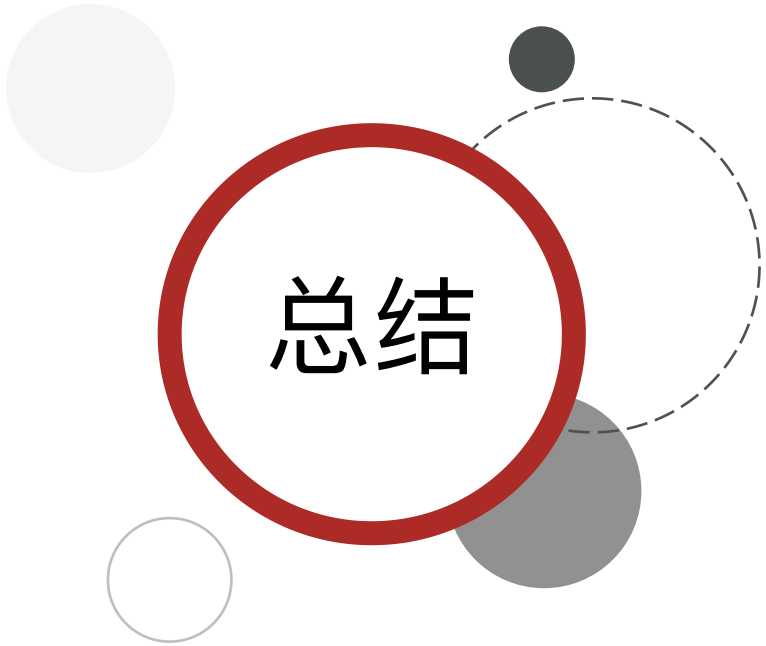


大数据技术框架

大数据软件生态

大数据体系内的软件种类还是非常多的。在后续的学习中我们都能够逐步的接触到它们。

小结



1. 大数据的核心工作：

- 存储：妥善保存海量待处理数据
- 计算：完成海量数据的价值挖掘
- 传输：协助各个环节的数据传输

2. 大数据软件生态

- 存储：Apache Hadoop HDFS、Apache HBase、Apache Kudu、云平台
- 计算：Apache Hadoop MapReduce、Apache Spark、Apache Flink
- 传输：Apache Kafka、Apache Pulsar、Apache Flume、Apache Sqoop

小结

大数据软件生态中，软件的名称都有：Apache

Apache是什么？

目录

Contents

学习目标

Learning Objectives

1. 了解Apache Hadoop框架
2. 了解Hadoop的发展

Apache Hadoop 概述

什么是Hadoop

Hadoop是Apache软件基金会下的顶级开源项目，用以提供：

- 分布式数据存储
- 分布式数据计算
- 分布式资源调度

为一体的整体解决方案。

Apache Hadoop是典型的分布式软件框架，可以部署在1台乃至成千上万台服务器节点上协同工作。

个人或企业可以借助Hadoop构建大规模服务器集群，完成海量数据的存储和计算。

Apache Hadoop 概述

为什么学习Hadoop

近10年来，大数据技术体系一词一直和Hadoop是划上等号的，提起大数据技术基本就是在提及Hadoop。

随着近些年的发展，越来越多的新技术框架的出现，给大数据技术体系带来了丰富的生态，但是拥有元老地位的Hadoop依旧非常重要。

为什么学习Hadoop有如下几个至关重要的原因：

- Hadoop是最早的一批大数据技术框架，在市面上拥有极高的占有率和庞大的用户群体。
- Hadoop在大数据体系内，技术难度相对较低，非常适合作为大数据学习的入门技术栈。

所以，学习Hadoop不仅仅因为其适合入门，同时也可以为大数据学习打下良好的基础。

Apache Hadoop 概述

Hadoop的功能

通常意义上，Hadoop是一个整体，其内部还会细分为三个功能组件，分别是：

- **HDFS组件** → HDFS是Hadoop内的分布式存储组件 可以构建分布式文件系统用于数据存储
- **MapReduce组件** → MapReduce是Hadoop内分布式计算组件。提供编程接口供用户开发分布式计算程序
- **YARN组件** → YARN是Hadoop内分布式资源调度组件。可供用户整体调度大规模集群的资源使用。

所以，我们会说Hadoop是一个集合了：**存储、计算、资源调度**为一体的大数据分布式框架

Apache Hadoop 概述

Hadoop发展

- **Hadoop创始人：Doug Cutting**
- **Hadoop起源于Apache Lucene子项目：Nutch**

  Nutch的设计目标是构建一个大型的全网搜索引擎。

  遇到瓶颈：如何解决数十亿网页的存储和索引问题

- **Google三篇论文**

  《The Google file system》：谷歌分布式文件系统GFS

  《MapReduce: Simplified Data Processing on Large Clusters》：谷歌分布式计算框架MapReduce

  《Bigtable: A Distributed Storage System for Structured Data》：谷歌结构化数据存储系统

Apache Hadoop 概述

Hadoop发展

2002

Nutch

开源搜索引擎框架雏形诞生

2004

Google论文

Google发布了《GFS》《MapReduce》相关分布式存储和计算论文

2006

Hadoop雏形

基于Google论文的内容，使用Java完成了Hadoop框架的初版开发

2006

Yahoo时代

创始人Doug Cutting加入雅虎公司，在雅虎的支持下继续开发Hadoop项目

2008

开源

Hadoop贡献给Apache软件基金会，成为Apache旗下顶级开源项目

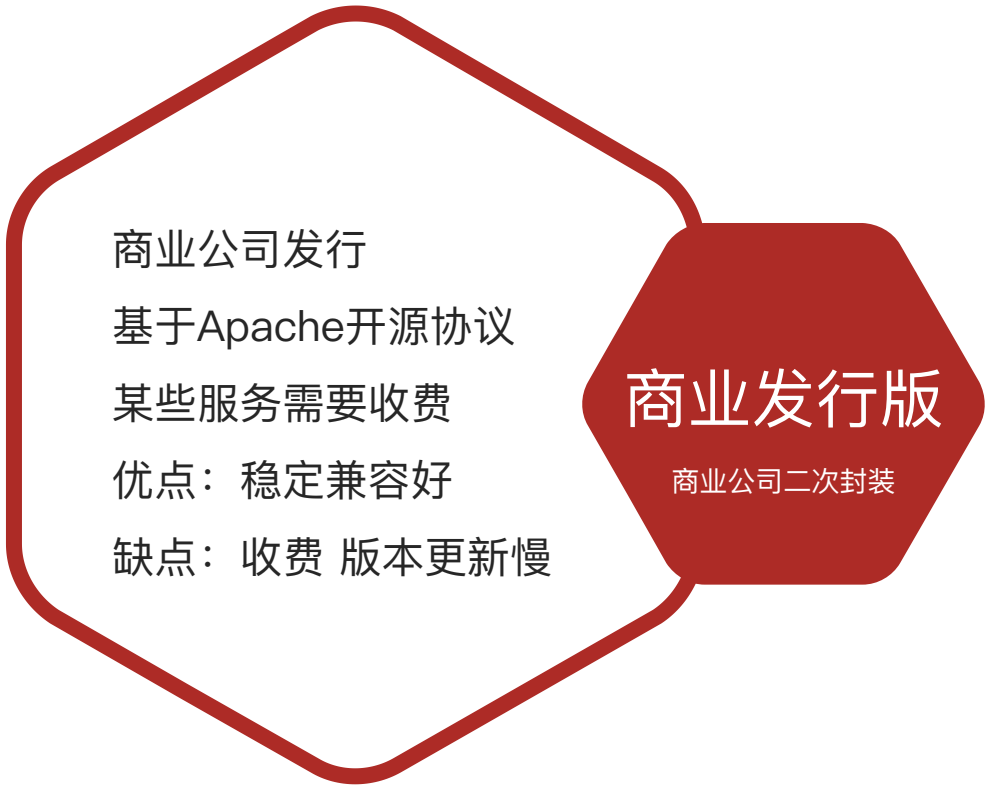
至今

稳步发展

从2008年开源以来到现在，Hadoop已更新到3.x版本，成为市面上知名的大数据框架

Hadoop发行版本


开源社区版
原生Hadoop体系
Apache开源社区发行
也是官方发行版本
优点：更新迭代快
缺点：兼容稳定性不周


商业公司发行
基于Apache开源协议
某些服务需要收费
优点：稳定兼容好
缺点：收费 版本更新慢
商业发行版
商业公司二次封装

Apache Hadoop 概述

Hadoop发行版本

Apache开源社区版本

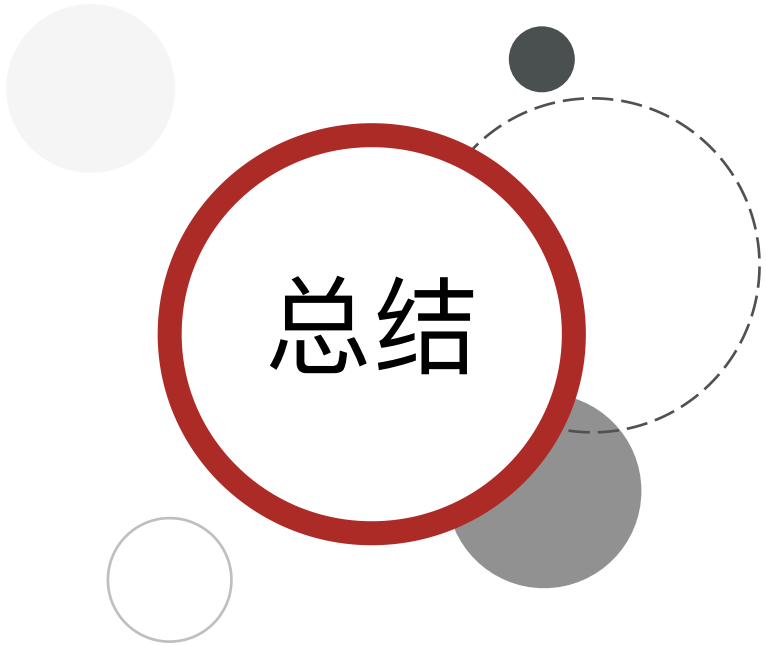
http://hadoop.apache.org/

商业发行版本

- CDH（Cloudera's Distribution, including Apache Hadoop） Cloudera公司出品，目前使用最多的商业版
- HDP（Hortonworks Data Platform），Hortonworks公司出品，目前被Cloudera收购
- 星环，国产商业版，星环公司出品，在国内政企使用较多

本课程中使用的是当前最新的Apache Hadoop（即开源版本），版本号为：**3.3.4**

小结



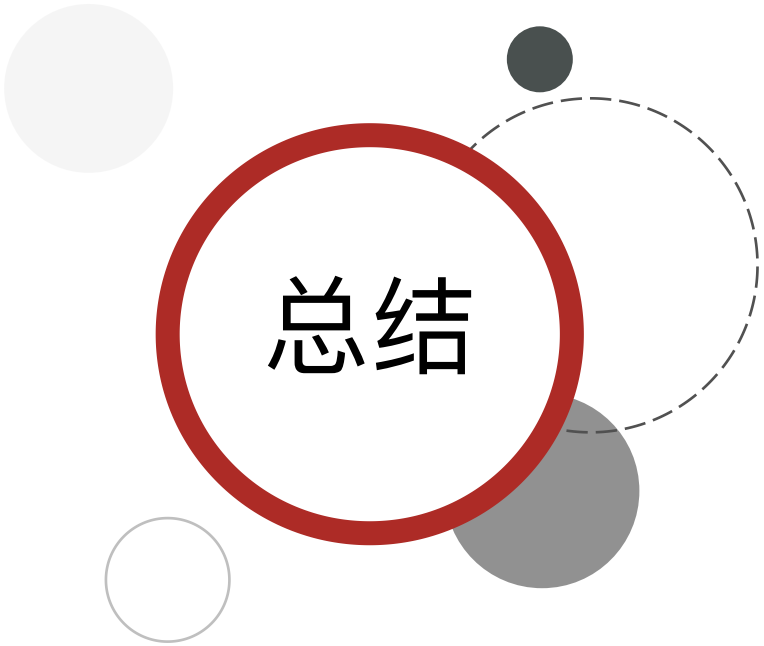
1. 什么是Hadoop

Hadoop是开源的技术框架，提供分布式存储、计算、资源调度的解决方案

2. 为什么学习Hadoop

- Hadoop诞生早，在企业中广泛应用
- Hadoop概念较为简单，适合大数据分布式入门

小结



3. Hadoop的发展

- 创始人Doug Cutting
- 基于Nutch搜索项目发展
- 发展受到Google三篇著名的论文影响

《The Google file system》：谷歌分布式文件系统GFS

《MapReduce: Simplified Data Processing on Large Clusters》：谷歌分布式计算框架MapReduce

《Bigtable: A Distributed Storage System for Structured Data》：谷歌结构化数据存储系统

4. Hadoop的版本

- Apache 开源社区版Hadoop（原生版本）
- 商业公司自行封装的商业版
  - •CDH
  - •HDP
  - •星环

目录

Contents

第六章

大数据与分布式（集群准备）

授课教师：姜婧妍

jiangjingyan@sztu.edu.cn

2023年