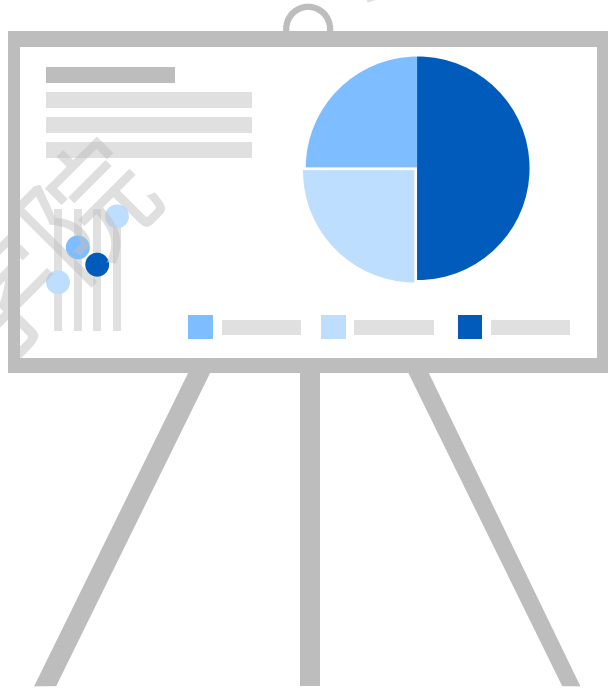
大数据存储与处理

大数据概述

目录



Part1.
大数据概述

数据的分类

结构化数据：包括预定义的数据类型、 格式和结构的数据，常见的比如关系型数据库中数据表里的数据 。

Id	Name	Age	Gender
1	lyh	12	male
2	liangyh	13	female
3	liang	18	male

半结构化数据：具有可识别的模式并可以解析的文本数据文件， 比如XML数据文件 。

```
<person>
  <name>A</name>
  <age>13</age>
  <gender>female</gender>
</person>
```

非结构化数据：没有固定结构的数据， 通常保存为不同类型的文件， 比如文本文档、 图片、 视频等。

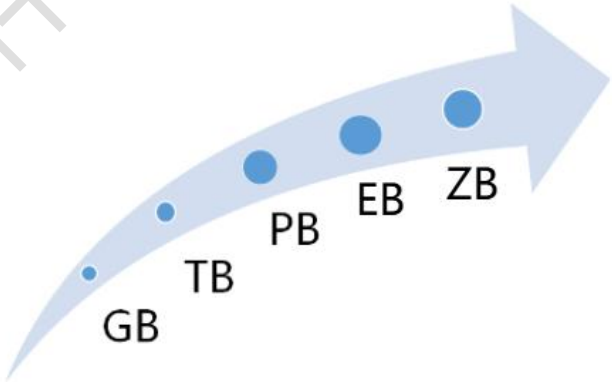
数据大爆炸

大数据摩尔定律 → **（当前大数据时代，数据发展的独特定律）**

- 根据国际组织IDC发布报告，人类社会数据每年50%的速度，**每两年就增长一倍。**
- 近两年产生的数据量之和相当于我们人类发展历史上所有的数据量之和。

International Data Corporation
国际数据公司

1ZB = 1000EB
1EB = 1000PB
1PB = 1000TB
1TB = 1000GB



数据生命周期管理

1 **分类**

对数据进行自动分类，分离出有效的数据，对不同类型数据制定不同的管理策略，并及时清理无用的数据

2 **存储**

构建分层的存储系统，满足不同类型的数据对不同生命周期阶段的存储要求，对关键数据进行数据备份保护，对处于生命周期末期的数据进行归档并保存到适合长期保存数据的存储设备中

3 **管理**

根据不同的数据管理策略，实施自动分层数据管理，即自动把不同生命周期阶段的数据存放在最合适的存储设备上，提高数据可用性和管理效率

大数据的概念

IDC组织：大数据是通过传统的数据管理工具难以处理的海量、复杂和多样化的数据集，通常需要新型的数据处理技术和分析工具来捕获、存储、管理和分析。

McKinsey & Company

麦肯锡公司：大数据指那些具有极大体量、快速生成且多样的数据，能够为企业和组织提供重要洞察，从而驱动业务决策和创新。

IBM公司：大数据是数据体量巨大，生成速度极快，并且数据类型多样。它包括来自各种来源的数据，能够为决策过程提供洞察，推动业务改进。

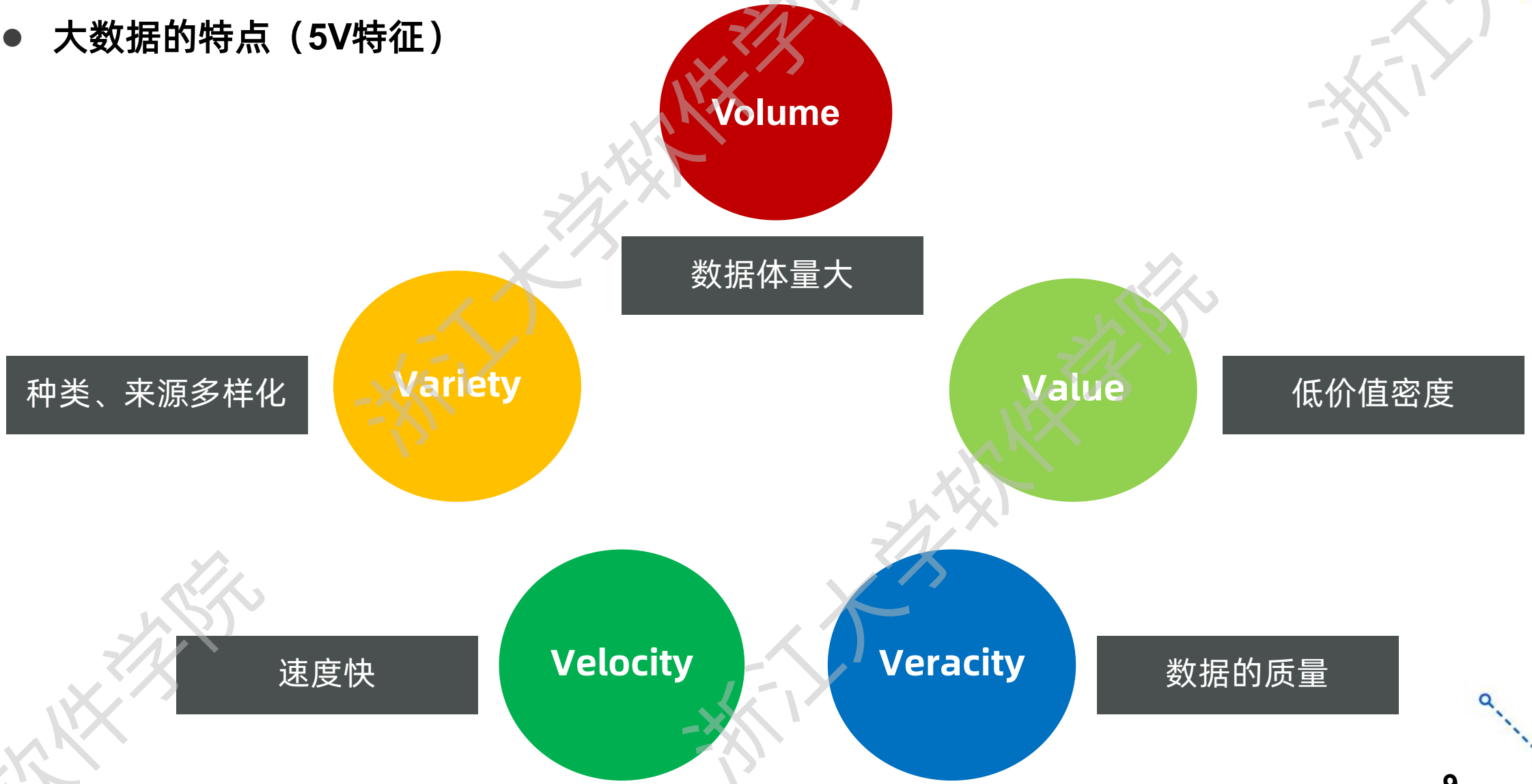
大数据的概念

通常，大数据（Big Data）指那些：

- 无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合
- 海量、高增长率和多样化的信息资产
- 需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力

大数据的概念

- **大数据的特点（5V特征）**



大数据应用场景

电商领域

精准广告位、个性化推荐、大数据杀熟

传媒领域

精准营销、猜你喜欢、交互推荐

金融领域

信用评估、风险管控、客户细分、精细化营销

大数据应用场景

交通领域

拥堵预测、智能红绿灯、导航最优规划

电信领域

基站选址优化、舆情监控、客户用户画像

安防领域

犯罪预防、天网监控

医疗领域

智慧医疗、疾病预防、病源追踪

大数据发展的阶段

阶段	时间	内容
第一阶段：萌芽期	上世纪90年代 至本世纪初	随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术开始被应用，如数据仓库、专家系统、知识管理系统等
第二阶段：成熟期	本世纪前十年	Web2.0应用迅猛发展，非结构化数据大量产生，传统处理方法难以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统两大核心技术，谷歌的GFS和MapReduce等大数据技术受到追捧，Hadoop平台开始大行其道
第三阶段：大规模应用期	2010年以后	大数据应用渗透各行各业，数据驱动决策，信息社会智能化程度大幅提高

Part2.
大数据存储技术

I. 数据规模与处理能力

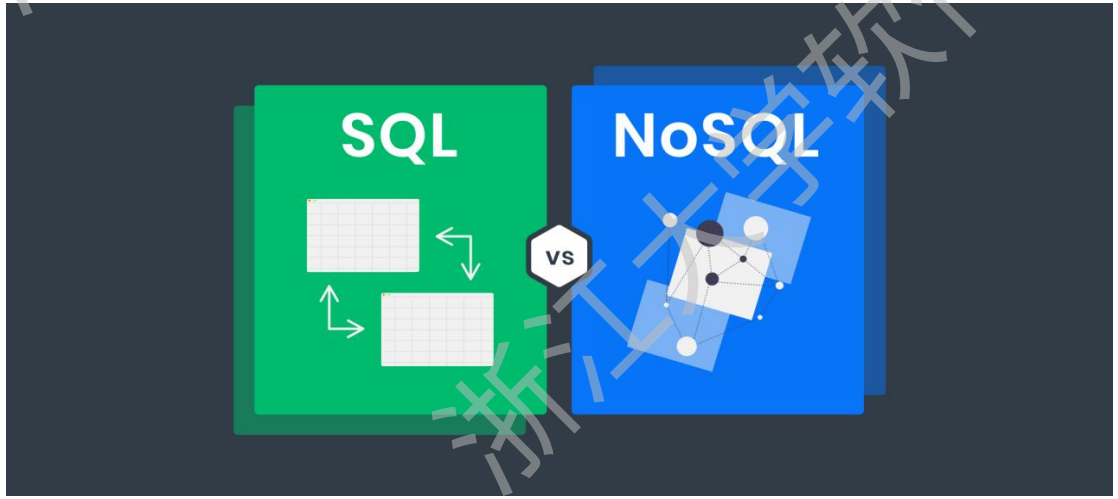
传统数据存储主要适用于较小规模的数据集，通常可以依赖关系型数据库管理系统（RDBMS）进行处理。这些数据集往往是结构化的，规模相对较小，可以方便地利用数据库的分析工具进行处理。

大数据存储则专为处理海量数据而设计，从几TB到几PB甚至更多的数据都能轻松应对。大数据存储技术需要具备高性能的读写能力，以支持对庞大数据集的高效访问和处理。

II. 数据类型与结构

传统数据存储主要管理结构化数据，这类数据具有固定的格式和预定义的模式，如关系型数据库中的表格。

大数据存储技术则更注重处理非结构化和半结构化数据，例如文本文件、图像、视频等。这些数据类型缺乏固定的格式和预定义模式，因此需要更为灵活的存储和处理机制。



III. 存储架构与扩展性

传统数据存储通常依赖于专用的硬件设备，其扩展性可能受限于硬件的物理容量和性能。当数据量增长时，可能需要更多的硬件投入来满足存储需求。

大数据存储采用分布式存储架构，数据被分散存储在多个独立的服务器上。这种架构提供了极高的可扩展性，能够根据数据量的增长动态地扩展存储容量，而无需进行大规模的硬件升级。

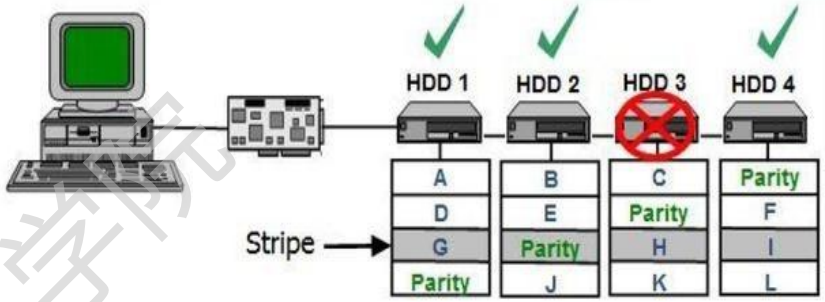
IV. 成本与效率

传统数据存储的初始成本可能较低，但随着数据量的增长，硬件和软件的升级成本可能会显著增加。同时，处理大规模数据时可能会遇到性能瓶颈。

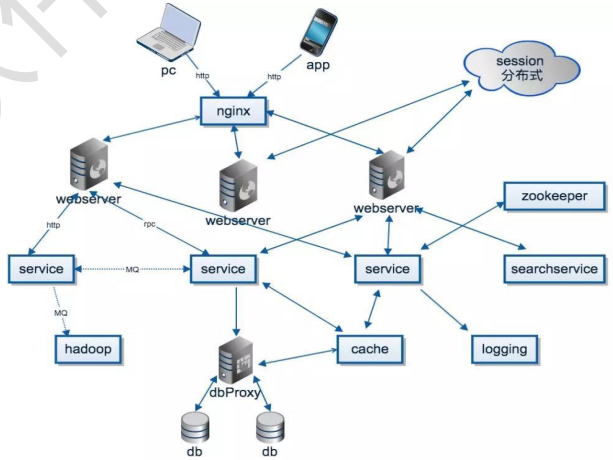
大数据存储技术通常以低成本的硬件和开源软件为基础，能够降低存储成本并提高存储效率。分布式计算模型允许并行处理和分布式存储，从而实现了高性能的数据处理和分析。

V. 数据安全性与可靠性

传统数据存储通常具有较高的数据安全性，数据备份和恢复策略也相对独立，有助于保障数据的完整性。



大数据存储技术同样重视数据的安全性和可靠性。通过采用分布式存储和冗余备份等策略，大数据存储系统能够在硬件故障时保证数据的完整性和可用性。



分布式技术概述

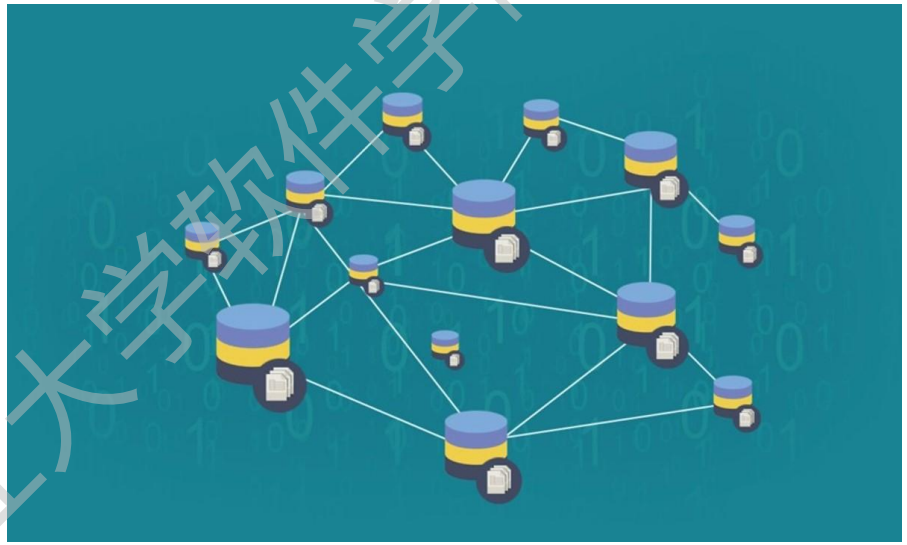
- **为什么需要分布式技术**
  - 数据存储要求的更新：传统存储迈向大数据存储
  - 应用和系统架构的变迁：单机单一架构迈向多机分布式架构

如何存储?

单机存储有瓶颈→多台机器分布式存储

如何计算?

单机计算能力有限→多台机器分布式计算



分布式系统概述

1. 分布式系统是一个硬件或软件组件分布在不同的网络计算机上
2. 彼此之间仅仅通过消息传递进行通信和协调的系统
3. 一群互相独立计算机集合共同对外提供服务
4. 对于系统的用户来说，就像是一台计算机在提供服务

分布式技术关键特性

- **负载均衡（Load Balance）**

将负载（工作任务）进行平衡、分摊到多个操作单元上进行运行，解决了单个无法处理所有任务，多个一起处理的问题

- **故障转移（Fail Over）**

当活动的服务或应用意外终止时，快速启用冗余或备用的服务器、系统、硬件或者网络接替它们工作。故障转移的核心是设置备份、出现故障时主备切换，而主备切换的前提是数据状态保持一致

- **伸缩性（Scalability）**

伸缩性也叫做弹性，可扩展性。指系统可以根据需求动态的扩容、缩容

分布式文件系统

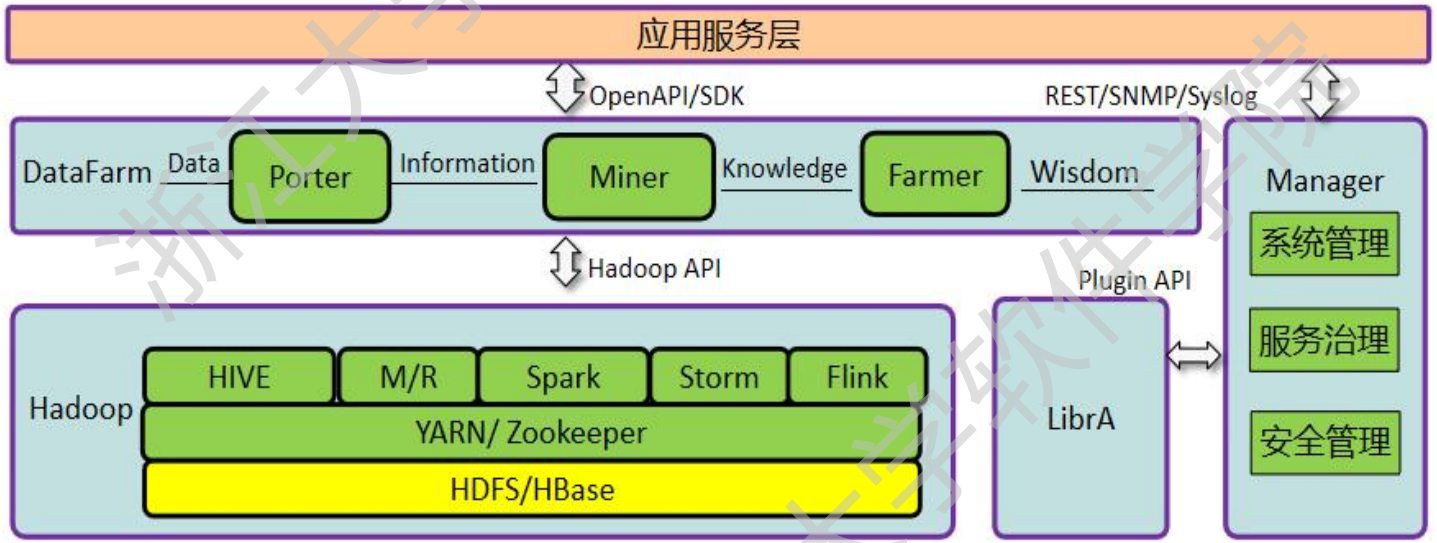
- 分布式文件系统（Distributed File System，DFS）是一种特殊的文件系统，其管理的物理存储资源并非直接连接在本地节点上，而是通过计算机网络与多个节点相连。

特点：

1. 可扩展性强：支持随时对数据服务器进行扩展，从而提升存储容量和访问带宽。这种扩展能力允许分布式文件系统轻松应对数据量的不断增长。
2. 统一命名空间：客户端看到的是一个统一的全局命名空间，使得用户操作起来就像是管理本地文件系统一样。
3. 高性能：由于文件被分成多个部分并保存在不同的数据服务器上，因此可以同时进行读取，从而提高访问性能。
4. 高可用性：分布式文件系统具有高容错能力，单点失效被有效避免，例如通过资源冗余技术或提供失效恢复服务，确保单个数据节点的故障不会影响整个集群的运行。
5. 弹性存储：业务需要可以灵活地增加或缩减数据存储，而不需要中断系统运行。

分布式文件系统

- **HDFS**(Hadoop Distributed File System)作为分布式文件系统的代表，基于Google发布的GFS论文设计开发。HDFS是Hadoop技术框架中的分布式文件系统，对部署在多台独立物理机器上的文件进行管理。



- 高容错性：认为硬件总是不可靠的。
- 高吞吐量：为大量数据访问的应用提供高吞吐量支持。
- 大文件存储：支持存储TB-PB级别的数据。

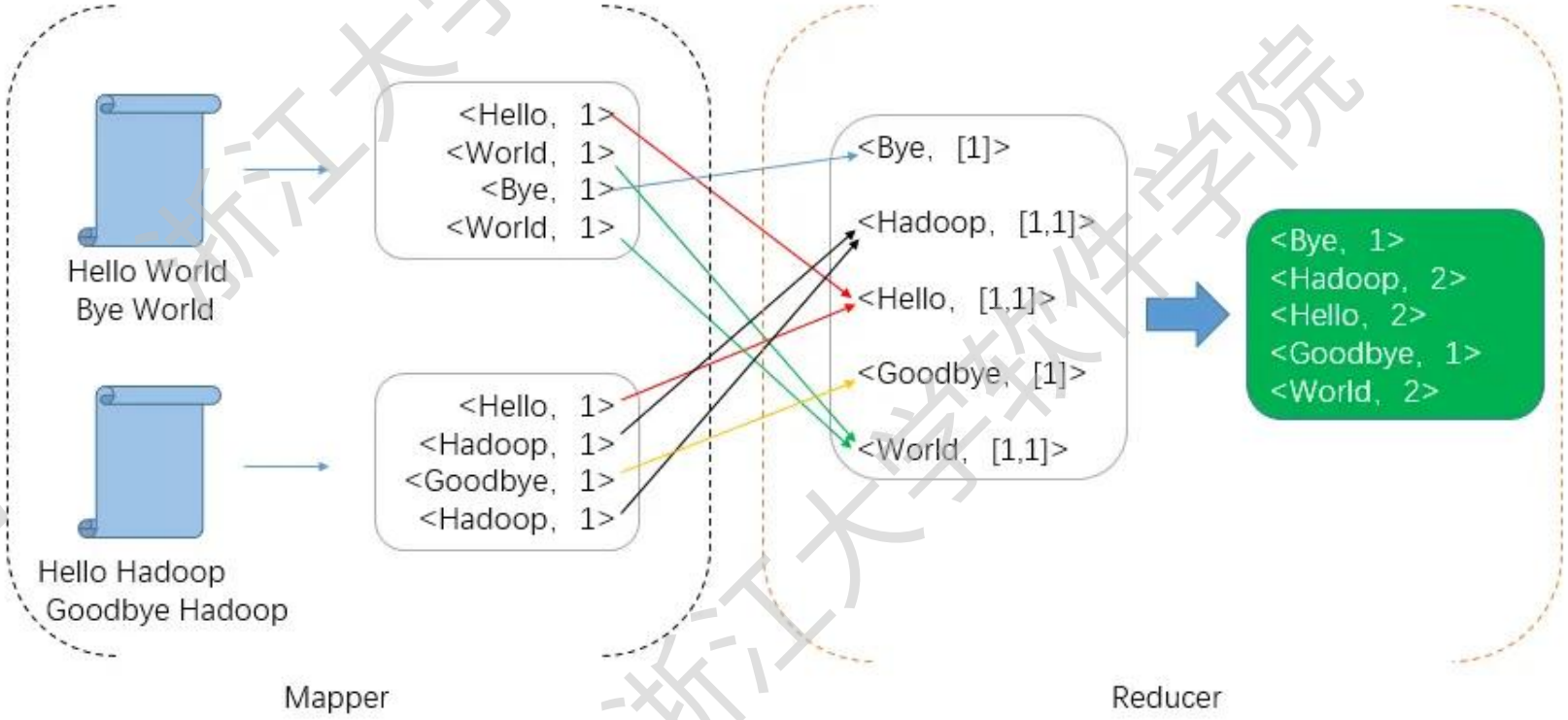
Part3.
大数据处理技术

批处理与流处理

	批处理（Batch Processing）	流处理（Stream Processing）
数据处理时机	据积累到一定量之后进行	数据一产生或到达系统就立即处理
数据处理方式	数据被视为一批静态的记录集合，处理过程通常是一次性的，处理完整个数据集后，任务结束。	数据被视为不断流动的数据流，系统持续不断地处理这些数据流。
适用场景	适合于不需要即时响应的场景，如日志分析、大规模数据集的ETL（Extract, Transform, Load）操作、复杂的数据转换和计算等。	适合于需要实时或近实时响应的场景，如实时监控、实时分析、在线推荐系统、实时欺诈检测等。
优点	可以优化处理过程、适合处理大量数据；通常容错性较好	能够实时处理数据，适应性强；可以快速响应事件
缺点	无法处理实时数据；对于数据处理有延迟；不适合需要快速响应的应用。	可能需要更复杂的系统设计来保证数据的准确性和处理的可靠性；对于系统资源的要求可能更高。

MapReduce编程模型

MapReduce编程模型是大数据处理技术中的一种核心工具，其设计理念和实现方式与大数据的处理需求紧密相关。它由 Google 的 Jeffrey Dean 和 Sanjay Ghemawat 在 2004 年提出，并广泛用于分布式计算环境中。MapReduce 模型主要包括两个核心操作：Map（映射）和 Reduce（归约）。



核心思想：分而治之

大数据处理框架

Hadoop是目前世界上使用最广泛的大数据工具。具有良好的跨平台性，并且可部署在廉价的计算机集群中，在业内应用非常广泛，是分布式计算架构的鼻祖。

Apache Spark 是一种用于大数据工作负载的分布式开源处理系统。它使用内存中缓存和优化的查询执行方式，可针对任何规模的数据进行快速分析查询。

Apache Flink 是一个框架和分布式处理引擎，用于在无边界和有边界数据流上进行有状态的计算。Flink 能在所有常见集群环境中运行，并能以内存速度和任意规模进行计算。

三大框架对比

	Hadoop	Spark	Flink
计算模型	MapReduce采用了面向批处理的模型，批处理静态数据。	Spark采用了微批处理。微批处理本质上是一种“先收集再处理”的计算模型。	Flink采用连续流式流传输模型，实时对数据进行处理，而不会在收集数据或处理数据时出现任何延迟。
性能	Hadoop仅支持批处理，不支持处理流数据，与Spark和Flink相比，性能会降低。	支持微批处理，但流处理效率不如Apache Flink。	Flink使用本机闭环迭代运算符，尤其在支持机器学习和图形处理方面，表现优异。
内存管理	提供可配置的内存管理，可以动态或静态地执行此操作。	提供可配置的内存管理，从Spark 1.6开始已朝着自动进行内存管理的方向发展。	有自己的内存管理系统，提供自动内存管理。

三大框架对比

	Hadoop	Spark	Flink
数据处理	专为批处理而生，一次将大量数据集输入到输入中，进行处理并产生结果。	定义是一个批处理系统，但也支持流处理。	为流和批处理提供了一个运行时，基于每个事件处理，支持毫秒级计算。
流引擎支持	Hadoop默认的MapReduce，仅面向于批处理。	Spark Streaming以微批处理数据流，实现准实时的批处理和流处理。	Flink是真正的流引擎，使用流来处理工作负载，包括流、SQL、微批处理和批处理。
数据流	MapReduce计算数据流没有任何循环，每个阶段使用上一阶段的输出，并为下一阶段产生输入。	尽管机器学习算法是循环数据流，但Spark将其表示为（DAG）直接非循环图或有向无环图。	Flink在运行时支持受控循环依赖图，支持机器学习算法非常有效。

Part4.
大数据挑战与未来趋势

挑战：企业内部数据孤岛严重

- 在很多企业中尤其是大型的企业，数据常常散落在不同部门，而且这些数据存在不同的数据仓库中，不同部门的数据技术也有可能不一样，这导致企业内部自己的数据都没法打通。

- 大数据需要不同数据的关联和整合才能更好的发挥理解客户和理解业务的优势。如何将不同部门的数据打通，并且实现技术和工具共享，才能更好的发挥企业大数据的价值。

挑战：数据可用性低，数据质量差

- 很多企业在大数据的预处理阶段很不重视，导致数据处理很不规范。甚至很多企业在数据的上报就出现很多不规范不合理的情况。以上种种原因，导致企业的数据的可用性差，数据质量差，数据不准确。

- Sybase的数据表明，高质量的数据的数据应用可以显著提升企业的商业表现，数据可用性提高10%，企业的业绩至少提升在10%以上。

挑战：数据安全

- 网络化生活使得犯罪分子更容易获得关于人的信息，也有了更多不易被追踪和防范的犯罪手段，可能会出现更高明的骗局。如何保证用户的信息安全成为大数据时代非常重要的课题。

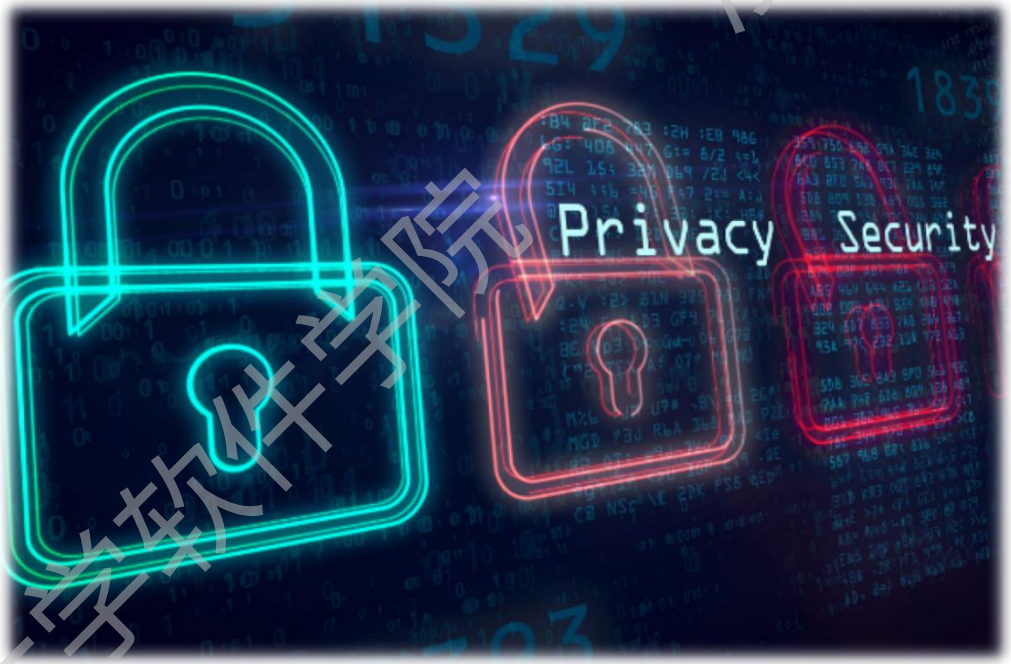
- 大数据的不断增加，对数据存储的物理安全性要求会越来越高，从而对数据的多副本与容灾机制也提出更高的要求。目前很多传统企业的数据安全令人担忧。

挑战：数据开放与隐私的权衡

- 由于政府、企业和行业信息化系统建设往往缺少统一规划，系统之间缺乏统一的标准，形成了众多“信息孤岛”，而且受行政垄断和商业利益所限，数据开放程度较低，这给数据利用造成极大障碍。

- 另外一个制约我国数据资源开放和共享的一个重要因素是政策法规不完善，大数据挖掘缺乏相应的立法。无法既保证共享又防止滥用。

- 同时，开放与隐私如何平衡，也是大数据开放过程中面临的最大难题。如何在推动数据全面开放、应用和共享的同时有效地保护公民、企业隐私，逐步加强隐私立法，将是大数据时代的一个重大挑战。



大数据未来趋势

- 趋势一：数据资源化，将成为最有价值的资产
- 趋势二：大数据在更多的传统行业的企业管理落地
- 趋势三：大数据和传统商业智能融合，行业定制化解决方案将涌现
- 趋势四：数据将越来越开放，数据共享联盟将出现
- 趋势五：大数据安全越来越受重视，大数据安全市场将愈发重要

Part5.
宁波工业大数据发展

在甬背景

- 大数据时代，数据成为一种生产资料。任何行业和领域都会产生有价值的数据，而对这些数据的统计、分析、挖掘，则会创造意想不到的价值和财富。

- 宁波是全国首个“中国制造2025”试点示范城市，也是第一批中小企业数字化转型试点城市，对掌握工业数字化技术的人才具有强烈的需求。在宁波市有大量的实际工业场景的数据处理和数据应用需求。

- 目前，宁波已培育多家数商企业，打造了一批大数据技术落地赋能的典型案例。

工业大数据应用场景

智能制造与优化

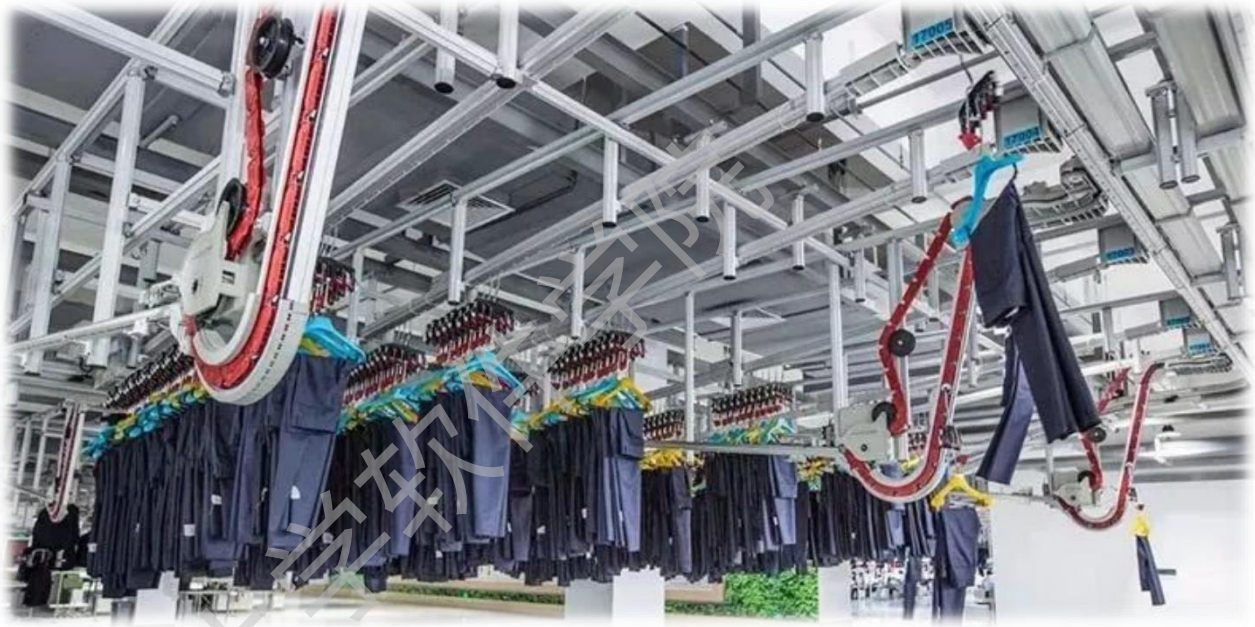
供应链管理与物流

市场分析与预测

.....

雅戈尔案例

- 走进雅戈尔的生产车间，生产线负责人熟练地打开屏幕，轻触设备，生产线的生产进度、设备状态等关键数据立刻显现。
- 雅戈尔总部生产基地产生的海量数据，通过抓取、收集与分析，再次作用于生产端，优化生产工艺。
- 如今雅戈尔总部生产基地的生产效率提高了至少25%。量身定制一件西服，用时从15天压缩至5天。



结 束