厚德 尚能 自强 鼎新

认识大数据

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

回顾：

上一节课了解什么是大数据？它是如何诞生的？它有哪些应用场景？

本节课我们将继续探讨大数据的特征与定义，大数据的处理过程。

学习目标

重点

1、掌握大数据具有的三个明显特征
2、理解大数据定义
3、理解大数据的 “4V”特性
4、理解并实践大数据处理过程

难点

大数据处理过程的理解与实践

课前测试 --测试线上学习情况

上节课，线上学习任务留下的一个思考题：
面对大数据，我们为什么不直接采取增加服务器节点每个节点上去承担部分的数据并提供服务，用这种方式去解决大数据的问题（即构建MPP集群）？

MPP集群会存在哪些问题：
1、热点数据与扩张性问题
2、性能问题
3、注意（结合业务），金融企业对数据的精确度要求比较高，一般都会采用Oracle数据库构建MPP集群的方式提供服务。
（这也是我们在面试过程中经常会遇到的问题）

课前测试 --测试线上学习情况

上节课，线上学习任务留下的一个思考题：
面对大数据，我们为什么不直接采取增加服务器节点，在不同的服务器上部署相同的数据库应用，每个节点上去承担部分的数据并提供服务，用这种方式去解决大数据的问题（即构建MPP集群） ？

1、热点数据与扩张性：通过改变架构方式构建MPP集群，虽然能解决燃眉之急，但扩展性和热点问题仍然得不到解决。

2、性能问题：传统的数据库一般只负责数据存储，提供良好的查询性能。在做数据处理时，需要编写额外的并发程序，先从数据库中读取数据，然后再进行处理。这种处理模式下，会涉及到大量的数据移动，对于磁盘和网络都是很大的消耗，进而影响处理效率。

3、注意（结合业务），金融企业对数据的精确度要求比较高，一般都会采用Oracle数据库构建MPP集群的方式提供服务。
（这也是我们在面试过程中经常会遇到的问题）

一、大数据综述

与传统数据的产生方式相比，大数据具有三个明显的特征：

数据量大：数据量大是大数据的明显特征，一般计量单位都是PB、EB甚至ZB。

非结构性：大数据既包含结构化数据也包含非结构化数据，而且通过特定的大数据技术从大量非结构化数据中提取有用的信息。

实时性：在互联网高速发展的背景下，我们所谈到的大数据不仅仅数量巨大，实时性、动态性成了大数据的另一重要特征。

二、大数据概念

大数据定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。即大数据是现有数据库管理工具和传统数据处理手段很难处理的大型、复杂的数据集，其涉及到采集、存储、搜索、共享、传输和可视化等方面。

二、大数据概念

大数据的特点可归纳为“4V”，即
Volume（容量），即海量的数据规模；
Variety（种类），即多样的数据类型；
Velocity（速度），即快速的数据流转和动态的数据体系；
Value（价值），即巨大的数据价值。

三、大数据处理过程

大数据采集：数据采集系统是结合基于计算机或者其他专用测试平台的测量软硬件产品来实现灵活的、用户自定义的测量系统。数据采集技术广泛应用在各个领域,比如摄像头，麦克风，都是数据采集工具。

大数据导入/预处理：虽然采集端本身会有很多数据库，但是如果要对这些海量数据进行有效的分析，还是应该将这些来自前端的数据导入到一个集中的大型分布式数据库，或者分布式存储集群，并且可以在导入基础上做一些简单的清洗和预处理工作。

三、大数据处理过程

大数据统计与分析：大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些数据进行专业化处理。可以说大数据分析是决策过程中的决定性因素，也是大数据时代发挥数据价值的关键环节。大数据分析技术帮助企业了解客户、锁定资源、规划生产、开拓新的业务。

三、大数据处理过程

大数据挖掘：从海量数据中发现有价值的信息，把这些数据转化成有组织的知识，这种需求导致了大数据挖掘的诞生。

数据挖掘主要是在现有数据上面进行基于各种算法的计算，从而起到预测（Predict）的效果，从而实现一些高级别数据分析的需求。

课后线上学习

线上学习内容将通过学习通下发，请同学在规定时间内容完成以下内容：

1、本节课课后作业，请按时提交。

2、线上内容学习并思考下面问题。

问题：

大数据的应用场景分为两种，离线处理场景和实时处理场景。请分析两种场景的处理流程以及两种场景的本质区别？ （下节课讨论的重点）

课堂小结

1、大数据具有三个明显的特征：数据量大、非结构性、实时性。

2、大数据定义：大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。

3、大数据的特点 “4V”：容量、种类、速度、价值。

4、**大数据处理过程：**大数据采集->大数据导入/预处理->大数据统计与分析->大数据挖掘

厚德 尚能 自强 鼎新

谢谢！

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

厚德 尚能 自强 鼎新

大数据技术基础

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

回顾：

上一节课了解什么是大数据？它是如何诞生的？它有哪些应用场景？只有了解了这些，才能窥视大数据的技术全貌。一个技术的诞生，是顺应时代的，是用于解决某些问题的，它的发展也一定是有内在逻辑的。

本节课我们将继续探讨****

学习目标

重点

1、基础架构支持:理解大数据技术架构组成。
2、数据采集的基本概论
3、数据采集的方法

难点

大数据技术架构理解与运用、数据采集方法

课前测试（上节课留下的问题） --测试线上学习情况

上节课，我们留下了这样一个思考题：
我们为什么不直接采取增加服务器服务器，部署多个数据库应用节点的方式去（即构建MPP集群）解决大数据的问题？

1、通过改变架构方式，组成的MPP集群，虽然能解决燃眉之急，但因为其扩展性和热点问题，无法满足日益增长的数据处理需要。

2、传统的NoSQL数据库一般只负责数据存储，提供良好的查询性能。但对于这类数据，用途较为灵活、广泛，在做数据处理时，需要编写额外的并发程序，先从NoSQL数据库中读取数据，然后再进行自定义处理。这种处理模式下，会涉及到大量的数据移动，对于磁盘和网络都是很大的消耗，进而影响处理效率。

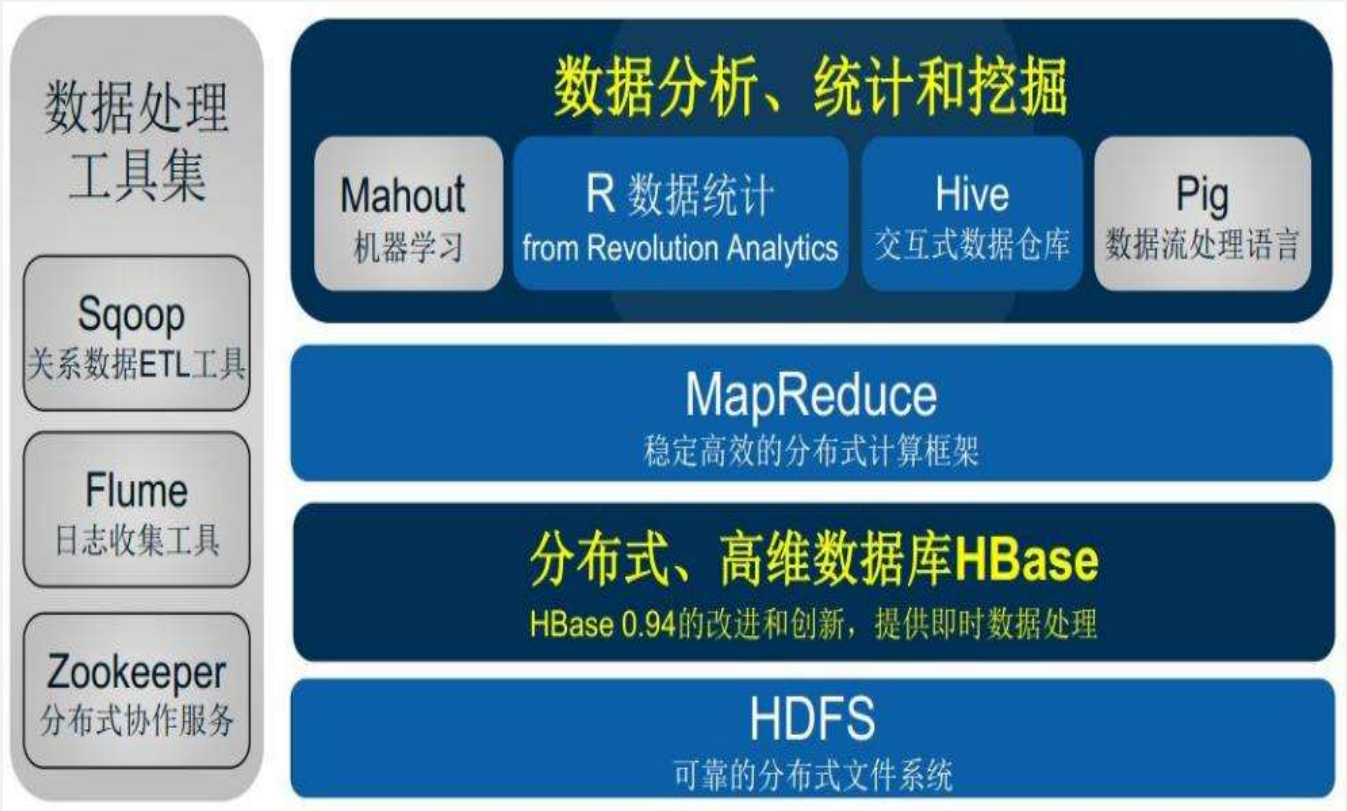
3、注意，目前大部分金融企业因为对数据有较高的实时性要求，且大部分数据为结构化的，仍主要采用Oracle数据库构建MPP集群的方式提供服务。

一、基础架构支持

Hadoop是一个由Apache基金会所开发的分布式系统基础架构。用户可以在不了解分布式底层细节的情况下开发分布式程序。Hadoop的框架最核心的设计就是：HDFS和MapReduce。HDFS为海量的数据提供了存储，则MapReduce为海量的数据提供了计算。

MapReduce是一种编程模型，用于大规模数据集（大于1TB）的并行运算。Map--映射，Reduce--归约。MapReduce采用"分而治之"的思想，把对大规模数据集的操作，分发给一个主节点管理下的各个分节点共同完成，然后通过整合各个节点的中间结果，得到最终结果。简单地说，MapReduce就是"任务的分解与结果的汇总"。

一、基础架构支持


数据处理
工具集
Sqoop
关系数据ETL工具
Flume
日志收集工具
Zookeeper
分布式协作服务
数据分析、统计和挖掘
Mahout
机器学习
R 数据统计
from Revolution Analytics
Hive
交互式数据仓库
Pig
数据流处理语言
MapReduce
稳定高效的分布式计算框架
分布式、高维数据库HBase
HBase 0.94的改进和创新，提供即时数据处理
HDFS
可靠的分布式文件系统

一、基础架构支持

HBase是运行在Hadoop上的NoSQL数据库，能够融合key/value存储模式带来实时查询的能力，以及通过MapReduce进行离线处理或者批处理的能力。HBase是基于列的而不是基于行的模式。

一、基础架构支持

Hive是建立在 Hadoop 上的数据仓库基础构架。可以用来进行数据提取转化加载（ETL），这是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制。Hive 定义了简单的类 SQL 查询语言，称为HQL，它允许熟悉 SQL 的用户查询数据。同时，这个语言也允许熟悉MapReduce 开发者的开发自定义的 mapper 和 reducer 来处理内建的mapper 和 reducer 无法完成的复杂的分析工作。

二、数据采集

数据采集是大数据价值挖掘中重要的一环，其后的分析挖掘都是建立在数据采集的基础之上。

各种类型信号采集的难易程度差别很大。实际采集时，噪声也可能带来一些麻烦。数据采集时，有一些基本原理要注意，还有更多的实际的问题要解决。

二、数据采集

数据采集的方法
1.基于物联网采集方法
2.系统日志采集方法
3.网络数据采集方法
4.其他数据采集方法

课后线上学习

线上学习内容将通过学习通下发，请同学在规定时间内容完成以下内容：

1、本节课课后作业，请按时提交。

2、线上内容学习并思考下面问题。

问题：

大数据的应用场景分为两种，离线处理场景和实时处理场景。请分析两种场景的处理流程以及两种场景的本质区别？ （下节课讨论的重点）

课堂小结

1、理解大数据技术架构组成（重点讲了Hadoop、mapreduce、hive、hbase基本理论）。

2、数据采集的理论与重要性

3、数据采集的方法：物联网、日志、网络、其它。

厚德 尚能 自强 鼎新

谢谢！

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

厚德 尚能 自强 鼎新

大数据管理

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

回顾：

上一节课了解什么是大数据？它是如何诞生的？它有哪些应用场景？只有了解了这些，才能窥视大数据的技术全貌。一个技术的诞生，是顺应时代的，是用于解决某些问题的，它的发展也一定是有内在逻辑的。

本节课我们将继续探讨****

学习目标

重点

1、数据质量的四大要素
2、数据清洗概念理解
3、数据清洗的方法
4、数据清洗的过程
5、数据类型与数据转换
6、大数据的提取和加载

难点

掌握大数据的清洗方法与过程、理解大数据的提取与加载

课前测试（上节课留下的问题） --测试线上学习情况

上节课，我们留下了这样一个思考题：
我们为什么不直接采取增加服务器服务器，部署多个数据库应用节点的方式去（即构建MPP集群）解决大数据的问题？

1、通过改变架构方式，组成的MPP集群，虽然能解决燃眉之急，但因为其扩展性和热点问题，无法满足日益增长的数据处理需要。

2、传统的NoSQL数据库一般只负责数据存储，提供良好的查询性能。但对于这类数据，用途较为灵活、广泛，在做数据处理时，需要编写额外的并发程序，先从NoSQL数据库中读取数据，然后再进行自定义处理。这种处理模式下，会涉及到大量的数据移动，对于磁盘和网络都是很大的消耗，进而影响处理效率。

3、注意，目前大部分金融企业因为对数据有较高的实时性要求，且大部分数据为结构化的，仍主要采用Oracle数据库构建MPP集群的方式提供服务。

一、大数据的清洗

数据质量具有四大要素：

①完整性 ②一致性 ③准确性 ④及时性

数据清洗就是按照一定的规则把“脏数据”“洗掉”，过滤不符合要求的数据，主要包括不完整的数据、错误的数据、重复的数据，然后将过滤的结果交给业务主管部门，确认是否过滤掉还是修正之后再进行提取。因此如何对数据进行有效的清理和转换，使之成为符合数据分析要求的数据源，是影响数据分析准确性的关键因素。

一、大数据的清洗

1. 数据清洗的方法

① 通过人工检查

② 通过专门编写的应用程序

③ 针对特定应用领域的数据清理

④ 针对与特定应用领域无关的数据清理

一、大数据的清洗

2. 数据清洗的过程

第一阶段：数据分析、定义错误类型

第二阶段：搜索、识别错误记录

第三阶段：修正错误

二、数据类型与数据转换

数据类型

①结构化数据 ②半结构化数据 ③非结构化数据

数据转换 数据转换是将数据从一种表示形式变为另一种表示形式的过程。由于每一个软件后台数据库的构架与数据的存储形式都是不相同的，因此就需要对数据进行转换。例如，对两个操作数进行运算，当操作数的类型不同，而且不属于基本数据类型时，经常需要将操作数转换为所需要的类型，这个过程即为强制类型转换。强制类型转换有两种形式：显式强制类型转换和隐式强制类型转换。

三、大数据的提取和加载

大数据的提取和加载是指将转换好的数据保存到数据仓库中去。大数据在加载时一般采用两种方式：

①完全刷新加载（全量加载）

从技术角度上说，完全刷新加载比增量提取和加载要简单得多，它适用于数据量不大并且时间代价和条件代价较小的情况。

②增量提取和加载（增量加载）

如何精准快速地捕获变化的数据是实现数据增量加载的关键。

（1）触发器方式；（2）时间戳方式；（3）全表比对方式

（4）日志表方式；（5）系统日志分析方式

课后线上学习

线上学习内容将通过学习通下发，请同学在规定时间内容完成以下内容：

1、本节课课后作业，请按时提交。

2、线上内容学习并思考下面问题。

问题：

大数据的应用场景分为两种，离线处理场景和实时处理场景。请分析两种场景的处理流程以及两种场景的本质区别？ （下节课讨论的重点）

课堂小结

重点

1、数据质量的四大要素：
2、数据清洗概念理解
3、数据清洗的方法
4、数据清洗的过程
5、数据类型与数据转换
6、大数据的提取和加载

厚德 尚能 自强 鼎新

谢谢！

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

厚德 尚能 自强 鼎新

大数据统计分析

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

回顾：

上一节课了解什么是大数据？它是如何诞生的？它有哪些应用场景？只有了解了这些，才能窥视大数据的技术全貌。一个技术的诞生，是顺应时代的，是用于解决某些问题的，它的发展也一定是有内在逻辑的。

本节课我们将继续探讨****

学习目标

重点

1、大数据分析的基本概念。
2、大数据分析的应用场景
3、大数据分析与传统数据分析区别
4、大数据分析与传统数据仓库分析区别
5、统计分析常见指标分类。

难点：

1、总量指标的理解与计算方法

课前测试（上节课留下的问题） --测试线上学习情况

上节课，我们留下了这样一个思考题：
我们为什么不直接采取增加服务器服务器，部署多个数据库应用节点的方式去（即构建MPP集群）解决大数据的问题？

1、通过改变架构方式，组成的MPP集群，虽然能解决燃眉之急，但因为其扩展性和热点问题，无法满足日益增长的数据处理需要。

2、传统的NoSQL数据库一般只负责数据存储，提供良好的查询性能。但对于这类数据，用途较为灵活、广泛，在做数据处理时，需要编写额外的并发程序，先从NoSQL数据库中读取数据，然后再进行自定义处理。这种处理模式下，会涉及到大量的数据移动，对于磁盘和网络都是很大的消耗，进而影响处理效率。

3、注意，目前大部分金融企业因为对数据有较高的实时性要求，且大部分数据为结构化的，仍主要采用Oracle数据库构建MPP集群的方式提供服务。

一、大数据分析概述

大数据分析指运用适的统计方法对收集来的大量数据进行分析，提取有用的信息以对数据加以详细研究和概括的过程。大数据分析可以帮助人们做出判断，以便采取适当行动。

案列：

1、企业可以通过对消费者爱好、需求以及对品牌忠诚度等因素进行大数据分析，来制定服务和营销智能决策。

2、通过对通信、金融活动记录的大数据分析，来精准地拓展业务和更好的服务客户。

一、大数据分析概述

大数据分析与传统数据分析区别

1、传统分析侧重于宏观、整体；而大数据分析可以分析微观、个体，实现个性化需求。

2、传统分析源于阶段性，针对性评估，采样过程可能误差较大；大数据分析来源于过程性、及时性记录，技术型的观察采样方式误差较小。

一、大数据分析概述

大数据分析与传统数据仓库分析区别

传统数据仓库采用精致的提取、转换和加载的流程以及使用数据库加以限制，意味着加载到数据仓库的数据是容易理解的、清洗过的，并符合业务的元数据

大数据分析针对的是非结构化数据，意味着不能保证输入数据是弯针的、清洗过的，使得分析过程更具挑战性。

一、大数据分析概述

应用：

1、大数据帮助能源公司设置发电机站点。
2、大数据帮助零售企业制定促销策略 。
3、大数据对交通行为进行预测。
4、大数据对疾病疫情的预测。
5、大数据帮助奥巴马大选连任成功。

二、统计分析常见指标

统计指标特点

总体性 ：反映的是客观事物的总体现象，而不是个体现象。

具体性 ：反映的总体数量数客观存在的，而不是主管抽象的概念和数字。

二、统计分析常见指标

统计指标的种类：

1、按统计指标所说明的总体现象内容不同，可分为数量指标和质量指标；

2、按统计指标按作用和表现形式不同，可分为总量指标、相对指标、平均指标、标志变异指标四类；

3、按统计指标的作用和功能的不同，可以分为描述指标、评价指标、监测指标和预警指标；

二、统计分析常见指标

总量指标

总量指标是反映社会经济现象在一定时间、空间条件下的总规模或总水平的最基本的综合指标，用绝对数表示，因此，总量指标又叫统计绝对数。

案列：如某企业去年总工资100万元，去年上半年总工资30万元，相减得去年下半年总工资。比如每年的政府工作报告都会公布关乎国计民生的重要总量指标。

二、统计分析常见指标

总量指标的计算方法 ：

直接计算法

它是对研究对象用直接的计数、点数和测量等方法，登记各单位的具体数值加以汇总，得到总量指标。如统计报表或普查中的总量资料，基本上都是用直接计算法计算出来的。

间接推算法

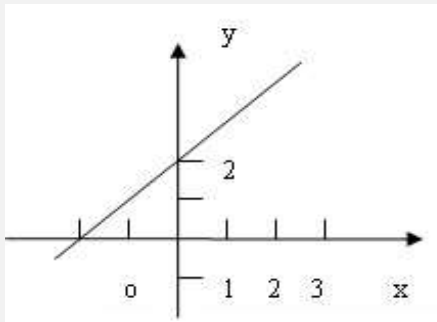
它是采用社会经济现象之间的平衡关系、因果关系、比例关系或利用非全面调查资料进行推算总量的方法。如利用样本资料推断某种农产品的产量，利用平衡关系推算某种商品的库存量等。

二、回归与预测

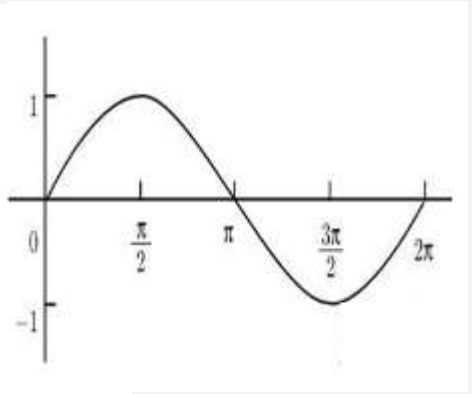
回归

一般说来，回归就是在分析自变量和因变量之间相关关系的基础上，建立变量之间的方程。

回归的本质是一种数学模型，通过建立变量间适当的依赖关系，以分析数据内在规律，并可用于预报、控制等问题



公式：y = a + b x



公式：y =sin(x)

二、回归与预测

预测

统计预测属于预测方法研究范畴，即如何利用科学的统计方法对事物的未来发展进行定量推测，并计算概率置信区间。是一种具有通用性的方法。最简单的预测方法就是回归预测，即将回归方程作为模型，根据自变量在预测期的数量变化来预测因变量值。

二、回归与预测

预测步骤：

1、确定变量：明确预测的具体目标，也就确定了因变量。

2、建立模型：依据自变量和因变量的历史统计资料进行计算，在此基础上建立回归分析方程，即回归分析预测模型。

3、进行分析：回归分析是对具有因果关系的影响因素（自变量）和预测对象（因变量）所进行的数理统计分析处理。只有当变量与因变量确实存在某种关系时，建立的回归方程才有意义。

课后线上学习

线上学习内容将通过学习通下发，请同学在规定时间内容完成以下内容：

1、本节课课后作业，请按时提交。

2、线上内容学习并思考下面问题。

问题：

大数据的应用场景分为两种，离线处理场景和实时处理场景。请分析两种场景的处理流程以及两种场景的本质区别？

课堂小结

重点

1、大数据分析的基本概念。
2、大数据分析的应用场景
3、大数据分析与传统数据分析区别
4、大数据分析与传统数据仓库分析区别
5、统计分析常见指标分类。

难点：

1、总量指标的理解与计算方法

厚德 尚能 自强 鼎新

谢谢！

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

厚德 尚能 自强 鼎新

数据挖掘

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

回顾：

上一节课了解什么是大数据？它是如何诞生的？它有哪些应用场景？只有了解了这些，才能窥视大数据的技术全貌。一个技术的诞生，是顺应时代的，是用于解决某些问题的，它的发展也一定是有内在逻辑的。

本节课我们将继续探讨****

学习目标

重点

1、******

2、*****

难点

课前测试（上节课留下的问题） --测试线上学习情况

上节课，我们留下了这样一个思考题：
我们为什么不直接采取增加服务器服务器，部署多个数据库应用节点的方式去（即构建MPP集群）解决大数据的问题？

1、通过改变架构方式，组成的MPP集群，虽然能解决燃眉之急，但因为其扩展性和热点问题，无法满足日益增长的数据处理需要。

2、传统的NoSQL数据库一般只负责数据存储，提供良好的查询性能。但对于这类数据，用途较为灵活、广泛，在做数据处理时，需要编写额外的并发程序，先从NoSQL数据库中读取数据，然后再进行自定义处理。这种处理模式下，会涉及到大量的数据移动，对于磁盘和网络都是很大的消耗，进而影响处理效率。

3、注意，目前大部分金融企业因为对数据有较高的实时性要求，且大部分数据为结构化的，仍主要采用Oracle数据库构建MPP集群的方式提供服务。

一、大数据挖掘基本概念

数据挖掘起源于多种学科，其中最重要的两门是统计学和机器学习，统计学起源于数学，因此，它强调数学上的精确，机器学习更多地起源于计算机实践。

数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

二、数据挖掘任务

数据总结目的是对数据进行浓缩，给出它的总体综合描述。通过对数据的总结，数据挖掘能够将数据库中的有关数据从较低的个体层次抽象总结到较高的总体层次上，从而实现对原始基本数据的总体把握。

最简单的数据总结方法是利用统计学中的传统方法，计算出数据库中各个数据项的总和、平均、方差、最大值、最小值等基本描述统计量。或者通过利用统计图形工具，对数据制作直方图、饼状图等。

利用OLAP(On Line Processing)技术（即联机分析处理技术）实现数据的多维查询也是一种广泛使用的数据总结的方法。

二、数据挖掘任务

分类的主要功能是使用一个分类函数或分类模型（也常常称作分类器），该模型能够根据数据的属性将数据分派到不同的组中。即：分析数据的各种属性，并找出数据的属性模型，确定哪些数据属于哪些组。这样我们就可以利用该模型来分析已有数据，并预测新数据将属于哪一个组。

分类应用的实例很多。例如，我们可以将银行网点分为好、一般和较差三种类型，并依此分析这三种类型银行网点的各种属性，特别是位置、盈利情况等属性，并决定它们分类的关键属性及相互间关系。此后就可以根据这些关键属性对每一个预期的银行网点进行分析，以便决定预期银行网点属于哪一种类型。

二、数据挖掘任务

关联分析的目的是找出数据库中隐藏的关联网，描述一组数据项目的密切度或关系。有时并不知道数据库中数据的关联是否存在精确的关联函数，即便知道也是不确定的，因此关联分析生成的规则带有置信度，置信度级别度量了关联规则的强度。

二、数据挖掘任务

聚类：当要分析的数据缺乏描述信息，或者是无法组织成任何分类模式时，可以采用聚类分析。聚类分析是按照某种相近程度度量方法，将用户数据分成一系列有意义的子集合。每一个集合中的数据性质相近，不同集合之间的数据性质相差较大。

统计方法中的聚类分析是实现聚类的一种手段，它主要研究基于几何距离的聚类。人工智能中的聚类是基于概念描述的。概念描述就是对某类对象的内涵进行描述，并概括这类对象的有关特征。概念描述分为特征性描述和区别性描述，前者描述某类对象的共同特征，后者描述不同类对象之间的区别。

三、数据挖掘流程

1、业务理解

确定业务目标、数据挖掘目标、制定实施计划

2、数据理解

数据收集、数据描述、探索性分析、数据质量检测

3、数据准备

选择数据、清洗数据、数据格式化

4、建立模型

选择建模技术、生成模型、评估模型

5、结果评价

对模型预测结果进行评估，主要模型准确性与模型说明。

四、数据挖掘常用方法

决策树方法就是利用信息论的原理建立决策树。该类方法的实用效果好，影响较大。决策树可高度自动化地建立起易于为用户所理解的模型，而且，系统具有较好地处理缺省数据及带有噪声数据等能力。

四、数据挖掘常用方法

表示：决策树是一树状结构，它从根节点开始，对数据样本（由实例集组成，实例有若干属性）进行测试，根据不同的结果将数据样本划分成不同的数据样本子集，每个数据样本子集构成一子节点。生成的决策树每个叶节点对应一个分类。构造决策树的目的是找出属性和类别间的关系，用它来预测将来未知类别的记录的类别。这种具有预测功能的系统叫决策树分类器。

四、数据挖掘常用方法

构造一个决策树分类器通常分为两步：树的生成和剪枝。决策树的生成是一个从上至下，“分而治之” 的过程，是一个递归的过程。设数据样本集为S，算法框架如下：

(1)如果数据样本集S中所有样本都属于同一类或者满足其它终止准则，则S不再划分，形成叶节点。

(2)否则，根据某种策略选择一个属性，按照属性的各个取值，对S进行划分，得到n个子样本集，记为 。再对每个迭代执行步骤1经过n次递归，最后生成决策树。从根到叶结点的一条路径就对应着一条规则，整棵决策树就对应着一组析取表达式规则。树构成步骤中，主要就是找出节点的属性和如何对属性值进行划分。

四、数据挖掘常用方法

特点：决策树是一种常用于预测模型的算法，它通过将大量数据有目的分类，从中找到一些有价值的，潜在的信息。它的主要优点是描述简单，分类速度快，特别适合大规模的数据处理。

请同学们提前预习遗传算法、神经网络、关联规则、粗糙集、判别分析等数据挖掘算法，这也是下节课的重点

课后线上学习

线上学习内容将通过学习通下发，请同学在规定时间内容完成以下内容：

1、本节课课后作业，请按时提交。

2、线上内容学习并思考下面问题。

问题：

大数据的应用场景分为两种，离线处理场景和实时处理场景。请分析两种场景的处理流程以及两种场景的本质区别？ （下节课讨论的重点）

课堂小结

厚德 尚能 自强 鼎新

谢谢！

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

厚德 尚能 自强 鼎新

大数据可视化

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

回顾：

上一节课了解什么是大数据？它是如何诞生的？它有哪些应用场景？只有了解了这些，才能窥视大数据的技术全貌。一个技术的诞生，是顺应时代的，是用于解决某些问题的，它的发展也一定是有内在逻辑的。

本节课我们将继续探讨****

学习目标

重点

1、******

2、*****

难点

课前测试（上节课留下的问题） --测试线上学习情况

上节课，我们留下了这样一个思考题：
我们为什么不直接采取增加服务器服务器，部署多个数据库应用节点的方式去（即构建MPP集群）解决大数据的问题？

1、通过改变架构方式，组成的MPP集群，虽然能解决燃眉之急，但因为其扩展性和热点问题，无法满足日益增长的数据处理需要。

2、传统的NoSQL数据库一般只负责数据存储，提供良好的查询性能。但对于这类数据，用途较为灵活、广泛，在做数据处理时，需要编写额外的并发程序，先从NoSQL数据库中读取数据，然后再进行自定义处理。这种处理模式下，会涉及到大量的数据移动，对于磁盘和网络都是很大的消耗，进而影响处理效率。

3、注意，目前大部分金融企业因为对数据有较高的实时性要求，且大部分数据为结构化的，仍主要采用Oracle数据库构建MPP集群的方式提供服务。

一、可视化分类

数据可视化是关于图形或图形格式的数据展示，它能够帮助人们快速地理解数据。

一、可视化分类

1、结构可视化：结构可视化反映数据的内在组织结构，比如构成数据的元素、部件以及构成关系等。

2、功能可视化是对数据所对应的功能的可视化描述，如汽车发动机的运转状态，可通过对发动机进行3D建模，形成一段动画来清晰地展示。

3、关联关系可视化在很大程度上都是反映数据之间的关联关系，比如层级关系、对比关系之类的社交图谱。

4、趋势可视化是对数据发展的走势、预测等进行可视化的一种方式。

二、大数据可视化表现形式

二维可视化的表现形式以平面的形式表达数据之间的关联。主要包括2D区域图、时间序列图、网络图等。

案列：

① 2D区域图方法使用GIS数据可视化技术，往往涉及到事物特定表面上的位置。

② 时间序列图是数据以时间轴的方式展示，例如展示某区域的温度变化

③ 网络图展示数据点之间的错综复杂的相互关系,它是一种常见的大数据展示方法。

二、大数据可视化表现形式

三维可视化：

3D渲染技术是近年来发展迅速和备受关注的行业，在数字娱乐、虚拟现实、工业设计、实时仿真、数字城市等各个领域都有着十分广泛的应用。

体感互动技术是通过硬件互动设备、体感互动系统软件以及三维数字内容，来感应站在窗口前的观看者，当观看者的动作发生变化时，窗口显示的画面同时发生变化。

增强现实是把原本在现实世界很难体验到的实体信息，通过电脑模拟仿真后再叠加，将虚拟的信息应用到真实世界，被人类感官所感知，从而达到超越现实的感官体验。

二、大数据可视化表现形式

仪表盘是模仿汽车速度表的一种图表，常用来反映预算完成率、收入增长率等比率性指标。

定制可视化：针对于不同企业和用户的需求，“魔镜”提供了多个增值和定制化模块，包括可定制化图表支持，跨数据库、数据源支持，行业数据分析（项目），可定制视化分析组合，定制化分析挖掘模型和解决方案等。

三、大数据可视化方式的选择

单一数据可视化

在展现数据的时候，有时我们只需要突出一个最重要的数据。我们需要直接将这个数据放大或通过简单的颜色对比反映数据

对比型数据的展示

在对比型数据表示过程中，一般通用的图表就是条形图或柱形图，长长短短一目了然。

比例型数据的展示

对于比例型数据的图表展示，我们一般会选择饼图或圆环图显示。

三、大数据可视化方式的选择

相关关系数据的展示

如果不清楚两个变量之间的关系， 散点图是一个不错的选择。

复合关系数据可视化

有的时候数据包含的信息太多太杂，单一的图表并不能够全面地传递信息。此时，就可以选择复合图表。

课后线上学习

线上学习内容将通过学习通下发，请同学在规定时间内容完成以下内容：

1、本节课课后作业，请按时提交。

2、线上内容学习并思考下面问题。

问题：

大数据的应用场景分为两种，离线处理场景和实时处理场景。请分析两种场景的处理流程以及两种场景的本质区别？ （下节课讨论的重点）

课堂小结

厚德 尚能 自强 鼎新

谢谢！

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

厚德 尚能 自强 鼎新

大数据的安安全性

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院

回顾：

上一节课了解什么是大数据？它是如何诞生的？它有哪些应用场景？只有了解了这些，才能窥视大数据的技术全貌。一个技术的诞生，是顺应时代的，是用于解决某些问题的，它的发展也一定是有内在逻辑的。

本节课我们将继续探讨****

学习目标

重点

1、******

2、*****

难点

课前测试（上节课留下的问题） --测试线上学习情况

上节课，我们留下了这样一个思考题：
我们为什么不直接采取增加服务器服务器，部署多个数据库应用节点的方式去（即构建MPP集群）解决大数据的问题？

1、通过改变架构方式，组成的MPP集群，虽然能解决燃眉之急，但因为其扩展性和热点问题，无法满足日益增长的数据处理需要。

2、传统的NoSQL数据库一般只负责数据存储，提供良好的查询性能。但对于这类数据，用途较为灵活、广泛，在做数据处理时，需要编写额外的并发程序，先从NoSQL数据库中读取数据，然后再进行自定义处理。这种处理模式下，会涉及到大量的数据移动，对于磁盘和网络都是很大的消耗，进而影响处理效率。

3、注意，目前大部分金融企业因为对数据有较高的实时性要求，且大部分数据为结构化的，仍主要采用Oracle数据库构建MPP集群的方式提供服务。

一、大数据的安全

大数据应用在创造价值的同时，也面临着复杂严峻的安全挑战，如大数据的产生使数据分析与应用更加复杂，难以管理。

数据的安全分为：物理安全、网络安全、应用安全、数据隐私

一、大数据的安全

物理安全是指为了保证计算机系统安全、可靠地运行，确保数据不会受到人为或自然因素的危害而使造成丢失、泄漏和破坏，对计算机系统设备、通信与网络设备、存储媒体设备和人员所采取的安全技术措施。

一、大数据的安全

物理安全包括环境安全，设备安全和媒体安全三个方面。

1、环境安全是对系统所在环境的安全保护，如受灾防护和区域防护等。

2、设备安全包括设备防盗、设备防毁、防止电磁信息泄露、防止线路截获、抗电磁干扰、电源保护等。

3、媒体安全包括媒体本身的安全及媒体数据的安全。

- 媒体本身的安全保护：指防盗（如数据被非法拷贝）、防毁（如防止意外或者故意的破坏）、防霉等
- 媒体数据的安全保护：指防止记录的信息不被非法窃取、篡改、破坏或使用。

二、大数据的安全行业应用

大数据在各行各业得到了广泛的应用，但不同领域的应用对大数据安全需求也有所不同。

互联网行业

可靠的数据存储、安全的挖掘分析、严格的运营监管，呼唤针对用户隐私的安全保护标准、法律法规、行业规范，期待从海量数据中合理发现和发掘商业机会和商业价值。

电信行业

确保核心数据与资源的保密性、完整性和可用性。在保障用户利益、体验和隐私的基础上充分发挥数据价值。

二、大数据的安全行业应用

金融行业

对数据访问控制、处理算法、网络安全、数据管理和应用等方面提出安全要求，期望利用大数据安全技术加强金融机构的内部控制，提高金融监管和服务水平，防范和化解金融风险。

医疗行业

数据隐私性高于安全性和机密性，同时需要安全和可靠的数据存储、完善的数据备份和管理，以帮助医生与病人进行疾病诊断、药物开发、管理决策、完善医院服务，提高病人满意度，降低病人流失率。

政府组织

政府组织对大数据安全的需求是：隐私保护的安全监管、网络环境的安全感知、大数据安全标准的制定、安全管理机制的规范等内容。

二、数据防护技术

镜像技术

镜像技术是将建立在同一个局域网之上的两台服务器通过软件或其他特殊的网络设备，将两台服务器的磁盘做镜像。

快照技术

快照技术是一种摄影技术，随着存储应用需求的提高，用户需要在线方式进行数据保护，快照就是在线存储设备防范数据丢失的有效方法之一 。

二、数据防护技术

持续数据保护

持续数据保护（CDP）是一种在不影响主要数据运行的前提下，可以实现持续捕捉或跟踪目标数据所发生的任何改变，并且能够恢复到此前任意时间点的方法。

用户管理

用户管理涉及到两个重要的问题：用户身份管理和用户权限管理。

身份管理为身份认证服务，只有身份管理，没有身份认证，那么身份管理是毫无意义的。

权限管理为访问控制服务，只有权限管理没有访问控制，则权限管理也是没有意义的。

大数据的应用

课后线上学习

线上学习内容将通过学习通下发，请同学在规定时间内容完成以下内容：

1、本节课课后作业，请按时提交。

2、线上内容学习并思考下面问题。

问题：

大数据的应用场景分为两种，离线处理场景和实时处理场景。请分析两种场景的处理流程以及两种场景的本质区别？ （下节课讨论的重点）

课堂小结

厚德 尚能 自强 鼎新

谢谢！

试讲人：5号　　时间：2025.06.7

湖北黄冈应急管理职业技术学院