

# 数据仓库与数据挖掘概述

张冬松

信阳学院  
大数据与人工智能学院

[dszhang@nudt.edu.cn](mailto:dszhang@nudt.edu.cn)





# 讲课前

课程性质

考查课

考试形式

1. 百分制

2. 总分 (100%) = 考勤 (20%) + 报告 (80%)

报告内容

根据每次课后布置的作业列表，多选一，自选一个主题，自拟题目，自己收集资料，按照《计算机工程与应用》投稿格式撰写研究报告，总页数不超过10页。

# 讲课前

报告格式



提交方式

1. **电子版1份**：请发到如下邮箱[dszhang@nudt.edu.cn]，邮件主题请注明 **“DWDM+最终版+学号+姓名”**，附件中上传电子版文件，文件格式只能是doc或docx或pdf。
2. **打印版1份**：请交给各班班长，由班长统一收齐后提交；或者，自行交到数计楼4楼。



# 讲课前

## 微信答疑群

DWDM2020答疑群



该二维码7天内(9月15日前)有效，重新进入将更新



# 目录 content



## 第一节

## 数据仓库概述

## 第二节

## 数据挖掘概述

## 第三节

## 数据仓库与数据挖掘结合

# 第一节

## 数据仓库概述

- 从数据库到数据仓库
- 数据仓库定义与特性
- 数据仓库的体系结构







## 1.1 从数据库到数据仓库

对比内容	数据库	数据仓库
数据内容	当前值	历史的、存档的、归纳的、计算的数据
数据目标	面向业务操作程序、重复处理	面向主题域、管理决策分析应用
数据特性	动态变化、按字段更新	静态、不能直接更新、只定时添加
数据结构	高度结构化、复杂、适合操作计算	简单、适合分析
使用频率	高	中、低
数据访问量	每个事务只访问少量记录	有的事务可能要访问大量记录
对响应时间的要求	以秒为计量单位	以秒、分钟、甚至小时为计量单位



## 1.1 从数据库到数据仓库

### A.数据库用于事务处理

- 1.数据库作为数据资源，用于管理业务中的事务处理。当前，数据库已经成为成熟的信息基础设施。
- 2.数据库中存储的数据主要是保存当前数据，然后随着业务的变化，随时更新数据库中的数据。
- 3.不同的管理业务需要建立不同的数据库。





## 1.1 从数据库到数据仓库

### B.数据仓库用于决策分析

- 1.数据库用于事务处理，而数据仓库用于决策分析。
- 2.数据库保存事务处理的当前状态，而数据仓库既保存过去的  
数据，又保存当前的数据。
- 3.数据仓库的数据是大量数据库的集成。
- 4.数据库的操作比较明确，操作数据量少。数据仓库操作不  
明确，操作数据量大。



## 1.1 从数据库到数据仓库

数据库	数据仓库
细节的	综合或提炼的
在存取时准确的	代表过去的数据
可更新的	不更新
一次操作数据量小	一次操作数据量大
面向应用	面向分析
支持管理	支持决策



# 第一节

## 数据仓库概述

- 从数据库到数据仓库
- 数据仓库定义与特性
- 数据仓库的体系结构





## 1.2 数据仓库定义与特性

### A.数据仓库定义

1. William H. Inmon 于1993年在所著《Building the Data Warehouse》中首先系统地阐述了关于数据仓库的思想和理论，为数据仓库的发展奠定了历史基石。

2. 该文将数据仓库定义为：

“一个面向主题的、集成的、随时间变化的、非易失性的数据集，用于支持管理层的决策过程。”



## 1.2 数据仓库定义与特性

### B.面向主题性

面向主题性表示数据仓库中数据组织的基本原则，数据仓库中的所有数据都是围绕着某个主题而组织的。

- 根据决策问题确定主题
- 确定主题后，需要确定主题应该包含的数据，不是所有业务数据都能进入数据仓库的主题中
- 不同主题之间可能会出现相互重叠的信息
- 主题在数据仓库中可以用多维数据库的方式来存储
- 主题的划分中，必须保证每一个主题的独立性



## 1.2 数据仓库定义与特性

### C.集成性

根据决策分析的要求，将分散于各处的源数据进行抽取、筛选、清洗、聚合等工作，最终集成到数据仓库中。数据来源广泛，可以是各种数据库，还可以是OLTP等系统。

### D.时变性

数据仓库中的数据有时限，如5~10年，应该随着时间的推移而发生变化，不断生成主题新数据。



## 1.2 数据仓库定义与特性

### E.非易失性

数据仓库包括了大量的历史数据，这些数据一旦进入数据仓库，不经常进行更新处理或根本不更新。

### F.集合性

集合性是数据的闭合性，能够提供主题分析的全部数据。数据仓库的数据量很大，相对于一个TB(1000GB)级别数据库。

### G.软硬要求高

需要一个巨大的硬件平台，以及一个分布式并行数据库系统。

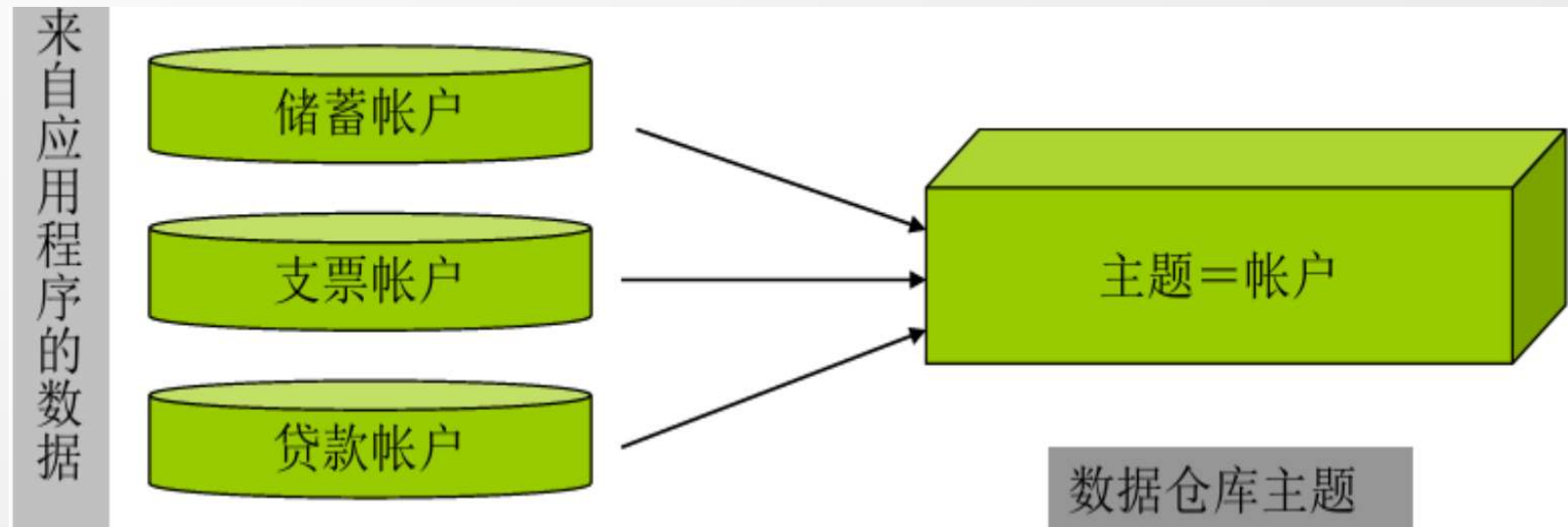


## 1.2 数据仓库定义与特性

### H.综合性

- 1.数据仓库中的数据来自不同的数据库、文件等数据源。
- 2.数据进入数据仓库之前，需要进行标准化工作：

- 命名规则
- 编码
- 数据特性
- 度量单位





## 1.2 数据仓库定义与特性

### 1.数据粒度

1.数据粒度是数据的细节程度。

2.数据仓库中，根据需求不同，需要不同层次的数据细节。

<u>每日数据</u>	<u>月汇总</u>	<u>季度汇总</u>
帐户	帐户	帐户
交易日期	月份	月份
数额	交易数	交易数
存款	取款	取款
取款	存款	存款
	期初结余	期初结余
	期末结余	期末结余

银行数据仓库的三个层次的数据粒度



# 第一节

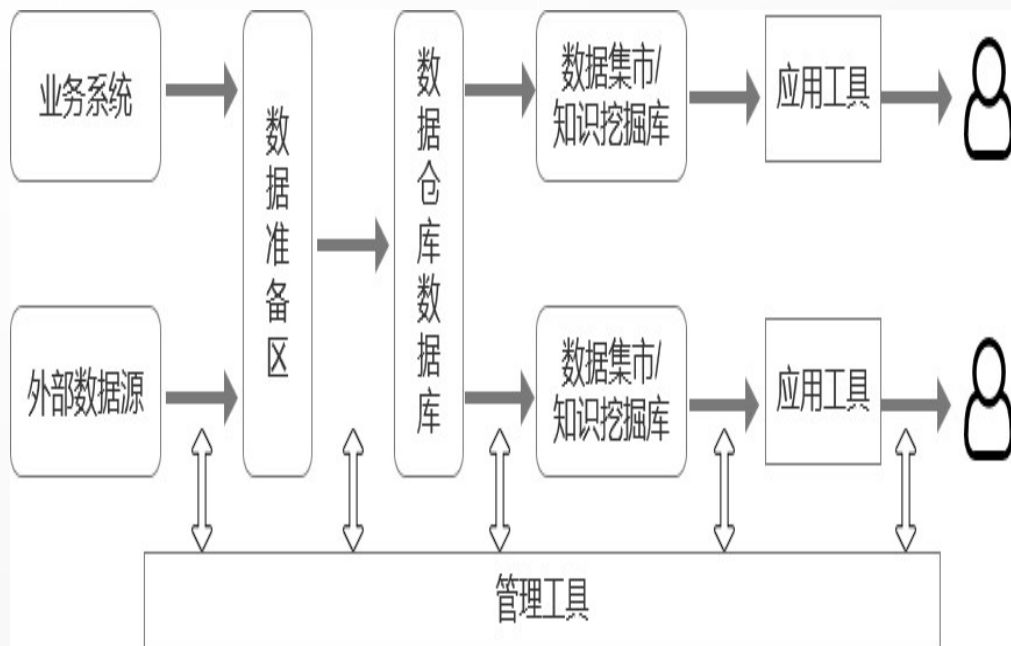
## 数据仓库概述

- 从数据库到数据仓库
- 数据仓库定义与特性
- 数据仓库的体系结构

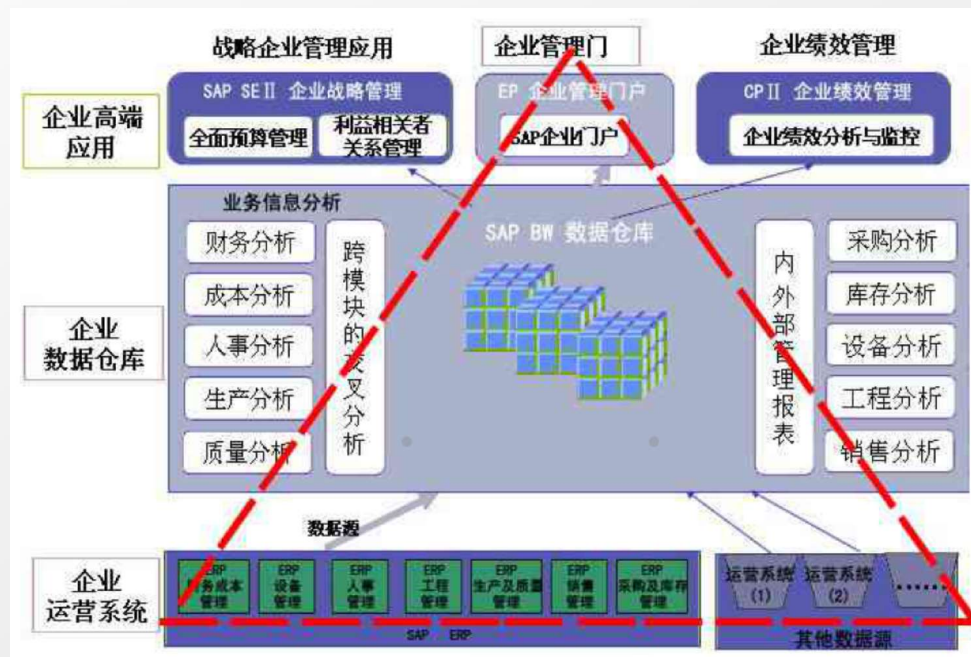


# 1.3 数据仓库的体系结构

## A.数据仓库体系结构



数据仓库的体系结构

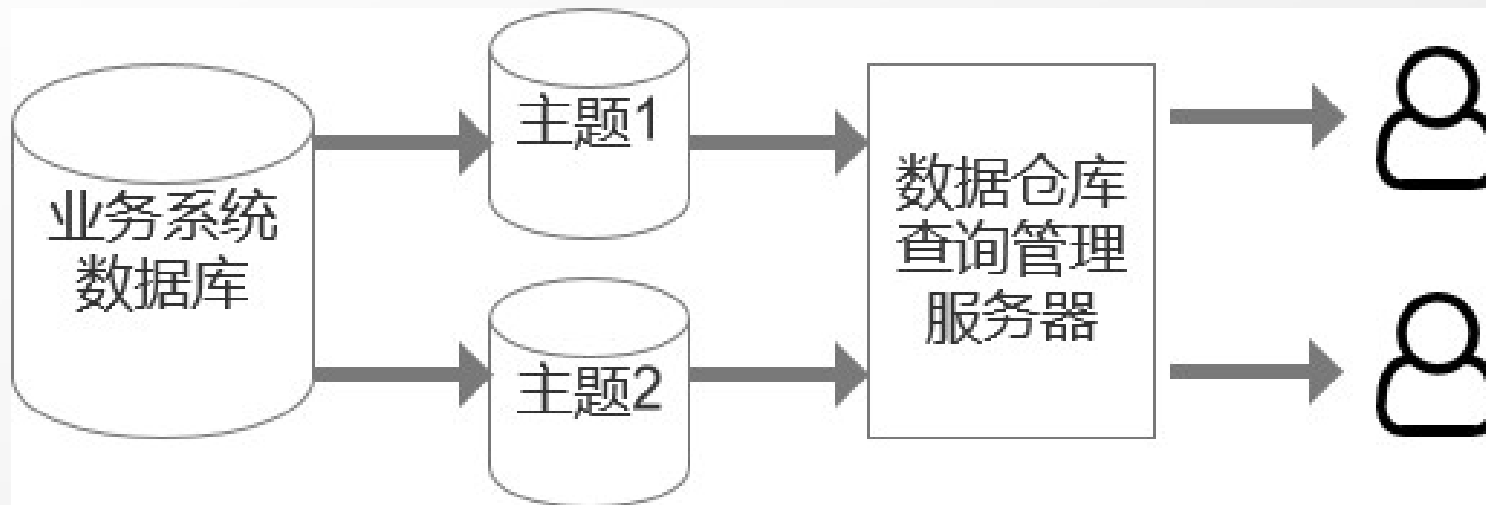


真实系统的体系结构

数据源，数据准备区，数据仓库中数据库，数据集市或知识挖掘库，管理工具，以及应用工具。

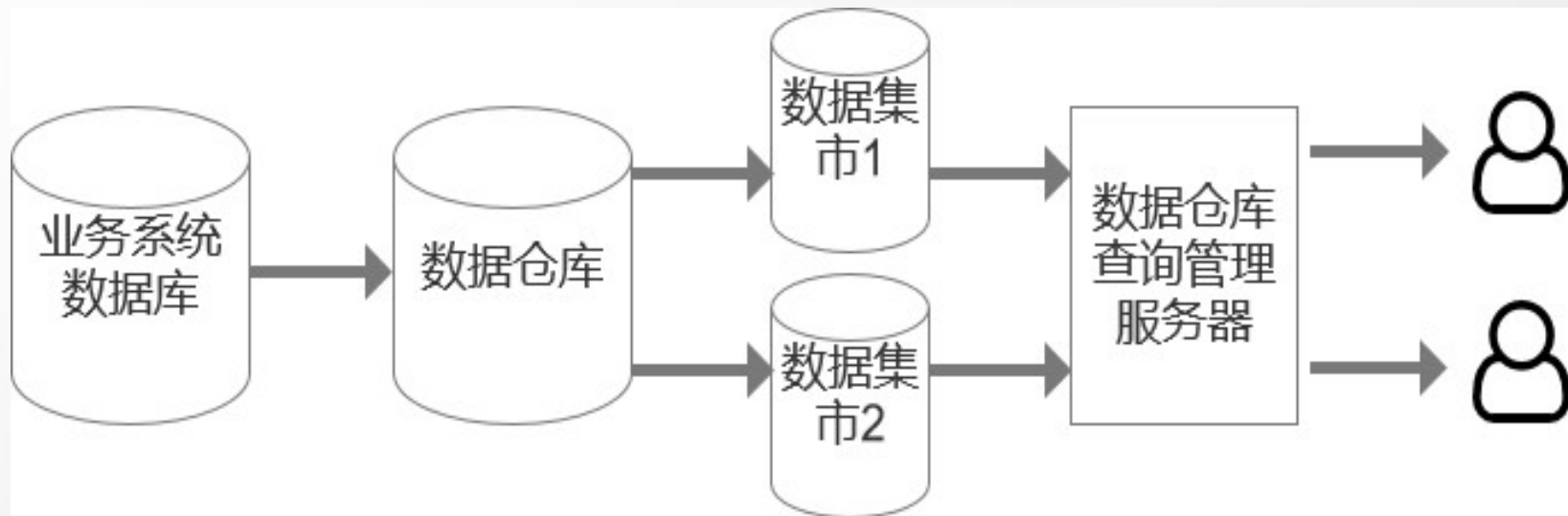
## 1.3 数据仓库的体系结构

### B.数据集市



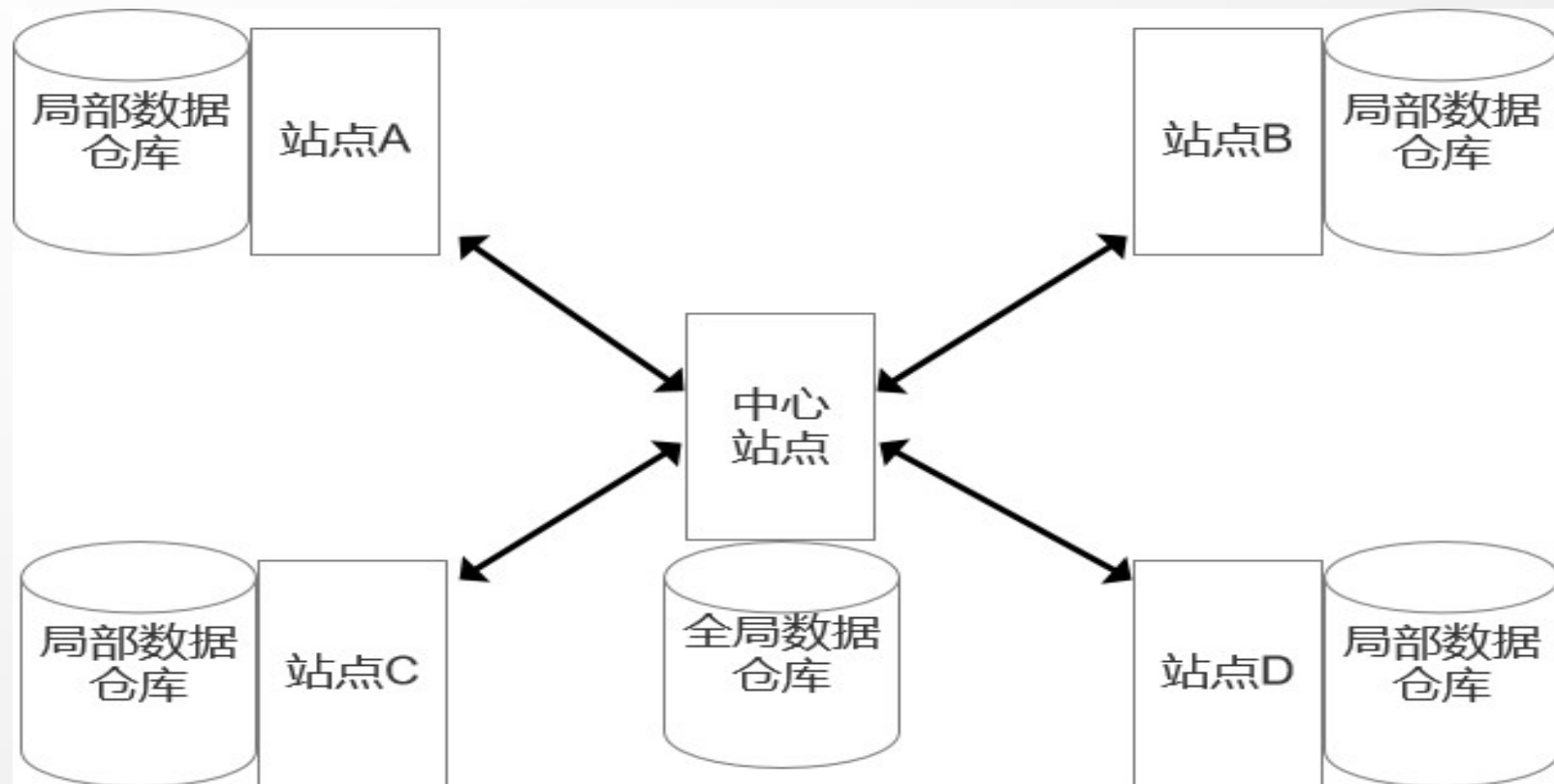
## 1.3 数据仓库的体系结构

### C.单一的数据仓库结构



## 1.3 数据仓库的体系结构

### D. 分布式数据仓库结构







## 第二节

## 数据挖掘概述

- 从机器学习到数据挖掘
- 数据挖掘定义
- 数据挖掘技术与工具





## 2.1 从机器学习到数据挖掘

- 1.学习是人类具有的智能行为，主要用于获取知识。
- 2.机器学习是研究使计算机模拟或实现人类的学习行为，即让计算机通过算法自动获取知识。
- 3.机器学习是人工智能领域中的重要研究方向。
- 4.机器学习研究始于20世纪60年代。



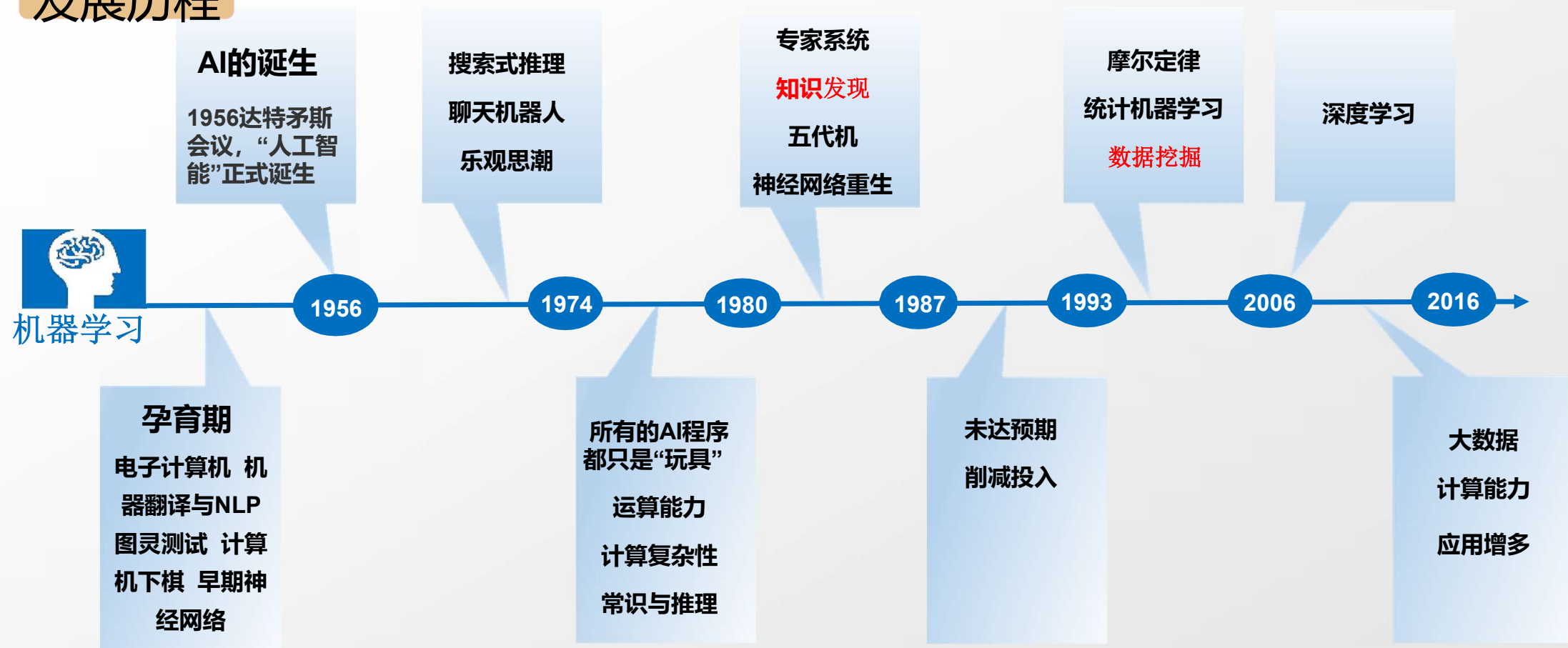
## 2.1 从机器学习到数据挖掘

- 1.1980年在美国召开第一届国际机器学习研讨会，明确了机器学习是人工智能的重要研究方向。
- 2.1989年8月于美国召开的第一届知识发现国际学术会议，首次提出知识发现概念。
- 3.1995年在加拿大召开第一届知识发现和数据挖掘国际学术会议，首次提出数据挖掘概念。
- 4.我国于1987年召开第一届全国机器学习研讨会。



## 2.1 从机器学习到数据挖掘

### 发展历程



## 2.1 从机器学习到数据挖掘

### 知识点



- 从以“推理”为重点到以“知识”为重点，再到以“学习”为重点
- 机器可以自动“学习”的算法，即从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。目前，机器学习=“分类”
- 人工智能 > 机器学习 > 深度学习



## 第二节

## 数据挖掘概述

- 从机器学习到数据挖掘
- 数据挖掘定义
- 数据挖掘技术与工具





## 2.2 数据挖掘定义

### A. 知识发现

从数据中发现有用知识的整个过程。

### B. 数据挖掘

Data Mining, 知识发现过程中的一个特定步骤, 采用专门的算法从数据中获取知识。

### 举例

从人类数据库中挖掘知识:

(头发=黑色)  $\vee$  (眼睛=黑色)  $\rightarrow$  亚洲人





## 2.2 数据挖掘定义

### C.数据挖掘的技术定义

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐含在其中的、人们所不知道的、但又是潜在有用的信息和知识的过程。

### D.数据挖掘的商业定义

数据挖掘是一种新的商业信息处理技术，主要特点是对商业数据库中的大量业务数据进行抽取、转化、分析和模式化处理，从中提取辅助商业决策的关键知识。



## 2.2 数据挖掘定义

### E.数据挖掘与统计学的区别与联系

- 1.统计学主要是对数值数据或连续值数据（如年龄、工资等），进行数值计算的定量分析，得到数量信息。
- 2.数据挖掘主要对离散数据（如性别、病症等）进行定性分析（如覆盖、归纳等），得到规则知识。
- 3.统计学与数据挖掘是有区别的，但是，两者之间是相互补充的。



## 第二节

## 数据挖掘概述

- 从机器学习到数据挖掘
- 数据挖掘定义
- 数据挖掘技术与工具





## 2.3 数据挖掘技术与工具

### A.常用数据挖掘技术

#### 1.传统数据分析类

包括：线性分析、非线性分析、回归分析、逻辑回归分析、时间序列分析、聚类分析等技术。

#### 2.知识发现类

包括：人工神经网络、决策树、遗传算法、规则发现、关联算法等。

#### 3.其他技术类

包括：文本数据挖掘、Web数据挖掘、数据可视化、空间数据挖掘、分布式数据挖掘等。



## 2.3 数据挖掘技术与工具

### B.常用数据挖掘工具

#### 1.按使用方式分类的数据挖掘工具

包括：决策方案生成工具、商业分析工具、研究分析工具。

#### 2.按数据技术分类的数据挖掘工具

包括：基于神经网络的工具、基于规则和决策树的工具、基于模糊逻辑的工具、综合性的数据挖掘工具等。

#### 3.按应用范围分类的数据挖掘工具

包括：专用型和通用型的数据挖掘工具等。



## 第三节 数据仓库与数据挖掘结合

- 数据仓库和数据挖掘的区别与联系
- 基于数据仓库的决策支持系统
- 集大成者，商业智能





## 3.1 数据仓库和数据挖掘的区别与联系

### A. 数据仓库和数据挖掘的区别

1. 数据仓库是一种存储技术，它要求不同用户对不同决策提供所需的数据和信息。
2. 数据挖掘研究各种方法和技术，从大量的数据中挖掘出有用的信息和知识。

### B. 数据仓库和数据挖掘的联系

1. 数据仓库是数据挖掘的基础，提供支撑平台。
2. 数据挖掘是数据仓库应用系统中重要工具之一。





## 3.1 数据仓库和数据挖掘的区别与联系

### C.两者的结合

数据挖掘作用于数据仓库，实现决策支持。

例如：

- 1) 预测客户购买倾向
- 2) 分析欺诈行为
- 3) 客户利润贡献度分析
- 4) 交通拥堵路段预测

等等



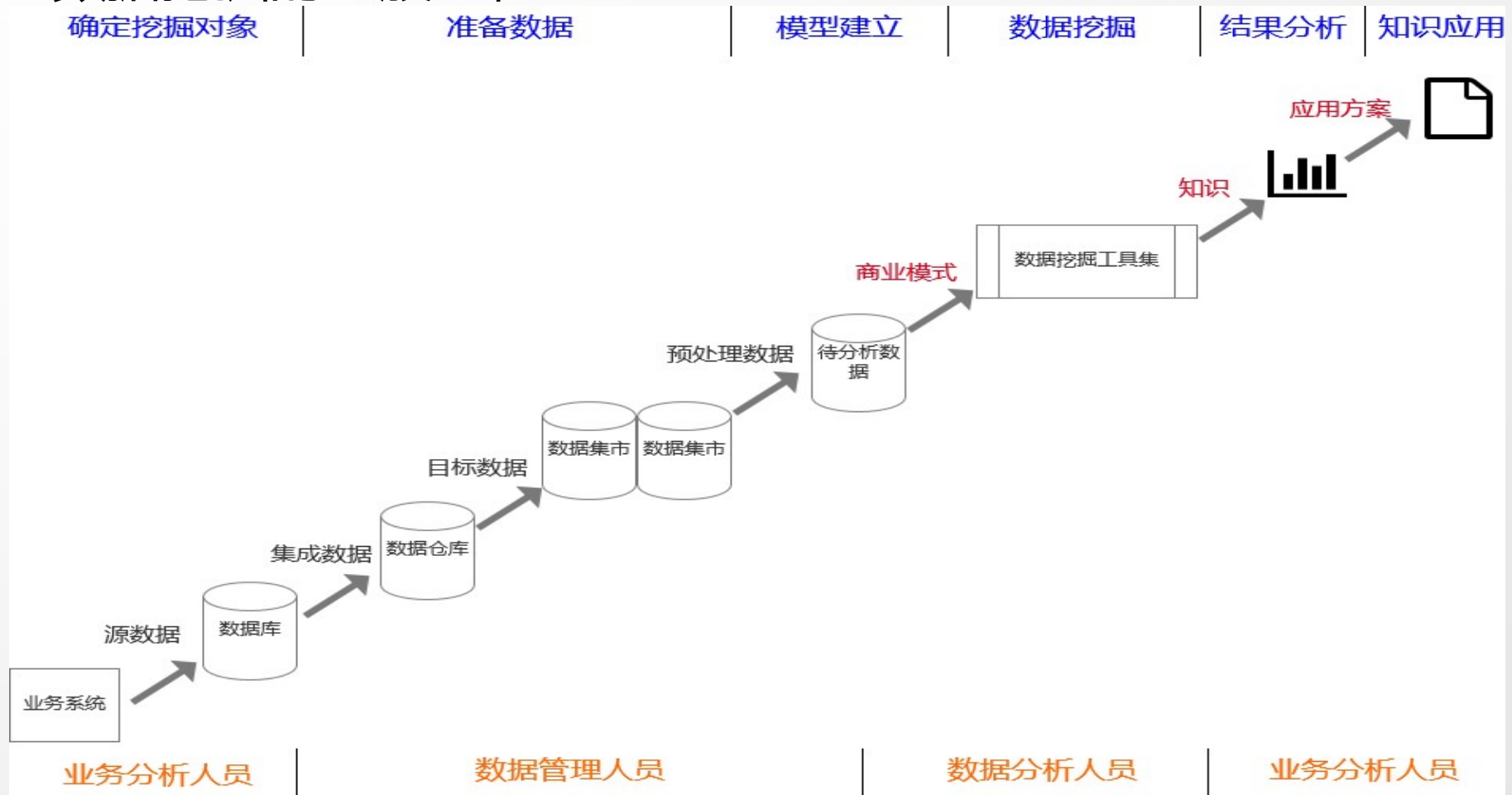
### 3.1 数据仓库和数据挖掘的区别与联系

D.数据仓库和数据挖掘的应用已经紧密结合在一起  
通常，数据挖掘工具需要在集成的、一致的、经过清理的数据上进行挖掘。

- 1) 数据仓库构造过程就包含：数据存储、数据集成、数据合并、数据库转换、数据库连接、管理和服务工具、数据分析功能等，用于数据处理和数据挖掘的基础设施。
- 2) 数据仓库中的在线分析处理工具（OLAP）完全可以为数据挖掘提供有关的数据操作支持。

## 3.1 数据仓库和数据挖掘的区别与联系

### E. 数据挖掘的一般过程



## 第三节 数据仓库与数据挖掘结合

- 数据仓库和数据挖掘的区别与联系
- 基于数据仓库的决策支持系统
- 集大成者，商业智能





## 3.2 基于数据仓库的决策支持系统

### A.决策支持系统包括哪些功能

- 1.对当前和历史数据完成查询和报表处理
- 2.可以用不同方法完成“如果满足什么条件，那么会如何”分析
- 3.从综合数据到细节数据，深入追踪、钻取和查询，找到问题原由
- 4.通过认识历史的发展趋势，用于预测未来的发展结果



## 3.2 基于数据仓库的决策支持系统

### B. 基于数据仓库的决策支持系统

1. **数据仓库**保存大量的综合数据和历史数据，通过预测模型计算可以得到预测结果。
2. **在线分析处理**（OLAP）工具可以对数据仓库中的数据进行多维数据分析，即多维数据的切片、切块、旋转、钻取等，得到更深层意义的信息和知识。
3. **数据挖掘**技术能获取关联知识、时序知识、聚类知识、分类知识等。

**决策支持系统** = 数据仓库 + 在线分析处理 + 数据挖掘 + ...



## 第三节 数据仓库与数据挖掘结合

- 数据仓库和数据挖掘的区别与联系
- 基于数据仓库的决策支持系统
- 集大成者，商业智能





## 3.3 集大成者，商业智能

### A.商业智能基本概念

1.商业智能是数据仓库、在线分析处理和数据挖掘等相关技术走向商业应用之后形成的一种应用技术。

- 该应用技术收集汇总了与商业活动有关的各种数据，将其集成进数据仓库中。
- 它采用在线分析处理技术对商业活动进行实时的监控、分析，便于及时采取有效的商业决策，提升商业活动的绩效。
- 它应用数据挖掘技术对描述商业活动的数据进行挖掘，以获取有效的商业信息，从中提取商业知识，为企业发展寻找新的增长点。




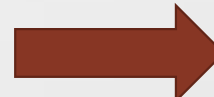
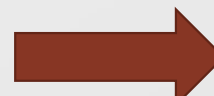


### 3.3 集大成者，商业智能

#### B.DW、OLAP、DM

举例：一位消费者在淘宝网上买了一本介绍金融知识的图书。

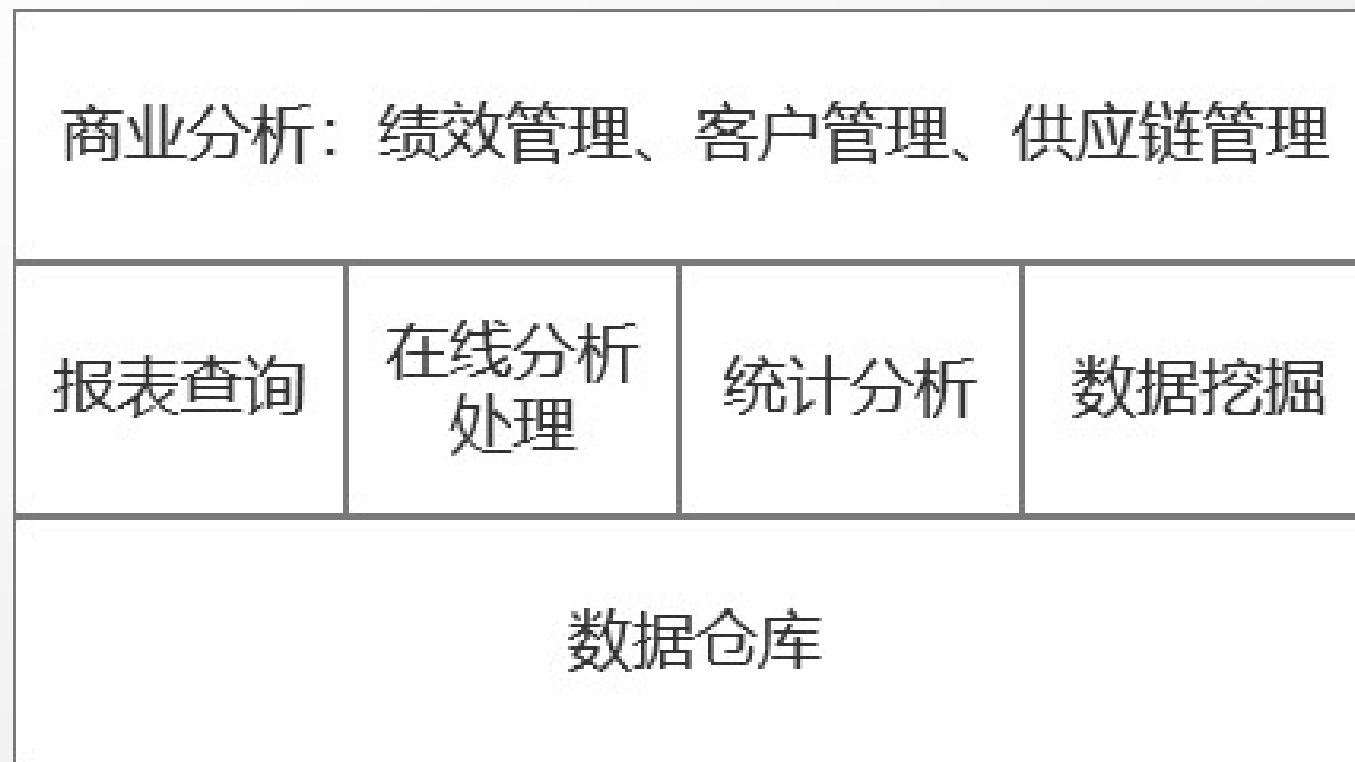
问题：

- 如何记录消费者的购买信息呢？  DW是基础
- 这本图书销售信息呢？其他金融类、证券交易类、经管类图书销售信息又如何？从时间维度、产品类型看呢？  OLAP是利器
- 如何向该消费者推荐其他图书呢？如何发现该类图书的潜在消费者群体呢？  DM是源泉



## 3.3 集大成者，商业智能

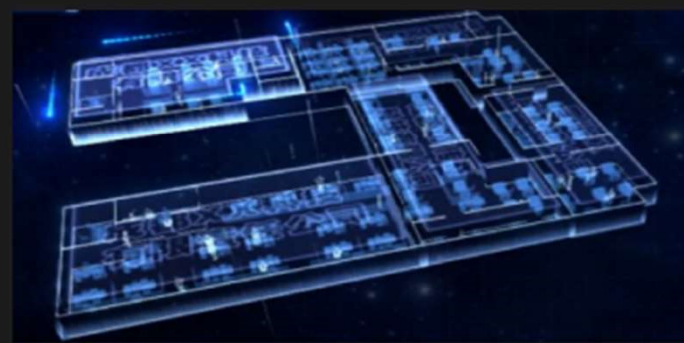
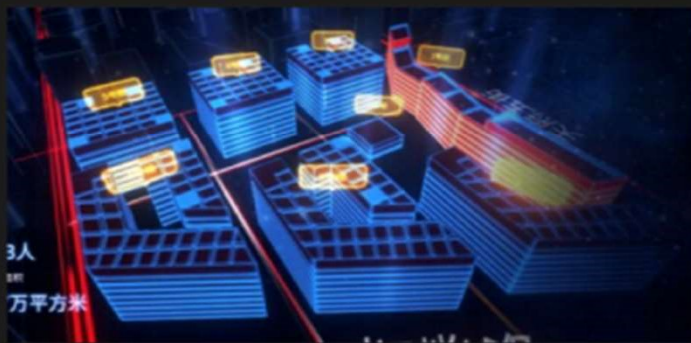
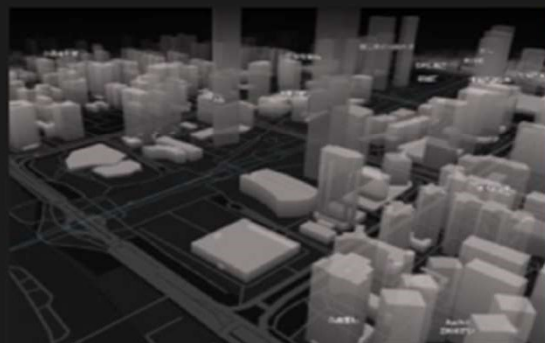
### C.商业智能体系结构





## 3.3 集大成者，商业智能

 DATAHUNTER





# 交通运输数据大屏

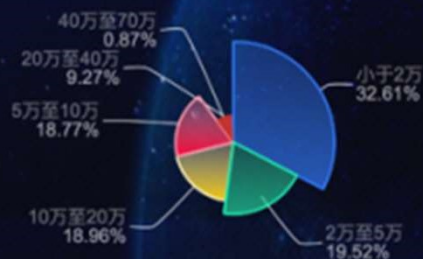
2020-07-22 17:36:09 星期三

本月新增用户  
8,238

本月活跃用户  
824



## ETC卡年金额消费区间车辆分布



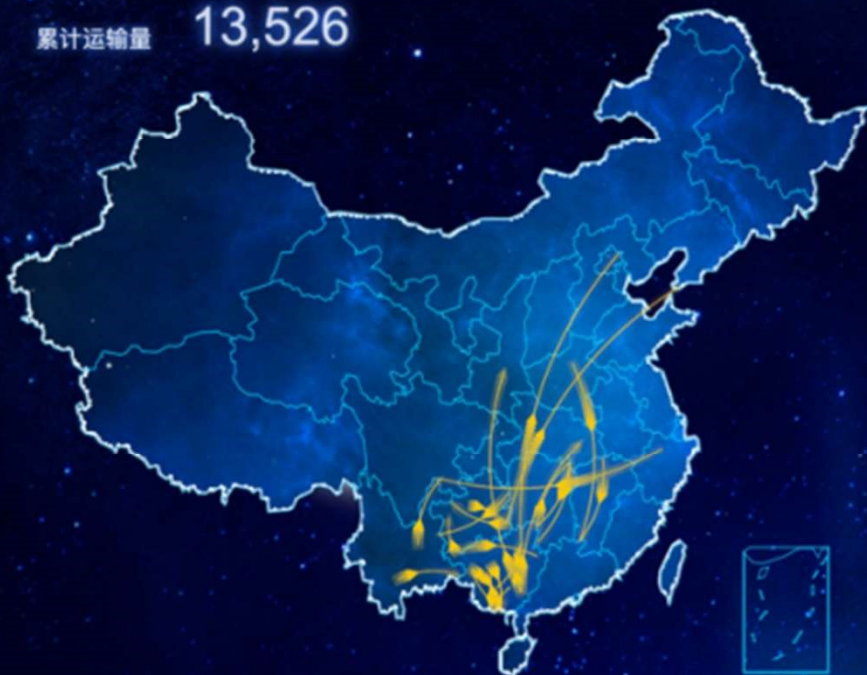
## ETC卡充值TOP企业

- NO 1 广西鑫盛物流有限公司
- NO 2 中国石油天然气运输公司广西分公司
- NO 3 广西华恒通能源科技有限公司
- NO 4 广西源盛仓储物流股份有限公司
- NO 5 广西盛源物流有限公司

物流线路

车源分布

累计运输量  
13,526

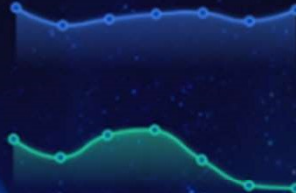


## IC卡道路运输证逐月申领量



本月订单量  
201 单

本月运单量  
132 单



## 线路运单情况

	运单分布	运输时效
玉林->湖北	2,021	20
玉林->西安	393	39
北京->湖北	2,521	25
上海->西安	395	39
西安->湖北	1,543	15
广州->西安	354	35

## 车辆类型统计

低速自卸货车  
轻型厢式货车  
重型仓栅式货车

# 智能工厂

18号厂房

2020年07月22日 17:34:20 星期三

本月计划

实际执行

526

323

本月计划执行准确率

96.20%

## 数控加工中心

## 泵车装配生产线

## 设备生产参数

SPM00031

立式加工中心  
SPM00031

操作员: 李建国

● 正在加工



当前加工产品

底板XPD10043TX.1.3A-1

当前刀具

3

当前坐标x

43.1850.644

当前坐标y

-2.4050.812

当前坐标z

-133.9650.767

进给速率

1002mm/min4

主轴速率

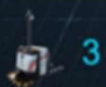
1002mm/min4

进给倍率

100%

## 物料配送

实时车辆



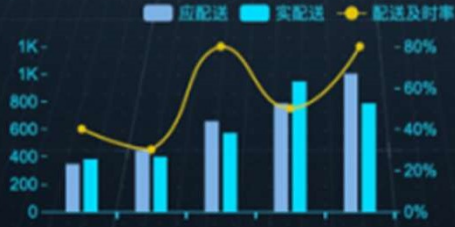
3

缺料呼叫

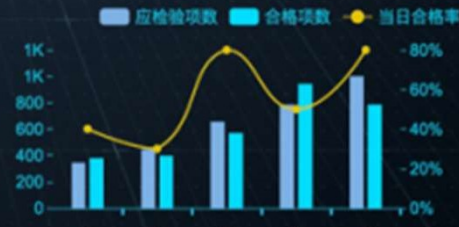


0

## 泵车线物料配送统计



## 泵车质检结果统计



## 质量控制

## 质量信息追溯

物料: 13806621

工位: 装配工位B

钢印: 4928382913

供应商: \*\*\*\*\*有限公司





### 3.3 集大成者，商业智能

SMARTBI EAGLE

自助数据分析平台

gz\_browser 

..... 

☒ 记住密码

登 录

效益，因管理而改变！管理，因我们(Smartbi)而改变！

&gt; 资源定制

数据管理

数据源

&gt; Demo数据源

&gt; JavaBean数据源

&gt; Javascript数据源

&gt; SYSTEM知识库

&gt; Session

&gt; Smartbi\_mpp

&gt; clickhouse

&gt; mongodb

&gt; mysql\_bigtable

&gt; vertica

跨库数据源

&gt; mongodb.test

&gt; mysql.northwind

&gt; 表关系视图

&gt; 计算字段

&gt; 过滤器

&gt; 业务视图

&gt; 联合数据源

&gt; 业务主题

数据集市

资源发布

公共设置

首页 定制管理 x



## 数据管理

数据源连接 业务主题  
数据权限 加载Excel数据  
导入模板



## 数据集

可视化查询 原生SQL查询  
SQL查询 存储过程查询  
Java查询 参数  
多维查询 X数据集



## 分析展现

透视分析 组合分析  
仪表分析 多维分析  
Web链接 X仪表盘



## 运维管理

用户管理 发布电脑主题  
新建任务 新建计划  
导入资源 导出资源  
备份知识库 恢复知识库  
系统日志 系统选项  
清空缓存 新建流程

更多 &gt;&gt;

数据分析任我行 管理决策更智能



张冬松  
[dszhang@nudt.edu.cn](mailto:dszhang@nudt.edu.cn)



# 谢谢! Q&A

THANKS FOR YOUR ATTENTION