大数据的大众化

孔宇华
Aster 业务部总监
大中华区

TERADATA. THE BEST DECISION POSSIBLE.

阻碍Big Data平民化的最大问题

- “阻碍Hadoop平民化的首要问题是其研究成本。不同的公司对hadoop有着不同方向的研究与优化，针对自己公司的业务又有着不同的改进方法。而对于绝大多数小公司来说，这样的研究与改进，无论从人力、物力还是时间等方面上，都是一笔巨大的成本。建议改进方案：首先要让更多的人知道hadoop，不仅仅了解他的名字，要让更多的人知道其架构、体系、原理，甚至具体实现等等...”

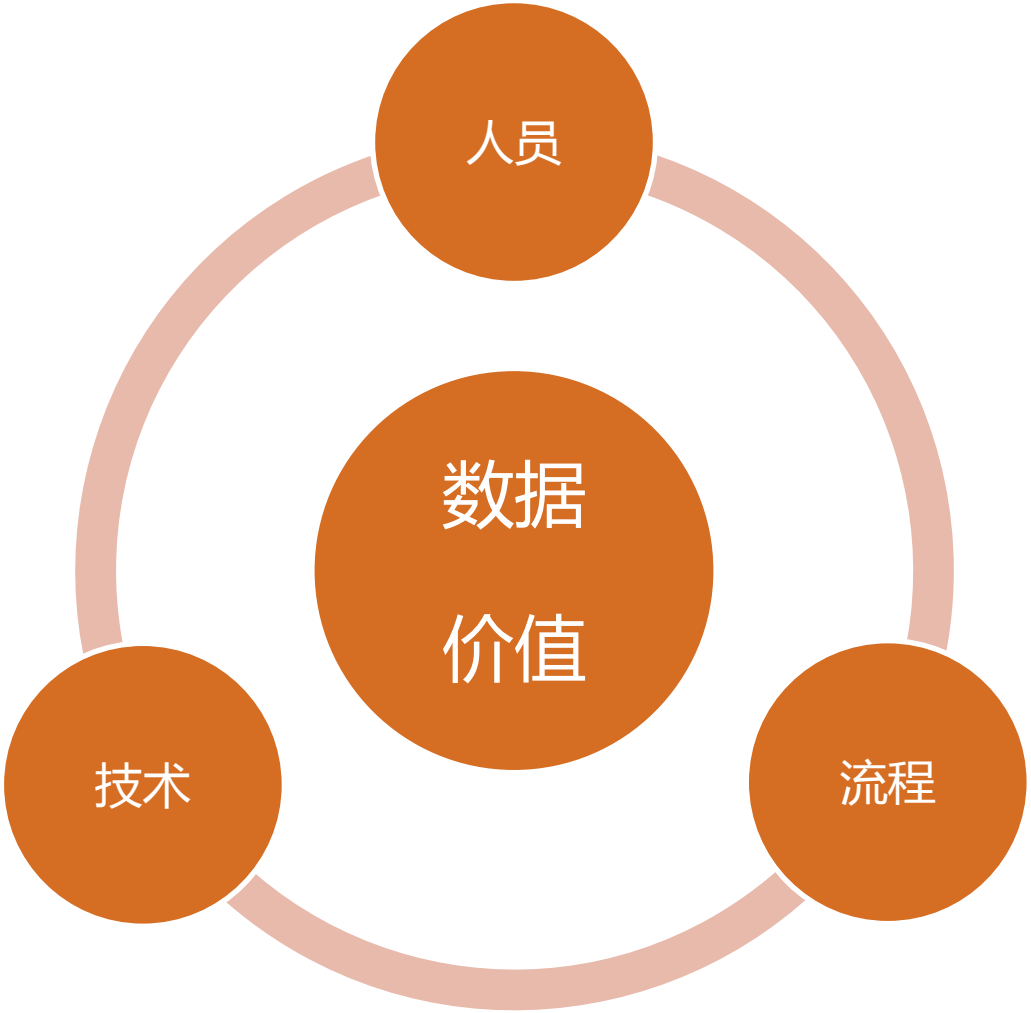
- 学习成本较高，出了Hive之外，其他的组件几乎都是一门全新的知识，Pig和HBase相对来说入门还比较容易，MapReduce程序简单的还好，复杂一点的程序编写程序是相当耗时间和非常考验编程能力的，Hive虽然支持SQL99的标准，但是很多数据库功能是没有的，用起来会非常的受限制，当然可以通过编写UDF函数来解决，但是据我自己编写UDF函数的经历告诉我，很耗时，而且不是那么容易上手的。

- 现阶段，Hadoop的实时性是不怎么好的，编程模型单一，比较适合离线的大数据的分析。

- 现有业务平台的迁移问题，比如：之前我在数据仓库、数据集市跑的一些报表啊之类的存储过程，到了Hadoop这里似乎就不是那么好迁移了，就算能迁移，付出的代价也是非常高的，作为企业的管理者，引入Hadoop是为了节省投入，目前阶段，肯定不会把大量逻辑复杂的存储过程放到Hadoop上面来跑，因为至少目前而言，是不会有这个想法的。

--CSDN 文章 (如何使Hadoop平民化?), 2012年11月

大数据的天时，地利，人和



用户情况

	技术公司	非技术公司
公司构成	许多计算机科学人才	计算机科学人才数量较少
人员技能	编程 (java, C, C++)	SQL
系统维护技能	编程人员	数据库管理员（DBAs）
关注点	灵活应用	易于使用

用数据库分析大数据

- 数据库一直都是作为数据分析的选择。
- SQL是高层次的,且易于重复使用
  - 适用于任何数据库结构
- 纯SQL可以用在大容量的数据
  - 已有许多上百TB级或PB级数据仓库

… 但这就够了吗?

- SQL 在一些问题上 匹配性能较弱
  - 有些问题用SQL繁琐，很难理解，或极难表达
  - 查询优化器做的选择比较低效
- User-defined functions（UDF）是一个不完整的修复
  - 不灵活
  - 不是并行设计
  - 跟数据模型关联很大，很难重复使用
- 大数据需求使得一些用户寻找其他平台

MapReduce “趋势”

- MapReduce 可扩展到巨大数据容量
  - 几个知名的互联网公司使用大规模MapReduce平台
  - 克服了传统数据管理系统缺乏的灵活性
- 优秀的编程模型
  - 容易理解
  - 具有平行处理数据的能力
- 但是，MapReduce需要操作者有良好的编程背景
  - 新的问题必须要写新的程序
  - 难以快速重复利用

...但我们失去了什么?

- 可重复使用的功能
  - 数据模型：模式，统计，局部优化
  - 通用算法:
    - 连接（joins）
    - 分组(grouping)
    - 排序(sorting)
- 为什么我们不能有
  - 可轻易重复使用的
  - 易用的
  - 能处理大容量的数据的分析工具?

SQL-MapReduce

- 可处理大容量的数据
  - 容易利用众多服务器的硬件资源
  - 容易管理，容易落地的系统
- 分析师易用
  - 用分析师熟悉的语言
  - 容易整合生态系统的工具
  - 现成的功能包，可以更简便地做出业务上的分析
  - 使软件开发人员能够创建可重复使用的功能给分析师
- 软件开发人员易用
  - 简单的编程模型
  - 以最大地自由提供有用的平台

SQL-MapReduce功能包（50+）列表（部分）

功能包	SQL-MR 函数	SQL-MR 函数说明
时间序列/路径分析	nPath	用于模式匹配的函数，使您可以在排序的行集合中指定模式，在匹配这些符号的行上指定其他条件，并从这些行序列中提取有用信息。
	Path generator	该函数将一组路径作为输入值，其中每个路径都是用户从头至尾使用的路线（页面浏览序列）。对于每个路径，该函数将生成正确格式化的序列和所有可能的子序列，以供“路径汇总器”函数进一步分析。路径中的第一个元素是用户访问的第一个页面。路径中的最后一个元素是用户访问的最后一个页面。
	Path starter	为特定父项生成所有子项，并计算其总数。
	Path summarizer	“路径生成器”函数的输出值是该函数的输入值。该函数用于计算节点总数。“节点”可以是普通的子序列，也可以是完成子序列。完成子序列是指序列与子序列相同的序列。完成子序列通过在序列末尾附加“$”来表示。
	Sessionization	用所有单机的数据建立相关的访问流程
统计分析	correlation	计算表中任意一对列之间的全局关联性。
	linear regression	输出由输入矩阵表示的线性回归模型的系数。
	logistic regression	为逻辑回归建立权重序列的一系列行函数和分区函数。
	weighted moving average	计算某时间序列中大量点的平均值，同时为较旧的值应用算术递减加权。
	histogram	为直方图提供生成分组数据功能的函数。
文本分析	text parser	处理文本字段的一个常规工具，可以拆分词语输入流，对它们进行阻止（可选），然后发出各个词语并计算每个词语的出现次数。
	Sentiment analysis	情感分析是从内容中提取用户观点（正面、负面、中立）的过程。它可以帮助客户分析各个用户对呼叫中心、社会媒体等内容的观点。
	Text categorization	这是根据一组文档培训朴素贝叶斯分类模型，并使用该模型预测新文档类别的一组函数。
关联分析	Affinity Analysis	生成同时购买的一系列商品或“购物篮”商品的数据记录，通常为交易记录或网页日志。
数据转换	pack	将多个列中的数据压缩到一个“打包”数据列中。
	unpack	获取一个“打包”列中的数据，并将其展开到多个列。
工具	XML Parser	XML 解析器函数是从 XML 文档中提取元素名称、属性名称和文本的一种工具。该函数的结果是一个展平表。

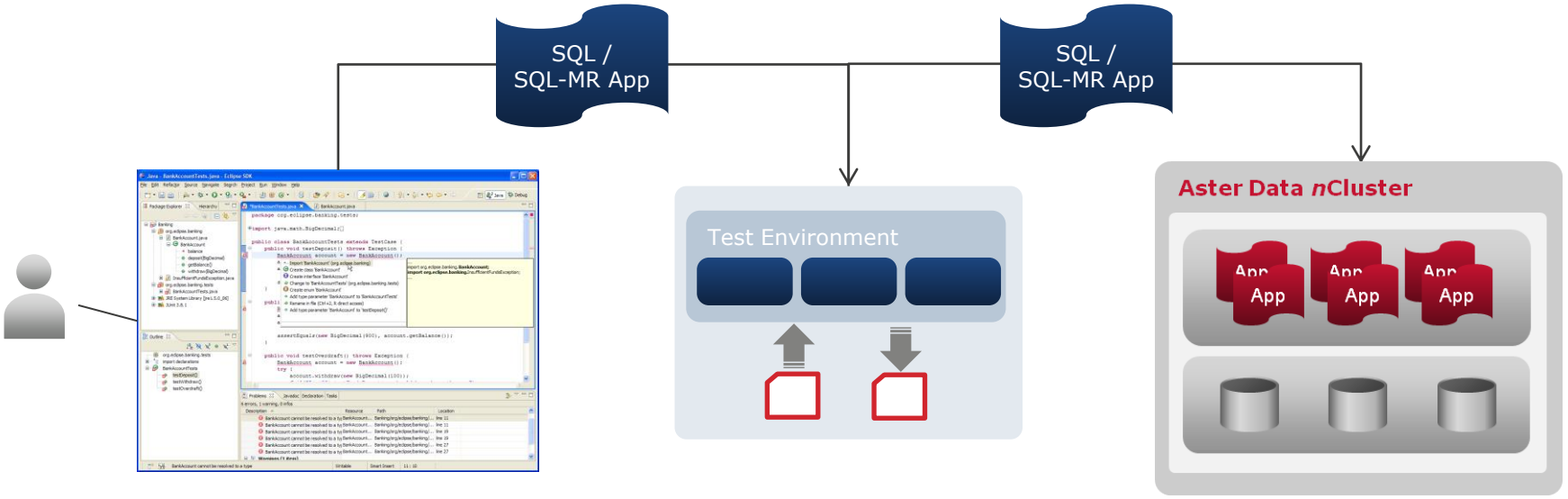
黄金通道分析

挑战：寻找十大最常见的从首页到购买页的路径.

```
SELECT path, count(*) as freq
FROM NPATH (
 ON page_event_fact
 PARTITION BY session_key
 ORDER BY page_event_timestamp
 MODE (NONOVERLAPPING)
 PATTERN ('^HOME.ANY*.PURCHASE$')
 SYMBOLS (page_key = 1 AS HOME,
                  TRUE AS B,
                  page_key = 20 AS PURCHASE  )
 RESULT( ACCUMULATE(page_key OF B) AS path  )
) T
GROUP BY path
ORDER BY freq DESC LIMIT 10;
```

Aster Developer Express: 简易的进行大数据功能包发展

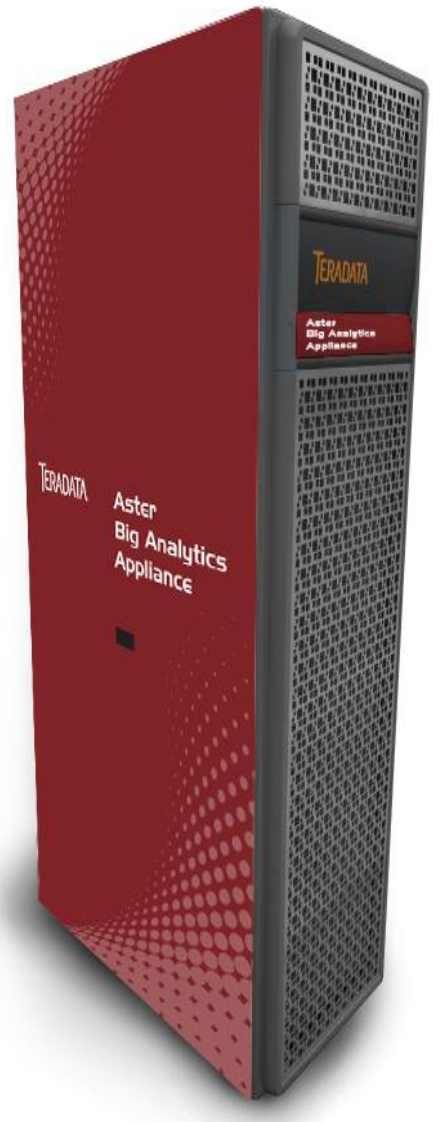
- 第一个集成MapReduce和SQL开发环境



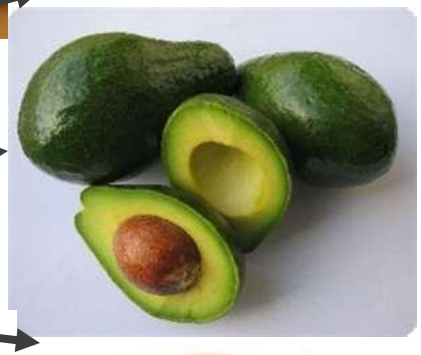
Teradata 统一数据架构

Teradata Aster大数据综合分析平台 3H领先优势

专利技术	SQL-MAPREDUCE专利技术是SQL和MAPREDUCE技术的完美结合，大规模并行处理，较MapReduce性能提升10-100+倍
大数据分析引擎	同时支持SQL和SQL-MAPREDUCE，预先集成了50+个SQL-MAPREDUCE功能包，同时允许对SQL-MAPREDUCE功能包的快速扩展
支持商用BI和ETL工具	Aster均支持商用的商业智能（BI）和转换加载（ETL）工具，如Tableau、Informatica等
开放、标准接口	支持ODBC、JDBC等标准访问接口
集成开发环境	基于Eclipse的可视化集成开发环境
高速互联技术	提供与Teradata等数据仓库进行高速互联的SQL-MAPREDUCE功能包（高速连接器）
集成Hadoop	行业内唯一同一机柜集成大数据分析平台和Hortonworks Hadoop的综合解决方案，并提供双向的高速互联通道，数据访问透明化
统一运维管理	提供统一的运维管理界面，同时可实现节点、交换机、磁盘、操作系统、数据库等集中化管理



Supervalu 案例- 关联分析



客户购买的主要商品是什么？对其他类别商品有何影响?

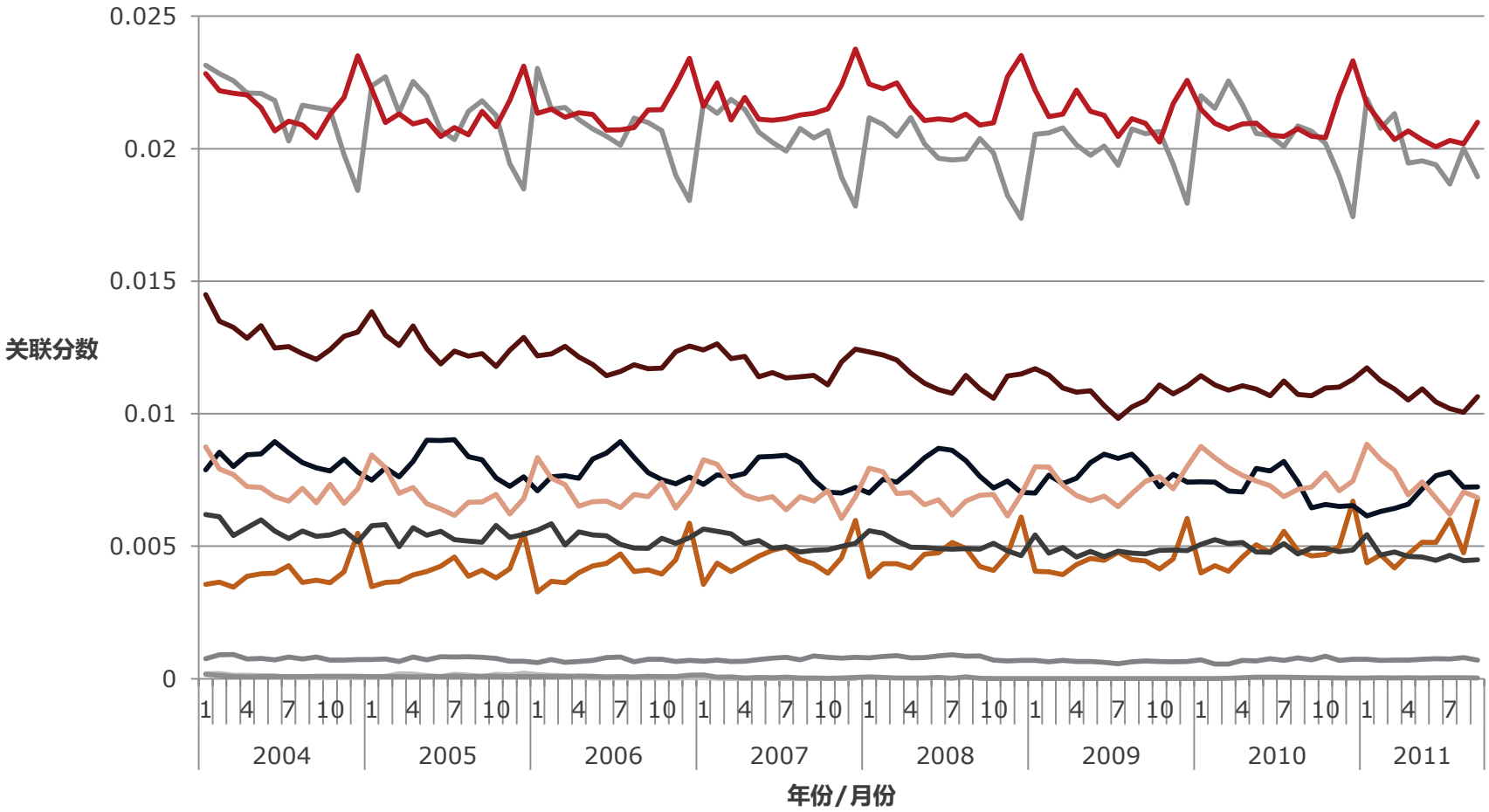
初期商用案例 – 关联分析结果

关联功能包的比较结果: 测试和生产

测试应用案例: 13周数据关联分析

测试分析应用案例: 8年数据关联分析

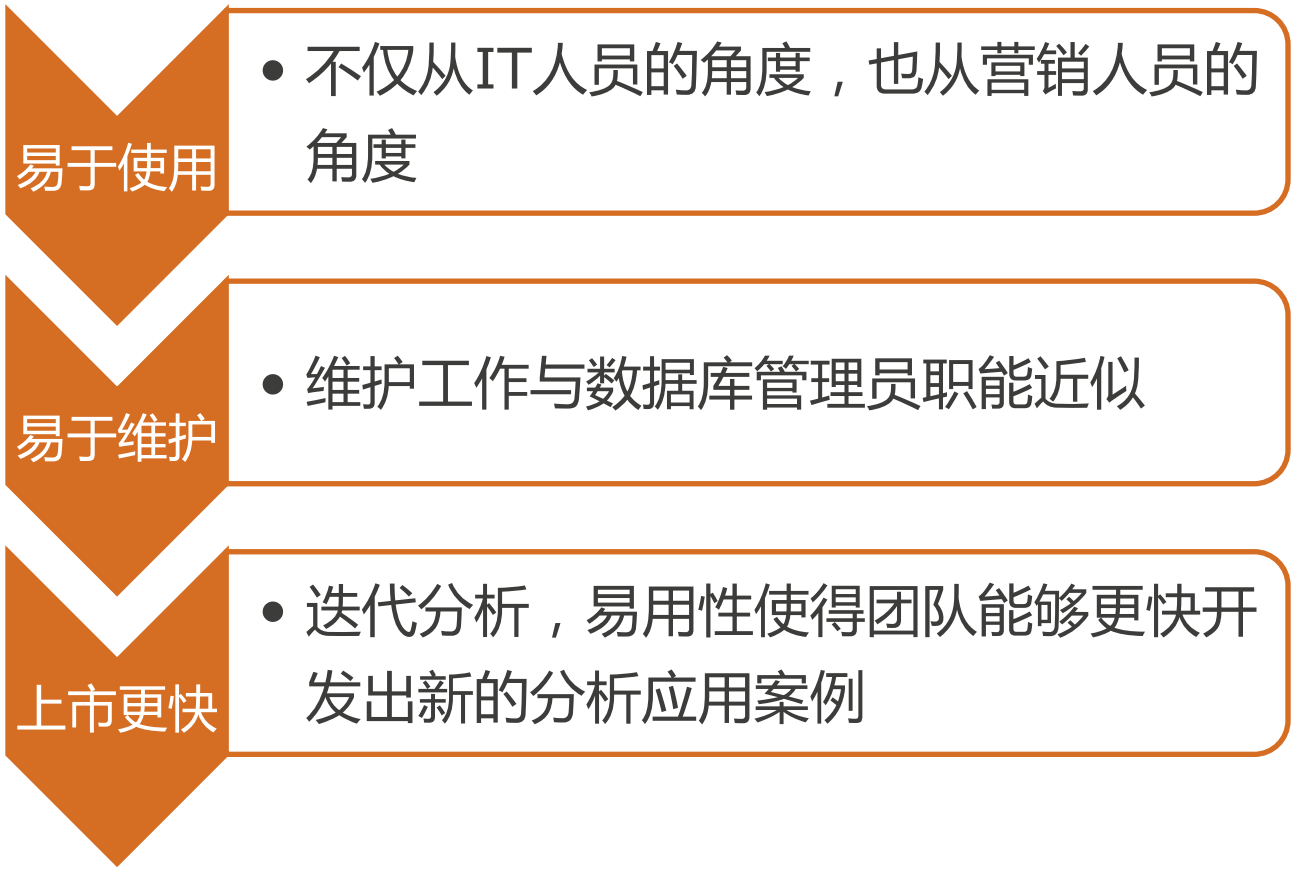
A类产品和其他类别产品随时间变化的关联



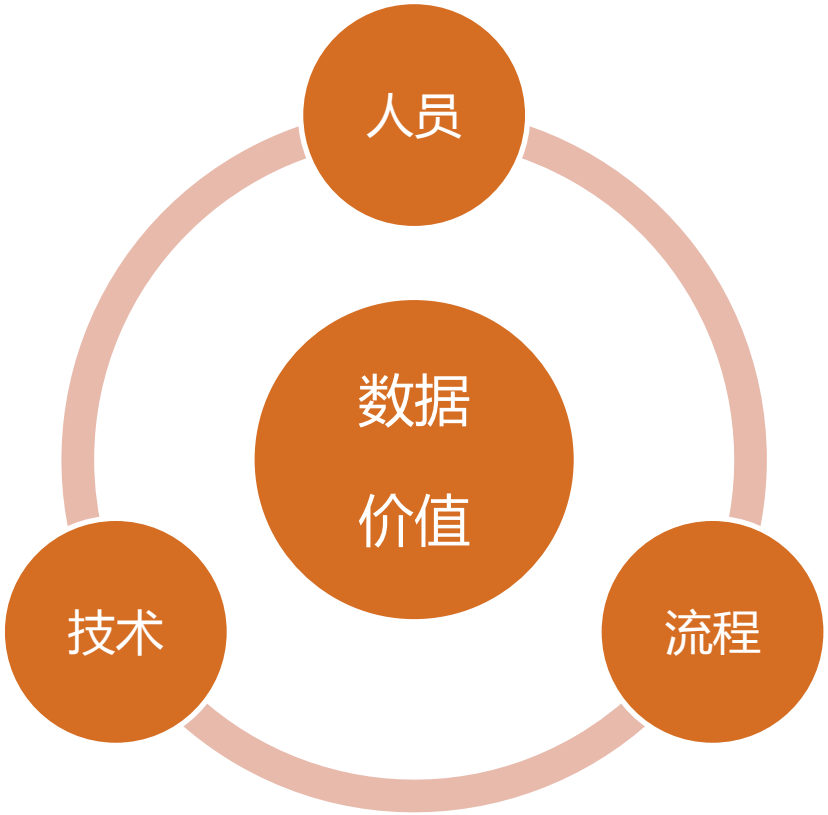
大数据令Supervalu实现

- 较长期分析
- 不同区域分析
- 不同年龄组分析
- 不同客户群分析
- 特别促销如何改善市场菜篮子组合分析
- 同一公司内部产品间的关联分析
- 跨类分析

Supervalu 大数据解决方案的优点



大数据的天时，地利，人和



- 哪些应用案例推动大数据的投资？大数据带来哪些价值？
- 从这一系统中，您有什么样的服务等级需求？响应时间？数据量？
- 使用人员有哪些技能？SQL？编程？
- 您需要什么技术支持这些应用案例？

TERADATA®

THE BEST DECISION POSSIBLE™

问题?