大数据分析挖掘技术在电商的应用

黄　晖　博士

上海天律信息技术有限公司

2014年6月

内容提要

1. 应对大数据：方法与趋势
2. 大数据分析挖掘技术
3. 大数据分析与电商应用

应对大数据-1： 公有云

阿里云 ODPS (Open Data Processing Service)

2010年2月第一版上线

集团内部生产机群规模18000台机器

单存储和计算机群最大规模5000台机器

日均处理3000万个作业请求，20万个计算任务

日均读3PB，写1PB数据；日均上传450TB，下载50TB数据

服务淘宝、支付宝、阿里金融等多项集团内部业务

支持淘宝贷款、数据模型、聚石塔等多款产品

目前处于公测阶段，今年2季度正式商用

应对大数据-2：自建分布式平台

硬件： PC服务器集群　（Google：　百万台服务器）

软件：

Hadoop	（分布式操作系统，管理服务器群）
HDFS	（分布式文件系统）
MapReduce	（分布式管理系统）
Hbase、 Cassandra	（分布式数据库）
Hive	（云端数据仓库）
Spark	（云端内存计算）
Markway	（分布式分析挖掘）
Pig Latin	（分布式数据处理语言）
Chukwa	（分布式数据采集）
ZooKeeper	（分布式协同工作和安全管理）

应用： 开店、存储、Email、OA、ERP、SCM、BI等等

应对大数据-3：虚拟化集群

硬件： 异构硬件的整合，大型机、小型机、PC机等等

软/硬件分离：

一个硬件运行多个不同操作系统

服务器虚拟化：

在一台物理服务器上
创建出多台虚拟服务器

系统虚拟化：

在一台物理机上同时运行
多个操作系统

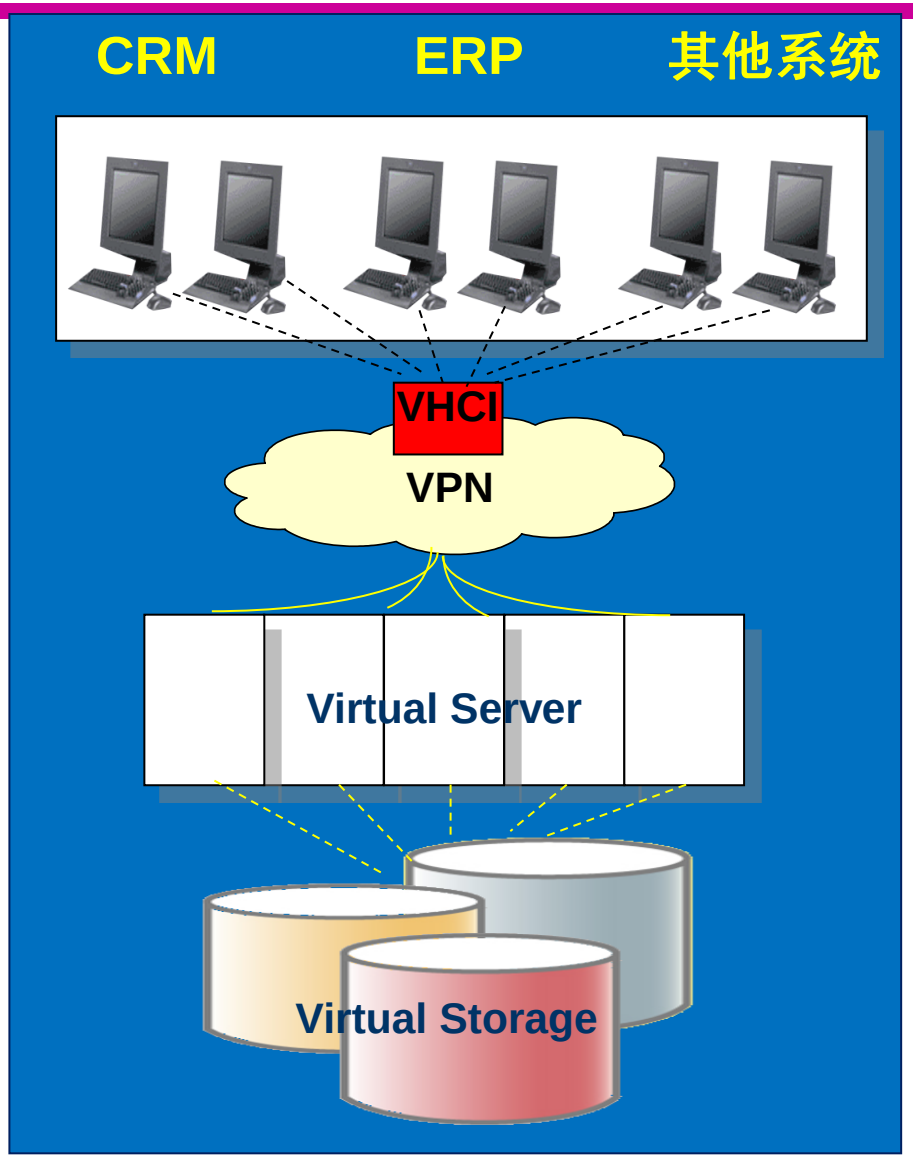
数据库集群：

多种或单种关系型数据库集群

应用虚拟化：

将应用程序与操作系统解耦合，为应用程序提供一个虚拟的运行环境

特点： 存储虚拟化、桌面虚拟化、应用虚拟化



应对大数据-4：内存计算

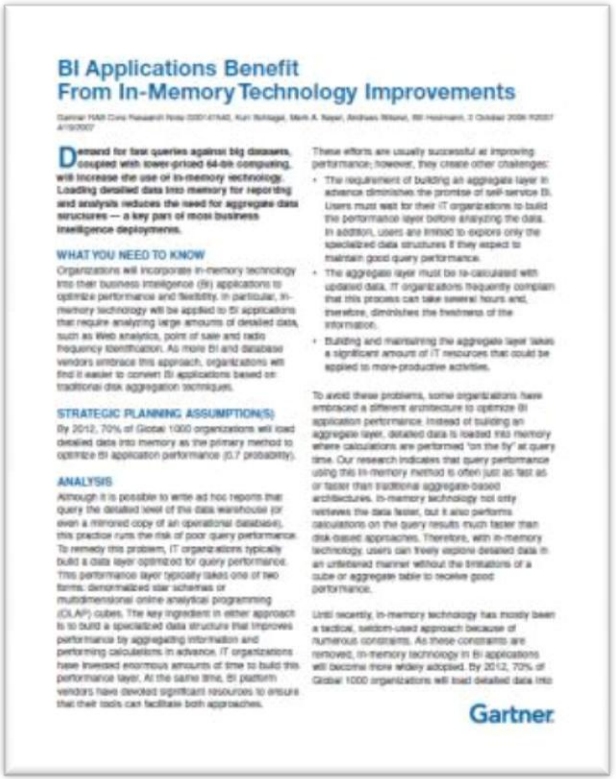
1.加速数据访问： 比磁盘快1,000,000倍

传统数据库

磁盘读取：5毫秒

内存数据库

磁盘读取：5纳秒



BI Applications Benefit From In-Memory Technology Improvements

Demand for fast queries against big datasets, coupled with lower-priced 64-bit computing, will increase the use of in-memory technology. Loading detailed data into memory for reporting and analysis reduces the need for aggregate data structures — a key part of most business intelligence deployments.

WHAT YOU NEED TO KNOW

Organizations will incorporate in-memory technology into their business intelligence (BI) applications to optimize performance and flexibility. In particular, in-memory technology will be applied to BI applications that require analyzing large amounts of detailed data, such as Web analytics, point of sale and radio frequency identification. As more BI and database vendors embrace this approach, organizations will find it easier to convert BI applications based on traditional disk aggregation techniques.

STRATEGIC PLANNING ASSUMPTION(S)

By 2012, 70% of Global 1000 organizations will load detailed data into memory as the primary method to optimize BI application performance (0.7 probability).

ANALYSIS

Although it is possible to write ad hoc reports that query the detailed level of the data warehouse (or even a mirrored copy of an operational database), this practice runs the risk of poor query performance. To remedy this problem, IT organizations typically build a data layer optimized for query performance. This performance layer typically takes one of two forms: denormalized star schemas or multidimensional online analytical programming (OLAP) cubes. The key ingredient in either approach is to build a specialized data structure that improves performance by aggregating information and performing calculations in advance. IT organizations have invested enormous amounts of time to build this performance layer. At the same time, BI platform vendors have devoted significant resources to ensure that their tools can facilitate both approaches.

These efforts are usually successful at improving performance; however, they create other challenges:

- The requirement of building an aggregate layer in advance diminishes the promise of self-service BI. Users must wait for their IT organizations to build the performance layer before analyzing the data. In addition, users are limited to explore only the specialized data structures if they expect to maintain good query performance.
- The aggregate layer must be re-calculated with updated data. IT organizations frequently complain that this process can take several hours and, therefore, diminishes the freshness of the information.
- Building and maintaining the aggregate layer takes a significant amount of IT resources that could be applied to more-productive activities.

To avoid these problems, some organizations have embraced a different architecture to optimize BI application performance. Instead of building an aggregate layer, detailed data is loaded into memory where calculations are performed "on the fly" at query time. Our research indicates that query performance using this in-memory method is often just as fast as or faster than traditional aggregate-based architectures. In-memory technology not only retrieves the data faster, but it also performs calculations on the query results much faster than disk-based approaches. Therefore, with in-memory technology, users can freely explore detailed data in an unfettered manner without the limitations of a cube or aggregate table to receive good performance.

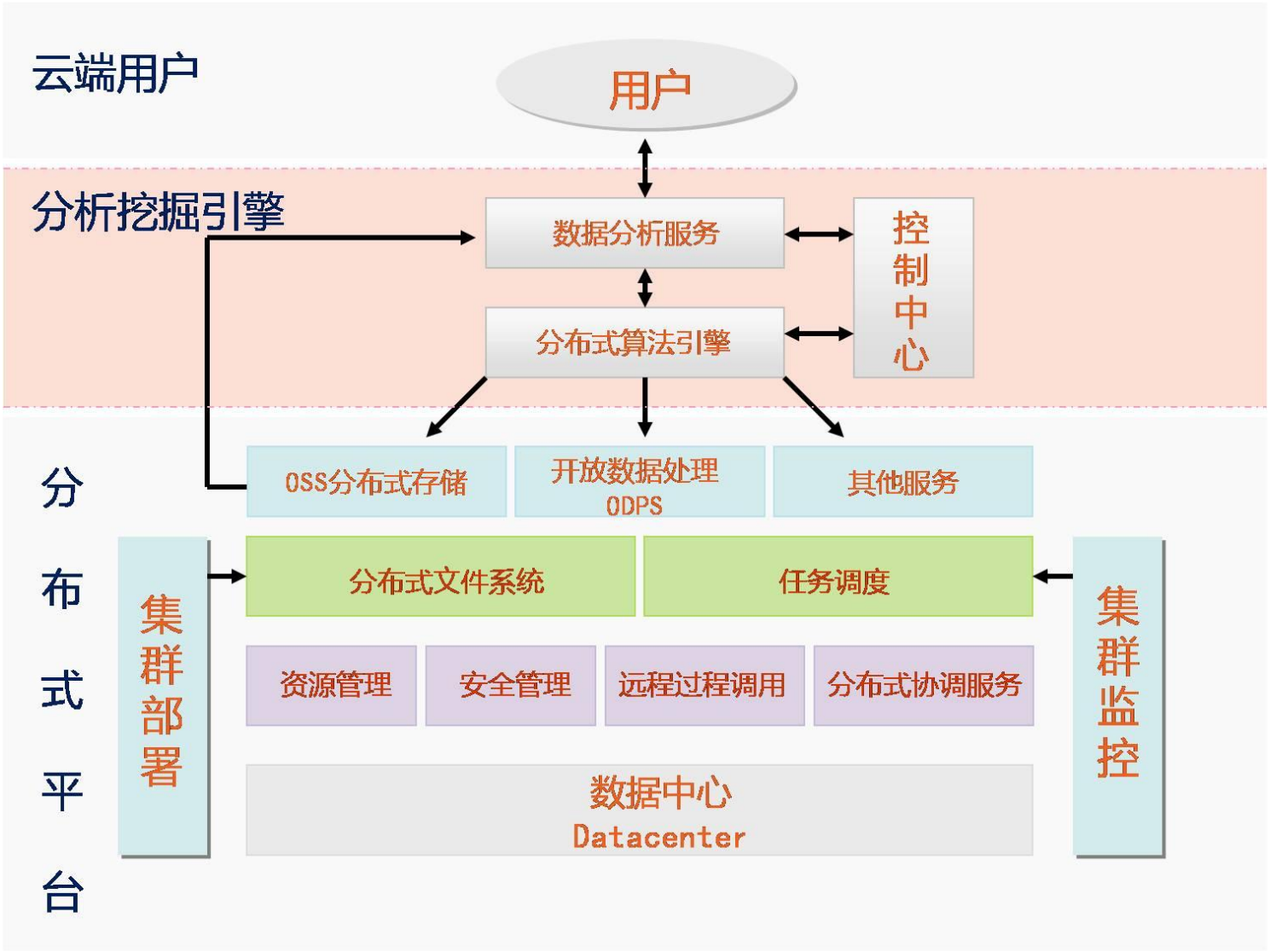
Until recently, in-memory technology has mostly been a tactical, seldom-used approach because of numerous constraints. As these constraints are removed, in-memory technology in BI applications will become more widely adopted. By 2012, 70% of Global 1000 organizations will load detailed data into

Gartner

“到2012年，70% 的全球1000强企业会将明系数据导入内存，以提升商务智能应用的性能。”

– Gartner

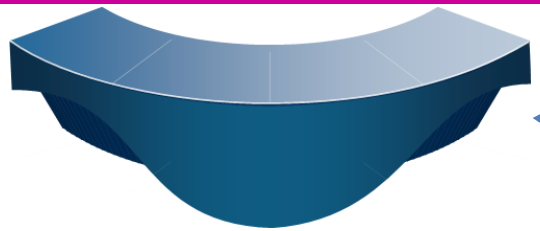
二、 大数据分析挖掘：马克威分布式算法



1、传统分析挖掘引擎

瓶颈：

- 无法应对大规模数据的挑战
- 无法利用多台机器资源
- 无法分析Internet数据源



分析挖掘引擎

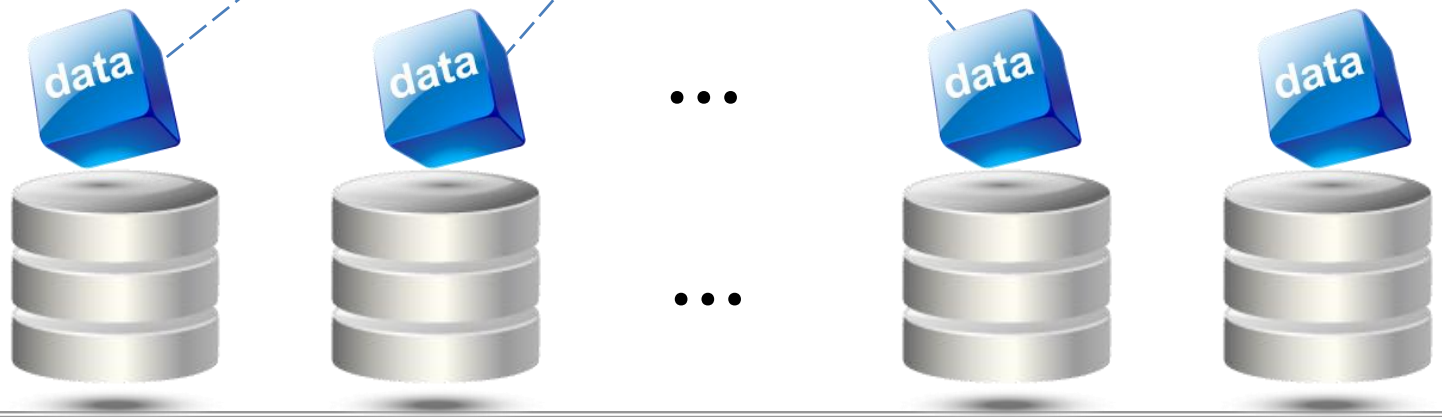
计算中…

等待计算

…

等待计算

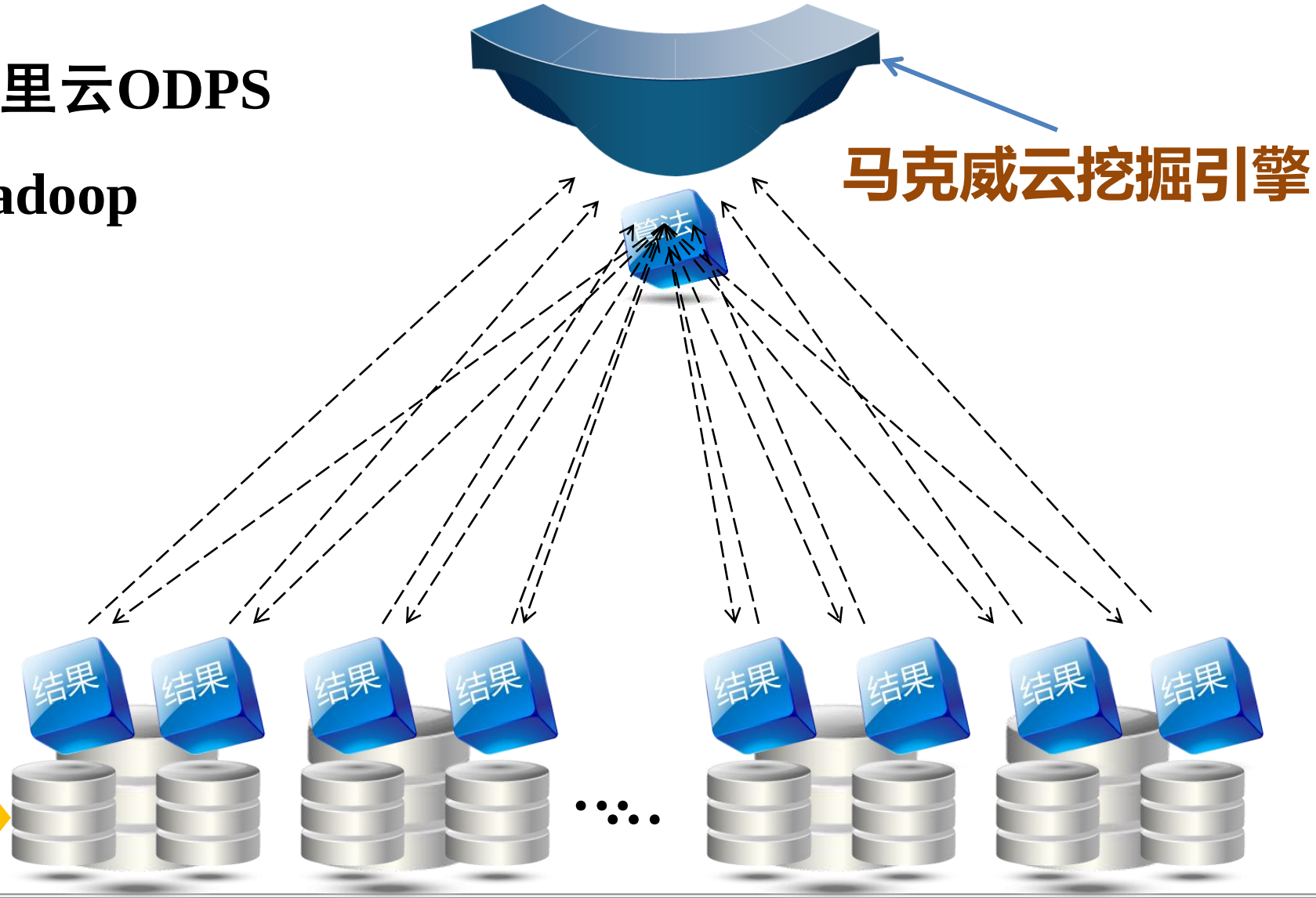
数据等待队列



数据源

2、分布式分析挖掘引擎

- ✓阿里云ODPS
- ✓Hadoop



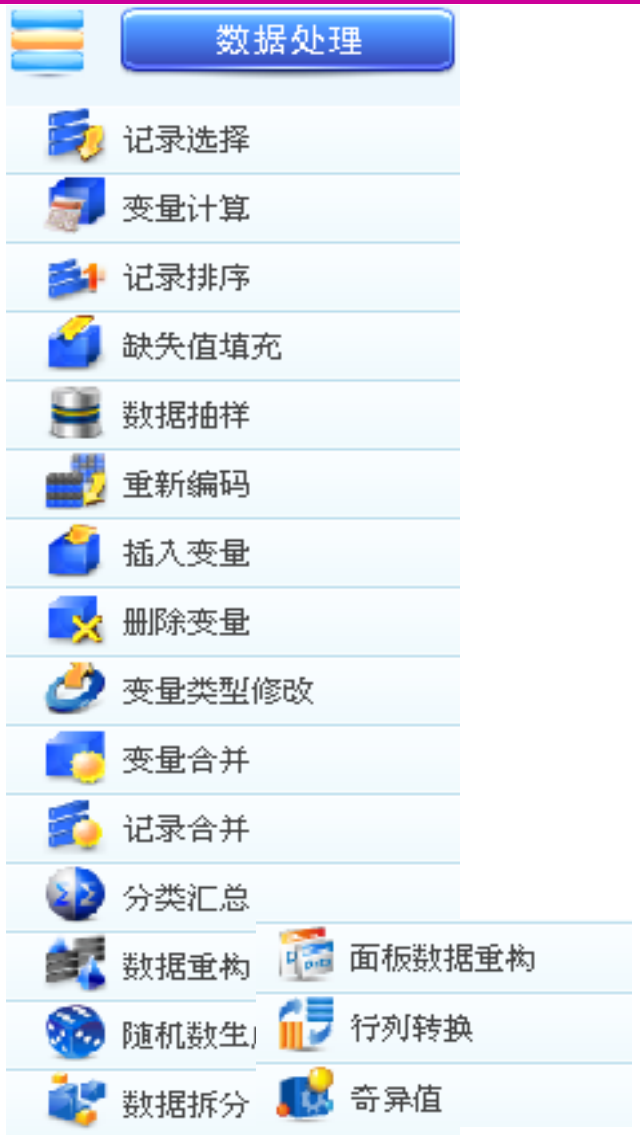
3、马克威云挖掘系统

- 基于阿里云飞天平台
- 基于Hadoop/MapReduce、
- 支持TB/PB级数据分析挖掘
- 可视化工作流操作模式
- 基于WEB服务的B/S架构

运行性能			
数据量	运行时间	服务器台数	Map数
10亿条记录，68个变量	25秒-5分钟	100台	736
176亿条记录，68个变量（3T）	36秒-30分钟	100台	11708

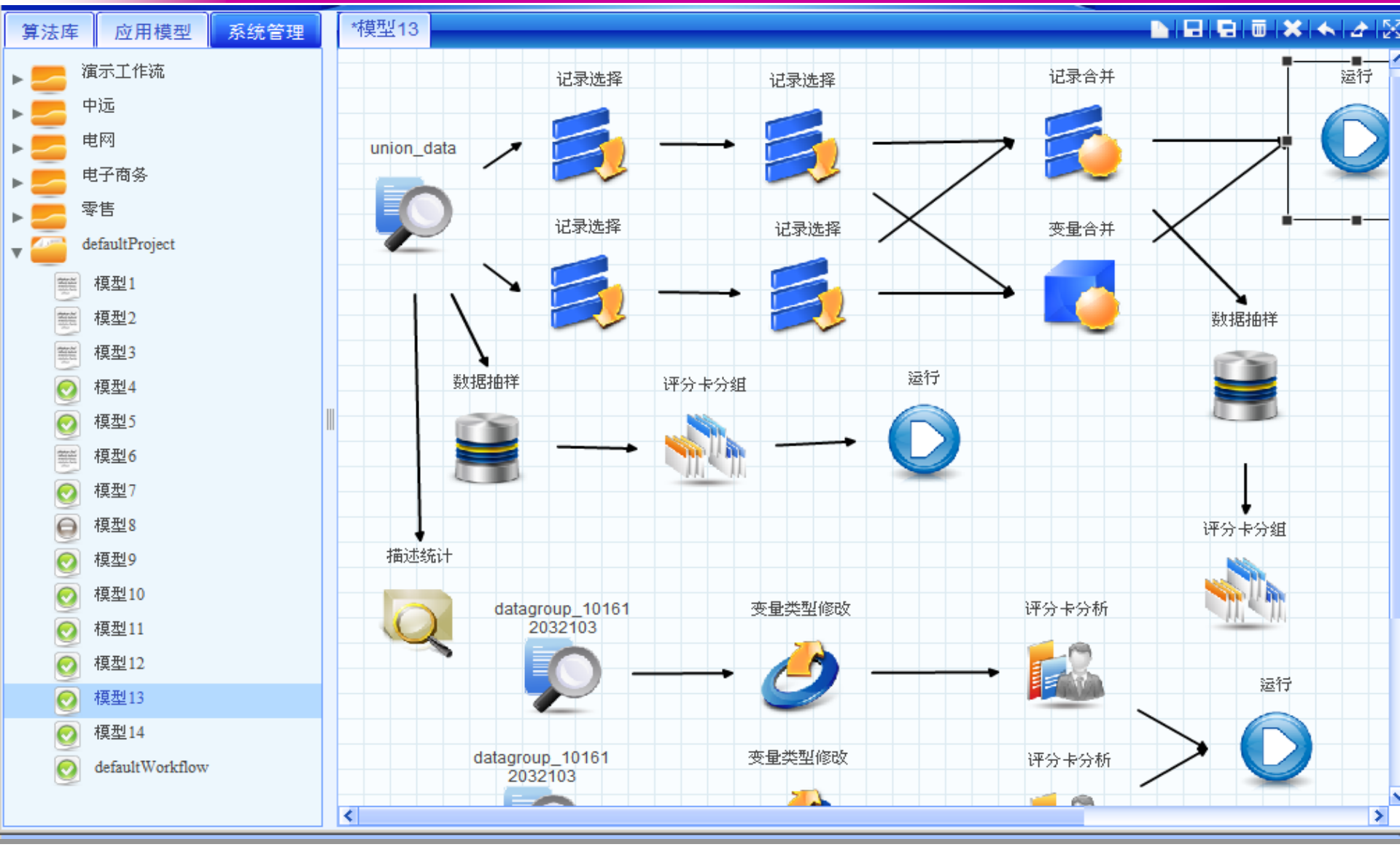
马克威云挖掘算法体系


数据源
单表数据源
多表数据源
HDFS数据源
HBase数据源


数据处理
记录选择
变量计算
记录排序
缺失值填充
数据抽样
重新编码
插入变量
删除变量
变量类型修改
变量合并
记录合并
分类汇总
数据重构
随机数生
数据拆分
面板数据重构
行列转换
奇异值


数据分析
决策树
孤立点分析
均值分析
频率分析
描述统计
曲线回归
快速聚类
二值逻辑回归
评分卡分組
评分卡分析
主成分分析
分层聚类
线性回归
模糊聚类
关联规则
变量聚类
支持向量机

马克威可视化工作流操作界面



4 主要客户

企业：阿里巴巴、余额宝、中信21世纪、国家电网、中国核电集团、上海宝钢集团、武汉钢铁集团、中国海运集团、中国远洋集团、海南航空、上海电信、中国移动（江苏）、重庆百货、上海广电集团、华氏医药等等

政府：国家统计局、国家海关总署、 2010上海世博会、中国人民解放军总参谋部、国家水利部、北京市发改委、上海市发改委、北京市统计局、上海市统计局、广州市统计局、福建省统计局、海南省统计局、云南省统计局、上海市公安局、上海市卫生局、上海市信访办、上海嘉定区政府、上海静安区商委等等

高校：华中科技大学、南京财大、中南大学、江西财大、上海金融学院、上海中医药大学、中央民族大学、新疆财大、解放军信息工程大学、东华大学、南京林业大学、山东曲阜师大、成都信息工程大学、哈尔滨理工大学、青岛理工大学、天津商业大学等等

三、大数据挖掘技术在电商的应用

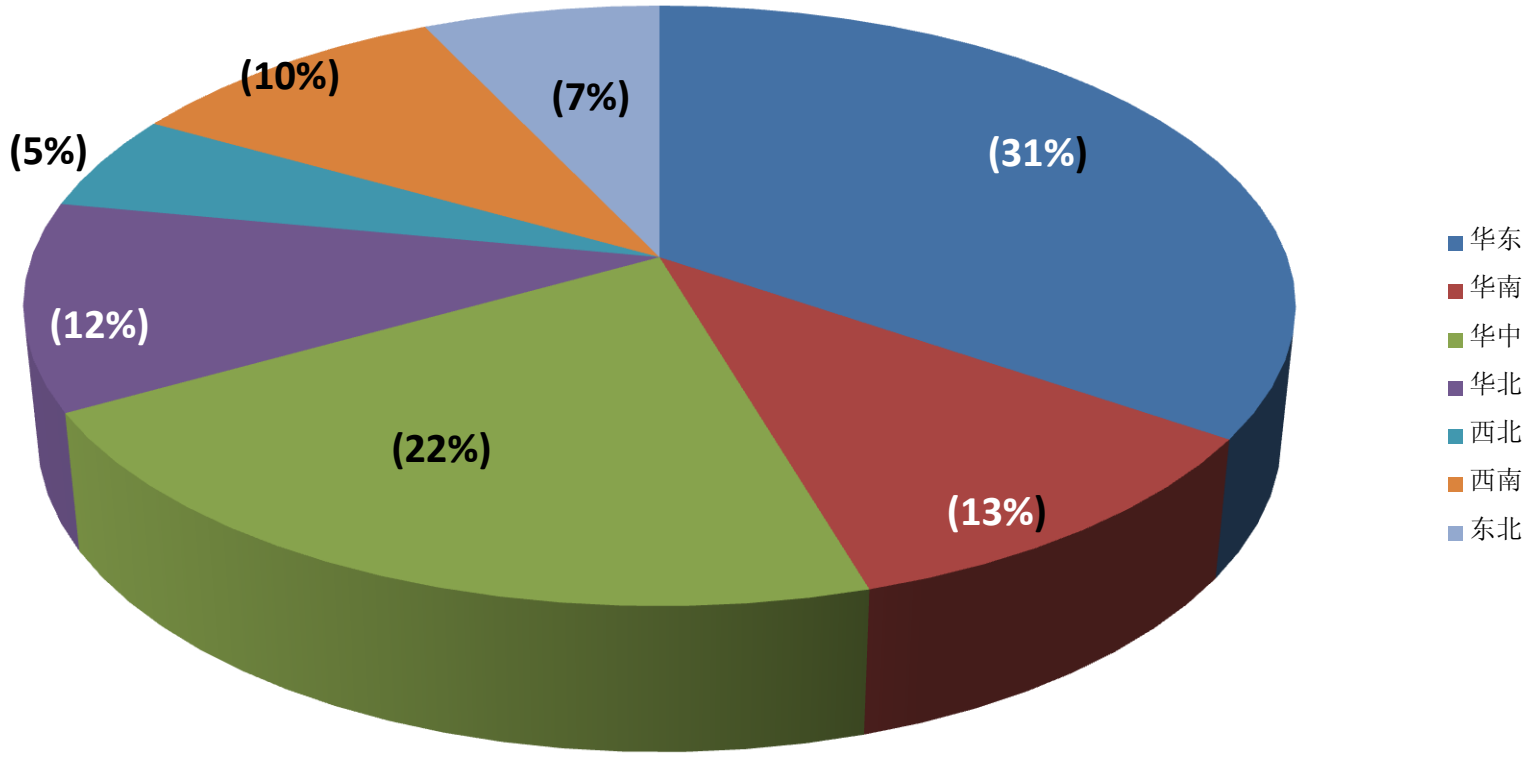
总量与构成	● 描述统计、频率分析、
趋势变化	● 时间序列、小波理论、比较
关联分析	● 聚类、回归、二值逻辑、关联规则、决策树
预测预警	● 神经网络、支持向量机、面板模型、贝叶斯网络

3.1 总量与构成

总量与构成

- 客户构成：地区、购买金额、频次、客单价
- 销量构成：品类数量结构
- 销售额构成：收入与品类贡献占比
- 利润构成：商品、客户对利润的贡献率
- 点击率和转化率：点击客户数，转化客户数

客户地区构成



交叉分析

客户价值分析： 谁贡献了多少

销售额与客户购买额分组：

销售总额 * 客户购买额 分组　　单位：万元

购买额	本组占比	本组客户价值（万元）	客户平均价值(万元)
《=1	9%	90	6
1--5	16%	180	21
6--10	40%	400	32
11-20	18%	210	17
21-50	12%	130	133
50-100	10%	108	133
》100	5%	52	140

搭配销售

时段选择： 昨天 | 最近7天 | 最近15天 | 最近30天 | 最近90天 | 本周 | 本月 | 上周 | 上月 | 2014-4-1日至2014-6-1

自定义时间： 开始时间 2014-4-1 愚人节 结束时间 2014-6-1 儿童节 确定

宝贝类目： 所有分类 > 服装 > 女装

（促销搭配建议清单）

2014年新款夏装情侣运动装

价格：76 元

相关联的宝贝	可能性	支持度
2014年新款红白经典搭配情侣装	100%	7. 8%
2014年新款浅灰加深蓝搭配情侣装	85%	5. 67%

2014年新款经典款夏装情侣运动装

价格：58 元

相关联的宝贝	可能性	支持度
2014年新款带帽子情侣装	95%	6. 5%
2014年新款绿色口袋情侣装	90%	5. 16%
2014年新款蓝色典雅情侣装	80%	4. 31%

使用算法：关联分析

2、趋势与对比

发展曲线：

- 销售额的趋势
- 销量趋势：品类
- 点击率和转化率的趋势
- 客户人数的趋势

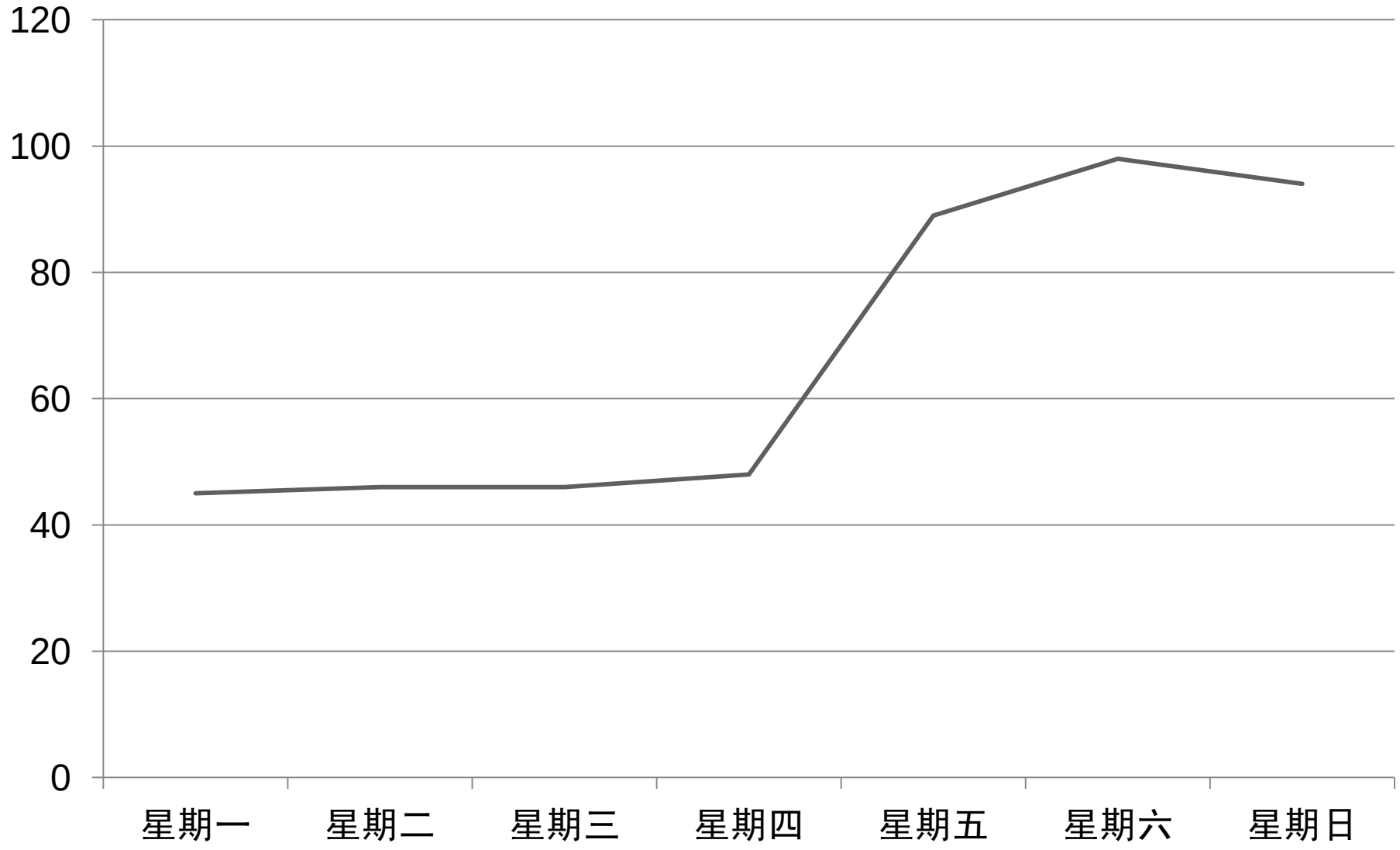
波动规律：

- 周期性：30天、60天、75天、、、
- 小波、大起大落
- 周变化规律、月、季度
- 节假日变化规律

价格趋势


趋势、预测： 澳粉价格(元/吨)
1,200
1,150
1,100
1,050
1,000
2010.01
2010.04
2010.07
2010.10
巴西精粉价格(元/吨)
预测值

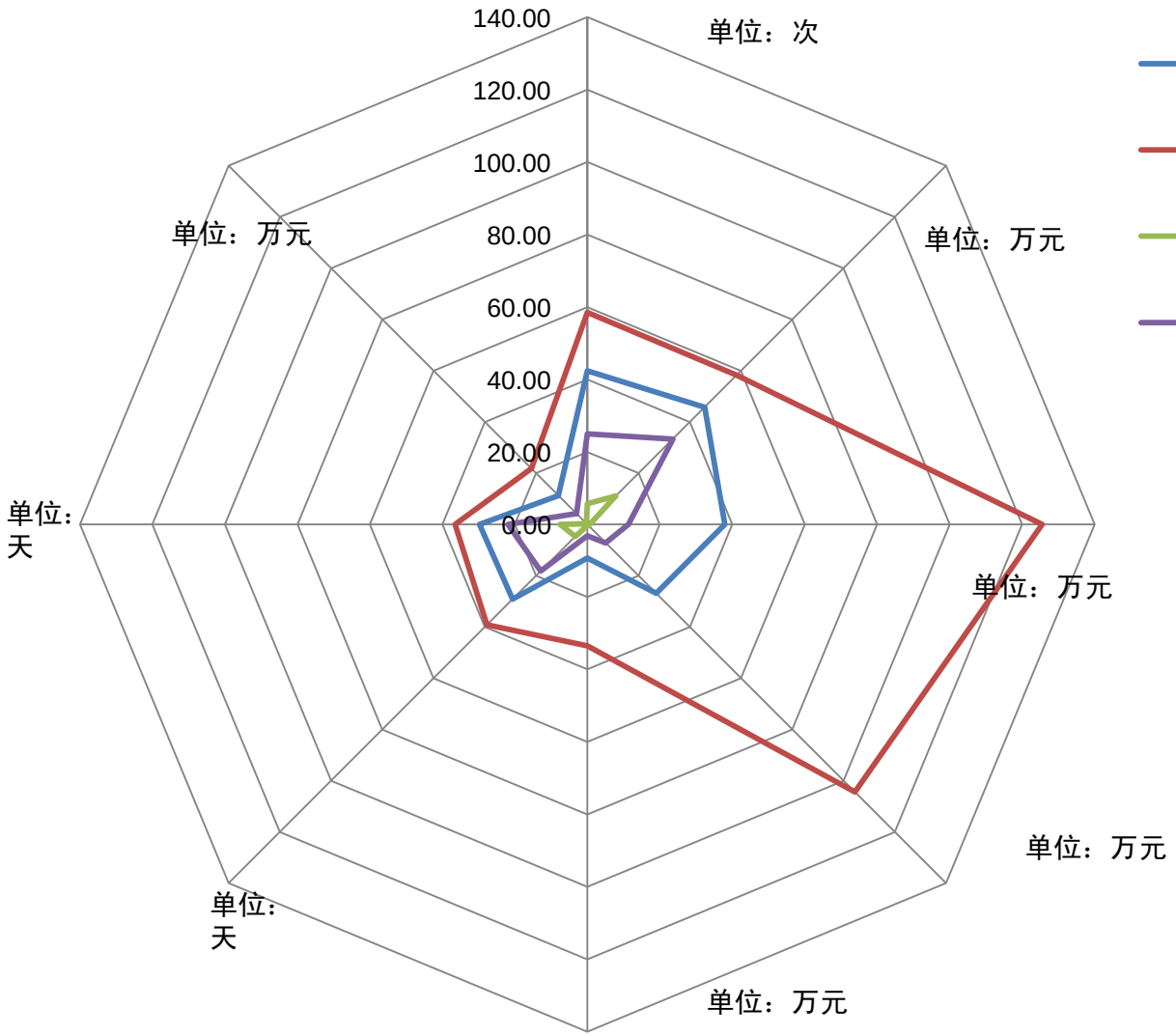
周内波动规律图



3.3 关联分析

1. • 客户行为聚类
2. • 客户价值分析
3. • 客户忠诚度分析
4. • 新产品与销量
5. • 促销与销量
6. • 广告与销量

客户聚类



类	人数	占比
类1	1449	5.56%
类2	236	3.42%
类3	49456	81.40%
类4	5442	10.62%

客户价值聚类与分析

。

- 具体分类为：

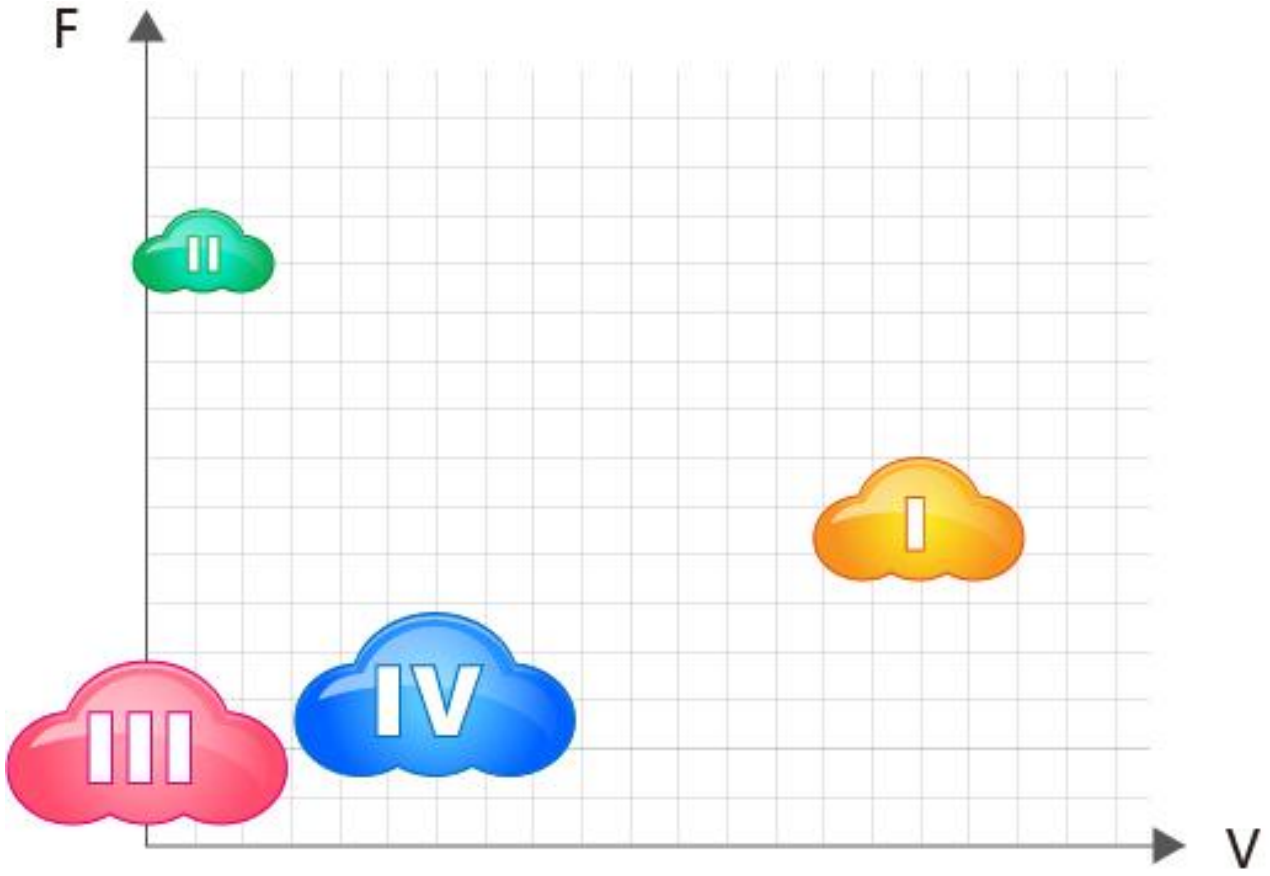
客户类型I：　价值高、购买频率高

客户类型II：　价值低、购买频率高

客户类型III：价值低、购买频率低

客户类型IV：　价值高、购买频率低

客户价值聚类分布示意图



图形说明

I：价高频高

II：价低频高

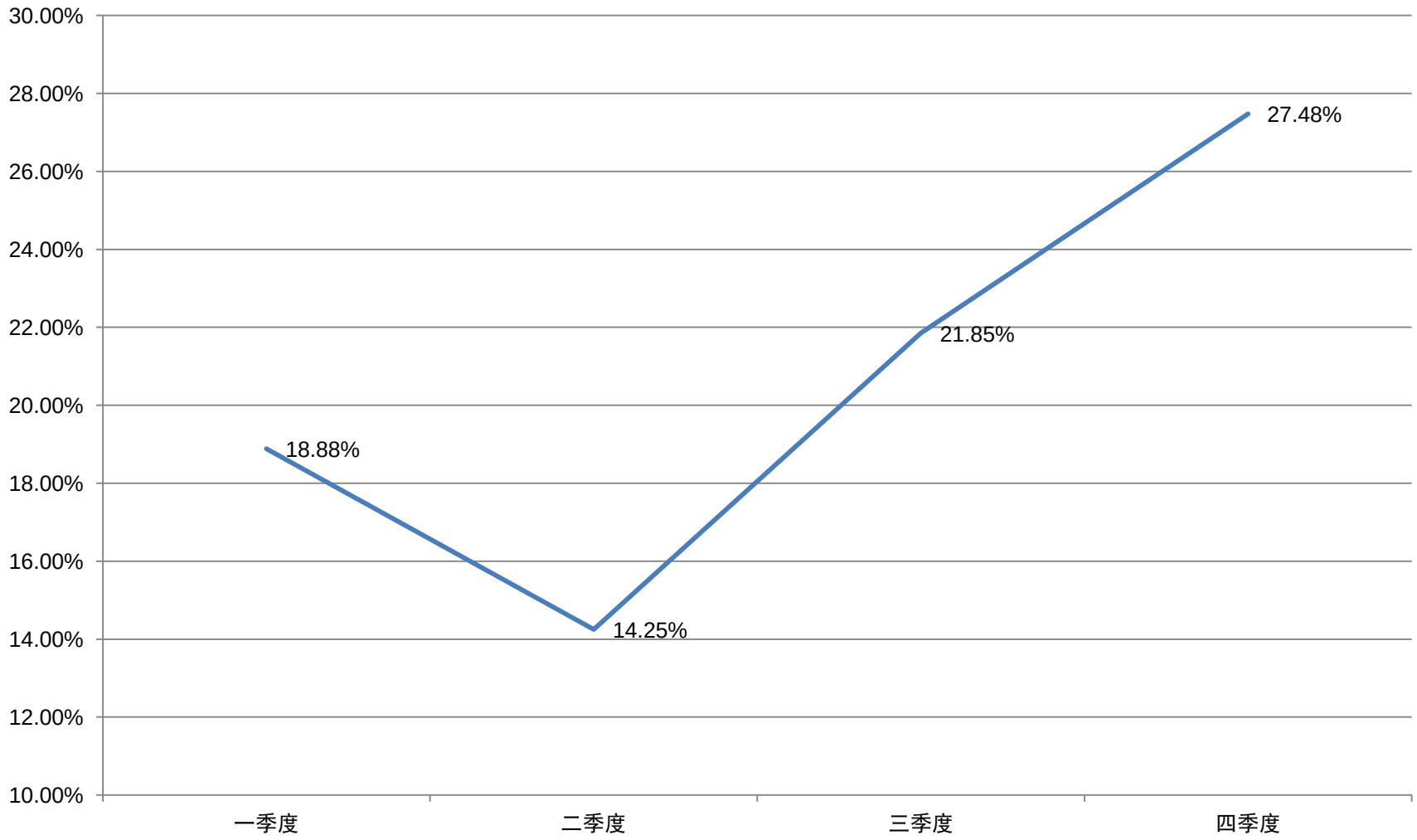
III：价低频低

IV：价高频低

3.4 预测

1. • 客户流失预测
2. • 销量预测
3. • 库存需求预测
4. • 新产品客户响应率预测

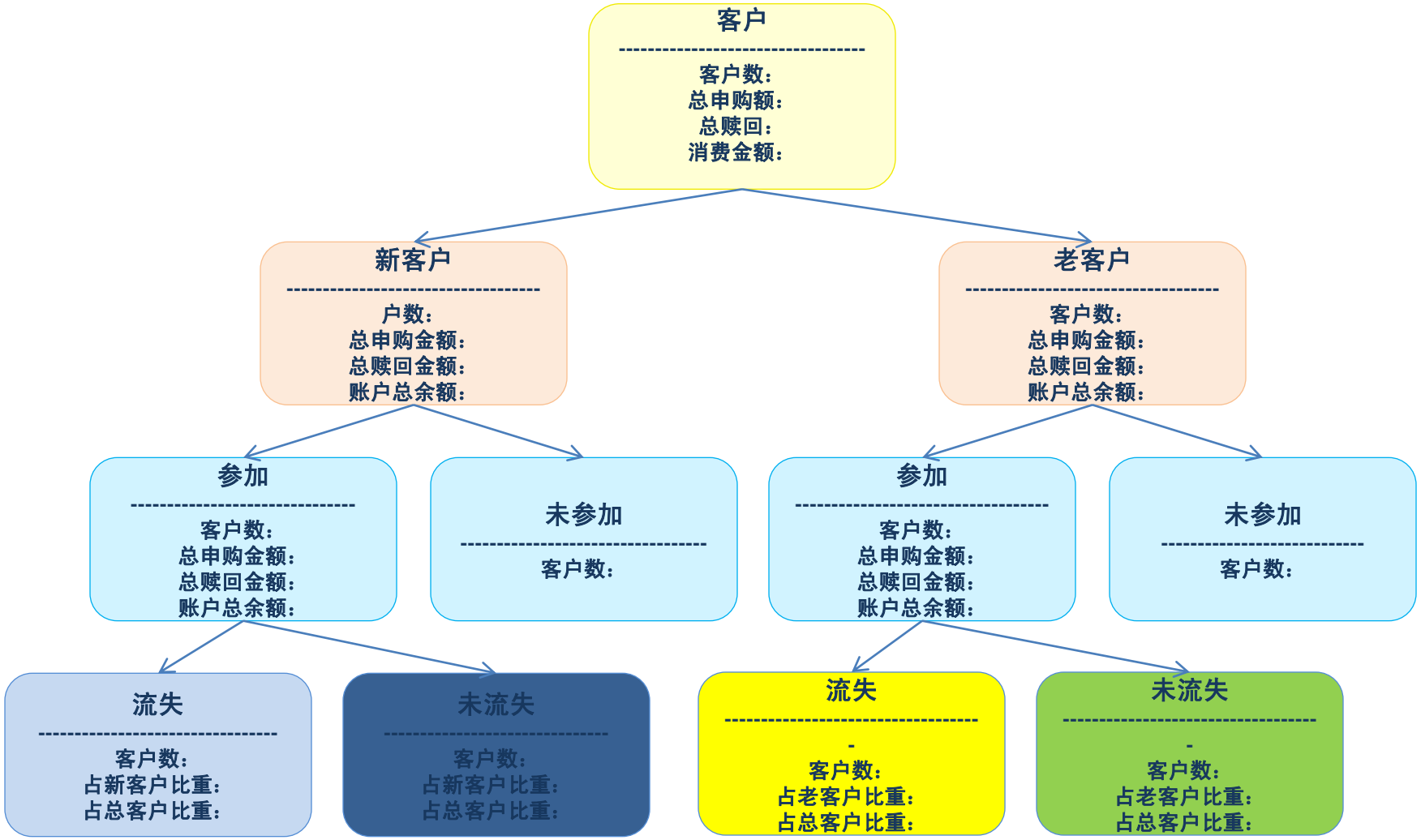
预测：客户流失


30.00%
28.00%
26.00%
24.00%
22.00%
20.00%
18.00%
16.00%
14.00%
12.00%
10.00%
18.88%
14.25%
21.85%
27.48%
一季度
二季度
三季度
四季度

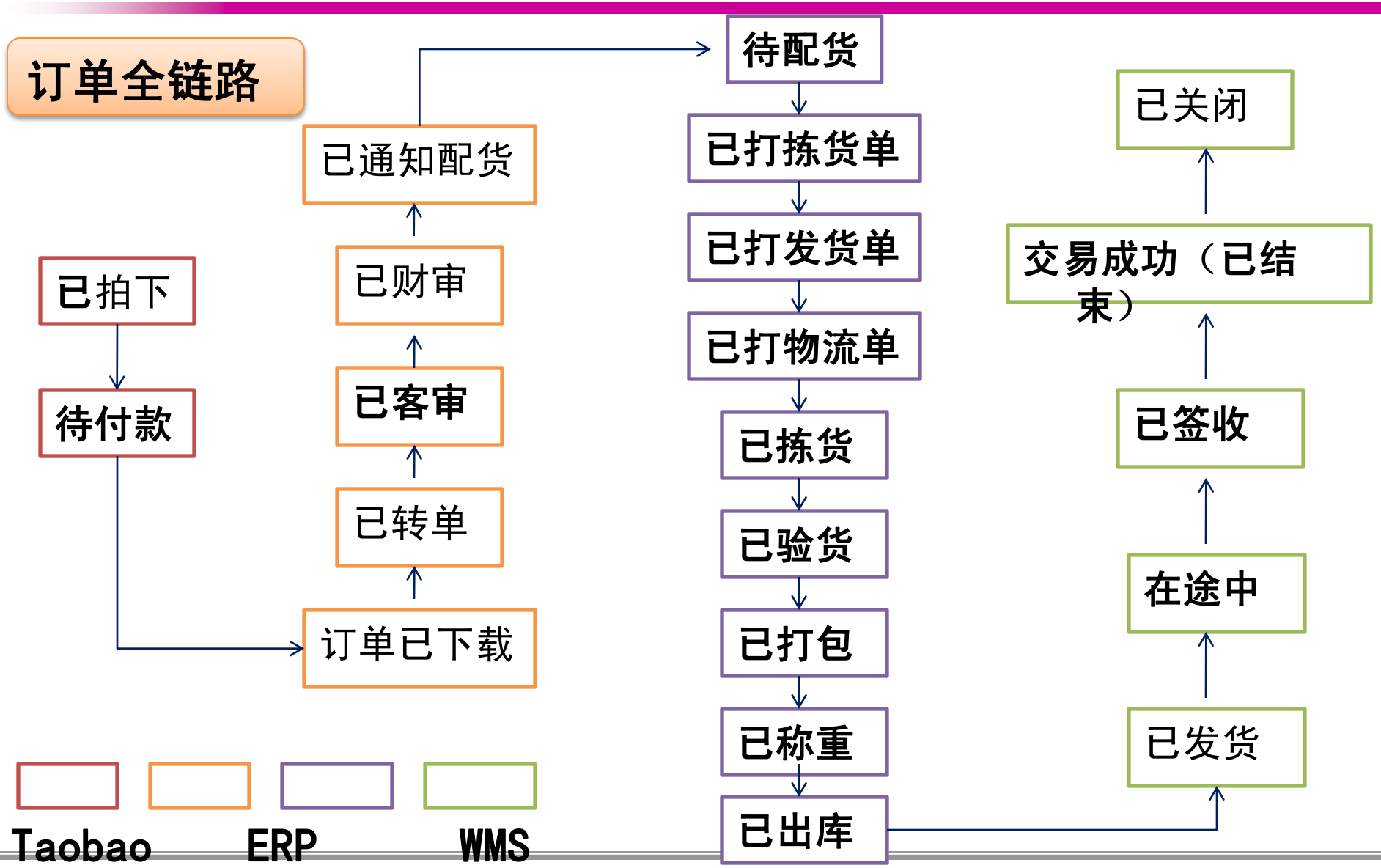
3.5 预警

- 重大事件分析：双十一
- 库存预警
- 竞争对手行动

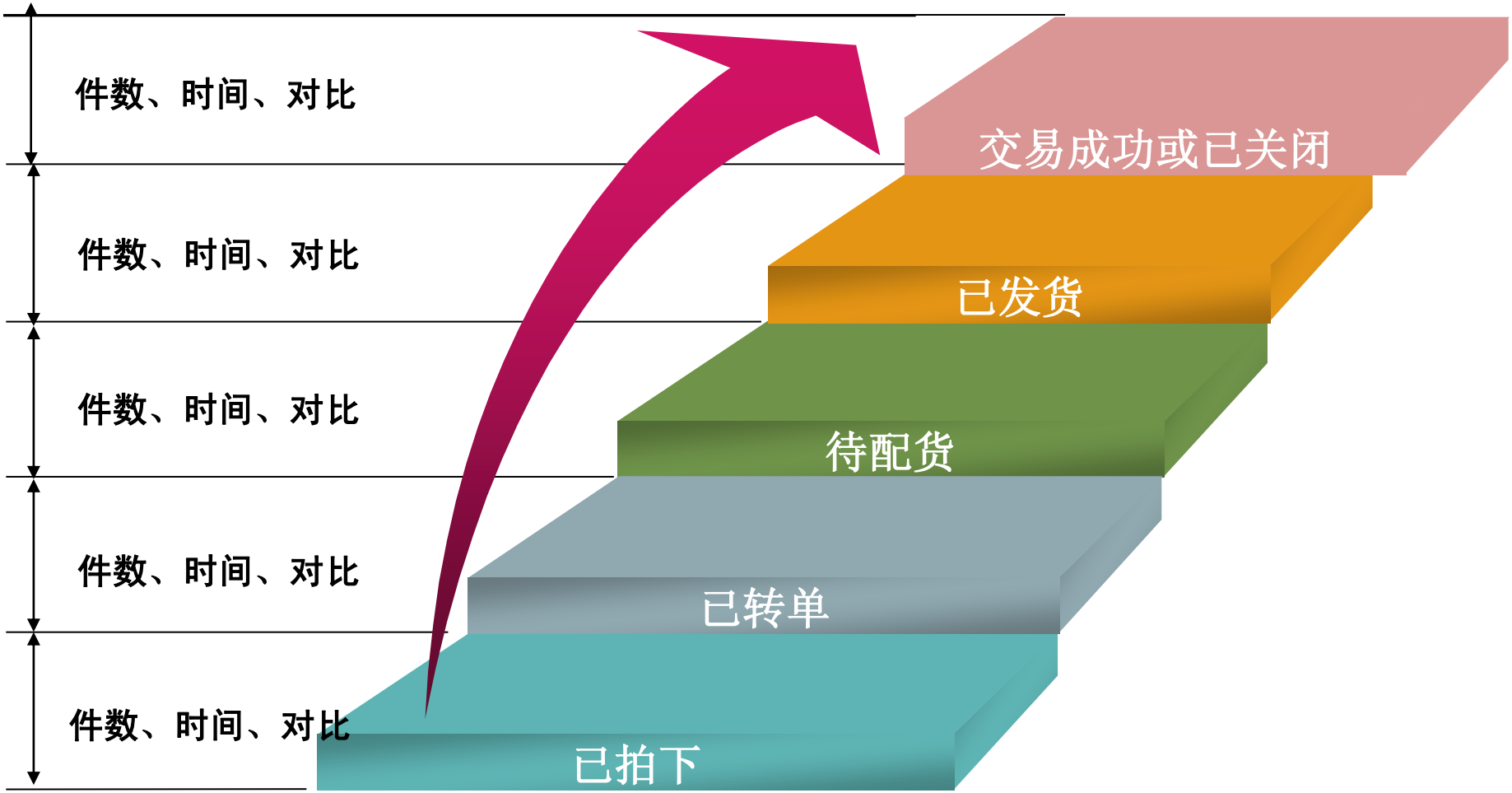
参加双十一活动客户分类分析结果展示



3.6 订单全链路分析

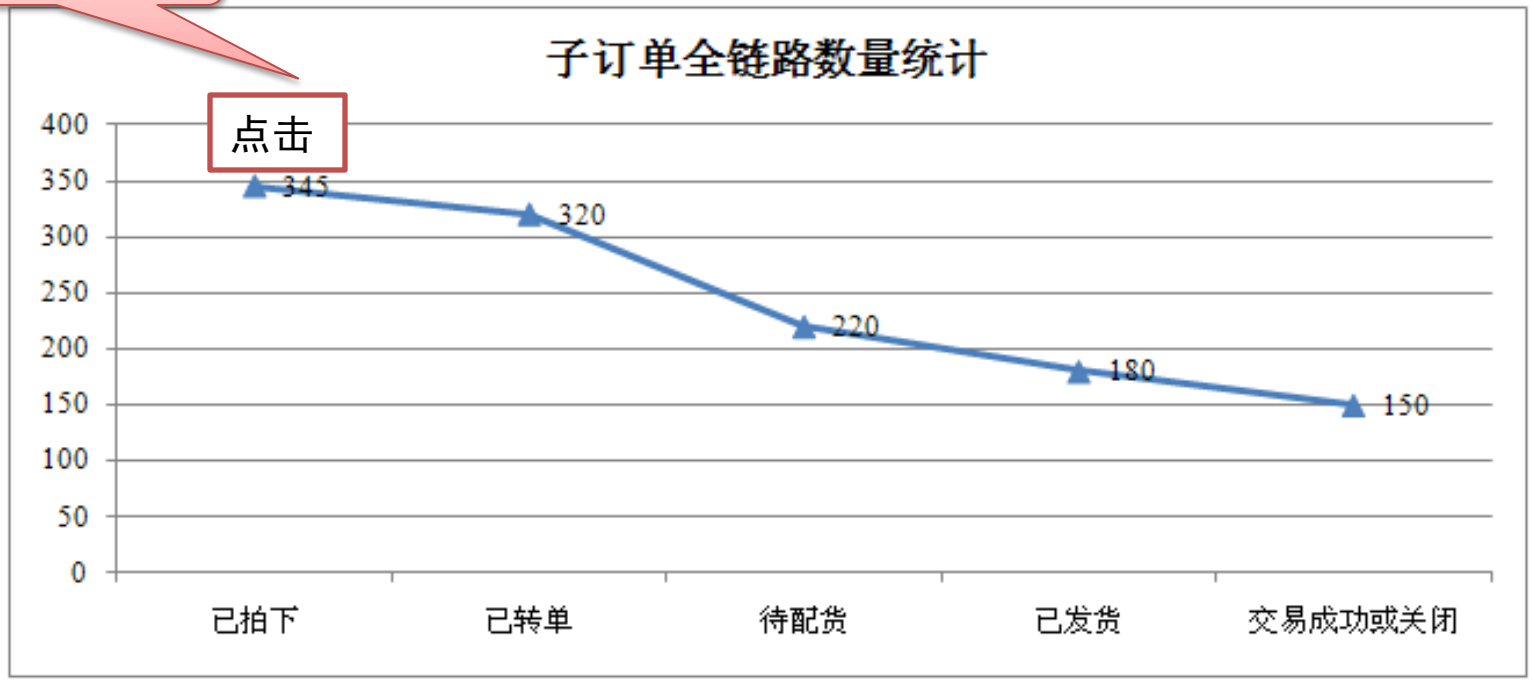


订单全链路分析


件数、时间、对比
件数、时间、对比
件数、时间、对比
件数、时间、对比
件数、时间、对比
交易成功或已关闭
已发货
待配货
已转单
已拍下

订单全链路数量统计

可查看子订单详情



通过子订单全链路的数量变化，帮助商家实时掌握子订单的状态信息，点击下钻可查看详情。

谢谢！

上海天律信息技术有限公司
地址：上海市浦东新区浦建路145号强生大厦1003室
电话：021-68763766
传真：021-58309596