


大数据社会对证券公司的挑战与机遇

- 资深项目经理
毛义彬



主要内容



一、迎接大数据社会

二、大数据的特性

三、大数据社会: **Ready?**

四、大数据的支撑

五、大数据的应用





一、迎接大数据社会

中国古典计数体系：

- 1、《孙子算经》中记载：“凡大数之法，万万曰亿，万万亿曰兆，万万兆曰京，万万京曰垓（**gāi**），万万垓曰秭（**zǐ**），万万秭曰穰，万万穰曰沟，万万沟曰涧，万万涧曰正，万万正曰载。”
- 2、由小到大依次为一、十、百、千、万、亿、兆、京、垓、秭、穰、沟、涧、正、载、极、……；
- 3、万以下是十进制，万以后则为万进制，即万万为亿，万亿为兆、万兆为京、万京为垓，……；



一、迎接大数据社会

1、2008年新产生数字信息的比特数：

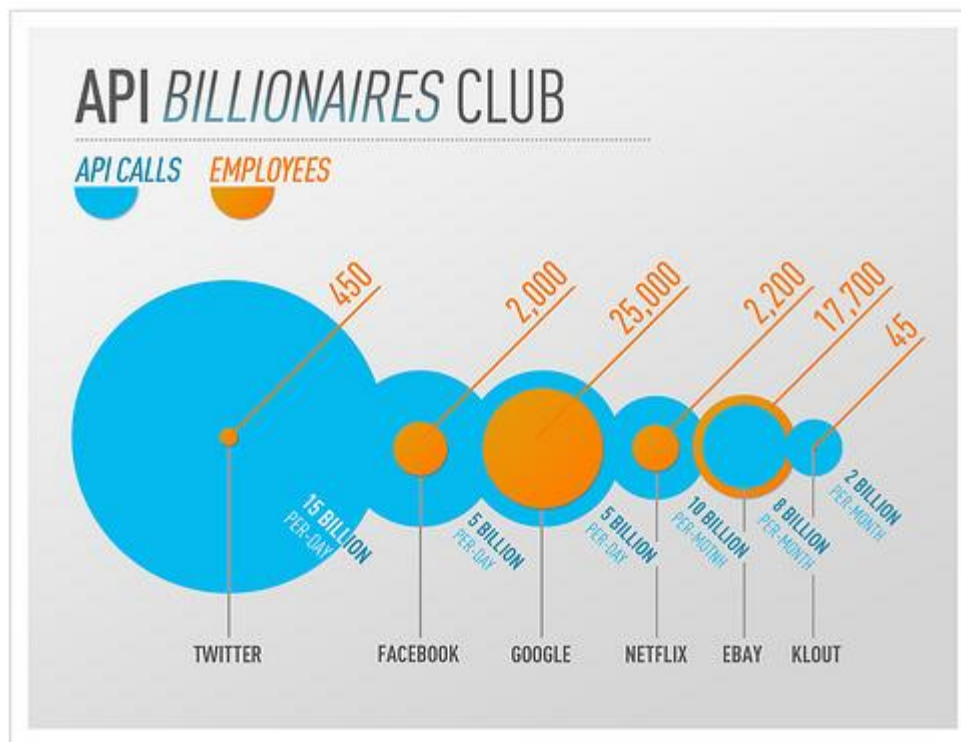
3,892,179,868,480,350,000,000

用中文表示为38垓9217京9868兆4803亿5千万

约等于39垓（音gāi）

也可计作38.9

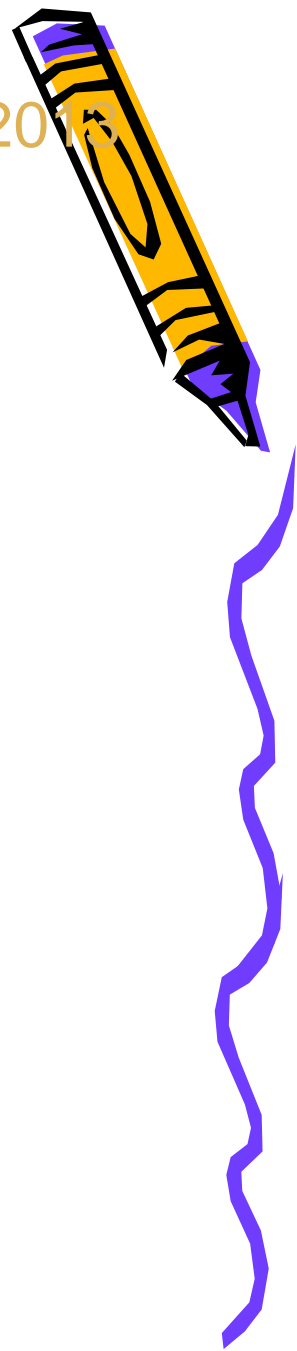
2、



一、迎接大数据社会

3、 Sysomos表示，在史蒂夫·乔布斯(Steve Jobs)辞世之后的13个小时内，Twitter用户发布的与乔布斯相关的信息多达250万条。

4、亚洲社交媒体的传播特性：分享——导致更多的信息传播





二、大数据的特性

IDC表示，首先必须成本低廉特征，其次是满足多样性（**variety**）、容量（**volume**）和速度（**velocity**）这三个标准中的两个。

- 1、Variety
- 2、Volume
- 3、Velocity





二、大数据的特性

1、互联网与Wiki

2、WikIT：指互联网技术的应用发展到今天，人们通过这个开放的环境进行协作，通过娱乐、交流和交易，形成的一种新型的关系，这样一种新型关系所潜在的巨大的社会价值我们所忽略，而去挖掘这样里的金矿，就是维基-IT(WikIT)的内涵。——进入的是维基-IT时代。WikIT-er

3、例证：web1.0招聘与facebook应用

从解决商业信息的不对称性到协同合作共赢模式

4、3Q大战——自觉与不知觉的开放



二、大数据的特性

SACC2013



1、改变了IT的生态环境

Appstore: 全民参与; prosumer

2、激活终端客户和partner共赢, 倒逼商业模式的变更云计算;

3、开放、有序产生价值



三、大数据社会：Ready?

什么是数据中心：

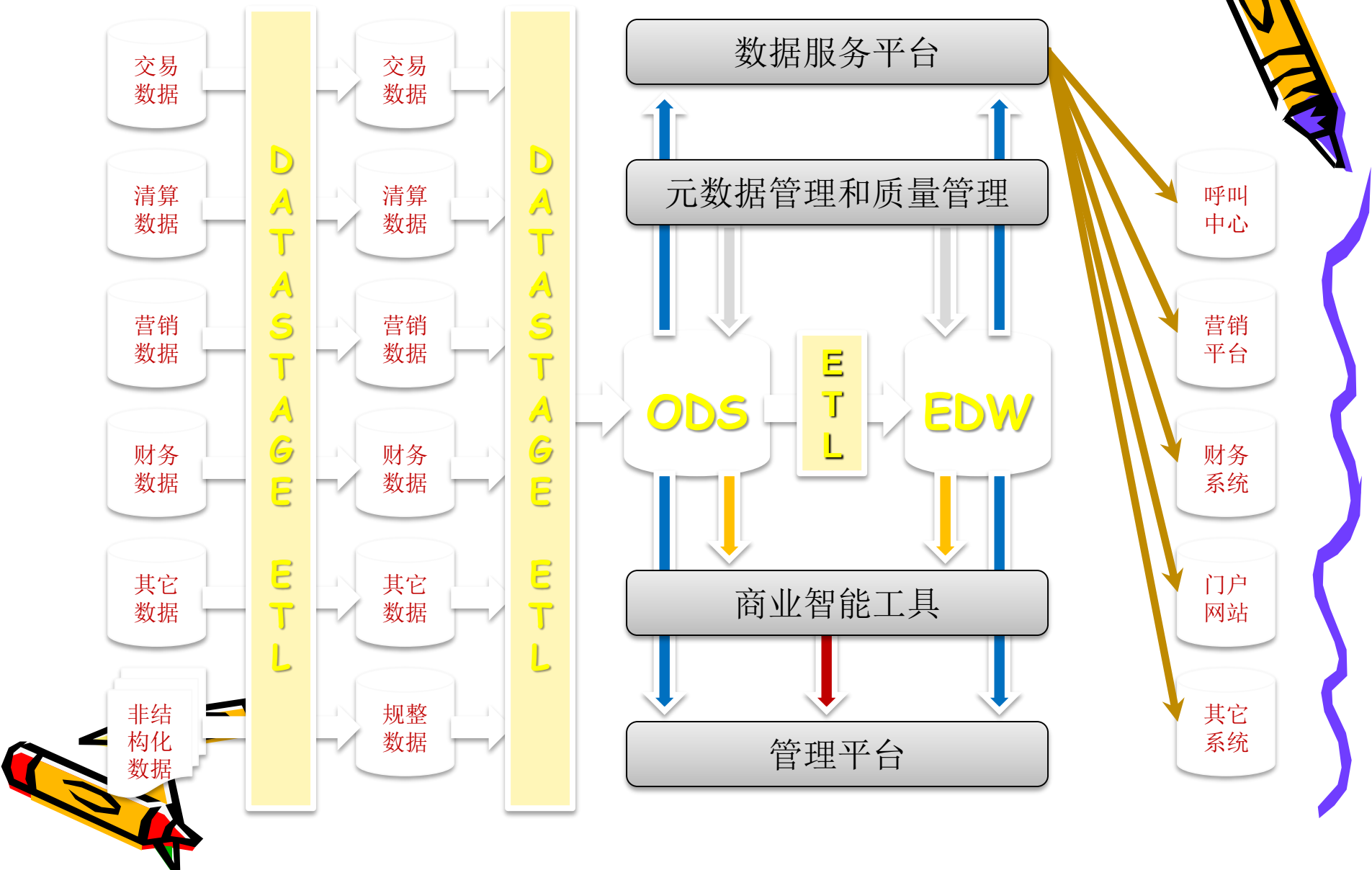
数据中心是企业的业务系统与数据资源进行集中、集成、共享、分析的场地、工具、流程等的有机组合。

- 1、从应用层面看，包括业务系统、基于数据仓库的分析系统；
- 2、从数据层面看，包括操作型数据和分析型数据以及数据与数据的集成/整合流程；
- 3、从基础设施层面看，包括服务器、网络、存储和整体IT 运行维护服务。



数据中心逻辑图(MPP架构)

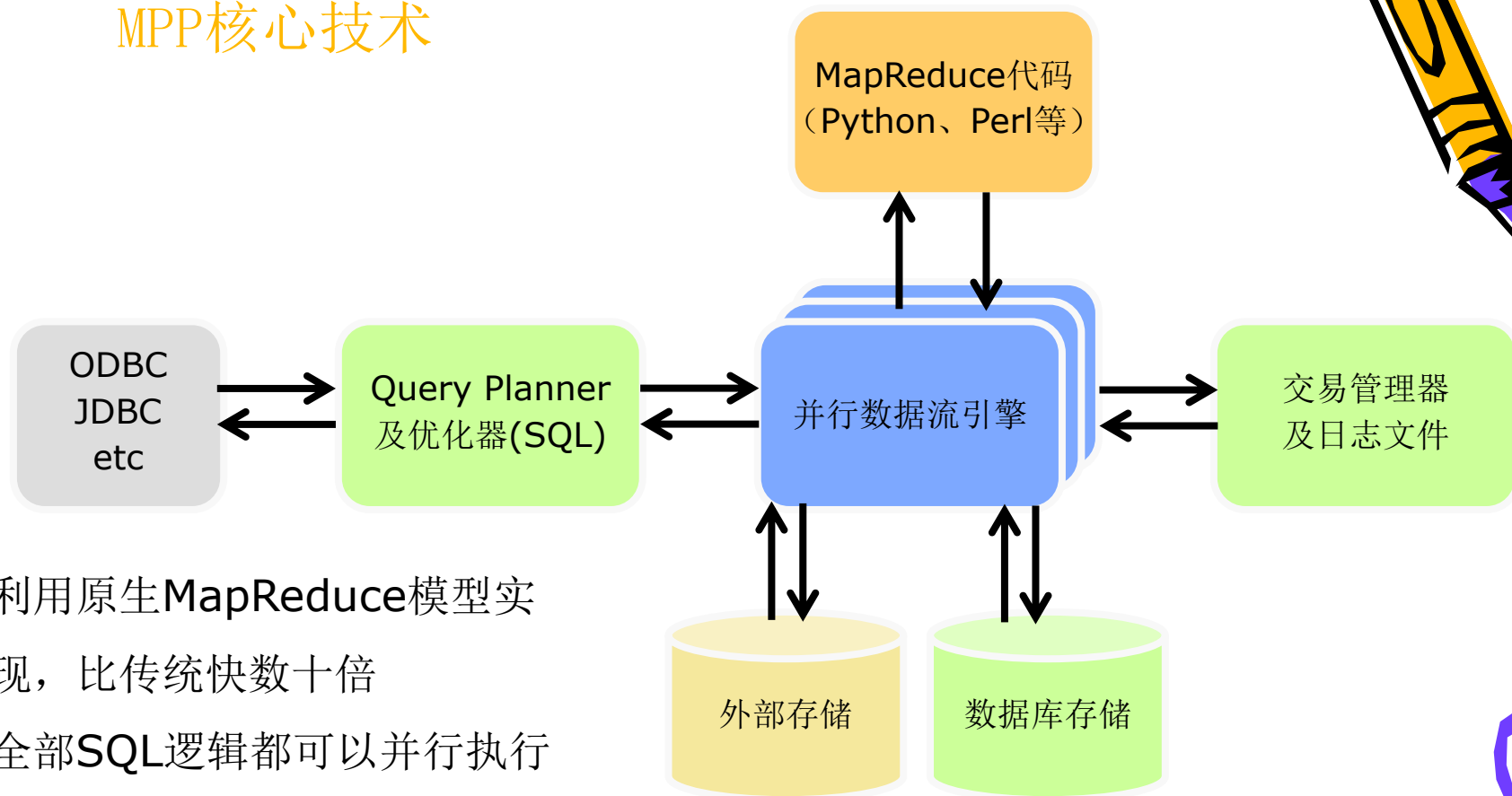
SACC2018



并行数据流引擎

SACC2013

MPP核心技术



- 利用原生MapReduce模型实现，比传统快数十倍
- 全部SQL逻辑都可以并行执行
- 并行技术加载和导出数据
- 并行数据备份和恢复

Master and Segment Node

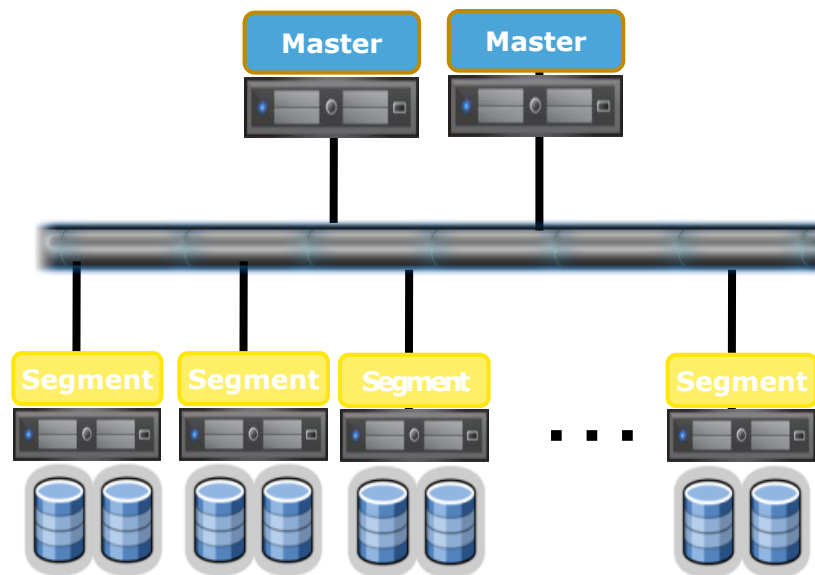
SACC2018

Master Node

- 建立与客户端的连接和管理
- SQL的解析并形成执行计划
- 执行计划向Segment的分发
- 收集Segment的执行结果
- Master不存储应用业务数据，只存储数据字典

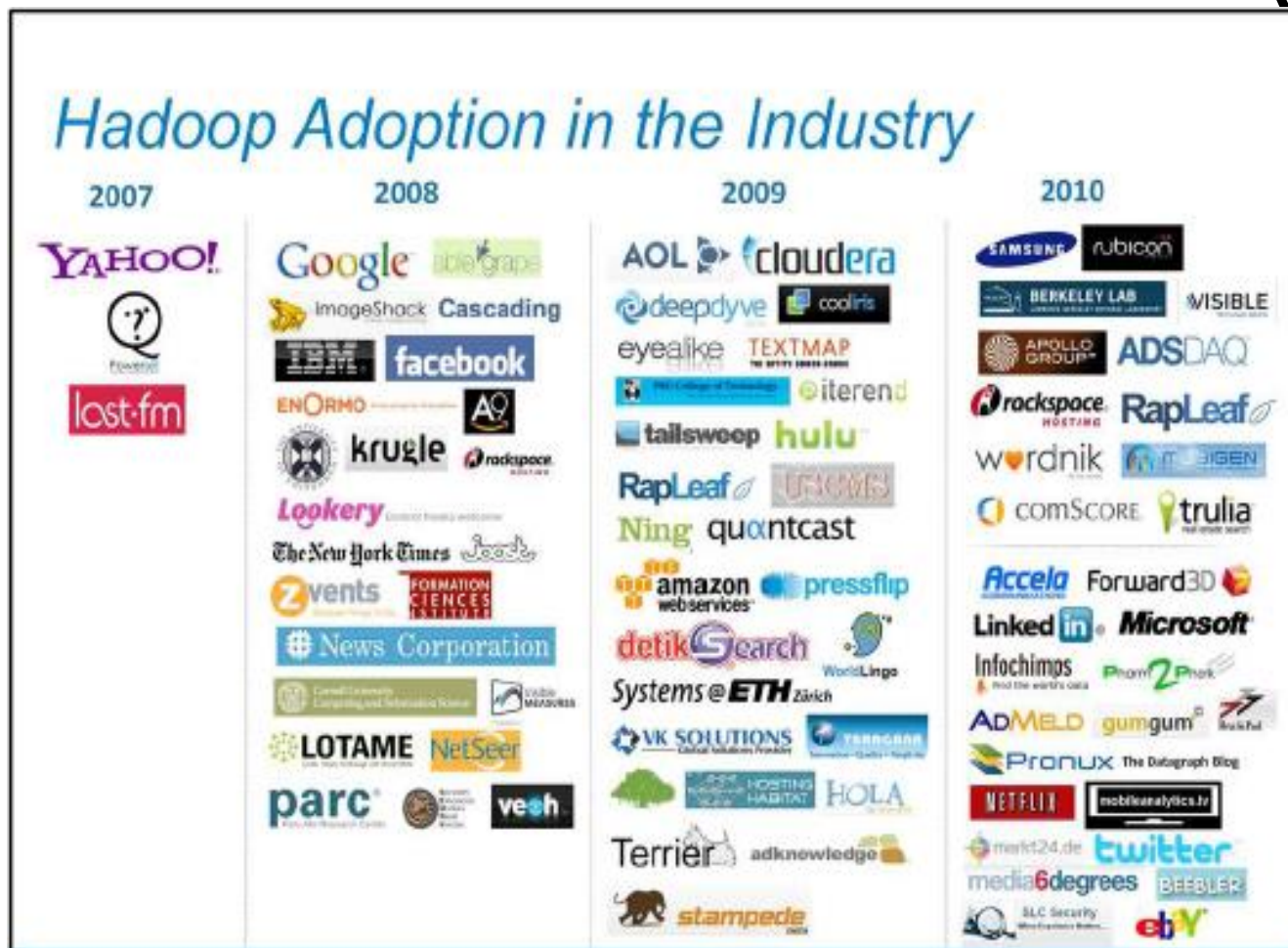
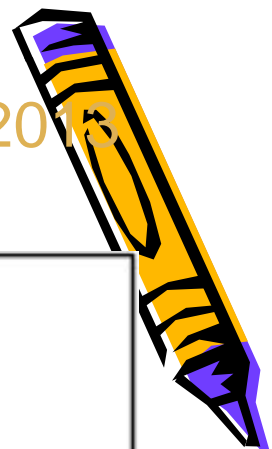
Segment Node

- 业务数据的存储和存取
- 用户查询SQL的执行



四、大数据的支撑

SACC2013

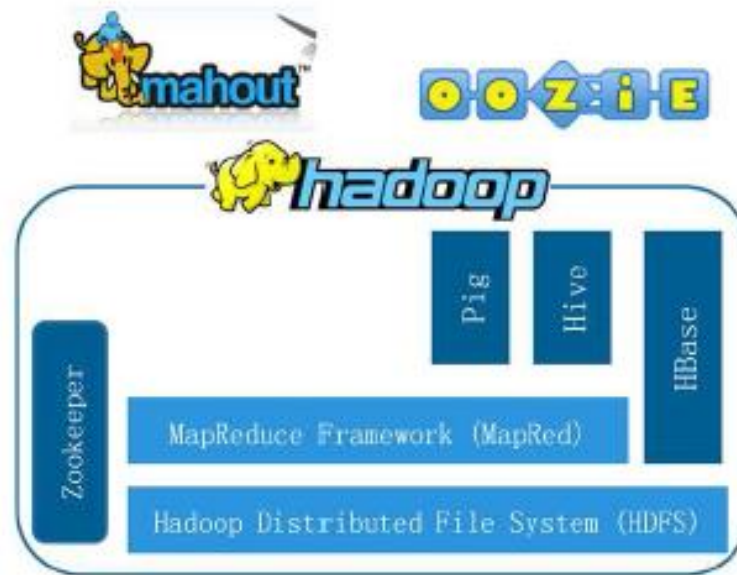


四、大数据的支撑

SACC2013

Standard Apache Hadoop Stack

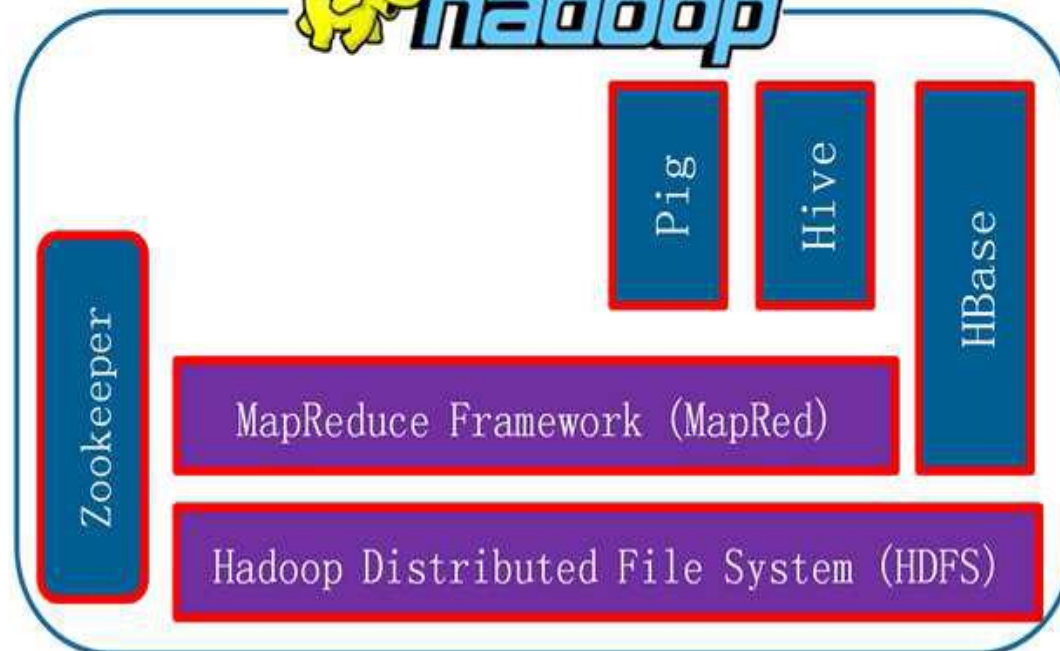
100%
APACHE



Greenplum HD: Community Edition Stack

SACC2013

100%
APACHE

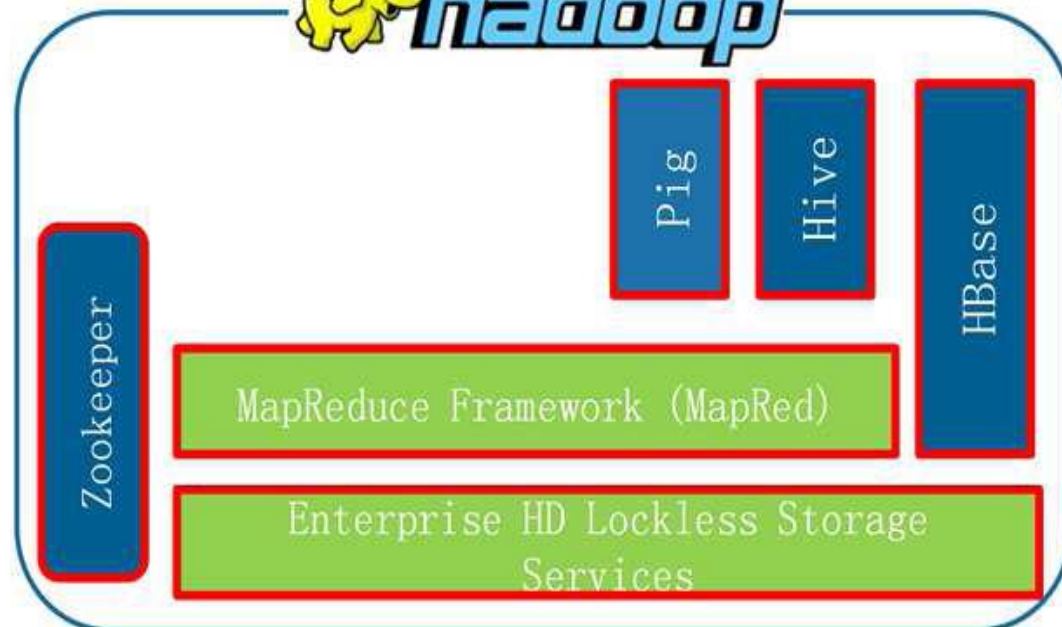


Currently

Greenplum HD: Enterprise Edition Stack

SACC2013

100%
APACHE
INTERFACE



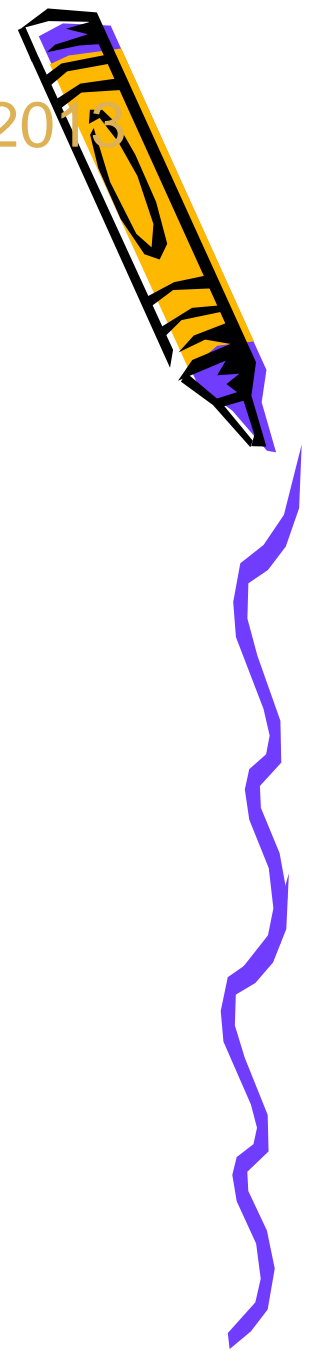
Enhanced Monitoring

Currently
supported

四、大数据应用探索

SACC2018

- 1、公众舆论与对冲基金
- 2、数据中心及数据挖掘的应用



构建细分模型的一般过程

SACC2018

方法论

Cross-Industry Standard Process for Data Mining 跨行业数据挖掘标准过程 (CRISP)

CRISP-DM 数据挖掘方法论用层次过程模型描述。包括四个抽象任务集合：

阶段(phase)

一般任务(generic task)

具体任务(specialized task)

过程实例(process instance)

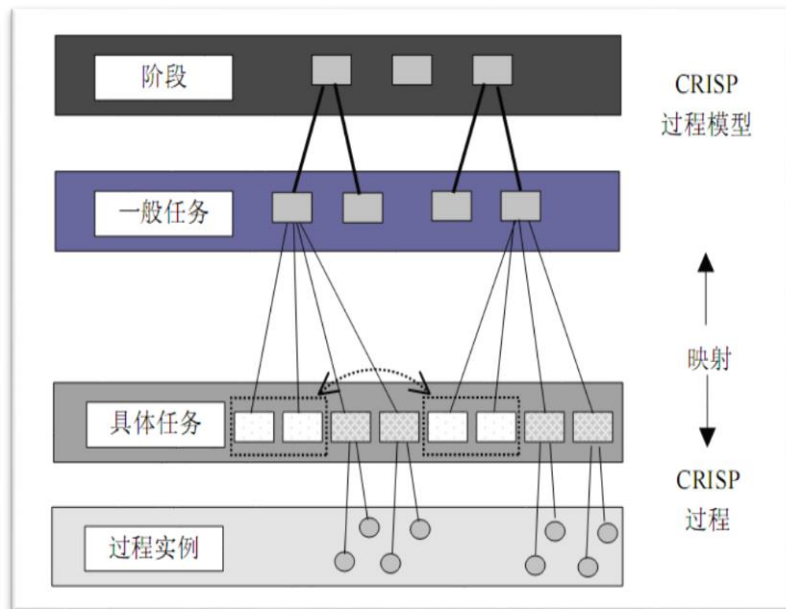
数据挖掘一般过程

第一层称为阶段，每个阶段包括若干个第二层的一般任务。

第二层称为一般任务，是因为计划把它设计得足够全面以涵盖所有可能的数据挖掘情况。“完全”意指涵盖数据挖掘的整个过程和所有可能的数据挖掘应用。“稳定”意指模型对于不可预见的发展比如新的建模技术也有效。

第三层称为具体任务层，描述一般任务层的活动如何在某一具体环境中实施。

第四层称为过程实例，是有关一次实际数据挖掘项目应用的活动、决策和结果的记录。



使用CRISP的一个例子

SACC2018

CRISP-DM



Business understanding

Data understanding

Data preparation

Modeling

Evaluation

Deployment

寻找潜在理财产品购买客户？

商业理解

截止数据日期，南京市信用卡用户141万，已经购买理财产品客户23933户，挖掘潜在的理财产品购买客户，分析理财业务……

数据理解

确认实体关系，设计数据挖掘宽表，进行基础的数据探索任务，撰写数据质量报告与数据探索报告。

数据准备

准备数据集、检查数据逻辑正确性、删除数据项、增加构造数据项、合并数据、格式化数据

建模

选择建模技术、设定假定命题、测试模型、参数调整、技术性模型评估

评价

依据商业知识评价模型、依据商业活动结果评价模型、核查模型稳定性、估计模型稳定周期

部署

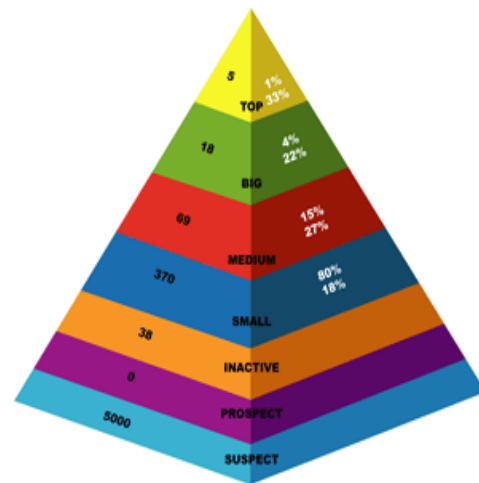
在合适的环境上部署模型，在稳定周期内循环使用

客户细分简介

SACC2018

客户细分的历史与发展：

- 客户细分是20世纪50年代中期由美国学者温德尔史密斯提出，其理论依据在于**顾客需求的异质性**和企业需要在有限资源的基础上进行**有效地市场竞争**。
- 发展至今，指企业在明确的战略业务模式和特定的市场中，根据**客户的属性，行为，需求，偏好、价值**等因素对客户进行分类，**并提供有针对性的产品，服务和销售模式**。



在快速发展业务的同时，是否需要更好的了解您的客户？

- 需要更详尽的了解用户群的构成情况；
- 需要更细致的了解不同用户群之间的差异情况；
- 需要更详细的了解用户群的消费行为和喜好；
- 需要更快速的了解用户行为的变化情况；

客户细分的价值：

- 业务人员的经验加上科学的细分方法使得细分结果更有效。
- 提供极大灵活性，快速建立市场细分模型。
- 确保企业及时的了解用户行为的变化情况。
- 为企业的策略制定提供数据支持。
- 为企业决策人员提供支持和帮助。

客户细分模型

SACC2015

提供了理解客户
的新思路

如何对待“Customer Segmentation”

- I. Not only a model;
- II. The logic thinking method;
- III. The starting point of the analysis;

如何实践“Customer Segmentation”

细分类型

具备的意义

战略细分

面向大市场，企业高层，定制市场战略等

价值细分

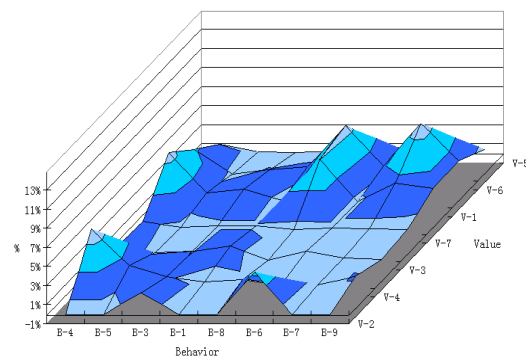
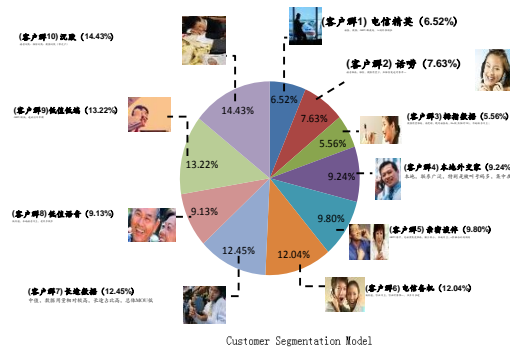
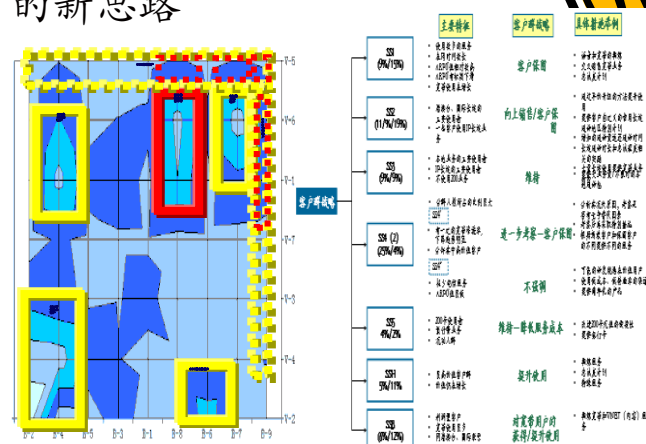
面向业务部门，制定营销倾向性策略等

行为细分

面向分析部门，了解客户行为特征、例如交易行为等

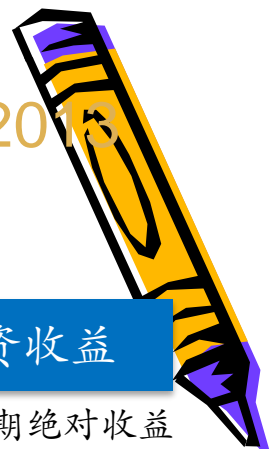
交叉细分

基于产品线的细分，以及细分子模型的组合



行为细分的目的：发现客户交易模式类型

SAC2018



账户状态	账户价值	交易习惯	投资偏好	投资收益
<ul style="list-style-type: none">✓ 有效性判断✓ 账户类型✓ 账户生命周期✓ 投资时间	<ul style="list-style-type: none">✓ 价值属性✓ 资产峰值✓ 资产均值✓ 交易量✓ 佣金贡献✓ 成本	<ul style="list-style-type: none">✓ 周转率✓ 市场关注度✓ 仓位✓ 平均持股市值✓ 平均持股时间✓ 单笔交易均值✓ 日均成交量	<ul style="list-style-type: none">✓ 偏好股票✓ 偏好品种✓ 下单渠道✓ 是否申购	<ul style="list-style-type: none">✓ 本期绝对收益✓ 本期相对收益✓ 今年绝对收益✓ 今年相对收益✓ 投资能力

账户状态中的变量用于圈定客户，其他变量可以用于数据分析或数据挖掘，同时区别对待连续变量与离散变量的使用方法；
使用原始变量分析经过计算后的变量之间是否存在共线性、相关性等因素，尽量获取独立性较强的变量进行依赖性分析；

扩充交易习惯类别的变量，获取原变量并按月汇总，进行衍生变量设计（比例型、业务组合型）；
此处建议不要考虑风险类字段，而将风险作为独立的题目进行设计；

建议使用3个月数据进行行为细分建模，使用6个月或12个月数据进行战略细分建模，针对时间范围内数据进行汇总



Alpine Miner中的聚类算法

SACC2018



K - Means

k-means 算法接受输入量k，然后将n个数据对象划分为k个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高，而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”（引力中心）来进行计算的。

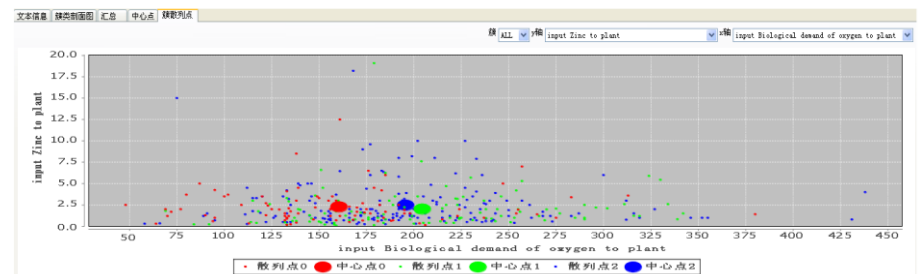
k-means 算法的工作过程说明如下：首先从n个数据对象任意选择k个对象作为初始聚类中心，而对于所剩下其它对象，则根据它们与这些聚类中心的相似度(距离)，分别将它们分配给与其最相似的(聚类中心所代表的)聚类，然后再计算每个所获新聚类的聚类中心(该聚类中所有对象的均值)，不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数。k个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。



o 中心点表

Cluster	input flow ...	input Zinc ...	input pH to ...	input Biolo...
0	46787.46	2.264301	7.8645163	160.51613
1	30319.686	2.0033333	7.763063	204.46848
2	36843.84	2.456875	7.84375	195.65341

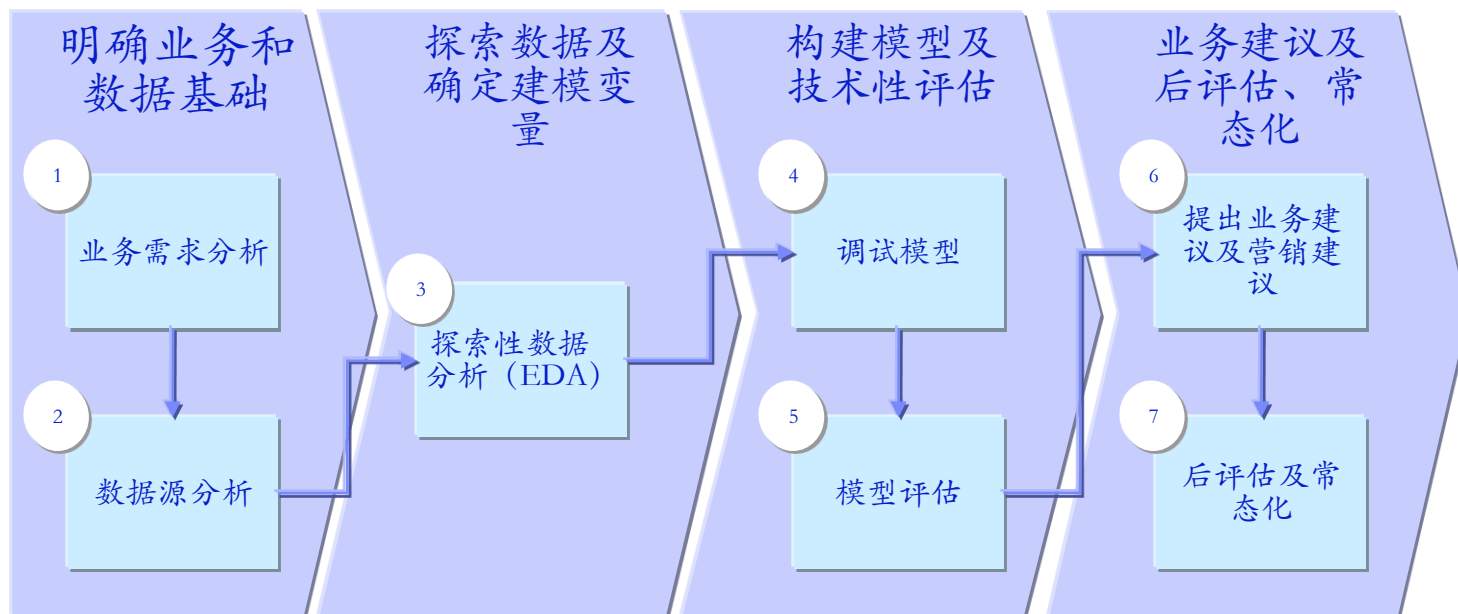
簇散列点图



支持9中距离计算函数方法

信用卡客户细分工作流程

SACC2018



1 获取客户的需求，并探讨想要的分析方向及分析重点及确立分析题目

2 对当前数据现状进行分析、诊断，确定具备分析工作能够展开的基本数据基础

3 数据质量检查、探索数据（业务统计）、变量探索、变量降维等工作

4 确定算法、参数，调试模型；模型比较

5 针对不同类型的模型使用不同参数进行评估LIFT、GINI等；模型解读，以业务能够理解的方式向业务人员解释成果

6 结合业务发展方向、EDA中的业务统计分析，以及当前模型结论给出相应的业务建议，操作建议；给出特定的营销活动策划建议；

7 收集模型测试数据，评估当前稳定性；提出常态化建设的意见、方法、运维思路；

信用卡客户细分建模过程

SACC2018



分析粒度：信用卡个人客户

产品范围：人民币贷记卡和国际卡，不包含准贷记卡产品，且过去6个月至少有一次主动交易（主动交易行为指消费及取现交易）

帐龄历史： ≥ 12 个月

数据窗口：待定

剔除客户：剔除黑名单客户，曾经有过M3逾期历史的客户将被剔除

地域：全行(30% Model, 70% Test)

Alpine Model Key Point(except Derivative variables)

平均信用限额

取现金额

消费总额

当前拖欠金额

还款总额

信用使用次数

小额循环信用

使用次数

最高额度使用

率

利润总额

消费总次数

日均帐户余额

使用算法：K-mean

距离函数：先用Euclidean，如果结果不如预期适尝试使用Manhattan方法(必须规范化数据)

聚类个数：7个

初始中心：3个

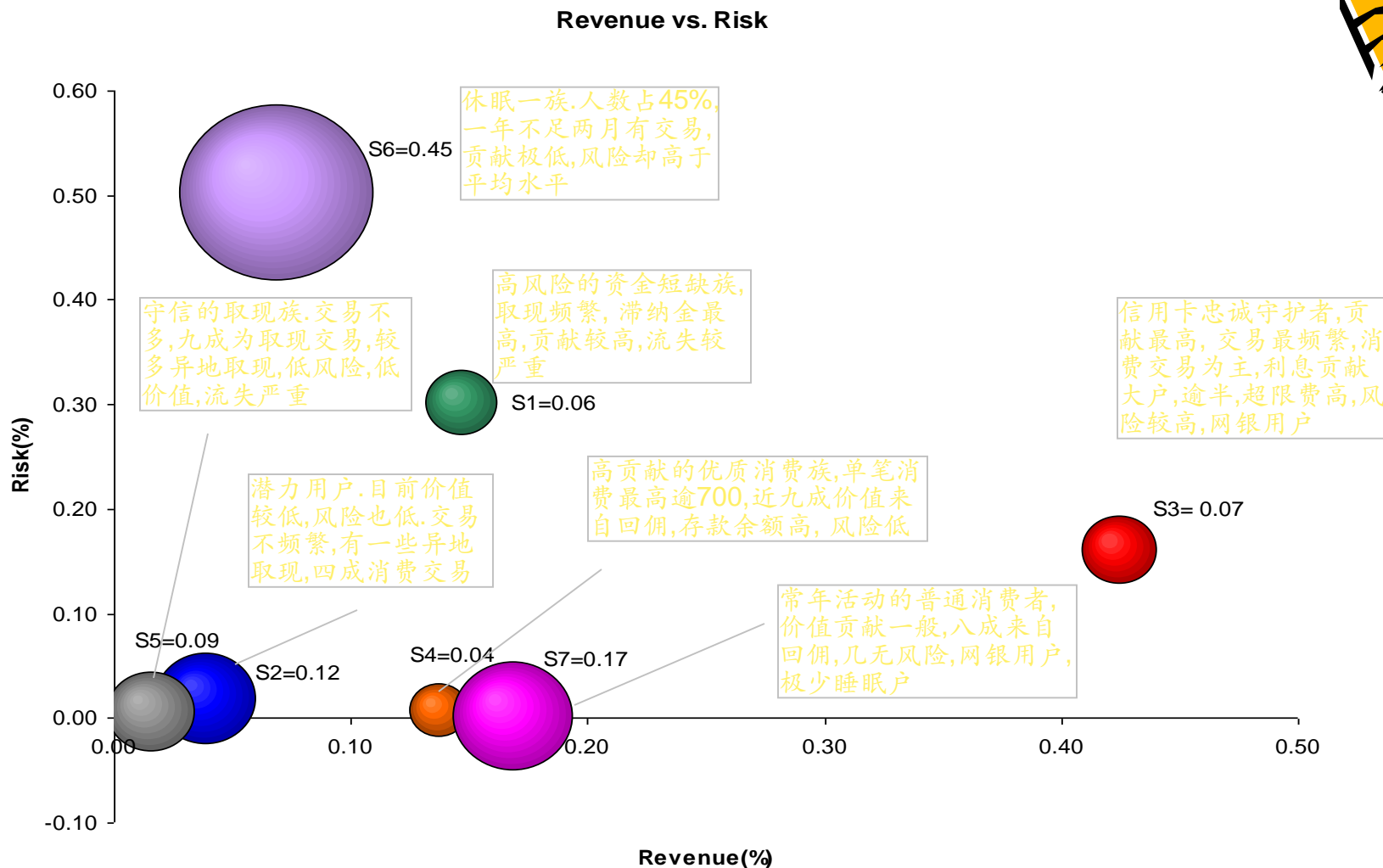
第一次训练模型使用5个变量开始，第二轮训练增加2个，第三轮训练在增加2个，直至增加到模型稳定,当最终变量超过11个，增加因子分析降维之后使用至少78%信息量的因子建模对比权衡使用模型

得到模型之后，第一种方法直接在聚类模型上分析特征；第二种方法将模型作为变量输入模型，例如使用决策树或分类回归树对聚类再次建模，获取规则路径。

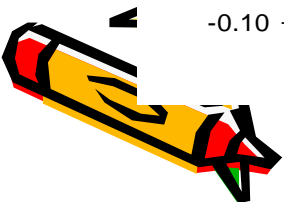


信用卡客户群特征总览(预期结果)

SACC2018



Note: 球体大小表示客户群大小; Revenue%表示收入贡献份额; Risk%表示M3+逾期人次占比



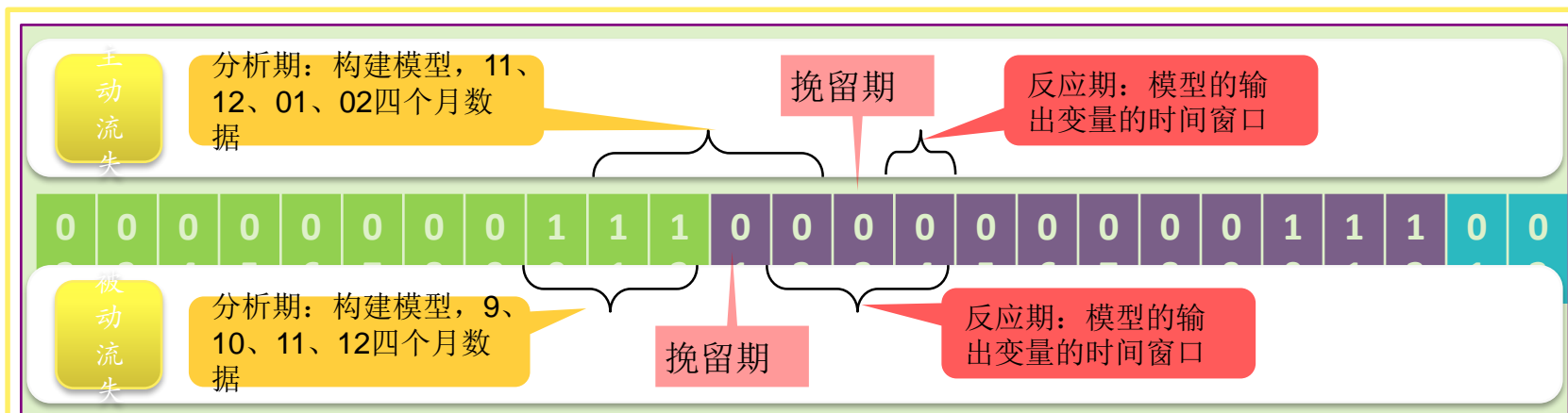




流失定义的重要性：“流失动作”与“流失意向”

SAC2018

- ✓ 主动流失建模预测的是**流失动作**发生的先决条件
- ✓ 被动流失建模预测的是**流失意向**产生的先决条件



主动流失标记条件：

主动销户

使用2010-11~2011-01月份共计4个月数据进行建模分析

被动流失标记条件：

交易量连续三个月下或资产转移等等……

使用2010-10~2010-12月份共计3个月数据进行建模分析



SACC2018

结束语

谢 谢

