互联网金融企业的大数据应用案例分享

联动优势 孟鑫

2013.9

主题

* 概述
* 大数据挑战
* 平台现状
* HBase应用
* 推荐系统
* 用户信用评分&支付交易监测

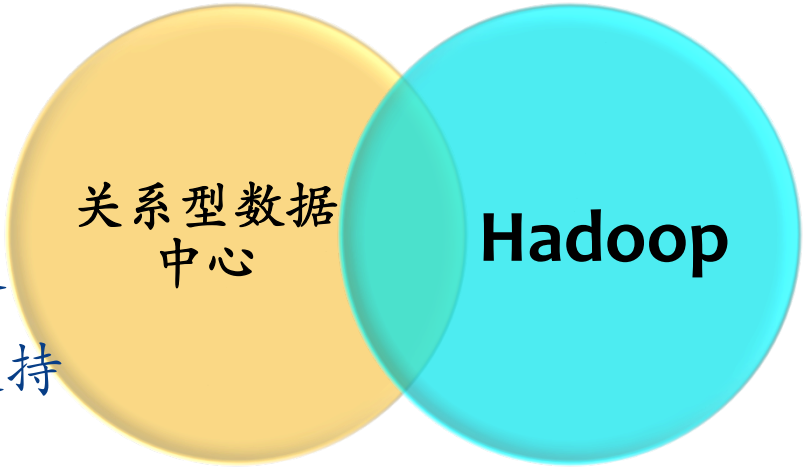
概述-背景

* 2013年第二季度第三方移动支付市场份额11.6%列第二位
* 某核心业务数据每日>1.5亿条，实际数据量每日>200GB
* 互联网支付交易每日>200万笔

概述-数据平台建设

* 关系型数据中心
  * 基于IBM Netezza和商业BI软件构建
  * 支持公司上百个重要业务指标计算和展现
  * 2011年上线

* Hadoop
  * 提供海量数据挖掘，实时访问服务
  * 为Netezza提供数据备份、ETL等支持
  * 2012年上线，规模50+

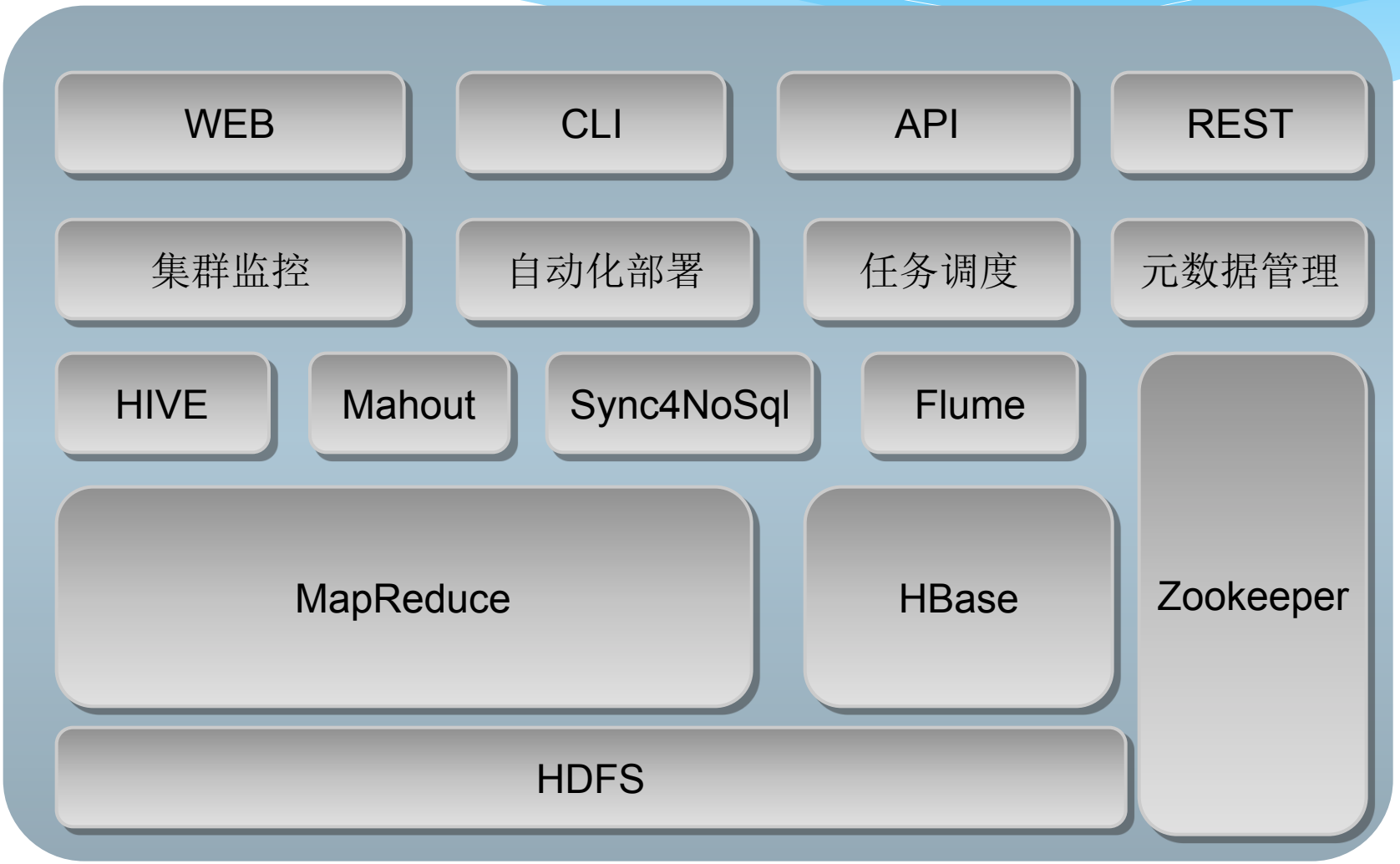


大数据挑战-长期诟病

大数据挑战-Hadoop平台的目标

* 数据恢复在线状态
* 承担大数据的离线统计分析
* 提供海量数据库给非OLTP系统
* 为智能应用提供数据挖掘支持

平台现状-架构


WEB
CLI
API
REST
集群监控
自动化部署
任务调度
元数据管理
HIVE
Mahout
Sync4NoSql
Flume
MapReduce
HBase
Zookeeper
HDFS

平台现状-线上系统

* 系统规模50+
  * 8核，128G或32G内存，SATA硬盘，单台16TB，多网卡绑定
* 平台基于CDH3U3版本
  * 公司内部开放HDFS、Hive、HBase
  * 基于共享存储的NameNode HA
  * Flume tail文件断点续传
  * Hive权限控制
  * 数据访问中间层

平台现状-测试系统

* 目前在测试环境进行Hadoop2.0新特性研究和开发
  * YARN
  * 基于QJM的HA
  * Hadoop安全

  * HBase 0.94
  * 二级索引
  * 类SQL支持
  * 事务支持

HBase应用-发展

* 2012年客服系统第一个尝鲜
* 2013年客服系统全部迁移到HBase上，通过Filter和数据访问中间层处理实现绝大部分功能

特点：数据量大，写多读少，查询条件简单

* 商户服务系统，用户服务系统逐步迁移到HBase，部分实现ANSI SQL92标准
* 数据同步由非实时向准实时过渡

特点：读多，查询条件复杂

HBase应用-简单查询

* 单张表数据>200亿，要求响应时间<1s，数据同步时间<3分钟
* RowKey:手机号+日期+唯一流水
* 查询条件非常简单，按rowkey查询可以搞定
* 查询特点是近日数据访问量大，历史数据访问量小
* 以手机号段切分region，转移到不同regionserver负载，预先加载昨日数据
* 缓存命中率极低，blockcache保存最近一天数据
* 通过pageFilter实现分页，数据中间层进行排序

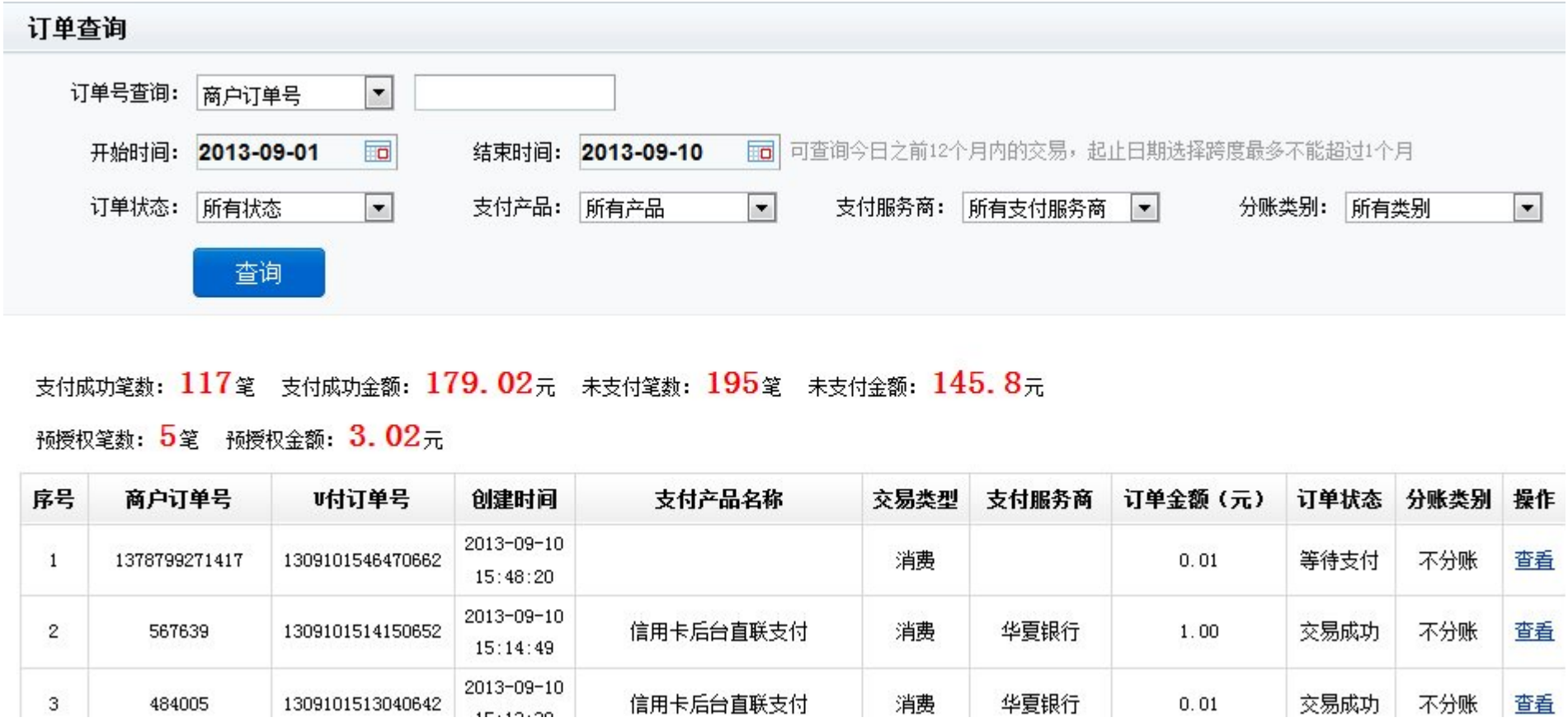
HBase应用-日志查询

HBase应用-复杂查询

* where条件字段较多
* 聚集函数count、sum、max、min、avg
* 需要Order By、分页、Group By等功能
* 支持Join
* 支持常见运算符：AND、OR、IN、=、>等

HBase应用-商户服务系统

HBase应用-商户服务系统



HBase应用-商户服务系统

4	20130910145757	1309101457570452	2013-09-10 14:58:30		消费		0.01	等待支付	不分账	查看
5	1378793860450	1309101416400432	2013-09-10 14:18:13		消费		0.01	等待支付	不分账	查看
6	1378793763886	1309101415040422	2013-09-10 14:16:36	借记卡无磁有密客户端	消费	中国农业银行	0.01	交易成功	不分账	查看
7	1378793673656	1309101413340412	2013-09-10 14:15:06		消费		0.01	等待支付	不分账	查看
8	1378793628738	1309101412490402	2013-09-10 14:14:21		消费		0.01	等待支付	不分账	查看
9	1378793519061	1309101410590392	2013-09-10 14:12:31	信用卡无卡支付（插件大额）	消费	中国农业银行	0.01	交易成功	不分账	查看
10	1378793487725	1309101410280382	2013-09-10 14:12:00		消费		0.01	等待支付	不分账	查看

下载查询结果

共32页 首页 上一页 1 2 3 4 ... 下一页 尾页 到第 页 跳转

HBase应用-商户服务系统

* 通过SQL解析器将SQL语句转换成HBase scan操作
* 通过Coprocessor执行聚合操作
* 在RegionServer端尽早过滤数据
* 自定义Filter

HBase应用-数据实时同步

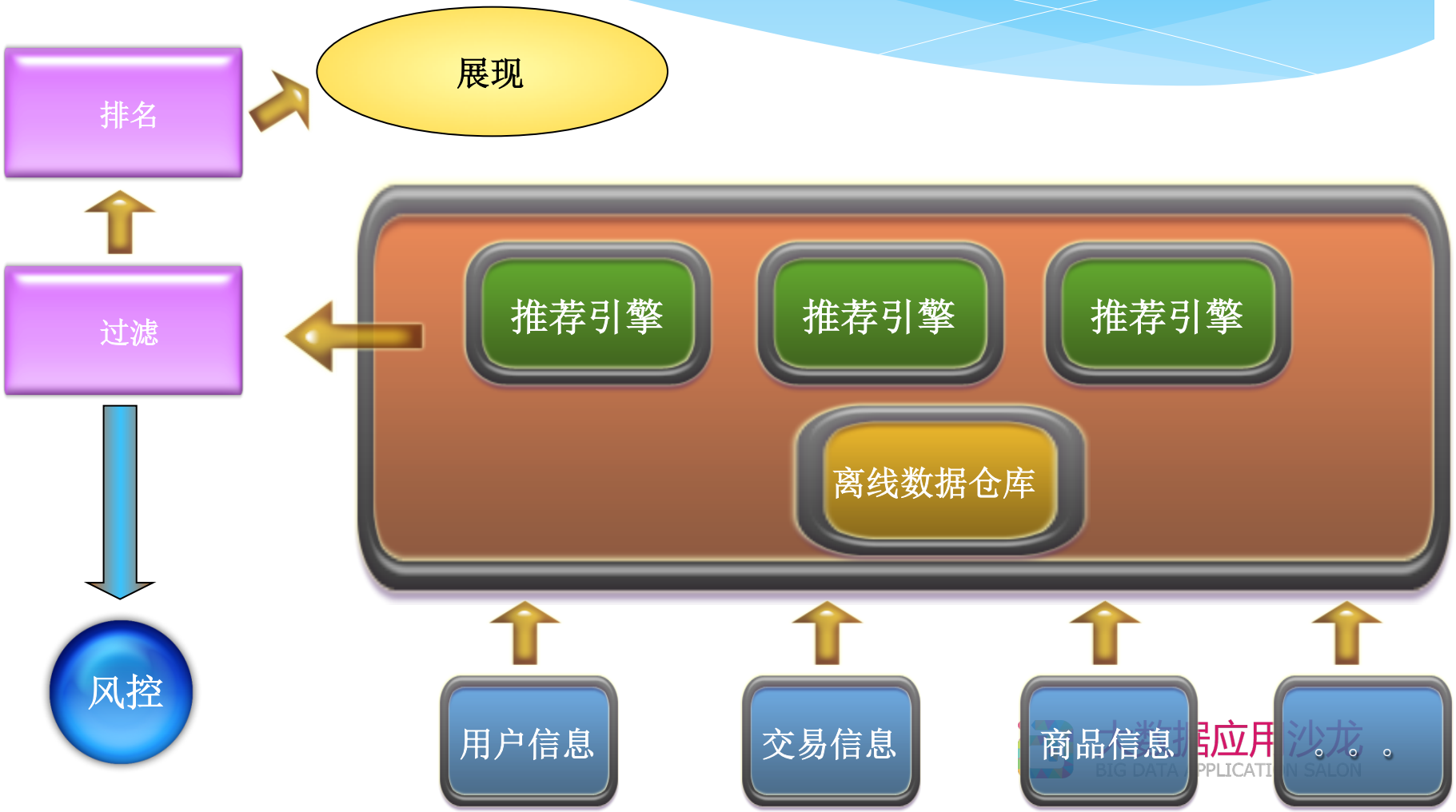
* Flume
  * 同步日志文件
  * 可靠性问题
  * 断点续传
* 公司自研的关系型数据库同步工具
  * 增加关系型数据库到HBase同步

数据同步实时性需求越来越多

推荐系统-起因

* 年交易增长率稳定在15%左右且很难有突破
* 传统营销方式成本太高、效果不佳
* 长尾商品

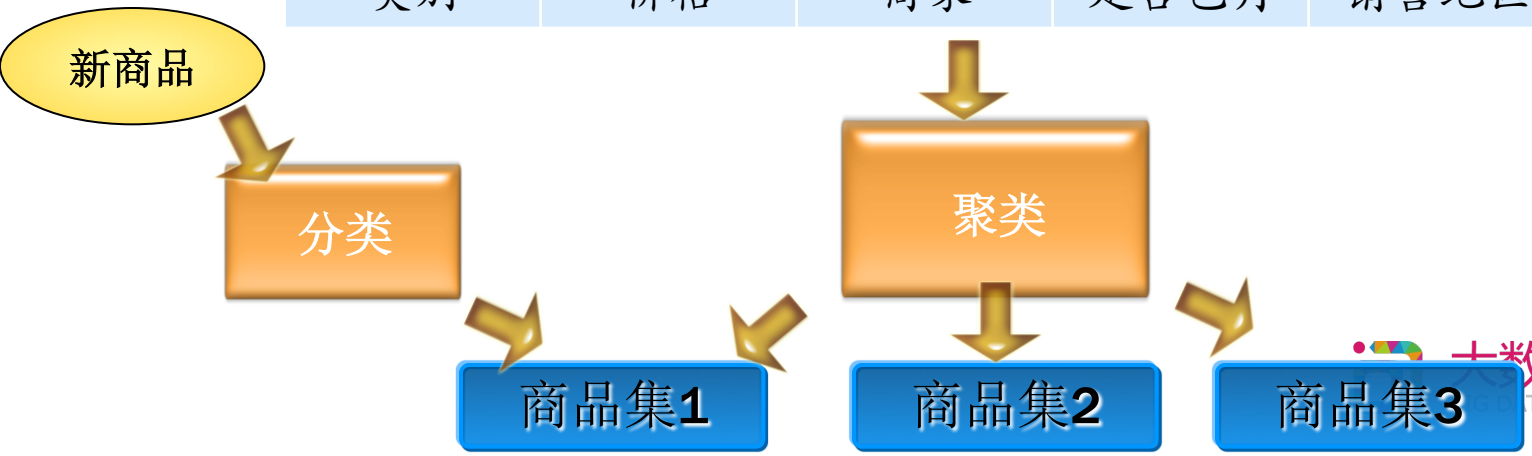
推荐系统-架构



推荐系统-默认推荐

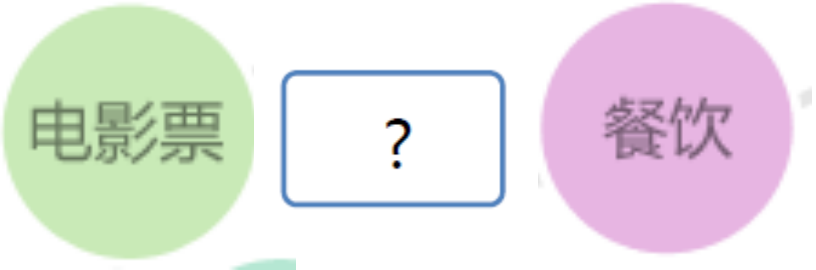
* 热门榜
  * 商品聚类、分类
  * TopN商品销售量
  * 过滤：违规商品、分地区、限额等
  * 适用于新用户，每个类别挑选一件商品进行推荐

商品信息				
类别	价格	商家	是否包月	销售地区



推荐系统-相关推荐

* 根据用户购买行为
  * 适用于有过交易的用户
  * ItemCF：协同过滤
  * 用户单一消费商品习惯?



推荐系统-制约因素

* 客户端商品信息不丰富
* 用户行为数据太少，无法做基于用户行为的推荐

用户信用评分-意义

* 发现优质用户
* 降低业务风险
* 预测用户好坏概率

用户信用评分-理论

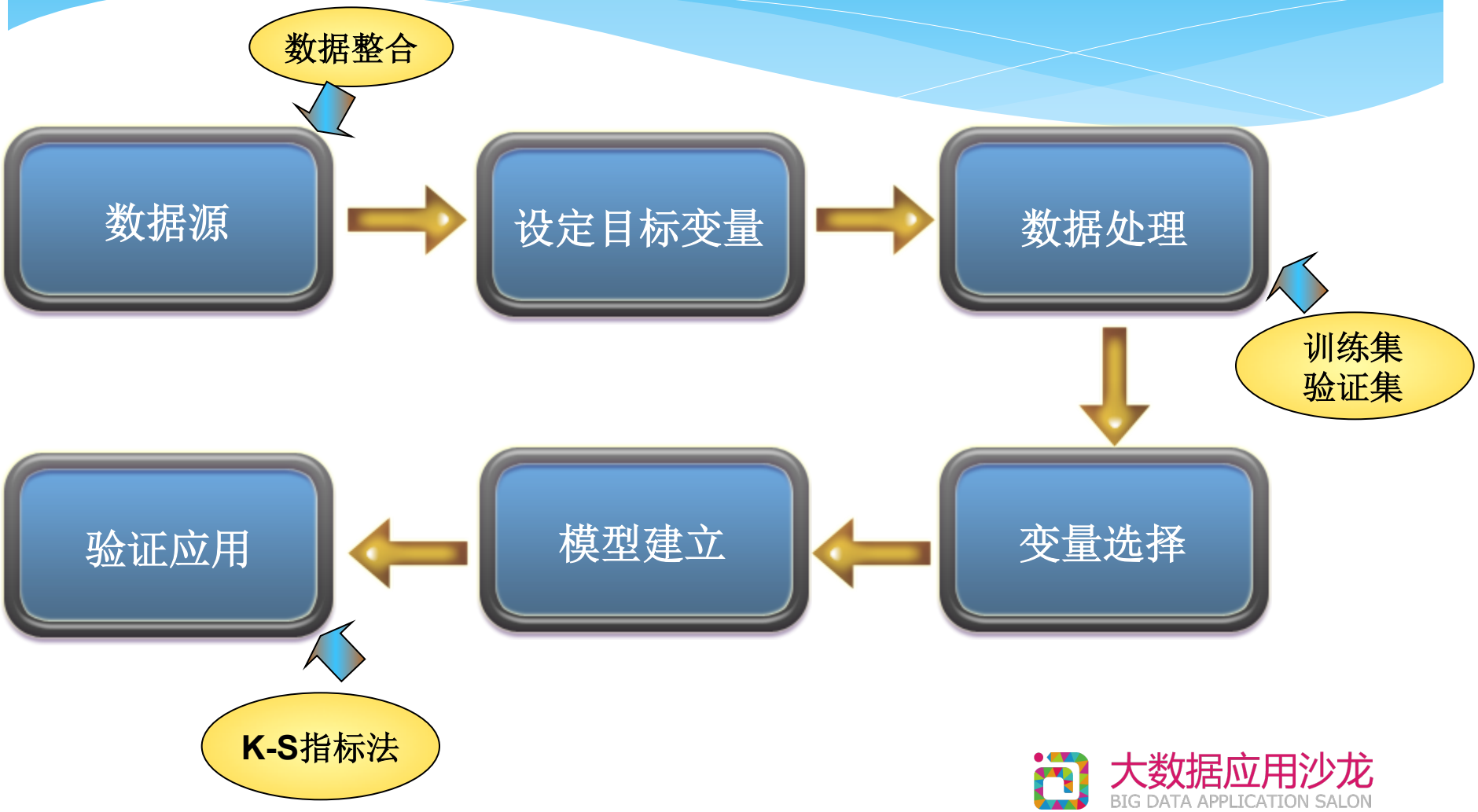
* 逻辑回归

$$p = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

$$x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots$$

求解系数，将用户特征属性值带入公式，计算概率

用户信用评分-流程



用户信用评分-结果

* 某省预测结果
  * 好用户8.18%
  * 关键变量：实名，准确率98.09%
  * 有9个变量对预测有重要影响

支付交易监测

* 立足于监测可疑支付交易，为打击洗钱和欺诈等犯罪活动提供信息支持。

从洗钱的一贯做法来看，通过银行支付结算，短期内转移巨额资金并使之从形式上合法化，是犯罪分子进行洗钱犯罪活动的主要特征之一。支付交易监测系统的建设，要根据洗钱的特征，对银行办理的大额支付交易进行有效监测，从中发现可疑支付交易线索，打击洗钱犯罪活动，促进支付结算业务健康发展。
支付交易监测由支付交易信息采集和支付交易信息分析两部分组成。
支付交易信息采集系统通过与业务处理系统连接，自动采集高频支付交易信息，形成高频支付交易数据库；通过开发和建立异常支付交易分析模型进行异常支付交易信息的搜集、接收、整理、监测和分析，形成异常支付交易数据库；通过与身份识别系统的连接和其他手段对异常支付交易信息进行进一步分析后，最终形成可疑支付交易数据库

监测分析模型

资金流动频繁的账号

模型一：分散转入，集中转出

设立该分析模型的目的，用于监测短期内资金分散转入，集中转出的情况。

模型二：集中转入，分散转出

设立该分析模型的目的：用于监测短期内资金集中转入，分散转出的情况

模型三：资金快速流动

设立该分析模型的目的：用于监测一笔资金通过某一账号迅速流动的情况

模型四：通过充值方式集中转入，分散转出资金

设立该分析模型的目的：用于监测短期内相近资金以充值的方式集中转入，分散转出的情况

监测分析模型

资金流动频繁的客户

模型一：同一客户短期内频繁发生收付

设立该分析模型的目的：用于监测短期内频繁发生收付业务的客户

模型二：同一客户在短期内以充值方式频繁发生收付

设立该分析模型的目的：用于监测短期内频繁发生大额充值的资金收付。

其他场景

资金流向分析：重点地区资金流向、重点行业资金流向、频繁且相近额度资金流向、季节资金流向、节假日资金流向、偶尔大额资金流向。


The End

玩数据，常联系！

* 大数据应用沙龙官网
* http://blog.linezing.com/salon

* 阿里技术沙龙官网
* http://club.alibabatech.org