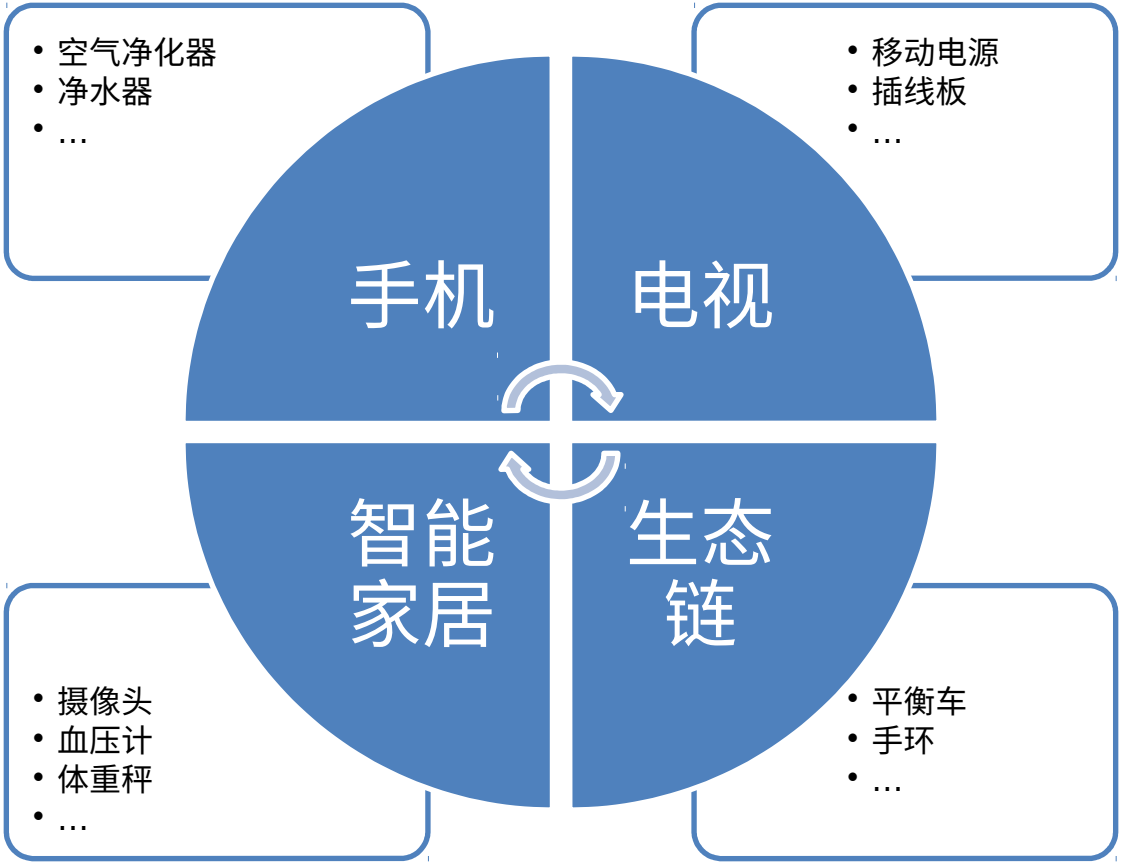
大数据和小米金融

方流，小米金融技术总监
fangliu@xiaomi.com

- 小米公司简介
- 小米金融
- DW 建设
- 用户金融画像
- 大数据反欺诈


• 空气净化器
• 净水器
• …
• 移动电源
• 插线板
• …
手机
电视
智能家居
生态链
• 摄像头
• 血压计
• 体重秤
• …
• 平衡车
• 手环
• …

2015年Q3
小米 国内第一 全球第五

排名	品牌	Q3 2015
1	小米	16.5%
2	华为	16.2%
3	苹果	12.1%
4	OPPO	10.2%
5	vivo	10.0%
6	三星	8.8%
7	魅族	6.7%
8	酷派	4.8%
9	联想	4.4%
10	中兴	2.7%
	其他	7.6%

数据来源IHS

排名	品牌	Q3 2015
1	三星	23.8%
2	苹果	13.5%
3	华为	7.5%
4	联想	5.3%
5	小米	5.2%
	其他	44.8%

数据来源IDC

- 信贷
- 保险
- 理财
- 证券

便捷

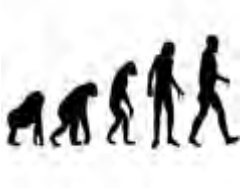
- 只需要一部小米手机
- 随时 / 随地

极速

- 1 分钟身份验证
- 1 分钟到账

灵活

- 第二天即可还款
- 根据信用不同，先息后本 / 等额本金等多种还款方式



进化

- 随着小米产品的使用 / 提交资料 / 使用贷款等方法可以提升信用
- 信用和新品公测 / 分期等结合


小米保险
业务平台
极致性价比的保险产品
支持保险业务快速成交、
快速服务快速理赔的 IT
平台
数据平台
基于智能设备与小米大数
据的保险业务风险建模
深度定制产品
保险公司
赢利点 2： 风控与
精算定价数据输出
赢利点 1： 产品
分销
健康险
车险
意外险
TO
C
TO
B
基于智能设备的
企业团体健康险

- 架构
- 工具

  -scribe

  -hadoop/hdfs

  -hbase

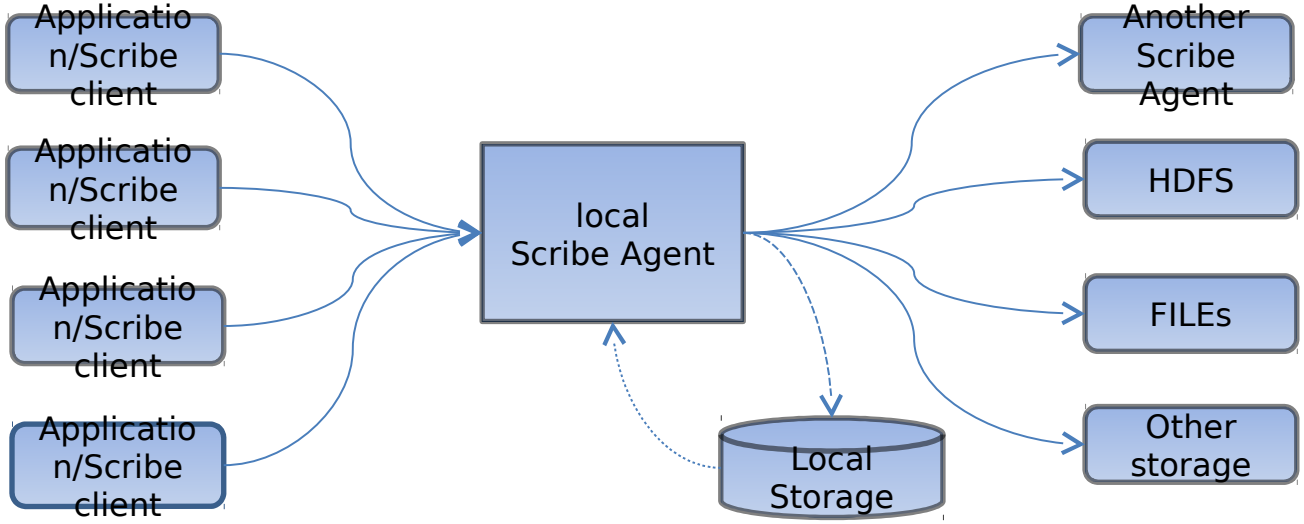
  -hive

  -impala

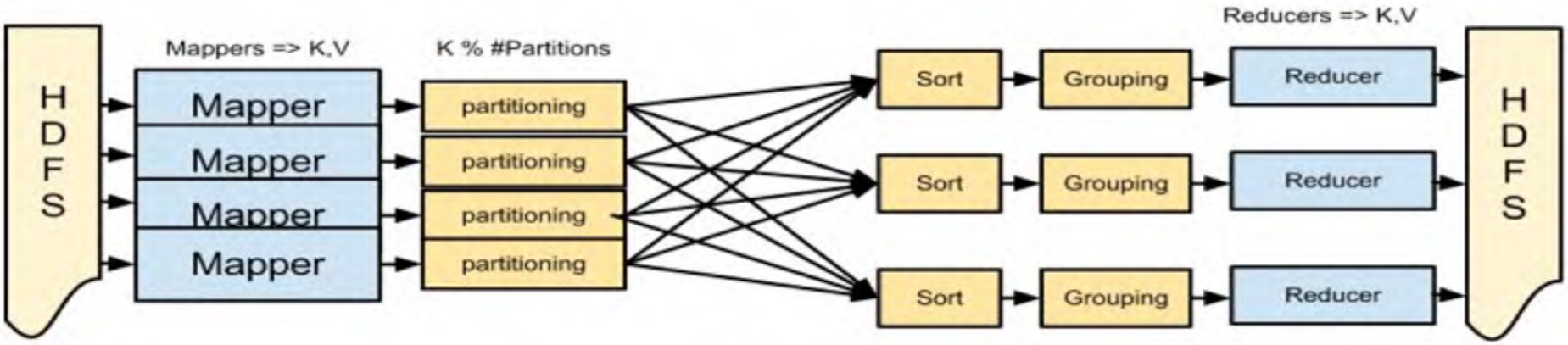
  -sqoop

  -spark


模型
信用模型
反欺诈模型
政策 / 运营
机器学习 / 数据挖掘
spark
Saiku
Hive
impala
R
数据仓库
KUDU
hbase
Hadoop/hdfs
scribe
sqoop
业务层
应用 / 主题商店
小米电视 / 盒子 / 视频
小米手环
智能家居
小米商城
多看阅读
小米血压计
外部数据（爬虫 / 合作数据）
小米金融 / 支付
小米生活
小米 push
等等
设备层



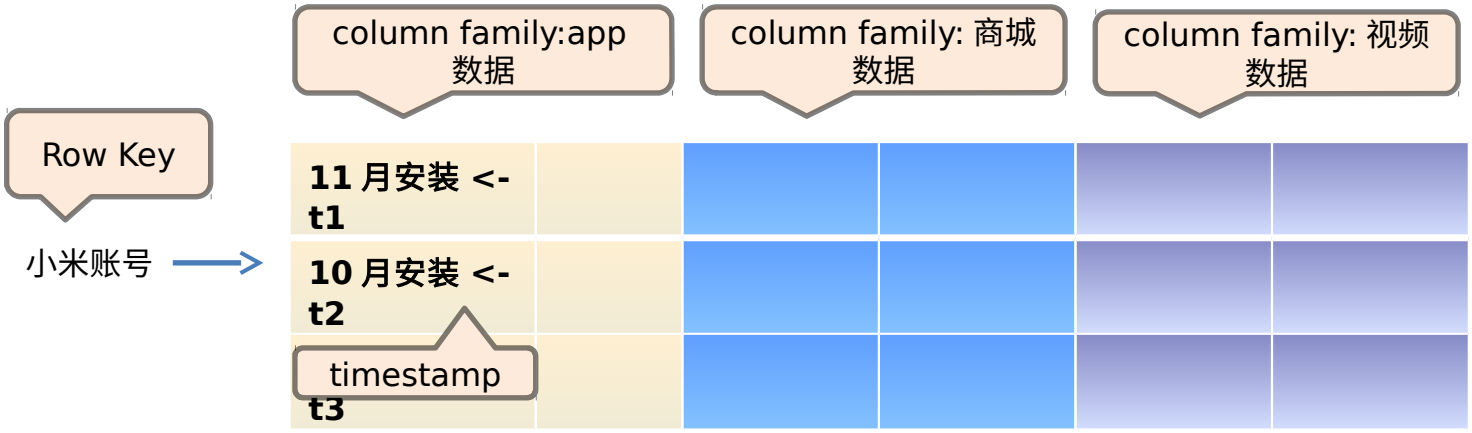
- 来自于 facebook
- 高性能
- 较好的容错性



The MapReduce Pipeline

A mapper receives (Key, Value) & outputs (Key, Value)
A reducer receives (Key, Iterable[Value]) and outputs (Key, Value)
Partitioning / Sorting / Grouping provides the Iterable[Value] & Scaling

- 每天上 T 数据
- ETL
- 批处理



- Column family 数量不能太多
- 线上服务（99% 读请求 10ms 左右，写请求 5ms 左右）
- 容易用 map/reduce 进行批处理

- 类 SQL 查询语言 / 易上手
- 无缝对接 hadoop/hdfs/hbase
- 使用 Sentry 进行权限控制
- 缺点：速度较慢

- 对应于 google 的 dremel
- 近实时 (分布式查询引擎 / 中间结果在内存 /LLVM/C++ 等)
- 类 SQL 查询
- 非常适合 OLAP

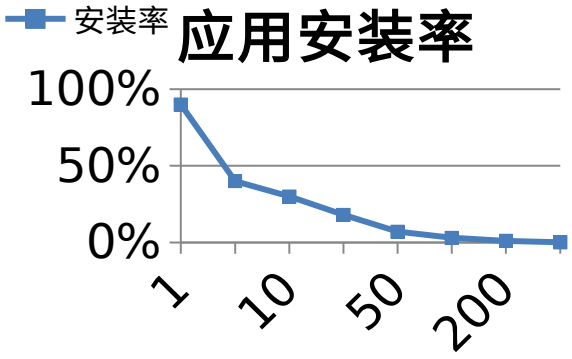
- 业务数据往往都在 mysql
- 从 mysql 到 hbase

- 比 hadoop 更通用（丰富的 API ）
- 高性能
- 良好的机器学习支持

- 目标
- 属性分析
- 数据稀疏性问题

- 金融属性
  - 收入 / 支出 / 资产 / 购物 / 理财 /…
- 行为属性 (人被行为所定义)
  -app/ 视频 / 图书 / 音乐 / 电话时间和次数 / 运动 /…
- 社交属性 (物以类聚，人以群分)
  - 居住区域 / 工作单位
  - 米聊 / 小米社区
  - 网络社交（微博 /linkedin ）
- 人口属性
  性别 / 年龄 / 学历 /…

- 基于 Spark
- 引入 GBDT+LR ， GBDT+FM 等方法自动发现、组合特征
- 采样：均衡性问题
- 去噪：部分业务数据可能有作弊数据



分类： SVM + 人工

- 应用元信息
- 用户评论评分
- 用户行为数据

定向

- 定向抓取
- 人工修正，需要一些领域知识

搜索引擎

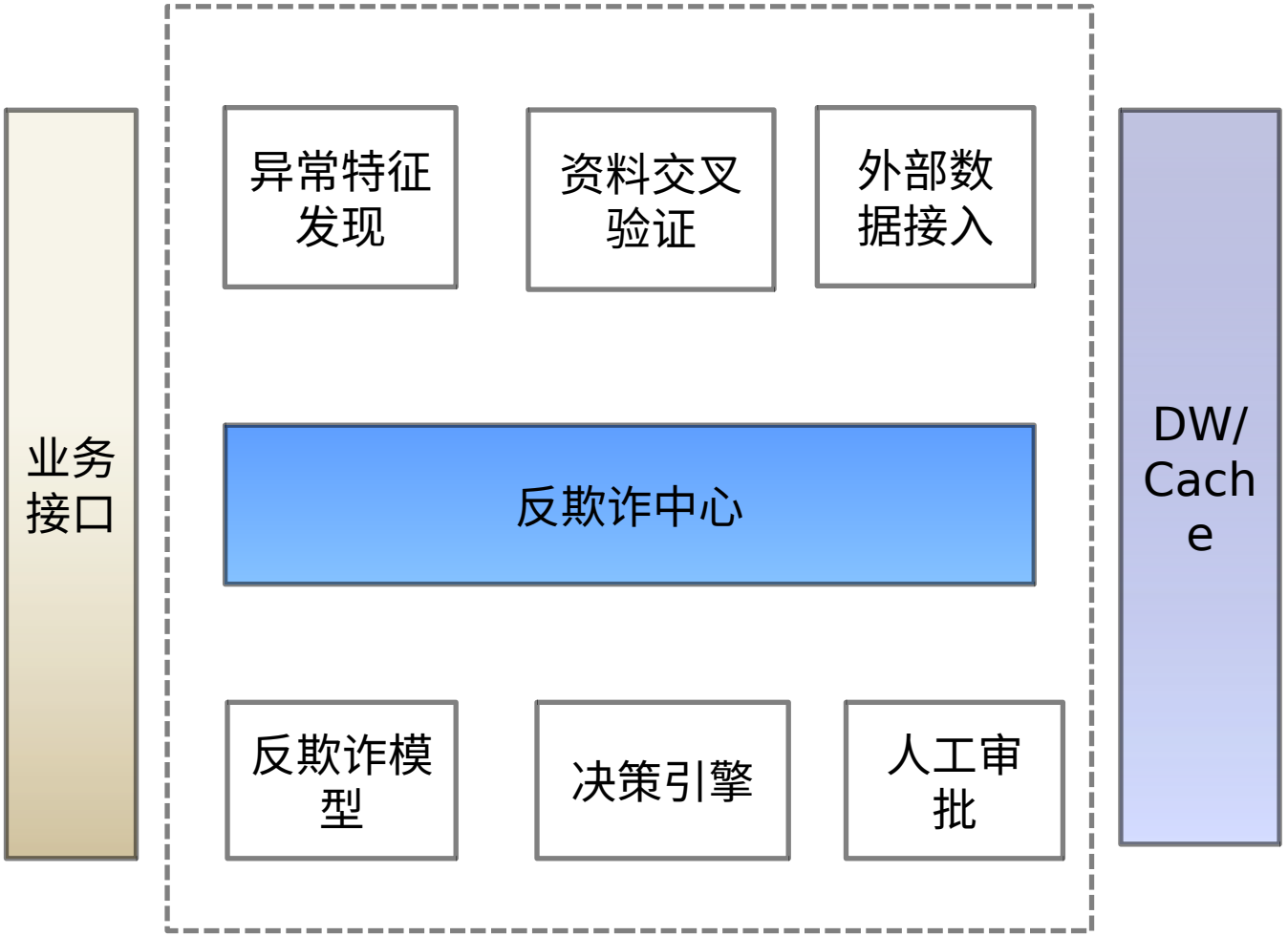
- 通过搜索引擎获得语义
- 机器学习分类

知识图谱

- 垂直搜索引擎
- 建立公司、职业的知识图谱

- 盗号 - 异常环境监测 / 手机验证
- 身份伪造 - 实名认证
- 虚假资料 - 交叉验证

大数据反欺诈 - 架构



- 反欺诈任重而道远—需要大家携手
- 如何衡量各自的价值?

Thanks