QCon

从“被虐”到“落地”——明略大数据产品演进实践

明略数据 刘诚忠

2015.4

目录

我们是谁

北京明略软件系统有限公司

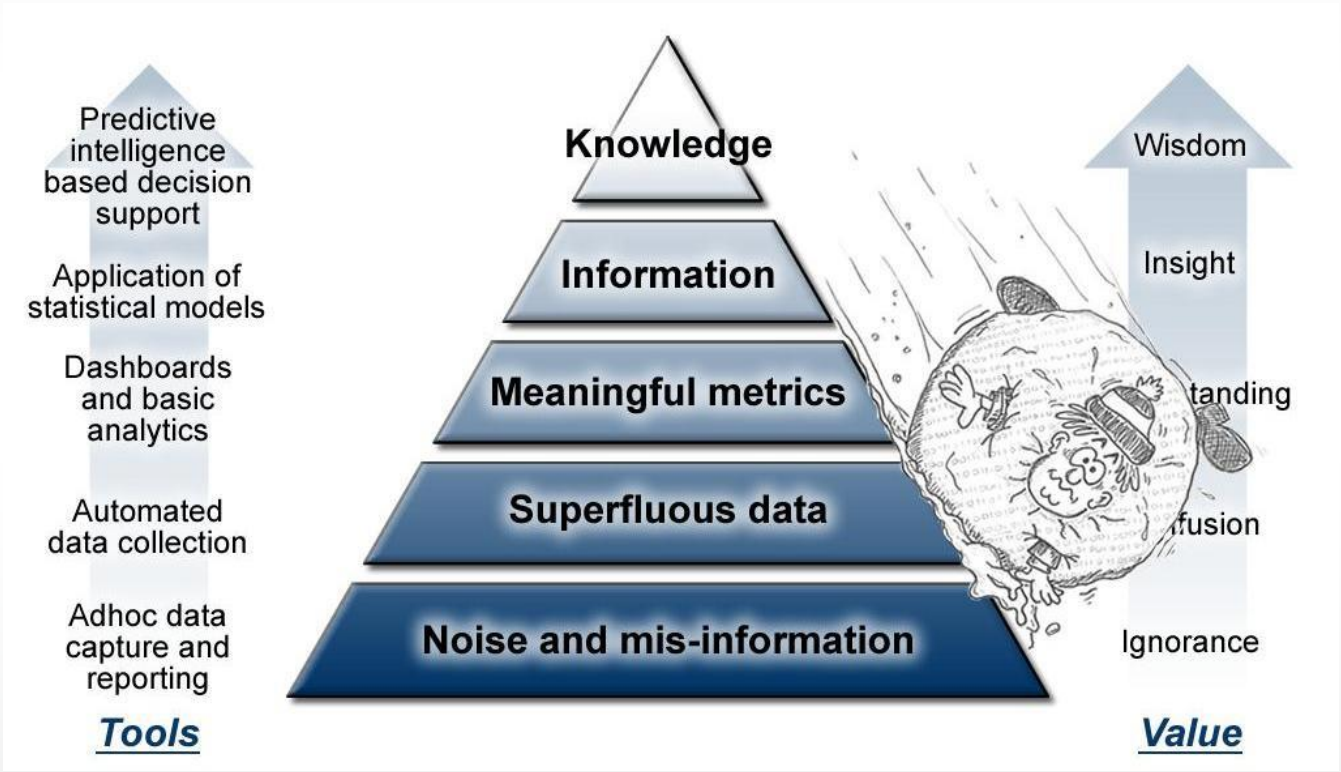
- 成立仅一年，66人
- 大数据平台，挖掘平台，数据工厂
- 国美，苏宁，北京台，银联，地税，邮储银行...




提供全行业的大数据整体解决方案，重点覆盖政府、金融、通信和零售四大支柱产业

理想和现实

期望到达



大数据的核心价值在于：挖掘隐藏在大数据背后的知识

目录

- WHY
- 大数据落地被虐实例
- 如何应对
- 案例分享

信心爆棚的进击

- Hadoop
- HBase
- Spark
- Storm
- Impala
- ML

很快感受到森森的恶意

丰富的数据源

权力的游戏

性能

更要命的问题——大数据？？

问题定义

整合

多源，异构，实时

保护

权限，集群，统一

分析

模型，效率，定制

交互

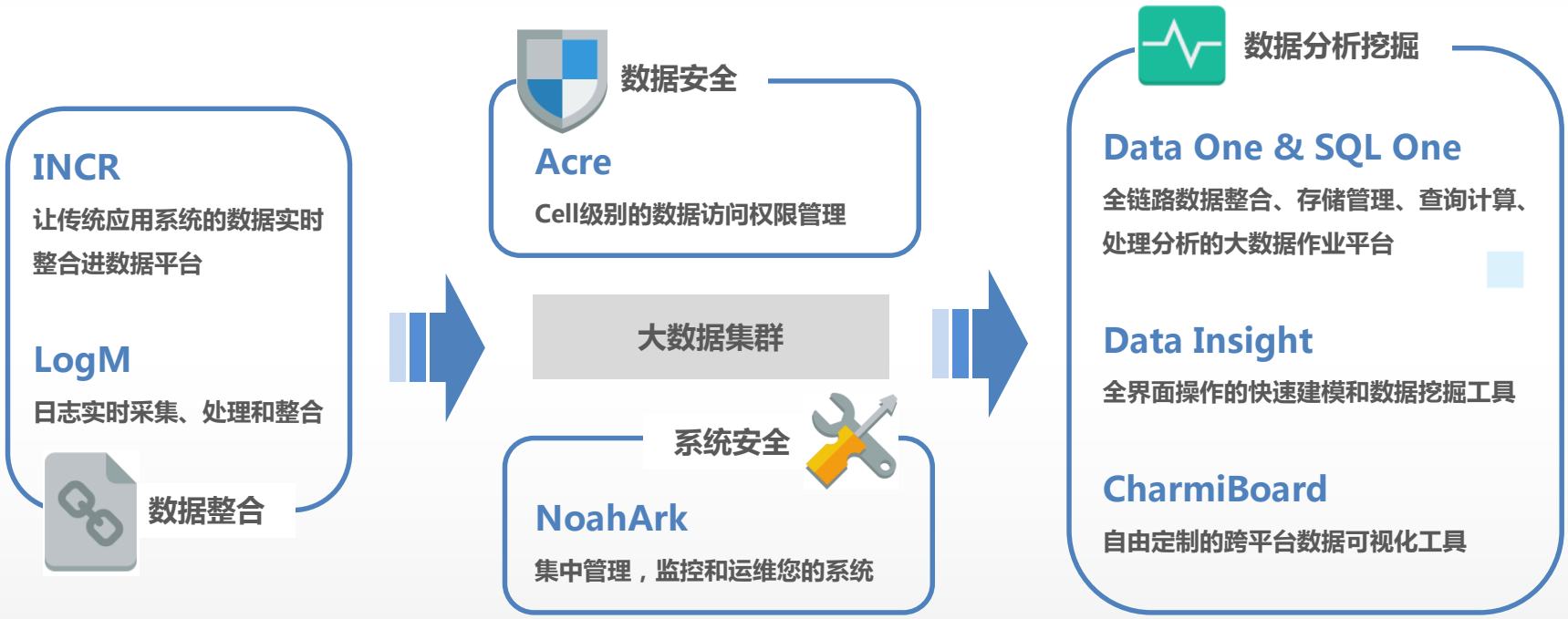
可视化，实时响应

目录

- WHY
- 大数据落地被虐实例
- 如何应对
- 案例分享

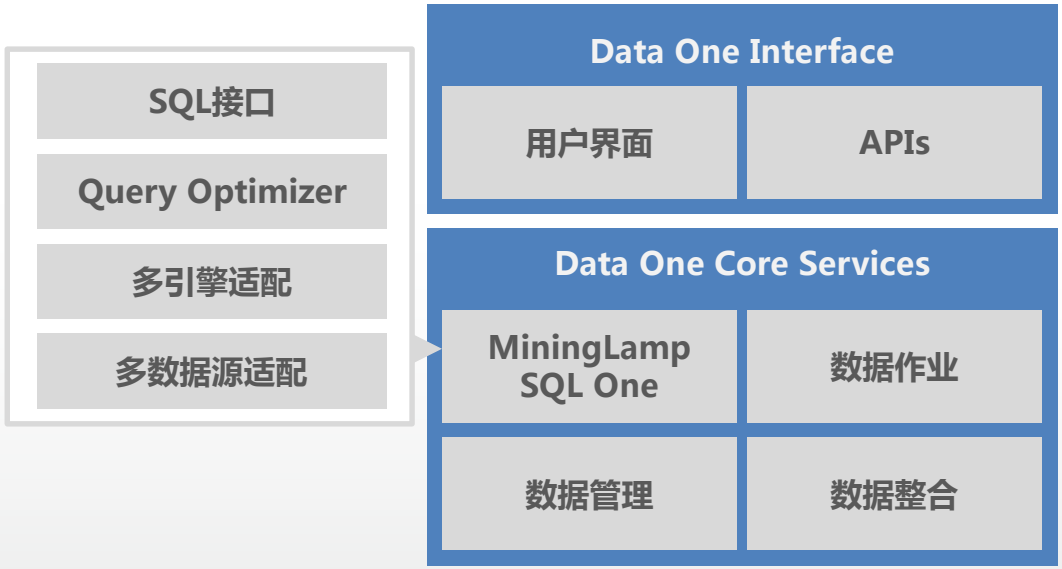
明略总体思路

核心产品组件


INCR
让传统应用系统的数据实时整合进数据平台
LogM
日志实时采集、处理和整合
数据整合
数据安全
Acre
Cell级别的数据访问权限管理
大数据集群
系统安全
NoahArk
集中管理，监控和运维您的系统
数据分析挖掘
Data One & SQL One
全链路数据整合、存储管理、查询计算、处理分析的大数据作业平台
Data Insight
全界面操作的快速建模和数据挖掘工具
CharmiBoard
自由定制的跨平台数据可视化工具

明略Data ONE

帮助企业实现数据治理的一站式大数据工作台
抽象设计，帮助业务方关注需求任务，不纠结底层技术
异构数据源混合查询的SQL引擎（专利申请中），可JOIN传统数据库，NoSQL，Hadoop数据



- **数据管理**

  管理平台中所有文件、结构化和非结构化数据

- **数据整合**

  依据分析场景，通过人机交互将异构的数据打通整合

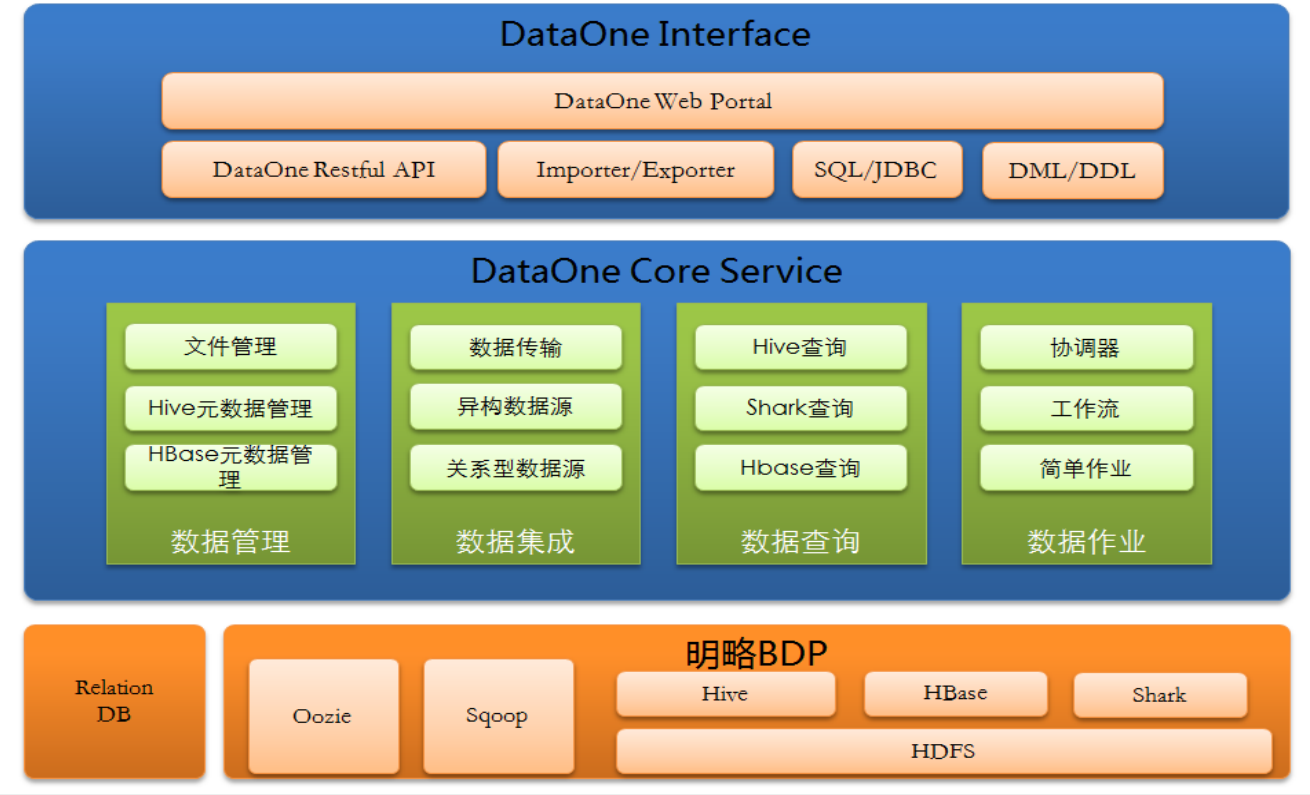
- **数据作业**

  实现不同复杂程度的数据处理和分析

- **人机交互**

  全界面操作，提供大量数据作业模板

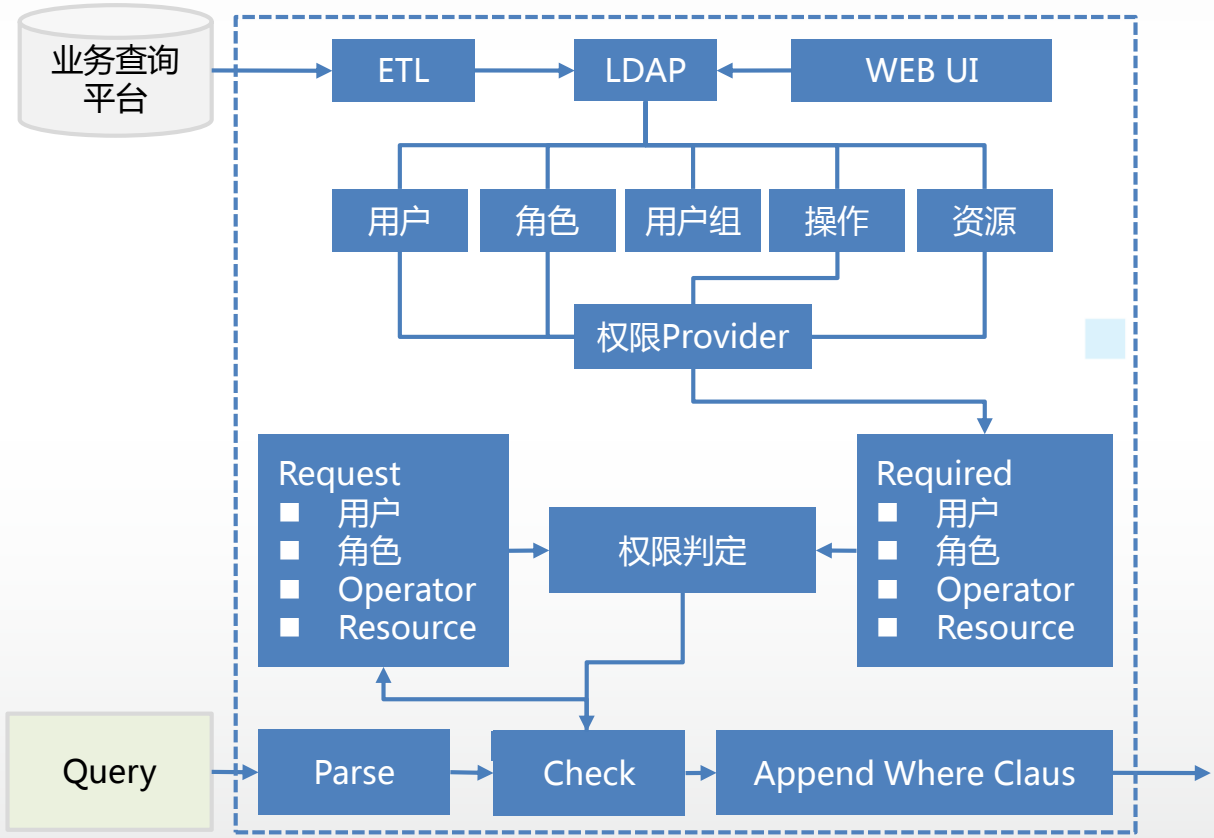
Data ONE系统架构



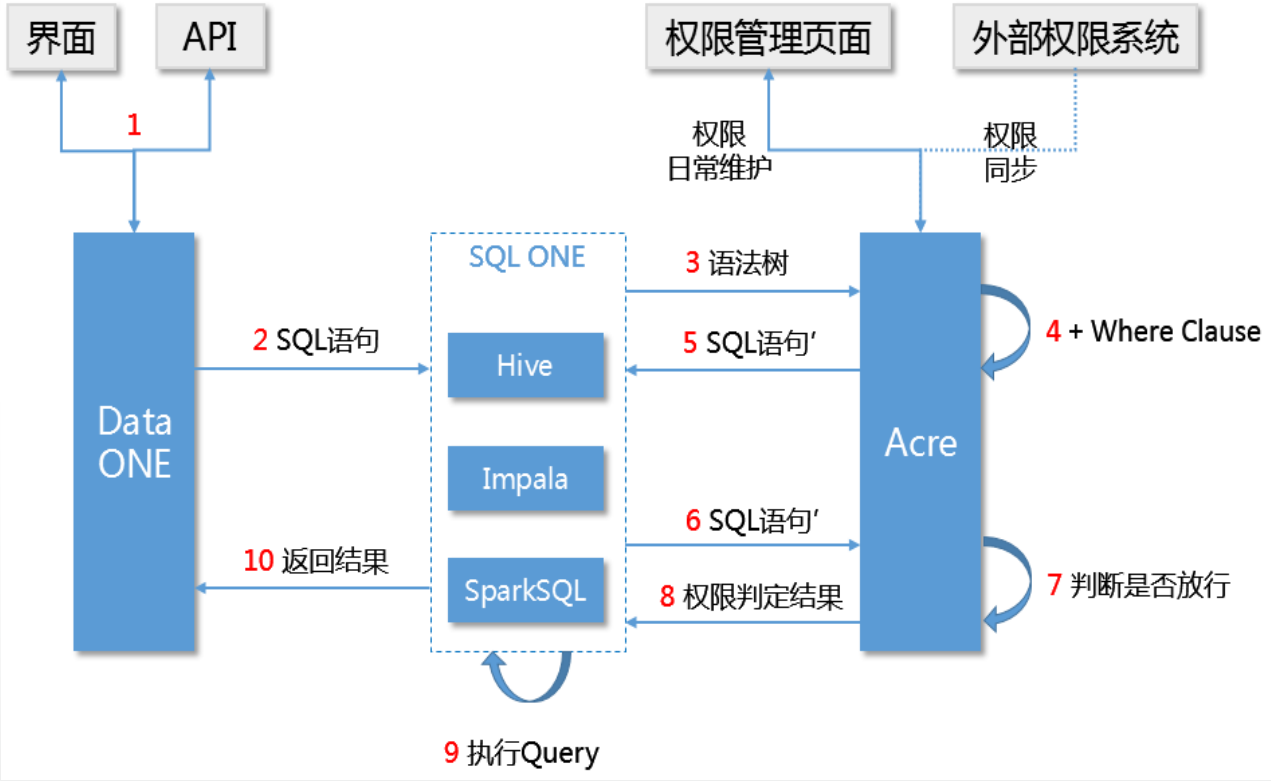
明略Acre

•支持Hive，Impala，MySQL多数据接口的统一授权管理（专利申请中）

•精确到cell级别的ACL/RBAC混合权限管理，超过市场同类产品（Cloudera列权限功能开发中）



Acre——系统架构


界面
API
1
权限管理页面
外部权限系统
权限
日常维护
权限
同步
SQL ONE
3 语法树
2 SQL语句
Hive
5 SQL语句'
4 + Where Clause
Data
ONE
Acre
Impala
6 SQL语句'
10 返回结果
SparkSQL
8 权限判定结果
7 判断是否放行
9 执行Query

明略Data Insight

带有调参反馈机制的可视化数据挖掘平台，为企业数据科学家打造的建模利器

集成所有主流数据挖掘算法

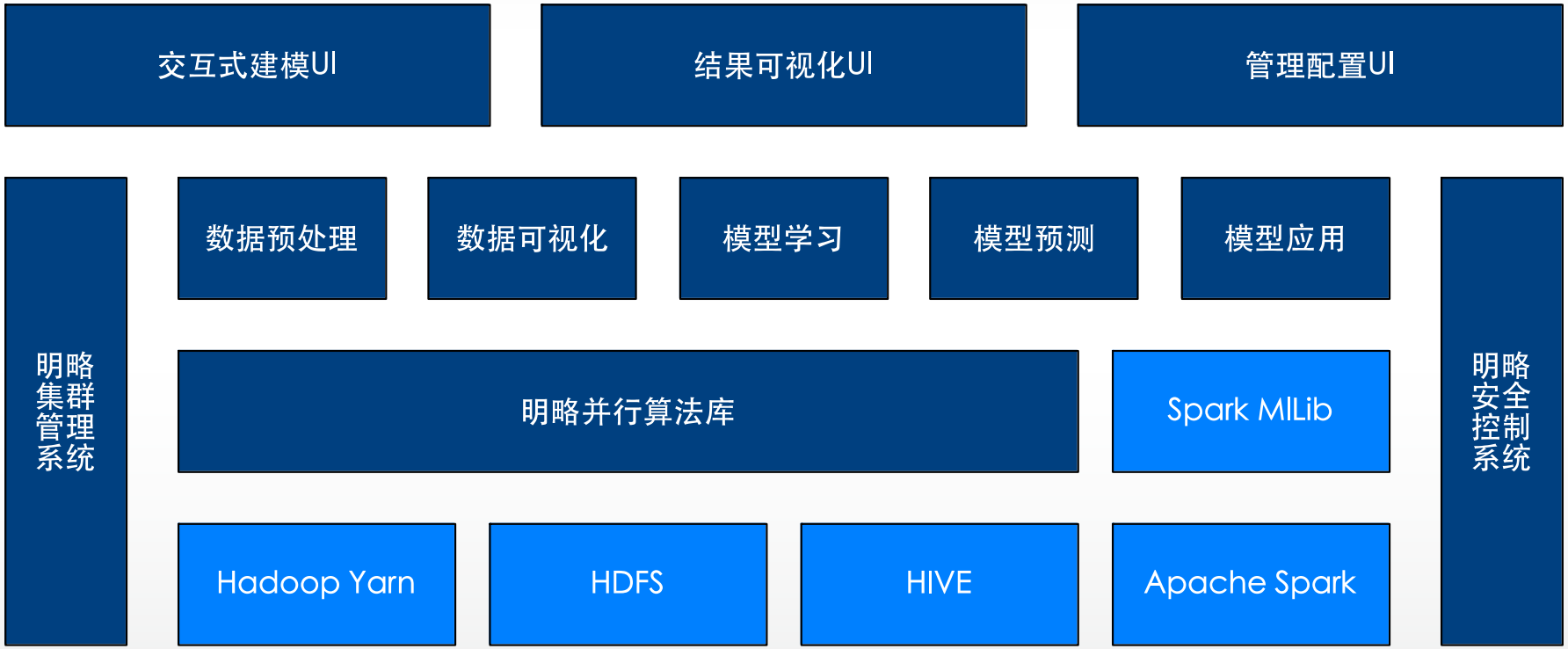
可视化快速建立数据挖掘模型

大大降低数据挖掘的交互复杂度

以反欺诈和征信为代表的垂直解决方案

- Decision Tree
- Logistic Regression
- Support Vector Machine
- Multinomial Naive Bayes
- Regression Tree
- Linear Regression
- Lasso Regression
- Ridge Regression
- K-Means
- …

DI——系统架构


交互式建模UI
结果可视化UI
管理配置UI
明略集群管理系统
数据预处理
数据可视化
模型学习
模型预测
模型应用
明略安全控制系统
明略并行算法库
Spark MILib
Hadoop Yarn
HDFS
HIVE
Apache Spark

DI——DSL支持

```
filter by $age > 20
group by name = $name
feature aage = AVG($age)
val tmp = $age + 1
feature a = first($tmp)
```

```
group by mzid = $mzid
feature transform = $transform
feature stable = $stable
feature caidCLK = sumWith($caid, 1 if $action == "CLK" else 0)
feature caidIMP = sumWith($caid, 1 if $action == "IMP" else 0)
feature spidCLK = sumWith($spid, 1 if $action == "CLK" else 0)
feature spidIMP = sumWith($spid, 1 if $action == "IMP" else 0)
```

- DataInsight自定义了简单的脚本语言，用来处理一些较为复杂的数据变换
- 目前DSL支持以下功能，基本满足常见的数据转换需求。
  - 数据过滤
  - GroupBy
  - 常用数学函数
  - 类型转换函数
  - 字符串操作
  - Map操作
  - 统计函数

DI——算法列表

分类算法

- SVM
- Logistic Regression
- Native Bayes
- Decision Tree
- Random Forest

聚类算法

- K-means
- DBScan

回归算法

- Lasso Regression
- Ridge Regression
- Linear Regression
- Gradient Boosted Regression
- Regression Tree

推荐算法

- Item based CF
- User based CF
- Alternating Least Squares

特征变换

- PCA

自然语言处理

- LDA
- Word2Vec

频繁模式

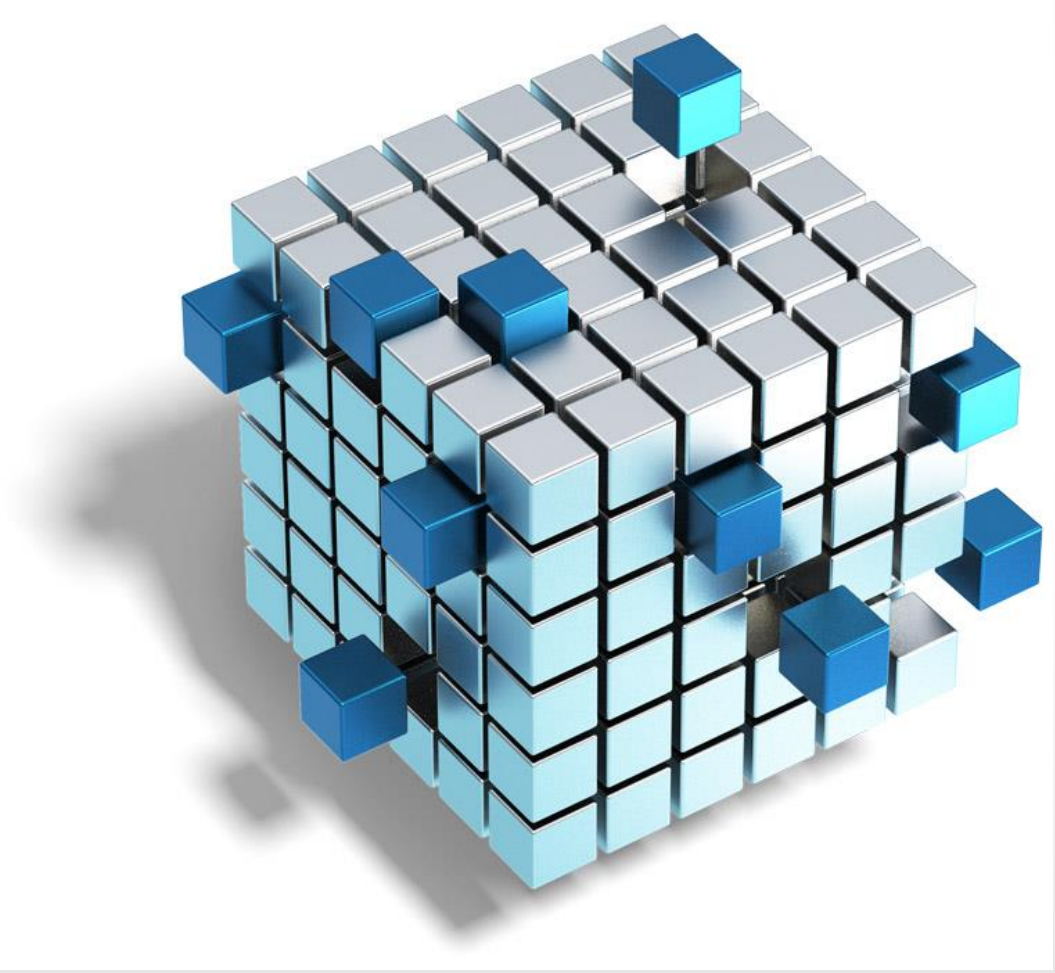
- FPGrowth
- BIDE

数理统计

- Correlation Analysis
- Distribution Statistics

基于Spark的并行化算法

新一代BI

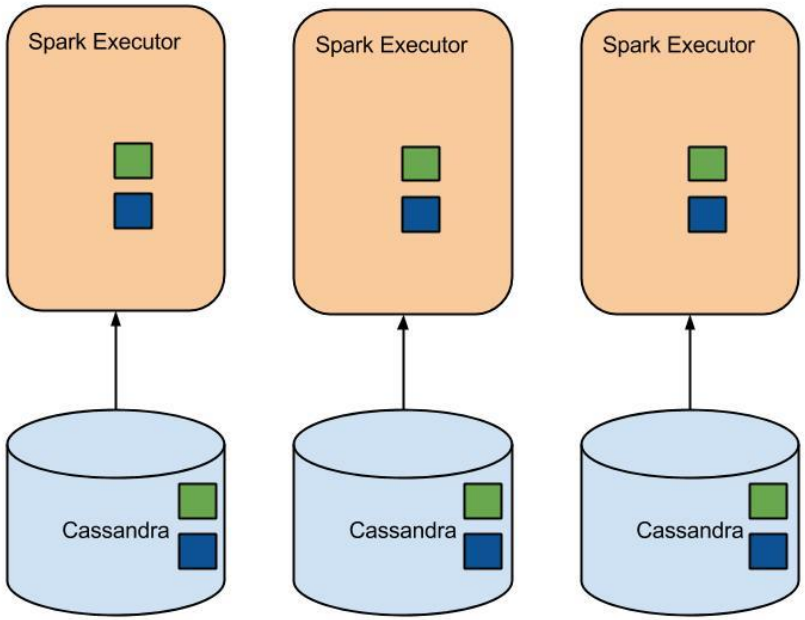


OLAP CUBE？

新一代BI

- 一站式的意义——DATA GRAVITY
- 要考虑到复杂分析可能——OLMP
- 分布式的痛永远在单点——最大限度去中心化

新一代BI



DATASTAX

Stratio

http://velvia.github.io/presentations/cassandra-spark-olap-2014/index.html#/25/2

新一代BI

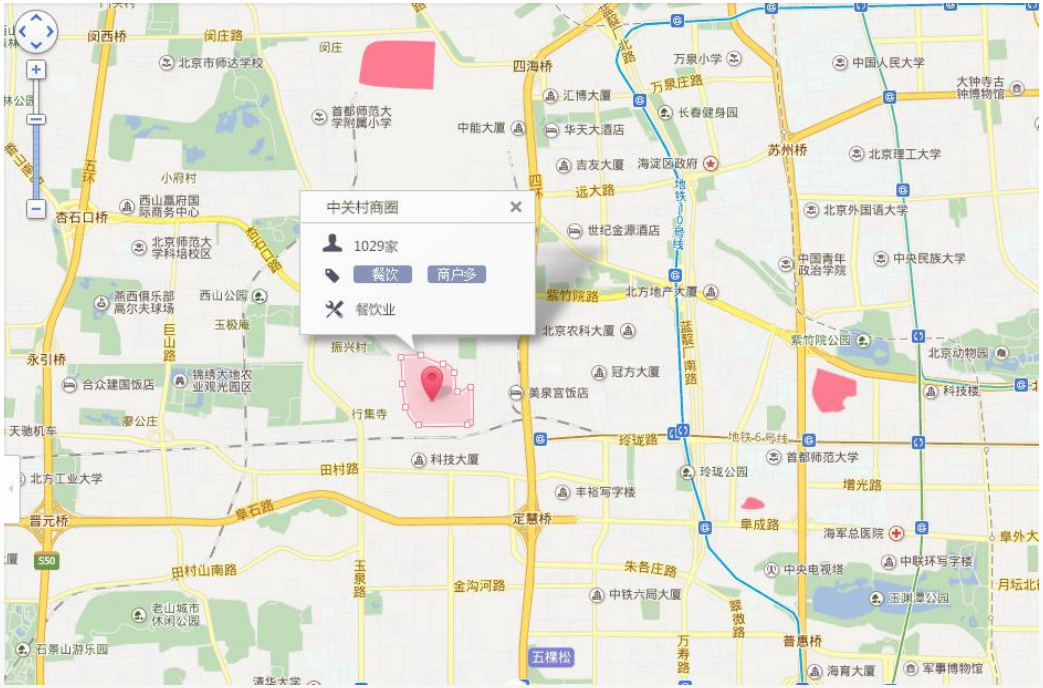
- GDELT dataset, 117 million rows, 57 columns, ~50GB
- Spark 1.0.2, AWS 8 x c3.xlarge, cached in memory
- Adhoc：　　0.49
- TOP K：　　1.51
- TOP Group By：　　2.69

http://velvia.github.io/presentations/cassandra-spark-olap-2014/index.html#/25/2

目录

- WHY
- 大数据落地被虐实例
- 如何应对
- 案例分享

案例——商圈聚类



- 通过商户的地理坐标，将全国所有城市中的商户聚集成商圈
- 使用DBScan算法进行商圈聚类
- 成功的将全国300个城市中的商户聚集成商圈，发现了很多人工未能标注的商圈。

案例——消费预测

序号	持卡人	交易时间	交易金额	交易商户	所在商圈
1	6226020000586980	上午	5,000	家乐福（学院路店）	五道口商圈
2	620049959996969	上午	5,000	家乐福（学院路店）	五道口商圈
3	6226020000586980	下午	5,000	家乐福（学院路店）	五道口商圈
4	620049959996969	晚上	5,000	家乐福（学院路店）	五道口商圈
5	620049959996969	晚上	5,000	家乐福（学院路店）	五道口商圈
6	620049959996806	晚上	5,000	家乐福（学院路店）	五道口商圈

- 通过用户的行为数据，对用户未来消费行为进行预测
- 使用基于概率转移矩阵的自定义算法对用户消费行为进行预测
- 预测结果包括：
  - 消费的商户
  - 消费次数
  - 消费平均金额
  - 消费时间属性

总结

- 大数据技术正在从互联网公司往传统行业飞速扩展，技术应用程度有gap但已经不大
- 应用更实时，更敏捷，更偏决策导向，IT层在变薄变轻，IT人需要重新定位
- 数据互联互通将成为全行业刚需，权限审计和行业规范是目前障碍
- 技术的进步并未逾越工具范畴，帮助人决策

www.infoq.com/cn

InfoQueue

@明略数据 明略数据

软件
正在改变世界!