荣之联大数据平台的应用实践

2017/10

议程

- 荣之联大数据平台的应用案例介绍
  - 商务中心大数据中心建设案例
  - 证券交易日志分析案例
  - 工业物联网大数据平台案例
- 荣之联大数据平台产品介绍
  - 产品架构及优势
  - 产品特色功能介绍

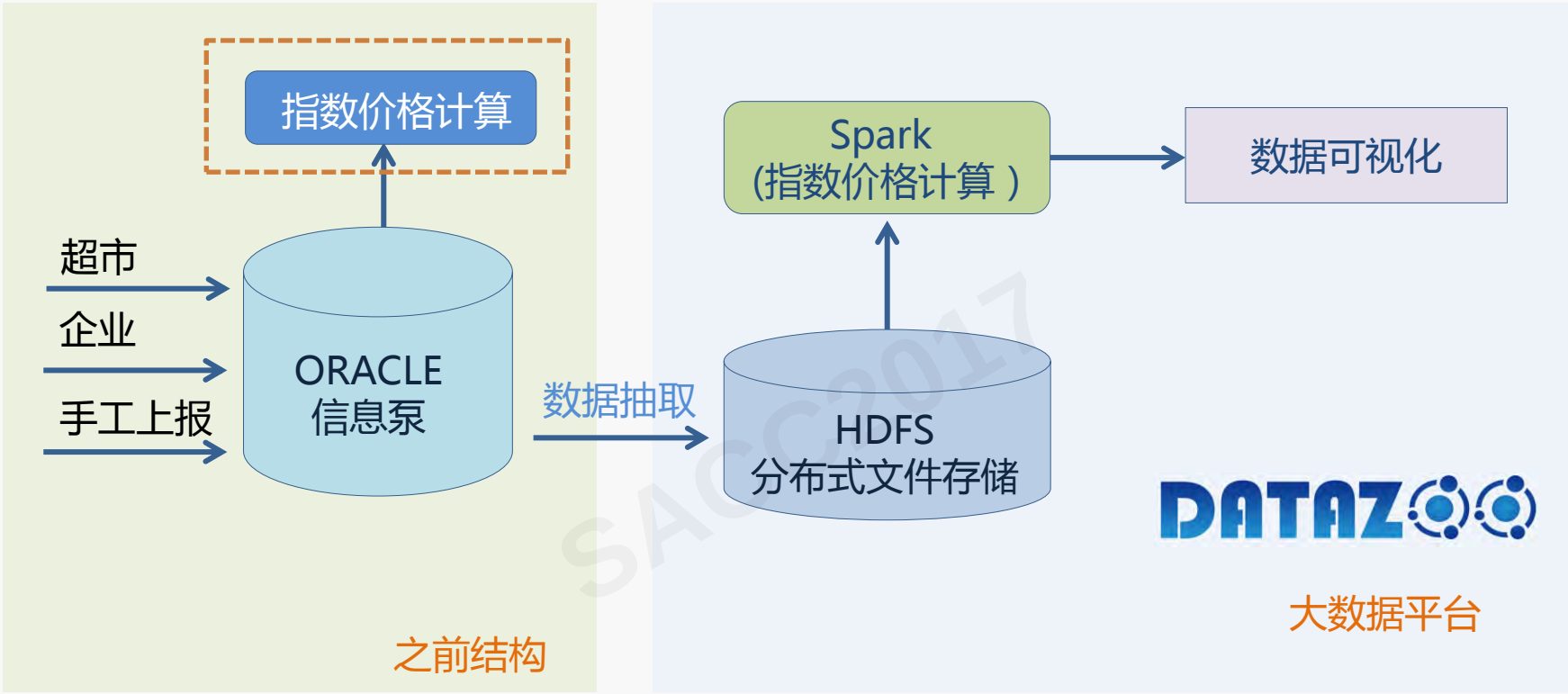
议程

- 荣之联大数据平台的应用案例介绍
  - 商务中心大数据中心建设案例
  - 证券交易日志分析案例
  - 工业物联网大数据平台案例
- 荣之联大数据平台产品介绍
  - 产品架构及优势
  - 产品特色功能介绍

商务中心大数据中心建设需求

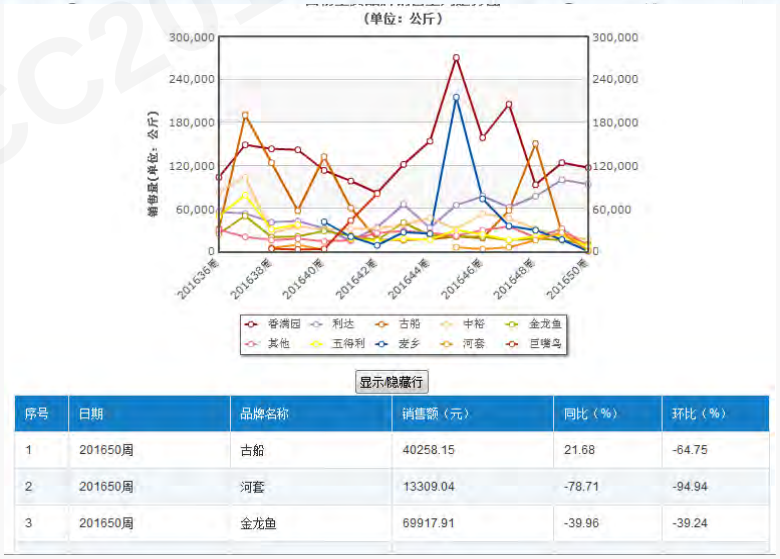
- 业务系统性能下降
- 运算速度缓慢
- 系统支撑能力面临增长瓶颈
- 非结构化数据缺乏管理和应用

商务中心大数据中心建设



客户价值

- ✓ **构建了大数据平台**
- ✓ 拥有处理更大数据量、更复杂格式、更**多样化**数据的能力
- ✓ 加速数据采集、使用和分析的时间，更快决策
- ✓ 处理的**数据量**：一个月1-1.3亿条数据，即10-13Gb
- ✓ 原来的处理+计算需要5个小时，现在需要2个半到3小时。



序号	日期	品牌名称	销售额（元）	同比（%）	环比（%）
1	201650周	古船	40258.15	21.68	-64.75
2	201650周	河套	13309.04	-78.71	-94.94
3	201650周	金龙鱼	69917.91	-39.96	-39.24

议程

- 荣之联大数据平台的应用案例介绍
  - 商务中心大数据中心建设案例
  - 证券交易日志分析案例
  - 工业物联网大数据平台案例
- 荣之联大数据平台产品介绍
  - 产品架构及优势
  - 产品特色功能介绍

X券商交易系统现状与问题

- 全国22个站点
- 每个站点约20台左右的网上交易事务机，共计近400台网上交易事务机
- 每天网上交易原始日志的日增量为90-120G

分散在全国的各个站点和服务器的状态如何监控？出现宕机或假死状态时能否及时判断和处理？

面对证监会的监管要求，或者客户的查询请求，如何对原始日志进行快速、精确的查询？

基于历史数据的统计分析报表的计算效率是否需要提高？

券商应用日志分析系统架构

UEC GROUP LTD

应用

- 实时运维监控
- 实时预警
- 明细毫秒级查询
- 实时报表
- 离线统计报表

数据处理平台

- 海量数据多维分析引擎
- 海量数据高并发检索引擎
- 实时数据流处理引擎

数据存储管理平台

- 日志数据
- 数据字典

数据采集平台

数据分发与解析处理层

- 解析、勾对、计算、入库

数据缓冲层

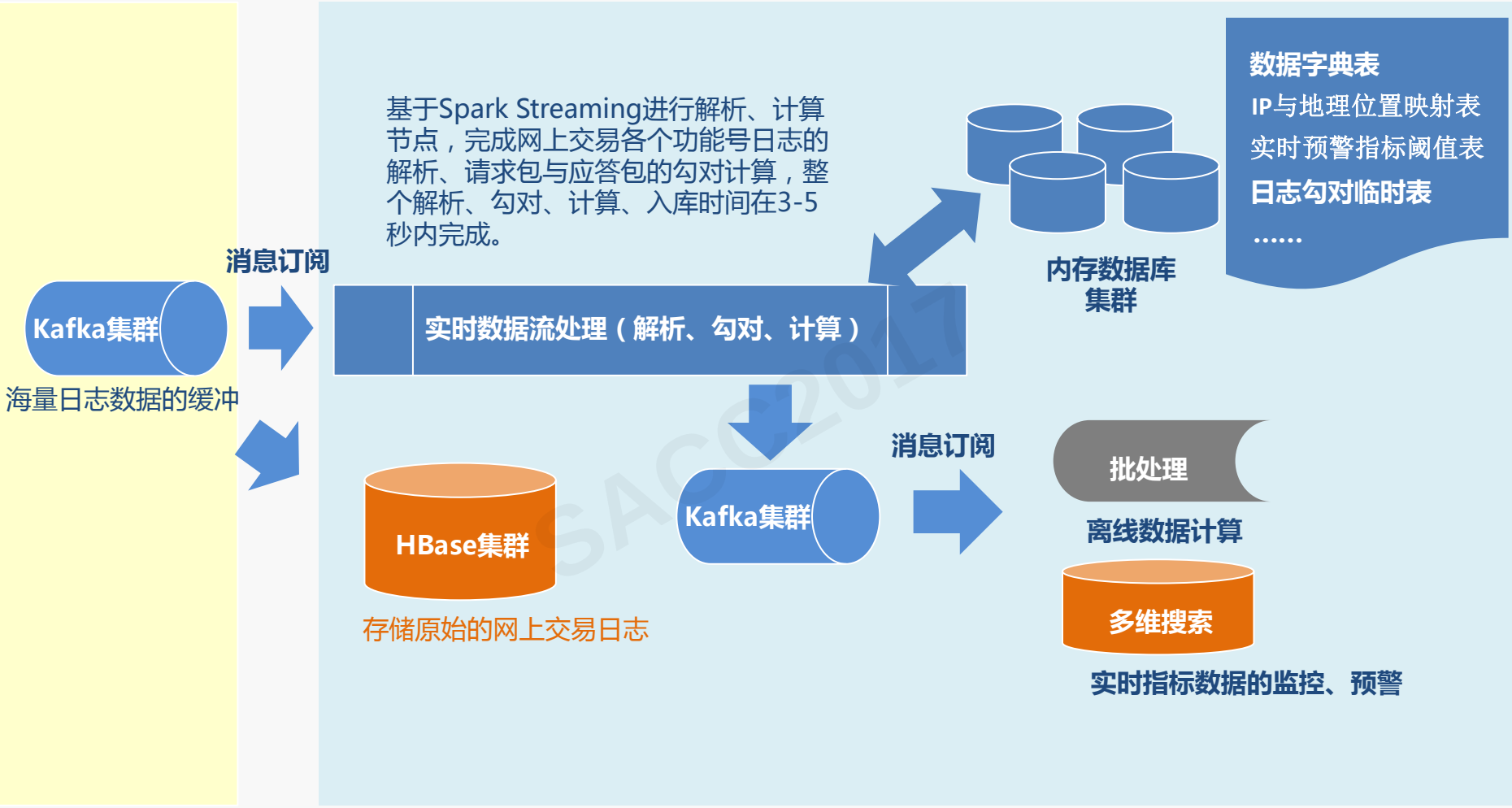
- 通达信topic
- 恒生topic
- 核新topic
- 网络设备topic

数据汇集层

- 通达信汇集集群
- 恒生汇集集群
- 核新汇集集群
- 设备数据汇集集群

监控及运维管理

券商应用日志分析系统 – 实时计算



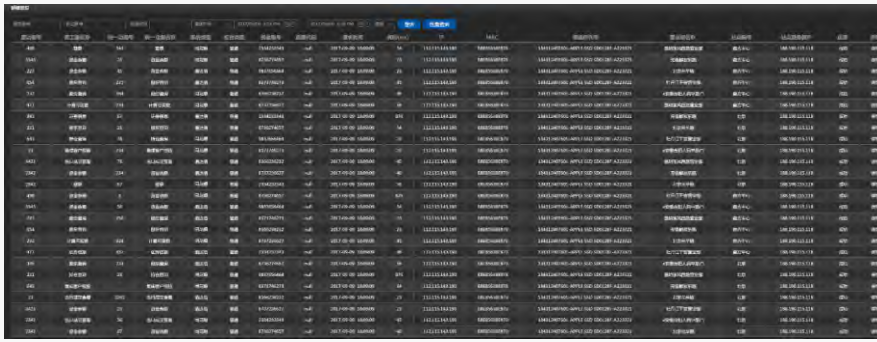
客户价值

- ✓ 为信息技术部**提供更完善的运维管理支持**，对站点/交易事务机的各技术指标做到实时监控，针对交易事务机的假死、宕机、恶意攻击等异常情况做到及时有效应对，防止影响到正常网上交易业务的正常运行。

- ✓ 针对资金账号的各种**异常情况的实时预警**，协助及时发现异常的资金账号，做出更加有效的管控和处理。

- ✓ **秒级的日志查询**，面对监管方的日志查询要求或客户请求，快速响应。

- ✓ 依托大数据技术的日志实时**采集与分析平台的搭建**，为某证券公司未来实现全系统、各交易品种的日志接入和大集中管理、基于海量数据的业务分析做了平滑的铺垫。



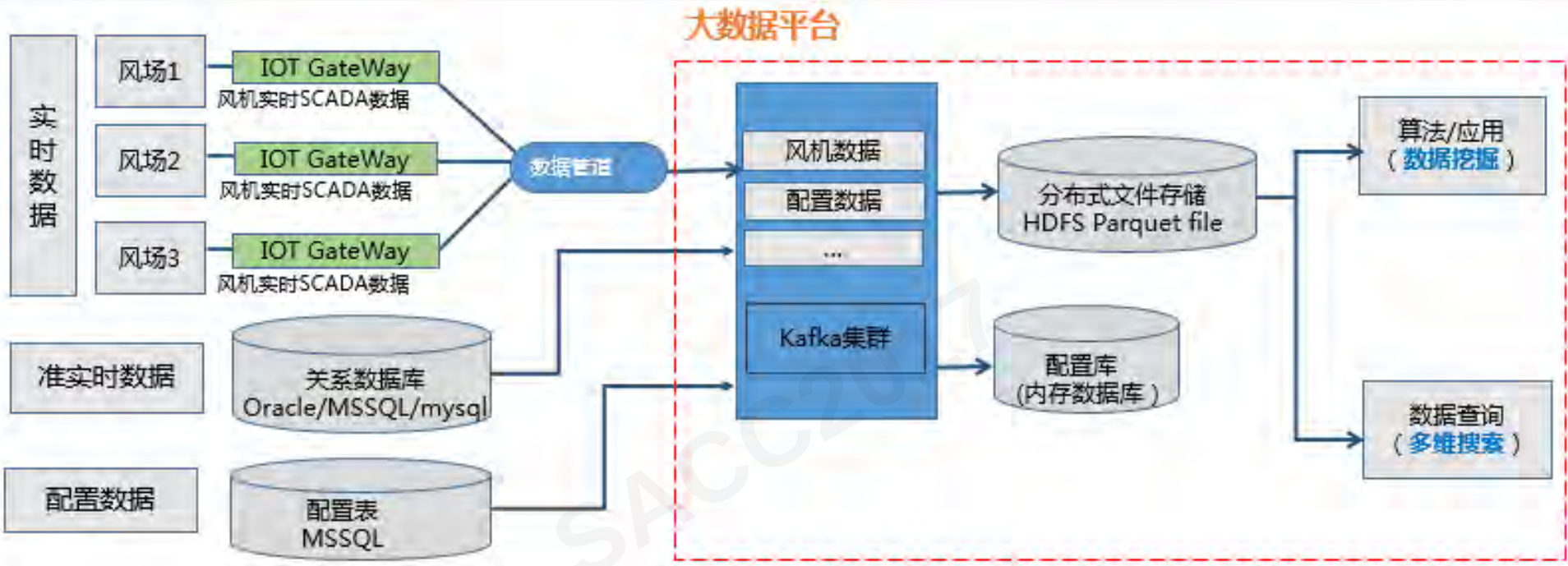
议程

- 荣之联大数据平台的应用案例介绍
  - 商务中心大数据中心建设案例
  - 证券交易日志分析案例
  - 工业物联网大数据平台案例
- 荣之联大数据平台产品介绍
  - 产品架构及优势
  - 产品特色功能介绍

风机数据资源现状

数据应用	数据管理	数据平台
• **数据存储分散**：统计数据、历史全量数据是云存储，本地存储载荷分析相关的数据 • **风资源数据分散**：散落到不同的业务系统中	• **子系统多而独立**：信息重复、信息堡垒和信息孤岛等问题出现 • **不能快速抓取到数据**：跨部门的业务及新业务需求需要数据支撑时，不能快速的抓取到所需要的数据（风机数据，业务数据，水务数据等等） • **集团级统计数据慢**：数周时间才能获得集团级统计分析数据	• **数据平台试验环境**：目标是生产环境 • **改进现有的数据处理业务**：云平台存在系统安全隐患及其他未知的风险

能源物联网大数据平台搭建



数据采集与处理平台

数据采集代理

数据流处理

数据入库

数据存储

分布式文件存储

采用Parquet 列存储

优化存储模型

计算优化

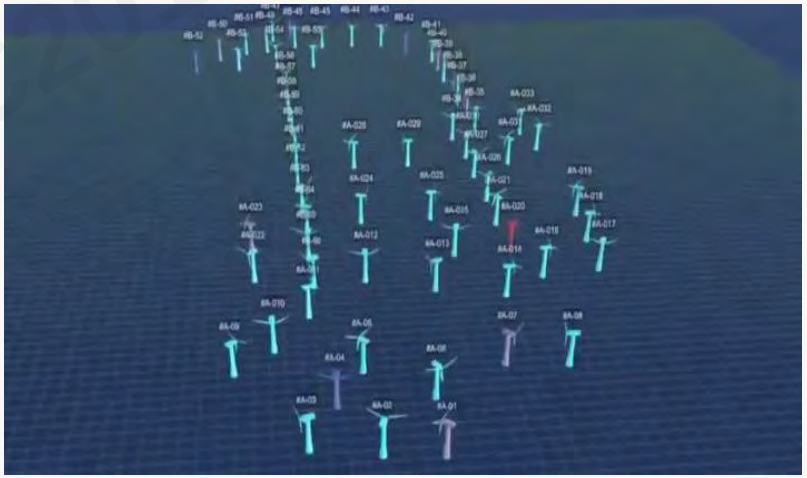
优化数据读取性能

算法本身优化

采用数据科学开发工具

客户价值

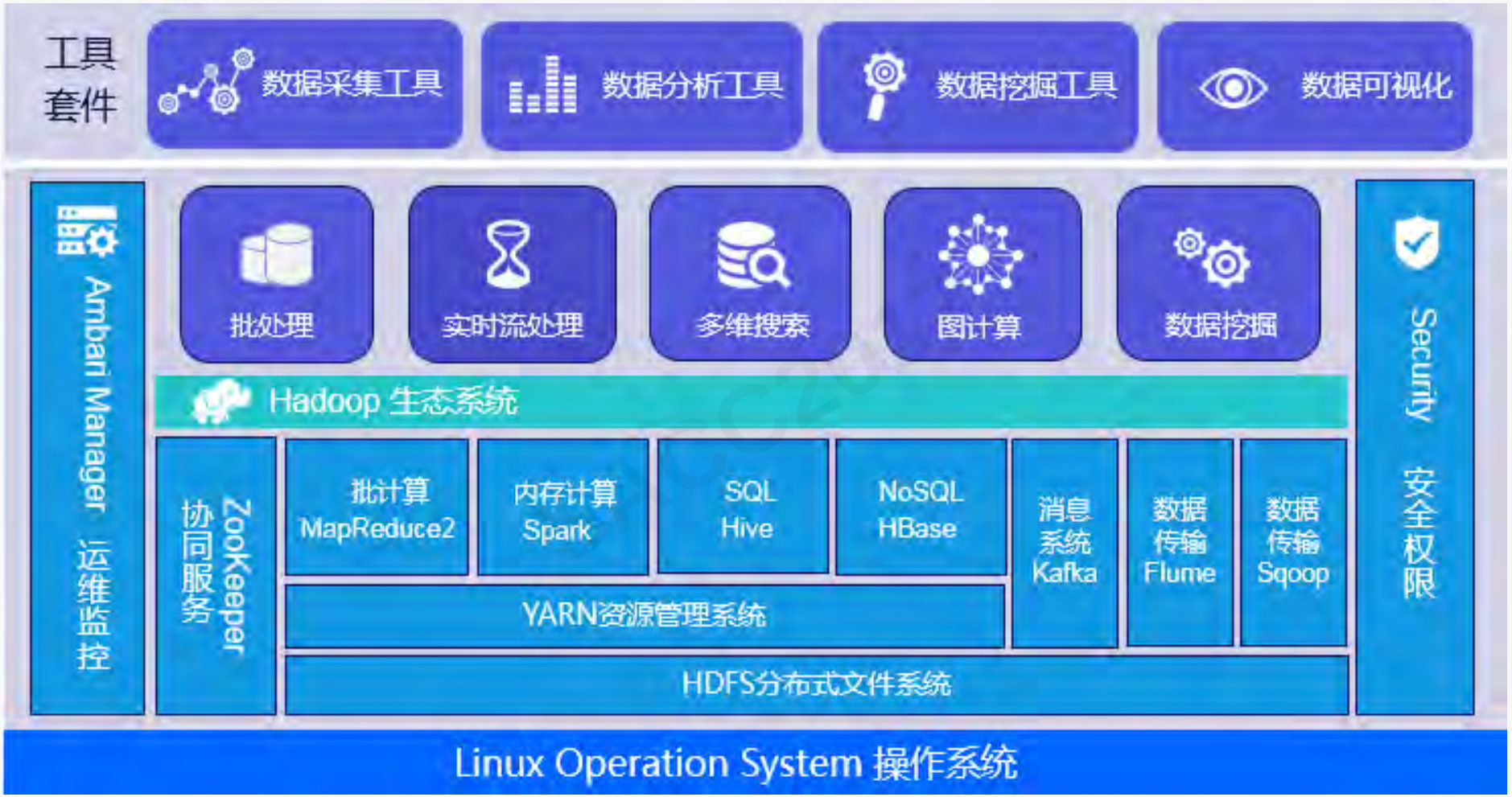
- ✓ 可弹性伸缩的大数据平台
- ✓ 数据集中存储，推动企业**数字化转型**
- ✓ 有效支撑**物联网**下大数据的应用
- ✓ 数据资产**集中化管理**，包含：数据采集和存储、数据查询、数据应用等
- ✓ 快速为管理层和业务层提供**数据服务**



议程

- 荣之联大数据平台的应用案例介绍
  - 商务中心大数据中心建设案例
  - 证券交易日志分析案例
  - 工业物联网大数据平台案例
- 荣之联大数据平台产品介绍
  - 产品架构及优势
  - 产品特色功能介绍

荣之联DataZoo大数据平台



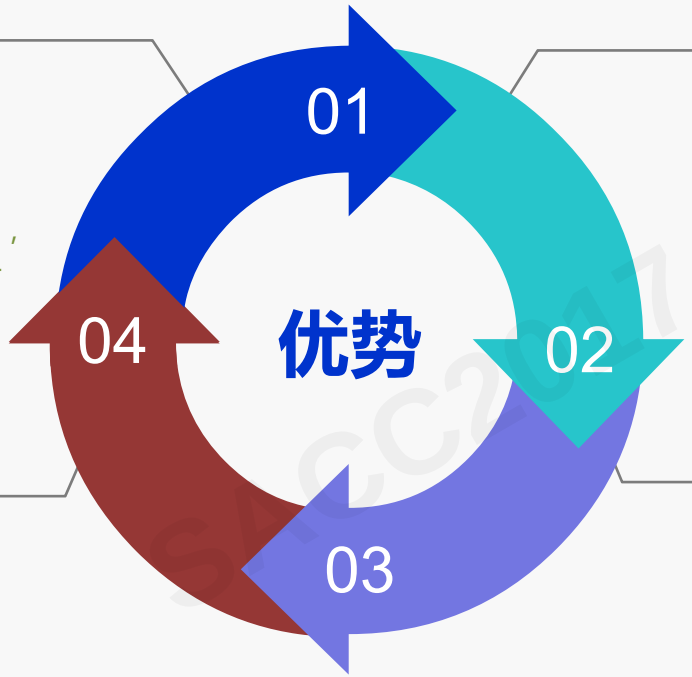
产品技术优势

完整的数据处理能力

- 新一代大数据平台，不仅仅只是Hadoop生态的集成
- 具备数据分析生命周期各方面能力，从数据采集、分析、挖掘到数据可视化都有相应的组件能力
- 平台自有特色功能

丰富的行业应用模型

- 驾驶行为分析
- 碰撞分析模型
- 用户画像
- 推荐模型
- 文本分析（分词、情感分析）



高性能的多维分析引擎

- 超低时延：分析结果一触即发
- 分析特性：任意维度组合分析
- 存储特性：嵌套列存储，计算过程不加载多余数据
- 强扩展性：支持横向纵向任意扩展
- 数据时效性：实时+离线数据

便捷的开发管理工具

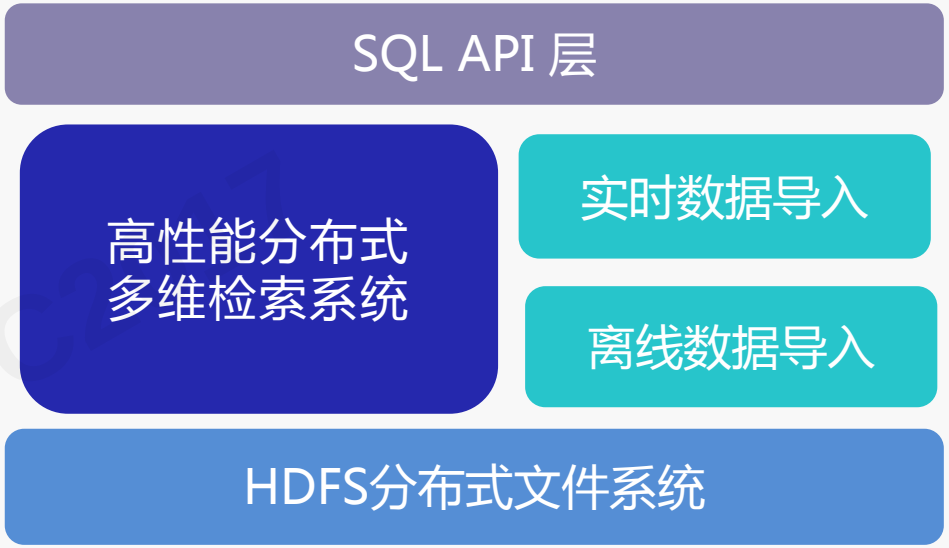
- 可视化管理、监控系统
- 统一的配置管理
- 从数据采集、预处理到数据分析与挖掘的工具套件

议程

- 荣之联大数据平台的应用案例介绍
  - 商务中心大数据中心建设案例
  - 证券交易日志分析案例
  - 工业物联网大数据平台案例
- 荣之联大数据平台产品介绍
  - 产品架构及优势
  - 产品特色功能介绍

特色能力-多维搜索

- 将数据存储在HDFS之上，基于HDFS做了磁盘与网络做了读写控速逻辑
- 与Spark深度集成，Spark对检索结果集直接分析计算，同样场景让Spark性能加快百倍
- 数据即可离线导入也可实时导入，索引即时生成，通过索引高效定位到相关数据



特色能力-行业模型

驾驶行为模型	用户画像	文本分析-分词
识别用户风险，改善车主驾驶行为，降低车险的赔付成本。	标签化用户模型，提供360度客户视图。	构建大量文本的切词模型，是文本分析的基础模型。
碰撞分析模型	**推荐模型**	**文本分析-情感分析**
通过识别事故真相减少车险的欺诈赔款，并提供更好的理赔服务。	基于用户的行为数据，做出精准式营销。	基于上下文语义，语句情感模型，应用于舆情监控、商品评估等场景。

Predictive 预测	Segmentation 分类	Associations 关联	Correlations 相关性	Similarities 相似度	分词模型	情感分析模型
Tress 树归纳法	DBSCAN	Apply Association Rule 应用关联规则	ANOVA Matrix 矩阵	Cross Distances 交叉距离	条件随机场 CRF Conditional random field	长短时记忆神经网络 (深度学习) Long Short Term Memory
Neural Nets 神经网络训练	Top Down Clustering	Create Association Rule 创建关联规则	Correlations Matrix 相关矩阵	Data to Similarity		
Functions 函数拟合	K-Means k-均值	FP-Growth	Covariance Matrix 协方差矩阵	Data to Similarity data		
Logistic Reg 逻辑回归	K-Medoids k-Medoids聚类	Generalized Sequential Patterns	Mutual Information Matrix	Similarity data		

平台监控管理工具


DATAZOO Ambari demo
仪表板
服务
主机
警告
管理员
admin
HDFS
YARN
MapReduce2
Tez
Hive
HBase
Pig
Sqoop
ZooKeeper
Flume
Ambari Infra
Ambari Metrics
Kafka
Log Search
Spark
行动
指标
热图
配置 历史
指标行为
最近1小时
HDFS 磁盘使用情况
6%
运行中的DataNodes
3/3
HDFS链接
NameNode
Secondary NameNode
3 DataNodes
更多...
内存使用率
9.3 GB
网络使用率
39.0 KB
19.5 KB
CPU 使用
100%
50%
集群负载
5
NameNode堆
14%
NameNode RPC
0.38 ms
NameNode CPU WIO
0.1%
NameNode 正常运行时间
22.81 mins
Thu Sep 21 2017
13:14:08
HBase Master 堆
2%
HBase 链接
HBase Master
3 RegionServers
Master Web UI
更多...
平均加载时间
1
HBase Master 正常运行时间
2.9 d

数据挖掘工具

UEC GROUP LTD

RStudio

mlfit.R

```
1  library(car)
2  plot(USPop)
3  lmfit = lm(population ~ year, data = USPop)
4  lines(USPop$year, predict(lmfit))
5  nlmfit1 = lm(population ~ I(year^2)+year , data = USPop)
6  nlmfit1
7  summary(nlmfit1)
8  lines(USPop$year, predict(nlmfit1), col = 'red')
9
10
```

Environment History

Global Environment

Values

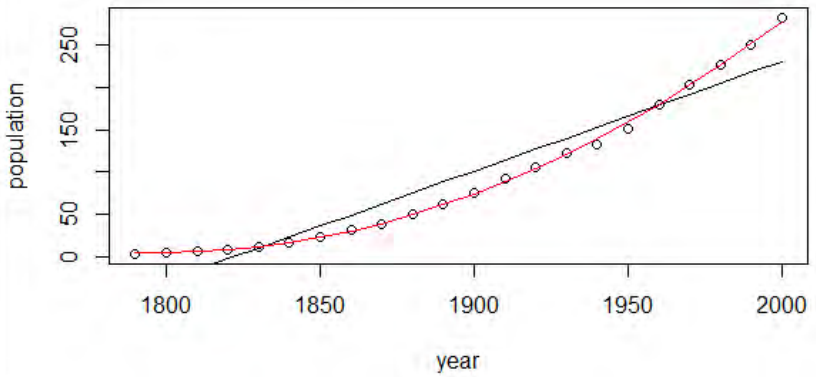
lmfit List of 12

nlmfit1 List of 12

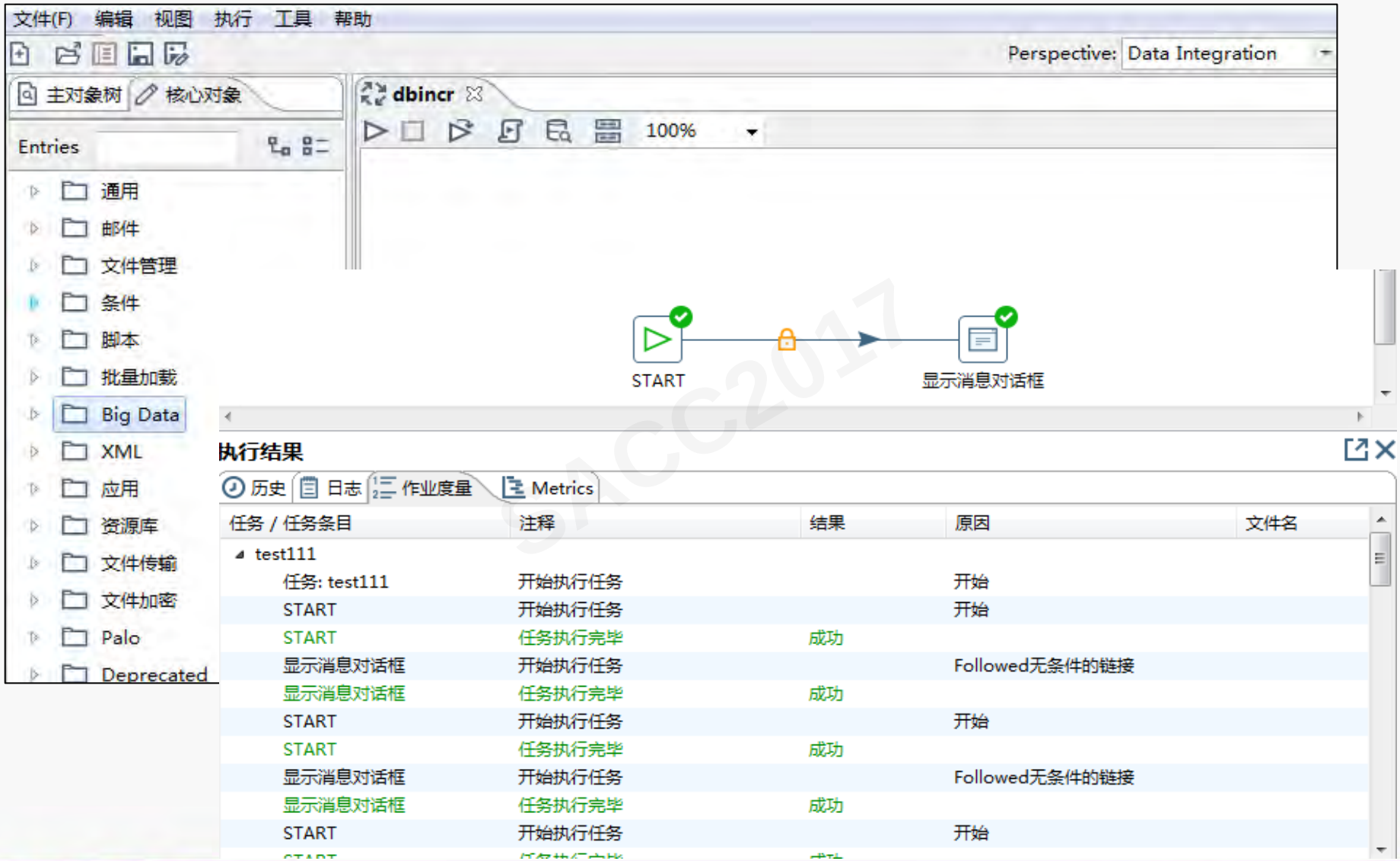
Console

```
  2.162e+04    6.681e-03   -2.403e+01

> summary(nlmfit1)

Call:
lm(formula = population ~ I(year^2) + year, data = USPop)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5557 -0.4308  0.6051  1.4230  4.6486

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.162e+04  6.389e+02   33.83   <2e-16 ***
I(year^2)    6.681e-03  1.780e-04   37.52   <2e-16 ***
year        -2.403e+01  6.749e-01  -35.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.997 on 19 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9988
F-statistic:  8892 on 2 and 19 DF,  p-value: < 2.2e-16

> lines(USPop$year, predict(nlmfit1), col = 'red')
>
```

Files Plots Packages Help Viewer



数据采集工具



数据可视化交互工具


图表设置
count
无
订单号
日期
顾客姓名
订单等级
订单数量
销售额
订单-运输方式
大卡
火车
空运
样式
数据
API：
Products Sales
获取
维度
订单号 日期 顾客姓名
订单等级 订单数量 销售额
折扣点 运输方式 利润
单价 运输成本 区域
省份 仓库 城市
产品类别 产品子类别
产品包装 运送日期
预览
编辑Dashboard，对接数据接口

荣之联
UEC GROUP LTD