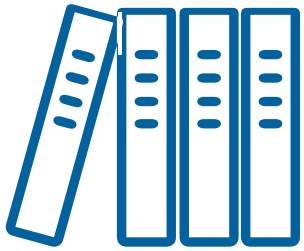
CCTC 2017 中国云计算技术大会
Cloud Computing Technology Conference 2017
CSDN

大数据核心技术在运营商的应用与实践

2017年5月19日



目录

CONTENT

运营商大数据架构演进方向分享

- 运营商原有大数据平台分为Iaas/Paas/Saas三层，未来运营商大数据平台将往更深层次方向演进，主要有如下六个方面的特征；

1、应用域：逐渐对传统的应用进行拆解解耦，实现应用微服务化；面向外部百花齐放的应用，逐渐开放应用能力；面向大数据应用变现，推进内外应用走向互联网化的运营方式；

2、数据域：即平台域中的数据平台域，从传统的采集、整合、服务转向数据资产化，数据资产化特征：数据资产治理、数据资产应用、数据资产经营；

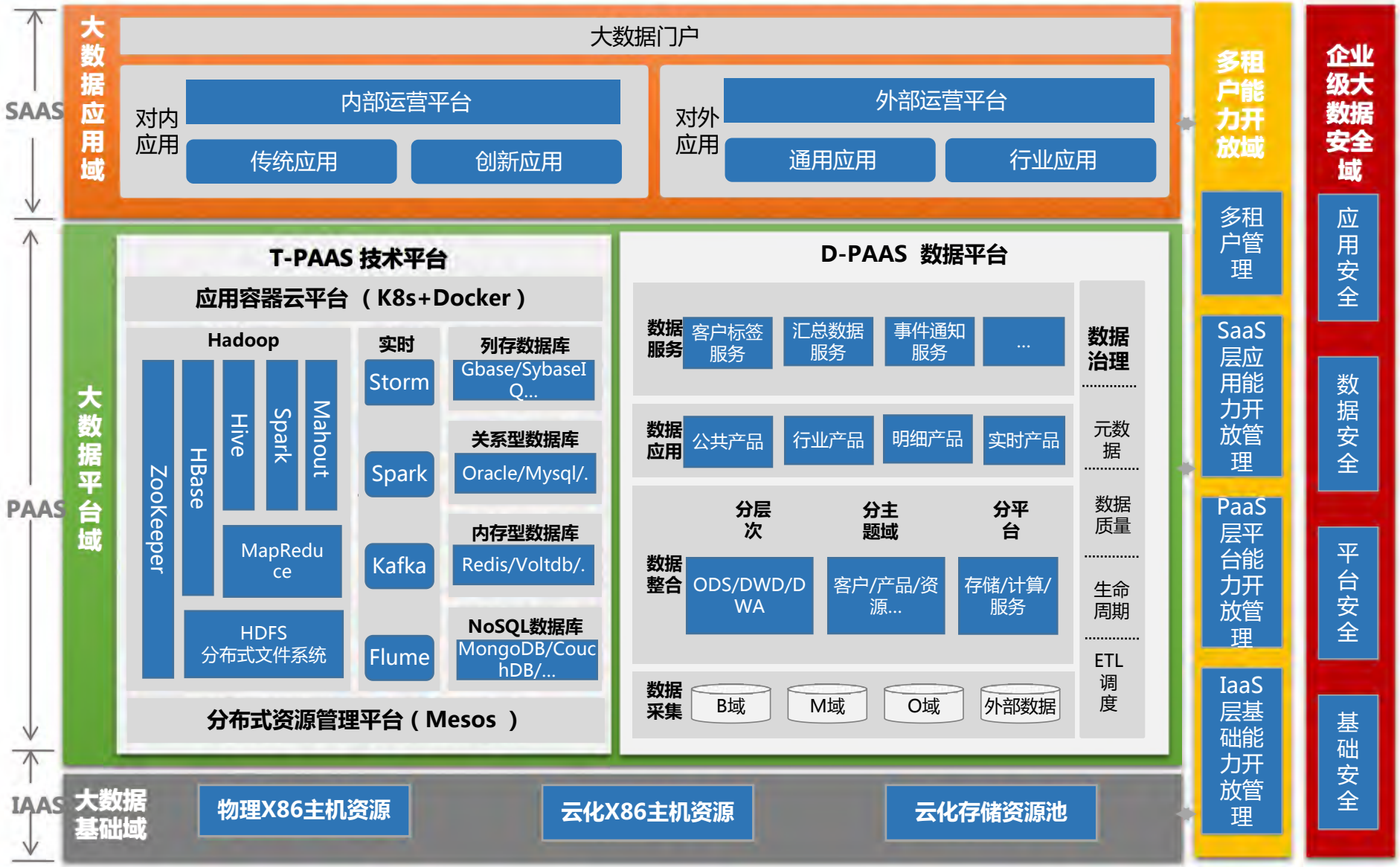
3、技术域：即平台域中的技术平台域，从Hadoop+ Oracle+实时流等的混搭架构逐渐演变为资源、应用、计算/存储的平台生态化；

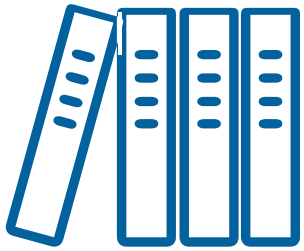
4、基础域：更进一步的去IOE化，X86及虚拟化基础设施更加弹性化；

5、安全域：从原来的数据安全走向企业级的大数据安全；

6、开放域：面向内部外部用户，在数据安全的基础上，提供逐层能力开放。

运营商大数据演进架构-LVL1



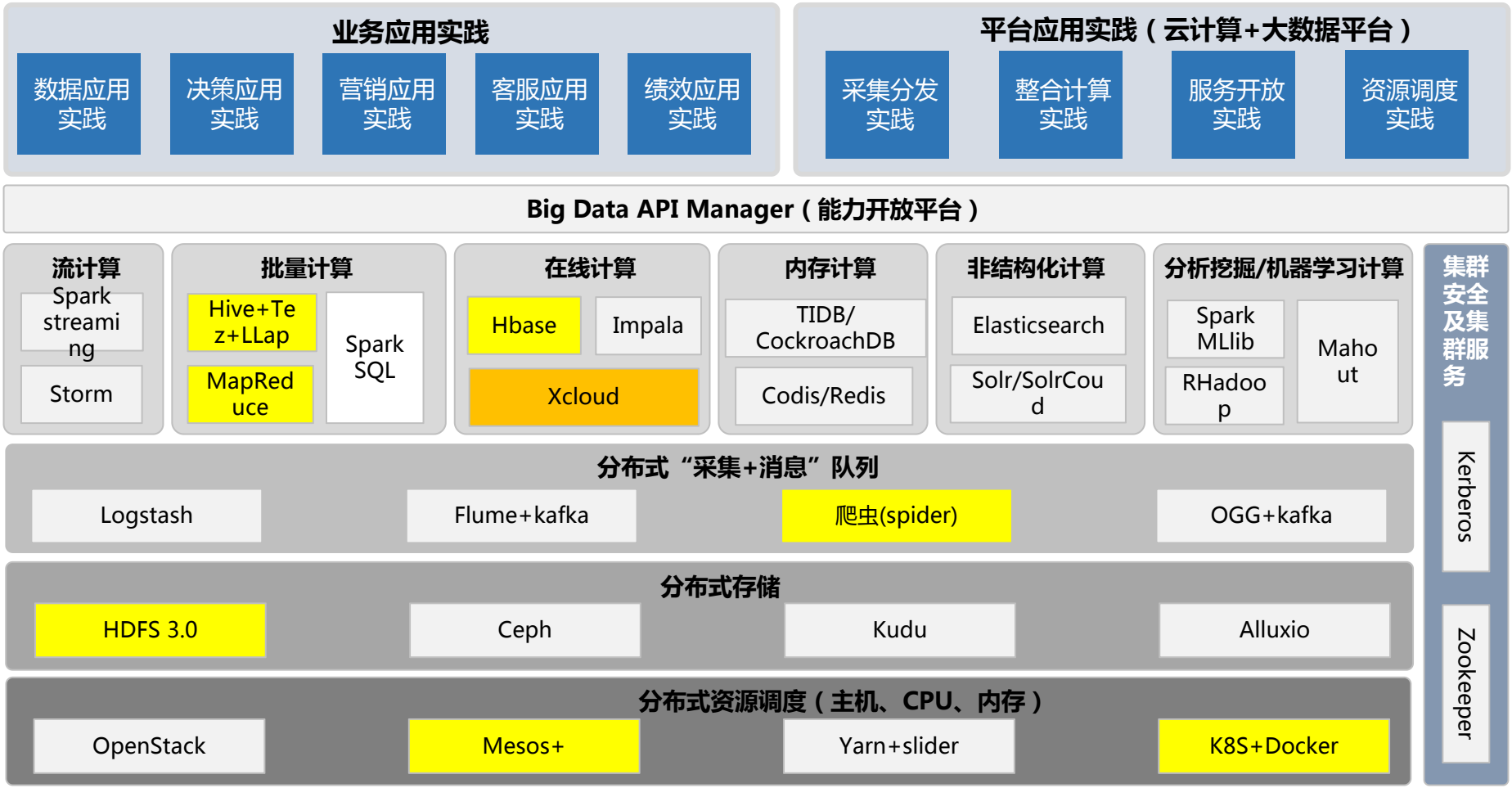


目录

CONTENT

构建 “开源+自主”大数据核心技术体系

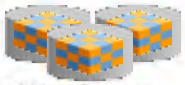
- 运营商经过多年的建设，当前技术主要基于“开源+自主”研发结合，利用大数据核心技术，构建面向业务应用和平台应用的实践；



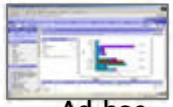
自主大数据核心技术-Xcloud

- XCloud是面向分析型应用领域，基于SQL on Hadoop，结合行列混合存储技术、大规模并行化计算技术、组合数据压缩算法及智能索引等技术构建的新型分布式数据库。

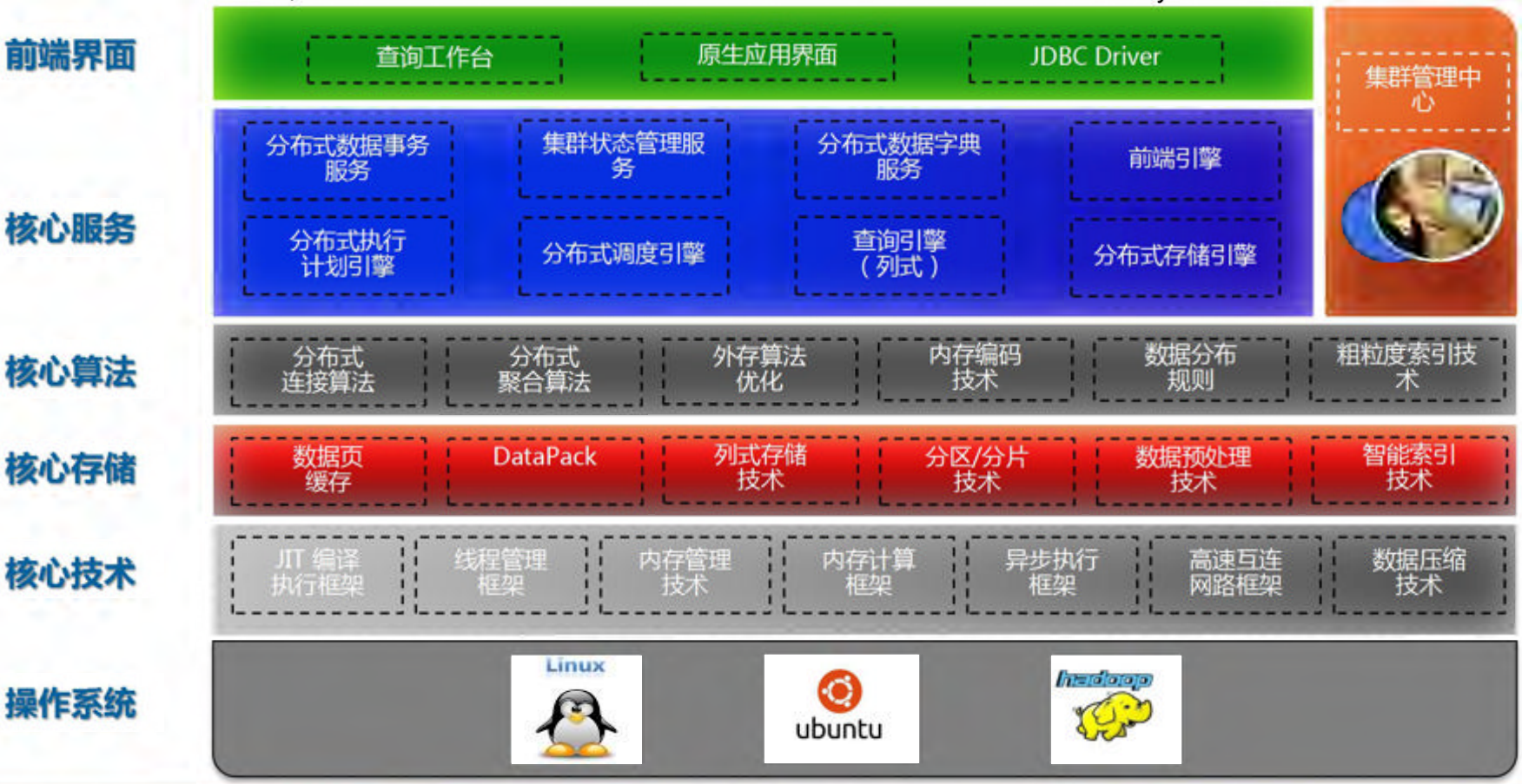
海量数据分析型应用



计算密集型分析应用

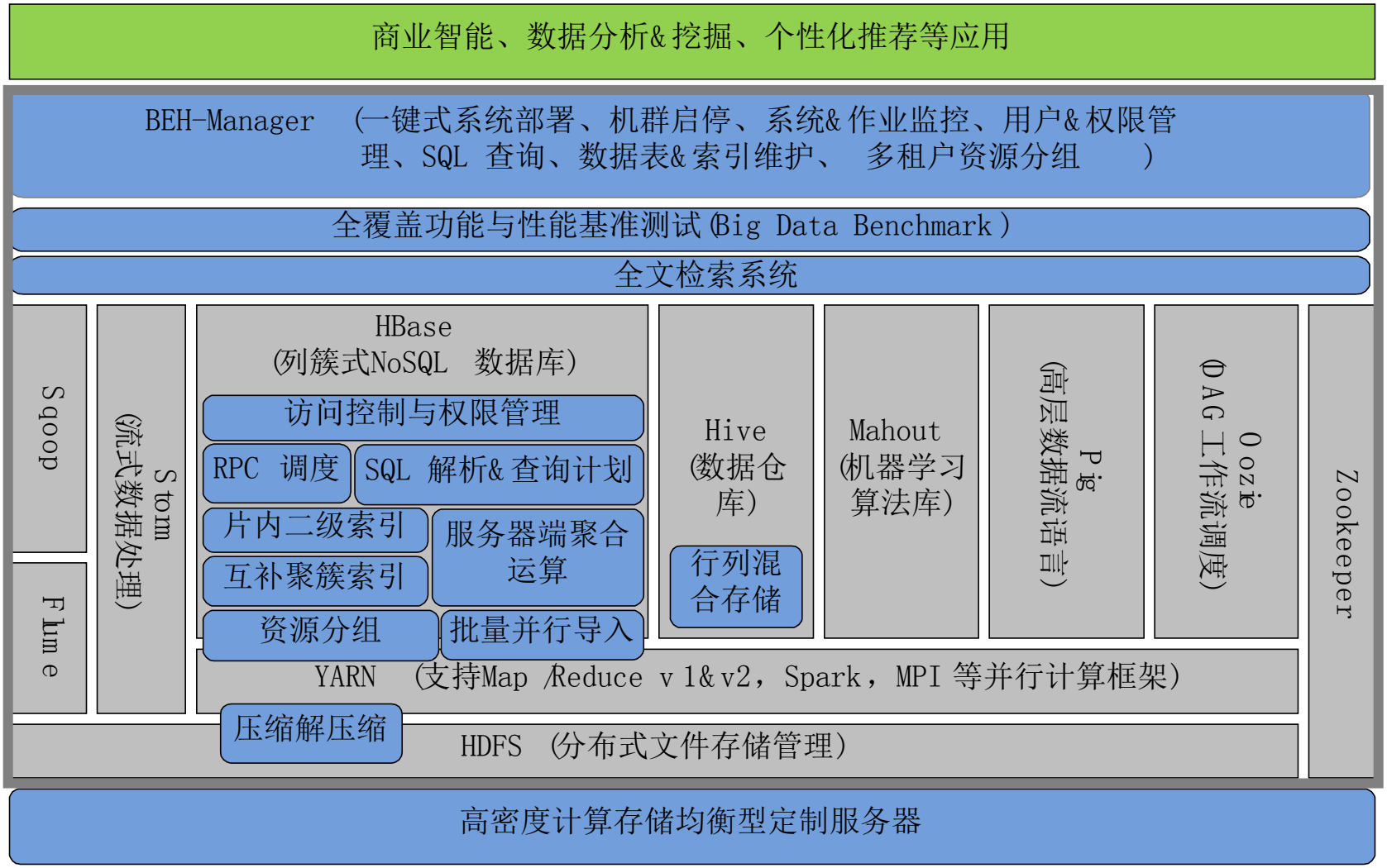


在线实时查询和分析



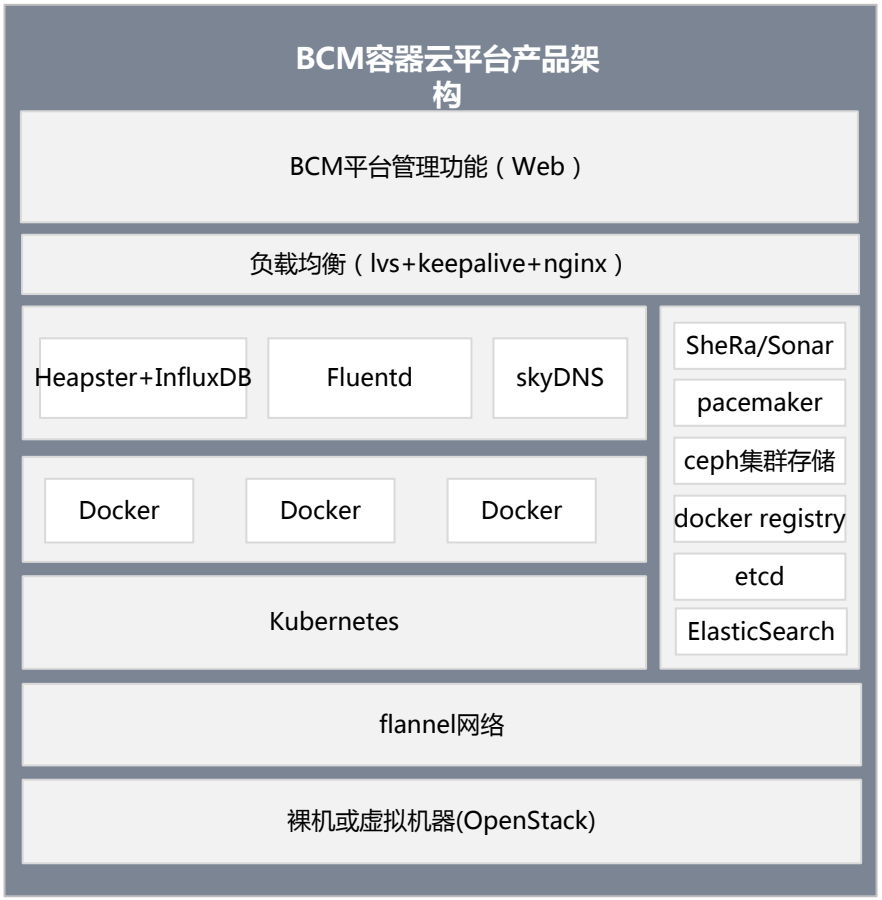
自主大数据核心技术-Hadoop BEH

□东方国信的Hadoop发行版本，是基于开源版本进行增强，兼容开源版本，能随着开源版本的升级而升级。



自主大数据核心技术-二次开发

- **基于开源社区源代码实现二次开发，转换为自己的核心技术，逐渐将应用与生产实践的验证部分代码提交给社区，比如： K8s+Docker底层源代码修订；**

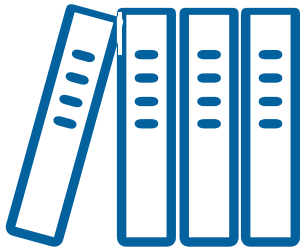


docker

- 改进docker的json-file格式日志的查询性能，tail容器日志行数较多，或者查询时指定since参数的情况下，官方提供的实现方法响应速度过慢，改进以后响应速度大大提高；
- 改进docker的json-file格式日志的查询方式，官方实现提供了对since参数的支持，我们添加了对until参数的支持，方便日志的查询；

kubernetes

- 改进kubelet里dns相关的代码逻辑，使其允许创建多个dns服务，便于实现k8s集群dns服务高可靠性部署方案；
- 添加kubelet新的参数，方便管理和配置docker日志文件的大小和个数，配合对docker日志功能的改进，方便更高效的查询管理容器日志；
- 添加独立的kubeng模块，实时监控服务的变化，和nginx一起实现服务的发现和访问代理；
- 改进kube-proxy里创建iptables的代码逻辑，使服务的外部访问到达一个node后，不再转发到其它node；
- 改进kubelet的代码逻辑，使cephfs可以关联到多个pod上；



目录

CONTENT

大数据核心技术在运营商业务应用实践

- **运营商的企业运营管理，围绕大数据为核心，面向客户和内部员工，实现企业业务运营和管理。业务应用实践包括数据、决策、营销、客户、绩效五个方面。**

数据应用实践

决策应用实践

营销应用实践

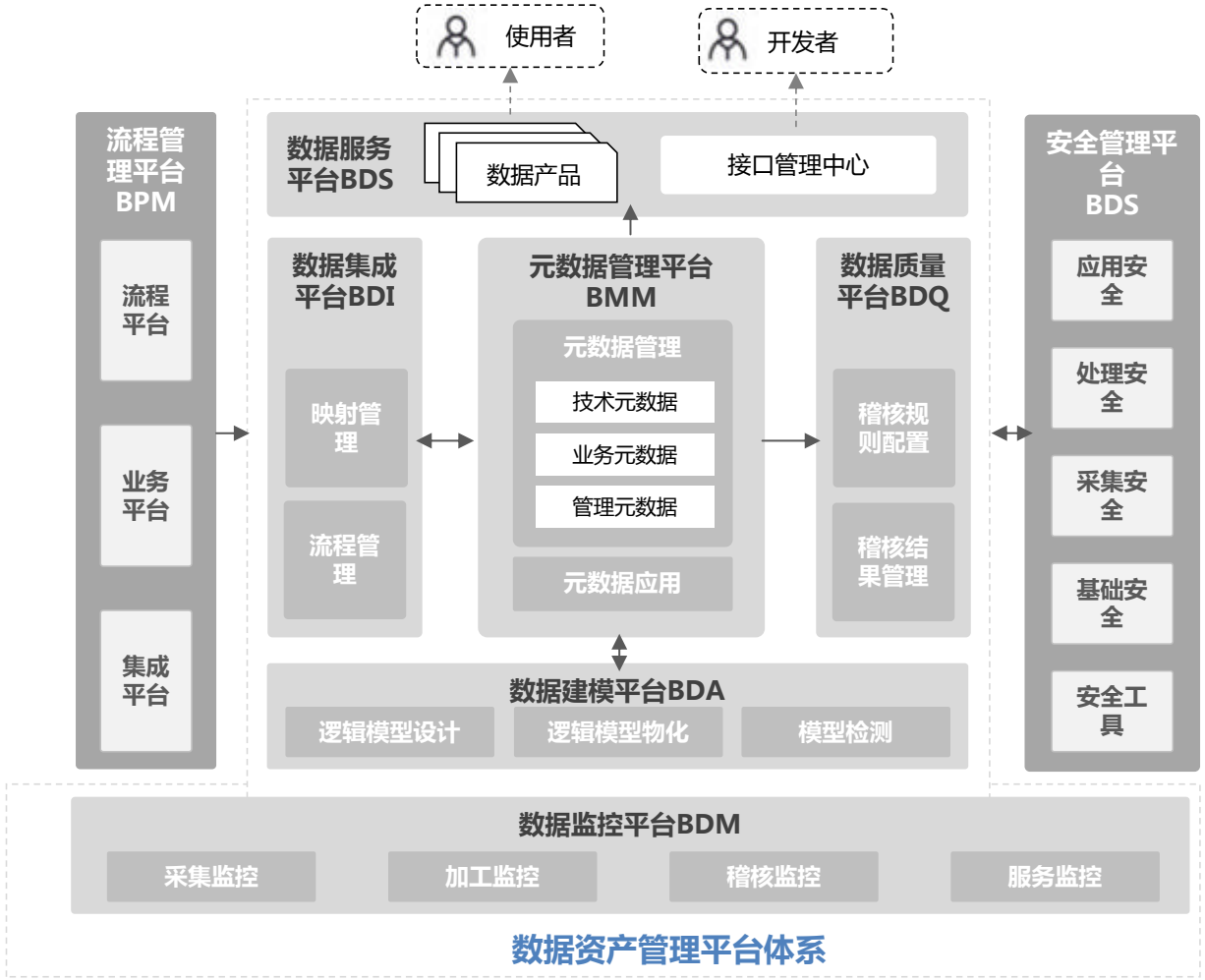
客服应用实践

绩效应用实践

- 数据应用：在生态化的大数据技术体系，以元数据为基础，实现异构数据管理，并构建统一的数据资产管理体系，实现数据资产统一采集、整合、服务、监控、安全等端到端的管理；
- 决策应用：基于大数据SparkR/R-Hadoop计算，结合数据检测、多维分析、离群检测等算法进行数据探索以及数据特征选取/聚类/关联，自动构建数据分类模型，为决策提供智能预警；
- 决策应用：基于大数据Xcloud海量列存技术，构建明细数据+数据定义+数据分析+组件接口的积木式数据微服务，针对数据实现深入下钻、多维、多面查看，为决策提供快速分析；
- 营销应用：基于大数据Xcloud+Strom/Spark Streaming技术，实现批量+实时结合的场景化营销；
- 客服应用：基于语音转文本/互联网爬取非结构化数据，实现互联网舆情和客户智能预判；
- 绩效应用：基于Spark SQL 小批量计算技术，实现准实时客户归属划配和客户绩效积分计算；

大数据核心技术-数据应用实践

- 构建大数据资产管理平台，实现数据全生命周期端到端透明化管控，实现“数据模型标准化、数据关系脉络化、数据加工可视化、数据质量度量化、数据服务自动化”，全业务流程的实时监控



数据资产管理平台体系

运营商规模最大的数据资产管理平台：已支撑：50197个数据模型，220746个元数据对象，日入库2870亿条数据，日稽核109988个任务，235个数据服务接口

紧耦合：

各产品相互联系，相辅相成，形成全面的数据资产质量一体化解决方案。

松耦合：

采用“组件化、微服务”的产品设计思想，每款产品可独立部署。

智能预警引擎—设计思路

- 智能引擎基于Hadoop/Spark主要分为数据探索、向导式数据建模和业务场景构建三大模块；支撑百万级用户的拖拉拽模式下多种挖掘算法；智能引擎可与SPARK连接应用于大数据分析；
- 与决策分析系统、智能预警等成熟系统衔接为一体构建决策分析类应用。

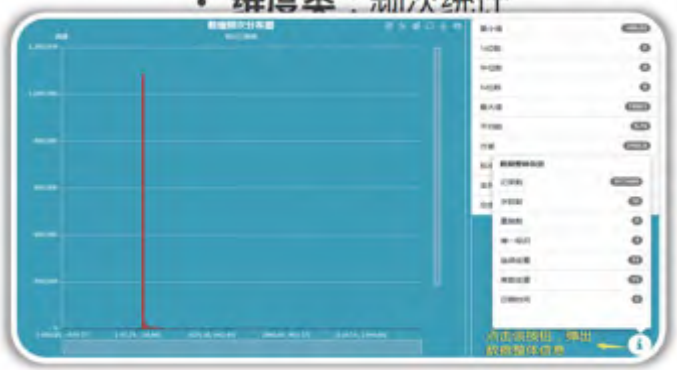
大数据核心技术-决策应用实践

智能预警引擎—数据探索

数据检测

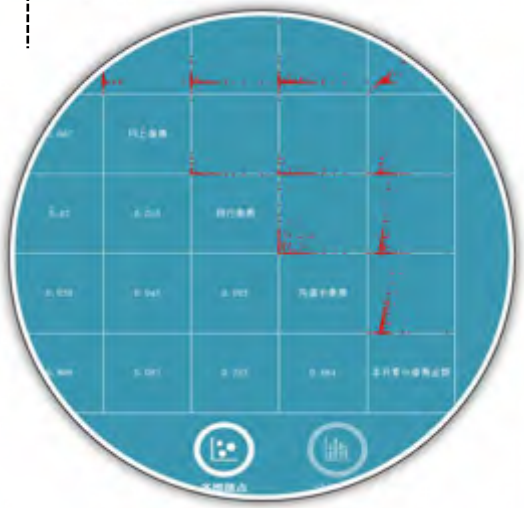
数据检测指单变量数学统计。

- **指标类**：最大/小值、四分位数、方差、平均值、变异系数等统计学统计结果
- **维度类**：频次统计



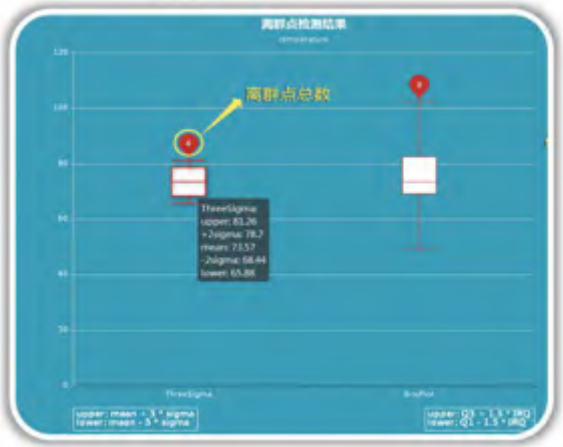
多维分析

- 多维分析是指对数据进行维度化分析后，完成多指标的散点矩阵图结果展示、多维度的分析、单维度多指标分析这三大功能。



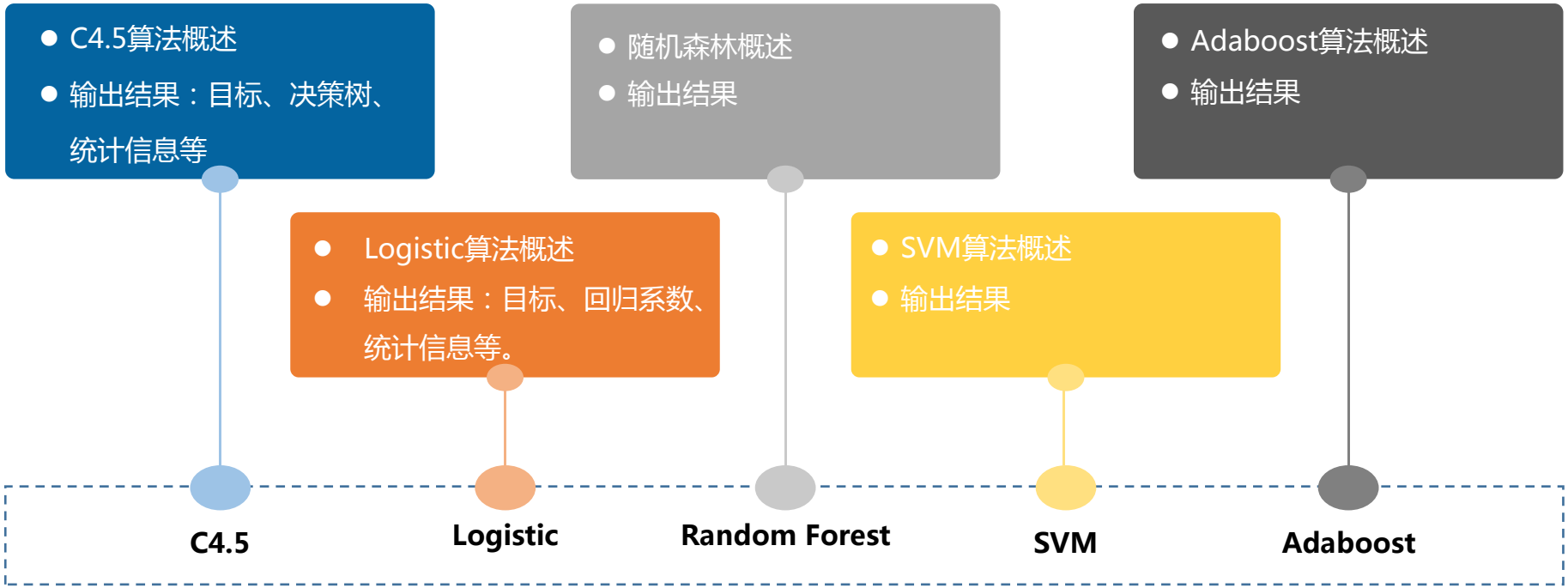
离群检测

- 用箱线图法和 3 Sigma 法对数据进行异常检测，将异常数据进行标记输出。



智能预警引擎—数据建模

- 智能引擎根据实际业务需要，提供分类建模，用于分类的算法有决策树、逻辑回归和随机森林等；
- 分类是一种有监督的机器学习，根据历史数据进行训练模型，然后根据其进行预测，最后将预测出的记录进行标记。



大数据核心技术-决策应用实践

智能预警引擎—场景应用

- 智能预警服务向下深钻挖掘分析一层，实现预警信息到预警用户群的聚焦，对聚焦的用户群进行多维分析，并将用户群与智能分析引擎衔接对预警用户进行深入聚类关联等挖掘探索。

- 常规多维分析+智能算法分析（聚类分析+关联分析）

- 提供预警信息用户级明细数据；

常规多维分析

智能算法分析

积木式数据微服务—设计思路

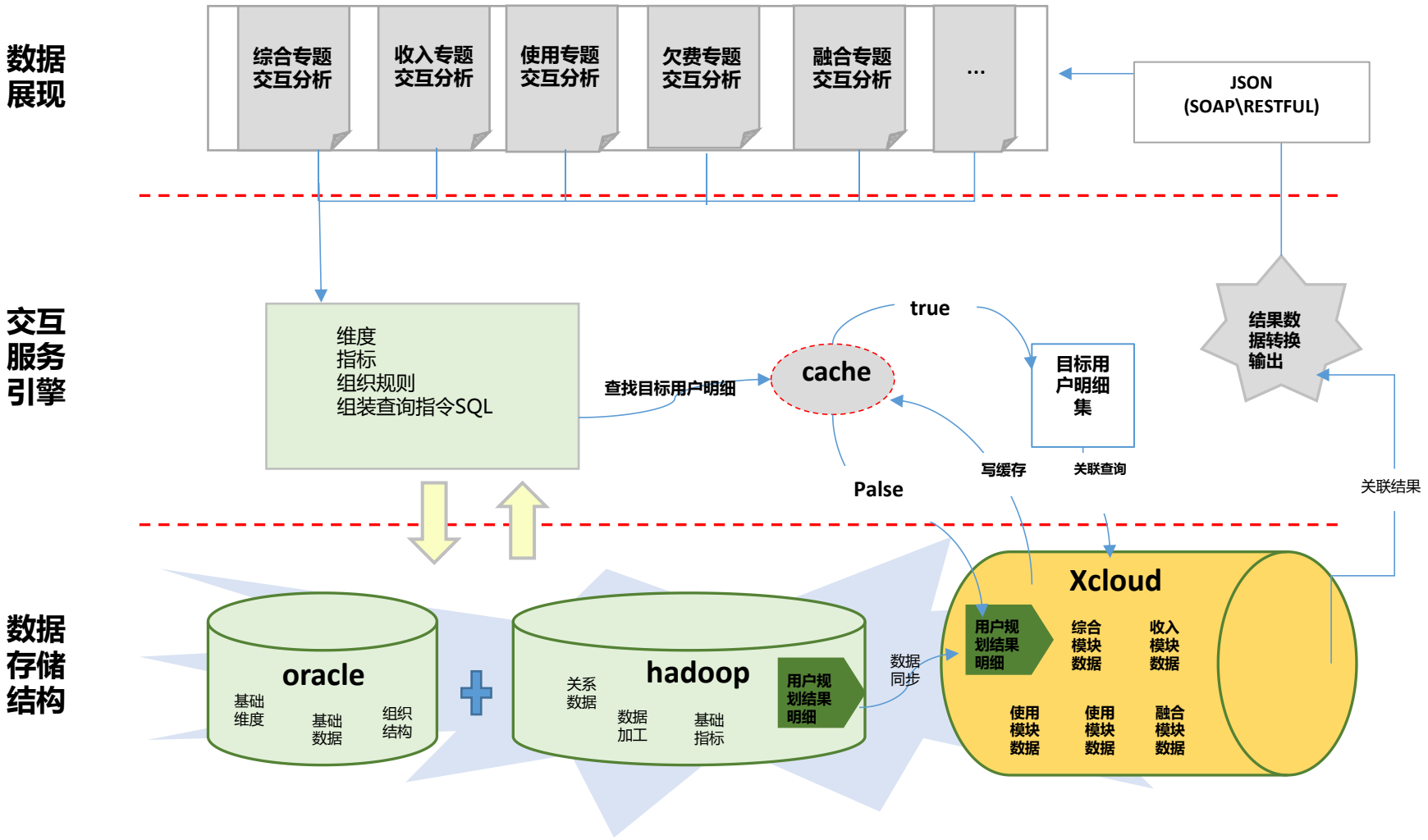
积木式组织机构

1. 让地市自行设计符合自己业务管理的组织架构
2. 提供规则引擎接口，让地市自行划分用户的归属区域
3. 对每个地市提供一套因子库，因子是划分用户的最小颗粒
4. 用户划分只需划分到最小组织架构上，向上汇聚就如同搭积木的概念，并且随着业务的变化可调整
5. 通过实时规则解析引擎，充分利用hadoop集群处理优势，将用户打标到每个组织机构节点上
6. 要能够支撑单个地市的机构或规则调整后数据的更新

交互式数据分析支撑

1. 由各种独立的服务或模块组成
2. 不仅支持单个微服务内任意组合、多维分析，还支持特定分析结果在其它服务的数据体现
3. 任意层级对标、任意维度对标、任意模块对标、页面任意指标用户群组合对标等
4. 后台依托混搭式大数据平台架构支撑
5. 引入交互式接口引擎技术，封装各种查询指令，利用大数据平台内存计算、缓存技术，提高页面查询效率

积木式微服务-系统架构


数据展现
综合专题交互分析
收入专题交互分析
使用专题交互分析
欠费专题交互分析
融合专题交互分析
...
JSON (SOAP\RESTFUL)
交互服务引擎
维度
指标
组织规则
组装查询指令SQL
查找目标用户明细
cache
true
Palse
目标用户明细集
写缓存
关联查询
结果数据转换输出
关联结果
数据存储结构
oracle
基础维度
基础数据
组织结构
hadoop
关系数据
数据加工
基础指标
用户规划结果明细
数据同步
Xcloud
用户规划结果明细
综合模块数据
收入模块数据
使用模块数据
使用模块数据
融合模块数据

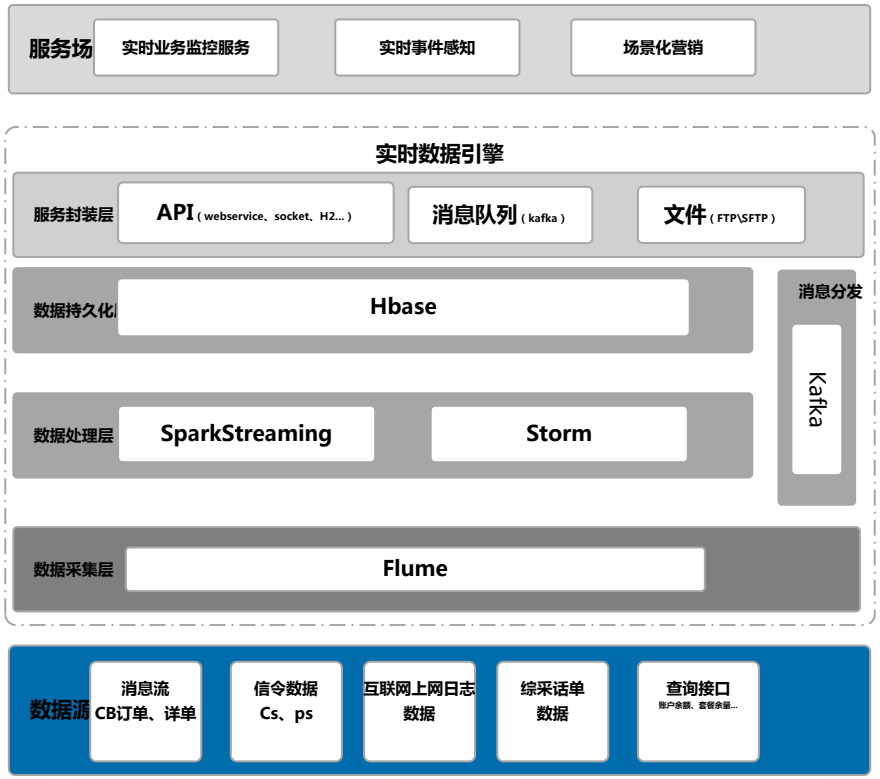
积木式数据微服务-场景应用

- **积木式微服务**=明细数据+数据定义+数据分析+组件接口
- **针对性深入下钻、多维、多面查看**：提供多种维度的灵活选择统计查询功能，针对于数据统计结果，进行关联查询，提供模块跳转下钻功能，使数据不光可以横向统计，还可以针对某一特定用户群纵向深入从多维、多面查看信息。

大数据核心技术-营销应用实践

- 采用strom/spark streaming等大数据技术，围绕用户全生命周期接触轨迹，通过对客户触点的价值挖掘，从超细分的客户实时行为事件捕获营销时机，并通过触点快速反应，实现基于位置、信令事件、客户接触、上网行为、异动特征等客户事件触发的**场景化实时营销能力**。

- **事件管理**：构建事件管理模块，使用大数据技术进行实时化精细处理，生成营销事件。通过独立的事件管理模块实现事件的定义、数据源管理、规则管理等，方便事件的灵活扩展和多场景、多活动共用。
- **场景管理**：构建“特定用户群+特定触点+特定事件”的用户触发式场景管理，支撑三级触点的触发式场景营销。

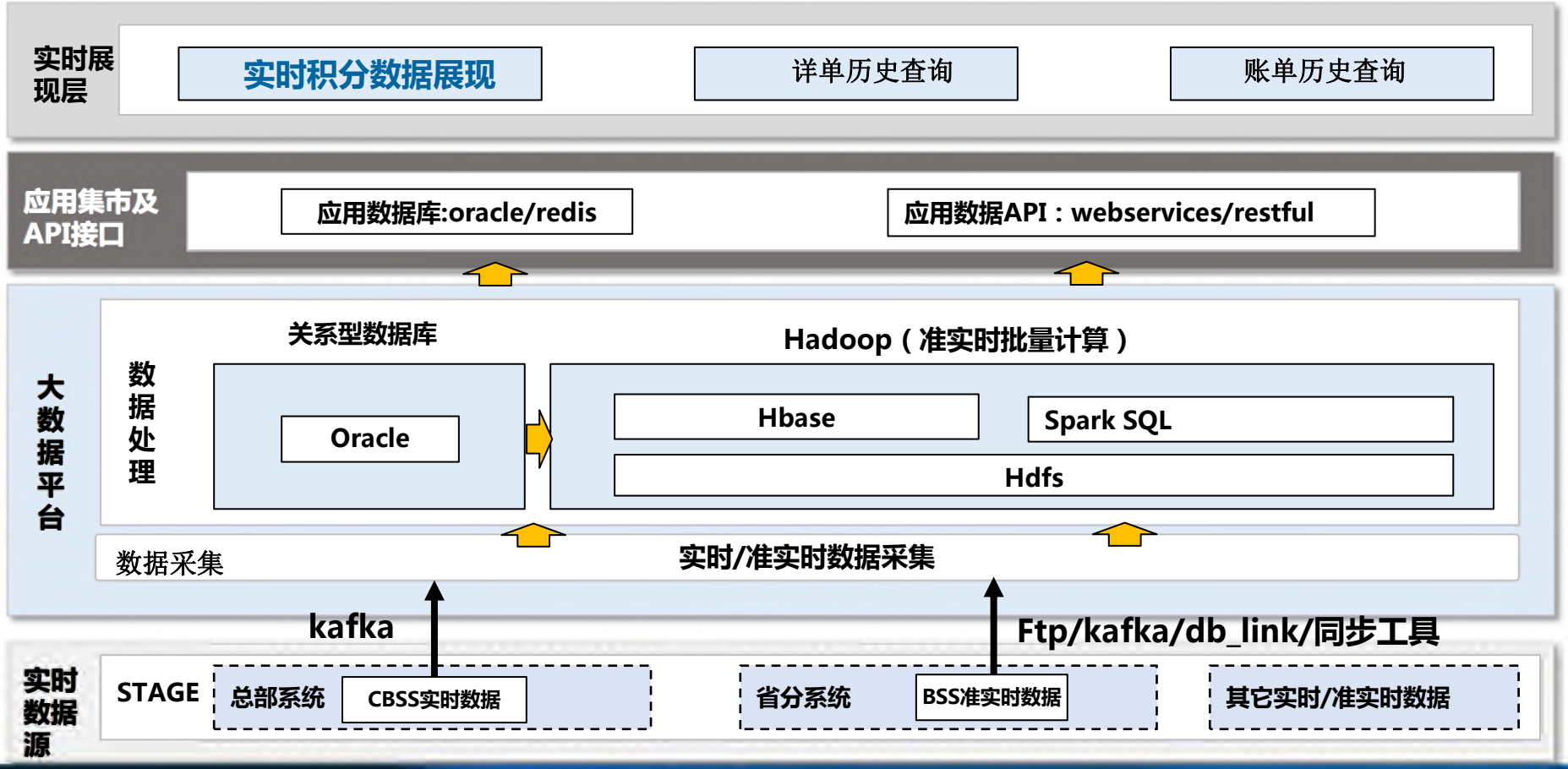


大数据核心技术-客服应用实践

- 采用爬虫和文本处理等大数据技术，整合互联网数据和DPI数据，面向客户提供服务应用。
  比如：基于语音转文本的客户智能投诉预判；

大数据核心技术-绩效应用实践

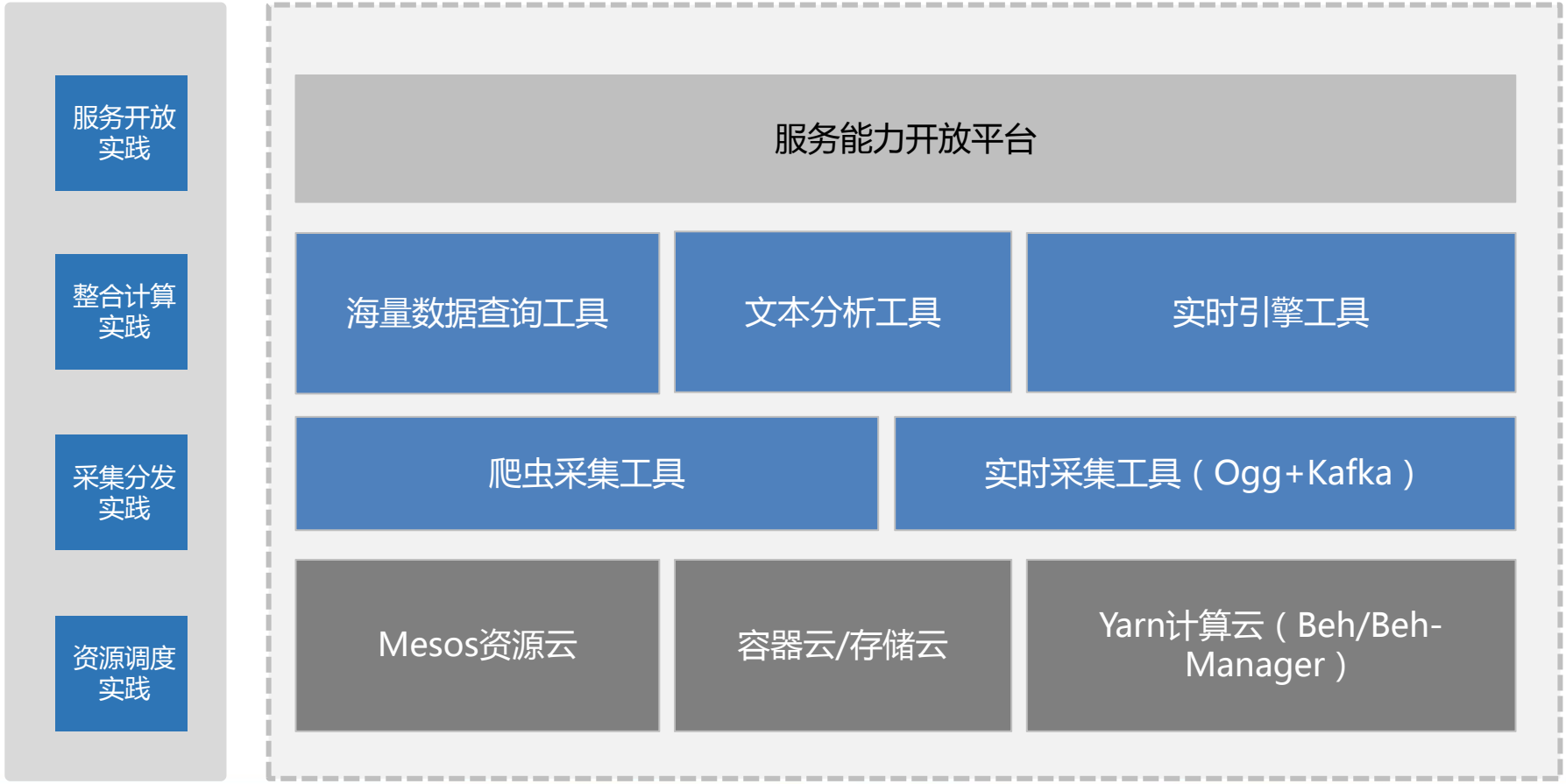
- 绩效应用项目中有大量需要实时计算考核的KPI和准实时查询的明细数据，比如实时积分计算，准实时详单查询；
- 该应用从关系型数据库中通过ogg+kafka准实时采集，通过oracle还原和Spark SQL准实时计算，Hbase做海量数据点查询，实现准实时明细数据和统计数据查询展现。



大数据核心技术在平台产品实践

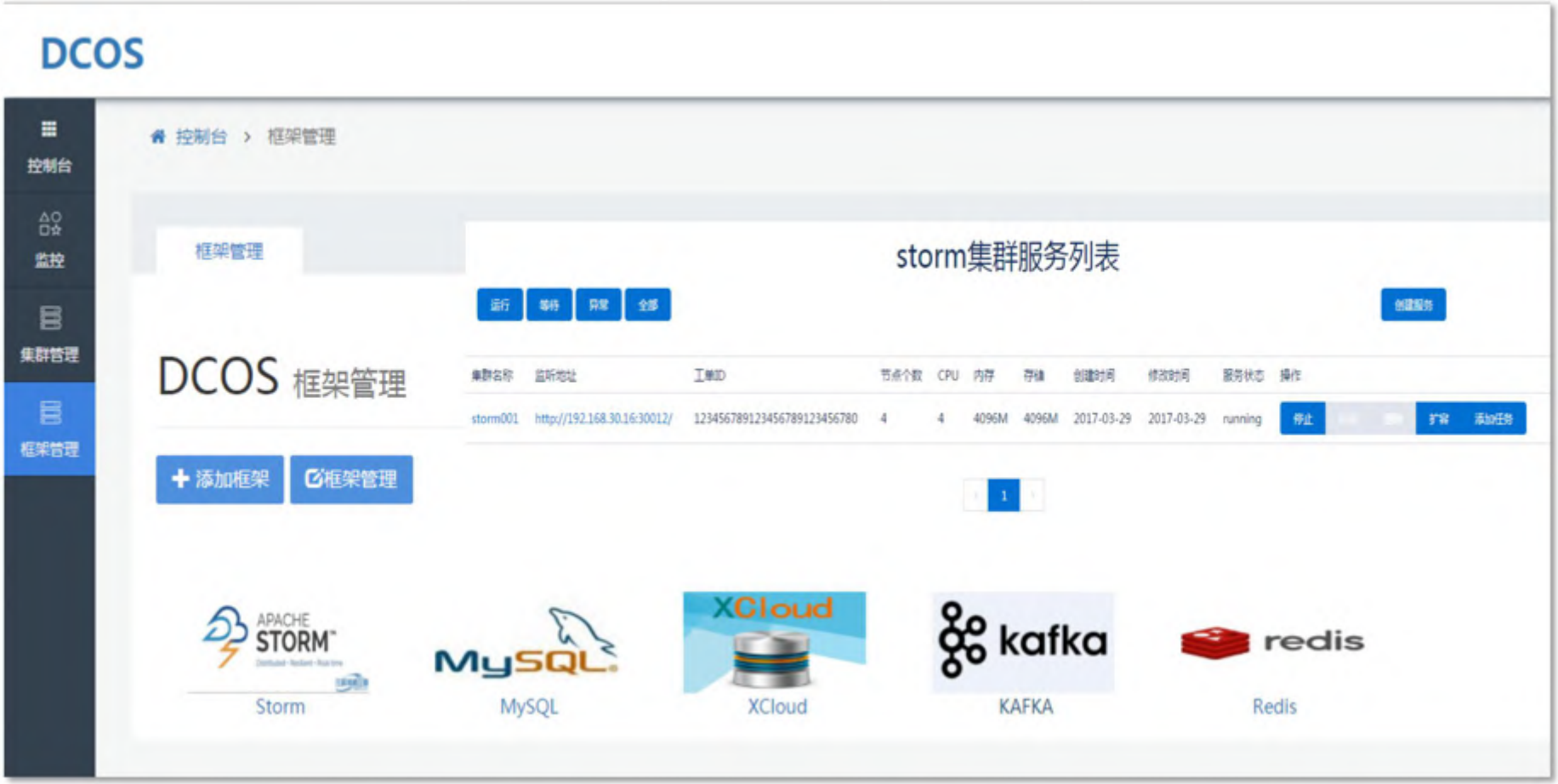
- 以大数据核心技术为基础，面向开发和使用场景，进行二次封装，持续构建从采集消息、存储、计算、安全、开放等方面可视化的平台产品及应用工具。

平台应用实践



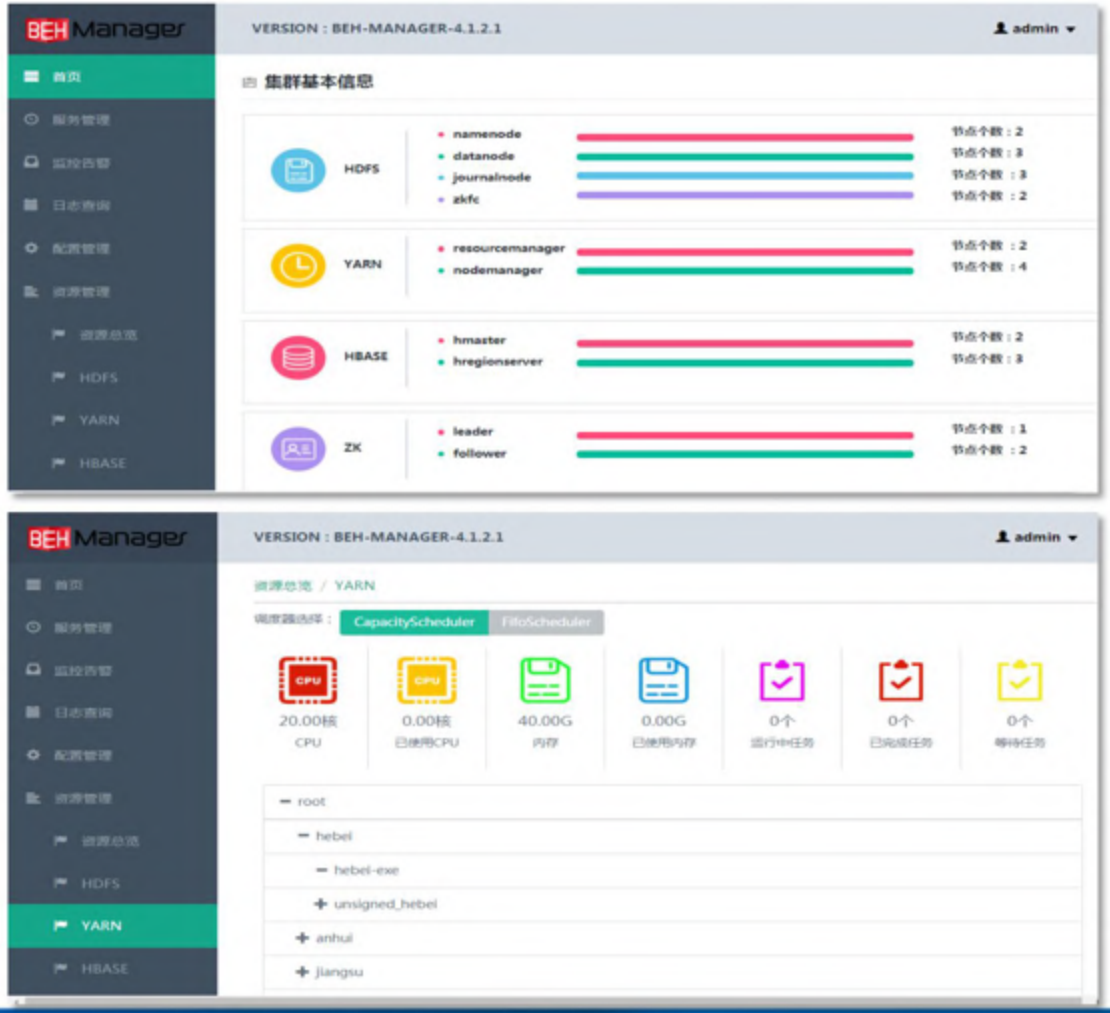
平台应用工具：DC/OS资源云

- 随着运营商去IOE化深入，需要管理越来越多的X86资源，如何弹性的使用资源和快速运维部署，对外提供多种计算能力的服务，运营商逐渐使用mesos进行资源统一管理；



平台应用工具：Yarn计算云

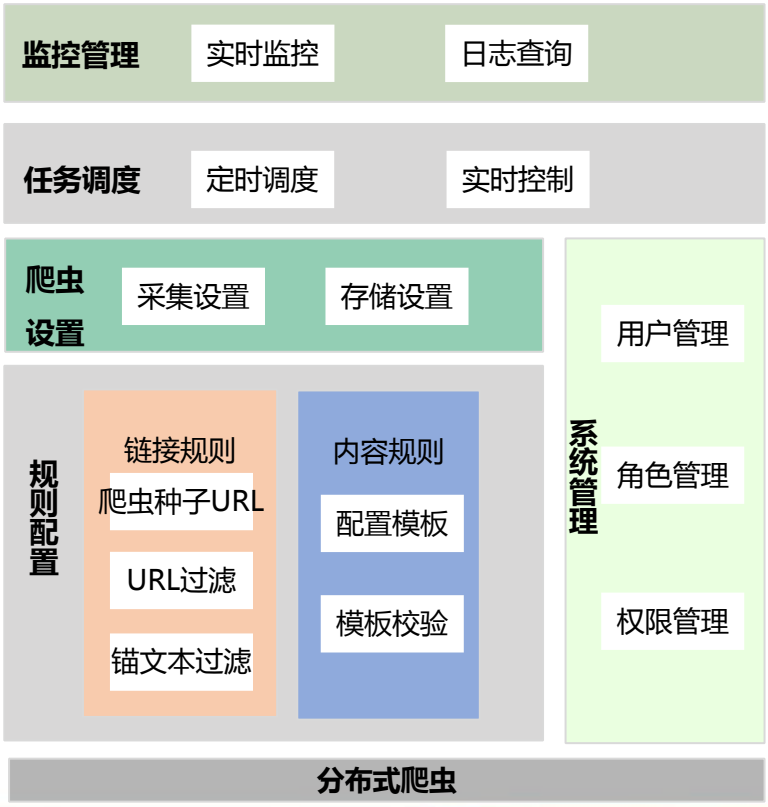
- 使用BEH可以帮助企业快速的搭建大数据平台，实现Hadoop平台的快速部署、组件管理、监控告警、多租户管理、配置管理、安全管理、开放管理，同时帮助企业减少开发和后期维护成本。



平台应用工具：互联网爬虫

- 通过灵活的配置可高效实时的对目标网站进行监测、采集，并从中提取链接、标题、时间、正文等，提高了信息采集速度并扩大了信息采集的规模。结构化数据，数据可选择存储到文件系统及各类关系型数据库中。

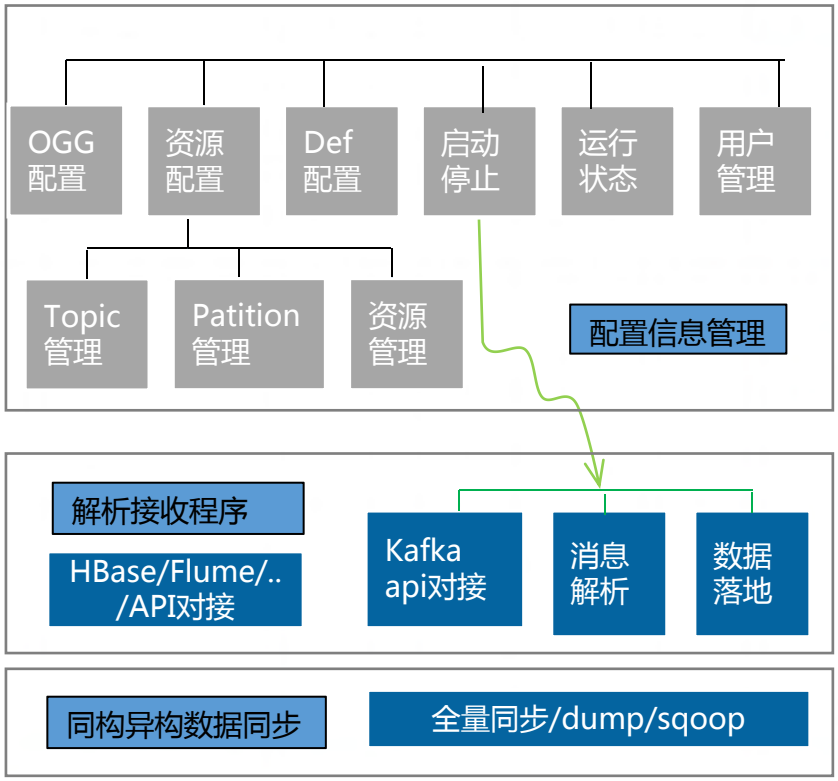
互联网爬虫功能架构



平台应用工具： OGG+Kafka实现增量和全量数据采集

- 实现关系型数据库准实时采集及落地，生产系统采用Ogg+Kafka的增量消息方案同步业务数据，把DML语句INSERT,DELETE,UPATE发布到Kafka集群中，外围系统根据需求自行解析接收数据。在整个流程中，外围系统不需要部署Kafka集群，解析接收生产侧Kafka集群消费即可。

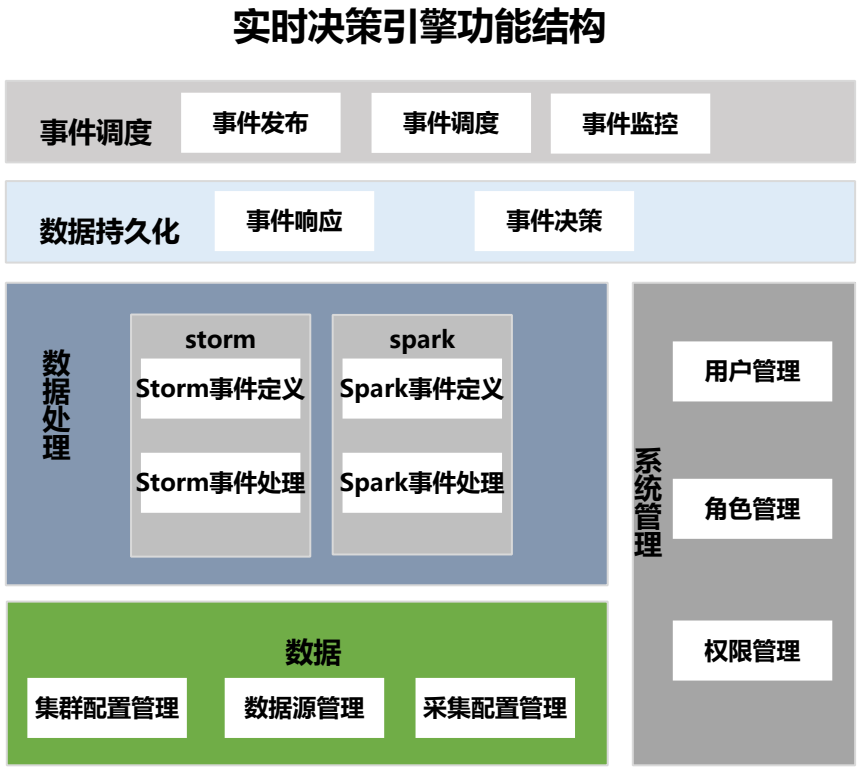
OGG+Kafka功能架构



平台应用工具：实时决策引擎

- 实时决策引擎后台组件集成了storm、spark streaming、kafka、flume、hbase等现阶段比较流行并且版本趋于稳定的实时流计算技术组件，通过将数据采集、事件定义、数据处理、事件调度等流程的界面可视化配置管理，用户可以非常方便的发布实时流场景应用程序，大大降低了用户的使用难度，提高了实时流应用程序的开发效率。

实时决策引擎功能结构

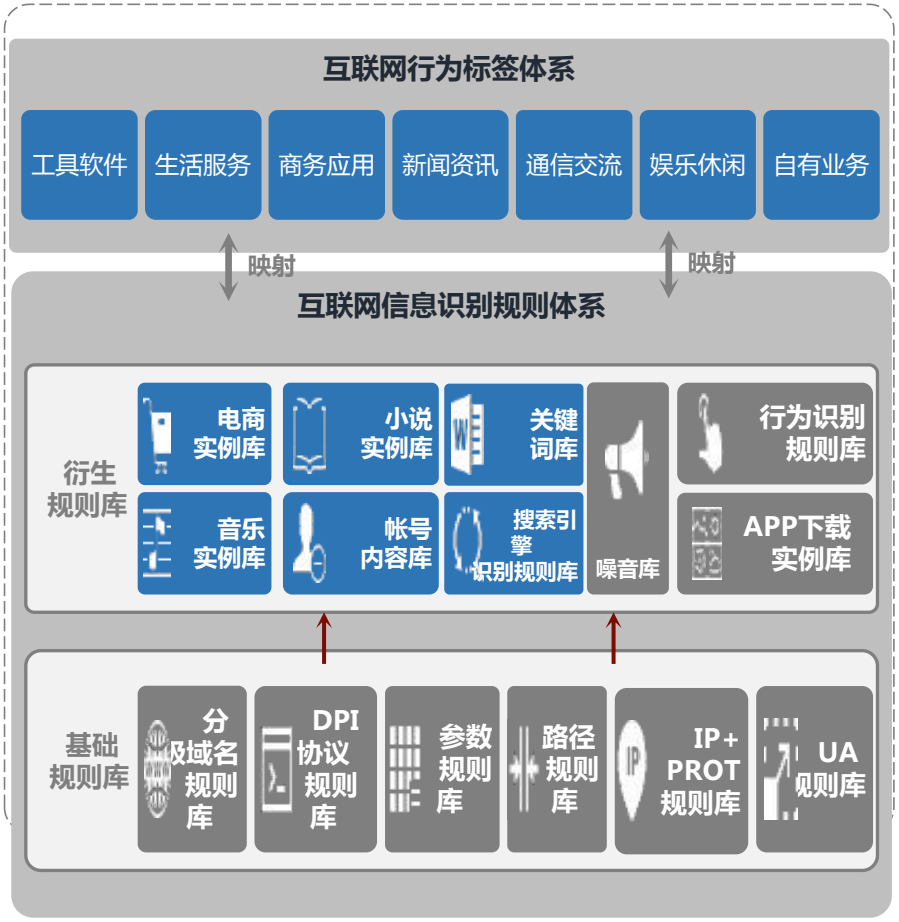


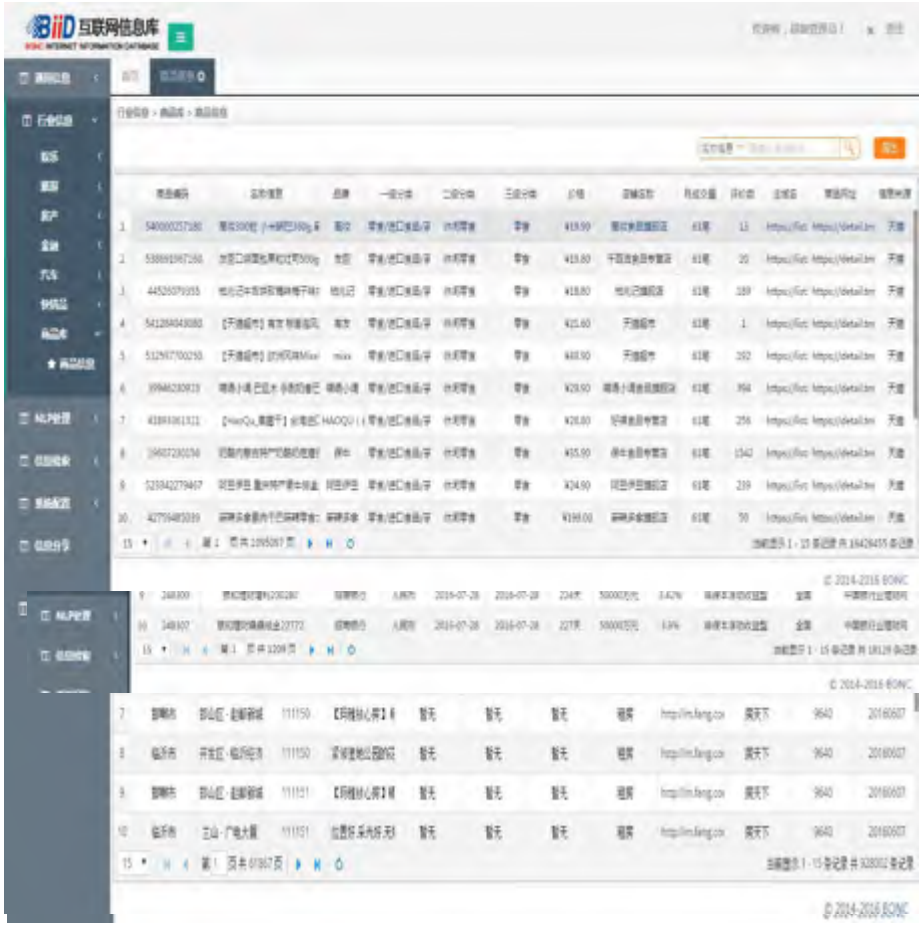
平台应用工具：文本分析工具

- 基于word2vec开源技术，主要以半结构或非结构的文本为处理对象，对大规模互联网文章或用户文本数据进行分析，提取出文本特征，然后采用各种文本挖掘方法对特征进行分析挖掘，以结构化和用户易于理解形式输出，指导用户更快、更好的利用数据。

平台应用工具：互联网信息工具

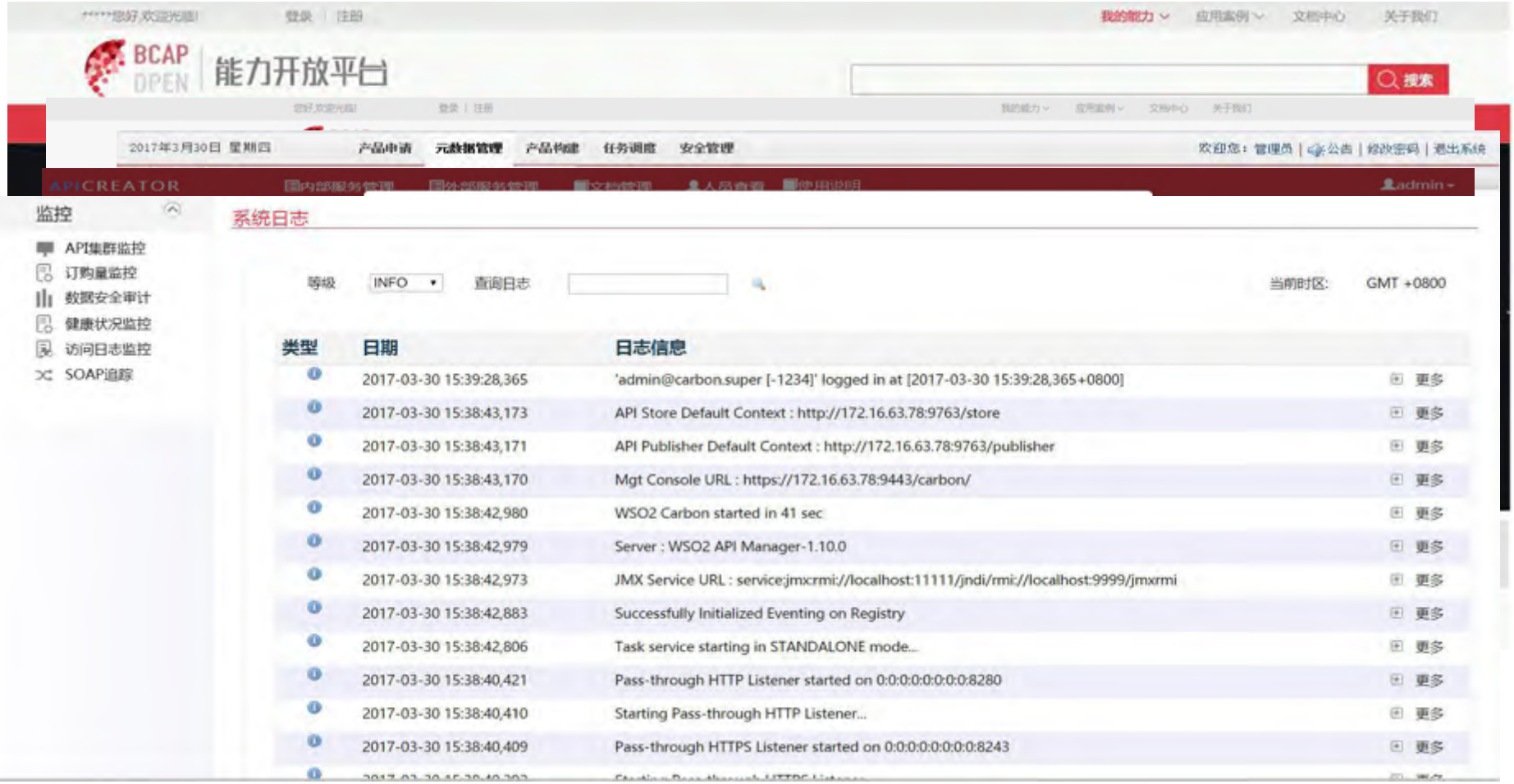
- 移动互联网的高速发展加速了通信网络整体IP化进程，数据流量蕴含的价值成为运营商收入提升新的着力点，为了助力联通走出流量贬值陷阱，走向流量内容经营，构建互联网信息库；





平台应用工具： API服务能力开放平台

- 借助基于SOA架构的WOS2开源技术，实现对数据、业务、应用、消息等原子资源的API封装，通过WOS2平台二次开发实现能力对外开放；



大数据核心技术运营商未来发展展望

- 实时能力：由于数据采集和处理更加实时，对实时数据应用能力提出更高要求；
- 在线能力：要求数据同时具备OLAP/OLTP能力，并且保证数据一致性，对基于明细数据的在线计算能力越来越高；
- 学习能力：从模型和算法的分析挖掘逐渐向半机器/机器学习演进；
- 开放能力：要求从资源到数据逐层对用户开放，对外服务高并发要求和安全性要求越来越多；
- 数据能力：数据管理越来越资产化，逐渐演进到软件驱动数据生产或者软件定义数据生产，提供全工具化的数据生产能力；
- 应用能力：内部应用更精细化和实时化，外部应用以运营商开放数据为基础，面向行业的跨行业应用；

谢谢大家！