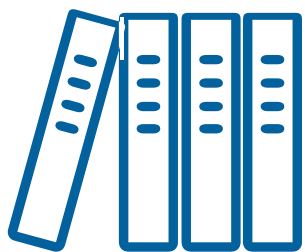


大数据核心技术 在运营商的应用与实践

2017年5月19日



目录

CONTENT



运营商的大数据架构

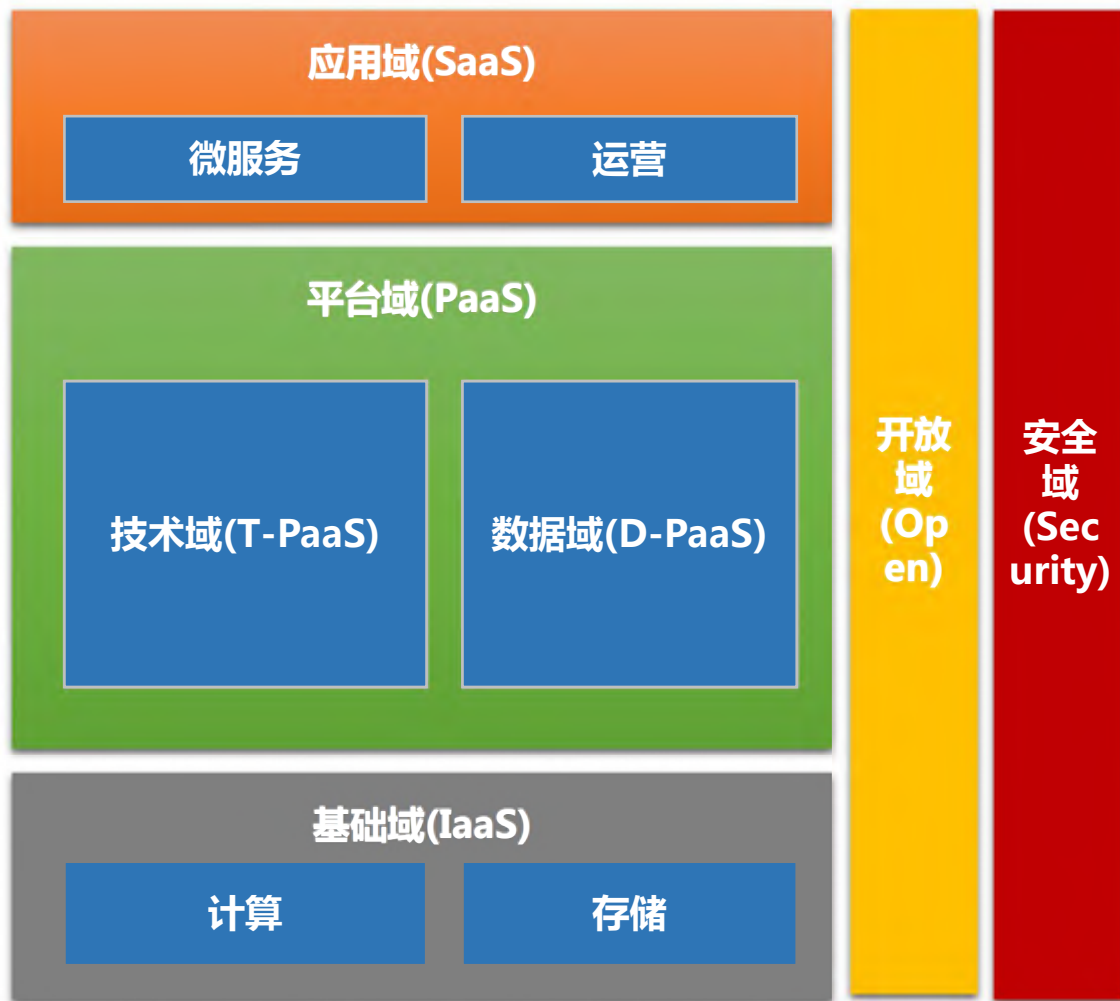


大数据核心技术体系

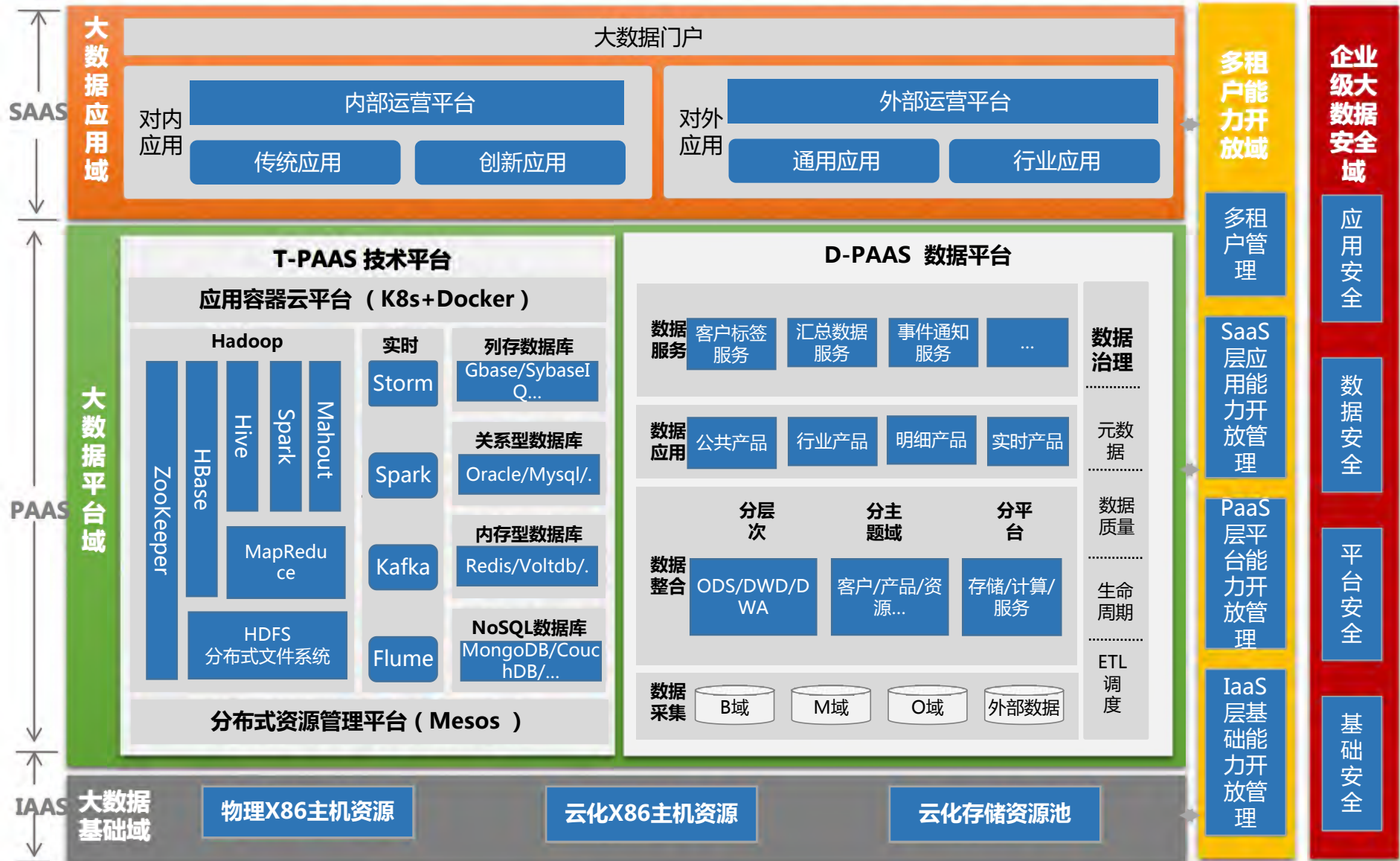


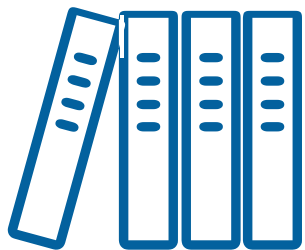
运营商的应用与实践

- 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；



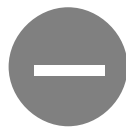
- 1、应用域：**逐渐对传统的应用进行拆解解耦，实现应用微服务化；面向外部百花齐放的应用，逐渐开放应用能力；面向大数据应用变现，推进内外应用走向互联网化的运营方式；
- 2、数据域：**即平台域中的数据平台域，从传统的采集、整合、服务转向数据资产化，数据资产化特征：数据资产治理、数据资产应用、数据资产经营；
- 3、技术域：**即平台域中的技术平台域，从Hadoop+ Oracle+实时流等的混搭架构逐渐演变为资源、应用、计算/存储的平台生态化；
- 4、基础域：**更进一步的去IOE化，X86及虚拟化基础设施更加弹性化；
- 5、安全域：**从原来的数据安全走向企业级的大数据安全；
- 6、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。





目录

CONTENT



运营商的大数据架构

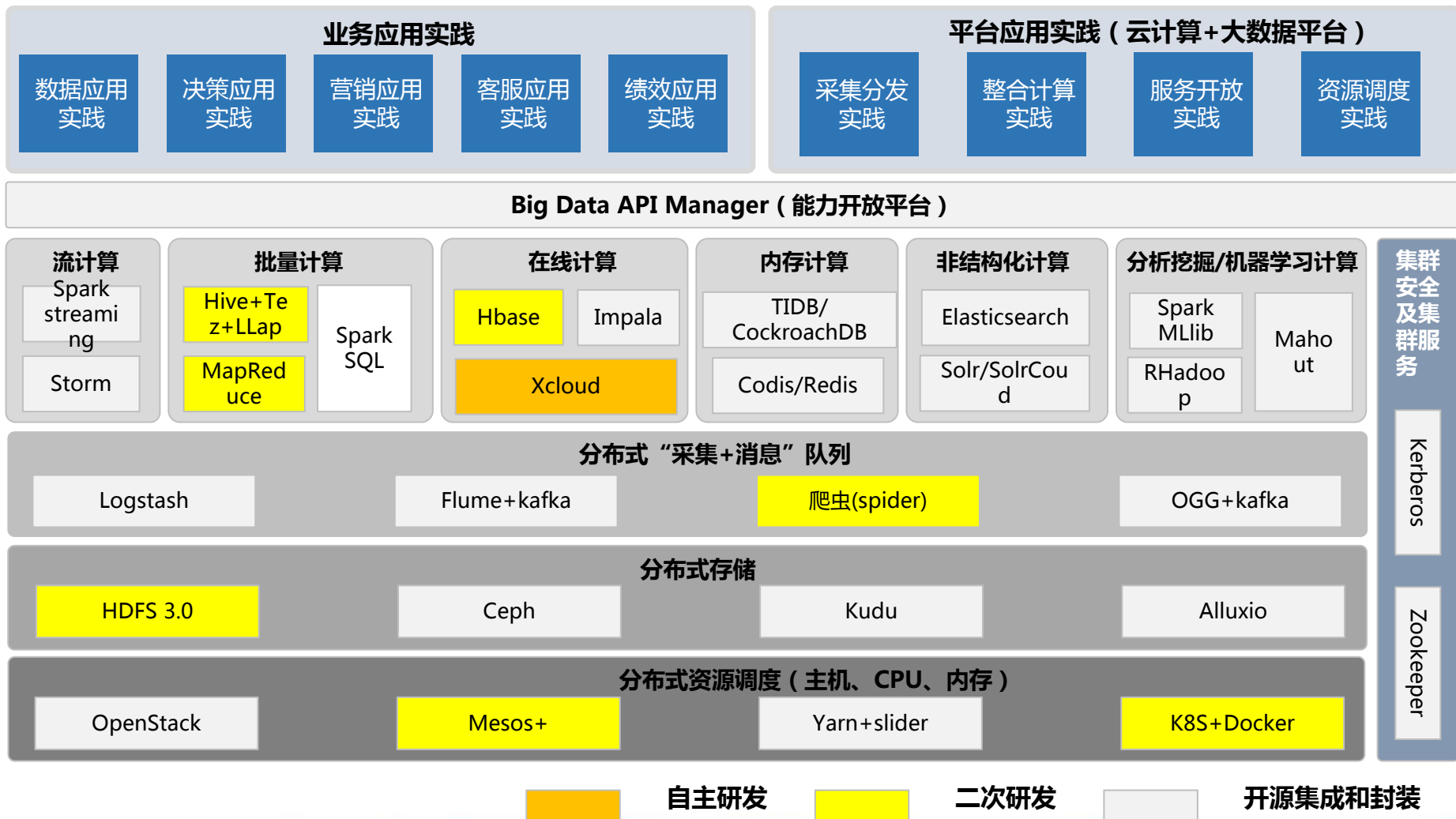


大数据核心技术体系



运营商的应用与实践

- 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；



- 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。



前端界面

查询工作台

原生应用界面

JDBC Driver

集群管理中心

核心服务

分布式数据事务服务

集群状态管理服务

分布式数据字典服务

前端引擎

分布式执行计划引擎

分布式调度引擎

查询引擎(列式)

分布式存储引擎

核心算法

分布式连接算法

分布式聚合算法

外存算法优化

内存编码技术

数据分布规则

粗粒度索引技术

核心存储

数据页缓存

DataPack

列式存储技术

分区/分片技术

数据预处理技术

智能索引技术

核心技术

JIT 编译执行框架

线程管理框架

内存管理技术

内存计算框架

异步执行框架

高速互连网络框架

数据压缩技术

操作系统

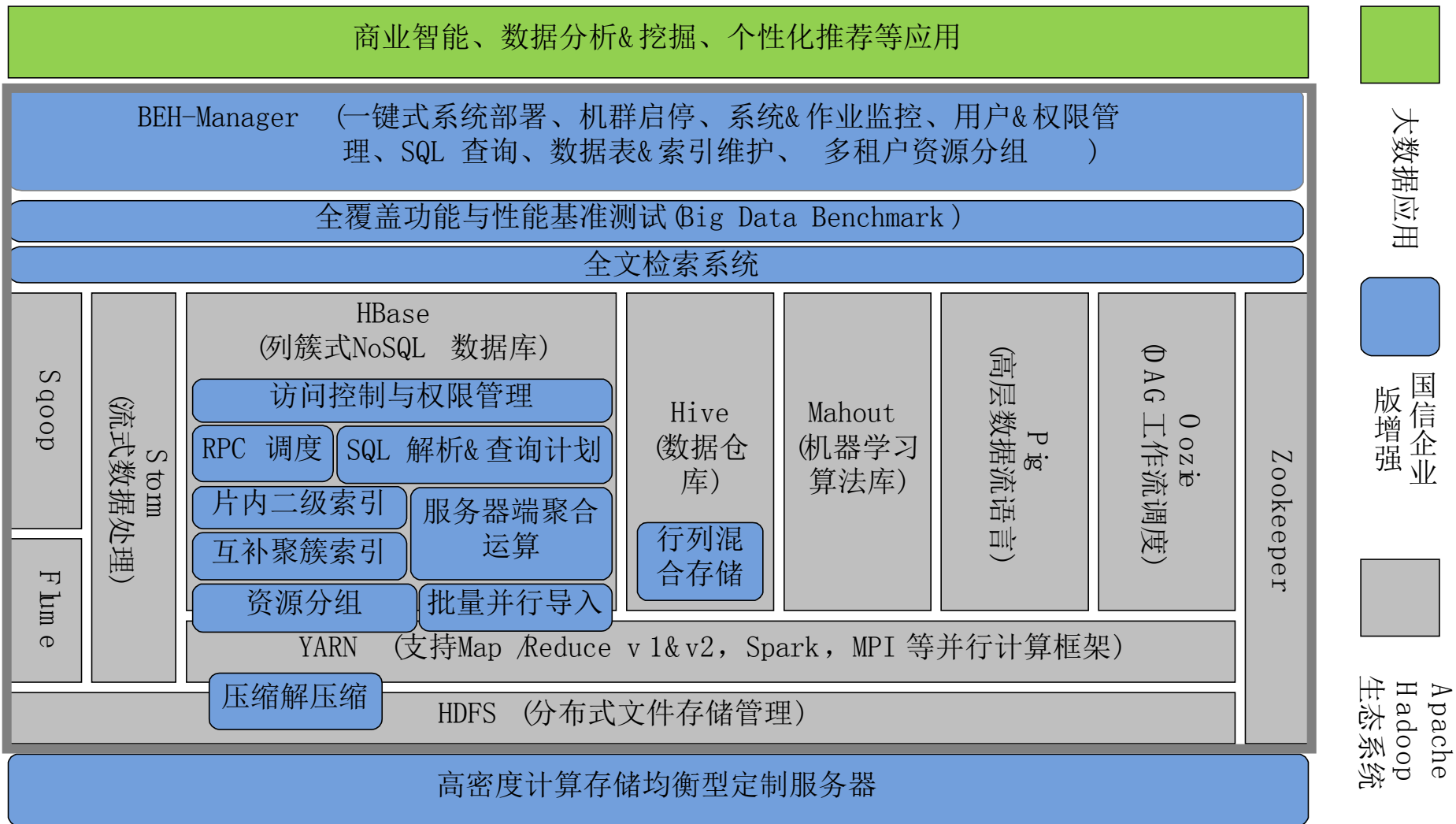
Linux



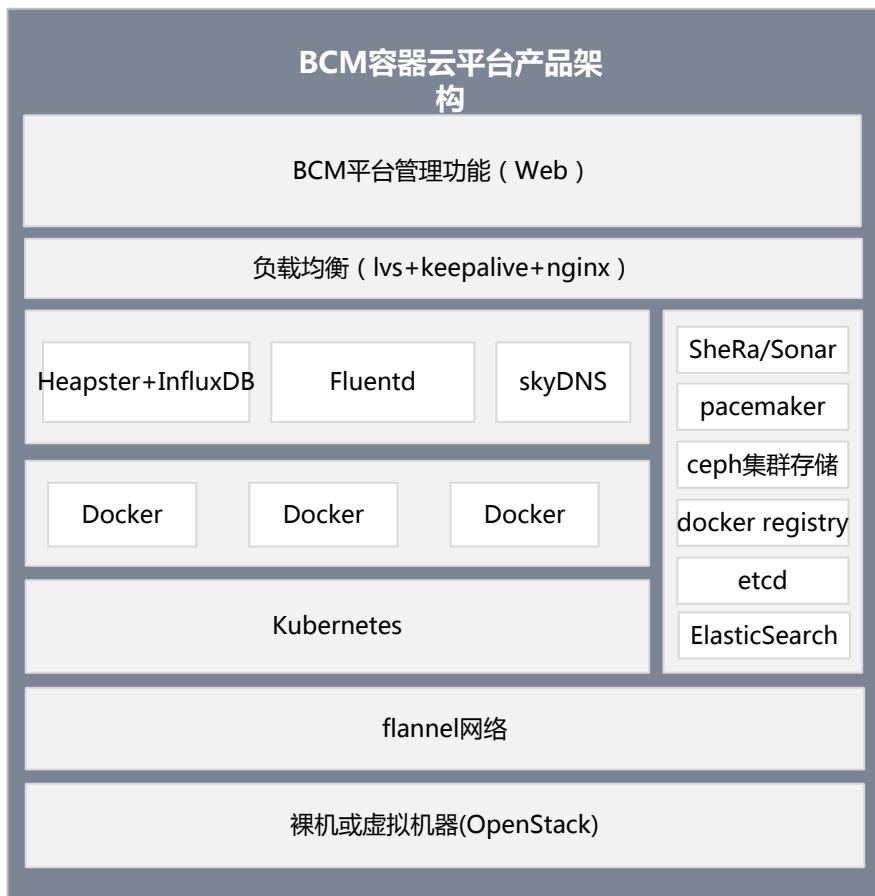
ubuntu



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。



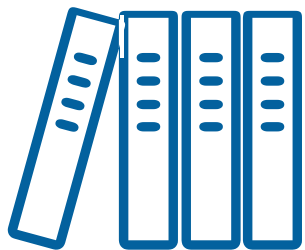
- 基于开源社区源代码实现二次开发，转换为自己的核心技术，逐渐将应用与生产实践的验证部分代码提交给社区，比如：K8s+Docker底层源代码修订；



- ◆ 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；

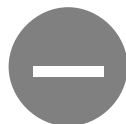


- ◆ 改进kubelet里dns相关的代码逻辑，使其允许创建多个dns服务，便于实现k8s集群dns服务高可靠性部署方案；
- ◆ 添加kubelet新的参数，方便管理和配置docker日志文件的大小和个数，配合对docker日志功能的改进，方便更高效的查询管理容器日志；
- ◆ 添加独立的kubeng模块，实时监控服务的变化，和nginx一起实现服务的发现和访问代理；
- ◆ 改进kube-proxy里创建iptables的代码逻辑，使服务的外部访问到达一个node后，不再转发到其它node；
- ◆ 改进kubelet的代码逻辑，使cephfs可以关联到多个pod上；



目录

CONTENT



运营商的大数据架构



大数据核心技术体系



运营商的应用与实践

- 运营的企业运营管理，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

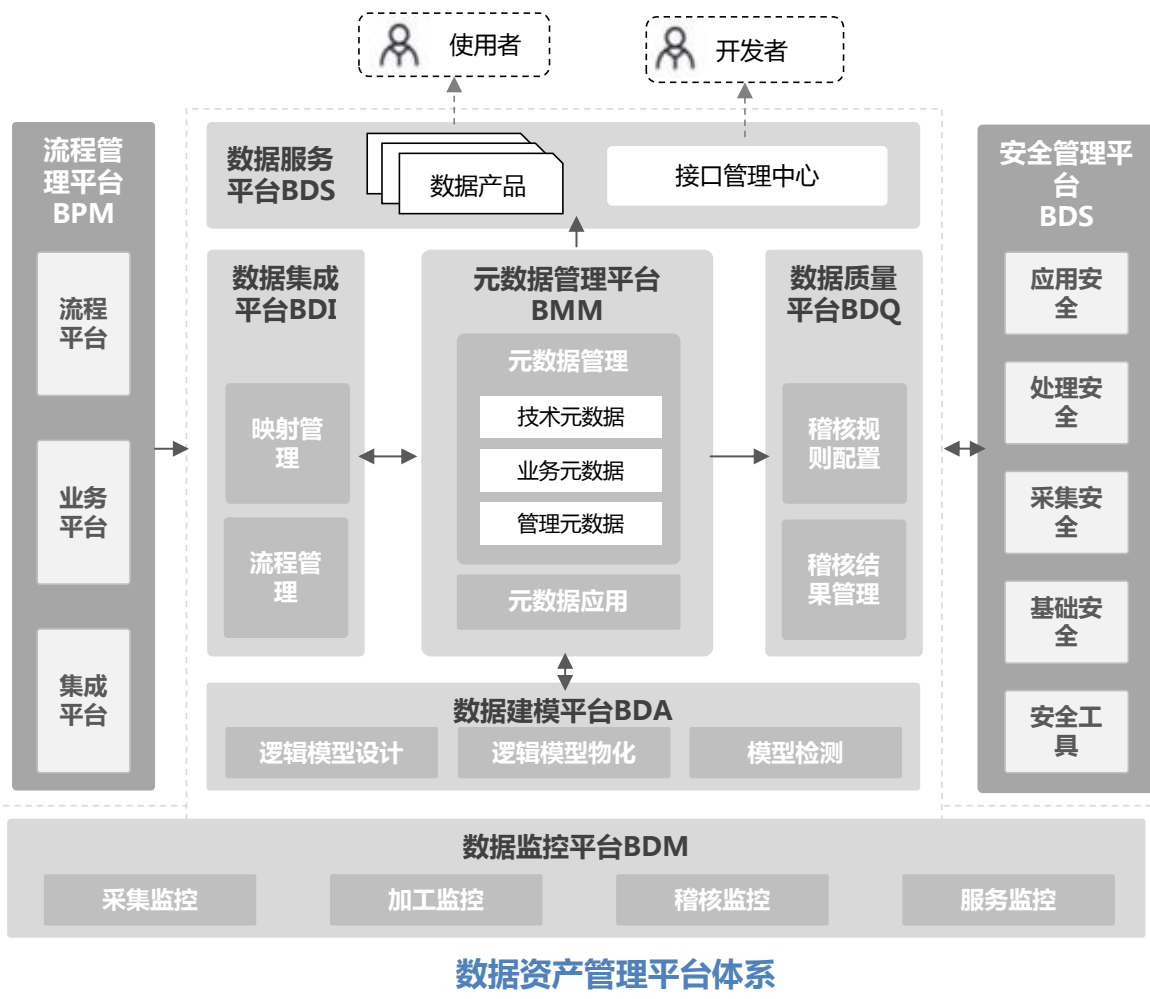
营销应用实践

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；

- 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



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。



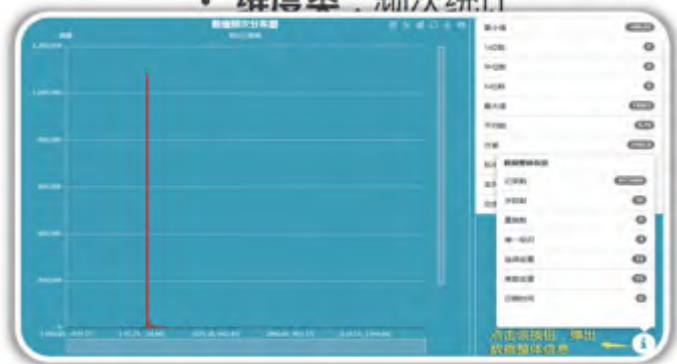
智能预警引擎—数据探索



数据检测

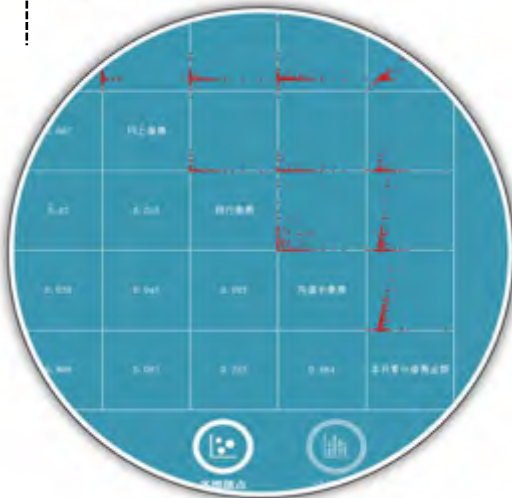
数据检测指单变量数学统计。

- **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



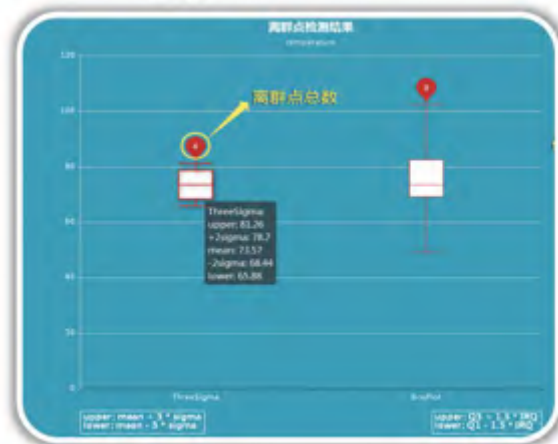
多维分析

- 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。



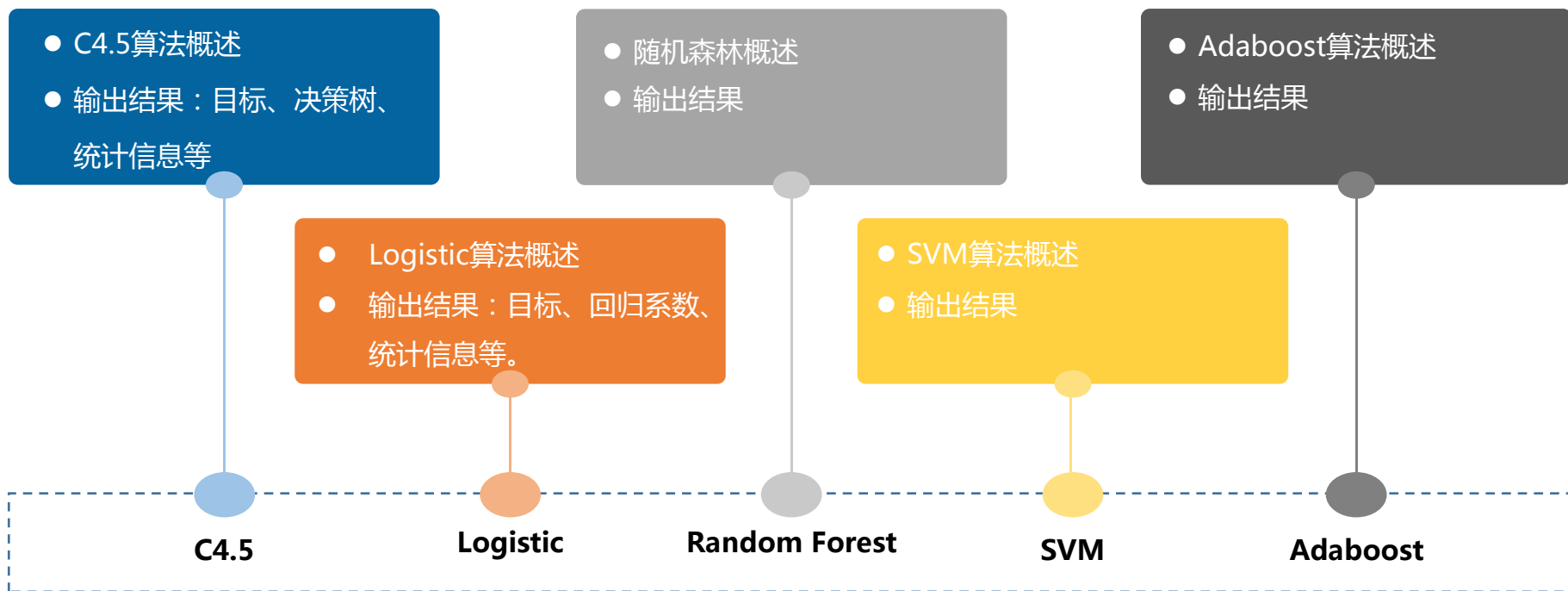
离群检测

- 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。



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。

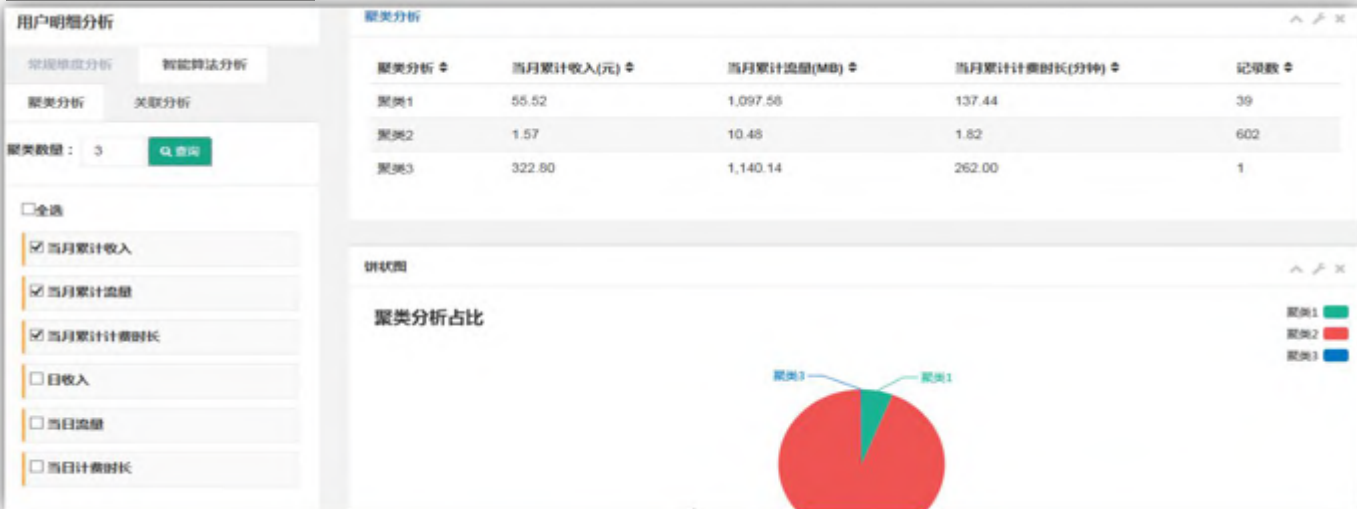
常规多维分析+智能算法分析（聚类分析+关联分析）

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

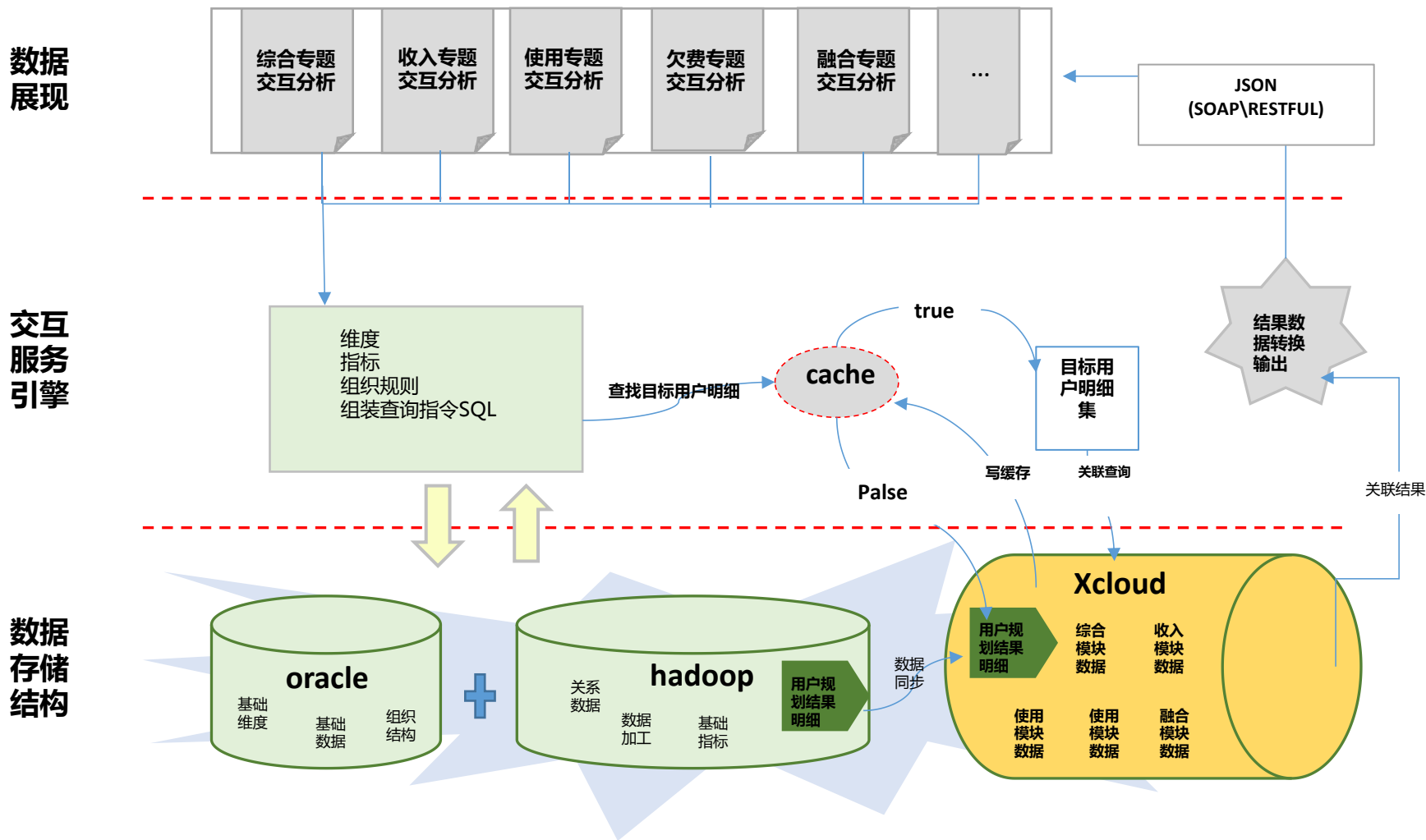
积木式组织 机构

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



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。

业务指标微服务

微服务数据查询

组织机构

- 河南省分公司
 - 郑州市分公司
 - 开封
 - 洛阳市分公司
 - 平顶山市分公司
 - 安阳
 - 鹤壁
 - 淇县
 - 浚县
 - 鹤壁市区
 - VIP客户营销部
 - 电子商务部
 - 集团客户事业部
 - 渠道服务中心
 - 山城营业部
 - 社区营销中心
 - 校园营销中心
 - 信息导航业务中心
 - 营业中心
 - 中小企业营销中心
 - 其它
- 新乡市分公司
- 焦作市分公司
- 濮阳
- 许昌
- 漯河
- 三门峡
- 商丘市分公司

维度条件选项

账期	用户数	网上用户数	活跃用户数	出账用户数	月折机用户数	欠费用户数	出账收入	欠费额
业务类型	964	6,152,761	5,513,454	6,675,640	201,513	4,541,663	366,769,266.30	625,970,231.35
客户类型	155	3,297,167	2,251,096	2,862,775	70,081	1,445,165	112,806,472.26	130,295,061.75
渠道类型	362	1,475,631	903,336	1,162,820	65,687	945,873	17,844,974.54	107,605,121.89
网络类型	726	1,638,197	1,246,120	1,512,200	48,679	903,397	5,841,316.89	90,102,357.47
合约类型	350	2,969,781	2,157,497	2,758,238	60,242	903,397	9,210,753.20	108,648,660.00
终端类型	169	1,874,415	1,231,482	1,612,752	56,654	903,397	5,080,107.91	78,873,031.16
城市类型	371	92	92	354	354	354	3,955,320.70	49,363,476.05
数据维度	135	2,1	2,1	149	149	149	9,925,768.76	125,208,679.57
周口	68,295	2,507,746	1,917,113	2,721,422	82,717	903,397	8,948,014.42	121,694,198.45
驻马店	44,193	1,918,279	1,372,589	1,848,629	59,608	903,397	5,469,436.52	89,546,821.57
三门峡	23,574	907,939	624,284	807,982	30,132	903,397	7,207,717.92	127,558,450.05
濮阳	42,560	1,275,979	949,487	1,185,368	29,137	903,397	6,838,608.28	109,355,357.91
鹤壁	23,444	844,773	662,001	780,449	21,631	903,397	2,816,851.52	47,632,020.12
济源	10,684	472,686	352,673	441,407	11,382	903,397	4,538,502.34	57,930,809.00

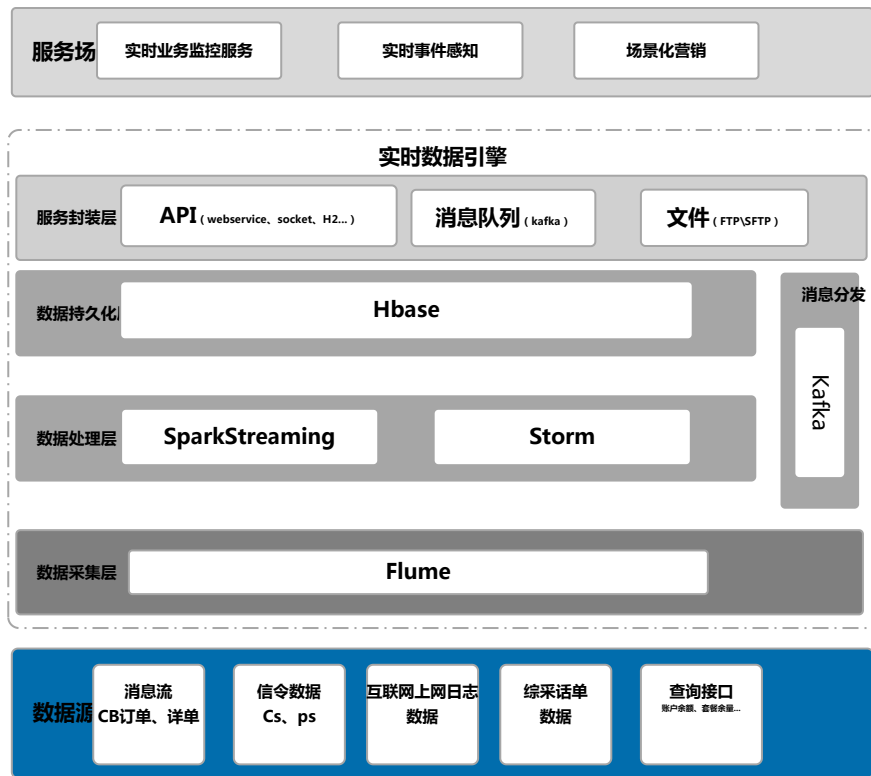
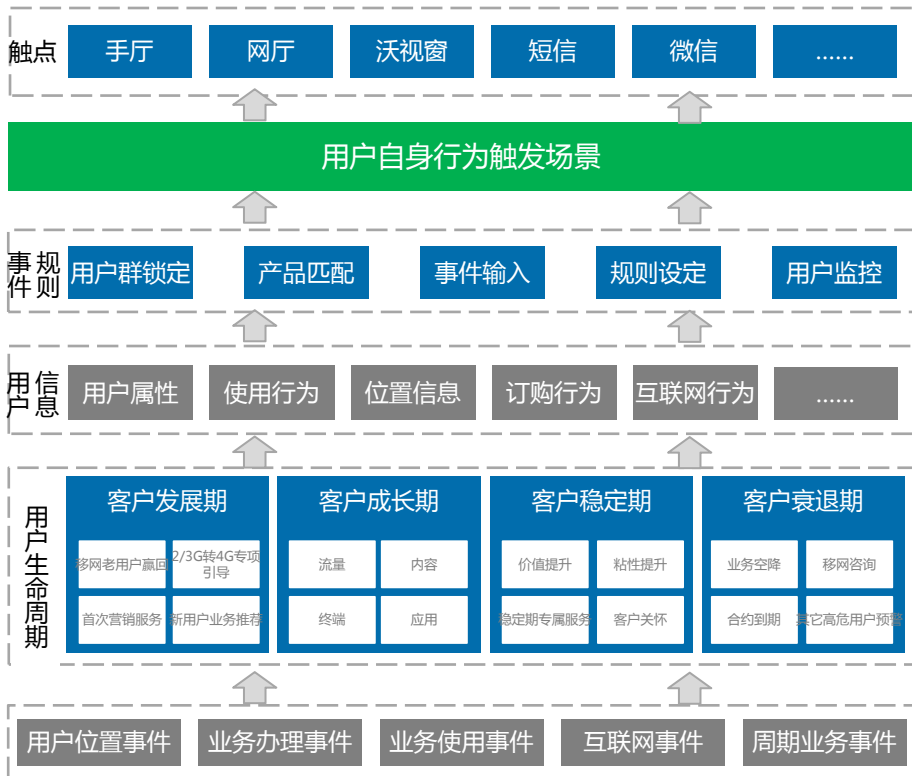
可在具体指标对应的特定用户群多维多面立体观察

积木式组织机构

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。



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；



舆情监测与舆情信息分析



需求认知与产品研究



渠道及媒体价值评估



精准服务标签



精准服务营销



竞争洞察

内容化运营时代 精准化信息认知

文本采集

标题 内容 转载 作者

时间 评论数 转发数

论坛

新闻

语音转文本

微博

公众号

APP

文本处理

提炼信息 分析信息 统计量化

找出关联 预测预警

关键词

实体

情绪

热度

摘要

聚类、分类

DPI关联其他标签

查看人数

查看人背后标签发布人

背后标签

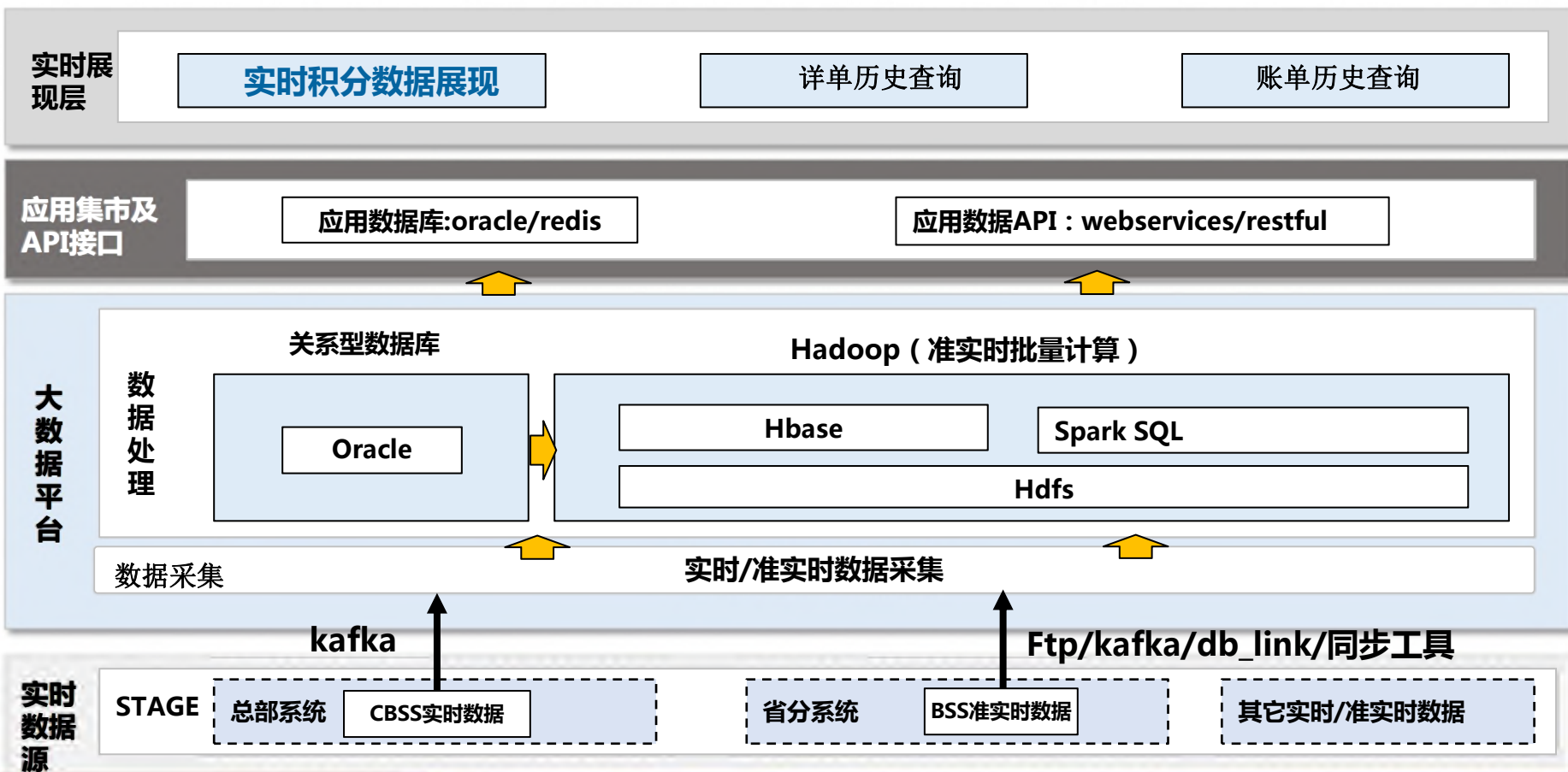
搜索引擎搜索数据

拓展了对服务需及投诉的了解和响应

拓展了对需求人群的精细认知和服务

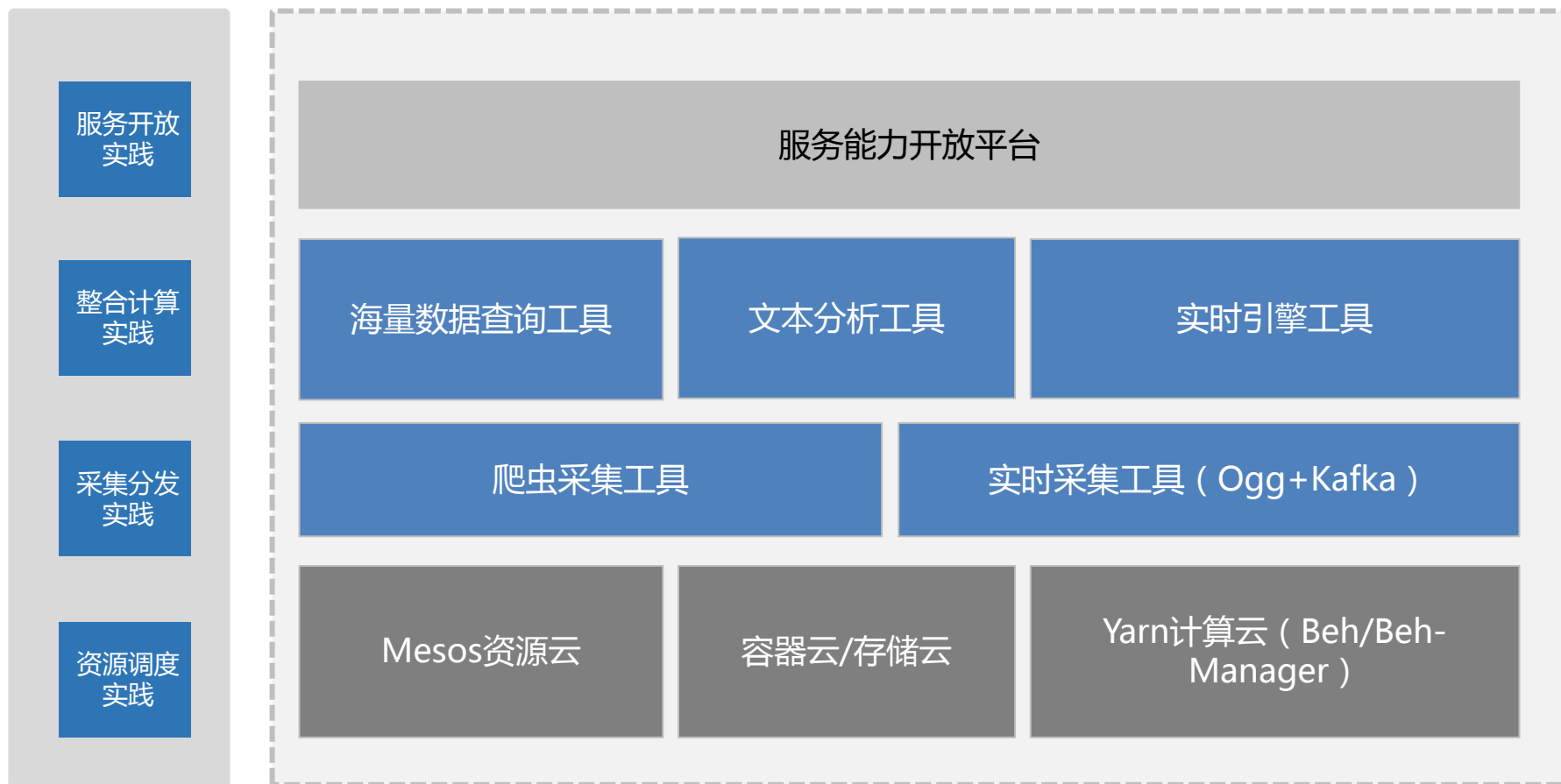
互联网大数据+DPI数据

- 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。



- 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；

DCOS

The screenshot displays the DCOS framework management interface. On the left is a navigation sidebar with options like '控制台', '监控', '集群管理', and '框架管理'. The main content area is titled 'storm集群服务列表' and includes a table of cluster information and a grid of service icons.

控制台 > 框架管理

框架管理

storm集群服务列表

运行 等待 异常 全部 创建服务

DCOS 框架管理

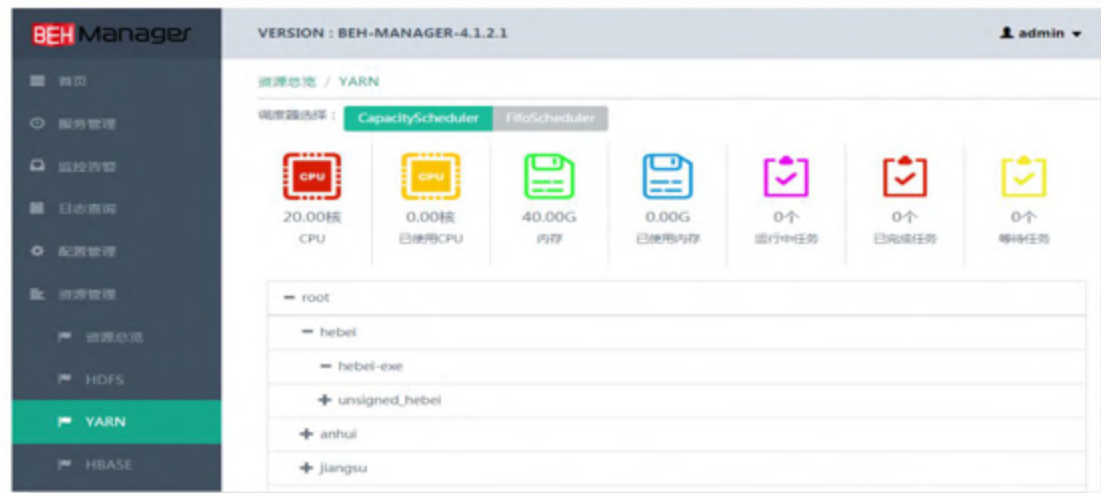
集群名称	监听地址	工单ID	节点个数	CPU	内存	存储	创建时间	修改时间	服务状态	操作
storm001	http://192.168.30.16:30012/	123456789123456789123456780	4	4	4096M	4096M	2017-03-29	2017-03-29	running	停止 扩容 添加任务

+ 添加框架 框架管理

Storm MySQL XCloud KAFKA Redis

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。

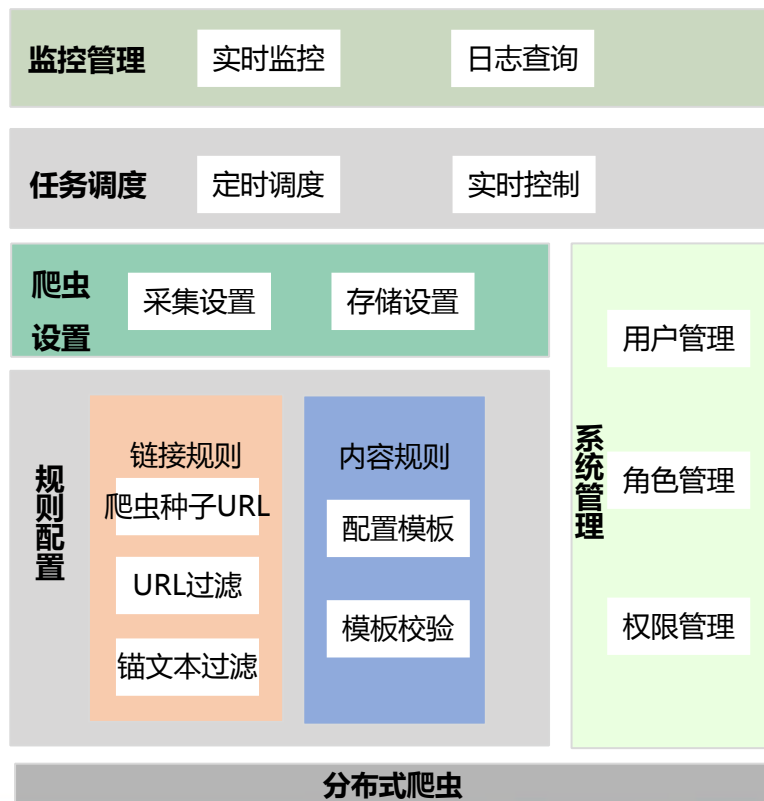


- 1、部署步骤繁琐，整合困难
- 2、部署固化
- 3、非直观、不可视，看不见、摸不着
- 4、运维工作复杂，对人员专业技能要求过高
- 5、管理工具分散，管理手段过多
- 6、无法满足多个租户合理使用平台
- 7、在多租户使用状态下，如何保证数据的安全性？
- 8、指标无监控和预警机制
- 9、平台行为记录可视化和无检索的问题

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。

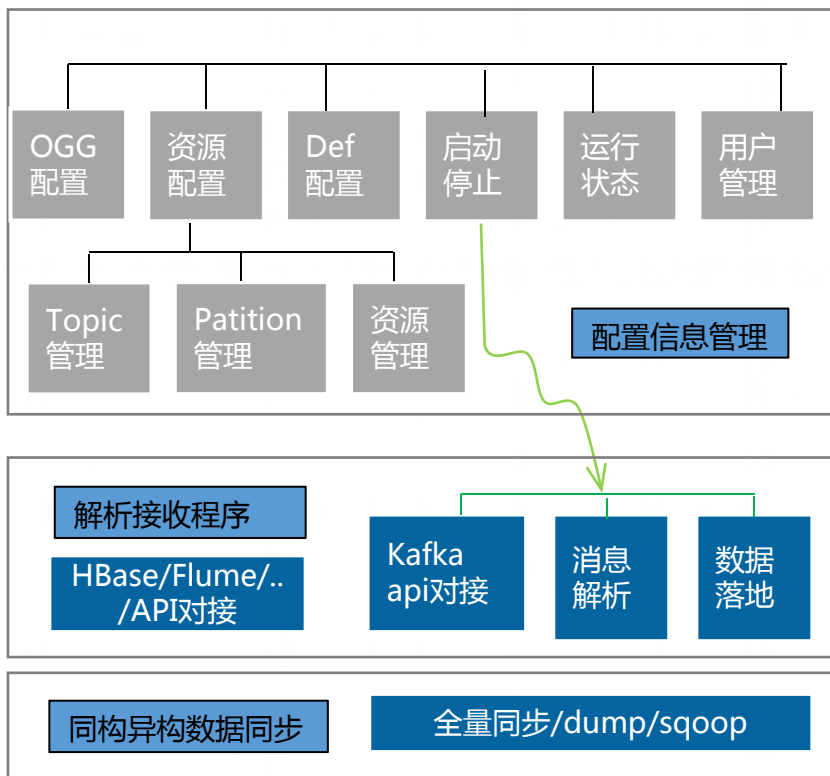
互联网爬虫功能架构



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。

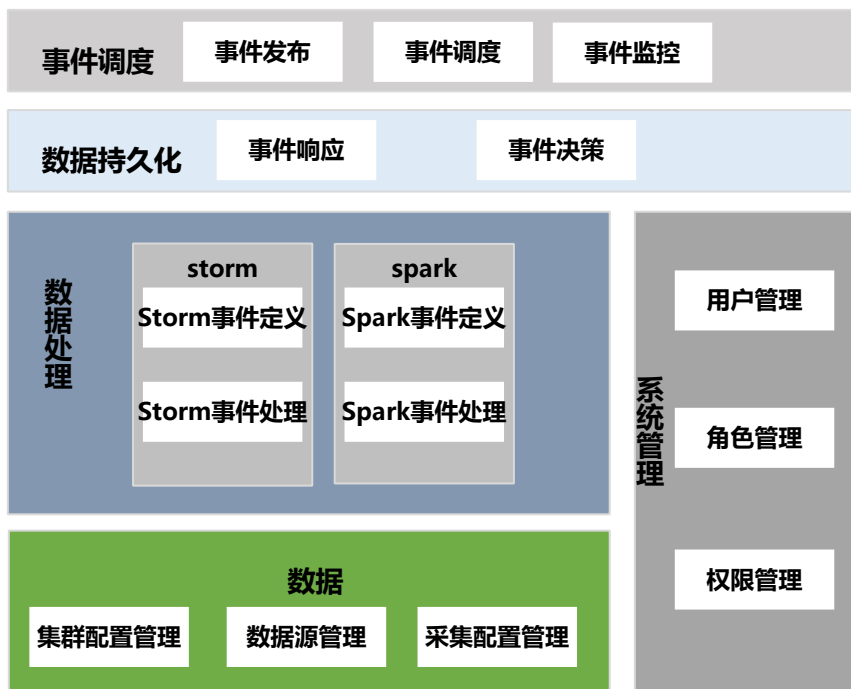
OGG+Kafka功能架构



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。

The screenshot displays a web-based text analysis tool interface. The main content area shows a document snippet with the title "山西煤企遭遇最冷寒冬 撤销县级煤运公司涉煤部门" and a date "2015-11-18". Below the text, a "分类" (Classify) button is visible. The interface also features a sidebar with navigation options like "自动分词", "我要关键词", "实体抽取", etc., and a "文本分类" (Text Classification) section. The classification results are presented in two formats: a circular gauge chart and a table.

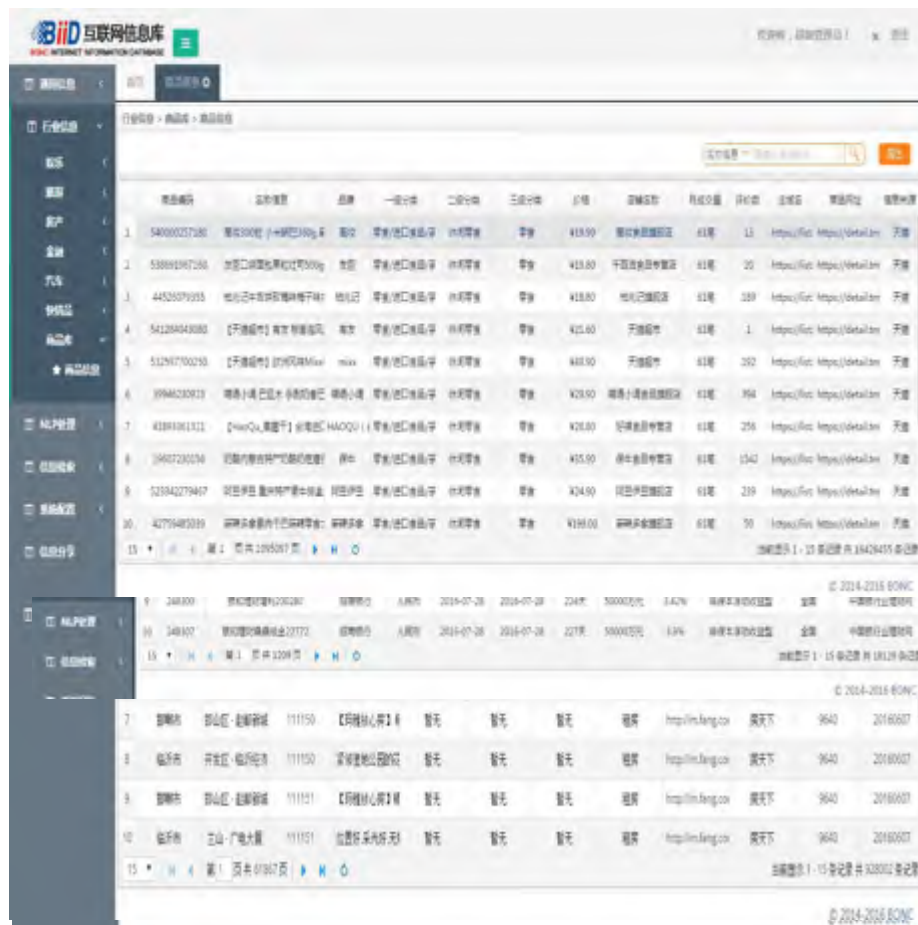
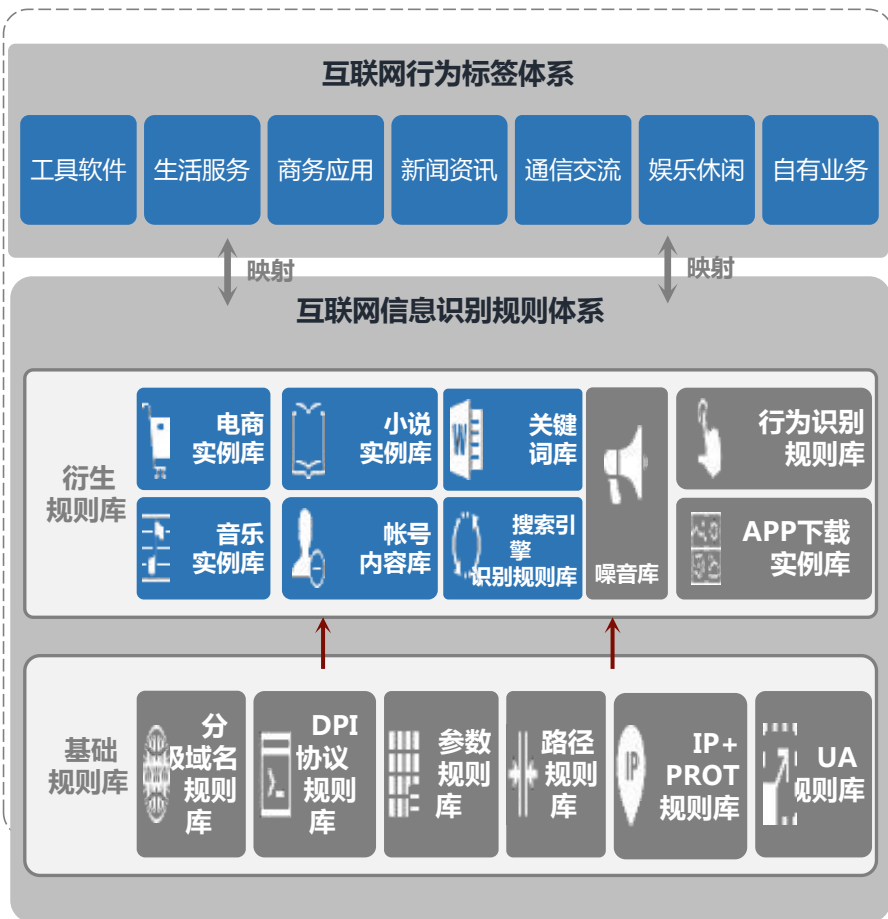
分类结果

置信度: 58.04%

分类	置信度
金融	0.5804
旅游	0.2867
IT	0.0404

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；

The screenshot displays the BCAP OPEN API Service Open Platform interface. The top navigation bar includes '我的能力', '应用案例', '文档中心', and '关于我们'. The main content area is titled '系统日志' (System Logs) and shows a list of log entries. The left sidebar contains various monitoring and management tools.

类型	日期	日志信息	更多
INFO	2017-03-30 15:39:28,365	'admin@carbon.super [-1234]' logged in at [2017-03-30 15:39:28,365+0800]	更多
INFO	2017-03-30 15:38:43,173	API Store Default Context : http://172.16.63.78:9763/store	更多
INFO	2017-03-30 15:38:43,171	API Publisher Default Context : http://172.16.63.78:9763/publisher	更多
INFO	2017-03-30 15:38:43,170	Mgt Console URL : https://172.16.63.78:9443/carbon/	更多
INFO	2017-03-30 15:38:42,980	WSO2 Carbon started in 41 sec	更多
INFO	2017-03-30 15:38:42,979	Server : WSO2 API Manager-1.10.0	更多
INFO	2017-03-30 15:38:42,973	JMX Service URL : service:jmxrmi://localhost:11111/jndi/rmi://localhost:9999/jmxrmi	更多
INFO	2017-03-30 15:38:42,883	Successfully Initialized Eventing on Registry	更多
INFO	2017-03-30 15:38:42,806	Task service starting in STANDALONE mode...	更多
INFO	2017-03-30 15:38:40,421	Pass-through HTTP Listener started on 0:0:0:0:0:0:8280	更多
INFO	2017-03-30 15:38:40,410	Starting Pass-through HTTP Listener...	更多
INFO	2017-03-30 15:38:40,409	Pass-through HTTPS Listener started on 0:0:0:0:0:0:8243	更多

- **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！