中国移动
China Mobile

中国移动“大云“大数据产品及应用

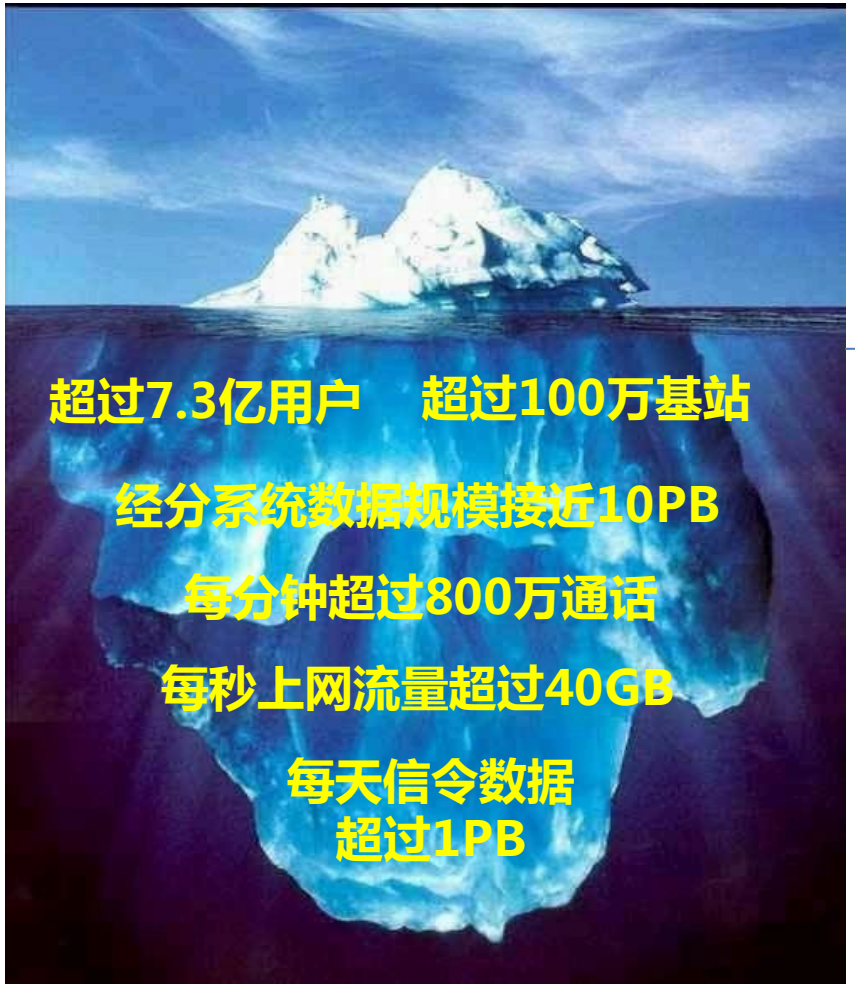
BDTC2014

徐萌

中国移动苏州研发中心

www.10086.cn

移动运营商的大数据有什么？

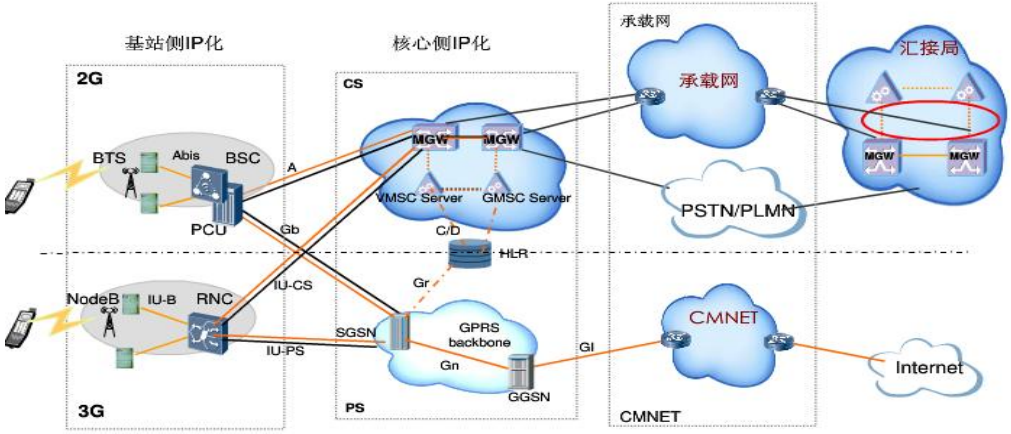


移动互联网服务商

专业SNS　博客　消息
电商　视频　优惠券
图片
新闻　点评　音乐　微博
地图　问答　SNS　签到　论坛

电信运营商

2G、3G、4G、WIFI

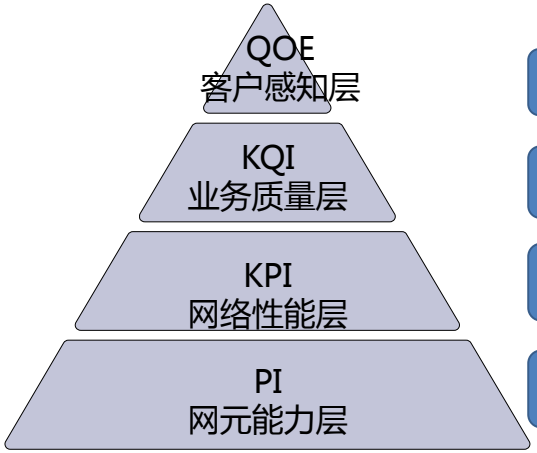


大数据成为网络优化、业务创新、精准营销和决策支持等工作的基础

电信运营商发挥管道优势，深入挖掘大数据的价值

借助云计算技术和运营商优势，将大数据信息转化为商业价值，促进业务创新

1. 优化网络质量：利用信令数据支撑终端、网络、业务平台关联分析，优化网络，实现网络价值最大化
2. 助力市场决策：充分挖掘用户的移动互联网行为特征，提升对用户消费偏好的精准把握，帮助市场营销等决策，实现精准营销。
3. 改善用户体验：智能语音门户通过知识库和语义搜索技术实现业务知识的机器智能回答



网络质量和效率分析

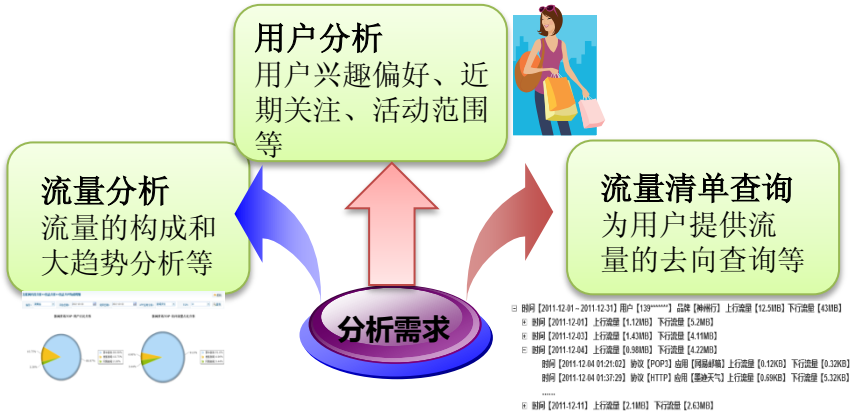
网络价值分析

客户感知及行为分析

业务端到端分析

优化网络质量

改善用户体验

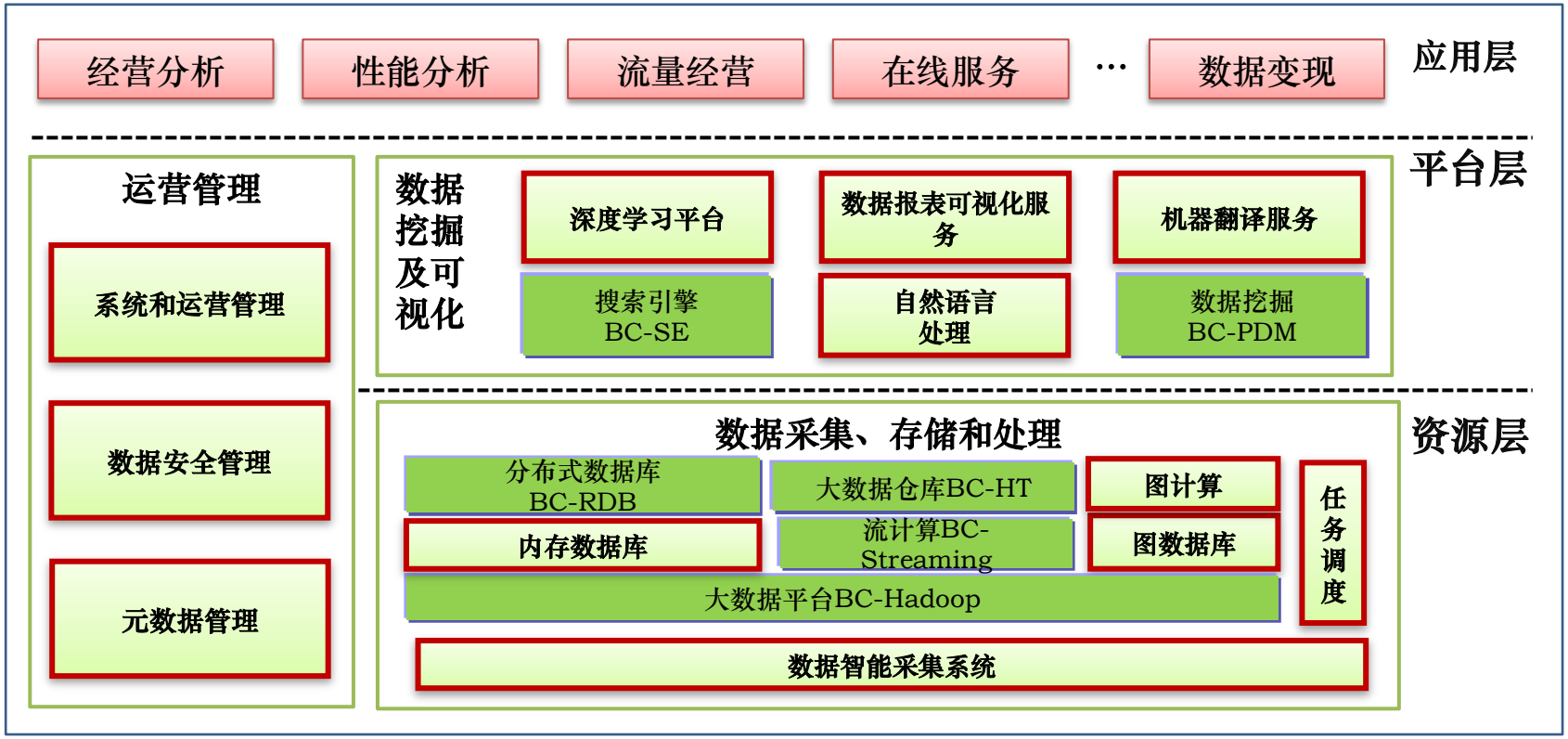


用户帐号	协议类型	业务类型	访问URL	终端信息	上行流量	下行流量	……

助力市场决策

中国移动“大云”大数据产品整体规划

大云大数据产品线为中国移动大数据应用提供三大领域的基础能力：**数据采集和处理、数据挖掘及可视化、运营管理三大领域。**



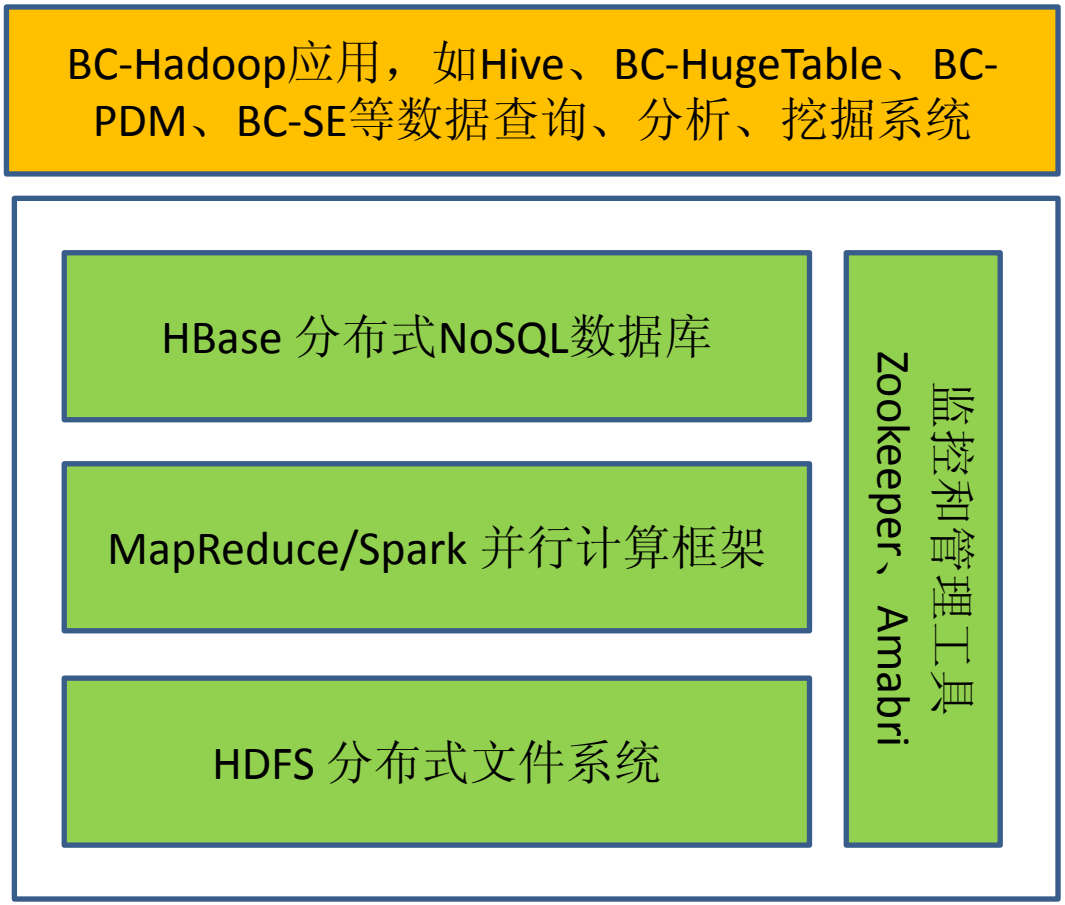
成熟产品　　新增产品

BC-Hadoop：“大云”大数据平台

各种不同版本的Hadoop软件目前已经在各个省公司广泛应用，但是由于产品化程度低，存在商业版本不开源、开源版本不统一、运维管理功能弱、多应用混合部署能力不足等问题。

BC-Hadoop 2.0主要特性

1. **开源开放**：核心系统是CDH5改进版本，代码开放，Patch反馈社区
2. **管理增强**：集成Ambari管理系统，支持BOMC、4A规范（在研），支持puppet自动部署系统
3. **资源共享**：利用YARN提供资源分配和调度方案
4. **多租户**：支持基于用户、队列的Hadoop多租户方案
5. **可靠性**：所有Hadoop组件没有单点问题
6. **服务化**：提供基于BC-EC弹性部署方案，支持弹性MapReduce计算



BC- HugeTable：“大云”大数据仓库系统

各省帐详单云主要采用开源HBase软件；云ETL主要采用开源Hive软件。难以解决对帐详单做分析，对ETL数据做查询的要求。一般需要建设两套系统，保存两份数据。BC-HugeTable针对同一份数据提供数据查询和数据分析功能。具有独特优势。

BC-HugeTable 5.1 主要特性

1. **交易和分析一体化**：通过集成Hive实现分析查询，集成Impala支持即席查询，研发OLTP引擎支持快速交互查询，实现智能路由。
2. **高可用**：支持BC-Hadoop提供的高可用能力；提高HTLoader的可用性
3. **高性能**：针对复杂索引查询在秒级别返回结果；复杂分析在分钟内完成
4. **管理增强**：支持资源池、运营管理平台集成；支持监控、告警、计量、统计接口；支持SNMP、OMI协议
5. **兼容性**：支持原生MapReduce和NoSQL接口；支持多数SQL92查询；兼容Hive、Impala、HBase数据操作API

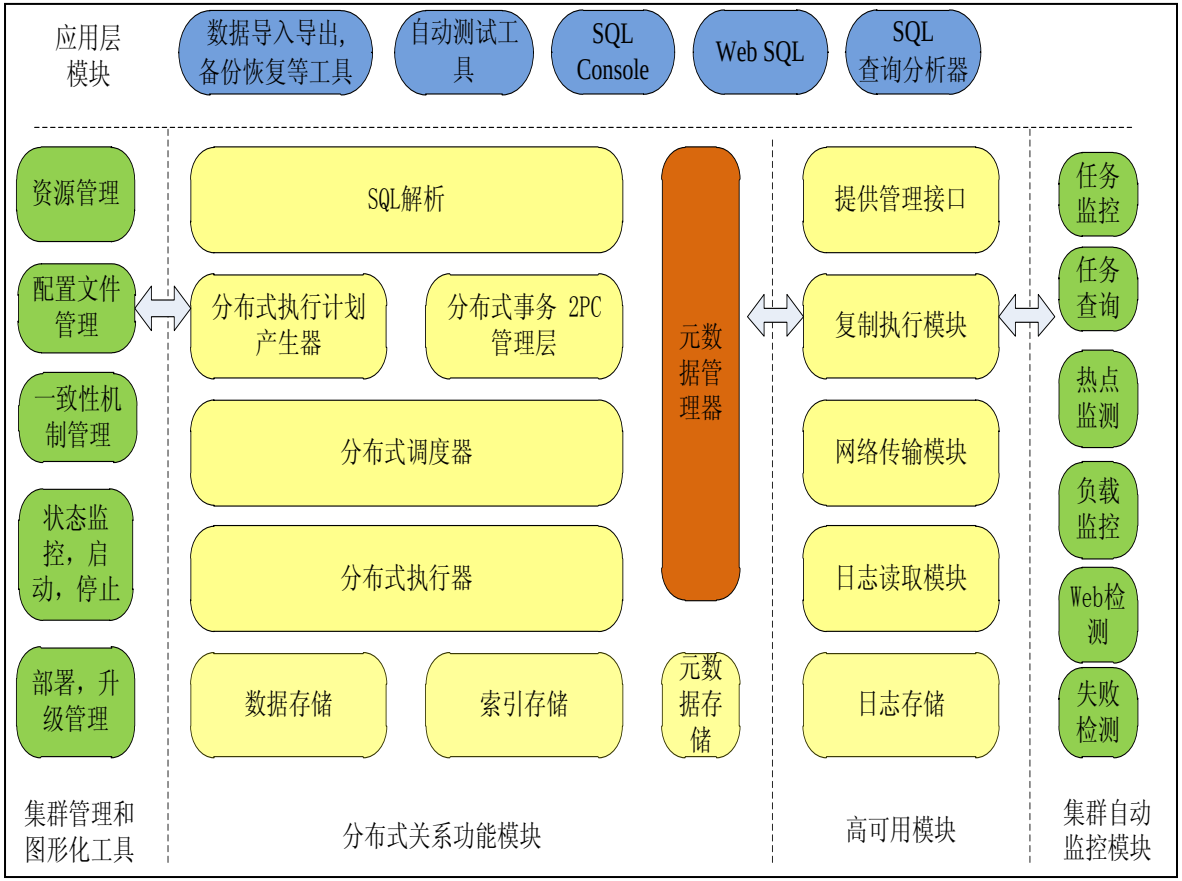
	HugeTable	商用MPP方案
单集群规模	复用Hadoop能力，支持5000节点规模	<300节点
SQL兼容性	支持主要SQL	完全支持
响应时间	部分SQL比MPP慢，部分相当	较短
优化能力	需要改进	较好
开放性	开源技术，社区非常活跃	私有方案
Hadoop支持	紧密集成	将Hadoop作为外部数据源
软件举例	Impala、Presto、Drill/mDrill、Shark、SparkSQL等	GreenPlum、Redshift、Asterdata、Vertica等
适用场景	大数据在线分析	中小型集市分析

BC-RDB：“大云”分布式关系数据库

传统OLTP数据库应用系统主要问题是采购和建设成本高、超许可使用，BC-RDB是基于X86服务器的、通过集群技术提供高可靠、高可用和高性能的分布式数据库系统，成为一种去IOE技术方案。

BC-RDB 2.2主要特性

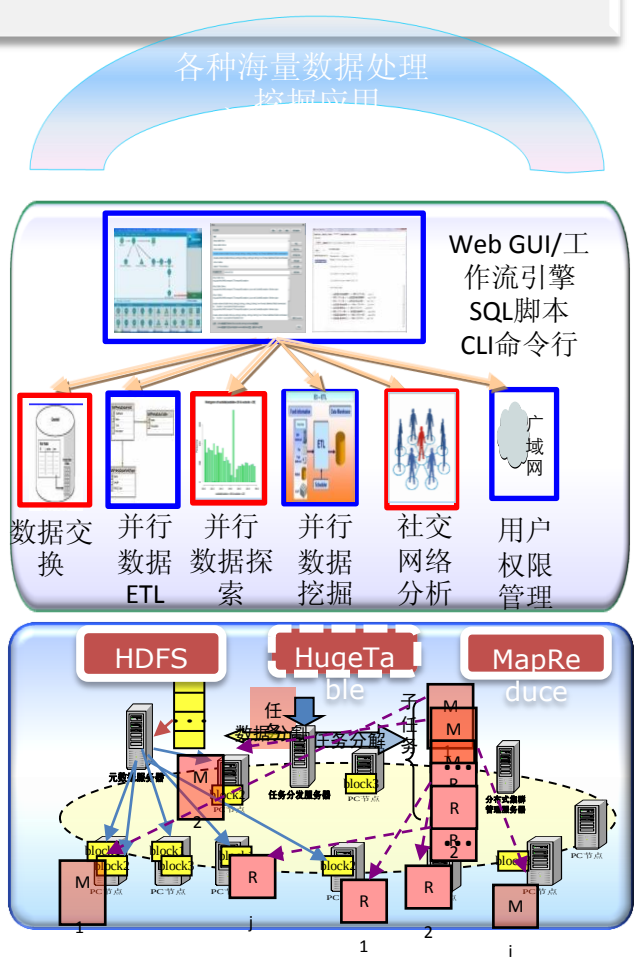
1. **高可靠**：数据在多个服务器上形成多副本，同步写完多个副本才成功。在存储引擎层保证一致性
2. **高可用**：集群节点互为备份，主备节点热备切换
3. **高性能**：在负载均衡环境，提供读写分离服务；可以采用高性能硬件优化
4. **兼容性**：完善SQL92兼容开发，仅子查询不支持，Join未经优化。提供Oracle 数据导入导出支持
5. **管理增强**：提供完善的统一监控、部署Portal；提供故障告警和数据一致性分析脚本



BC-PDM：“大云”数据挖掘和ETL工具

相比开源数据挖掘软件，BC-PDM具有功能强大、简单易用、多应用支持，服务完善等优势；相比传统数据挖掘工具，BC-PDM处理能力强、性能优秀、可以完成数据全量挖掘、定制化程度高，具有明显优势。

	大云BC-PDM	传统数据挖掘产品	开源软件（Mahout等）
数据处理规模	1000TB以上级别，各种应用的全量数据挖掘，集群	10TB级别，各种应用的抽样数据挖掘，单机	1000TB以上规模，部分应用的全量数据挖掘，集群
算法支持	21种挖掘算法，非结构化算法、SNA	算法种类同左，有更多细分算法	算法种类较少，集中在推荐算法
数据预处理支持	45种ETL操作	支持	不支持
数据来源	各种格式文件、数据库	各种格式文件、数据库	各种文件
使用方法和定制开发能力	友好，界面拖拉拽、SQL、定制化算法插件	友好，界面拖拉拽	不友好，命令行程序
用户群	数据分析工程师、第三方工具开发者	数据分析工程师	程序员
产品服务	培训、现场、远程、升级、定制化开发	培训、现场、远程、升级	无服务



客户评价：“大云BC-PDM领先业界同类产品一年”

“大云”大数据应用实践

中国移动 China Mobile

- **对内支撑精细化运营**：支撑客户体验提升、精细营销、产品创新、网络优化、企业管理水平提升。
- **对外寻求新业务增长点**：支撑行业大数据解决方案、数据变现及社会化洞察等对外服务模式。

大数据对外服务

- 行业大数据解决方案
- 数据变现
- 社会化洞察

市场营销

- 基于位置的实时推荐
- 实时互动个性化推荐
- 基于设备的实时个性化推荐
- 竞品分析
- 产品引入分析
- 产品优化
- 产品设计和开发
- 客户对产品的购买概率分析

客户体验

- 垃圾短信拦截
- 搜索业务优化
- 客户离网风险预测
- 个性化的实时交互人工服务
- 实时的客户接触关怀
- 客户离网原因预测
- 沉默用户（服务）主动关怀
- 客户体验差的时候主动关怀
- 个性化的挽留营销活动

网络优化

- 用户投诉故障定位
- 客户掉话率分析
- 实时WIFI转移（四网协同）
- 基于价值的网络规划
- 网络故障检测和恢复
- 基于价值的实时网络拥塞管理

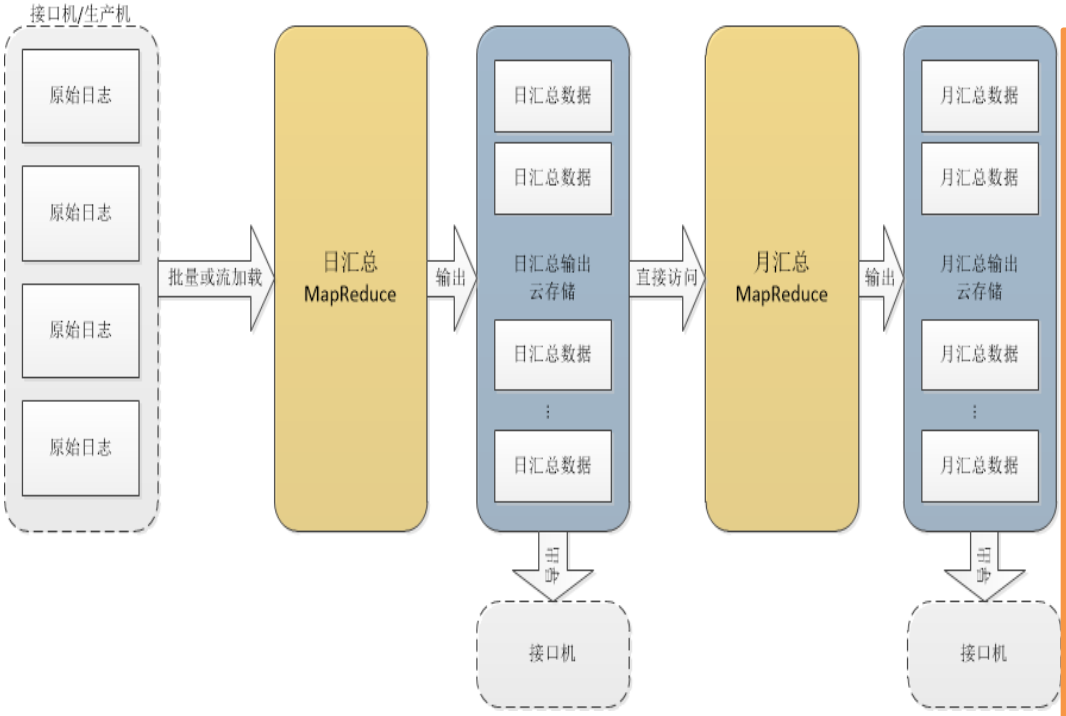
IT系统优化

- ETL云化
- 帐详单查询
- 终端进销存系统
- 上网行为分析
- 运营管理分析
- 指标库查询

云ETL

现状：以数据仓库的E-L-T模式为主，海量原始数据直接入库，对数据仓库产生极大压力，数据仓库扩容压力大，影响其他分析业务正常运行。

解决方案：以BC-Hadoop、BC-HugeTable为基础，基于BC-PDM工具针对结构化、非结构化数据实现ETL操作，包括从各种数据源获取数据，并进行清洗、转换、去重、缺值补充等操作，进而实现上报一经各类数据分析及汇总。



例图：分时段汇总的业务场景

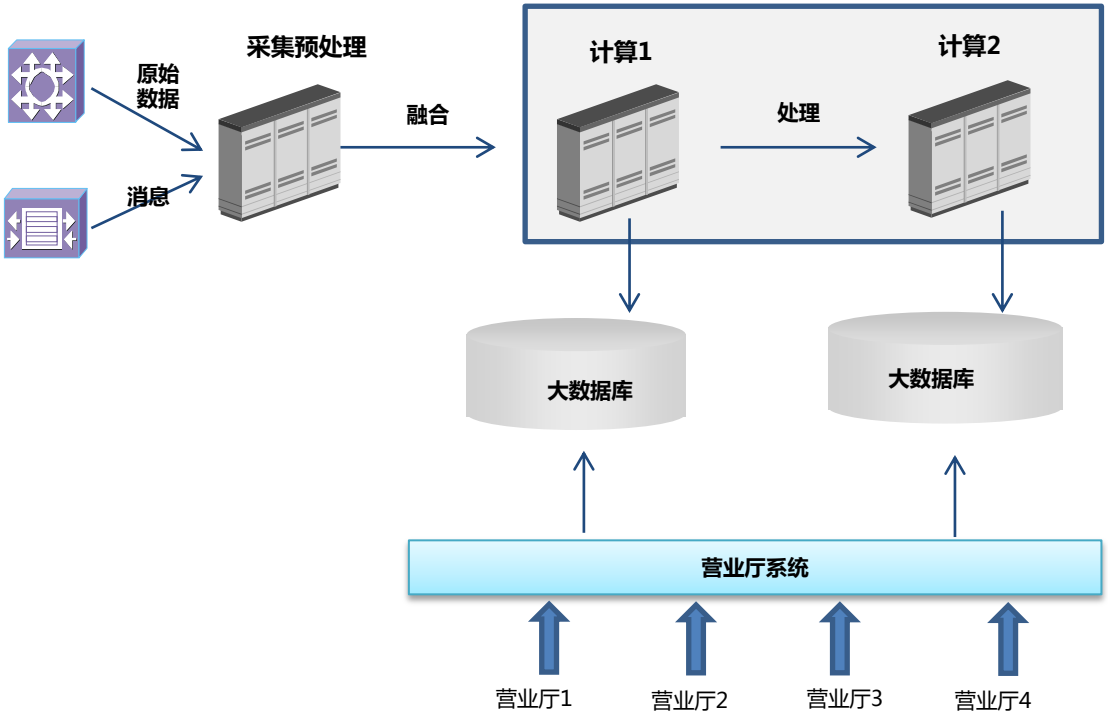
效果：

- ✓省公司：完成一经wap日志30天存储、分析、汇总，减少原有数据库25%的长时间负载占用，性能提高3-5倍。
- ✓省公司：存储60天数据，6PC
  - ✓受业务复杂度影响，性能提升比例不同，但绝对时间上提升明显；
  - ✓对于而且对于数据量大、逻辑相对简单的业务提升比例更高，日调度提升平均3倍以上，月调度提升部分可达5倍以上

详单类数据查询分析

现状：数据库承载详单类型数据的查询及分析操作，随着用户及4G业务增多，数据库压力大响应延迟增加。

解决方案：以BC-Hadoop、BC-HugeTable为基础，仅保存一份数据，以标准SQL支持对详单类数据的查询与分析统计，包括支持客服的详单查询、上网日志查询、网络数据查询及分析等。



例图：帐详单查询系统

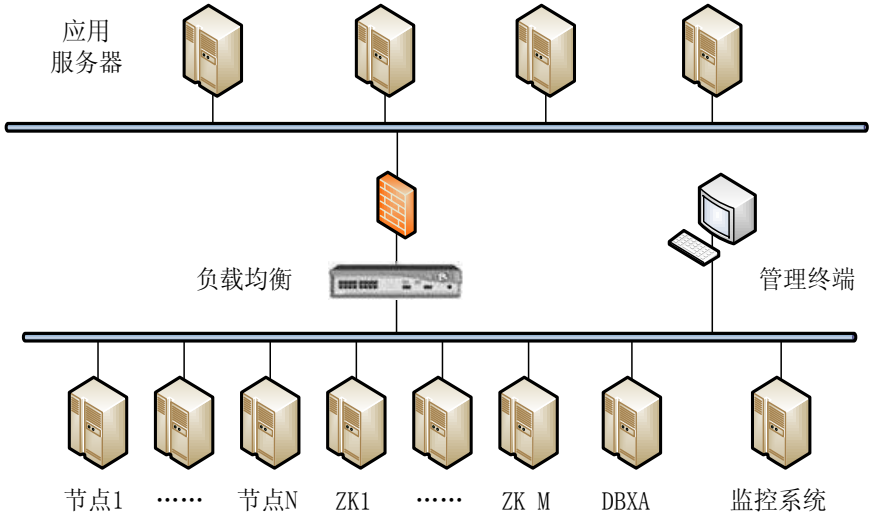
效果：

- ✓**省级帐详单查询系统**：30个节点满足6个月数据供150多T数据存储，秒级支持客服及用户的详单查询
- ✓**省级运营管理大数据平台**：12节点存储7天详细日志1年汇总数据，支持管信客户感知专题及CRM防绕行审计，基于大云实现大数据平台自动安装部署、监控及管理，同时支持ooize、pig等组件。

交易数据库应用（去IOE）

现状：对于海量数据的事务处理需求，现网小型机系统在扩展性方面遇到瓶颈，开源单机数据库性能支持不够，只能采用分库的方案，而在跨库查询时给应用改造带来一定复杂度。

解决方案：基于分布式数据库BC-RDB系统实现分布式事务和统计分析功能，支持标准SQL接口，提供高并发和高可靠性的数据库系统，传统数据库可平滑迁移。



集群可以部署于自带硬盘的x86服务器，不需要小型机和磁盘阵列

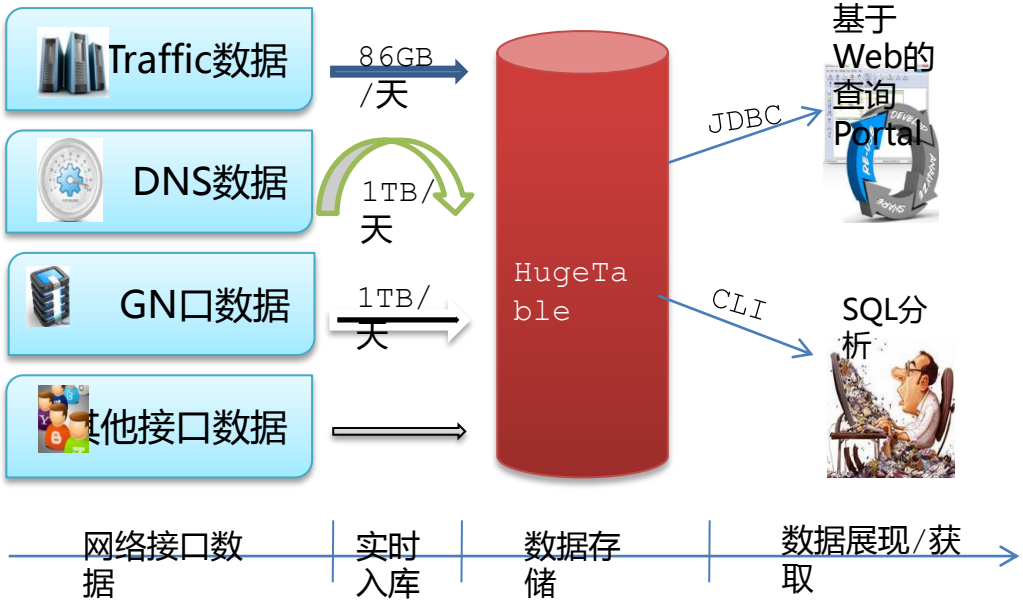
效果：

- ✓**省级终端进销存系统**：6节点支持3000人并发服务于各终端网点、经销商、各级公司以及供应商的销售支撑业务管理平台。
- ✓**省级BC-RDB一体机测试**：完成4节点集成测试，实现数据库基准功能，验证了BC-RDB在多种平台的可集成性。

用户投诉故障定位

现状：传统网络数据保存周期短，4G业务增多，网络数据巨大日增数十TB（省），无法应对网络优化需求。

解决方案：以BC-Hadoop、BC-HugeTable为基础，支持各种网络数据存储，包括traffic/Gn/Gb/wlan等数据，支持网络投诉的迅速定位、掉话率分析等等



例图：分时段汇总的业务场景

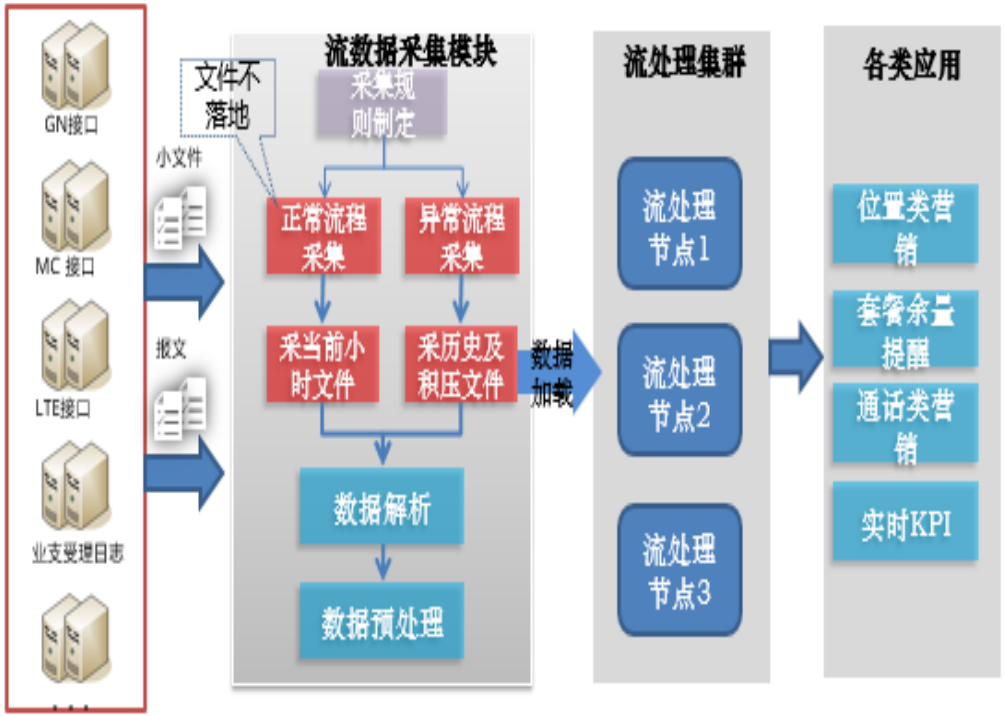
效果：

- ✓省公司
  - ✓LTE与2G/3G信号共存干扰现象是影响无线通信网络质量的关键因素之一，当接到用户投诉时，采用传统方案，平均需要5-7个工作日完成故障定位，现在故障定位时间缩短到分钟级别。
- ✓省公司
  - ✓感知掉话率是传统话统掉话率4倍，感知掉话率与用户实际投诉匹配度更高

基于流计算的位置精准营销及实时PKI

现状：面向实时类市场营销需求以及实时信息决策需求，目前对于实时数据的快速响应和处理，目前现网系统还难以支持。

解决方案：基于开源流计算系统支持实时数据处理及响应，提供实时数据缓存、数据分析、事件累积及触发等能力。



例图：客户分类识别应用

效果：

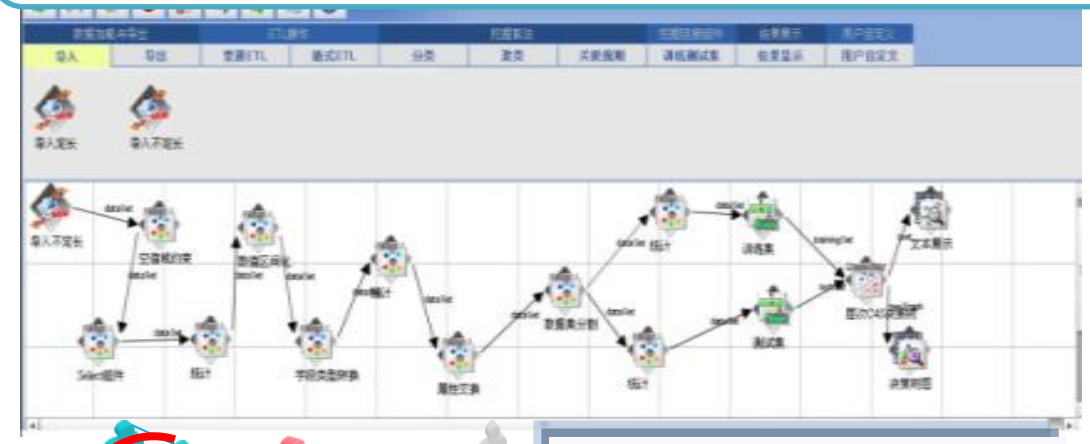
✓可支持的应用包括：

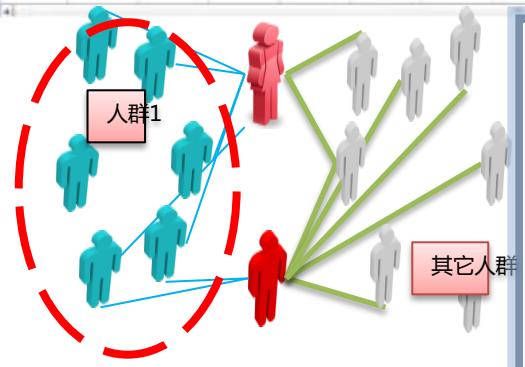
- ✓实时营销类：基于位置的营销
- ✓实时体验类：套餐余量提醒
- ✓实时KPI监控：用户数、通话成功率、短信数等指标监控

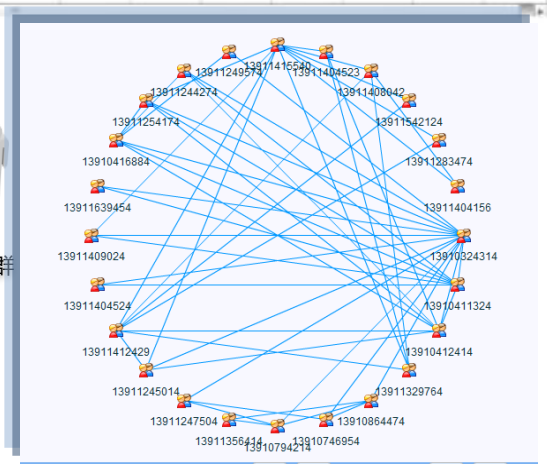
数据挖掘和专题分析

现状：对于海量数据的挖掘与分析，由于传统数据挖掘工具系统性能有限，通常采用抽样手段来完成，不能对全网数据分析，尤其针对社交网类型应用，抽样不能解决根本问题。

解决方案：基于BC-PDM实现数据挖掘专题，提供面向结构化、非结构化数据分析挖掘，支持分类、聚类、关联规则、社交网分析等近20种算法，实现了数据探索、数据流程可视化、数据结果展示及流程调度等功能。







例图：客户分类识别应用

效果：

- ✓**数据挖掘试点**：分别在福建泉州、河南商丘、上海公司实现了BC-PDM试点，可以对全网数据进行分析，效果良好
- ✓**省数据挖掘专题**：实现了无锡融合套餐用户流量适配模型、家庭宽带专题等数据挖掘流程，正在验证环节
- ✓**电信交网圈应用**：客户影响力分析、客户重入网分析、家庭客户识别、集团客户行为趋势分析

POI搜索系统

现状：位置基地选择商用POI搜索系统，难以支持移动业务的定制化需求，例如基于运营数据的系统优化，POI数据扩充及检索排序需求等。

解决方案：以BC-SE为基础，实现对POI母库及关键词库的多重索引机制，提供灵活的与公交查询系统集成接口，提供类别排序定制化需求。

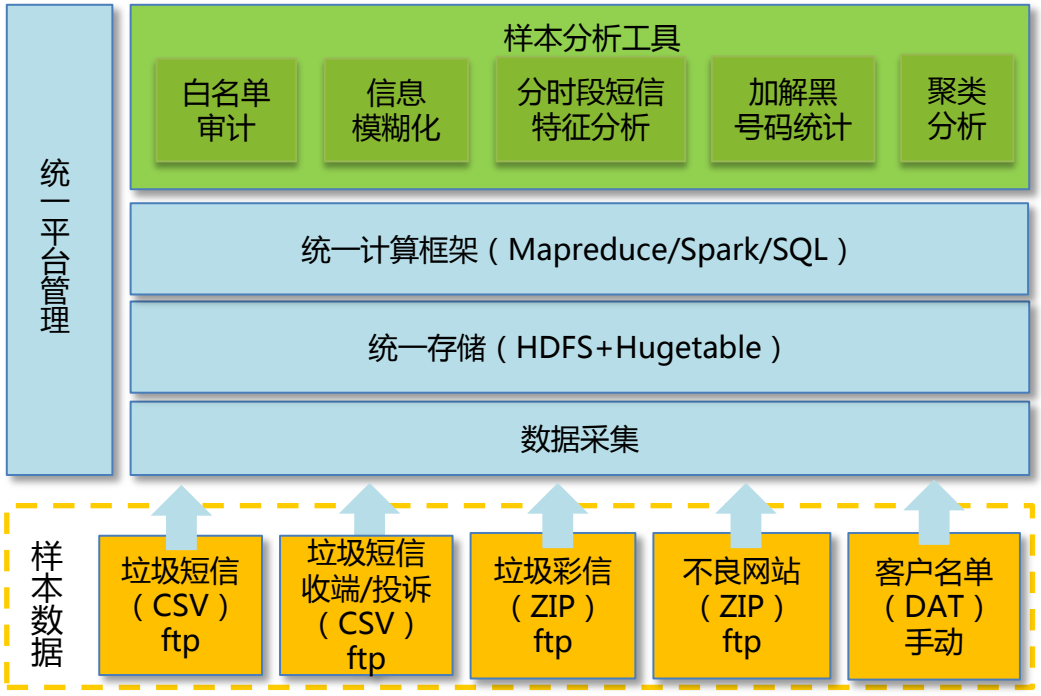
效果：

- ✓支持运营数据（点击率）对搜索结果的优化
- ✓支持按类别进行POI搜索及排序
- ✓提供系统的可运营可管理能力
- ✓支持定期手动和自动更新POI信息库

垃圾短信拦截

现状：对于海量垃圾短信、彩信等各种不良信息，传统基于规则的方式过滤，过滤结加以人工筛查，由于规则判定特征，人工工作量巨大，难以满足现网需求

解决方案：基于BC-Hadoop实现不良信息统一存储，提供统一计算框架，BC-PDM实现各种文本分析算法，包括聚类、分类等，支持不良信息自动化决策支持。



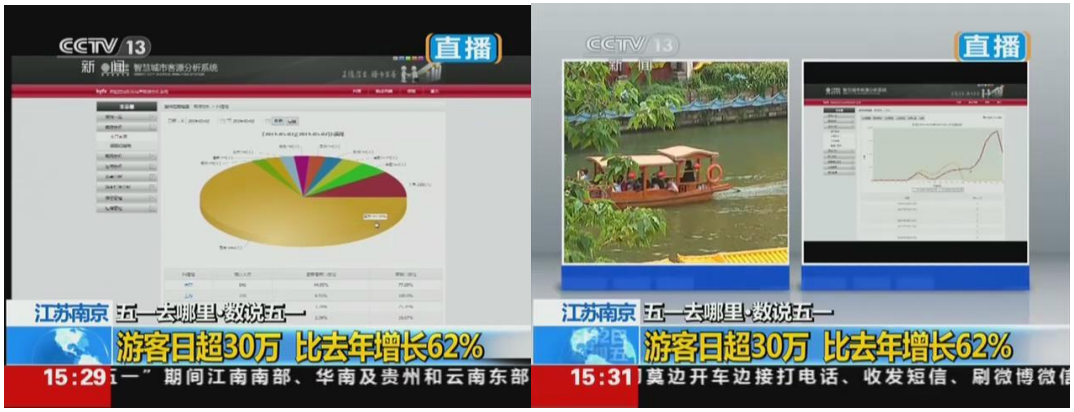
不良信息样本存储与分析平台

效果：

- ✓**规则的优化**：对垃圾短信行为模式的发现，例如频繁发送，只发不收等基本规则优化
- ✓**基于文本内容的识别**
  - ✓针对人工校验结果的不精确性，采用聚类算法方式，给出纠错建议
  - ✓利用人工校验结果作为训练集，采用指纹算法等方式实现垃圾短信识别
  - ✓有效减轻人工校验工作量，经过测试，系统验证违规短信与人工判定违规误差10%

行业解决方案

“智慧洞察”（Smart Insights）对外数据服务平台。平台依托企业数据中心强大的处理能力与海量数据，基于完全匿名和聚合的移动数据，利用统计分析、数据挖掘等技术，向客户提供标准化数据产品、大数据分析报告、高效Open API服务。为社会、政府、企业以及家庭、个人客户提供经过分析挖掘而形成的价值产品与服务，实现数据价值提升与共享。



效果：

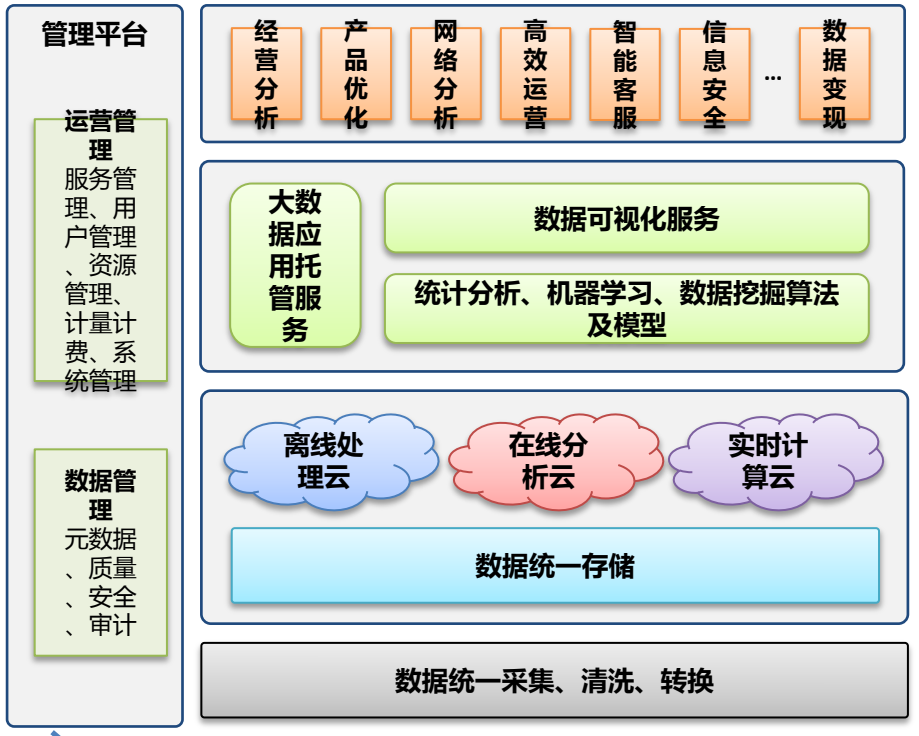
客源分析标准产品：客源构成分析：分析人群构成，区分出真正游客人员。

- 景区景点人员密度：通过基站分析各景点实时客流情况，便于疏导与管理。
- 流量监控与预警：提高景区管理职能、服务能力及安全保障能力

数据开放API：在2014年度江苏省智慧旅游推进会上，此项目被江苏省旅游局评为“江苏省智慧旅游优秀项目”

中国移动大数据应用展望

规划中国移动的大数据中心，提供数据获取、存储、处理等服务能力以及提供大数据应用创新平台



- 对外能力开放：能力开放及应用托管
- 可视化服务：可视化展示及人机交互
- 数据分析模型：文本、统计、机器学习

- 存储：大规模（百PB）、热数据、温数据、冷数据
- 离线处理：非结构化数据转为结构化知识、数据归一化、离散数据的整合
- 在线处理：大规模数据高性能即席查询、SQL支持
- 实时计算：流式数据实时计算
- 统一调度：混合部署和统一资源调度

- 多种数据源采集，内部数据、外部数据
- 采集效率：数据压缩、增量采集
- 带宽建设

- 数据管理：元数据管理、数据质量、安全访问、审计
- 运营管理：资源管理、计量

中国移动
China Mobile

谢谢！