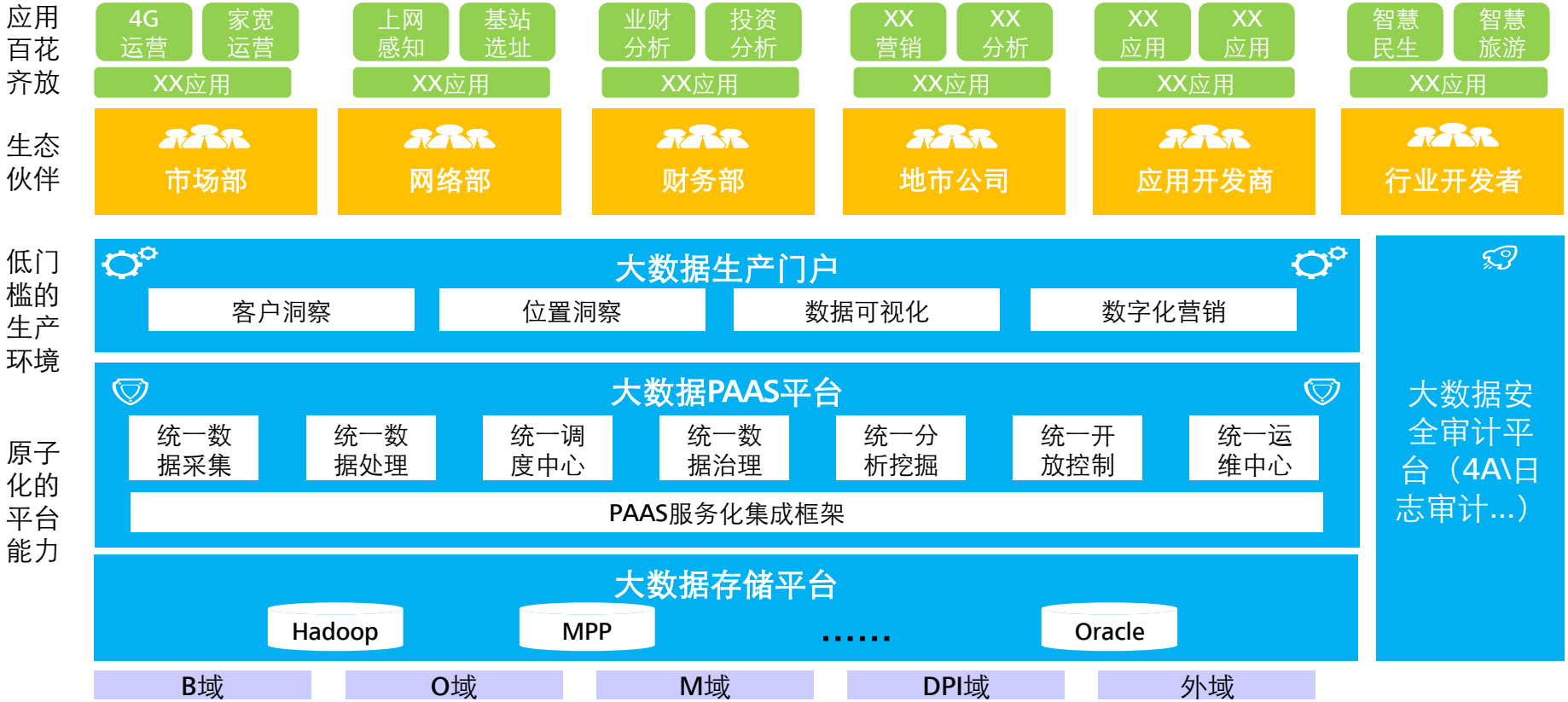
企业级大数据平台架构

www.huawei.com

HUAWEI

目录

1.1 企业级大数据运营架构总体介绍

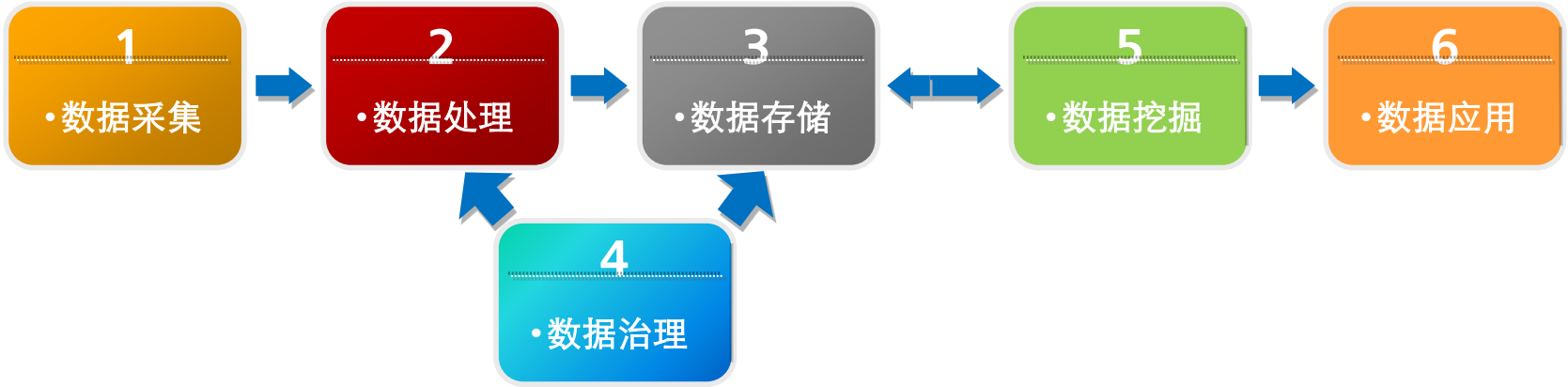


1.2 企业级大数据平台关键技术

分类	大数据技术	相关工具与产品	
数据分析与服务	图谱处理与机器学习	特征/标签库	
	数据预测与挖掘	关联/聚类/分类/回归 Mahout(Hadoop) / Mllib(Spark)	
	数据查询/统计与分析	不作一一列举	
数据存储与计算（结构化/半结构化/非结构化）	数据计算框架	MapReduce-批处理离线	Storm 实时流处理
		Spark-批处理实时	
	内存数据库	Redis	
	NoSQL / NewSQL	HBase/MongoDB/VoltDB	
	关系型数据库（RDBMS/SQL/MPP）	Oracle/DB2/MySQL Greenplum/HP Vertica	
数据采集与处理	ETL（抽取-转换-加载）	不作一一列举	
	数据采集	Flume/Kafka	
基础架构支持	云计算、云存储、虚拟化	不作一一列举	

1.3 企业级大数据运营流程

- 针对企业内外部的各类数据源，通过数据采集、预处理、数据存储及数据加工处理，为不同应用及使用对象提供多种数据服务。

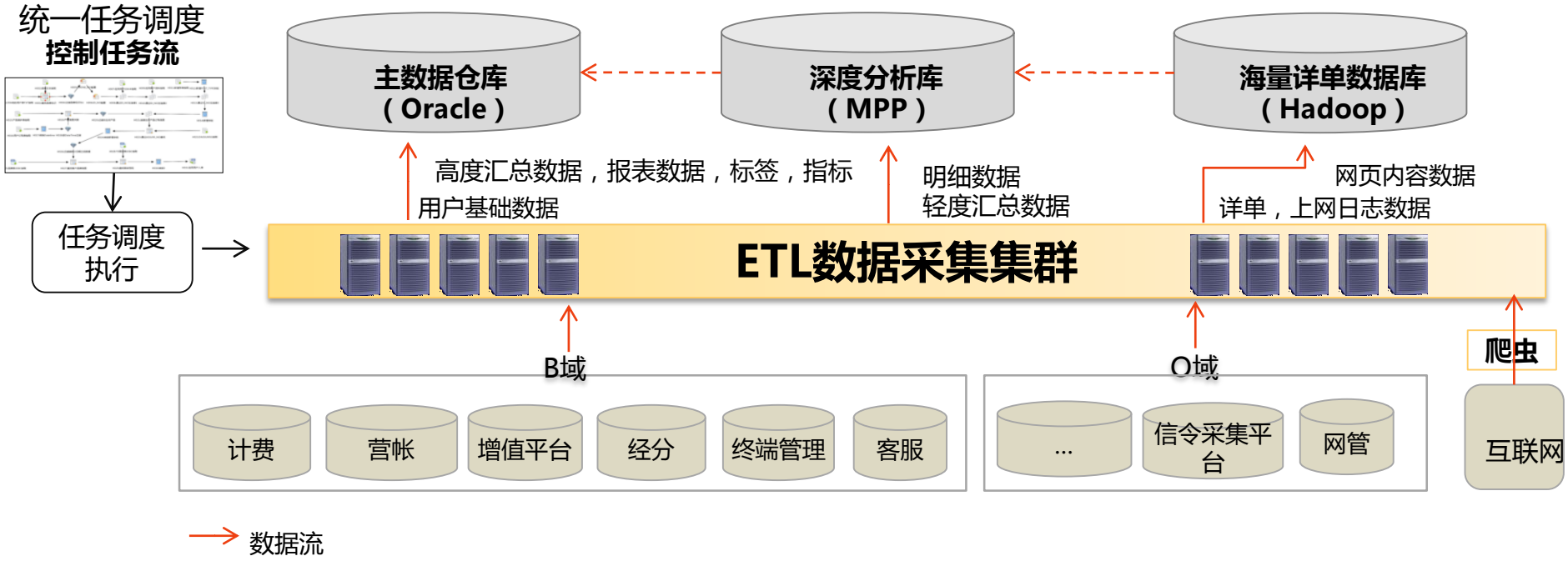


1.3.1 数据采集

- **ETL**:是BI领域形成的专有名词。数据抽取、转换以及加载。之前一直是聚焦于数据批量处理的功能。
- **Crawler:** 随着业务的发展，数据分析不再局限于内部数据分析，根据指定的URL位置提取对应的网页内容叫**爬虫**。
- **流处理（Streaming)**：对实时发送过来的数据进行实时处理，用于支撑实时、准实时分析，如营销和监控、告警场景等。

1.3.1 数据采集（续）

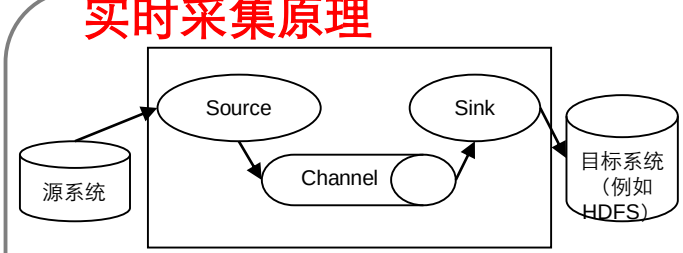
- 离线数据采集

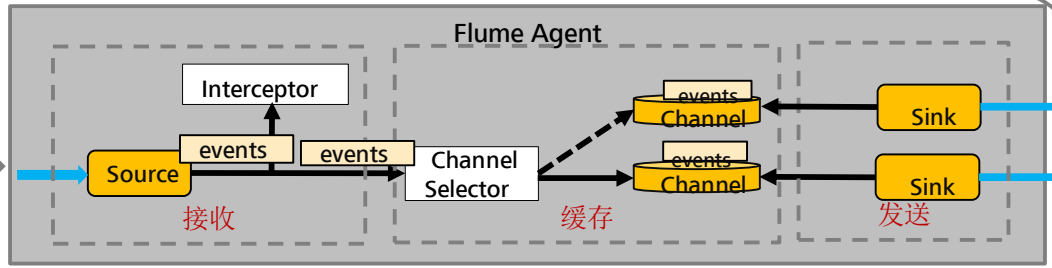


1.3.1 数据采集（续）

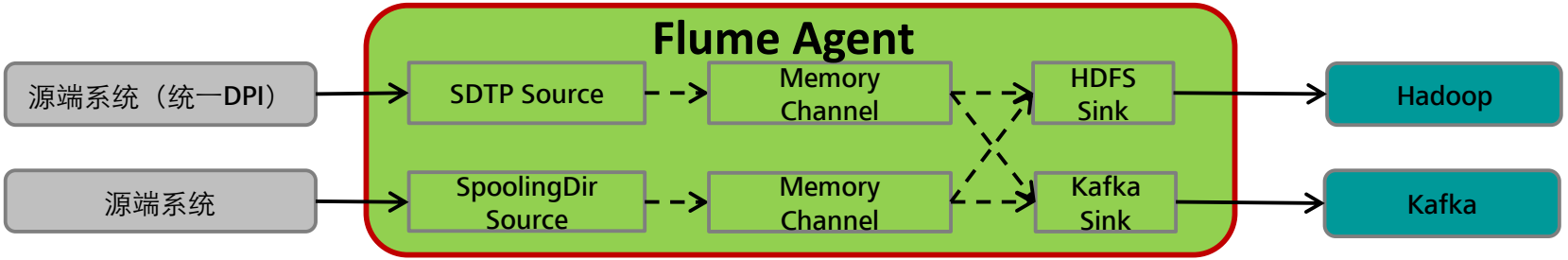
- 实时数据采集

实时采集原理





Flume是分布式、可靠、和高可用的海量数据聚合的系统，支持在系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。一个独立的Flume进程称之为Agent,包含流计算Source、Channel、Sink ,能支持常见的数据源协议和目标系统。



1.3.2 数据处理

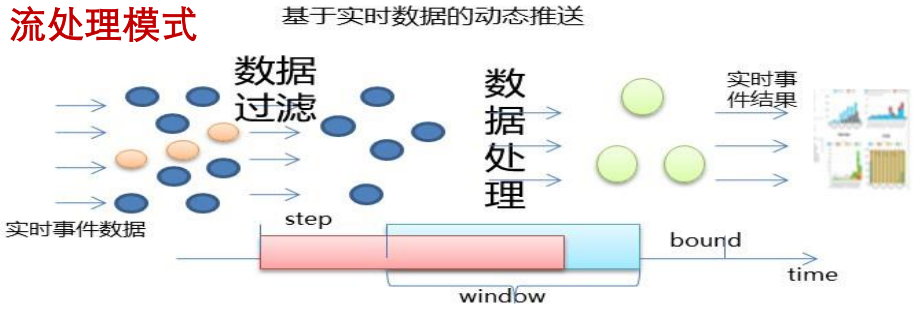
- 数据处理模式分类及适用场景

批处理模式

- 可用于传统数据库（DB2、Oracle等）或分布式数据库（Hadoop、MPP等）来实现；
- 支持结构化与非机构化数据的处理；
- 支持大量数据的处理需求；

适用场景：时效性要求不高，同时数据处理规模较大的场景，如传统报表分析和数据挖掘；目前大部分数据处理采用批处理模式；

流处理模式



- 支持流式数据的处理与计算；
- 处理时效性较高；
- 处理过程数据不落地；

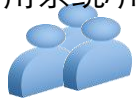
适用场景：针对数据处理结果需要高效地延迟的场景，如基于信令、位置等数据的实时营销或实时服务；例如XX移动企业级数据中心项目采用流处理模式进行用户行为信息的处理；

1.3.2 数据处理（续）

- 数据处理的方式转变：

库内计算

应用系统/用户



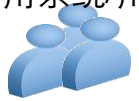
标准SQL请求/相应

内部处理：
SQL解析
任务调度
数据搜索定位
数据过滤、排序
压缩解压缩、加密
……

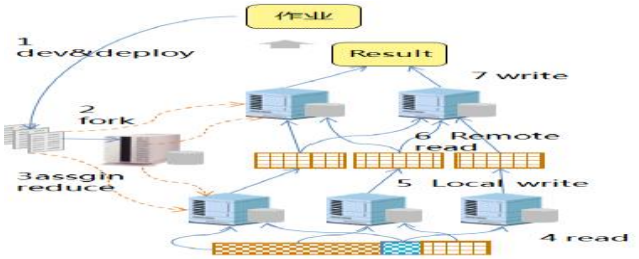
数据仓库

库外计算

应用系统/用户



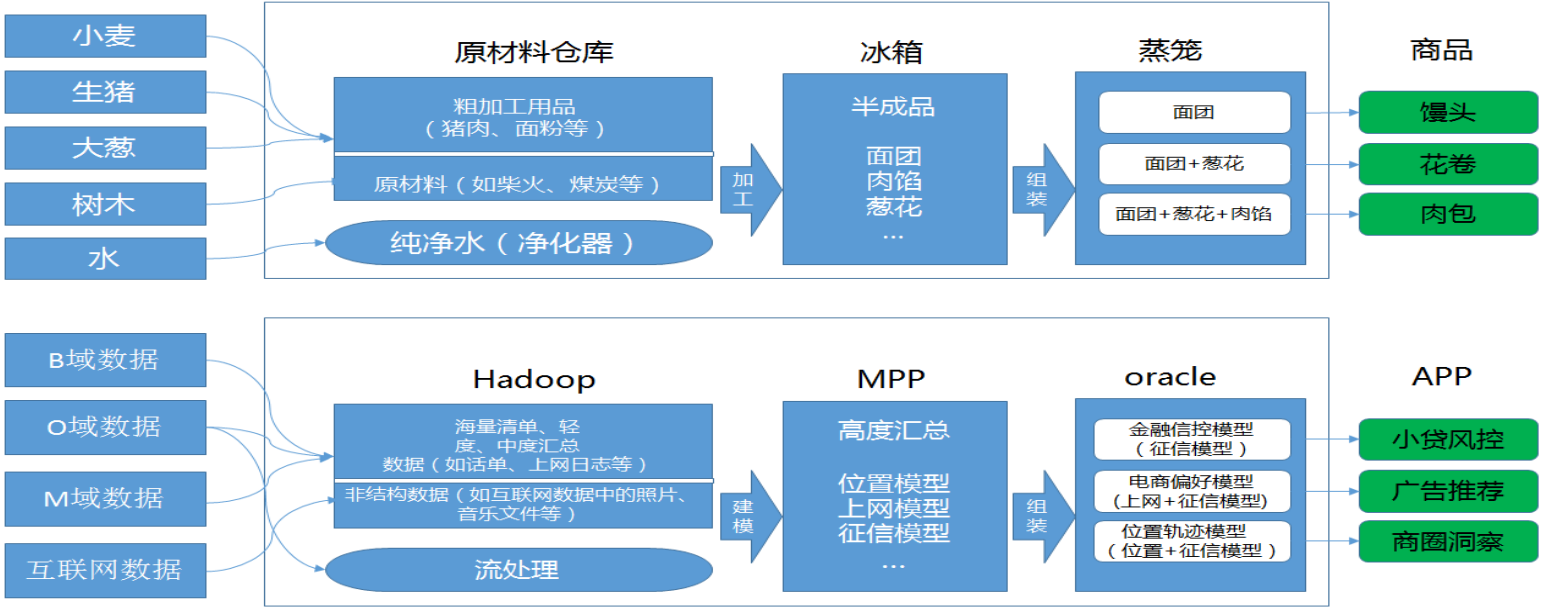
API请求/相应



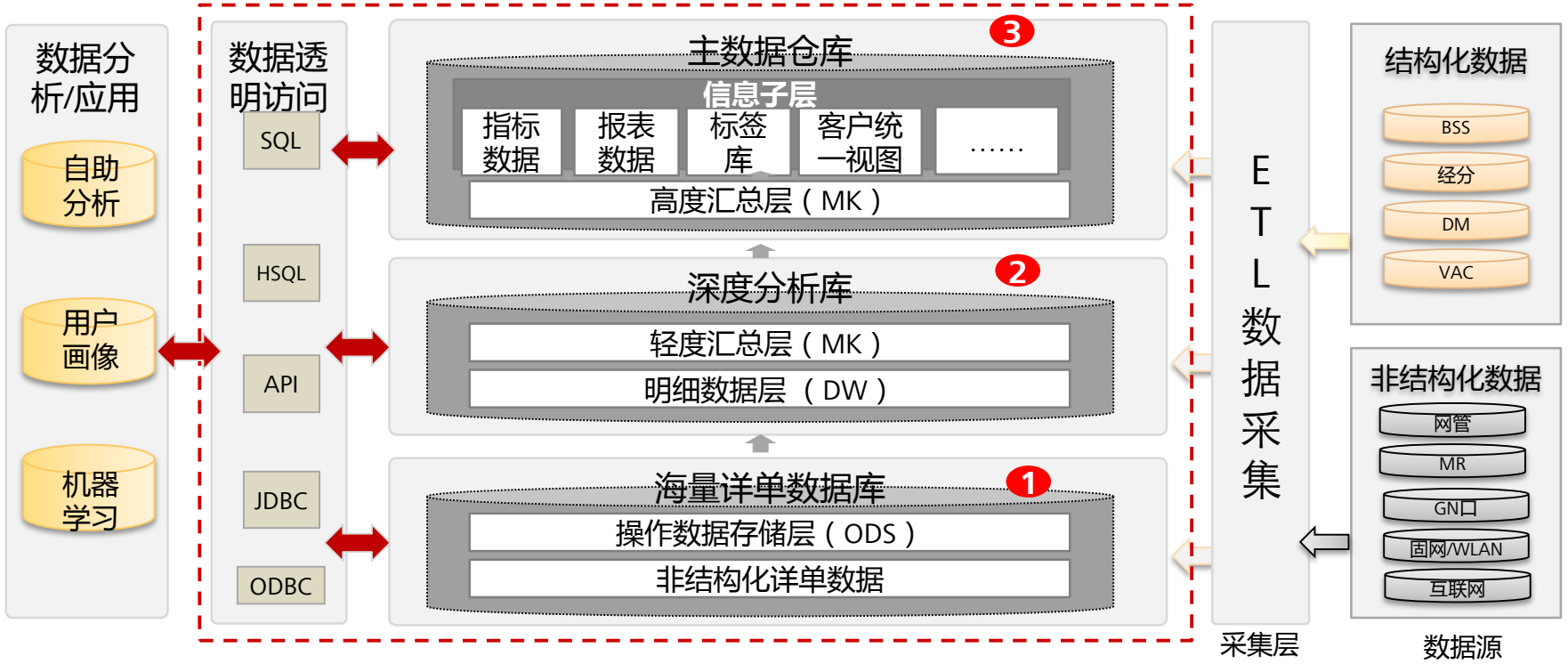
Hadoop集群

1.3.3 数据存储

- 如何理解大数据存储架构

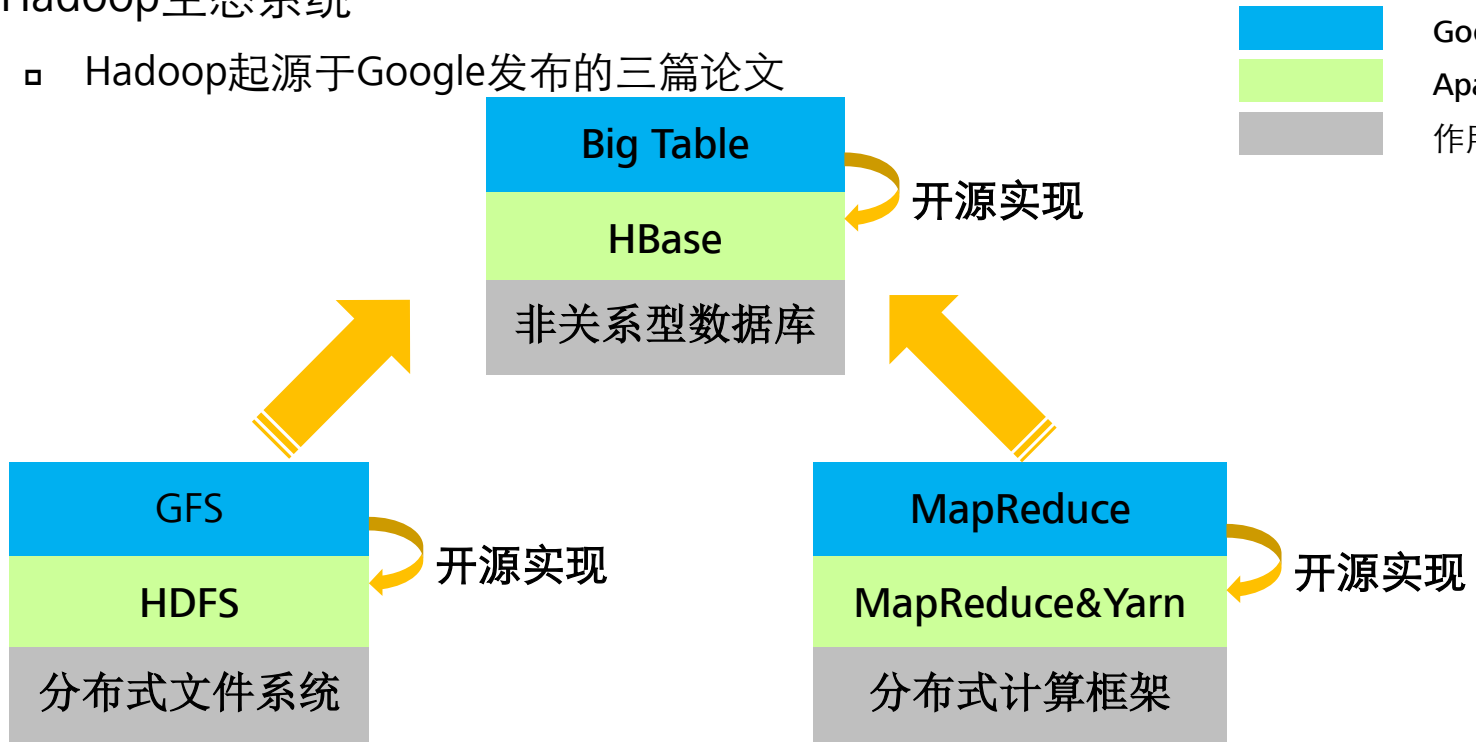


1.3.3 数据存储（续）

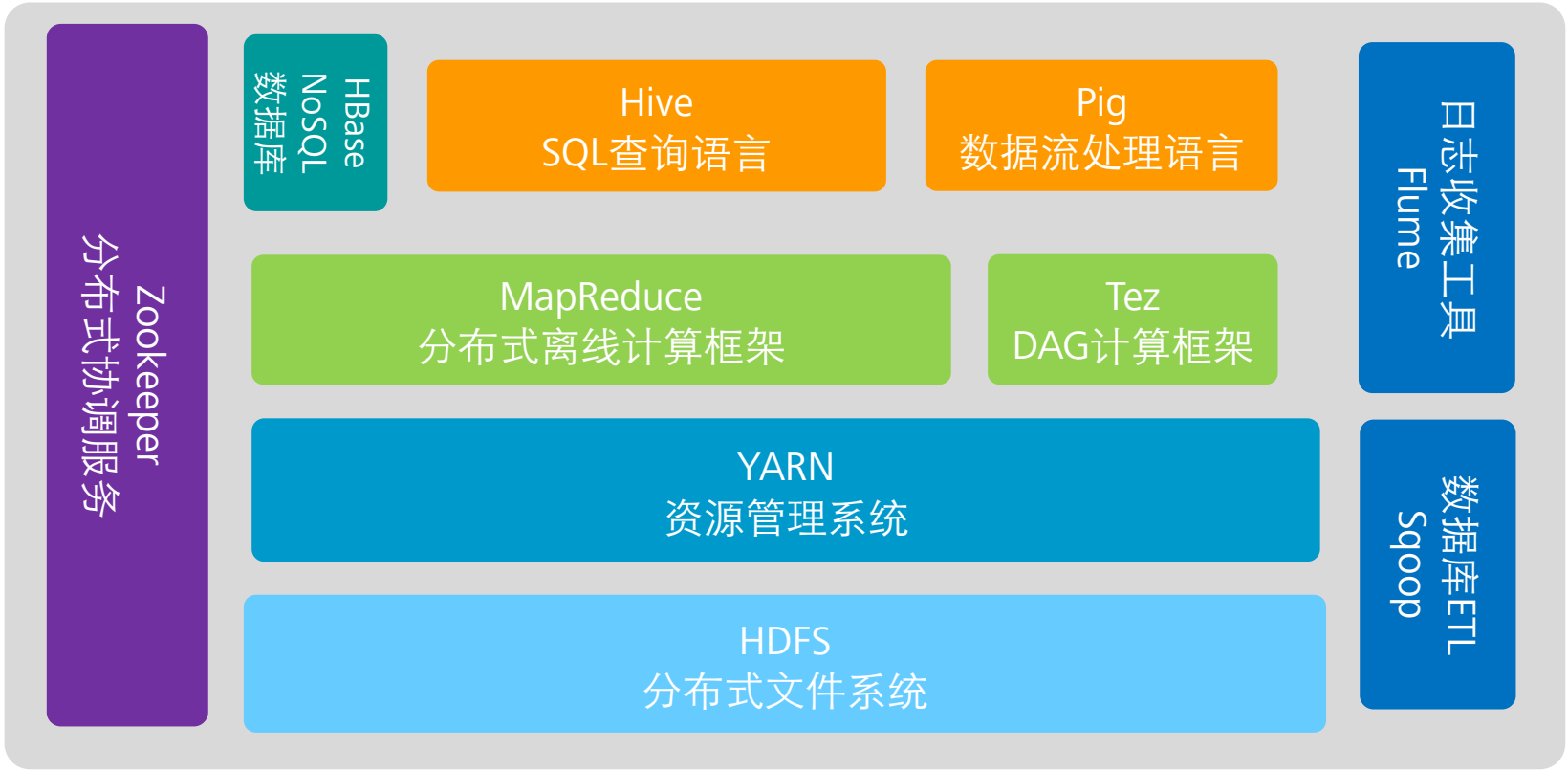

数据分析/应用
自助分析
用户画像
机器学习
数据透明访问
SQL
HSQL
API
JDBC
ODBC
主数据仓库
3
信息子层
指标数据
报表数据
标签库
客户统一视图
……
高度汇总层（MK）
深度分析库
2
轻度汇总层（MK）
明细数据层（DW）
海量详单数据库
1
操作数据存储层（ODS）
非结构化详单数据
ETL数据采集
结构化数据
BSS
经分
DM
VAC
非结构化数据
网管
MR
GN口
固网/WLAN
互联网
采集层
数据源

1.3.3 数据存储（续） - Hadoop

- Hadoop生态系统
  - Hadoop起源于Google发布的三篇论文

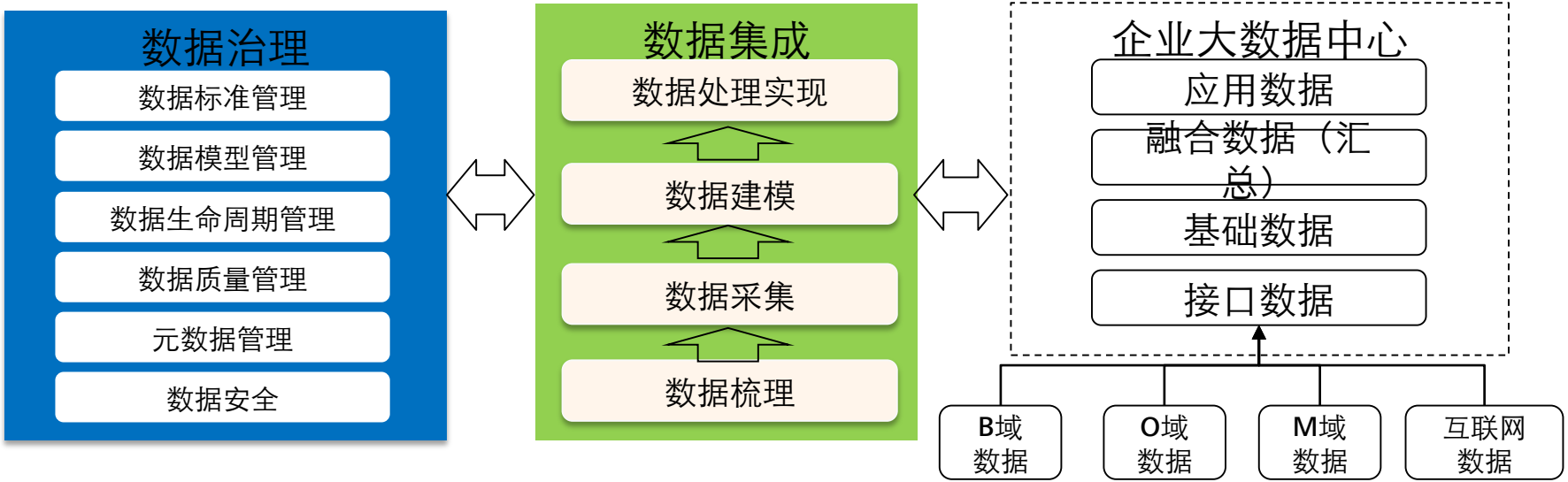


1.3.3 数据存储（续） - Hadoop



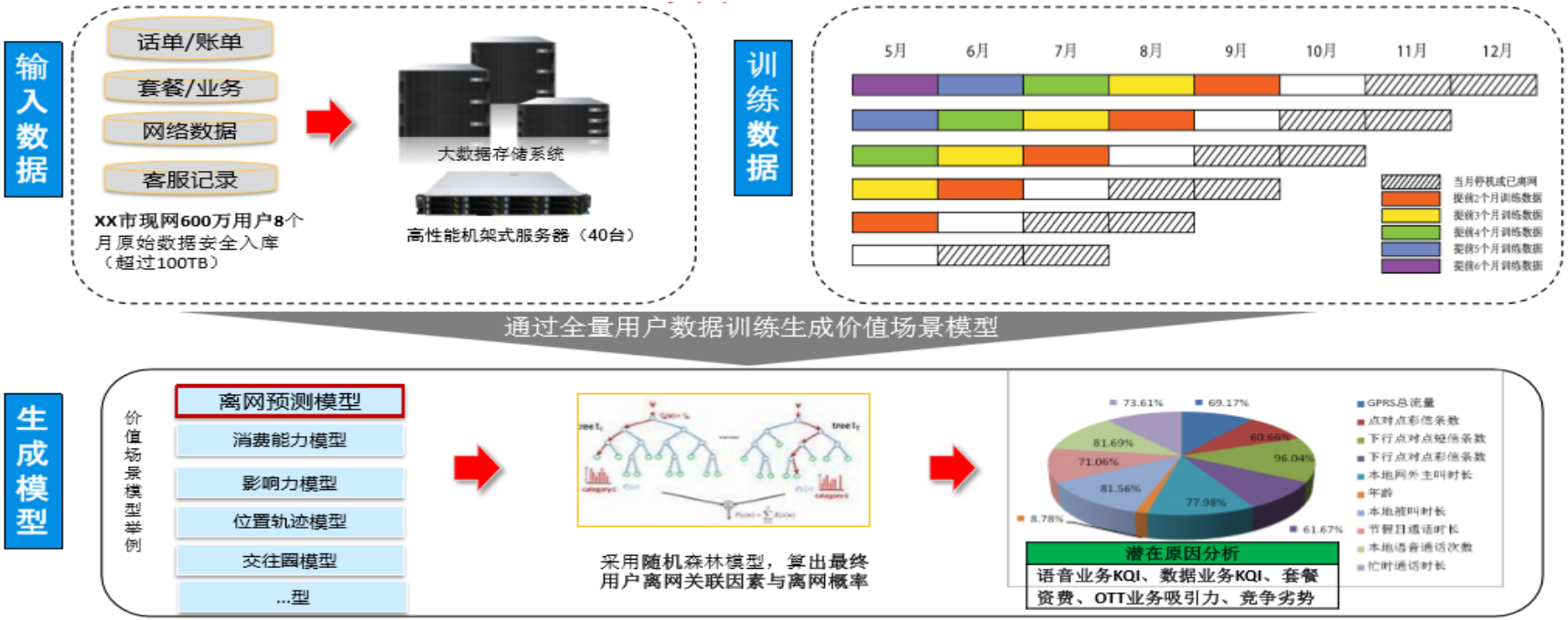
1.3.4 数据治理

- 在大数据系统中，只有制定了合理的标准和管理流程，才能有序、高效的进行数据的组织和使用。通过数据集成来实现数据的组织和生成，而数据集成的关键在于数据治理。



1.3.5 数据分析与挖掘

- 基于价值应用场景的大数据建模



1.3.6 数据应用

- 客户画像

用户位置行为：工作地、居住地、OD分析、出行方式等

用户通信行为：用户通话特征、短信特征、字冠分析等

移动基本电信属性：品牌、业务订购情况、UP值、性别、年龄等

互联网生活搜索数据：金融、商旅、酒店、票务等机构电话、位置等

用户细分

社会角色：

在校学生、教师、都市白领、医生、商旅人士、农民工、公务员等

聚类细分群组：

有车一族、上下班比较远用户、消费能力强用户、电信业务活跃用户等

业务关联群组：

根据业务特征关联性建立系统增强模型，如：虚拟网关联群组、亲情网关联群组等

1.3.6 数据应用（续）

- 实时营销

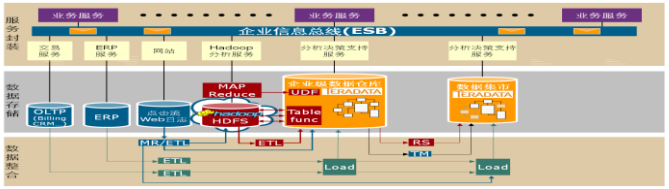
1.3.6 数据应用（续）

- 实时监控及热力图

目录

2.1 互联网企业大数据平台架构

割裂式混搭架构

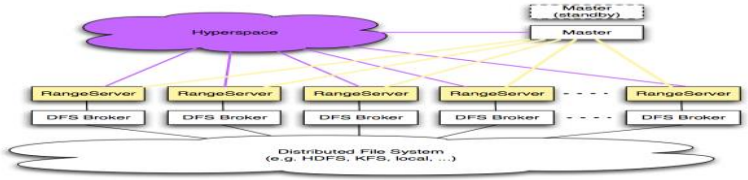


架构模式：Hadoop + MPP RDB /SMP RDB；
如：eBay , KDDI

混搭架构+深度定制化部件

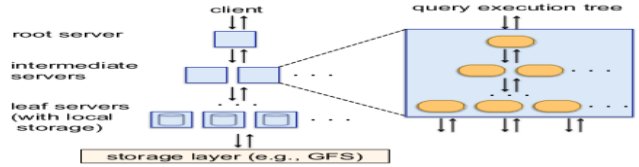
架构模式：
Hadoop + MPP RDB + NoSQL & MyFox /Prom /glider
如：阿里巴巴，淘宝系

Hadoop深度定制架构



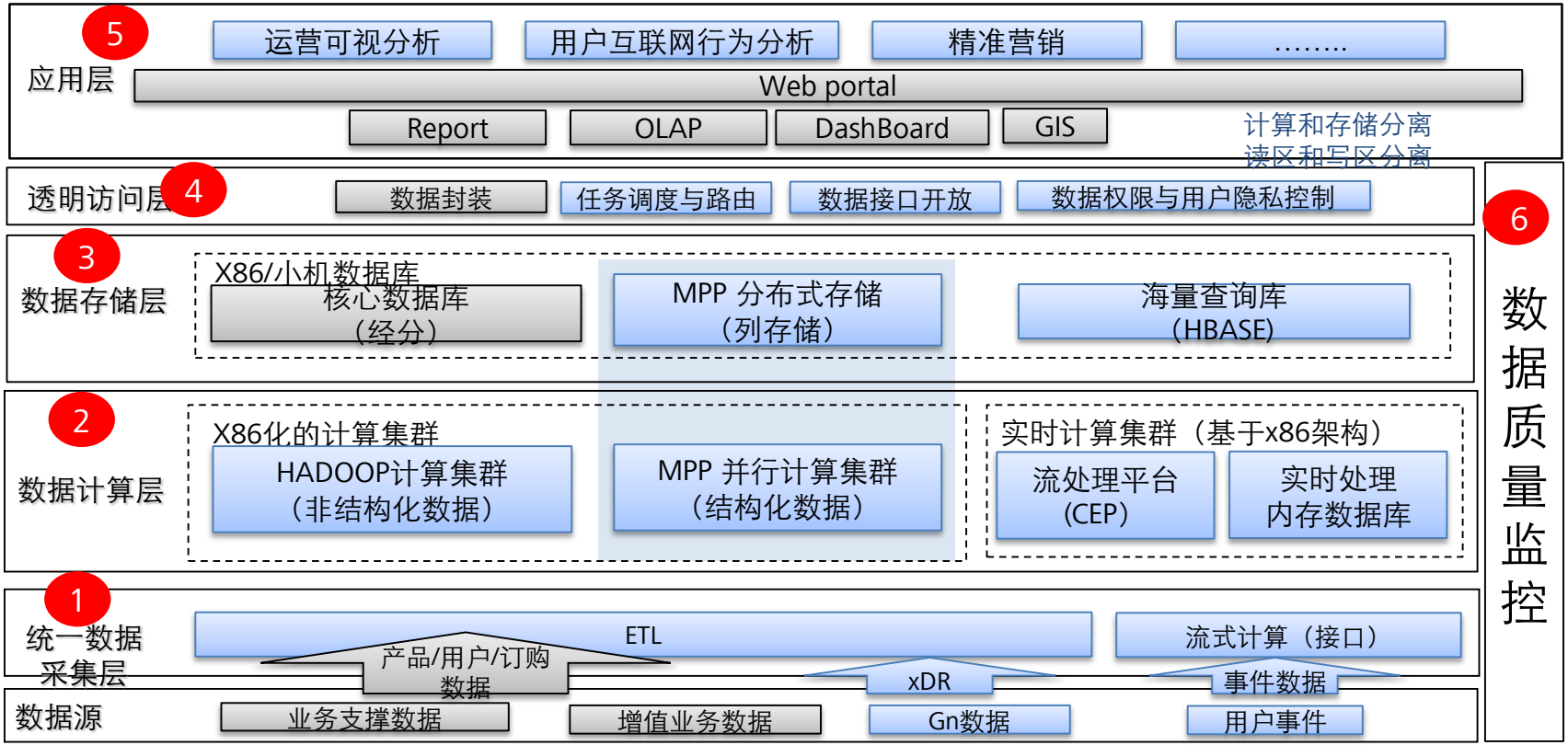
架构模式：Hadoop Enhanced；
如：腾讯、百度

自主研发架构



架构模式：Caffeine，Pregel，Dremel，Power Drill,Storm,Qubole,RCFile；
如：Google, Twitter，Facebook

2.2 结合运营商需求的混搭架构



总结

- 大数据平台整体架构
- 大数据管理流程及关键技术
  - 数据采集
  - 数据处理
  - 数据存储
  - 数据治理
  - 数据挖掘
  - 数据应用

Thank you

www.huawei.com