
QINGCLOUD青云
QINGCLOUD青云
IT大咖说
知识分享平台
实践课堂 / 第3季 / 深圳站
QingCloud Workshop Season 3 / Shenzhen
第1期 / 2017.4.22

QINGCLOUD青云

QINGCLOUD青云 | IT大咖说 知识分享平台

大数据平台架构实践 & 思考

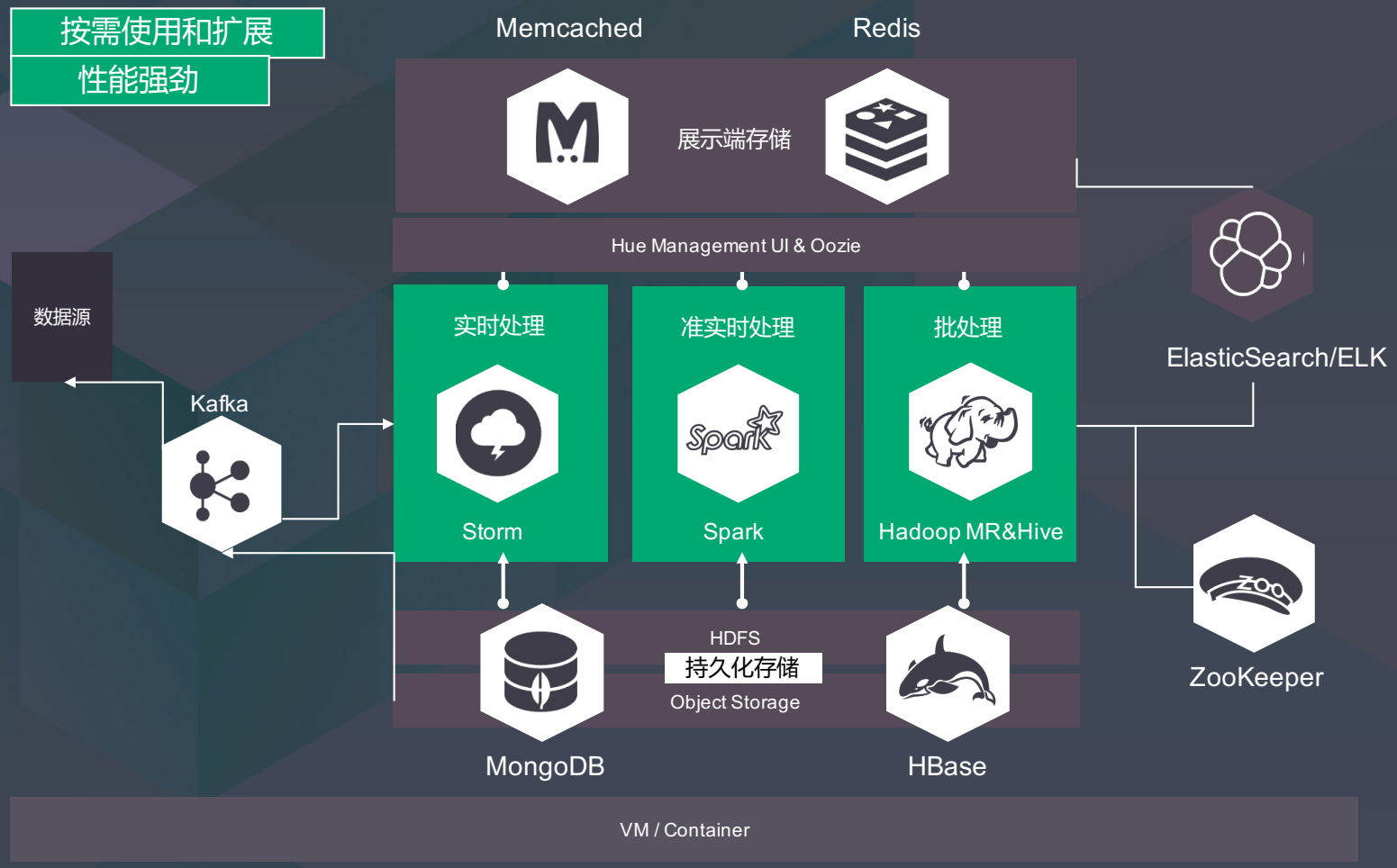
李威 I 青云QingCloud 大数据平台架构师

Agenda

- 青云基础技术架构
- 大数据平台架构演进
- 主流产品技术对比
- 计算与存储的关系
- 案例简介
- 大数据平台的 Roadmap

青云提供一个完整的基础架构云和技术平台云

完整的企业级大数据平台



实时流处理引擎对比

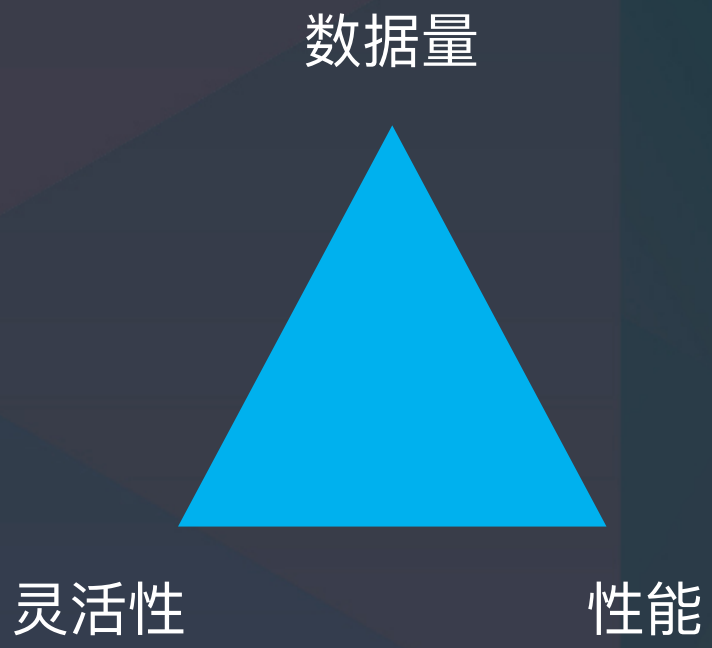
					
Streaming Model	Native	Micro-batching	Micro-batching	Native	Native
API	Compositional		Declarative	Compositional	Declarative
Guarantees	At-least-once	Exactly-once	Exactly-once	At-least-once	Exactly-once
Fault Tolerance	Record ACKs		RDD based Checkpointing	Log-based	Checkpointing
State Management	Not build-in	Dedicated Operators	Dedicated DStream	Stateful Operators	Stateful Operators
Latency	Very Low	Medium	Medium	Low	Low
Throughput	Low	Medium	High	High	High
Maturity	High		High	Medium	Low

存储 - HBase vs Cassandra

	HBase	Cassandra
一致性	基于 row de 强一致性	最终一致性，可平衡调整
稳定性	多 HMaster，Namenode HA	去中心化，没有单点故障
分区策略	主键有序排列的范围分区	一致性 Hash 排列，可自定义策略
可用性	Down 掉一台短暂不可读写	Down 掉可继续读写
依赖	ZooKeeper、HDFS	无

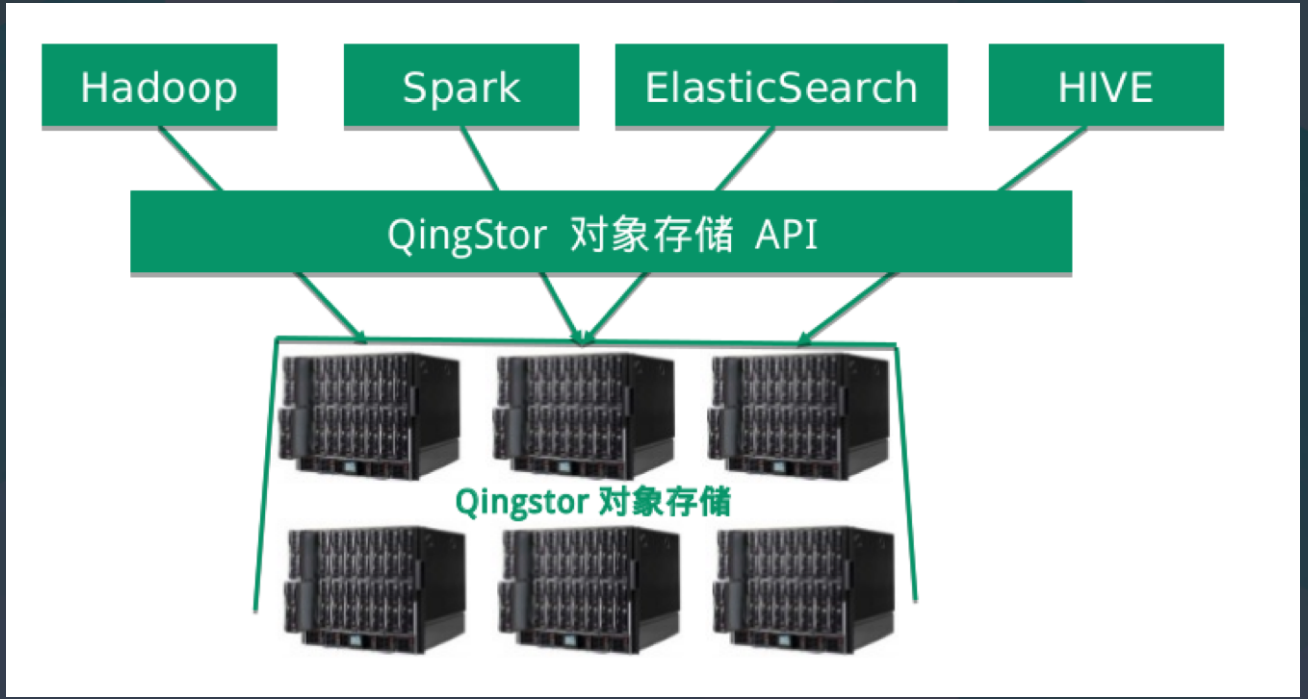
Ad-hoc & OLAP 查询分析产品对比

- Hive
  - 海量数据、查询灵活、性能低
- Phoenix+HBase
  - 海量数据、性能高、RowKey查询
- ElasticSearch
  - 查询灵活、性能高、T级数据
- Kylin
  - 海量数据、性能高、预聚合查询
- Druid
  - 海量数据、性能高、查询灵活、时序数据
- HashData (GreenPlum)
  - 查询灵活、性能高、结构化数据

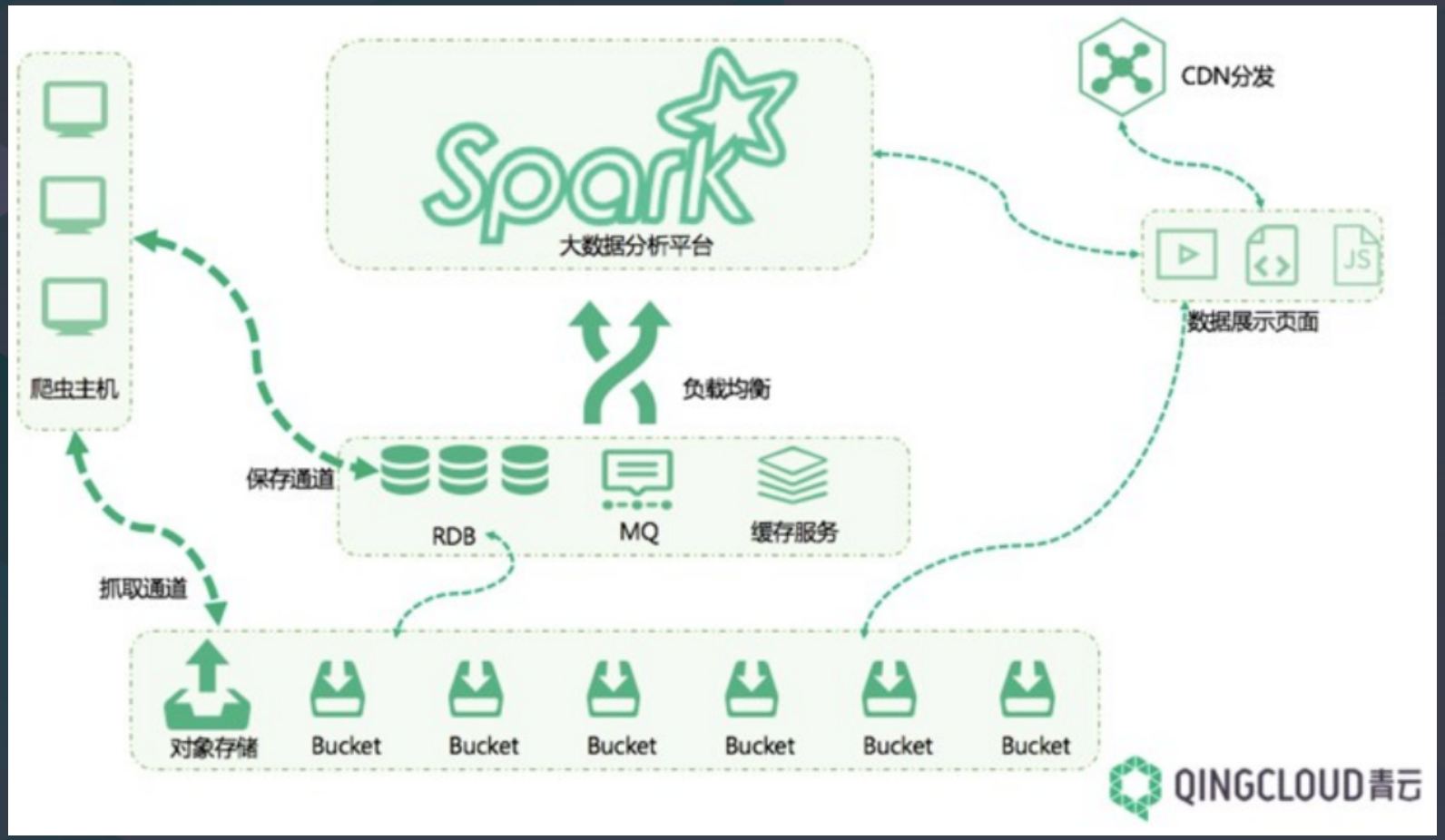


与 QingStor 对象存储无缝集成

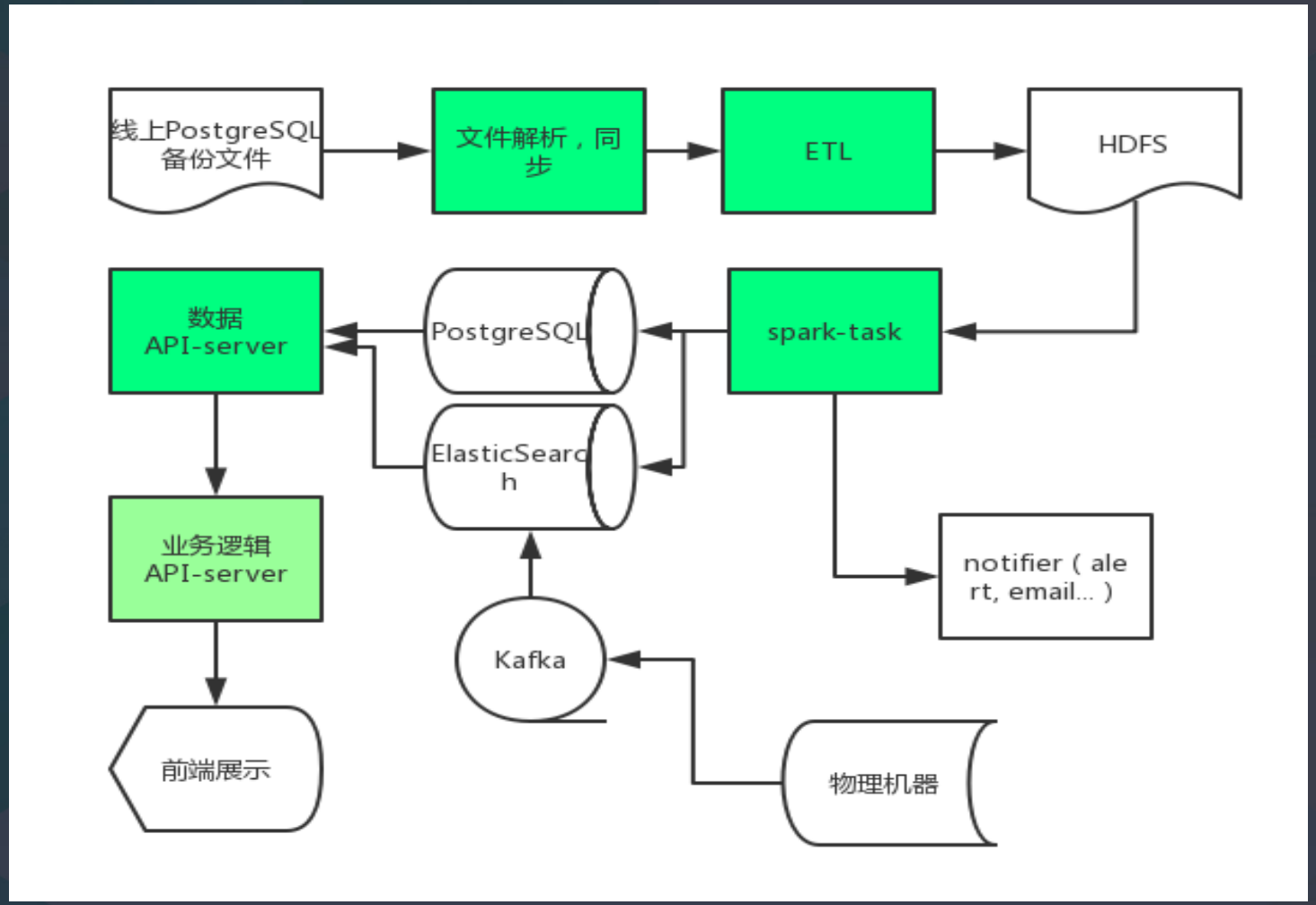
- Hadoop: HDFS <—-> S3
- Spark: w/wo HDFS & Standalone/Yarn
- ES: Backup & Restore



某大型家电集团 - 基于海量数据的舆情分析系统

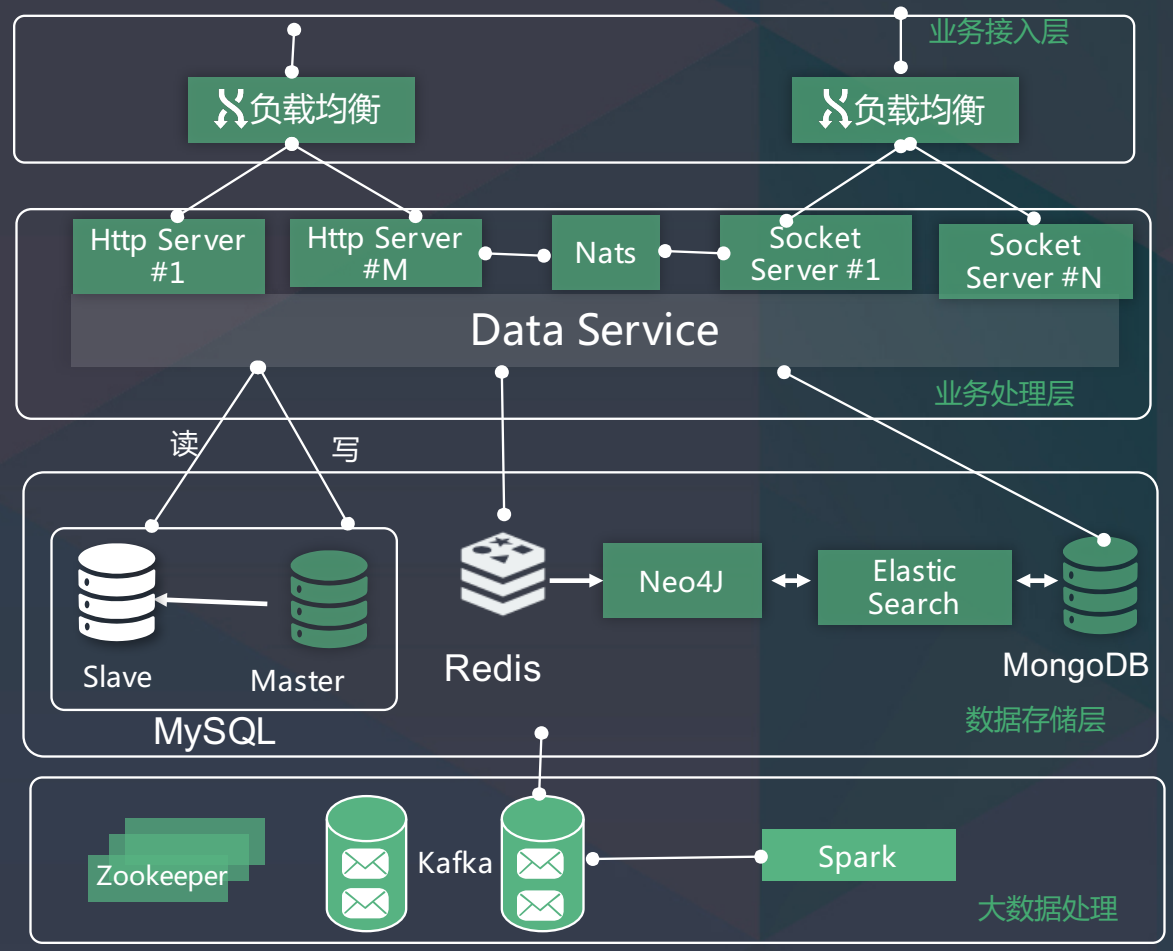


青云 - 在线业务大数据分析流程图

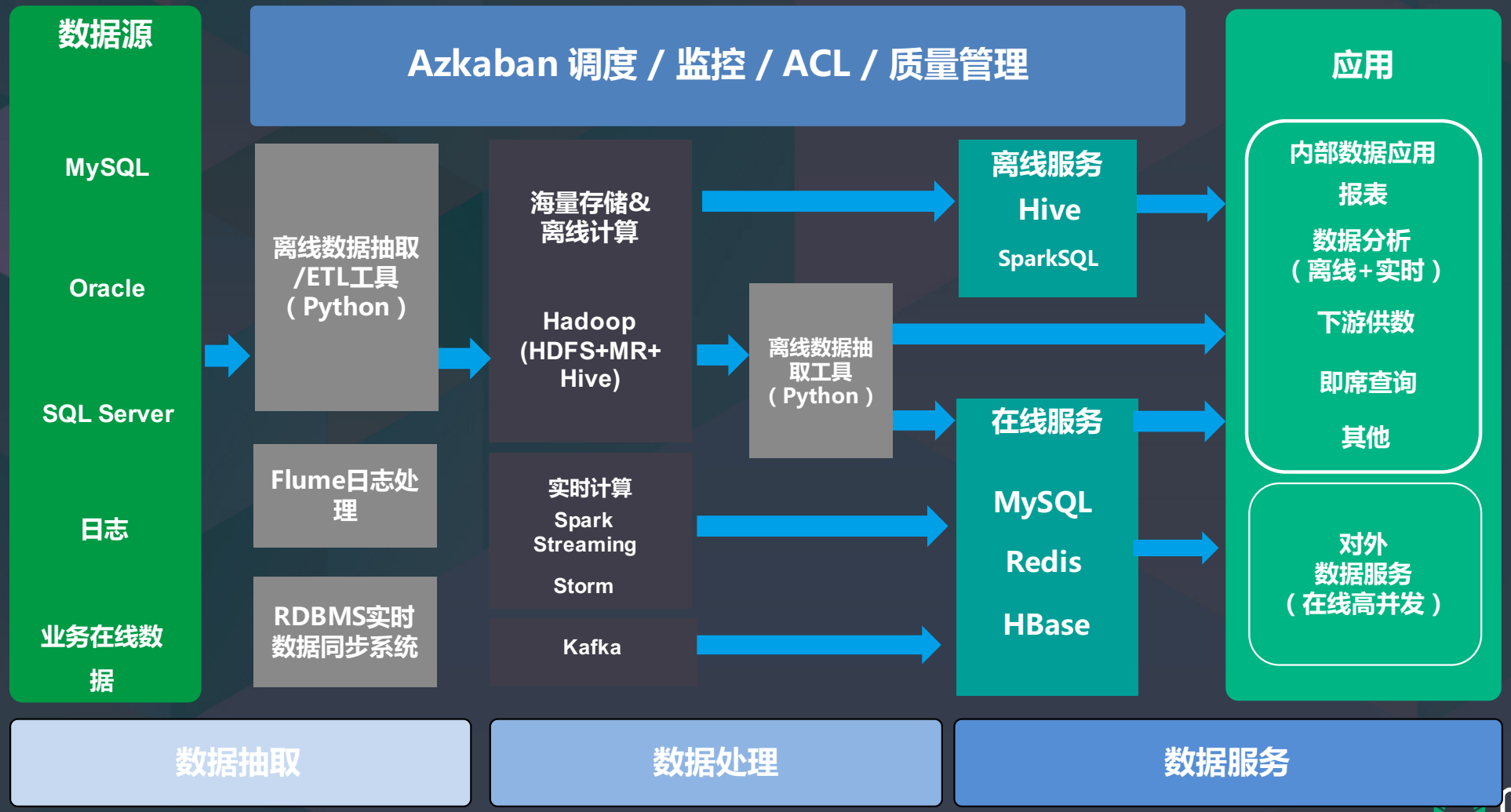


某大型互联网社交平台

- 秒级创建：可构建快速弹性伸缩的后端系统，灵活应对高并发的业务访问压力。
- 监控报警：自定义时间段监控云内的 IaaS 以及 PaaS 资源,灵活定义监控项以及监控周期
- 负载均衡器集群：可构建一个能支持高并发读写、高可用的后端IT架构。
- 大数据平台构建：应用及系统日志等信息通过 Kafka 和 MongoDB 输入到 Spark 平台上进行分析如:对于系统推荐的好友,用户是否添加了好友等。



某创新型综合金融公司

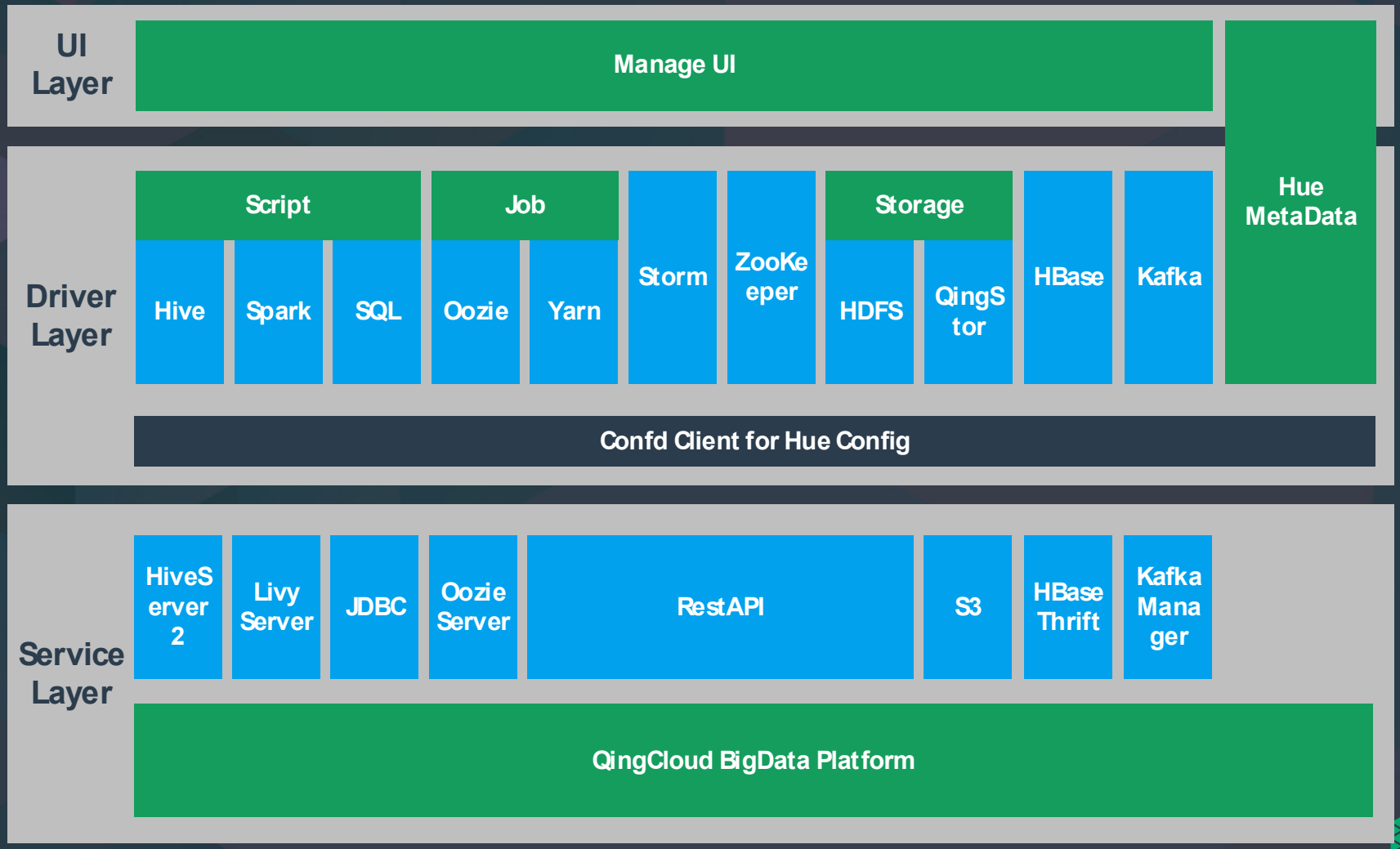

数据源
MySQL
Oracle
SQL Server
日志
业务在线数据
Azkaban 调度 / 监控 / ACL / 质量管理
离线数据抽取/ETL工具（Python）
Flume日志处理
RDBMS实时数据同步系统
海量存储&离线计算
Hadoop (HDFS+MR+Hive)
实时计算
Spark Streaming
Storm
Kafka
离线数据抽取工具（Python）
离线服务
Hive
SparkSQL
在线服务
MySQL
Redis
HBase
应用
内部数据应用
报表
数据分析（离线+实时）
下游供数
即席查询
其他
对外数据服务（在线高并发）
数据抽取
数据处理
数据服务

其他大数据案例

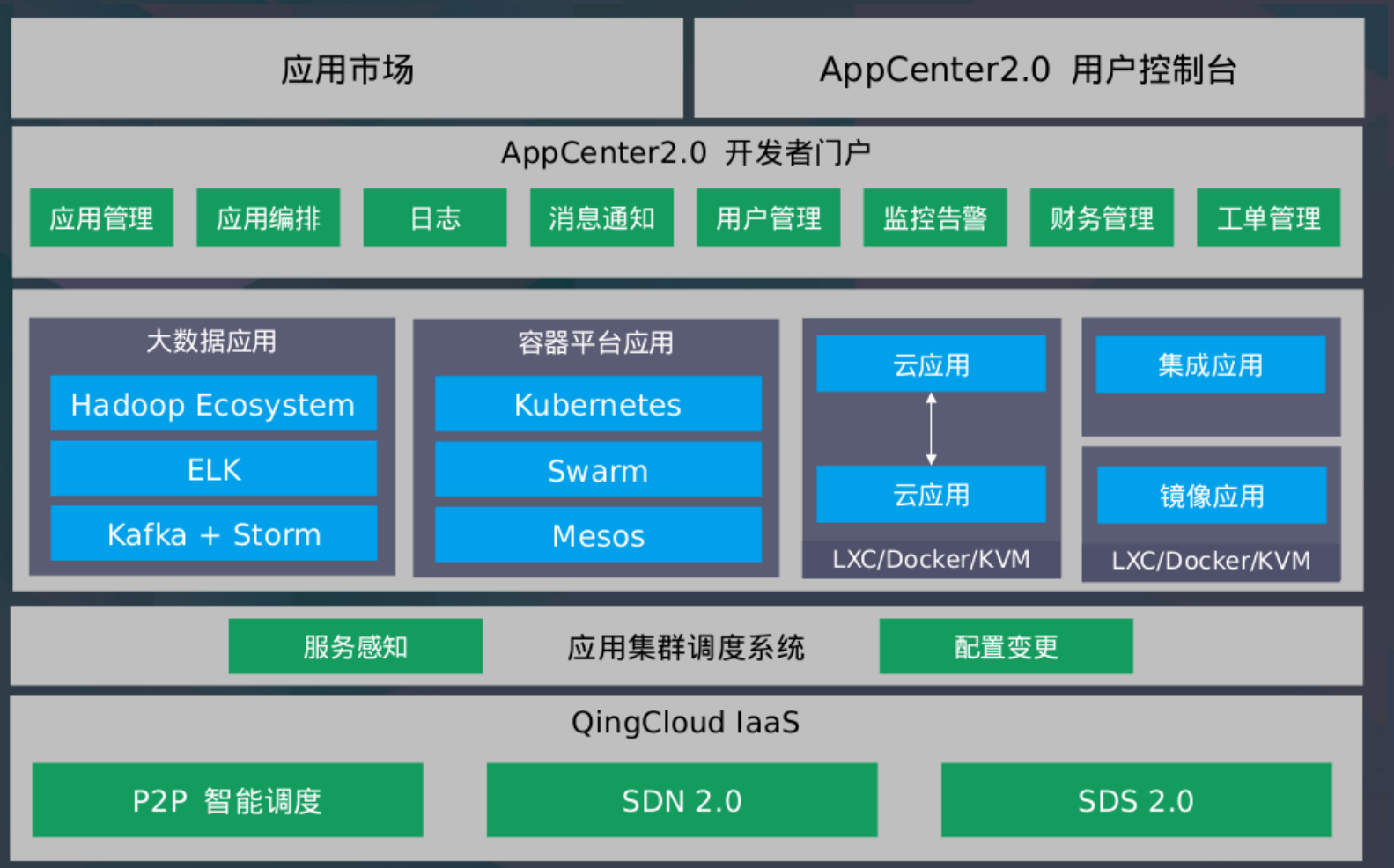
- 某政府部门——人特征信息查询（Kafka + ElasticSearch 集群）
- 某家电集团——海量互联网数据挖掘&分析（上百亿小文件+60 节点 ElasticSearch 集群 + 20 节点 HBase 集群）
- 某生物医药公司——生物基因检测（27 节点 Hadoop 集群 +27 节点 Spark 集群）
- 某智能医疗公司———移动&可穿戴设备海量数据收集分析（ZooKeeper + Kafka + ElasticSearch + Spark + Sqoop + HBase 集群等）

.....

Roadmap - 大数据平台管理架构



Roadmap - 大数据平台 + AppCenter 2.0



What's Next?

Kylin
Flink
TensorFlow
Caffe
Cassandra
Flume
Solr
Neo4j
Druid
.....