中国大数据分析平台行业研究报告

摘要

行业界定：大数据分析平台逐渐由产品态转向集成态，行业边界模糊。在技术架构上，主要包含数据采集与存储、计算、分析与决策三个层级。在 OLAP 之上融合了深度学习等技术，在提升数据分析深度和广度的同时，也极大增加了数据服务在业务侧的低门槛和友好性，满足用户运用数据分析驱动业务发展的需求。

市场情况：尽管行业边界泛化，市场参与者众多，但按照部署模式、架构分类及能力补给，可分为以下五类：1）以云上数据湖方案为主的公有云厂商；2）以本地化大数据分析平台为主的传统软件服务商；3）提供轻量化数仓架构的数据库/数仓厂商；4）为数据应用层提供服务能力的软件供应商；5）提升数据应用能力的人工智能厂商。行业市场整体呈现竞合状态。

架构选型：搭建平台前用户首先需要明确自身的数据体量和业务场景需求。在明确大数据分析平台需要具备的基本功能后，再决定平台搭建过程中使用的大数据处理框架和工具。在分层架构中，数据分析层的组件选型和整体搭建十分关键，尤其是存储引擎的选型直接决定了离线、在线、实时三大场景的支撑和算力效率的高低。

趋势洞察：传统架构下的湖仓分体引发数据孤岛，造成实施、运维和成本问题。湖仓一体架构在数据和查询层面形成一体化架构，突破实时性和并发度、集群规模受限、非结构化数据无法整合、建模路径冗长、数据一致性弱等瓶颈。同时，平台融合 AI 自主学习和自适应能力，增强用数人员的分析和决策能力。

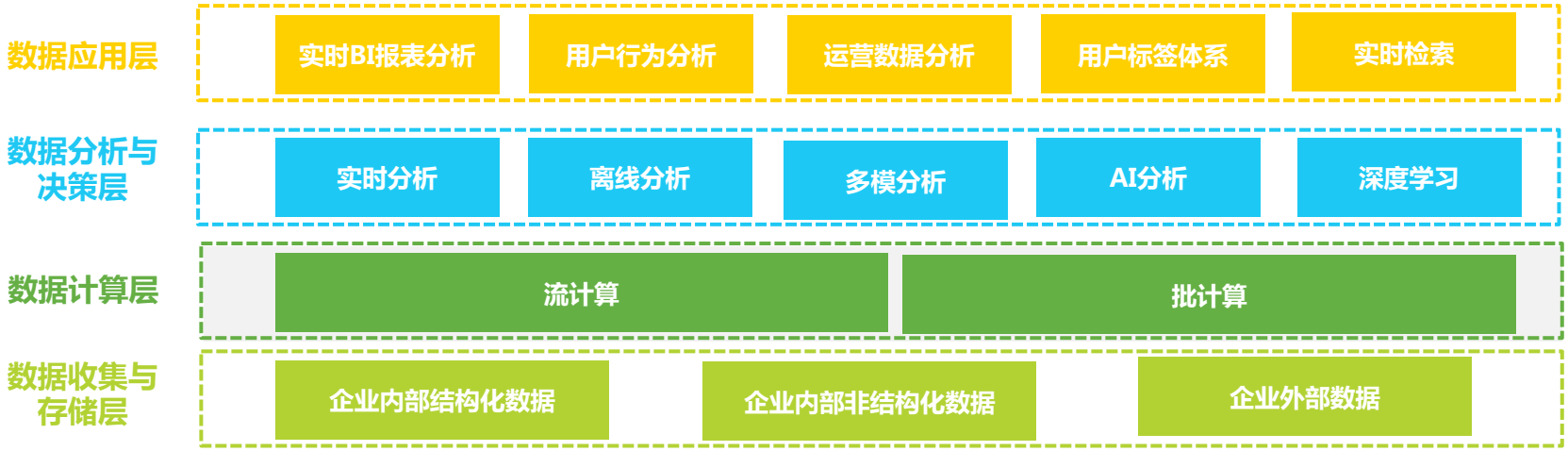
来源：艾瑞咨询研究院自主研究绘制。

行业界定

驱动业务的全场景数据分析平台，提供实时、多维的数据分析和智能决策

大数据分析平台，是企业用户在大数据环境下用于分析与决策的平台。按技术架构划分，主要包含数据收集与存储、数据计算、数据分析与决策三个层级。从服务边界来看，大数据分析平台概念小于数据中台，强调平台的数据分析与决策能力，弱化了数据本身的规划、治理与服务；在 OLAP 之上，又融合了深度学习等技术，在提升数据分析深度和广度的同时，也极大增加了数据服务在业务侧的低门槛和友好性。企业通过构建大数据分析平台，聚拢各业务系统数据，打通全渠道组织各业务维度，用数据分析驱动业务，满足企业级宽表实时分析、实时 BI 报表分析、用户行为分析、自助分析、 AI 智能分析等全方位需求。

大数据分析平台技术框架及核心组件

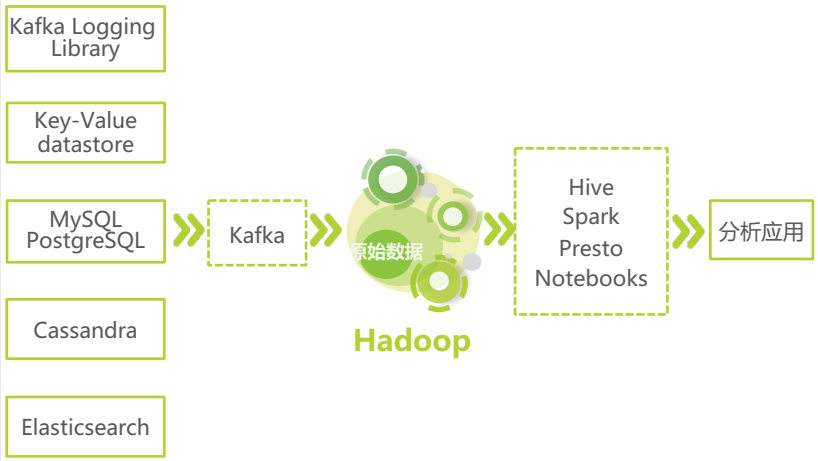


来源：艾瑞咨询研究院自主研究及绘制。

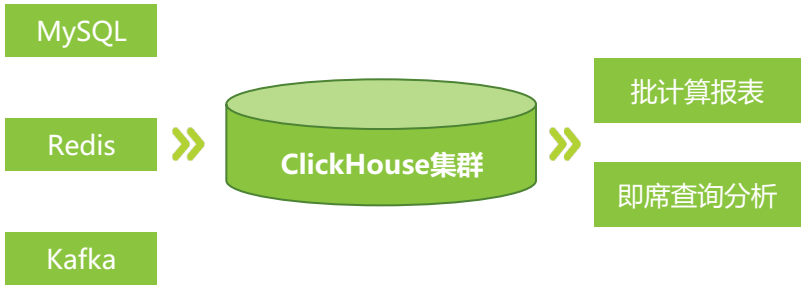
技术沿革（一）

平台技术架构持续更新迭代，由离线处理向实时分析演进

基于Hadoop的分析架构



数据库结合AP分析引擎架构（以ClickHouse为例）



架构剖析

- **基于 Hadoop 分析架构的流程原理：**
  各类结构化数据通过采集管道进入 Kafka，Spark 实时消费 Kafka 的数据，写入集群内的 HDFS，RDS 数据库中的数据通过 Spark 每天一次全量扫表同步至 HDFS。HDFS 存储汇总用户数据，对数据库数据定期执行 snapshot。
- **基于 Hadoop 分析架构的优缺点：**
  优点：借助 Hadoop 集群的高并发能力，实现百 TB 到 PB 级数据的离线计算和处理，同时数据存储在 HDFS 上，存储成本低。
  缺点：数据定期入库，数据计算的时效性通常是 T+1。

来源：艾瑞咨询研究院根据公开资料整理及绘制。

架构剖析

- **数据库结合 AP 分析引擎架构的流程原理：**
  将平台架构引入 TP 引擎结合 AP 引擎实现实时分析平台，各类结构化数据同步至分析引擎后便可进行交互分析。
- **数据库结合 AP 分析引擎架构的优缺点：**
  优点：舍弃了传统离线大数据架构，实现实时批量计算，在 GB 到100TB 级别的计算有了很大提升，BI 人员无需等待 T+1的离线计算后得到最终结果，大幅提升数据资产的商业价值。
  缺点：在处理百 PB 级以上数据时，ClickHouse 架构的扩展能力、复杂场景计算和存储成本相对 Hadoop 方案较弱。

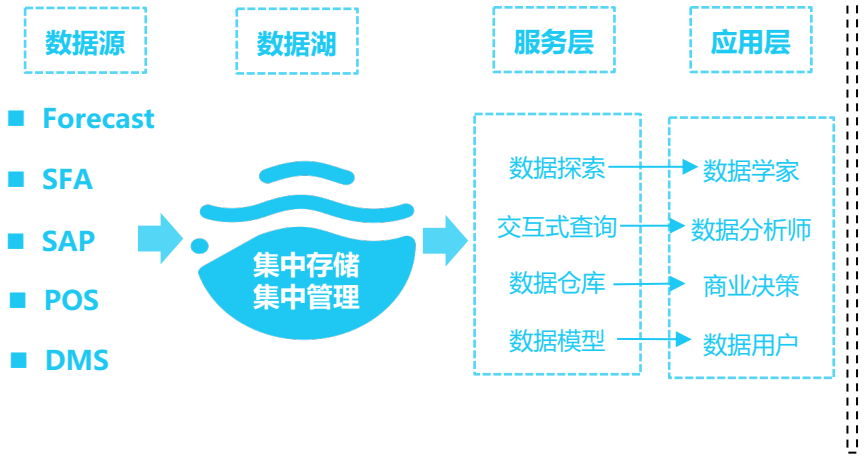
来源：艾瑞咨询研究院根据公开资料整理及绘制。

©2022.12 iResearch Inc. www.iresearch.com.cn

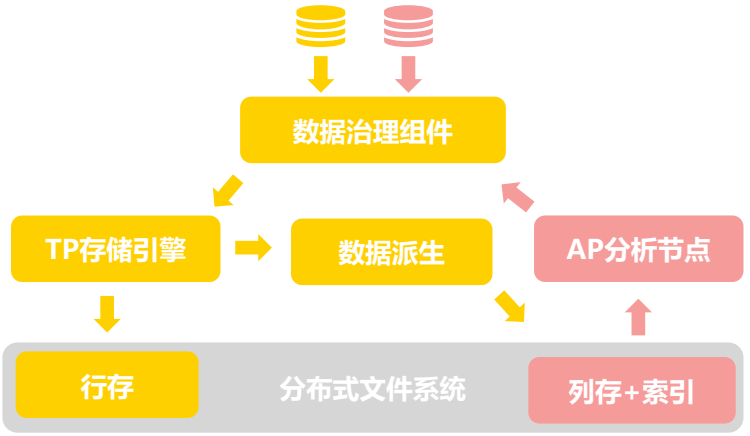
技术沿革（二）

平台技术架构持续更新迭代，产品在云上落地和升级

基于云上数据湖的分析架构



基于云原生结构化存储引擎的分析架构



架构剖析

- **基于云上数据湖的分析架构的流程原理：**
  可理解为借助云原生存储引擎，基于传统 Hadoop 方案的云上落地和升级，保留自建 HDFS 集群的分布式存储可靠性和高吞吐能力，借助数据湖降低传统方案的运维和存储成本。
- **基于云上数据湖的分析架构的优缺点：**
  优点：对大数据平台的使用者做了区分和定义，针对不同的使用场景，数据的使用方式，分析复杂度和时效性也会有不同。
  缺点：数据湖方案本身并没有解决传统方案的所有痛点。

来源：艾瑞咨询研究院根据公开资料整理。

架构剖析

- **基于云原生结构化存储引擎的分析架构的流程原理：**
  将类似第二阶段和第三阶段的融合，在线库和分析库隔离，不依赖在线库数据；全量数据支持高效批量计算，分析结果集支持即席查询，支持实时写入实时流计算。
- **基于云原生结构化存储引擎的分析架构的优点：**
  优点：在具备宽表合并高吞吐低成本存储的同时，可以提供 TB 级别数据即席查询和分析的能力，无需过度依赖额外的计算引擎，实现高效实时分析能力。

来源：艾瑞咨询研究院根据公开资料整理。

核心应用

流批一体：统一开发、统一计算、逻辑一致、降低成本

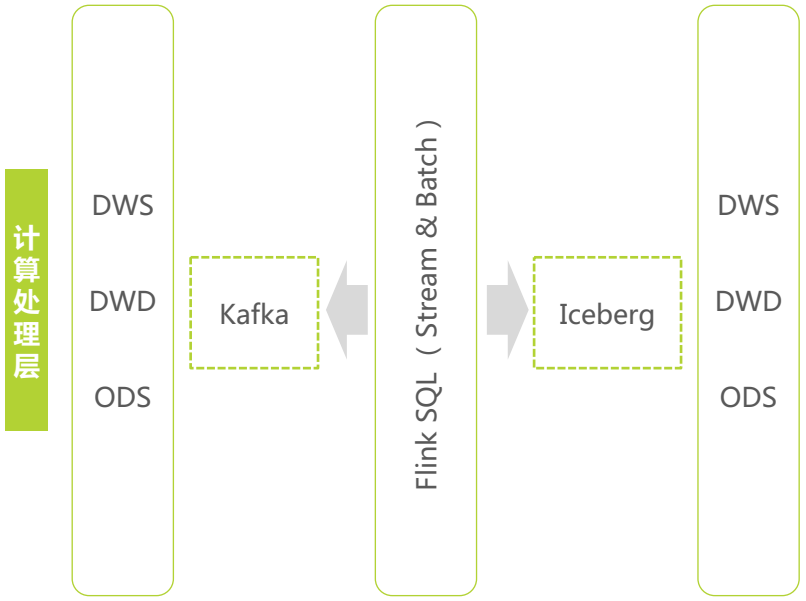
采用流计算+交互式分析双引擎架构，流计算负责基础数据，交互式分析引擎是中心，流计算引擎对数据进行实时 ETL 工作，与离线相比，降低了 ETL 过程的 latency。交互式分析引擎自带存储，通过计算存储的协同化，实现高写入 TPS、高查询 QPS 和低查询 latency，从而做到全链路的实时化和 SQL 化，实现用批的方式做到实时分析和按需分析，并能快速响应业务变化，两者配合实现1+1>2的效果。流批一体实现了建立一套统一的系统，由同一个开发团队开发，同时支持流式计算和批量计算，提供一致的编程环境，降低开发和运维成本，减少资源浪费，提高数据口径的一致性。

流批一体的技术趋势及行业应用

数据存储层面的湖仓一体	将数据仓库的便捷管理、高质量数据特性与数据湖的可探索、高灵活性相融合，完成数据仓库到湖仓一体的转变，实现对未知数据先统一存储再灵活探索的能力，在数据架构层面更进一步。
分析应用层面的工具生态	广义的流批一体包含存储一体（湖仓一体）、计算一体（即狭义流批一体）、分析应用一体（数据同步、数据血缘等生态），目前分析应用部分还处于相对早期阶段，不论是成熟型厂商还是新兴型厂商都在积极探索。
行业应用	目前已在工业、农业、能源、金融、零售、教育、传媒等行业落地应用，极大推动了实时计算，提升数据价值，节省计算资源。

来源：艾瑞咨询研究院自主研究及绘制。

流批一体的技术框架



来源：艾瑞咨询研究院自主研究及绘制。

©2022.12 iResearch Inc. www.iresearch.com.cn

核心产品（一）

商业智能 BI：通过数据整合分析实现商业价值

商业智能（BI，Business Intelligence）是大数据分析最典型应用领域，是由数据库、数据仓库、数据湖、湖仓一体、ETL、OLAP、数据挖掘、机器学习和人工智能等技术组成的一套完整解决方案。随着大数据处理技术的发展，商业智能的洞察和分析能力进一步提升，数据分析和可视化的门槛不断降低，企业实现不同层级的拖拽式自助分析和多种类型的图表展示，并在统一平台进行整合和共享，获得不同层级的数据洞察，最终用于商业决策。机器学习和人工智能在商业智能中扮演越来越重要的角色。

BI 的技术发展趋势

BI SaaS化	云上落地是商业智能最大的技术发展趋势
一站式平台化	商业智能趋于集成数据仓库提供存储功能，集成 python 及 R 语言提供数据挖掘，延伸范围越来越广
BI 与新技术融合	商业智能与流程自动化 RPA 和人工智能等新技术深度融合

BI 的行业应用及典型企业