QCon+ 案例研习社

笑谈天下库，煮酒论引擎

——湖、仓、OLAP引擎面面观

郭炜（郭大侠）

Apache Software Foundation Member
ClickHouse中国社区创始人
Apache DolphinScheduler PMC
Apache SeaTunnel（incubating） Mentor

OLAP、数据湖、数仓谁主沉浮？

ICEBERG

Apache hudi

Greenplum

ClickHouse

DORIS

DATA FUSION

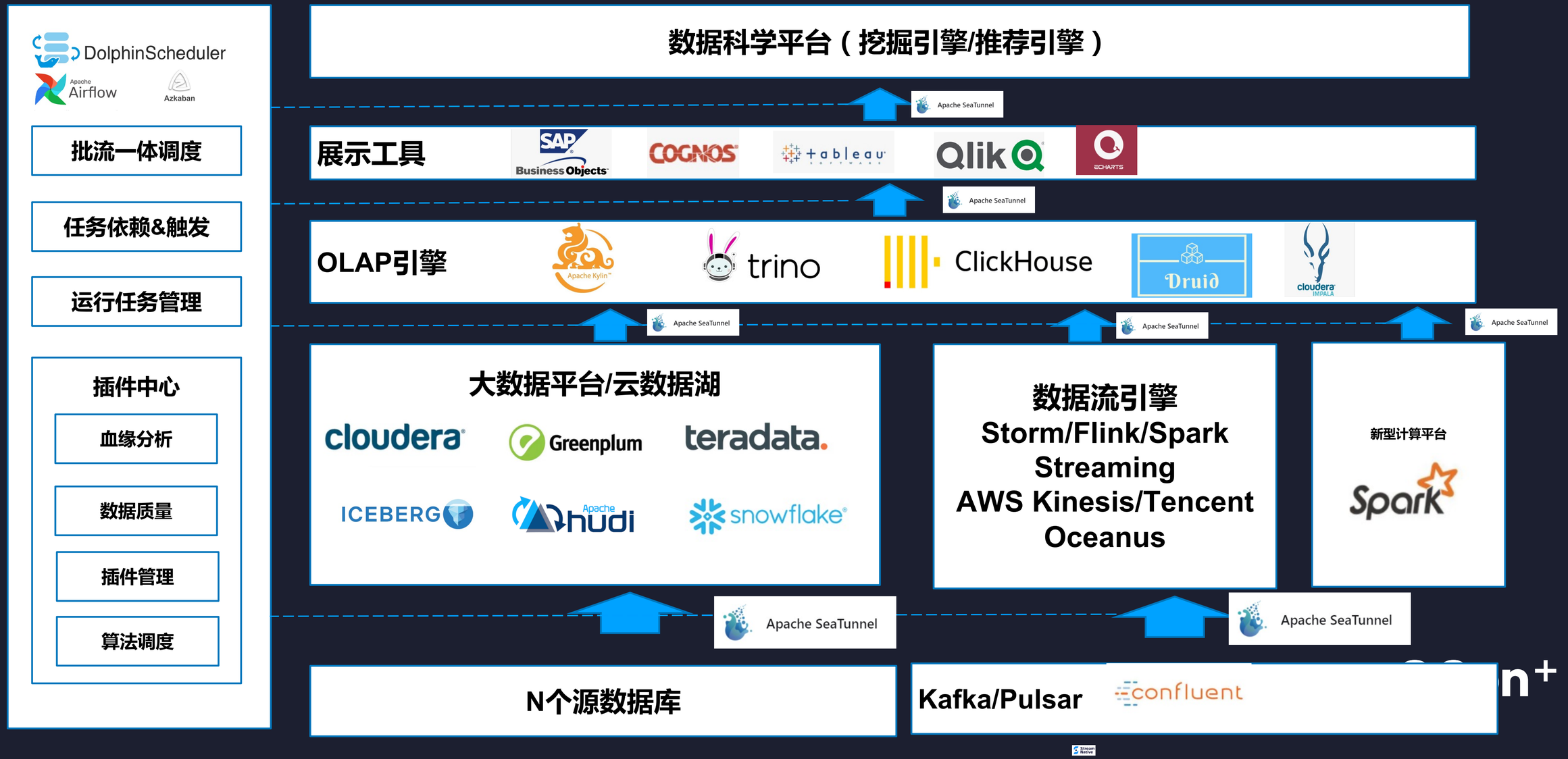
……？

煮酒论引擎仅代表个人观点

QCon+

湖、仓、OLAP 一个都不能少

——常见的企业大数据架构



OLAP 引擎各自特点 —— Kylin

Apache Kylin 多维数据库，擅长高并发确定性查询指标

最新发布的 Metrics store 指标中台用户很多

不擅长的场景：即席查询、临时查询

OLAP 引擎各自特点 —— ClickHouse

著名的向量计算引擎数据库，擅长单表复杂查询，配有 Projection 预计算和针对日志场景大量函数和相关功能，适应超大规模计算

不擅长的场景：Join 场景、高并发指标查询、手动挡开不起来的…

OLAP 引擎各自特点 —— Doris

国内开源大数据计算 OLAP 引擎，具备完备的分布式管理框架和分布式查询层，运维成本较低

不擅长的场景：超大规模计算，性能还在追赶中

仓引擎各自特点 —— Greenplum

开源最稳定的数据仓库之一，支持 SQL-2021，适合数据仓库 3NF，分层数据仓库建设，企业运维方便，直接库内分析，是传统稳定的数据仓库之一

不擅长的场景：即席查询、Cube 计算、实时计算

湖引擎各自特点—— Hudi

Uber 开源，支持 Fast Upsert/delete，严格 Schema 定义，支持增量抽取

不擅长的场景：早期只能有一个 writer 写一个表，后期依赖 ZooKeeper，大规模集群，引擎“手动挡”比较明显

湖引擎各自特点 —— IceBerg

轻量级 lake，和 Hive 整合操作简单，主要解决 Hive MetaStore 大大的问题，读写引擎、文件存储，Schema 自定义都可以换（例如B站）

不擅长的场景：upsert 和 delete 支持不够好

手动引擎各自特点 —— DataFusion

基于 Apache Arrow 的计算引擎，Rust 变现，天然支持向量化，做引擎执行部分非常方便（Presto worker, Spark excutor）,根据场景自定义开发，想象空间非常大

https://github.com/apache/arrow-datafusion

不擅长的场景：只是执行引擎，限于骨灰级玩家来玩

加DF群：公司-职位-实名

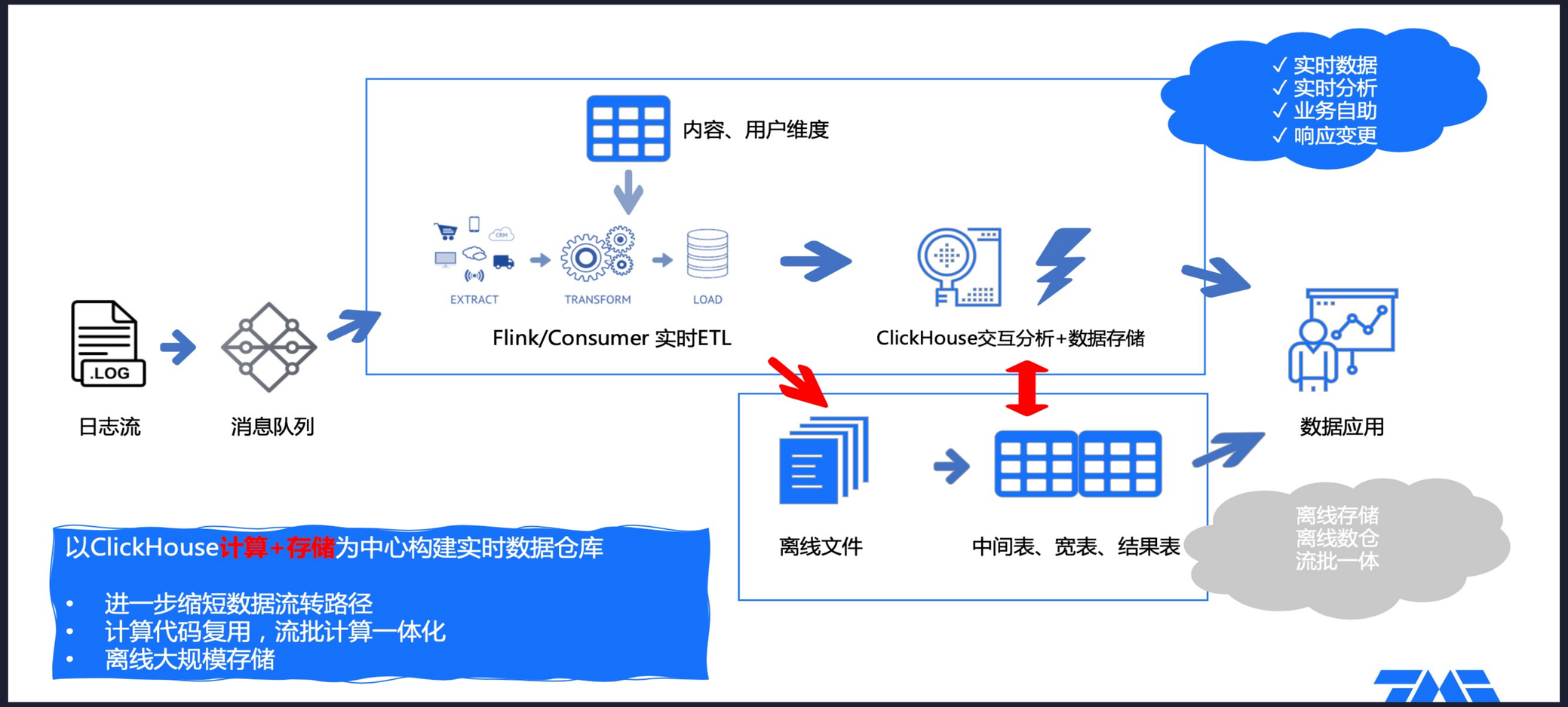
QCon+

未来引擎的趋势 —— 场景化

数据量 * 场景需求 → 通用引擎技术能力

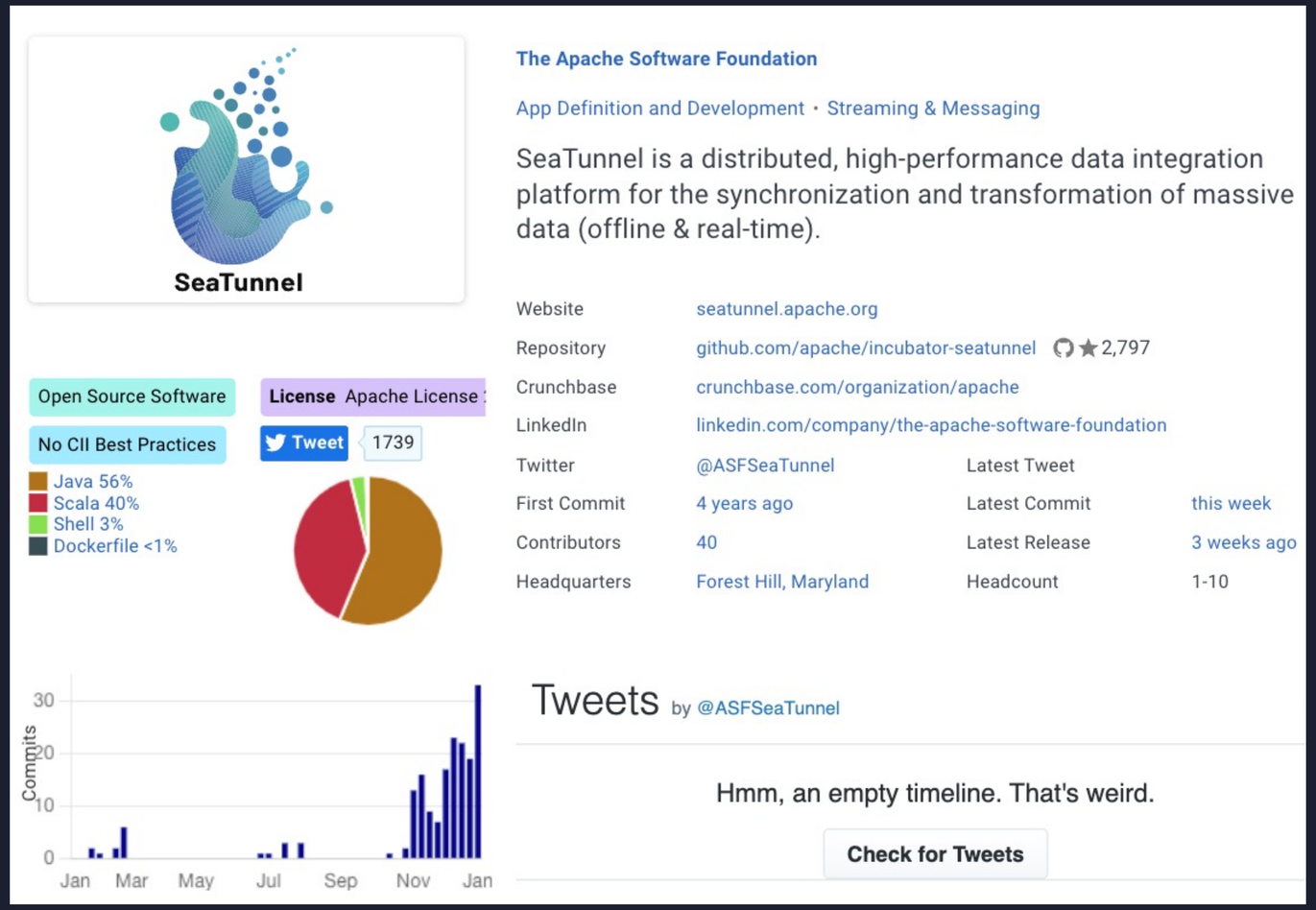
ClickHouse 不是万能的，没有 ClickHouse 是万万不能的

ClickHouse 可以做实时数据仓库么？



一个好汉三个帮，ClickHouse 生态 —— SeaTunnel

https://seatunnel.apache.org/



- "There is high demand for convenient, easy to use and powerful tools for transferring and transforming large amounts of data. I'm happy to see that SeaTunnel has joined Apache incubator and I will follow its growth!"

——Alexey Milovidov, ClickHouse CTO.

- 恭喜 SeaTunnel 成功进入 Apache 孵化器，很高兴看到围绕着数据流转又有一个新的优秀的开源项目出现，现在这个时代，说数据作为业务的核心一点不为过，而且数据存储方面的技术又在这个时代高度的细分化，数据库之间的同步和转化非常有必要，希望 SeaTunnel 成为打通数据孤岛的「桥梁」！

—— PingCAP 联合创始人 & CTO 黄东旭

- 祝贺 SeaTunnel 进入 Apache 孵化器，越来越多的来自中国的孵化器项目表明了中国开源社区的活跃和技术贡献，非常高兴看到 SeaTunnel 社区在数据处理方面的新思考，期待再孵化过程中看到 SeaTunnel 社区的成长

—— Apache Kylin PMC Luke Han

- 恭喜 SeaTunnel 成功进入Apache孵化器。现在是一个异构数据的时代，各种数据库、大数据平台之间需要一个开源、高效的连接器，希望 SeaTunnel 成为这个细分领域的领军者！

—— 涛思数据 TDengine 创始人陶建辉

- 可喜可贺，恭喜 SeaTunnel 成功进入Apache孵化器，预祝团队再创辉煌！作为同是Apache 基金会的 Cassandra 项目，期待与SeaTunnel深度整合。

—— DataStax(Cassandra) China总经理 卢东明

- 恭喜 SeaTunnel 进入 Apache孵化器，SeaTunnel 是一个简单易用的数据同步组件，通过SeaTunnel 可将数据更方便导入 Apache Hudi 数据湖中，也期待两个社区后续进行更深度的合作！

—— Apache Hudi PMC 李少锋

- 恭喜 SeaTunnel 加入Apache孵化器！SeaTunnel 作为一款简单易用、性能突出的海量数据处理产品，今年我们也实现了 SeaTunnel 的 Doris Spark/Flink Sink，希望打通从数据处理到数据分析的通路，能更好服务所有开源用户。我们也相信 SeaTunnel进入孵化器后，在Apache 之道的指引下社区可以进一步发展，有更多热爱开源的企业和个人开发者一同参与进来！最后预祝SeaTunnel 可以早日毕业！

—— Apache Doris PPMC 陈明雨

一个好汉三个帮，ClickHouse 生态—— CKman

https://github.com/housepower/ckman

Clusters › test › tables

Table Metrics

过去24小时查询成功的数量　过去7天的耗费统计

Table Name 表名	Columns 列	Rows 行数	Parts 分区数	Disk Space 占用硬盘空间	Completed Queries in last 24h	Failed Queries in last 24h 过去24小时查询失败的数量	Queries Cost(0.5, 0.99, max) in last 7 days
Kktest_01	16	52	34	78506	205	50	0.026s,0.838919999...
Tmp_k_01	3	0	0	0	0	0	0s,0s,0s
Yh118	7	0	0	0	0	0	0.0305s,0.03985000...
Yh118$	7	0	0	0	0	0	0s,0s,0s
Yh118我	7	0	0	0	0	0	0s,0s,0s

QCon+

QCon+ 本周最新专题

ClickHouse 集群版深度实践

郭炜 ClickHouse 中国社区发起人

案例 1

微信 ClickHouse 海量数据接入的探索和实践

孙弘毅 腾讯微信 大数据高级研发工程师

案例 2

如何高效一致同步海量数据进入 ClickHouse

王玉 唯品会 OLAP 资深工程师

案例 3

ClickHouse 去 ZooKeeper 集群使用方案

贺钰城 联想 开源优化架构师

案例 4

中国移动基于 ClickHouse 提升数据处理效率的应用实践

高天铎 中国移动 云能力中心高级系统架构师

案例 5

ClickHouse 在网易用户行为分析中的实战

虞李凯 网易 高级数据研发工程师

案例 6

ClickHouse 在云智慧 AIOps 场景的应用实践

孔文 云智慧 研发总监

视频列表 共6个视频

微信 ClickHouse 海量数据接入的探索和实践

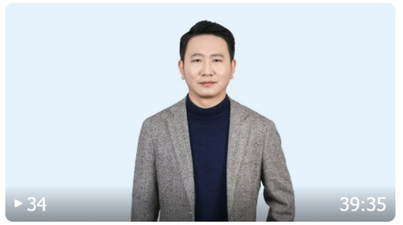
孙弘毅 腾讯微信大数据高级研发工程师

如何高效一致同步海量数据入 ClickHouse

王玉 唯品会 OLAP 资深工程师

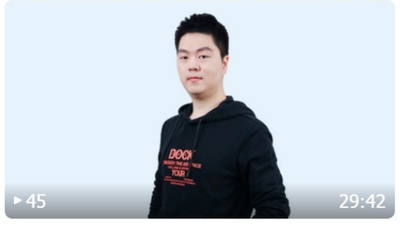
ClickHouse 去 ZooKeeper 集群使用方案

贺钰城 联想开源优化架构师



中国移动基于 ClickHouse 提升数据处理效率的应用实践

高天铎 中国移动云能力中心高级系统...



ClickHouse 在网易用户行为分析中的实战

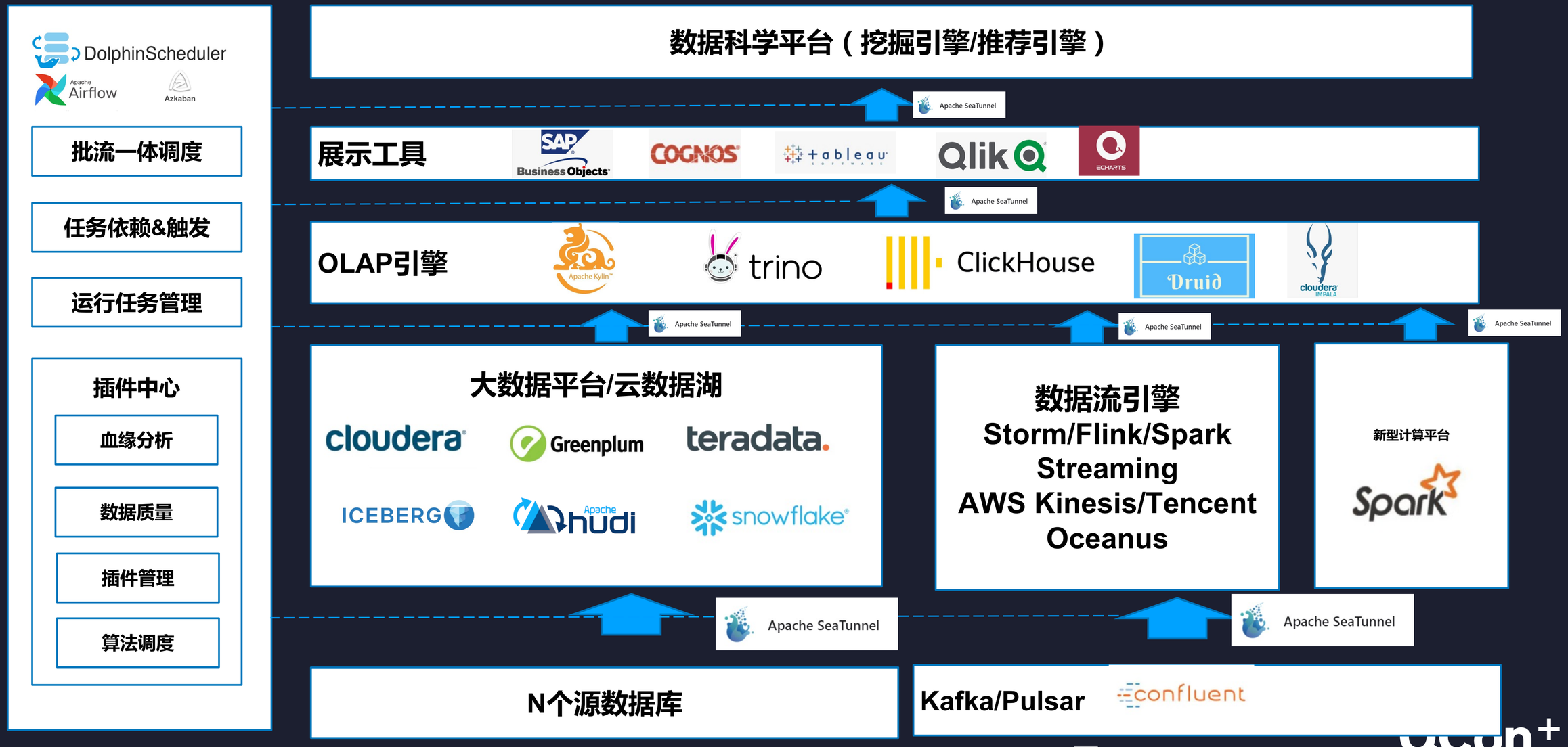
虞李凯 网易高级数据研发工程师

ClickHouse 在云智慧 AIOps 场景的应用实践

孔文 云智慧研发总监

QCon+

笑谈天下库，煮酒论引擎


DolphinScheduler
Apache Airflow
Azkaban
批流一体调度
任务依赖&触发
运行任务管理
插件中心
血缘分析
数据质量
插件管理
算法调度
数据科学平台（挖掘引擎/推荐引擎）
Apache SeaTunnel
展示工具
SAP Business Objects
COGNOS
tableau
Qlik
ECHARTS
Apache SeaTunnel
OLAP引擎
Apache Kylin
trino
ClickHouse
Druid
cloudera IMPALA
Apache SeaTunnel
Apache SeaTunnel
Apache SeaTunnel
大数据平台/云数据湖
cloudera
Greenplum
teradata.
ICEBERG
Apache hudi
snowflake
数据流引擎
Storm/Flink/Spark Streaming
AWS Kinesis/Tencent Oceanus
新型计算平台
Spark
Apache SeaTunnel
Apache SeaTunnel
N个源数据库
Kafka/Pulsar
confluent

数据引擎不要被各种评测忽悠，

最适合你场景的引擎“们”，才是最好的。

多引擎的时代，来了！

视频号：郭大侠说开源

联系我注明：公司-职位-实名

QCon+

THANKS

QCon+ 案例研习社