DTCC 2010 2022

第十三届中国数据库技术大会

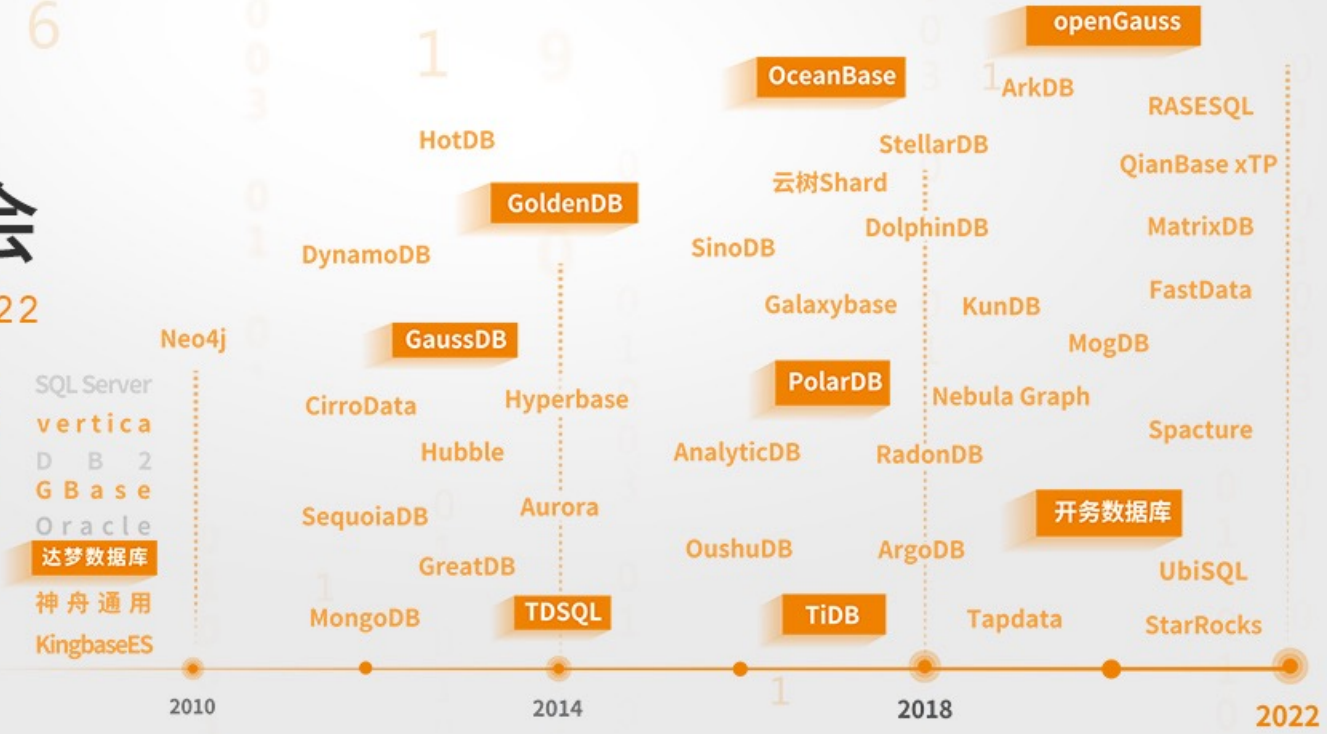
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



阿里云数据湖与湖仓架构设计与实践

范佚伦（子灼）
阿里云 技术专家

1 数据湖与湖仓一体

什么是数据湖

WIKIPEDIA
The Free Encyclopedia

A data lake is a system or repository of data stored in its natural/raw format, usually object blobs or files. A data lake is usually a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning. A data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video).

aws

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

Microsoft Azure

Azure Data Lake includes all the capabilities required to make it easy for developers, data scientists, and analysts to store data of any size, shape, and speed, and do all types of processing and analytics across platforms and languages.

阿里云

数据湖是统一存储池，可对接多种数据输入方式，您可以存储任意规模的结构化、半结构化、非结构化数据。数据湖可无缝对接多种计算分析平台，直接进行数据处理与分析，打破孤岛，洞察业务价值。同时，数据湖提供冷热分层转换能力，覆盖数据全生命周期。

■ **相比于传统数仓内置存储，数据湖通过存算分离实现统一存储**

- 统一的存储，解决数据孤岛问题
- 灵活性强，存算分离，开放的数据，多种引擎分析

■ **相比于传统数仓事前建模，数据湖可以保存原始数据**

- 事后建模
- 云上大规模、高可用、低成本的中心化存储
- 数据类型丰富，支持结构化、半结构化、非结构化数据类型
- 存储原始数据，避免数据丢失

■ **相比于传统数仓，数据湖缺乏数据治理、性能**

- 安全，权限
- 事务性
- 数据质量
- 性能

数据湖与数仓融合，实现湖仓一体

■ **云上数据湖优势**

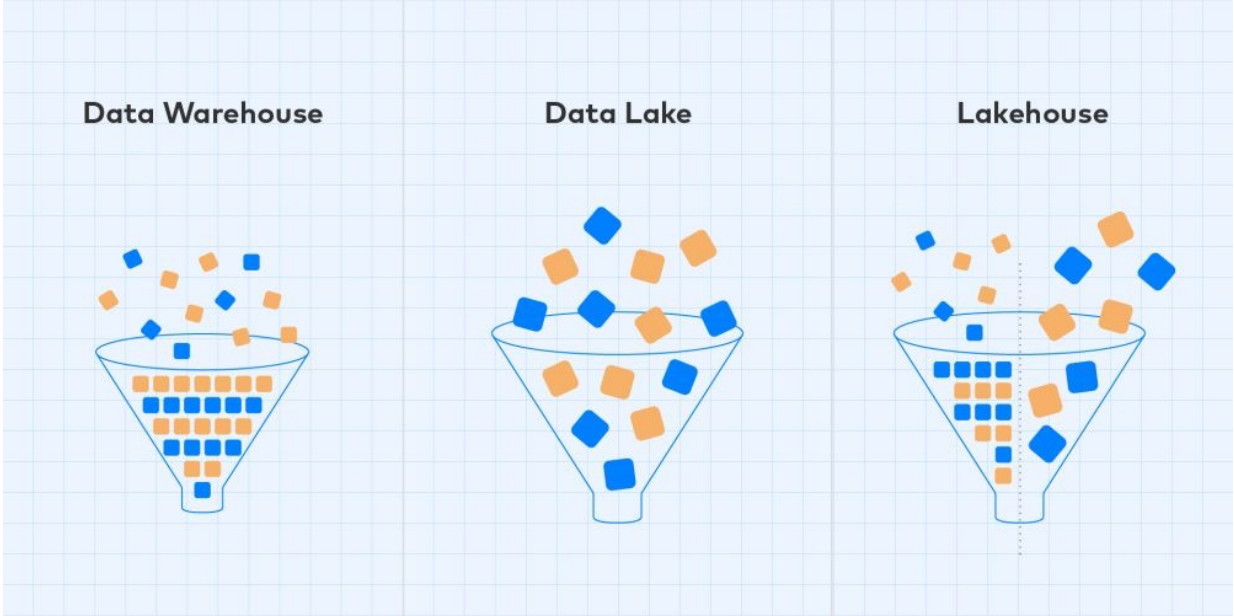
- 利用开源软件，启动成本低，灵活性强
- 计算存储分离，一定的成本优势
- 计算引擎丰富，多种场景覆盖

■ **云上数据仓库优势**

- 全托管，免运维
- 统一的权限身份认证
- 针对特定场景，高度优化的引擎

■ **通过打通湖仓数据互通实现湖仓一体**

- 将数据湖、数仓的元数据打通
- 数仓支持读写云上对象存储
- 数仓通过cache加速数据湖读写



数据湖逐渐支持数仓能力，实现湖仓一体

■ 元数据统一与数据湖管理

- 统一元数据层查询和定位数据
- 开放的数据格式支持多引擎直接读取
- 统一的元数据 / SQL API
- 统一权限提供企业级数据管理和安全的能力

■ 数据湖查询优化

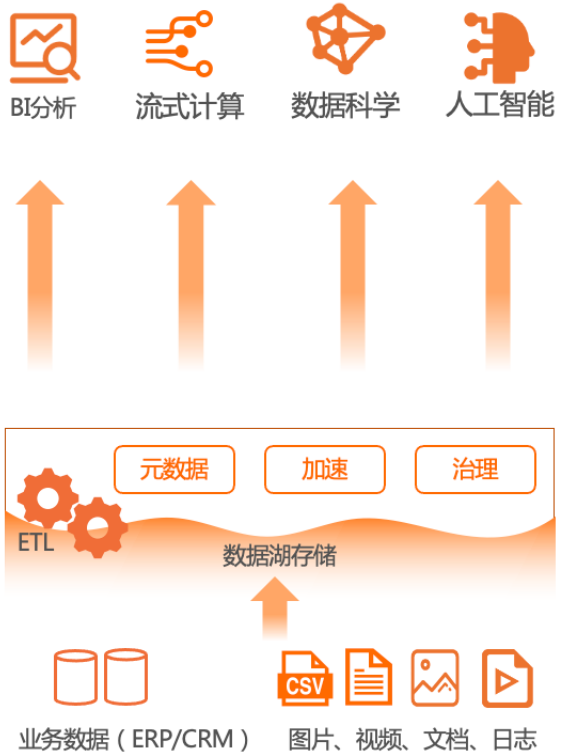
- 数据缓存加速
- 数据索引加速

■ 利用数据湖格式实现事务层

- 支持ACID事务隔离，解决读写冲突
- 支持多版本时间旅行，指明每个Table版本所包含的数据对象
- 同时支持流批混合读写
- DELTA LAKE Apache hudi ICEBERG

■ 统一的数据湖存储层

- 利用云上对象存储
- 存储原始数据，支持半结构化和非结构化数据



2 阿里云数据湖架构

阿里云数据湖架构

■ **支持多引擎计算分析**
- E-MapReduce (Hive/Spark/Presto/Impala/Starrocks)
- MaxCompute、Flink、Hologres

■ **通过DLF进行数据湖管理优化**
- 统一元数据管理，支持跨引擎分析
- 统一权限管理
- 搭配Jindo Cache数据湖缓存加速
- 数据湖存储管理和自动冷热分层

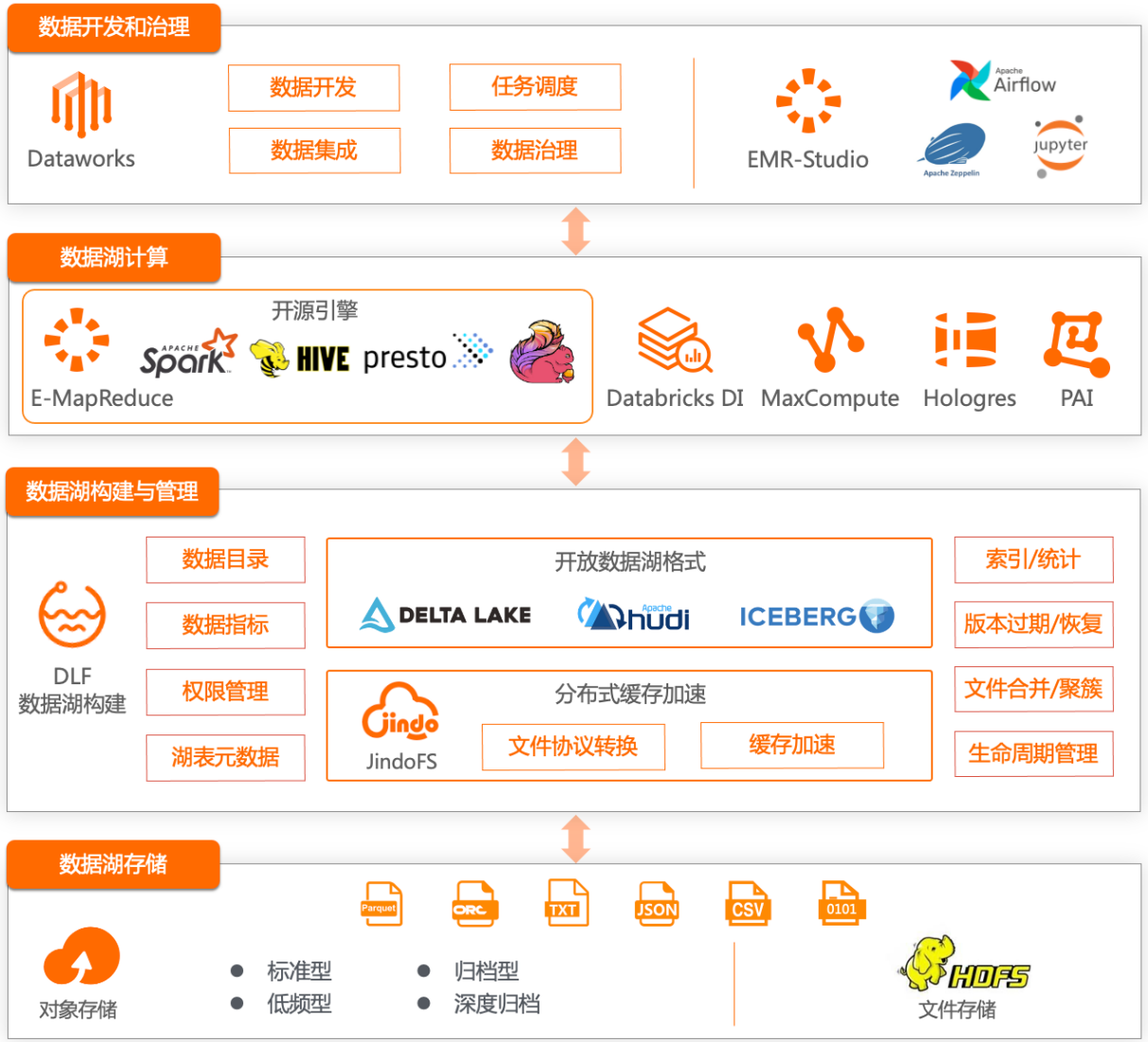
■ **使用OSS作为数据湖存储**
- 低成本，高可靠性，无限扩展，高吞吐，免运维
- 开启OSS-HDFS，支持POSIX，高性能rename/list
- 通过冷归档降低成本

开发层（数据开发与治理）

计算层（弹性计算引擎）

管理层（数据管理与优化）

存储层（数据湖统一存储）



使用数据湖架构的挑战


元数据服务
HMS搭建复杂
扩展性、稳定性、性能一般
对接多个计算引擎难度大
数据查找不方便
数据安全
权限控制方案复杂、成熟度低
无法覆盖所有计算引擎
开放存储、权限难控制
湖管理和优化
数据治理体系和工具不完善
容易退化为数据沼泽
存储成本高
分层存储使用难度大
湖格式使用难度大
DLF 数据湖构建
统一元数据中心
• 高稳定性、扩展性和性能，免运维
• 无缝对接多种计算引擎
• 元数据发现、迁移、备份
企业级权限控制
• 对接RAM/LDAP账户体系
• 字段级读写权限控制
• 提供Location权限托管
• 不同引擎统一权限控制
智能数据管理与优化
• 精细化存储统计与分析
• 湖格式优化、无效数据清理
• 自动化数据冷热判断和分层

阿里云DLF（Data Lake Formation）简介

统一元数据服务

- 存算分离架构下，提供全托管的有状态服务
- 高可用、高性能、可扩展、免运维
- 兼容开源HMS协议，无缝对接开源/自研引擎
- 总量支持超过10万DB，1亿Table，10亿Partition

权限与安全

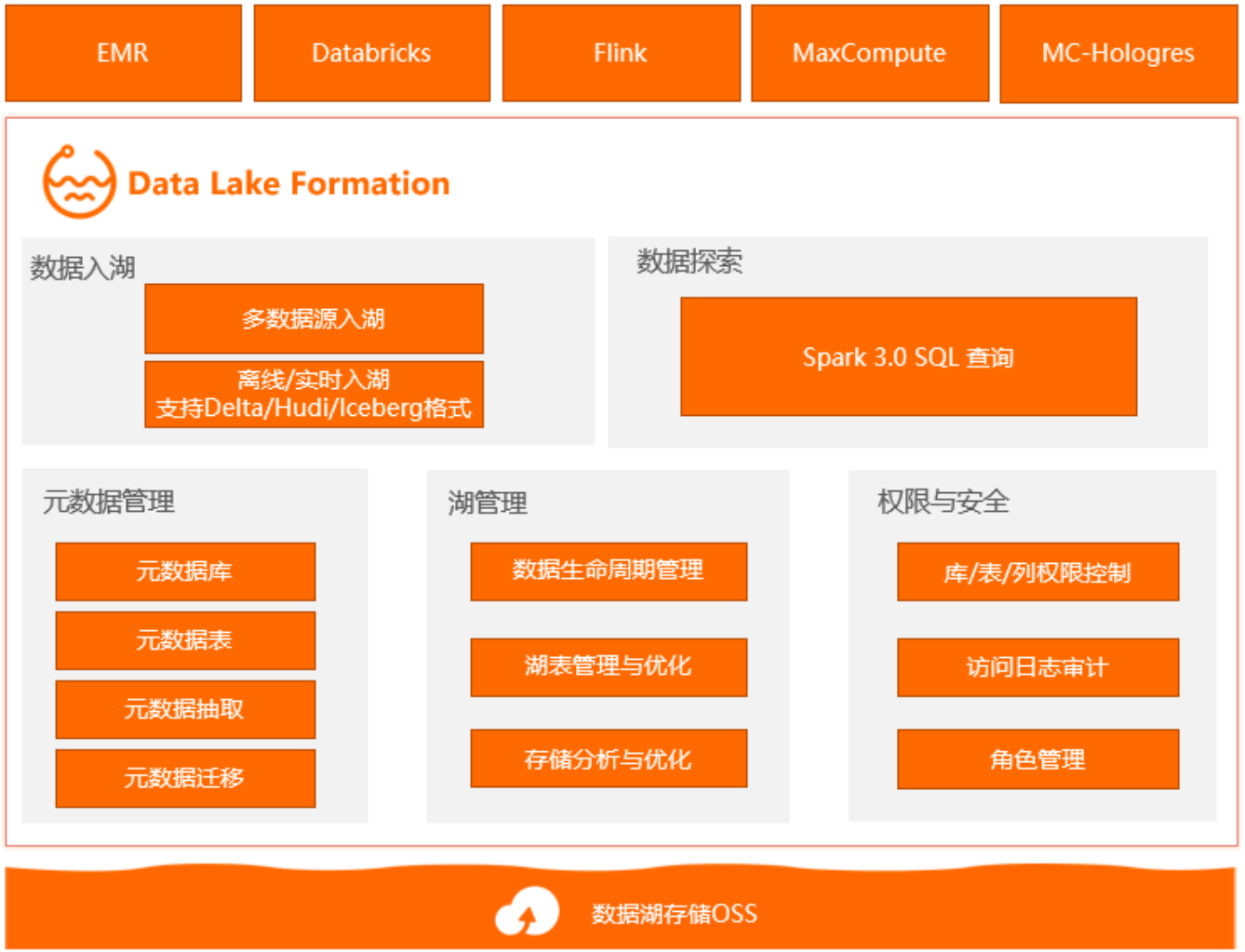
- 支持按库/表/列对湖内数据进行权限配置
- 支持数据访问日志审计

数据入湖与探索

- 模板化支持多种数据源入湖，MySQL、SLS、OTS、Kafka等
- 离线/实时入湖，支持Delta/Hudi/Iceberg等多种数据湖格式
- 提供便捷的数据探查能力，快速对湖内数据进行探索和分析

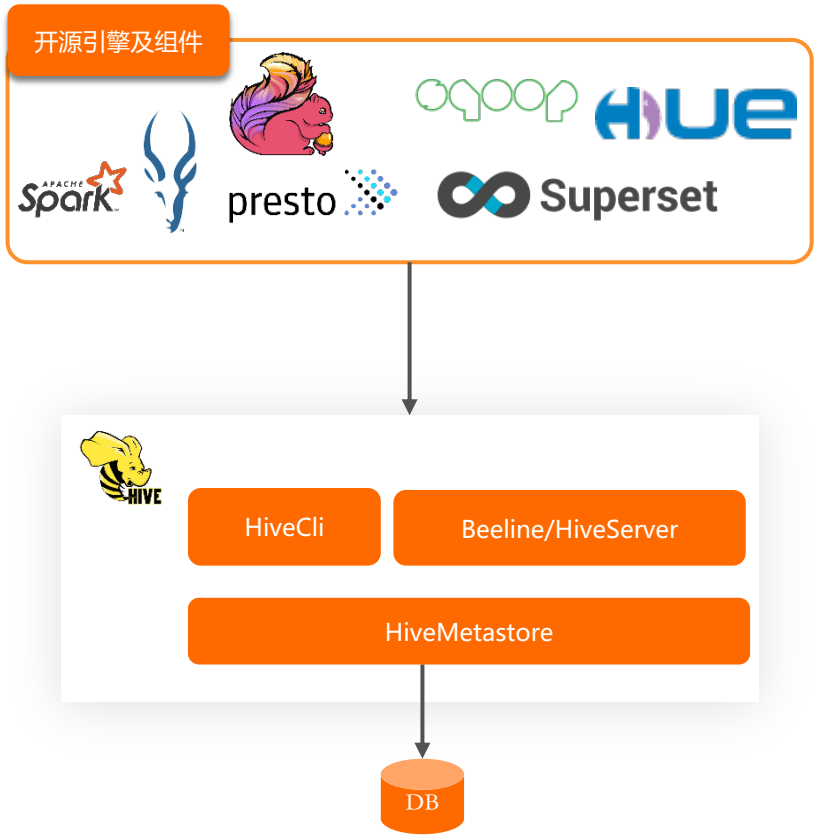
数据管理与优化

- 存储分析与成本优化
- 湖表数据分布与索引加速
- 数据生命周期管理



3 DLF统一元数据设计与实践

开源元数据体系和问题



■ **Hive 是开源数仓的事实标准**

- 各个引擎逐渐形成了围绕着Hive Metastore的元数据体系

■ **高级特性支持有限**

- 不能通过Time-Travel查询数据/元数据的历史快照
- ACID特性和Hive引擎绑定

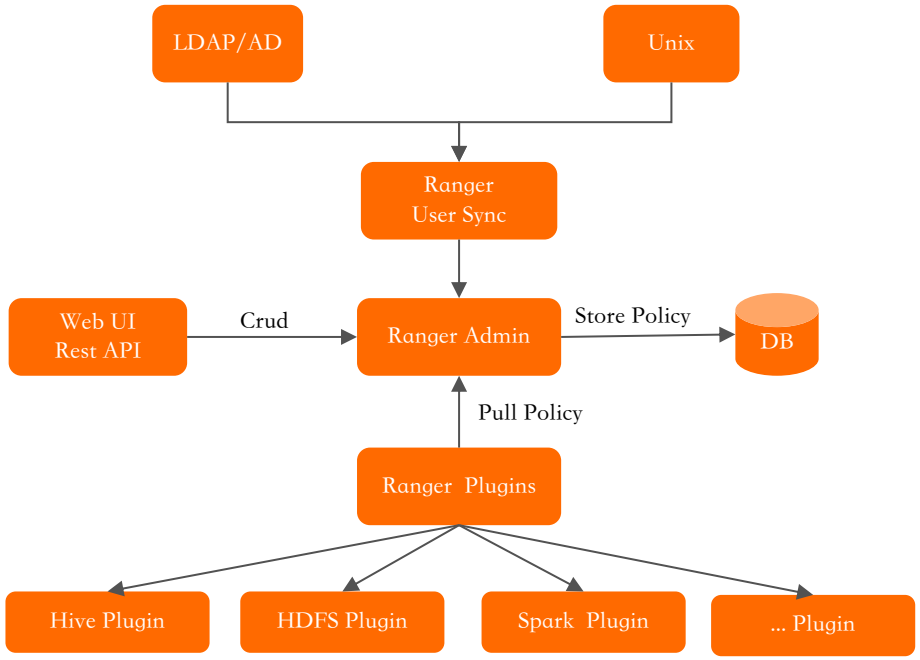
■ **不易于对接内部自研引擎/云上数仓**

- Hive Meta store额外部署运维，单点问题，需要网络直连
- 引擎需要实现Thrift协议接入

■ **受限于单个数据库瓶颈**

- 单个MySQL数据量瓶颈
- 高可用问题

开源权限体系和问题



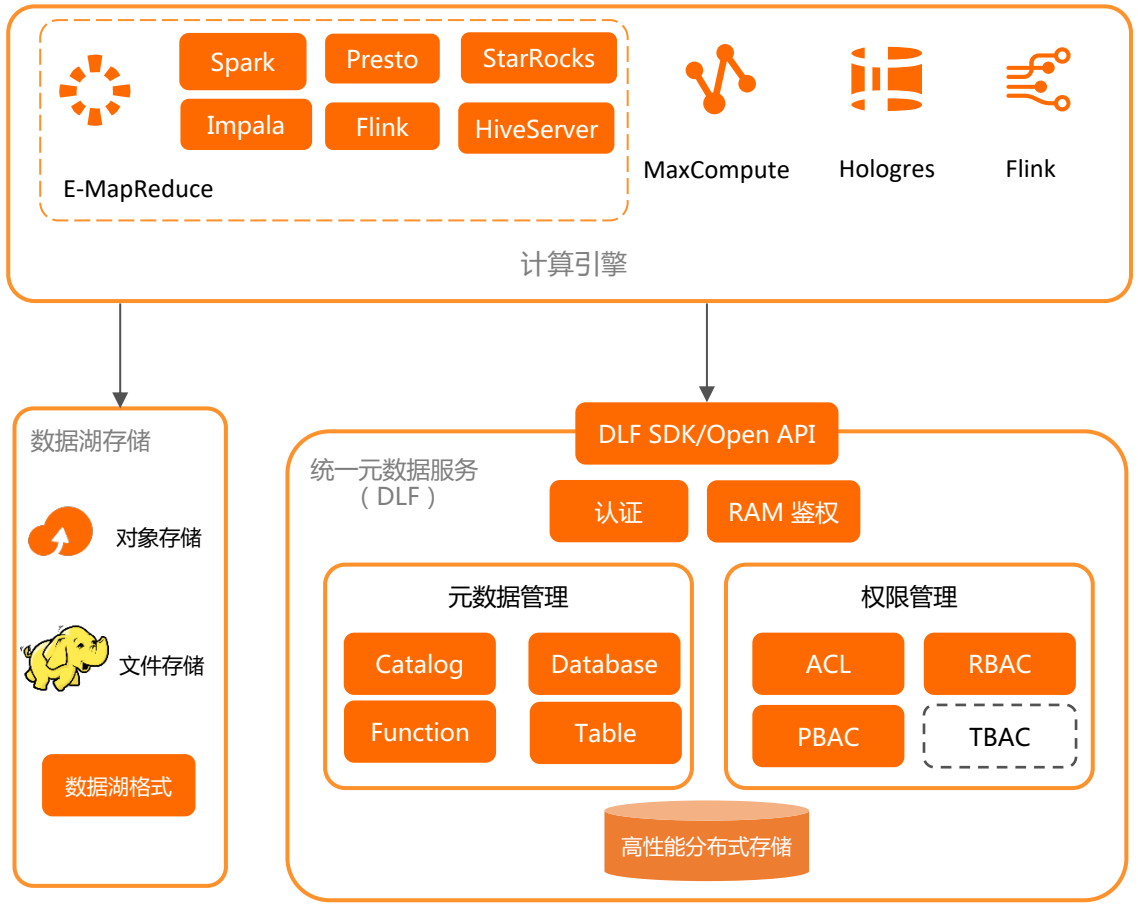
■ Hive Authentication

- Storage-Based Authorization
  - 元数据操作权限映射到底层文件的权限
  - 不支持细粒度鉴权
- SQL-Standard Based Authorization
  - 支持GRANT/REVOKE，对库、表鉴权
  - 依赖HiveServer2

■ Apache Ranger

- 中心化的权限控制方案，支持很多Hadoop生态组件，支持PBAC
- 官方没有提供SparkSQL权限插件
- 元数据接口与权限接口分离
- 数据湖格式不兼容

阿里云DLF统一元数据架构



■ **统一元数据，多引擎支持**

- SDK兼容HMS协议，无缝对接开源与自研计算引擎
- 标准OpenAPI，支持客户自建集群及系统集成

■ **全托管增强型元数据服务**

- 高可用、高性能、可扩展、免运维
- 支持多Catalog多租户
- Schema多版本
- Table/Partition Column Statistic统计

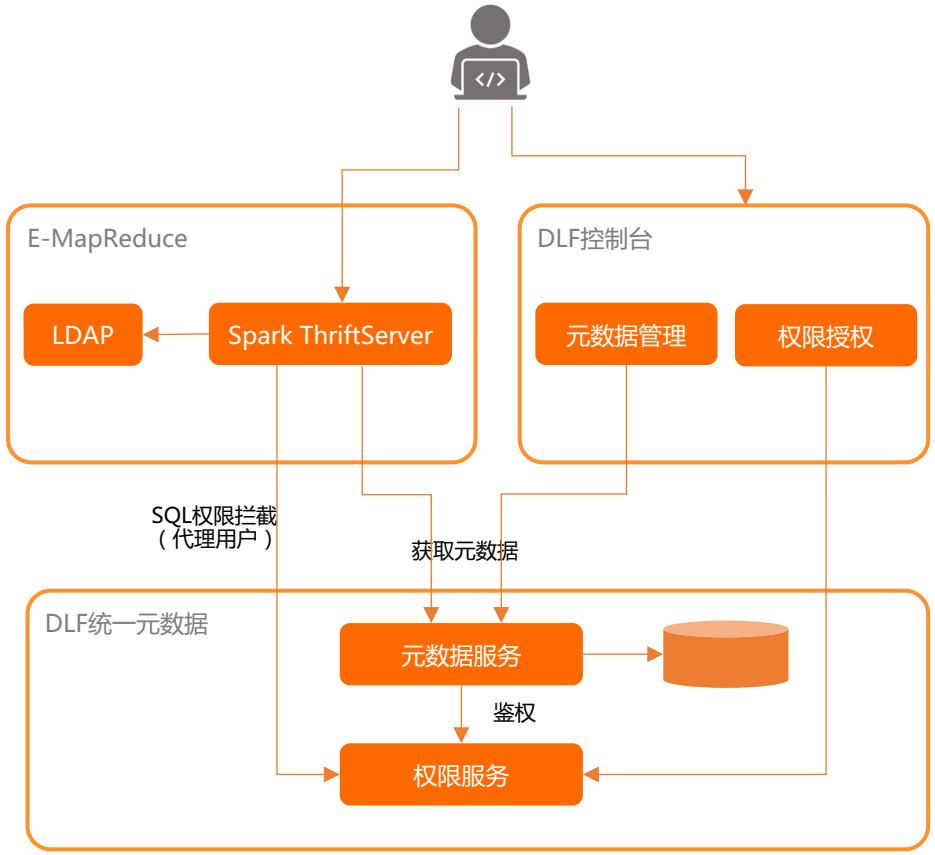
■ **统一权限控制**

- 一套配置，多引擎统一管控
- 支持RAM和LDAP账号体系

■ **元数据实时检索**

- 实时消费元数据变更，写入ES进行全文检索

阿里云DLF统一权限鉴权流程



以 Spark为例

■ **用户自定义权限开关**

- Catalog级别权限设置

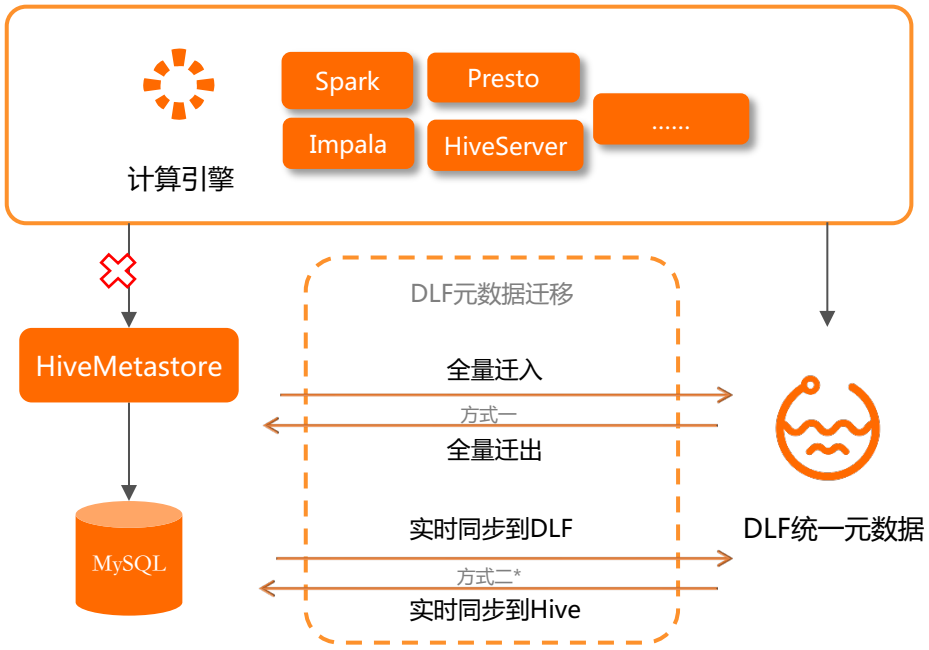
■ **多种鉴权方式**

- 通过API/SDK/控制台访问元数据鉴权
- 计算引擎通过代理用户的方式进行SQL鉴权

■ **细粒度权限控制**

- 授权主体可以是RAM用户或自定义角色
- 授权资源包括Database、Table、Column、Function
- 访问资源方式包括Describe、Alter、Drop、Select、Update等

元数据迁移



■ **产品化元数据迁移功能**

- 页面配置一键迁移，兼容Hive2/Hive3
- 自动处理网络打通
- 免运维迁移任务集群和资源

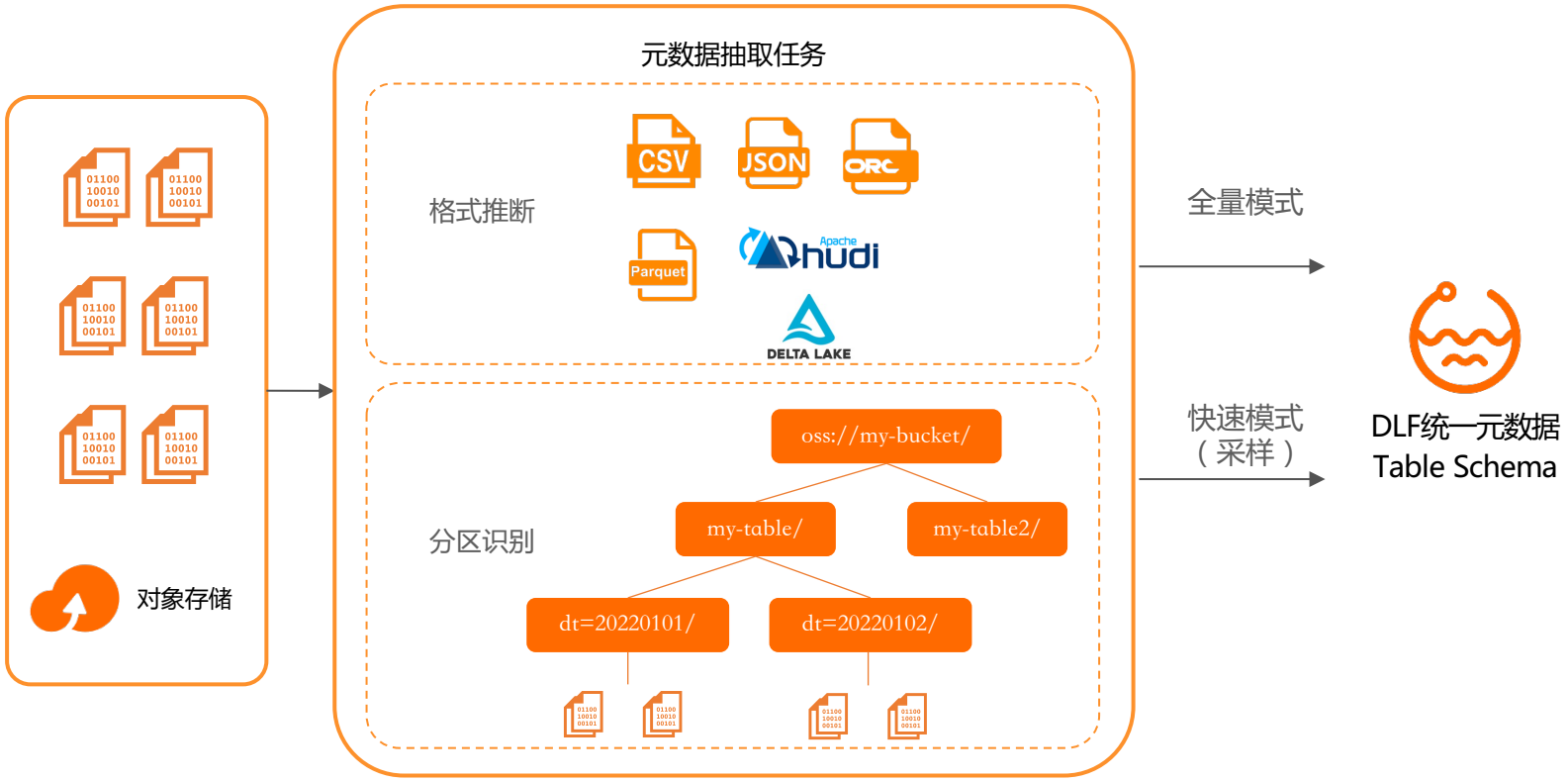
■ **支持全量元数据迁移**

- 一次性全量迁移，支持迁入/迁出
- 提供元数据比对工具，用于校验两边数据差异
- 提供元数据补偿工具，用于迁移后增量数据迁移

■ **支持元数据双写同步*(规划中)**

- 实时异步复制Hive元数据到DLF
- 实时异步复制DLF元数据到Hive

元数据抽取



■ 自动发现数据湖文件schema

- 人工维护的csv文件
- 导入的数据集文件
- ……

■ 格式与分区自动识别

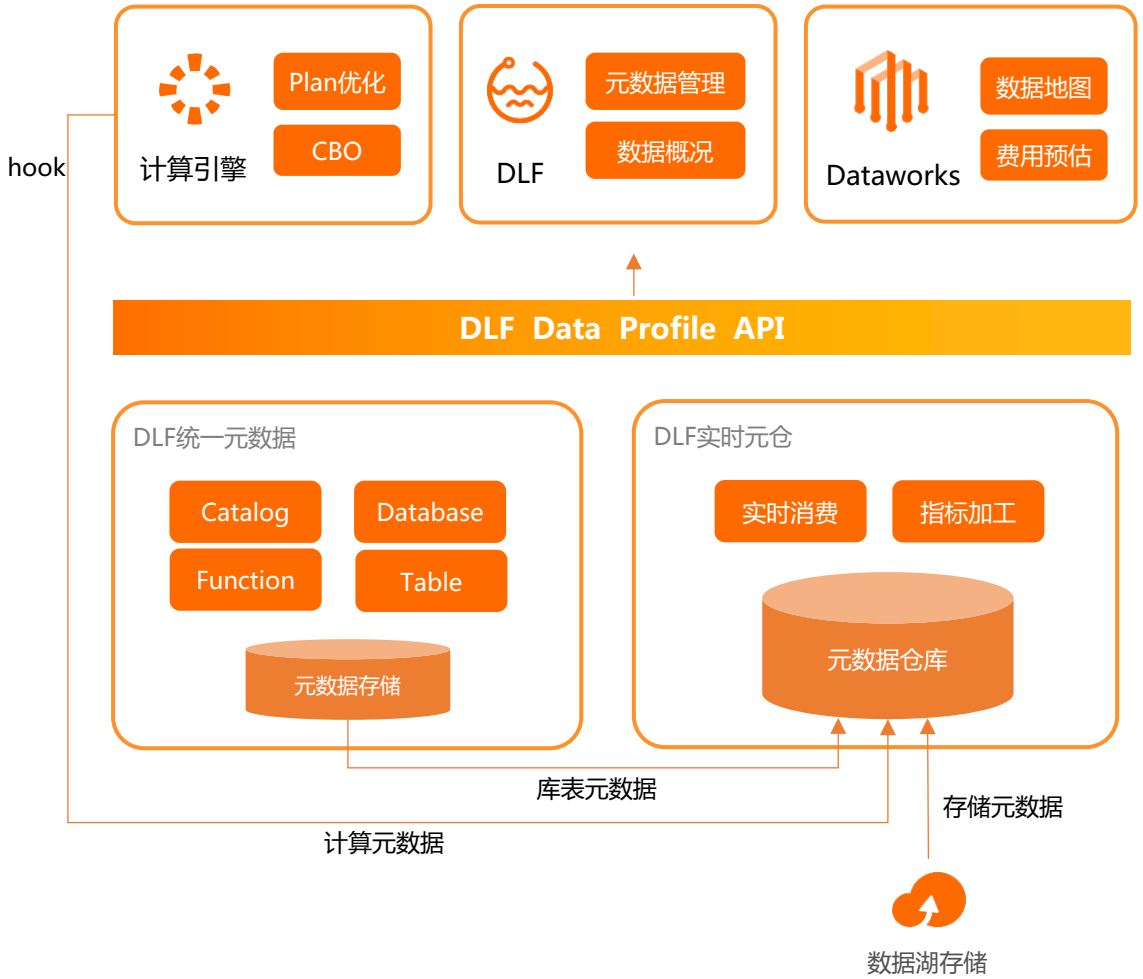
- 文件格式自动识别，包括 csv/json/parquet/orc/hudi/delta
- 自动识别出符合Hive分区的目录结构并创建分区
- 支持全量或采样抽取

■ 多种抽取策略

- 发现表字段更新时，全量更新表结构或者仅新增列
- 发现文件已被删除时，是否删除表的元数据
- 支持cron定时执行

4 DLF数据湖管理与湖格式优化

通过元仓加强元数据管理



■ 实时元仓架构

- 基于Hologres的元数据仓库，用于补充构造元数据相关的额外指标
- 数据源包括计算引擎hook消息、元数据变更消息、存储分析数据等

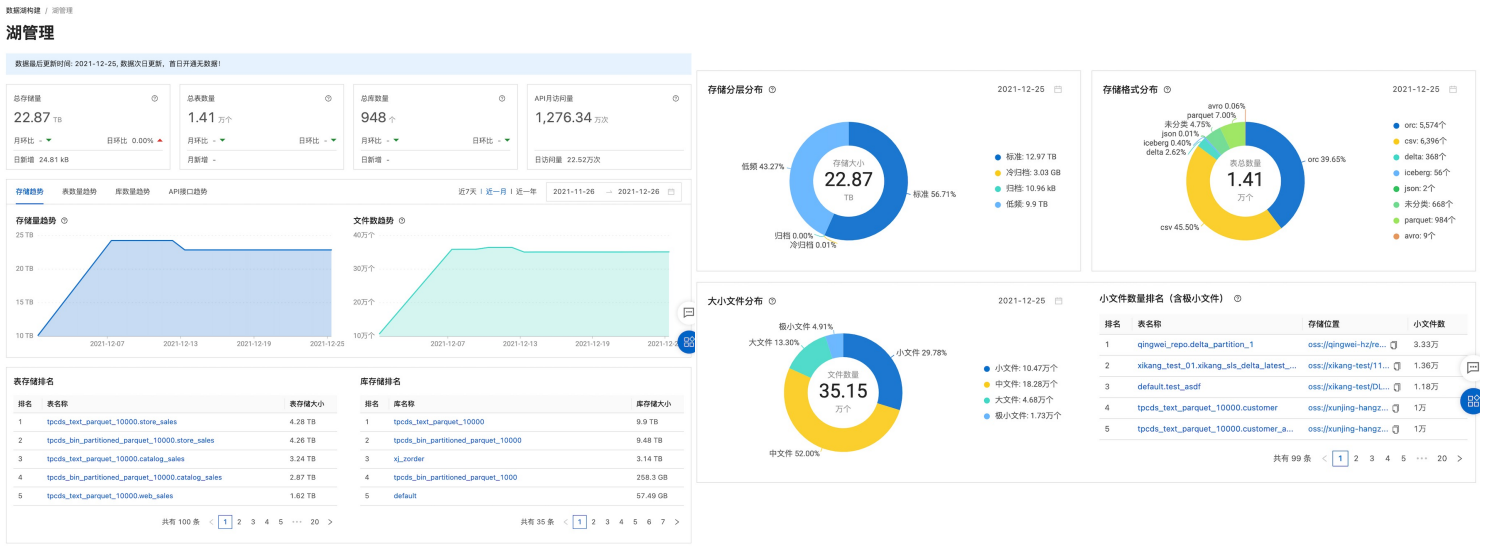
■ 丰富的Dataprofile指标

- 表大小、分区大小、行数、文件数
- 小文件数、小文件占比、文件冷热度
- 湖格式有效文件数、无效文件数

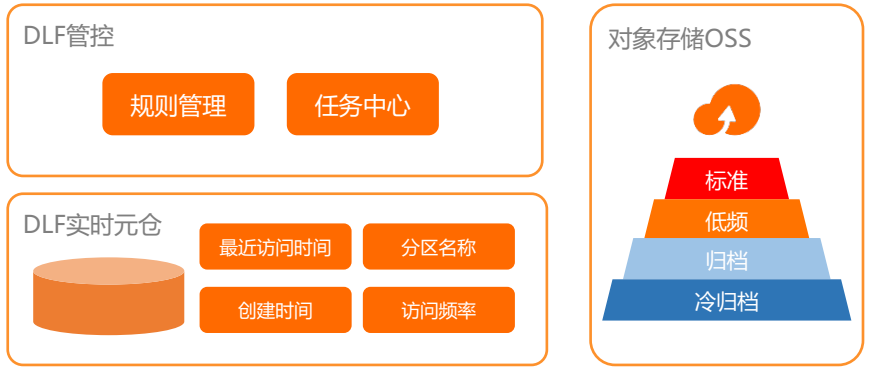
■ 元仓指标应用

- 为计算引擎提供tableSize等信息
- DLF元数据管理
- 提供OpenAPI供用户分析

数据湖存储分析与优化



存储分析



生命周期管理

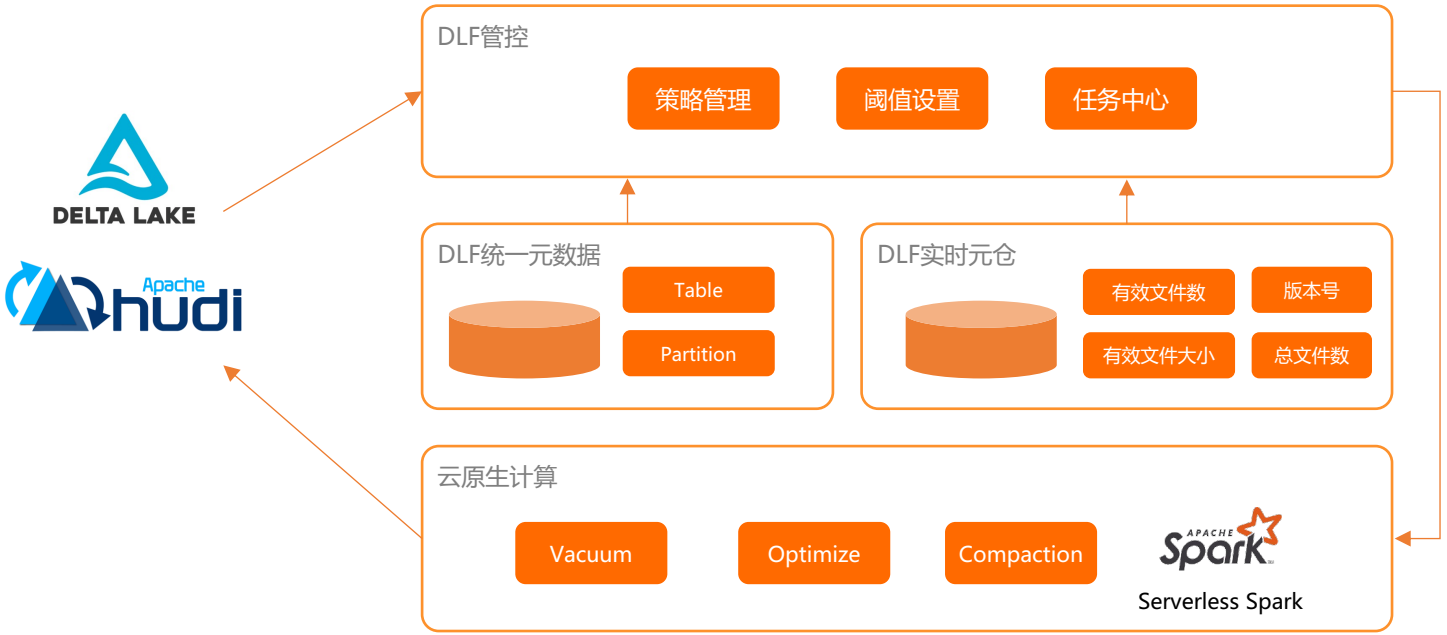
■ 存储分析

- 利用元仓进行数据资产统计与分析
- 支持结构化、非结构化数据
- 实时统计&历史分析
- 包含库/表/分区 存储趋势，大小文件分布，存储格式分布等信息

■ 生命周期管理

- 表/分区级别冷热度计算及分层
- 多种规则配置自动归档
- 一键解冻

湖格式自动管理优化



■ 阿里云EMR Delta Lake特性

- 支持元数据自动同步
- 拓展Spark SQL支持Streaming SQL语法
- 支持G-SCD，无需变更表结构，通过savepoint实现历史快照永久保存.

■ DLF湖格式管理

- 全托管入湖，支持Delta lake、Iceberg、Hudi
- 基于规则策略，实现湖格式表自动优化：
  - 自动Vacuum清理历史版本
  - 自动Optimize合并小文件
  - 自动Z-Order排序优化查询
  - 自动识别新分区


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
SequoiaDB
开炁数据库
OushuDB
ArgoDB
GreatDB
UbiSQL
神舟通用
MongoDB
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS