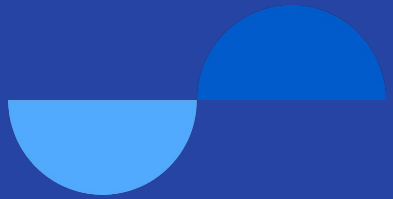

数据技术嘉年华
// Data Technology Carnival
开源 · 融合 · 数智化 — 引领数据技术发展 释放数据要素价值
中国DBA联盟
All China DBA Union
墨天轮

Databend

Databend

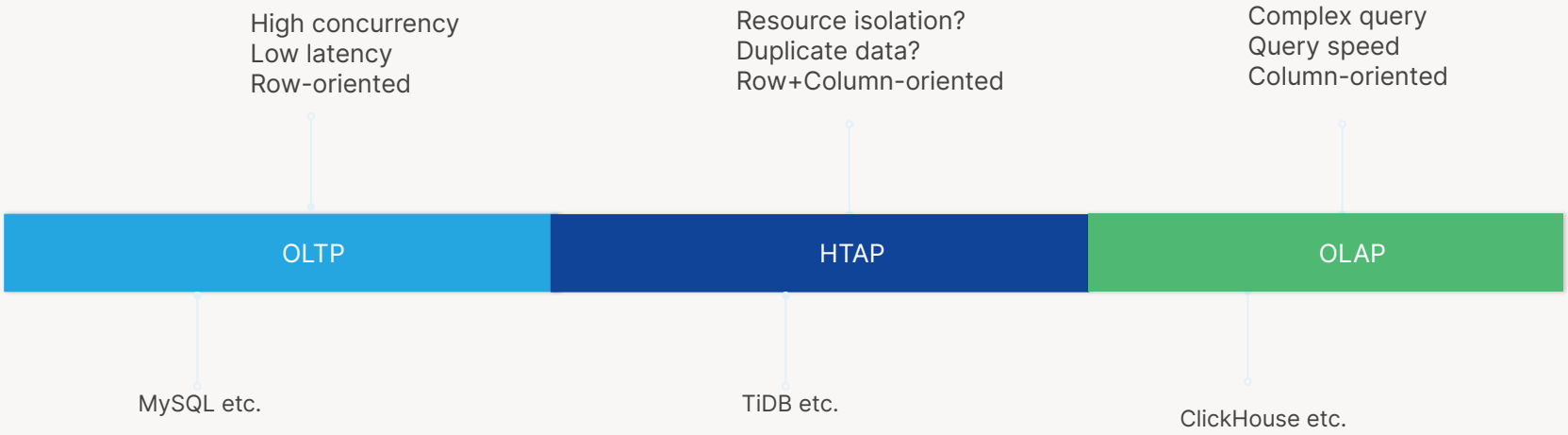
A modern cloud warehouse with Rust for your massive-scale analytics

https://github.com/datafuselabs/databend



Database and Data Warehouse

Databend



大纲

- 云端大数据分析的”新”问题
- 传统数仓解决方案的局限性
- 新一代云原生数仓设计原则
- 新一代云原生数仓核心功能与能力
- 使用 Rust 从零开始研发一款数仓是种什么体验？

Databend

Bohu TANG (张雁飞)

Co-Creator of Databend: https://github.com/datafuselabs/databend
ClickHouse and MySQL(TokuDB) 重度贡献者

Database Kernel | Distributed Database | Data Warehouse

https://bohutang.me/

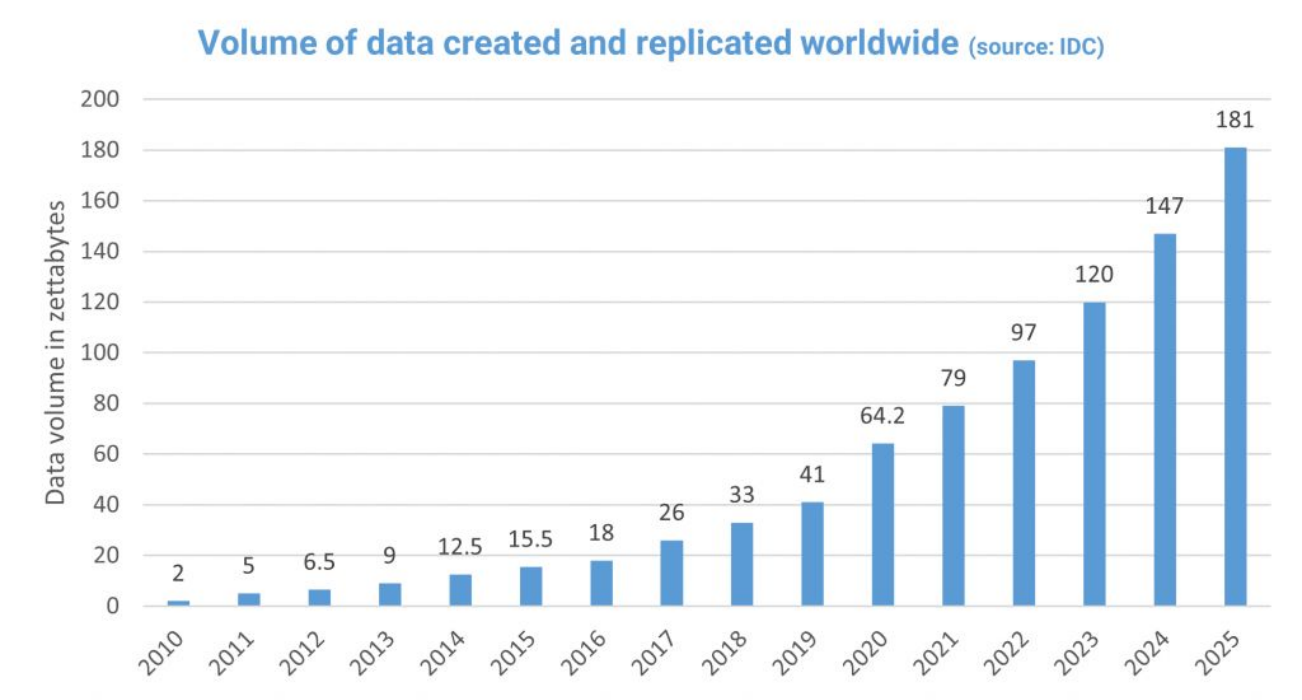
Databend

01

当今(2023)大数据分析新问题

全球数据指数级增长

Databend



1024PB = 1EB, 1024EB = 1ZB

当今大数据新问题 1

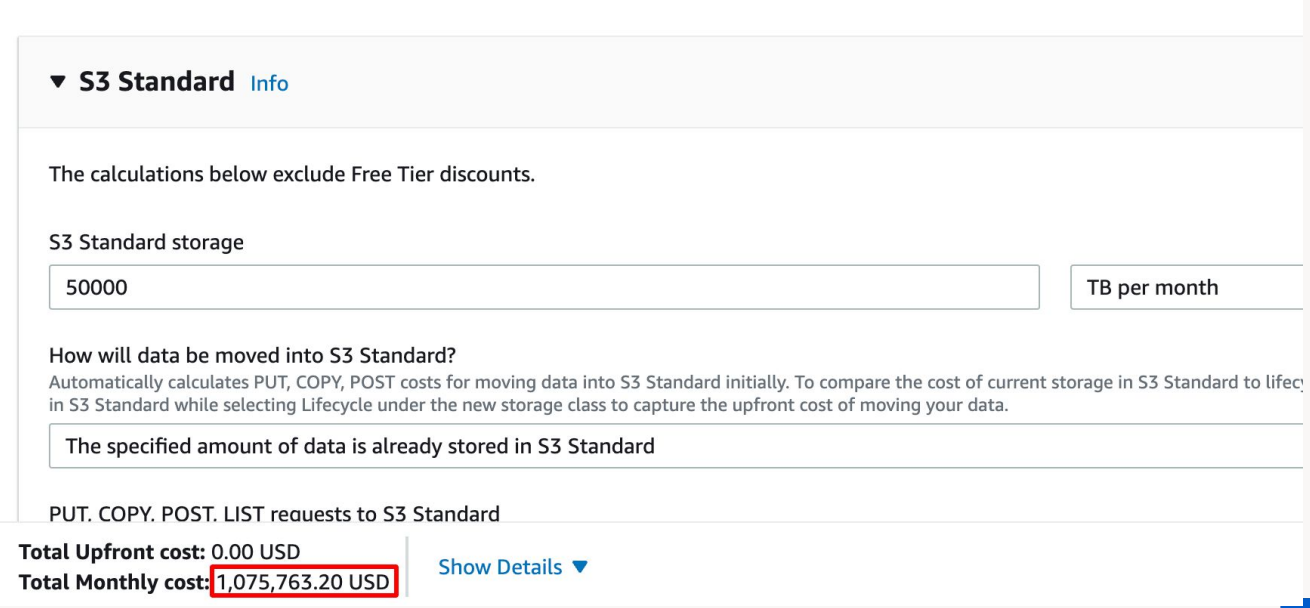
Databend

- 大数据量下的资源利用率问题：< 50%
- 物理资源常驻问题：受限于架构，资源无法充分利用
- 面向固定资源做调度问题
- 大数据分析，波峰、波谷问题

当今大数据新问题 2

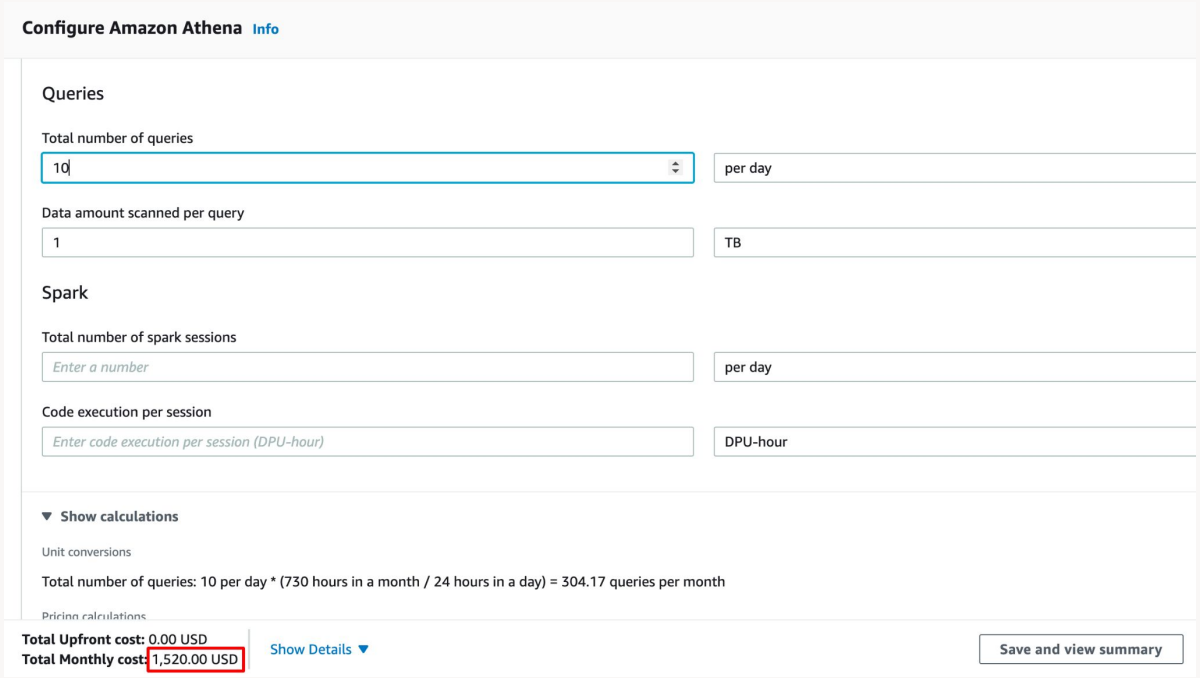
Databend

- 大数据量下的存储成本问题
- PB 级数据，每月存储成本百万美金！



当今大数据新问题 3

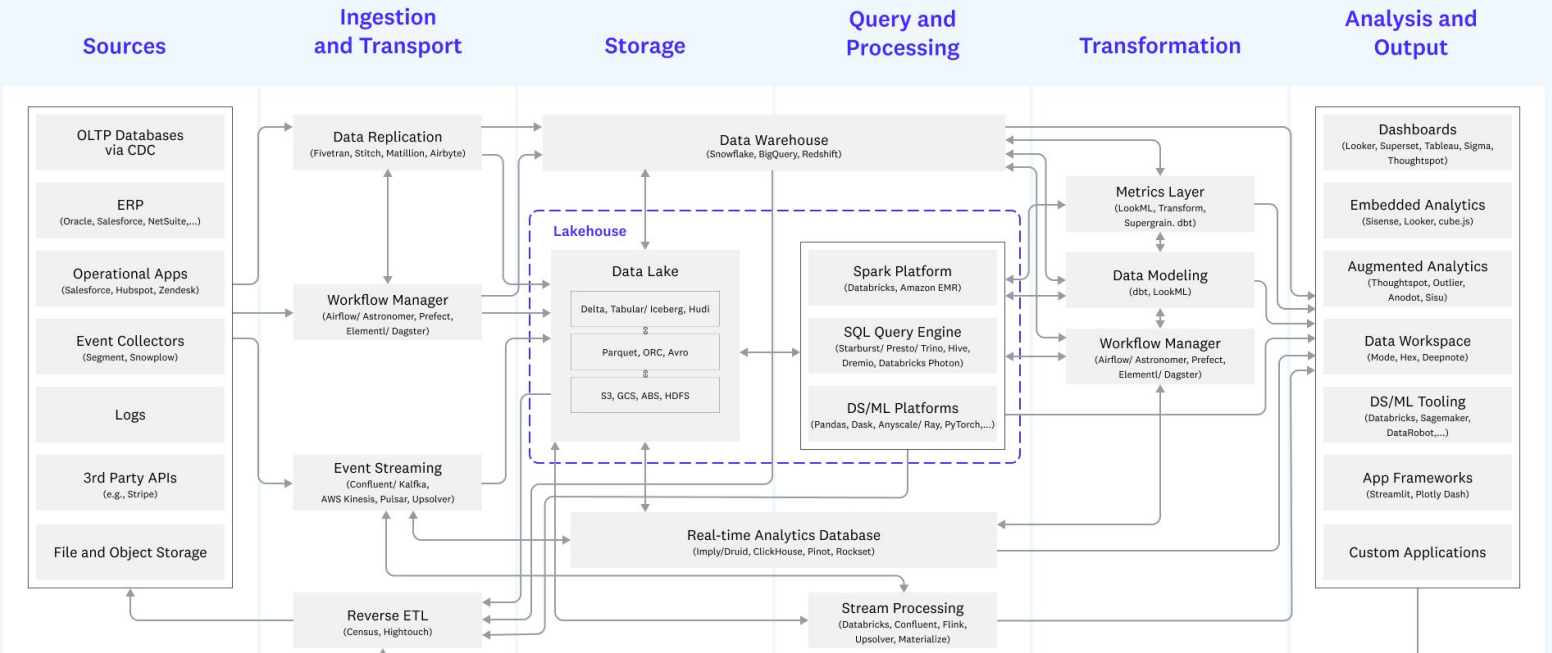
- 大数据量下的计算成本问题
- 对扫描数据量要求非常高，容易破产



当今大数据新问题 4

- 随着数据量的增长，数据平台的复杂度逐渐上升

Unified Data Infrastructure (2.0) (From a16z)



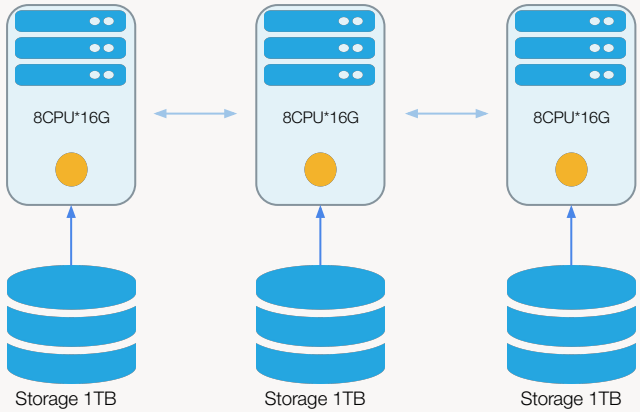
Databend

02

传统数仓架构 vs. 弹性数仓架构

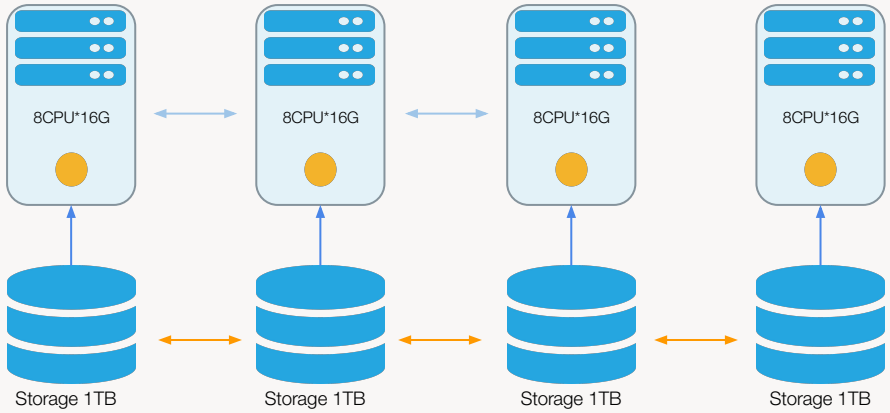
传统数仓架构

- Shared-Nothing
- 存储、计算一体
- 资源固定(Fixed-Set)式调度
- 资源控制粒度粗



传统数仓架构

- Shared-Nothing
- 存储、计算一体
- 资源固定(Fixed-Set)式调度
- 资源控制粒度粗



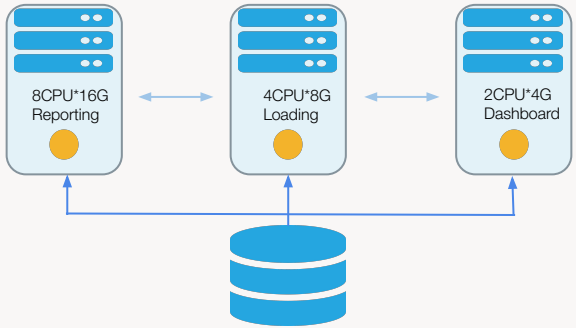
传统数仓架构

- Shared-Nothing - 弱弹性
- 存储、计算一体 - 弱弹性
- 资源控制粒度粗 - 成本高

成本(高) = Resource * Time

新一代弹性数仓架构

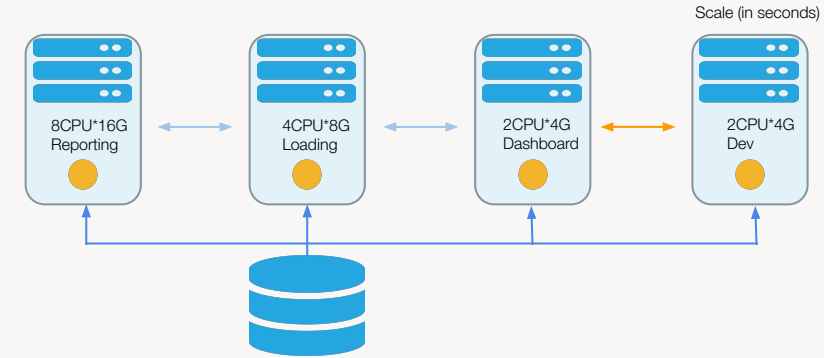
- Shared-Storage (Amazon S3, Azure Blob …)
- 真正存储、计算分离
- 实时弹性扩容和缩容
- 资源按需(Workload-Based)式调度
- 资源控制粒度细



AWS S3, Azure Blob, Google GCS, 阿里云 OSS, 腾讯云 COS, 华为云 OBS, IPFS ...

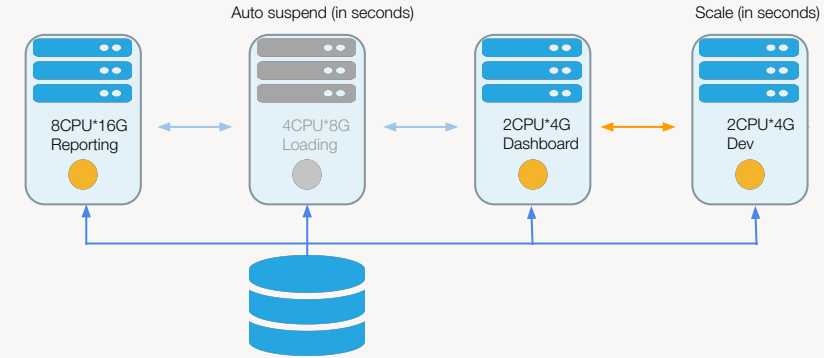
新一代弹性数仓架构

- Shared-Storage (Amazon S3, Azure Blob …)
- 真正存储、计算分离
- 实时弹性扩容和缩容
- 资源按需(Workload-Based)式调度
- 资源控制粒度细



新一代弹性数仓架构

- Shared-Storage (Amazon S3, Azure Blob …)
- 真正存储、计算分离
- 实时弹性扩容和缩容
- 资源按需(Workload-Based)式调度
- 资源控制粒度细



新一代弹性数仓架构

Databend

- Shared-Storage (Amazon S3, Azure Blob …) - 高弹性
- 真正存储、计算分离 - 高弹性
- 实时弹性扩容和缩容 - 高弹性
- 资源控制粒度细 - 成本低

成本(低) = Resource * Time

Databend

03

Databend 新一代云原生数仓设计

ClickHouse

Databend

- OS Warehouse
- 向量化计算，细节优化到位
- Pipeline 处理器和调度器
- MergeTree + Wide-Column 存储引擎
- 单机性能非常强悍
- 缺点：分布式能力较弱，运维复杂度高，不是为云设计

[ClickHouse Group By 为什么这么快]：https://bohutang.me/2021/01/21/clickhouse-and-friends-groupby/

[ClickHouse Pipeline 处理器和调度器]：https://bohutang.me/2020/06/11/clickhouse-and-friends-processor/

[ClickHouse 存储引擎技术进化与MergeTree]: https://bohutang.me/2020/06/20/clickhouse-and-friends-merge-tree-algo/

Snowflake

- Cloud Warehouse
- 多租户，存储、计算分离
- 基于对象存储便宜介质
- 弹性能力非常强悍，面向云架构设计
- 缺点：单机性能一般，重度依赖分布式

Databend

Databend = ClickHouse + Snowflake + Rust

- 借鉴 ClickHouse 向量化计算，提升单机计算性能
- 借鉴 Snowflake 存储、计算分离思想，提升分布式计算能力
- 借鉴 Git，MVCC 列式存储引擎，Insert/Read/Delete/Update/Merge
- 全面支持 HDFS/Cloud-based Object Storage 等 20 多种存储协议
- 基于便宜的对象存储也能方便的做实时性分析
- 完全使用 Rust 研发（20w+ loc），Day1 在 Github 开源
- 高弹性 + 强分布式，致力于解决大数据分析**成本**和**复杂度**问题

影响云原生数仓架构设计的因素与挑战

Databend

- Ingest 海量数据网络费用问题
  传统 INSERT 模式费用昂贵，需要一套基于 S3 的免费方案
- 对象存储不是为数仓而设计，延迟和性能如何平衡？
  Network-Bound -> IO-Bound -> CPU-Bound
- 如何让系统更加智能，根据查询模式自动创建索引？
  如何让某些场景的 Query 越跑越快...
- 如何面向 Warehouse + Datalake 双重需求设计？

如何解决 Ingest 数据网络费用问题

Databend

- 云端网络费用繁多，影响数仓架构设计
- 基于对象存储做 Stage 中转站: Internal Stage, External Stage
- Flink/Kafka 等数据链路面向 Stage，与 Databend 解耦
- 通过 COPY 命令实时从 Stage 实时 Ingest 数据到 Databend

```
COPY INTO mytable
  FROM 's3://mybucket/data.csv'
  CONNECTION = (
        ENDPOINT_URL = 'https://<endpoint-URL>'
        ACCESS_KEY_ID = '<your-access-key-ID>'
        SECRET_ACCESS_KEY = '<your-secret-access-key>')
  FILE_FORMAT = (type = CSV field_delimiter = ','  record_delimiter = '\n' skip_header = 1)
```

https://databend.rs/doc/sql-commands/dml/dml-copy-into-table

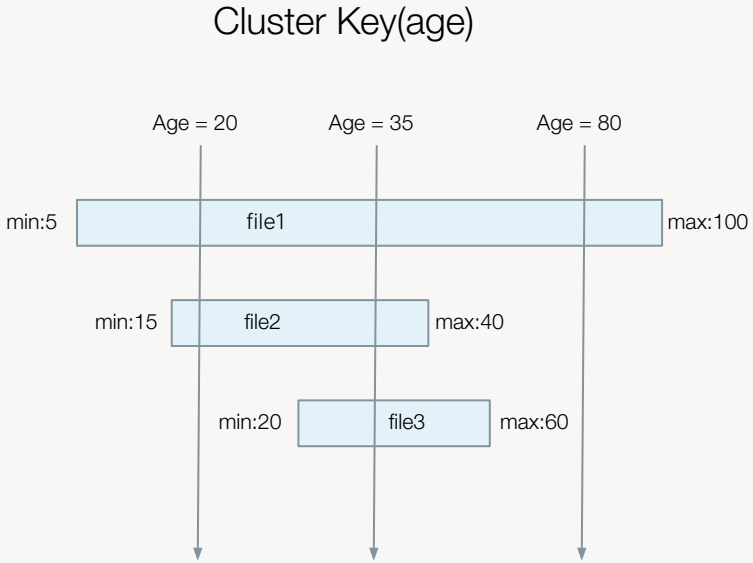
如何解决对象存储延迟与性能问题

Databend

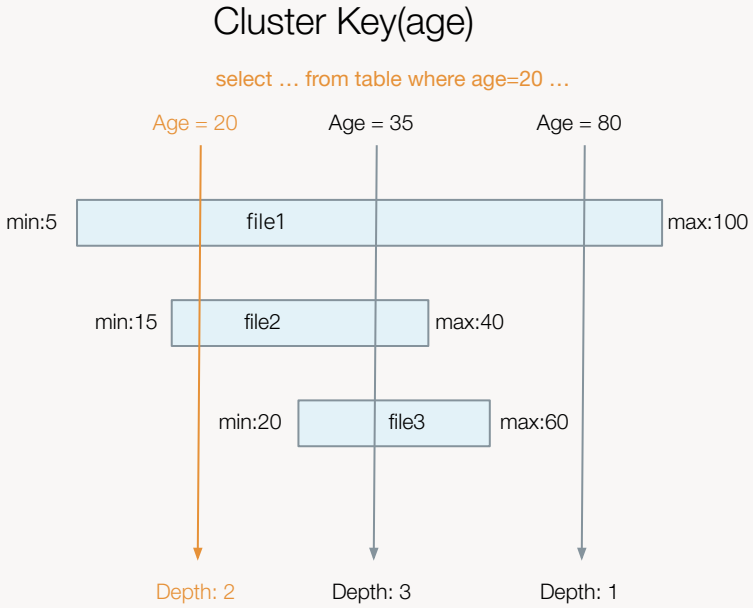
- 本地盘架构磁盘 IO 延迟低、带宽高
- AWS S3 延迟 10ms+、带宽 1GB/秒，但是便宜！
- 执行器(Execution)负载感知(Workload-Aware)，运行时动态扩展
- 索引加速：MinMax 索引，Bloom 索引，MAP/JSON 虚拟列索引
- 缓存分级：Meta&Index 缓存本地磁盘，Data 缓存可以到 Redis
- 复杂查询：GroupBy/OrderBy 中间结果“溢出”到对象存储

Automatic Tuning

Databend

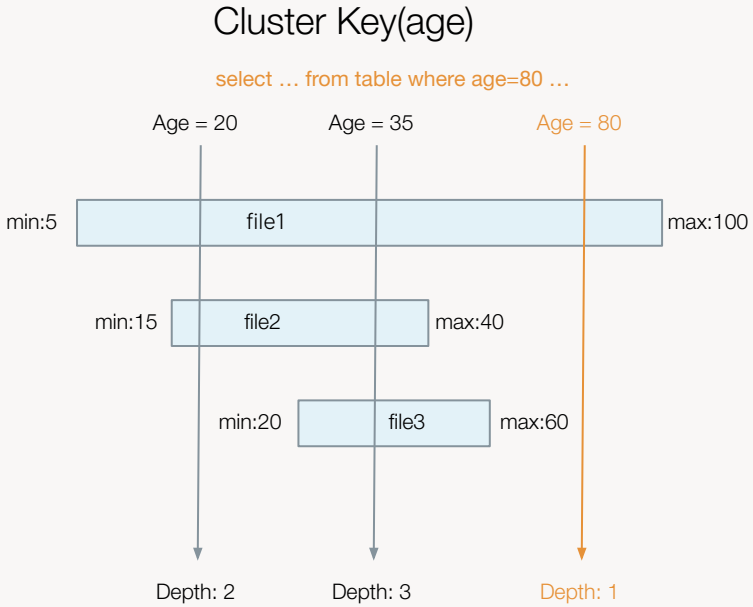


Automatic Tuning



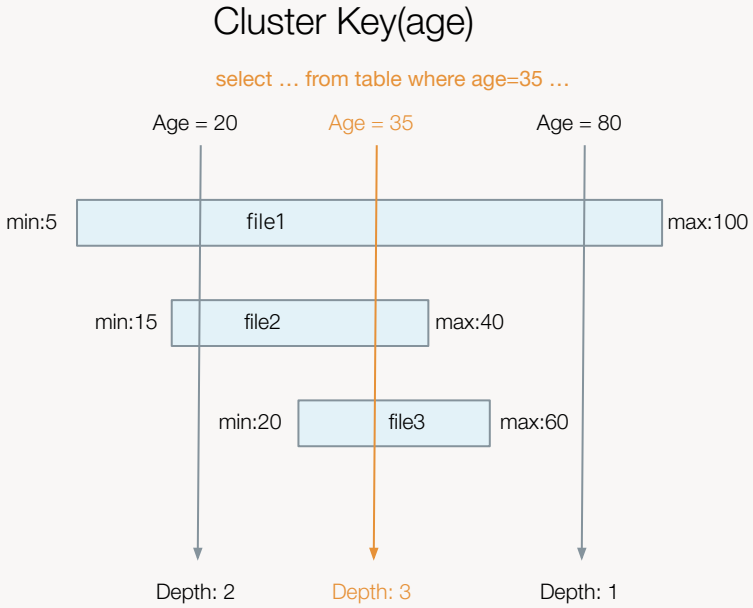
Automatic Tuning

Databend



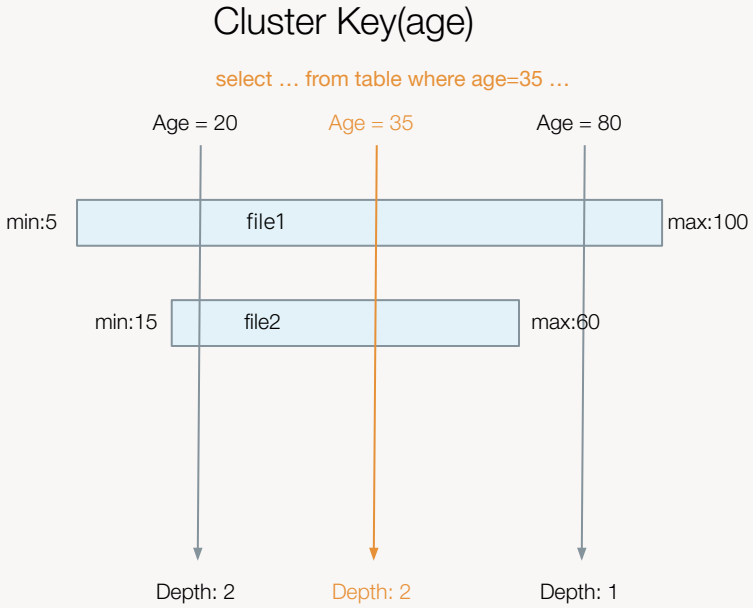
Automatic Tuning

Databend



Automatic Tuning

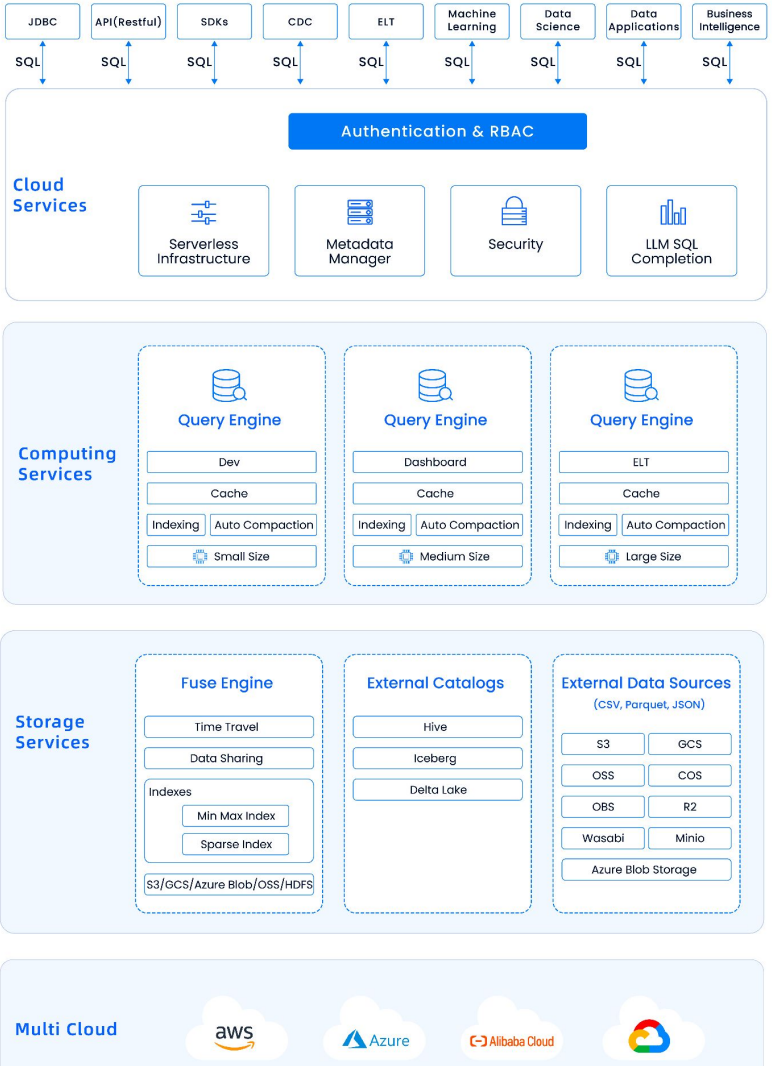
Databend



Databend Catalog: Hive + Iceberg

Databend

```
CREATE CATALOG my_iceberg
  TYPE="iceberg"
  URL='s3://path/to/iceberg'
  CONNECTION=(
    ACCESS_KEY_ID=...
    SECRET_ACCESS_KEY=...
    ...
  )
```


JDBC
API(Restful)
SDKs
CDC
ELT
Machine Learning
Data Science
Data Applications
Business Intelligence
SQL
Authentication & RBAC
Cloud Services
Serverless Infrastructure
Metadata Manager
Security
LLM SQL Completion
Computing Services
Query Engine
Dev
Cache
Indexing
Auto Compaction
Small Size
Query Engine
Dashboard
Cache
Indexing
Auto Compaction
Medium Size
Query Engine
ELT
Cache
Indexing
Auto Compaction
Large Size
Storage Services
Fuse Engine
Time Travel
Data Sharing
Indexes
Min Max Index
Sparse Index
S3/GCS/Azure Blob/OSS/HDFS
External Catalogs
Hive
Iceberg
Delta Lake
External Data Sources
(CSV, Parquet, JSON)
S3
GCS
OSS
COS
OBS
R2
Wasabi
Minio
Azure Blob Storage
Multi Cloud
aws
Azure
Alibaba Cloud

Databend

Databend 生产降本增效效果

- 替换 Trino/Presto 场景成本降低了 **75%**
- 替换 Elasticsearch 场景成本降低了 **90%**
- 归档场景成本降低了 **95%**
- 日志和历史订单分析场景成本降低了 **75%**
- ~700TB/天（2023.3 统计）在使用 Databend 写入公有云对象存储
- 用户来自欧洲、北美、东南亚、非洲、中国等地，节省数百万美元
- 开源、开放，运维简单、分钟级部署，为云端海量数据分析而设计

Databend

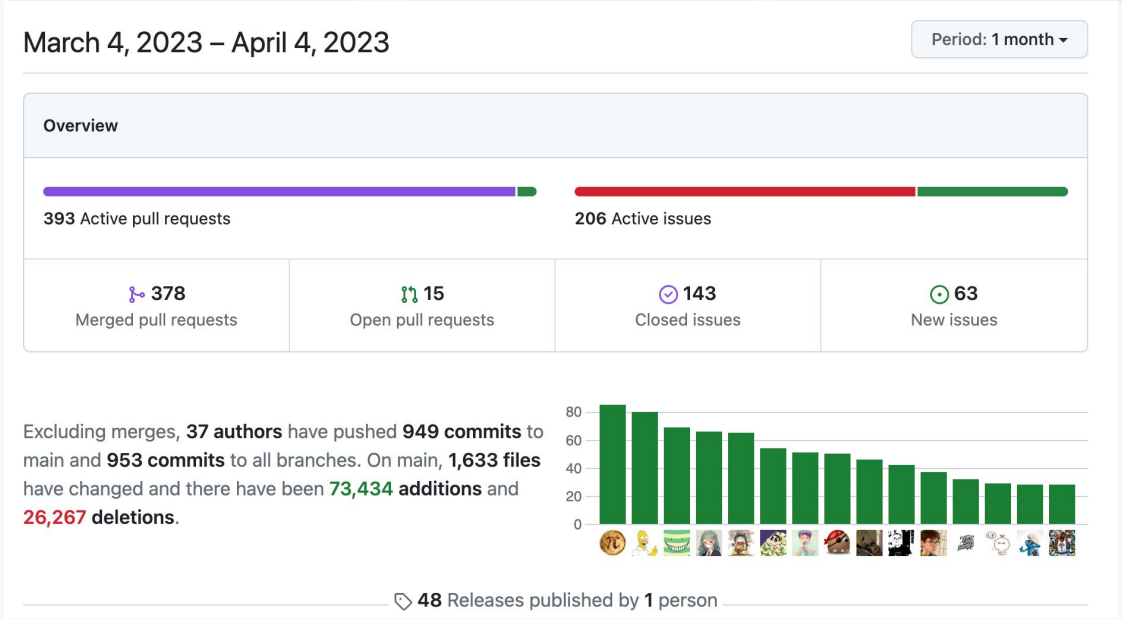
04

Databend 开源社区

Databend 开源社区

Databend

5.7K Stars	**150+** Contributors	迭代非常快



https://github.com/datafuselabs/databend

Databend 开源社区

社区贡献者：

SAP
Yahoo
Fortinet
Shopee
PingCAP
Alibaba
Tencent
ByteDance
EMQ
快手

Databend 社区被**顶级需求、顶级场景驱动**

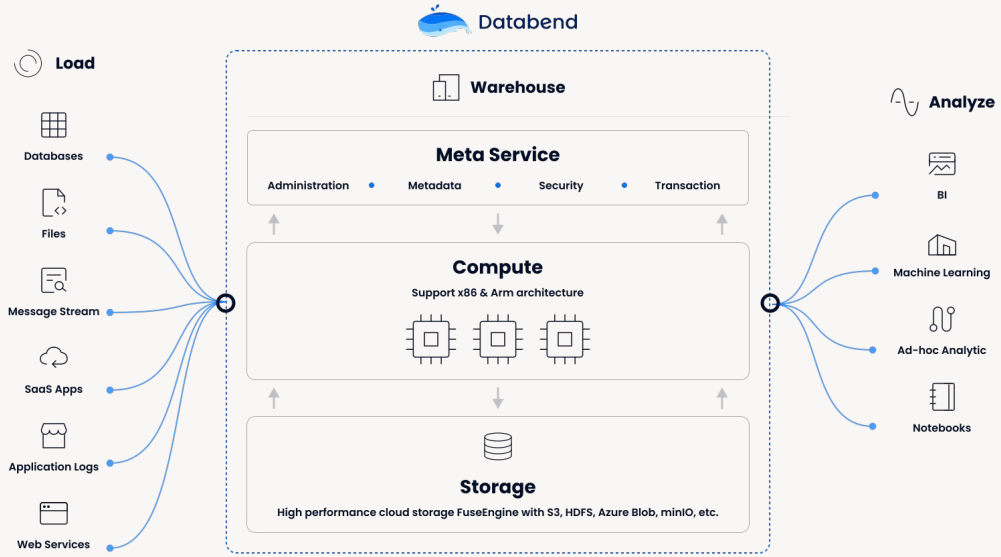
Databend 部分生产用户

Databend

More …

Databend 体验：On-Premises, Serverless

- On-Premises
  社区版：https://databend.rs
- Serverless Cloud
  海外(AWS)　https://app.databend.com
  国内(阿里云) https://app.databend.cn



Databend

谢谢观看

THANKS FOR WATCHING