DTCC 2010 2022

第十三届中国数据库技术大会

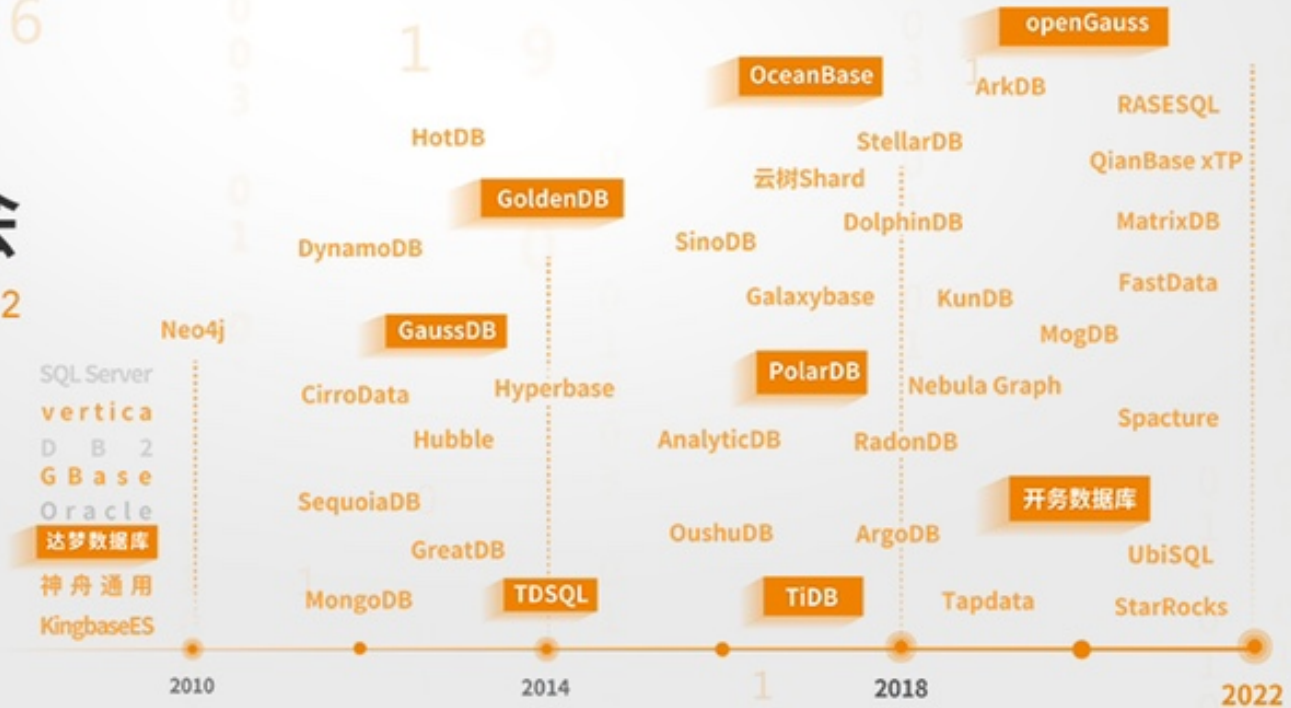
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



CnosDB 2.0 云原生时序数据库

随着万物互联时代的发展，时序数据库成为了物联网行业的底层基础架构。传统的时序数据库因为时间线膨胀和数据采样频率提高，产生了比较大的系统瓶颈；同时，时序数据库在云原生环境上的部署与资源管理也成为企业面临的挑战。CnosDB 2.0 是一款云原生时序数据库，具有高可用、高性能、高压缩比的特点。本报告主要讲述在云原生时代时序数据库面临的挑战、构建云原生时序数据库的技术以及时序数据库未来的发展前景。

郑博　北京诺司时空科技有限公司　CEO

主讲人介绍

郑　博

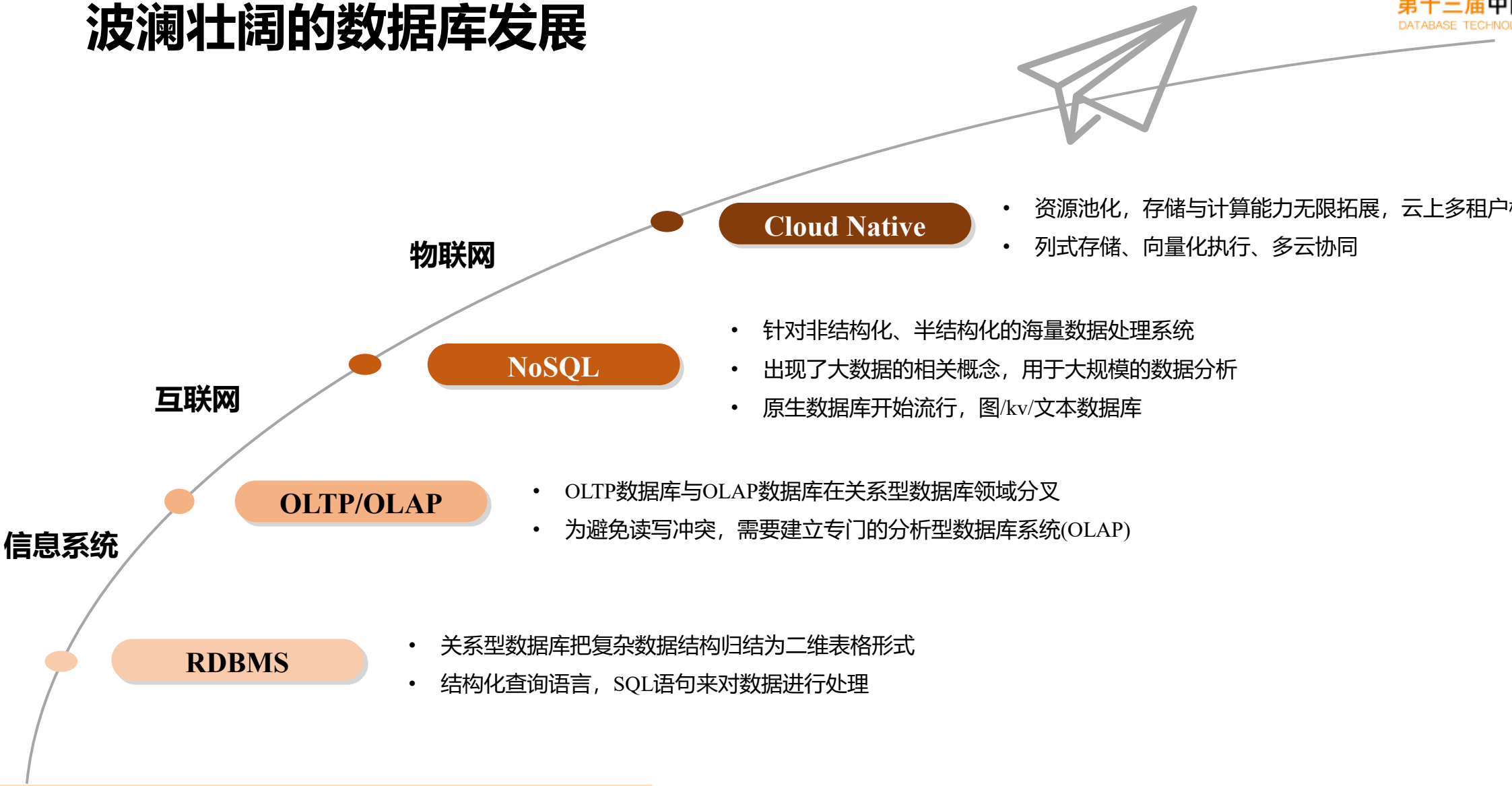
CnosDB 时序数据库开源社区发起人

北京市高层次人才专家，北京市朝阳区高层次人才专家，CCF数据库专委会执行委员。卡内基梅隆大学计算机科学硕士。在国内外重要会议及期刊发表论文10余篇，授权发明数十项。在美国曾参与工业制造业OEE平台创业，负责系统架构与时序数据平台的研发；参与多个开源社区如MongoDB、Alfresco的组织与代码开发贡献。回国后曾作为主要成员参与企业服务及人工智能相关行业创业。中国企业服务联盟特聘专家、中国CIO联盟特聘专家、中国云体系产业创新战略联盟理事。

Part One

概念解析

波澜壮阔的数据库发展



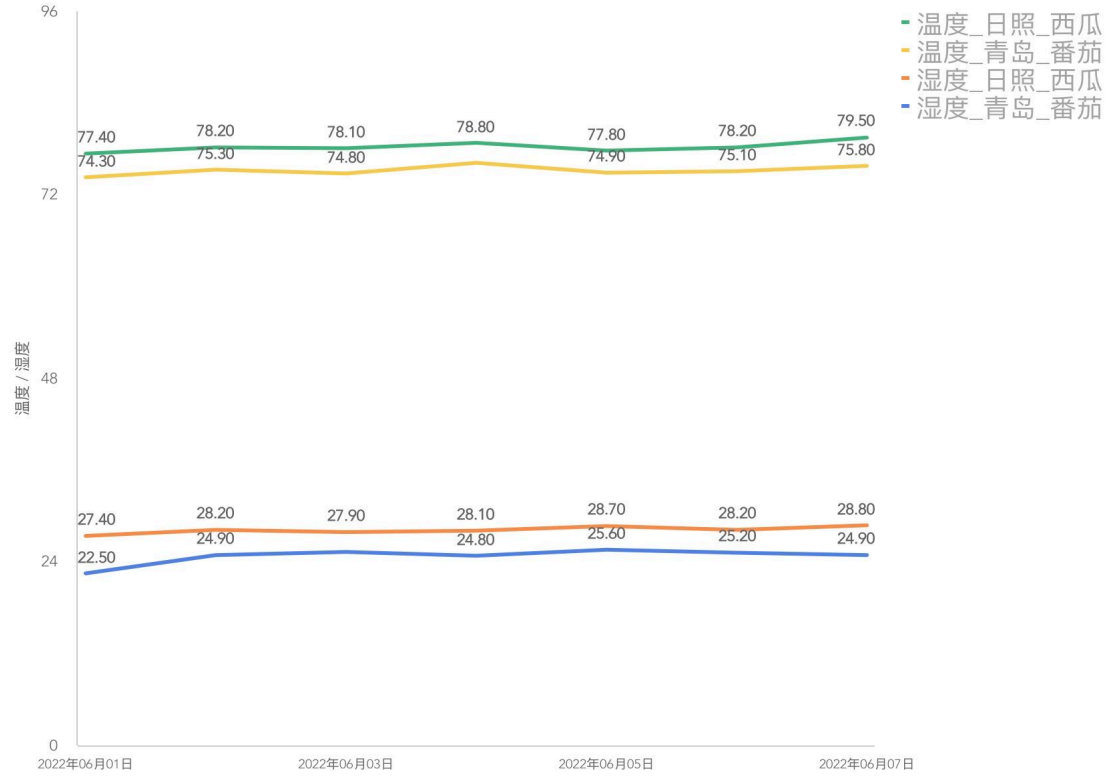
- 资源池化，存储与计算能力无限拓展，云上多租户模型
- 列式存储、向量化执行、多云协同

- 针对非结构化、半结构化的海量数据处理系统
- 出现了大数据的相关概念，用于大规模的数据分析
- 原生数据库开始流行，图/kv/文本数据库

- OLTP数据库与OLAP数据库在关系型数据库领域分叉
- 为避免读写冲突，需要建立专门的分析型数据库系统(OLAP)

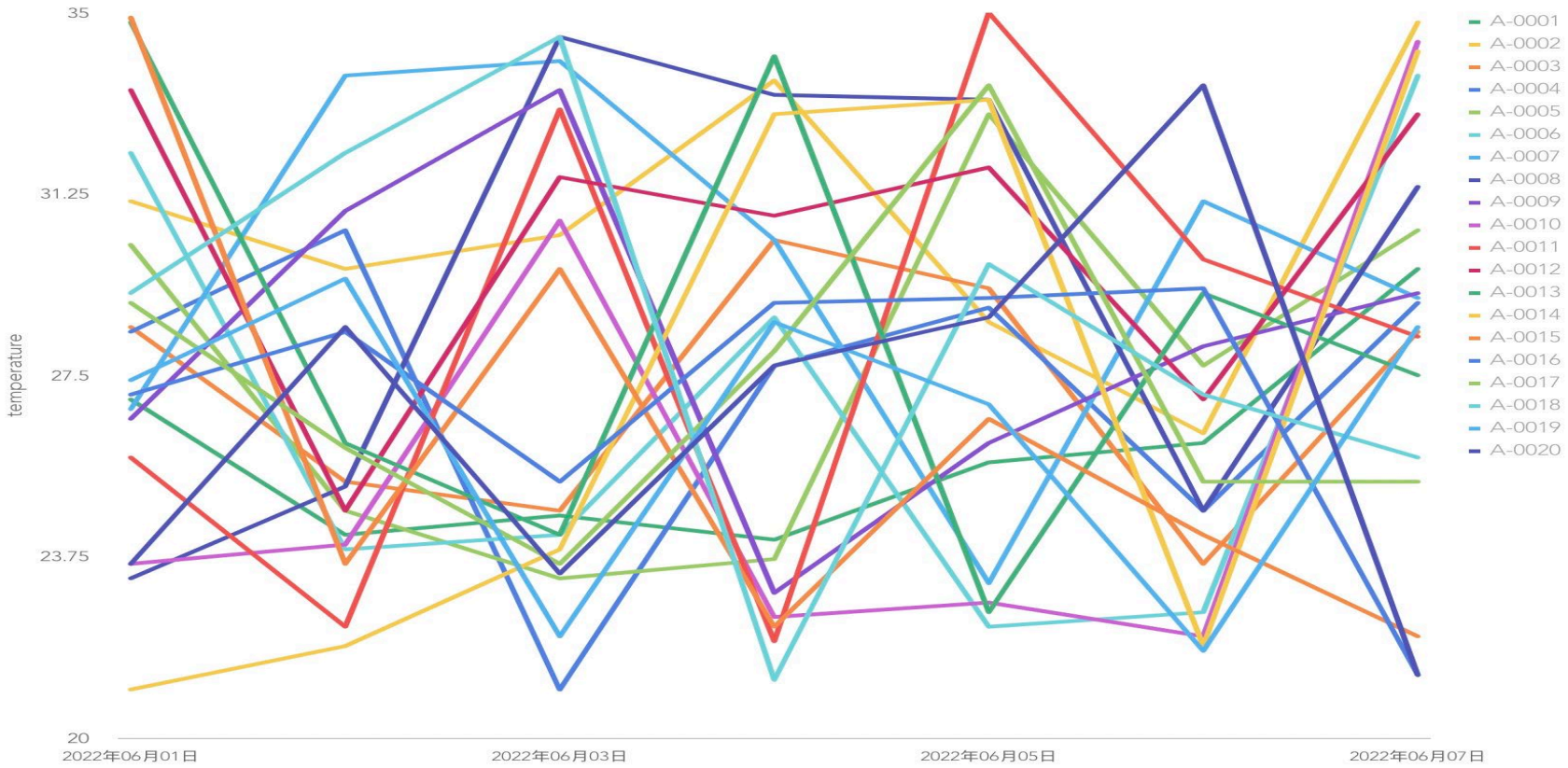
- 关系型数据库把复杂数据结构归结为二维表格形式
- 结构化查询语言，SQL语句来对数据进行处理

什么是时序数据



时间	地区	作物	温度	湿度
2022/6/1	青岛	番茄	74.3	22.5
2022/6/2	青岛	番茄	75.3	24.9
2022/6/3	青岛	番茄	74.8	25.3
2022/6/4	青岛	番茄	76.2	24.8
2022/6/5	青岛	番茄	74.9	25.6
2022/6/6	青岛	番茄	75.1	25.2
2022/6/7	青岛	番茄	75.8	24.9
2022/6/1	日照	西瓜	77.4	27.4
2022/6/2	日照	西瓜	78.2	28.2
2022/6/3	日照	西瓜	78.1	27.9
2022/6/4	日照	西瓜	78.8	28.1
2022/6/5	日照	西瓜	77.8	28.7
2022/6/6	日照	西瓜	78.2	28.2
2022/6/7	日照	西瓜	79.5	28.8

时序数据形态


35
31.25
27.5
23.75
20
temperature
2022年06月01日
2022年06月03日
2022年06月05日
2022年06月07日
A-0001
A-0002
A-0003
A-0004
A-0005
A-0006
A-0007
A-0008
A-0009
A-0010
A-0011
A-0012
A-0013
A-0014
A-0015
A-0016
A-0017
A-0018
A-0019
A-0020

什么是时序数据库

物联网时代数据存储基础设施

信息系统

RDBMS

SQL

Microsoft SQL Server

ORACLE

IBM DB2

MySQL

互联网

BIG DATA

NoSQL

mongoDB.

APACHE HBASE

cassandra

物联网

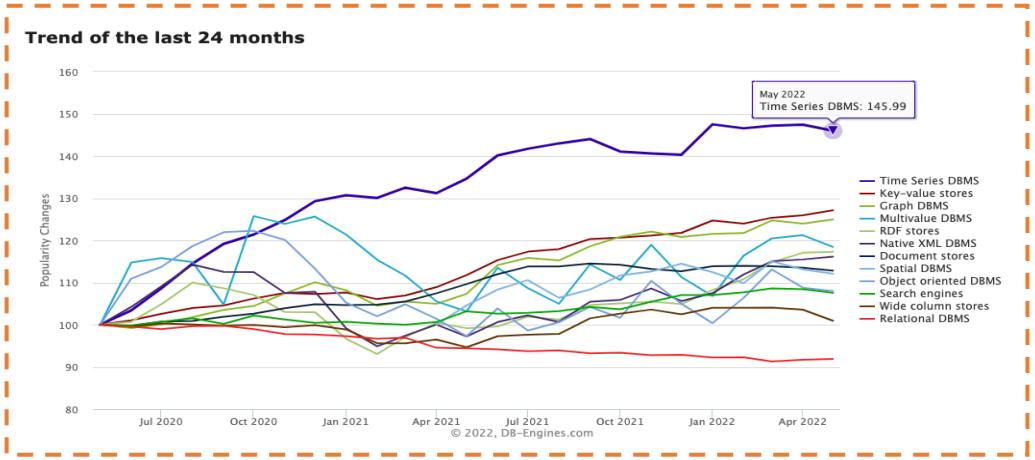
TSDB

CLOUD NATIVE

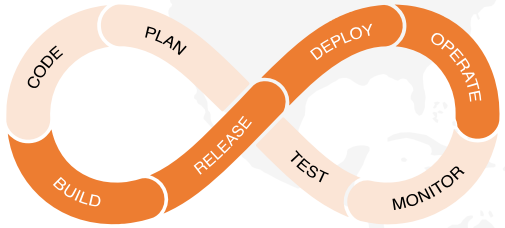
influxdb

CnosDB

Timescale



时序数据库的场景跃迁



监控从规划、开发、集成和测试、部署到运营的整个开发流程

互联网

物联网

铁路地铁

轨道监测/车辆监测
路网优化

桥梁

桥梁结构监控
安全预警

核电站

核电DCS安全监控

智慧电力

电网监控
电网调度自动化

数控机床

精密仪器监控

燃气

漏气预警/用气趋势
派单效率分析

锅炉

工业锅炉远程监控
供热锅炉优化

化工

油田生产管理
灾害预警

水务海洋

雨量监测
水质预警

万物智联

WEB 3.0/4.0/5.0

区块链上交易

车联网与无人驾驶

VR/AR/可穿戴

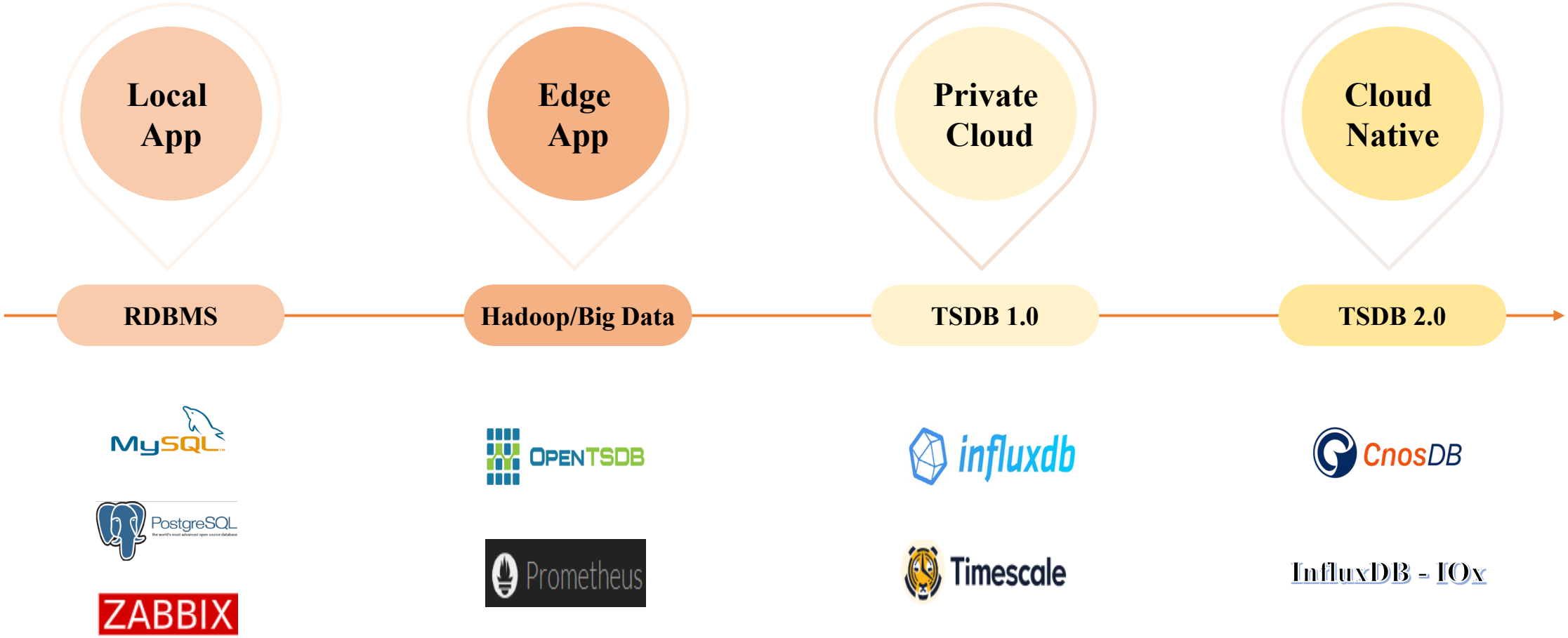
消费级电子设备

脑机接口

Part Two

时序数据库挑战

时序数据库发展


Local App
Edge App
Private Cloud
Cloud Native
RDBMS
Hadoop/Big Data
TSDB 1.0
TSDB 2.0
MySQL
PostgreSQL
ZABBIX
OPENTSDB
Prometheus
influxdb
Timescale
CnosDB
InfluxDB - IOx

时序数据库面临的问题

问题一

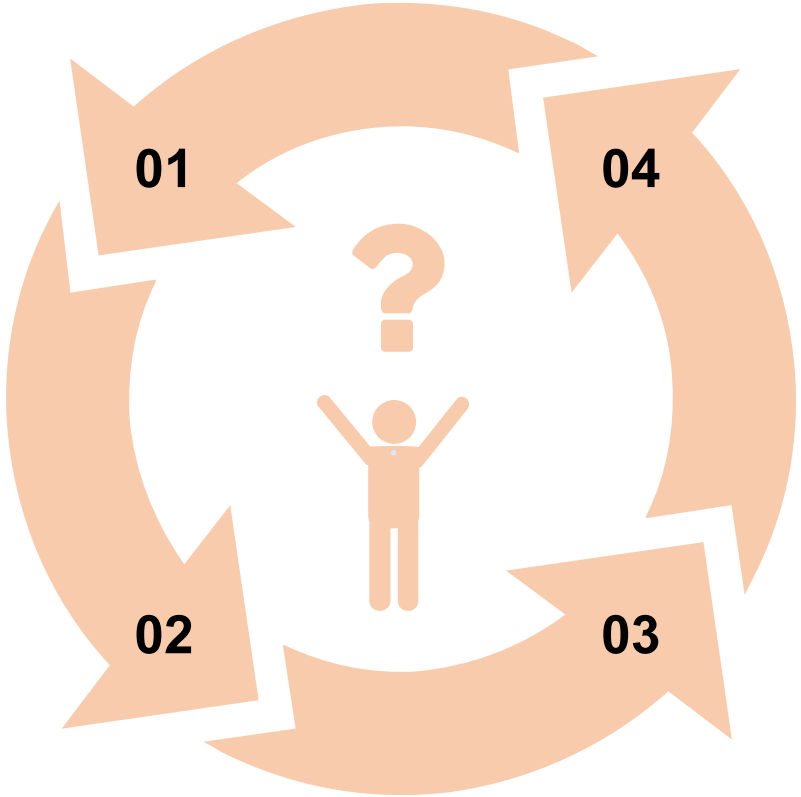
时间线膨胀和数据采样频率

提高带来双重写入性能要求

问题二

大量时间线筛选和复杂

聚合带来查询延迟过大



问题四

在时间序列维度上数据

偏斜带来的负载不均衡

问题三

高昂学习成本

使用成本和部署成本

云环境下需要解决的问题

问题一

云环境下资源的管理

问题二

数据格式无法与第三方适配

问题三

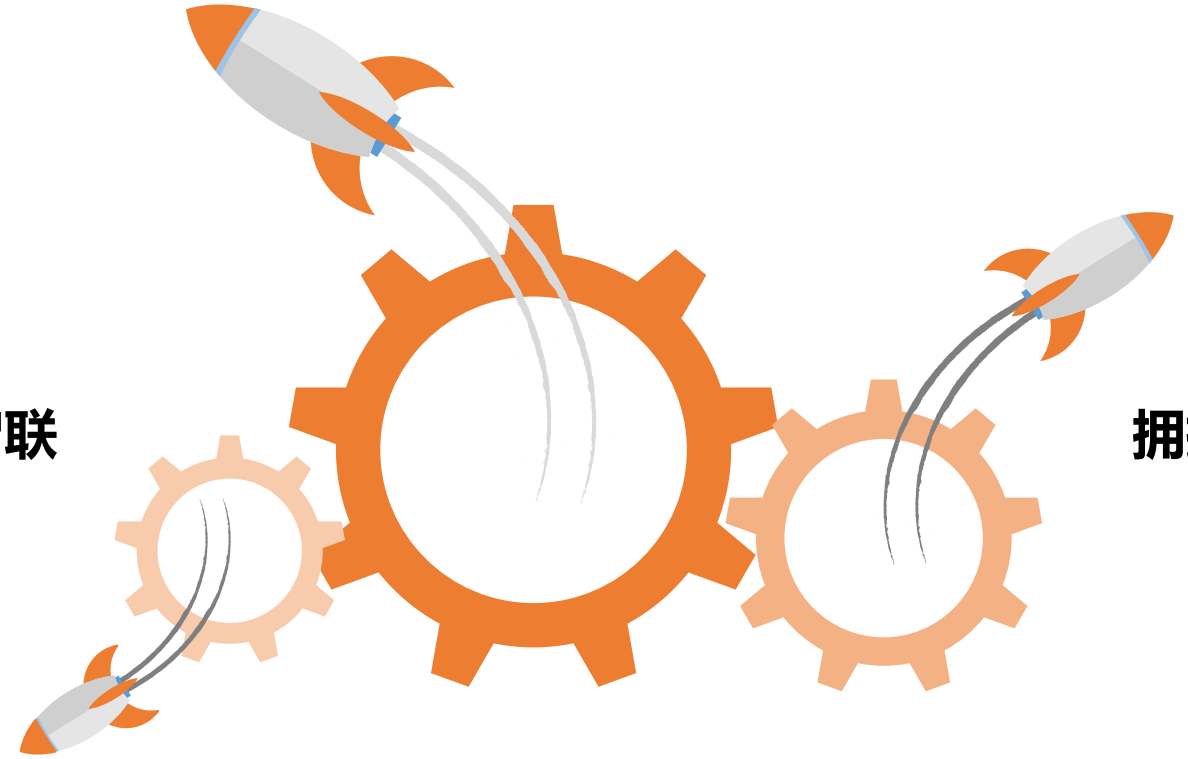
云环境下支持高可用性

问题四

云上查询的实时性要求和
日益复杂的数据分析需求

面向下一代的时序数据库

开启万物智联



拥抱云原生

需要更好的时序数据库

Part Three

CnosDB

云原生时序数据库

CnosDB 2.0 设计理念

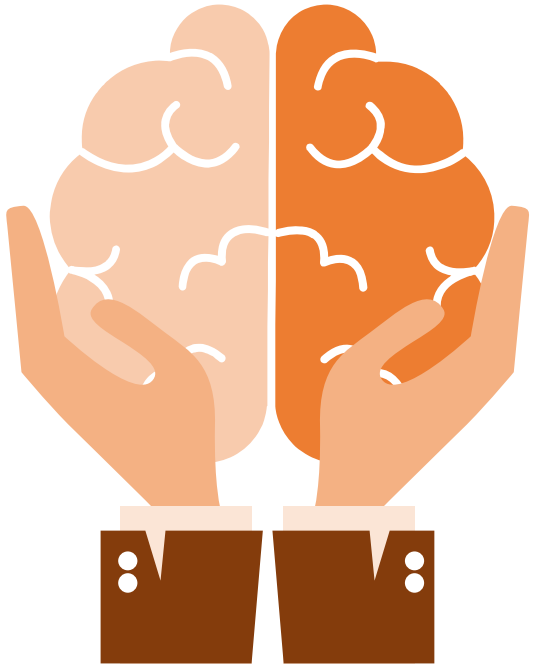
产品

Serverless

DBaaS

Everything on Cloud

边端与云融合



非产品

没有Silver Bullets

专业的人做专业的事儿

承认自己菜是进步的先决条件

不相信All in One

面向未来，不破不立

CnosDB 2.0 设计目标

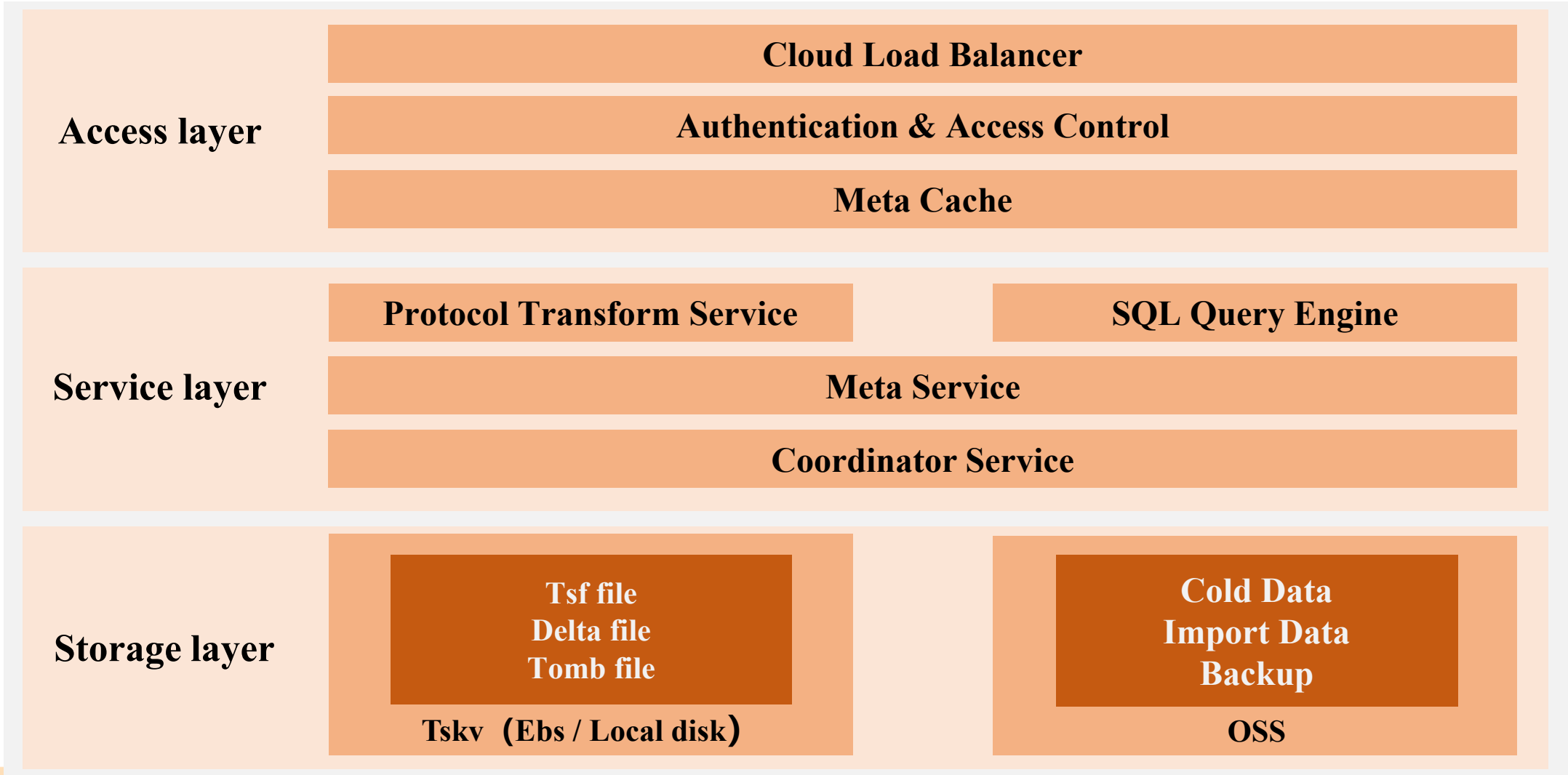
数据库

- 超强的扩展性
- 计算存储分离
- 平衡存储性能与成本
- 查询引擎支持矢量化查询
- 支持多种时序协议
- 支持外部全生态

云原生

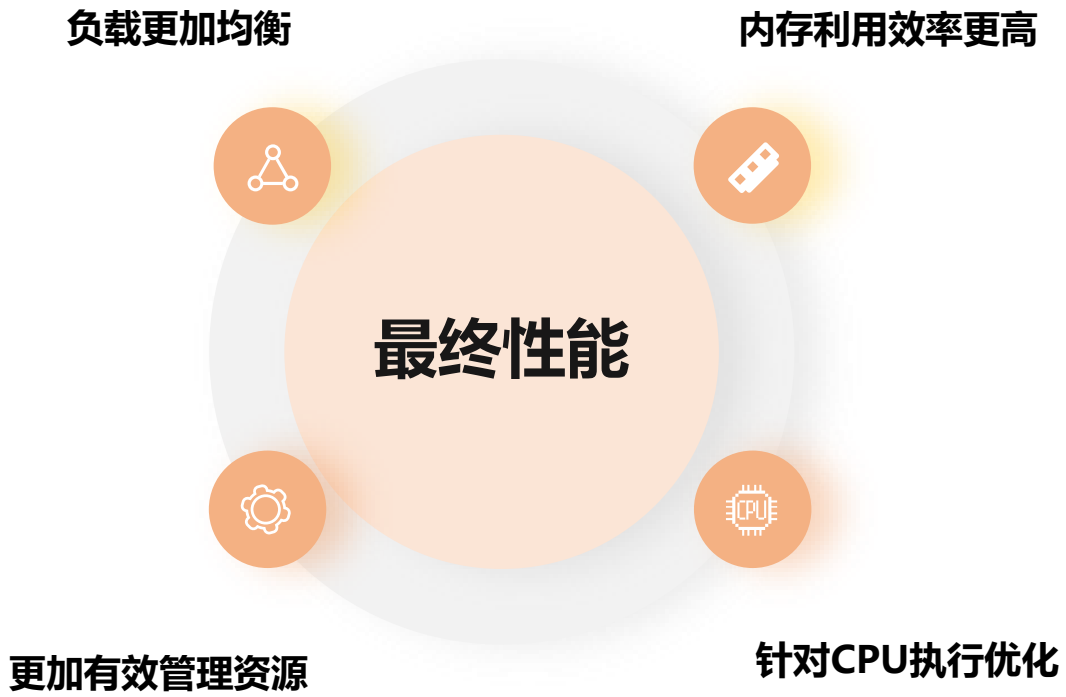
- 支持云原生
- 云上高可用
- 原生支持多租户, 按量付费
- CDC
- 需求可配置
- 云边端协同
- 云上生态融合

整体框架


Access layer
Cloud Load Balancer
Authentication & Access Control
Meta Cache
Service layer
Protocol Transform Service
SQL Query Engine
Meta Service
Coordinator Service
Storage layer
Tsf file
Delta file
Tomb file
Tskv（Ebs / Local disk）
Cold Data
Import Data
Backup
OSS

语言选择“借 Rust 东风”

语言及模型选择	
Rust	无GC，获得对内存生命周期更精细的控制能力。性能高，表达能力强：它零成本抽象，工程效率高，又不影响运行性能。智能编译器，在编译前发现问题，而不是把问题留在运行时。
Tokio	一个异步IO的运行时，提供了I/O、网络、调度、定时器等等异步编程所需的功能和工具，性能和功能异常强大。
Glommio	一个基于thread-per-core级别并使用 IO_uring 实现的专用运行的工具。可以配合磁盘直读与直写操作。
Run-To-Completion	RTC是一种调度模型，其中每个任务运行到完成或者将控制权交还给调度器，可以优化CPU的利用效率，让负载更加均衡。



Rust 优点

Rust编写的程序表现力和性能好，其拥有高级函数式语言的大部分特性

Rust编译期可以在编译时跟踪程序中资源的变量，并在没有GC的情况下完成所有这些操作。

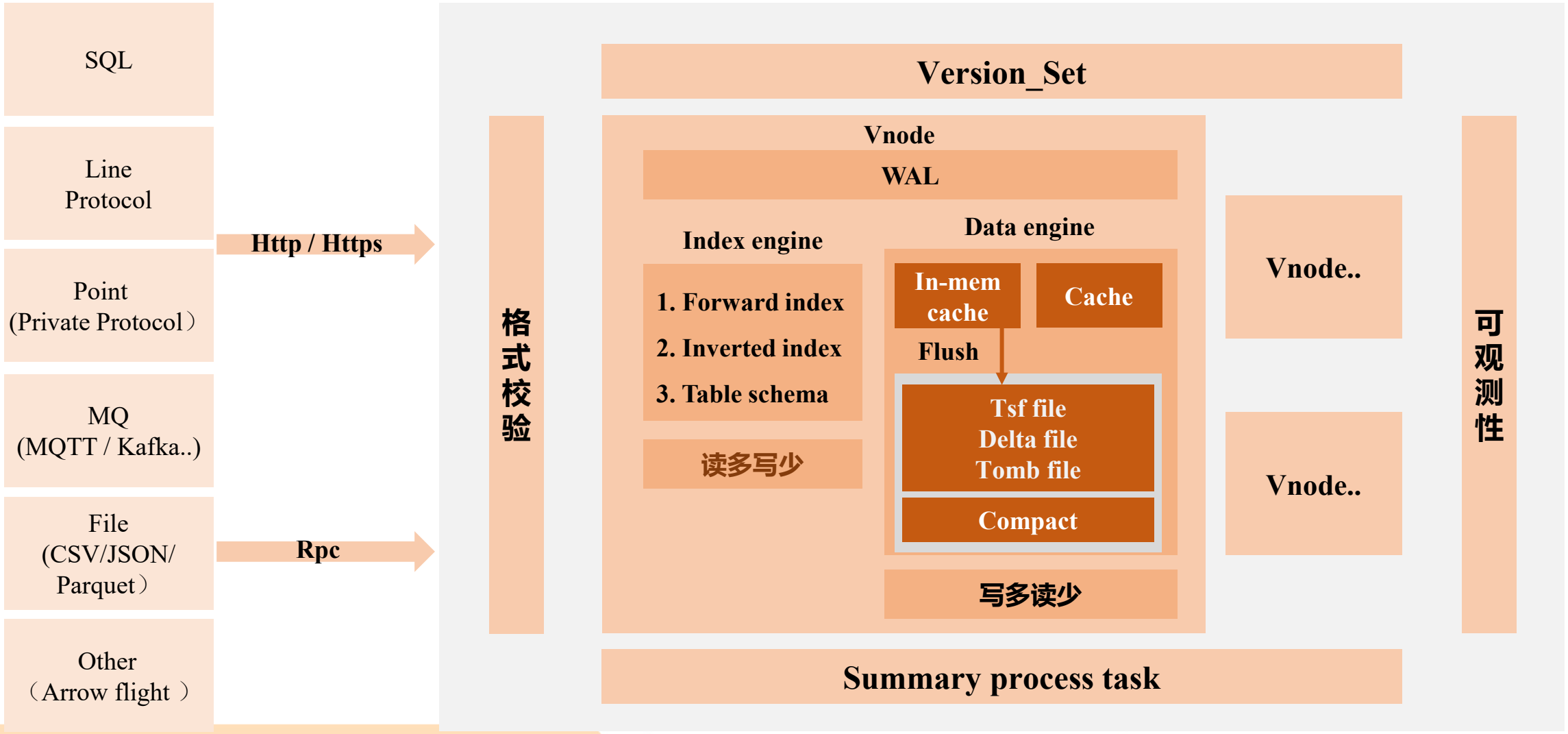
抽象在运行时没有任何成本，只在编译时。

Rust是并发安全的，其具有、API和抽象能力，使得编写正确和安全的并发代码变得非常容易。

Rust 的一些组件

序号	名称	说明
1	arrow	Apache Arrow 列式内存分析库的 Rust 实现
2	arrow-datafusion	Apache Arrow 的扩展查询执行框架，支持 SQL 和 DataFrame API
3	futures	异步编程的一些特性
4	parking_lot	提供更好的 Mutex / RwLock / Condvar / Once，并提供Reentrant Mutex
5	parquet	Apache Arrow 的数据读写部分
6	sqlparser	解析 ANSI:SQL 2011 为抽象语法树
7	tikv-jemalloc-ctl	对 jemalloc::mallctl*() 的安全包装
8	tikv-jemalloc-sys	为 Rust 创建 C 语言库 jemalloc 的绑定
9	libc	支持通过 FFL 直接绑定平台的系统库
10	tokio	事件驱动的异步网络传输库
11	tokio-stream	为 tokio 库添加流处理功能
12	tonic	grpc 的 Rust 实现
13	tracing	跟踪、收集诊断信息

存储引擎


SQL
Line Protocol
Point (Private Protocol）
MQ (MQTT / Kafka..)
File (CSV/JSON/ Parquet）
Other （Arrow flight ）
Http / Https
Rpc
格式校验
Version_Set
Vnode
WAL
Index engine
1. Forward index
2. Inverted index
3. Table schema
读多写少
Data engine
In-mem cache
Cache
Flush
Tsf file
Delta file
Tomb file
Compact
写多读少
Vnode..
Vnode..
可观测性
Summary process task

存储引擎与压缩

压缩算法考虑压缩比
和压缩效率的均衡

类型支持的
压缩算法（如右表）

用户的场景不同
用户灵活地选择压缩

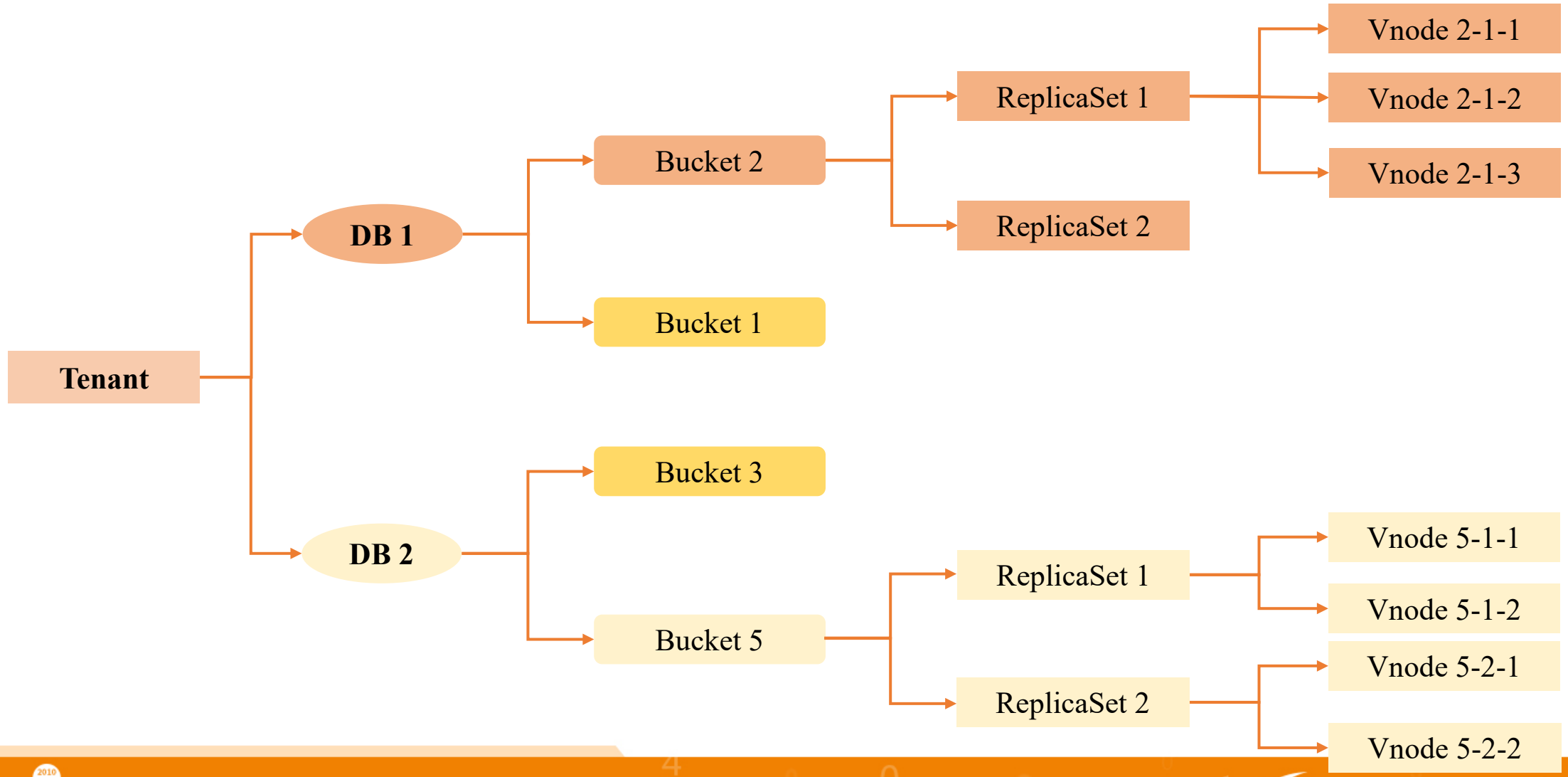
在创建列不指定压缩算法的情况下
我们使用默认的压缩算法

可以设置为无压缩算法
可以适用压缩比不敏感的情况

文档链接
https://docs.cnosdb.com/guide/design/compress.html

类型	压缩算法	最高压缩比
timestamp/u64/i64	Delta, Quantile	40x
f64	Gorilla, Quantile	1.5x
string	Gzip, Bzip, Zstd, Snappy, Zilb	11x
bool	Bitpack	8x

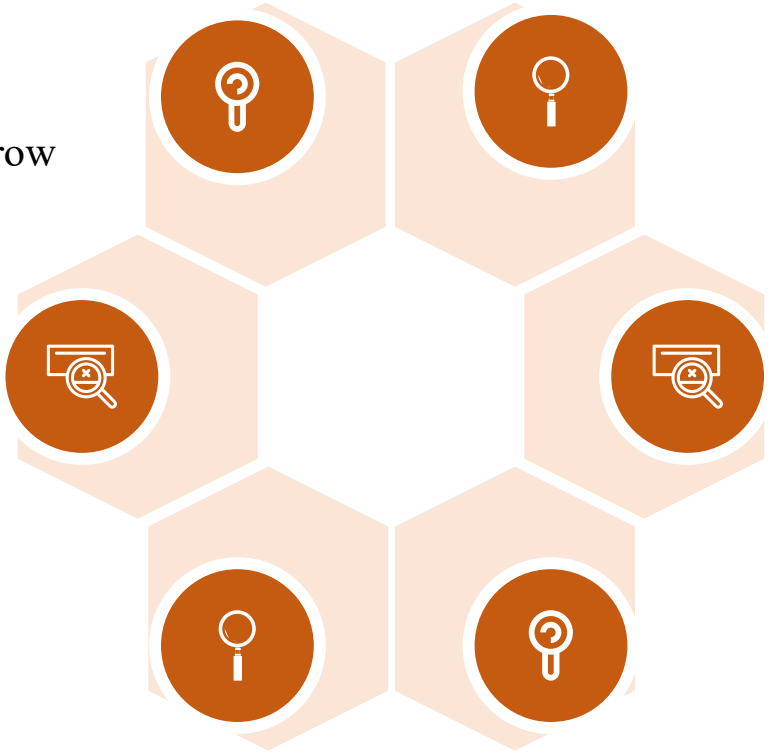
数据目录与结构


Tenant
DB 1
DB 2
Bucket 2
Bucket 1
Bucket 3
Bucket 5
ReplicaSet 1
ReplicaSet 2
ReplicaSet 1
ReplicaSet 2
Vnode 2-1-1
Vnode 2-1-2
Vnode 2-1-3
Vnode 5-1-1
Vnode 5-1-2
Vnode 5-2-1
Vnode 5-2-2

查询引擎

基于RUST实现

融合 DataFusion 与 Apache Arrow



支持大部分的SQL-92

高性能

高扩展性

支持多种数据源

支持各种函数

甚至是复杂的分析函数

拥抱 DataFusion

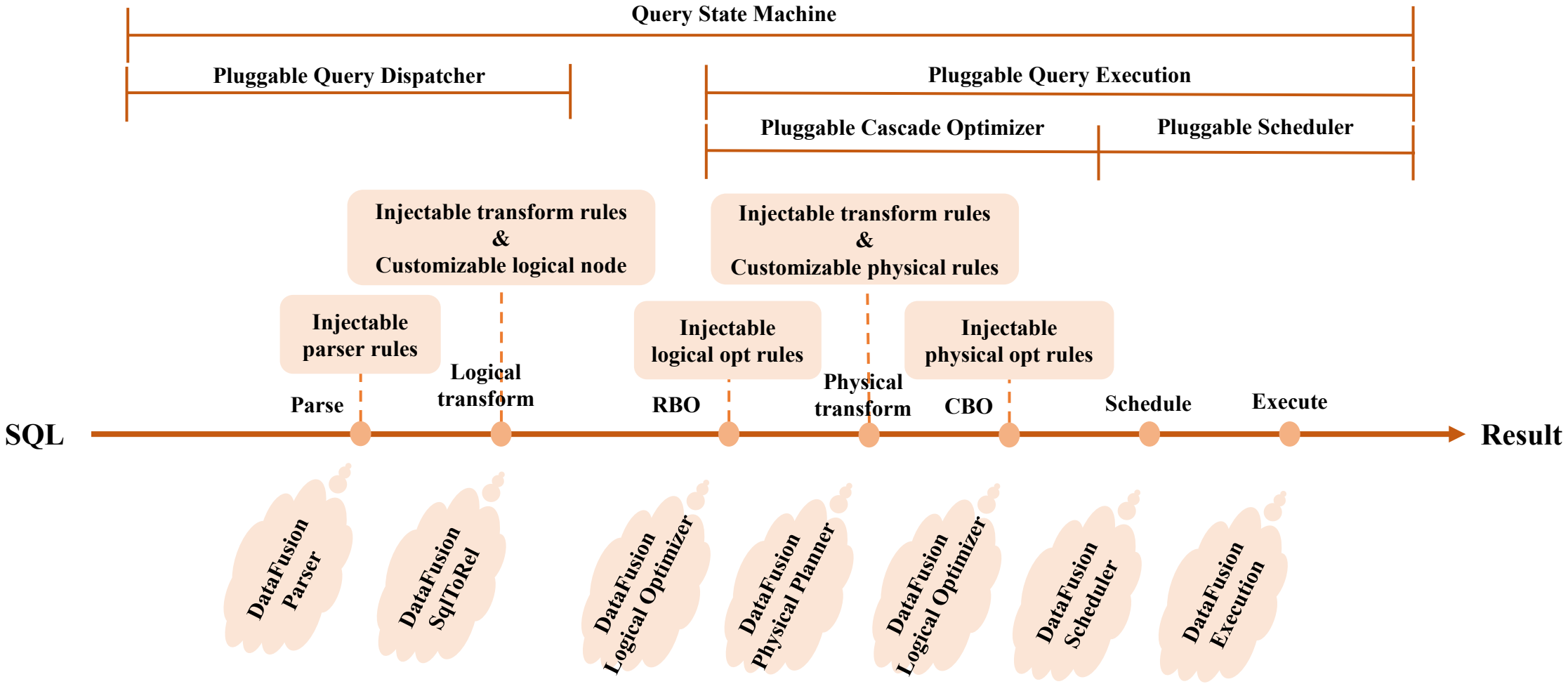
高性能

易于嵌入

支持对象存储

高质量

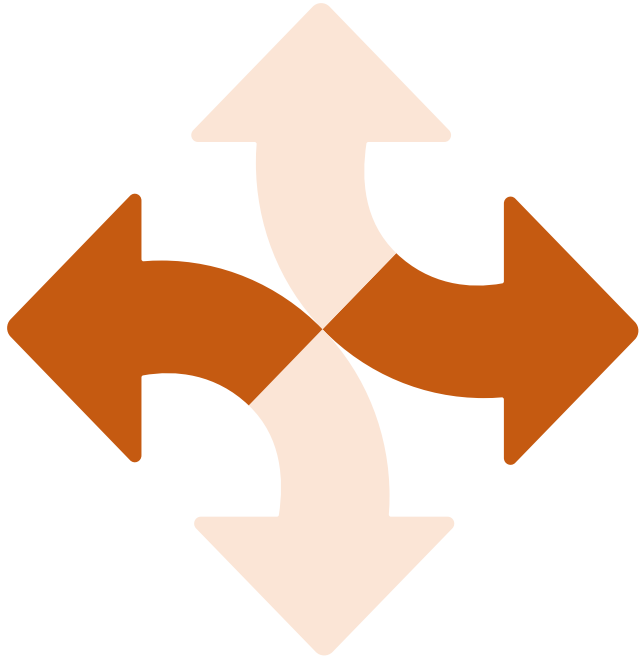
使用 DataFusion



分布式

基于Quorum NRW 实现可调一致性级别

满足不同场景用户的需求



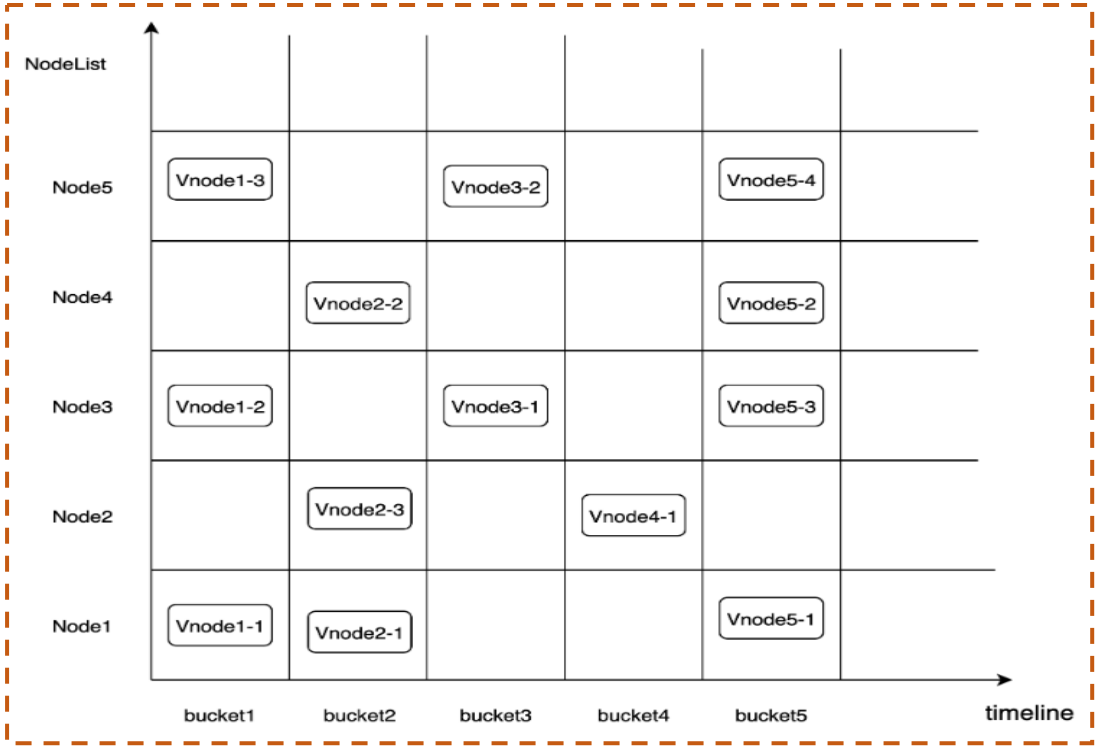
实现Hinted-handoff机制

节点对等写入无中心化节点

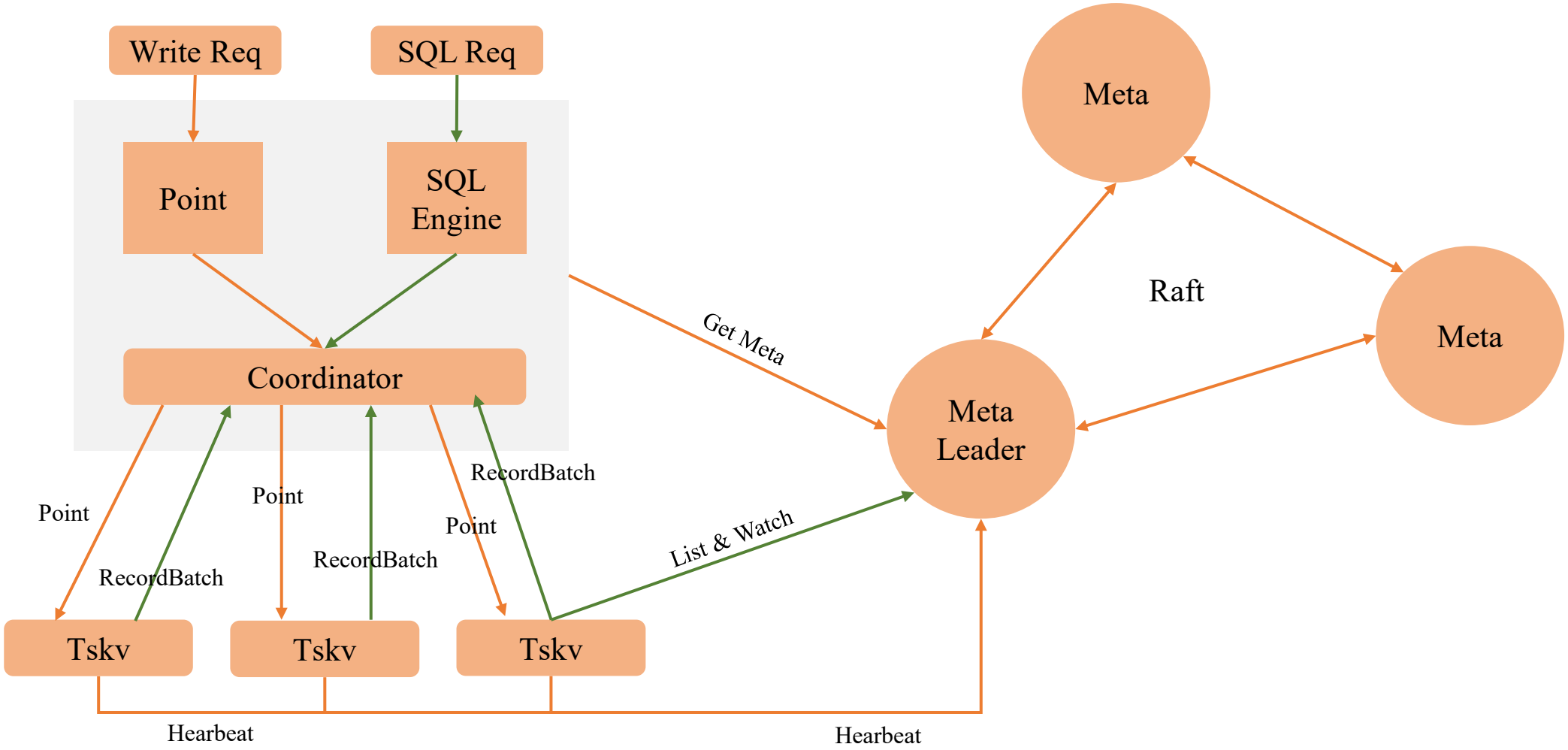
实现Anti-entropy反熵机制

分布式数据分片规则

- CnosDB 基于 time-range 的分片规则通过 db+time_range 来切分成不同的bucket
- Bucket是一个虚拟逻辑单元，bucket分为不同的复制组，每个复制组有一组Vnode。
- Vnode是一个具体运行单元，分布到一个具体的Node上，每个Vnode是单独Lsm tree。



数据流


Write Req
SQL Req
Point
SQL Engine
Coordinator
Get Meta
Meta
Raft
Meta
Meta Leader
Point
RecordBatch
Point
RecordBatch
Point
RecordBatch
List & Watch
Tskv
Tskv
Tskv
Hearbeat
Hearbeat

Part Four

云原生

什么是云原生

云原生是在云计算环境中构建、部署和管理现代应用程序的软件方法。

高扩展性

云原生分布式数据库与底层的云计算基础设施分离，所以能够灵活及时调动资源进行扩容缩容，以从容应对流量激增带来的压力，以及流量低谷期因资源过剩造成的浪费。生态兼容的特点，也让云原生数据库具备很强的可迁移性。

易用性

云原生分布式数据库易于使用，它的计算节点在云端部署，可以随时随地从多前端访问。因其集群部署在云上，通过自动化的容灾与高可用能力，单点失败对服务的影响非常小。当需要升级或更换服务时，还可以对节点进行不中断服务的轮转升级。

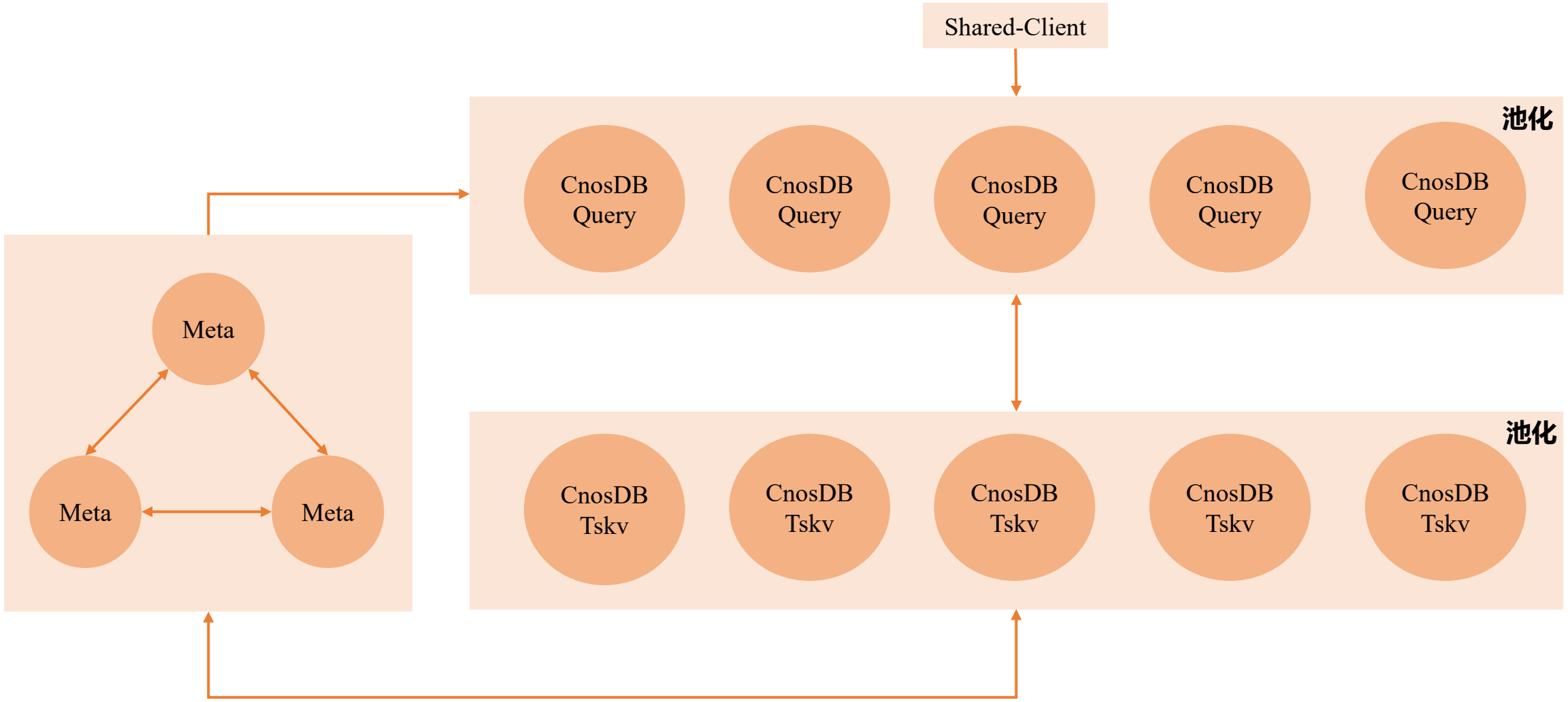
快速迭代

云原生分布式数据库中的各项服务之间相互独立，个别服务的更新不会对其他部分产生影响。此外，云原生的研发测试和运维工具高度自动化，也就可以实现更加敏捷的更新与迭代。

节约成本

建立数据中心是一项独立而完备的工程，需要大量的硬件投资以及管理和维护数据中心的专业运维人员，持续运维也会造成很大的财务压力。云原生分布式数据库以较低的前期成本，获得一个可扩展的数据库，实现更优化的资源分配。

整体架构


Shared-Client
池化
CnosDB Query
CnosDB Query
CnosDB Query
CnosDB Query
CnosDB Query
Meta
Meta
Meta
池化
CnosDB Tskv
CnosDB Tskv
CnosDB Tskv
CnosDB Tskv
CnosDB Tskv

云原生：高可用

多云多区

支持AWS

支持GCP

支持跨云灾备

支持跨区灾备

热点自动迁移

自主识别热点

热点数据自动缓存

热点智能迁移

秒级故障恢复

自动化增量备份

恢复自动化

云端备份不变性

备份实时重做传输

专家7×24支持

热点迁移

SQL调优

高可用性

99.999%

云原生：低成本

01

灵活的Dedicate套餐

支持 Spot

支持 Arm

02

支持多租户

计费方式灵活

支持按子租户独立计费

03

秒级伸缩

根据业务量弹性伸缩

04

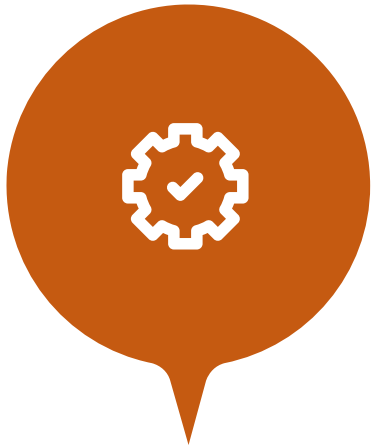
支持EBS + S3混合存储体系

05

支持 Serverless

根据访问次数和存储量计费

云原生：高性能



支持PB级数据

Dedicate套餐可以支持PB级数据量

无感知在线扩容

在线扩容无感知，消除业务抖动

云原生：生态

兼容 AWS

01

Lamdba

CloudWatch

兼容 AP分析平台

02

Snowflake

Aws Redshift

Flink

兼容开源时序数据库

03

InfluxDB

Promethues

兼容 Cloud AI 生态

04

PyTorch

兼容ML平台

Severless vs Dedicate

Severless VS Dedicate

01

需要自动弹性扩展时

- 自动扩大/缩小规模，对需求的变化即时作出反应。
- 对工作负荷剧增或不可预测的企业尤为关键。

02

需要最小化操作时

- 团队可以花更少的时间来担心数据库，而把更多的时间用于构建你的应用程序。

03

测试、实验或评估时

- 适合轻量级应用、原型、测试和开发环境、辅助项目等，因为它们是自助式的、快速的。

04

需要最小化成本时

- 根据实际存储和计算使用量收费。
- 数据库所分配的资源会随着需求自动增加和减少。

01

需要对硬件进行控制时

- Severless 是基于云的，不能控制硬件。
- 出于安全或监管的原因需要可以控制硬件的解决方案。

02

需要一个深入的功能集时

- Serverless 得功能目前还相对较少。
- 有些公司需要一个具有多区域功能的数据库。

03

安全问题排除在多租户之外时

- Severless 归根结底租户仍共享同一台机器。
- 高安全性工作负载， Dedicate有其优势。

04

提供更好的性能或成本更低时

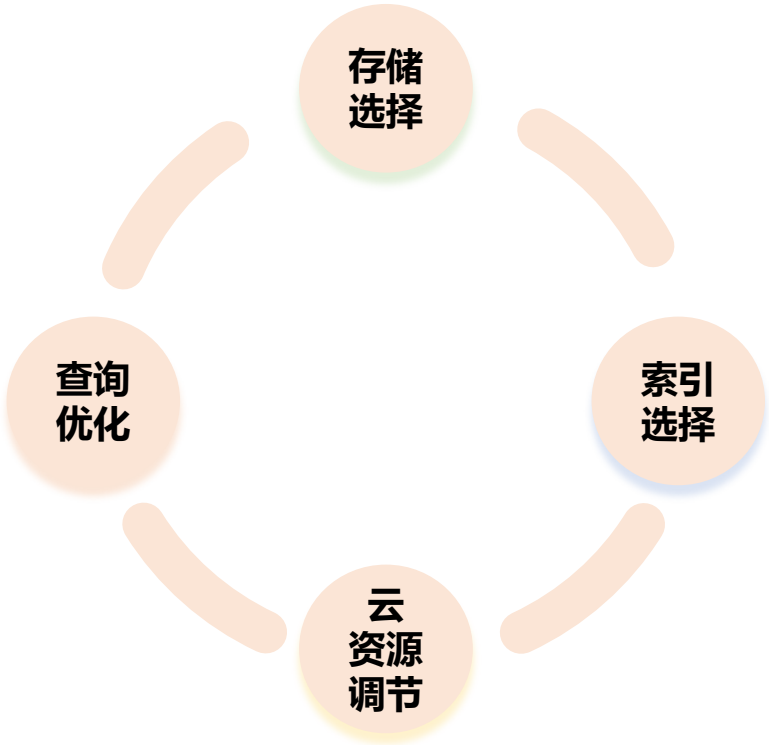
- Severless 是许多用例的最佳选择，但没有"完美"的数据库解决方案能满足所有可能的用例/工作负载。

Part Five

What else

AI for DB

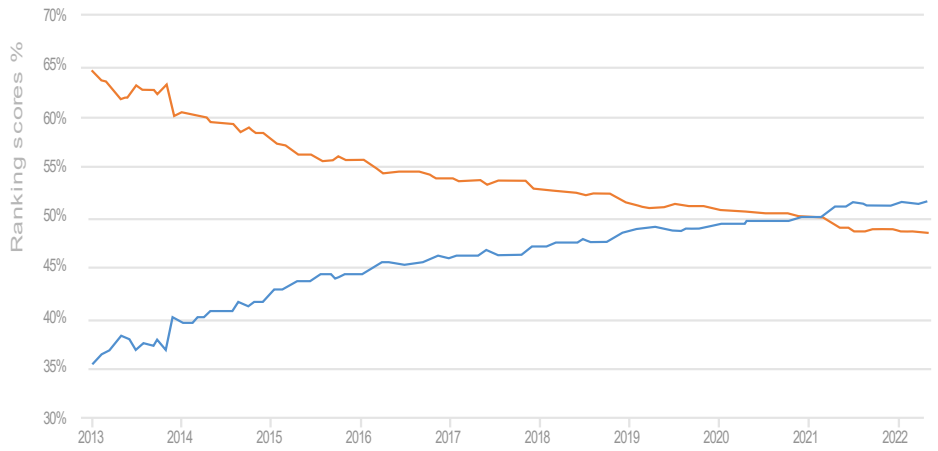
人工智能学习数据库相关得体系化经验，实现智能化运维管理数据库



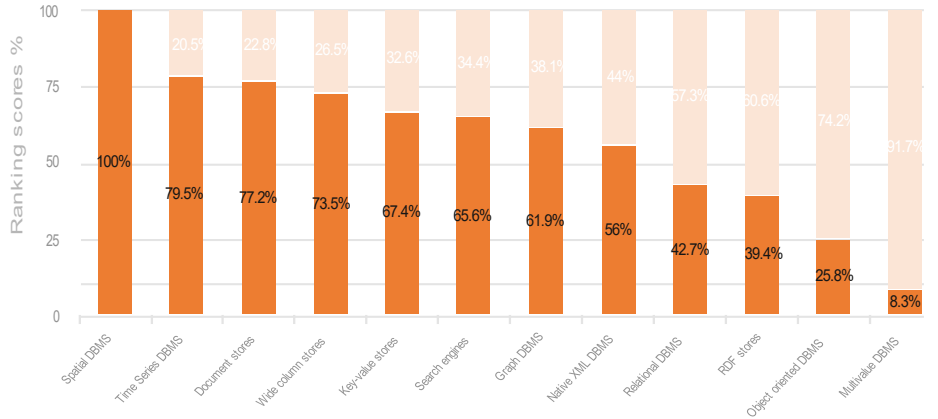
序号	研究成果
1	基于机器学习的知识图谱双存储结构
2	APRIL: 基于强化学习的图自动管理
3	基于强化学习的RDF图数据分表存储方法
4	基于Seq2Seq模型的SparQL查询预测
5	PreKar: 知识图谱存储结构性能智能评估
6	基于代价的轻量级存储自动决策
7	基于代价的数据库内机器学习轻量级存储自动决策
8	基于机器学习的自动化文档管理
9	基于强化学习的NoSQL数据库索引选择技术
10	基于卷积神经网路的通用索引推荐模型
11	智能数据库事务并发控制算法

开源：基础软件最优商业模式

Popularity trend



Popularity brken down by database model, May 2022



信任基础

- 开源软件可以进行白盒测试，无法测试作弊。
- 公司可能倒闭，但开源社区代码永存。

人才杠杆

- 大厂不再996释放了程序员的时间。
- 开源社区贡献是已成为IT人员的重要评判标准。
- 互联网寒潮下，更多程序员思考自己的未来。

市场杠杆

- 白盒且免费的产品易于快速传播，形成网络和规模效应。
- 社区反馈获得产品未来的研发和市场方向。
- **开源数据库的商业化被市场成功验证过。**

我们的开源社区

使用手册

https://docs.cnosdb.com

Content 学术研究、行业前沿、工程实践、使用手册、小白入门

Channel B站、知乎、CSDN、公众号、CCF、抖音、Twitter, Linkedin, Youtube

Data 全面监控全域数据，A/B test，提高社区活跃度，扩大社区影响力

产品预报

2.0 分布式

2022 Q4

Cloud Version

2023 Q1

Enterprise Service

2023 Q2


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