Day7 云上玩转KV存储-CloudTable服务介绍

CloudTable：基于Apache HBase的NoSQL数据库服务

基于Apache HBase提供的分布式、可伸缩、全托管的NoSQL数据存储服务，天然集成OpenTSDB和GeoMesa，可提供毫秒级随机读写能力，适用于海量(半)结构化数据存储，如消息日志、时序、时空、用户画像类数据。可被广泛应用于物联网、车联网、金融、智慧城市等行业。

高性能：响应**时延毫秒级**，**TPS支持千万级**，横向扩展

生态开放：**兼容原生接口** HBase/OpenTSDB/ GeoMesa

时序数据库：读写性能提升**30%-60%**，支持插值、降精度、聚合强大分析能力，10:1**高压缩**比，成本更低

时空大数据：帮助物联网存储和分析海量位置、轨迹、时空(spatio-temporal)数据，超越传统GIS

应用场景

适用于海量KeyValue数据高吞吐量写入和实时查询场景。

- **车联网**：车辆产生的属性数据和地理位置变化数据
- **IoT**：燃气、水务、电力、化工设备、计算机设备、智能家居，等IoT设备监控分析
- **在线教育/零售/旅游**：销售的区域分析，为优化区域销售提供决策支撑
- **气象**：支持气象数据的三维空间+时间的数据存储、查询和分析

某电力平台毫秒级响应效果：从一年207亿的数据中点查一条用时6ms，查询一天96条数据用时12.8ms， 一个星期数据672条20.4ms，一个月数据2880条60ms，从一年45亿条数据中实时查出最新的一条数据用时1.1ms。

CloudTable功能描述

兼容原生HBase，内核及架构深度优化

- 兼容HBase原生接口，支持KeyValue数据模型。
- 架构高可用。Master为两个节点，主备模式，HA实时检测；计算单元的故障，region可以秒级转移，保证业务的高可用。
- 存储和计算分离。存储采用多备份机制，同时计算和存储分离保证数据的高可靠。

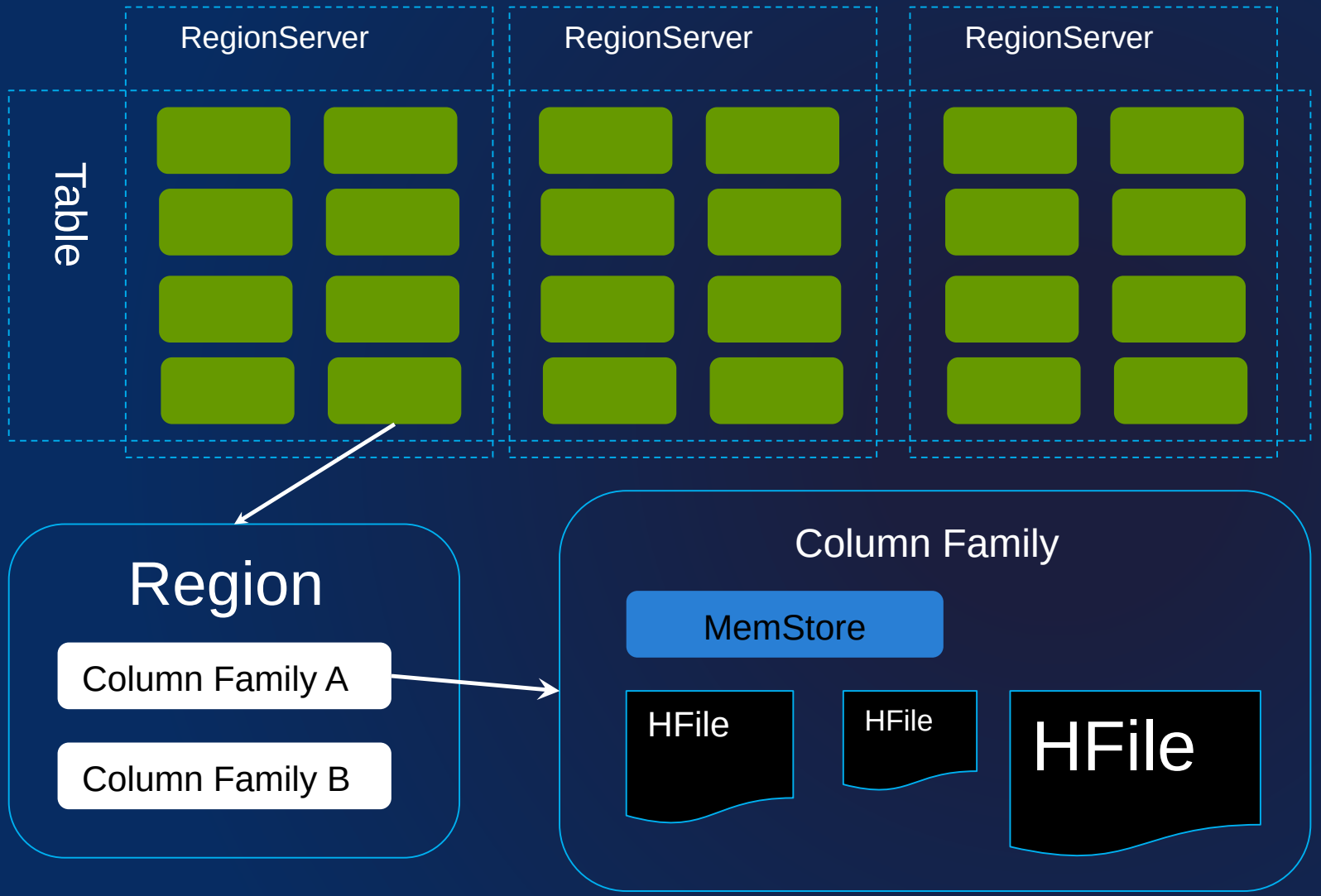
支持时序数据存储，集成OpenTSDB

- 集成OpenTSDB，提供时序数据高效读写、时序数据查询、计算能力
- 高压缩技术，单数据点平均2字节，压缩率高达90%

支持时空数据存储，集成GeoMesa

- 集成地理大数据处理套件GeoMesa，可以帮助物联网存储和分析海量时空(spatio-temporal)数据，提供轨迹查询、区域分布统计、区域查询、密度分析、聚合、OD分析等功能

HBASE 基础概念



- **Table**

可理解成传统数据库中的一个表，但因为Schema-Less的设计，它较之传统数据库的表而言，更加的灵活

- **Region**

Region可理解成将Table按RowKey横向切割的一个子表

- **RegionServer**

HBase的数据服务进程。Region必须被部署到某一个RegionServer中才可以提供读写服务

- **ColumnFamily**

一些列的集合。同一个Region中不同的Column Family的数据被存储在不同的地方

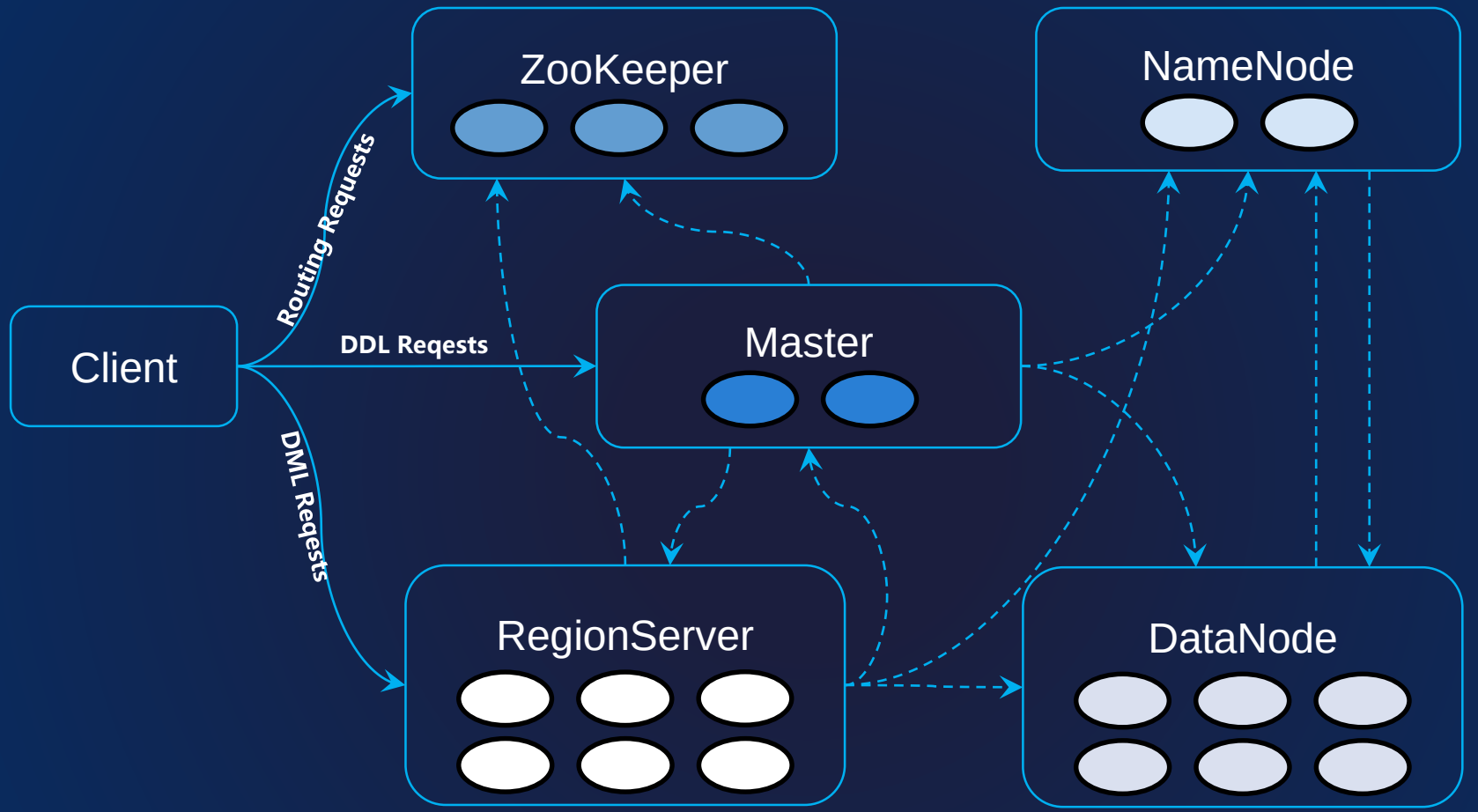
- **MemStore**

用户数据被写入Region中时，首先存在于各个Column Family的内存中，当达到一定大小之后Flush到磁盘。这部分位于内存中的用户数据被称之为MemStore

- **HFile**

HBase表数据以HFile形式存在于文件系统中，它拥有特定的索引组织结构来加速数据读取

HBASE 关键进程角色

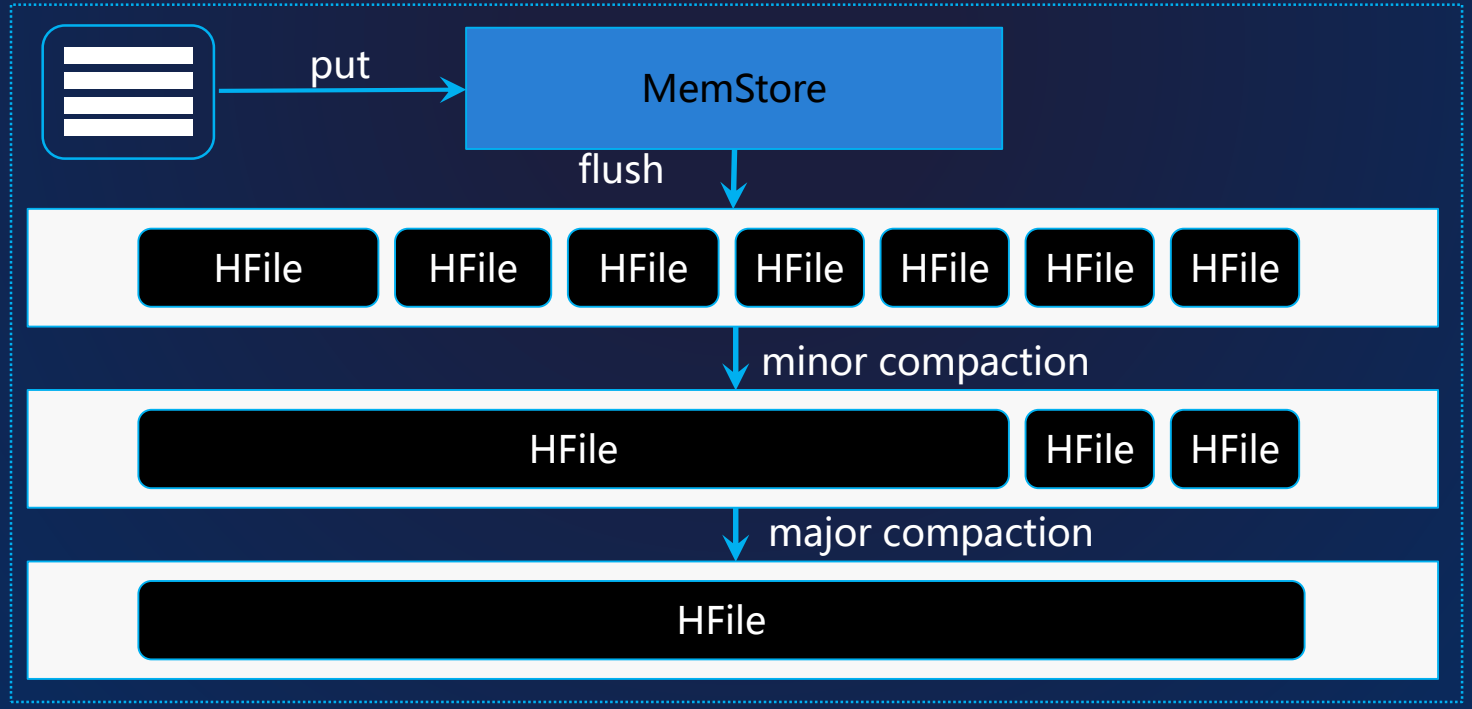


Meta表的路由信息在ZooKeeper中；Master负责表管理操作，Region到各个RegionServer的分配以及RegionServer Failover的处理等；RegionServer提供数据读写服务。HBase的所有数据文件都存放在HDFS中。

HBASE 写入流程

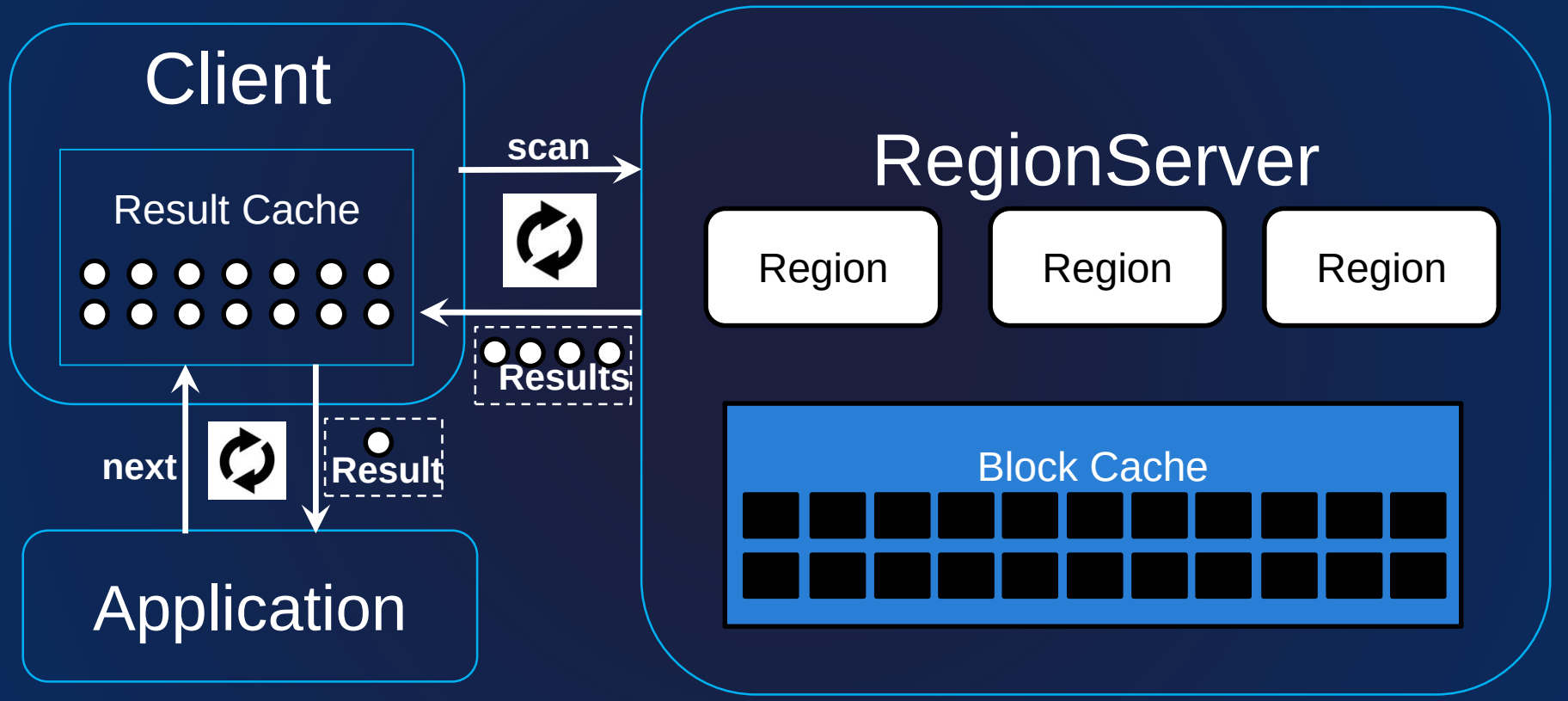
假设In-memory Flush&Compaction未开启

RegionServer侧写入

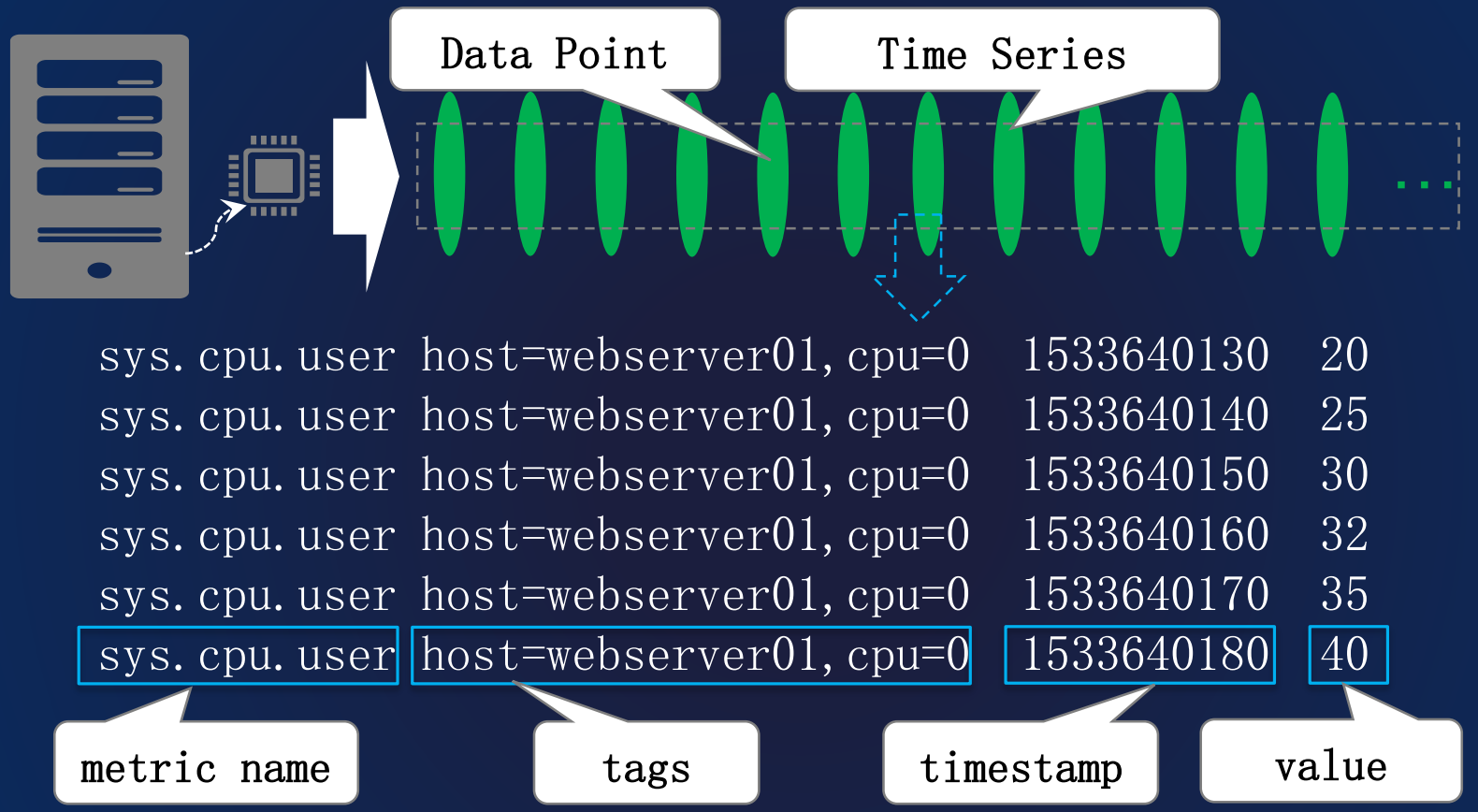


Flush & Compaction

HBASE 读取整体流程



OpenTSDB – 数据模型



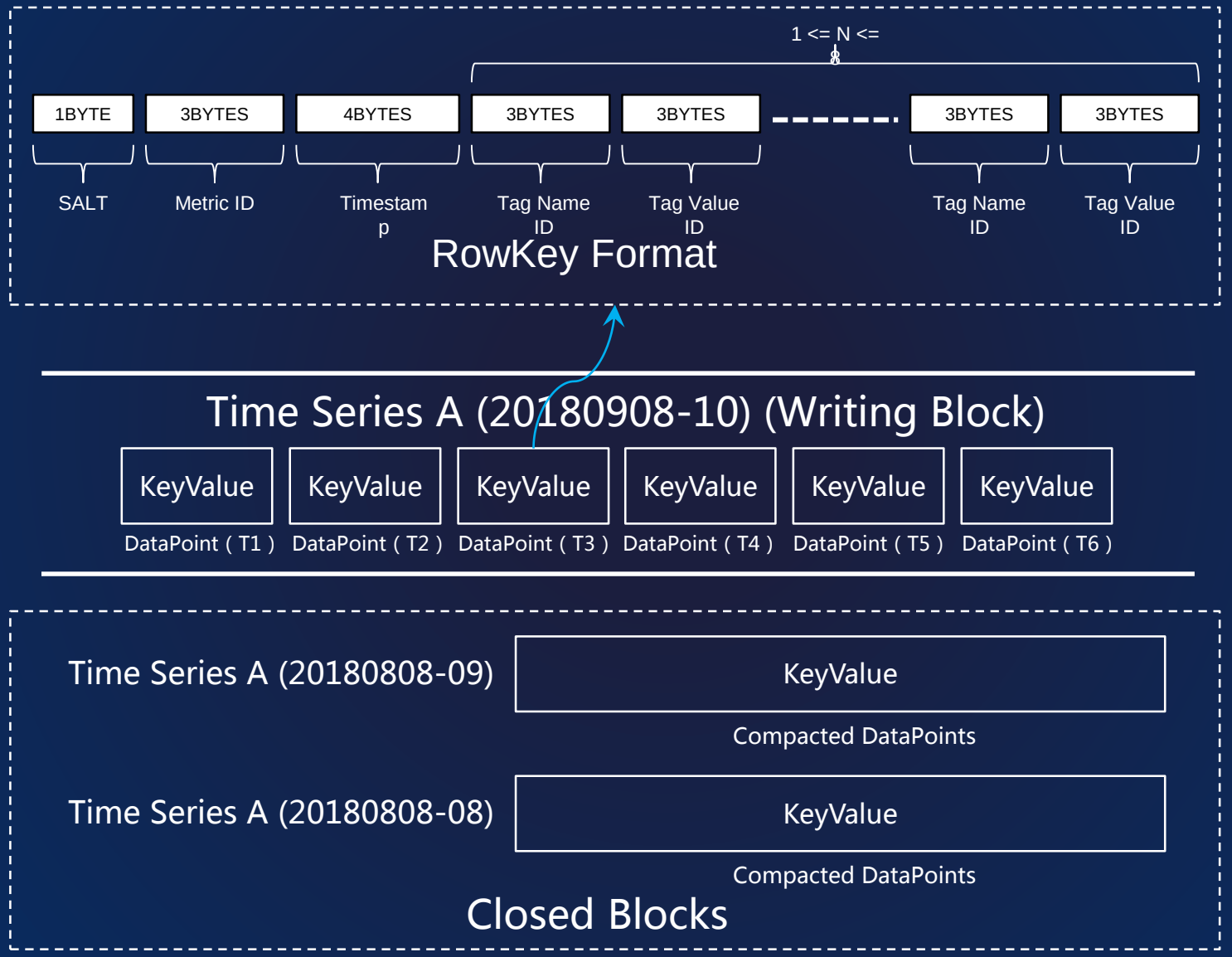
一个Time Series可以理解成是一个数据源的一个指标按时间产生的指标数据序列，每一个指标称之为一个Data Point。OpenTSDB使用一个Metric Name以及一组Tags信息来唯一确定一个Time Series。

OpenTSDB – 典型查询场景

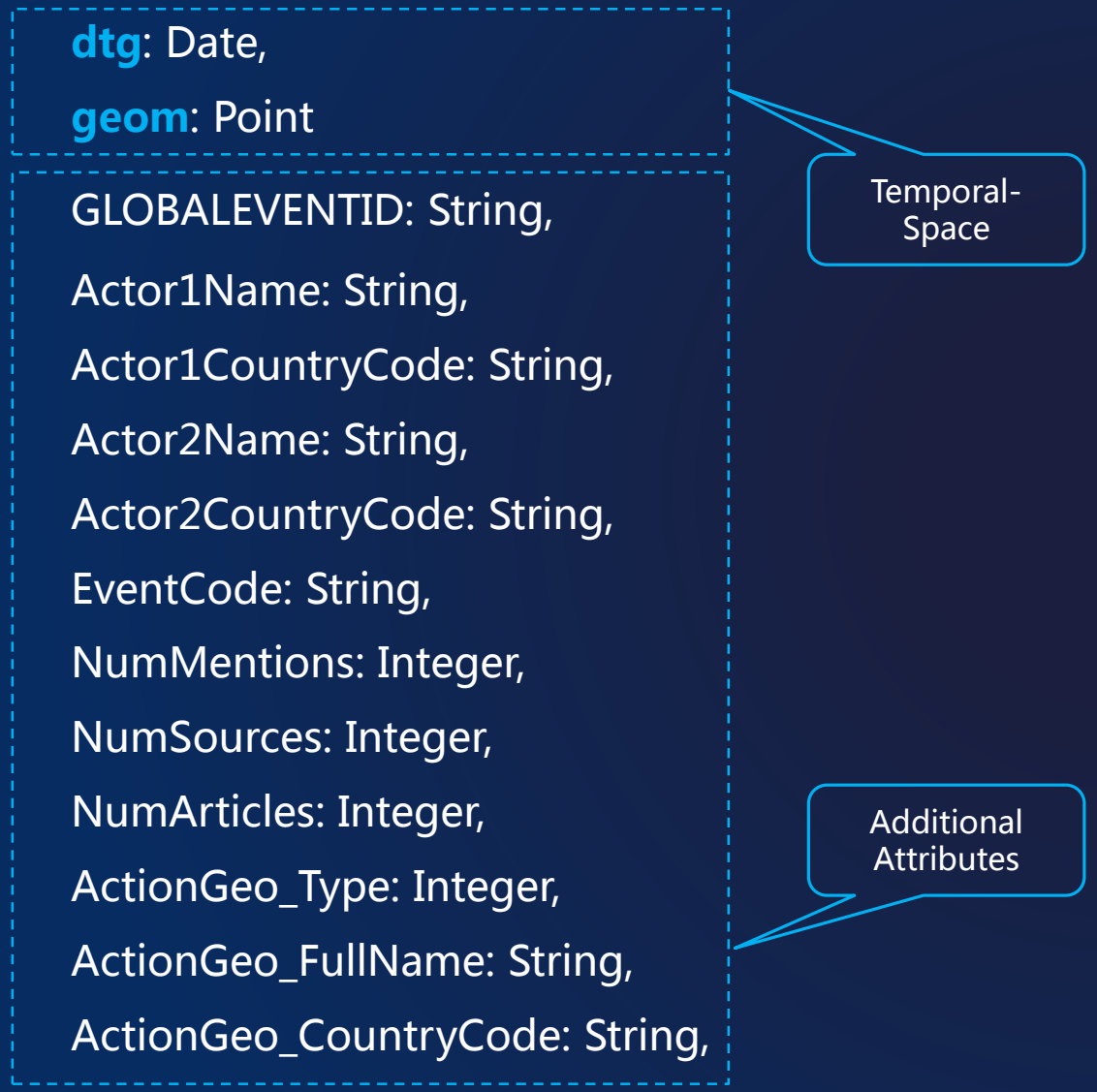
- 给定Metric Name以及一组Tags信息，查询某时间范围的所有的Data Points
- 给定Metric Name以及一组Tags信息，查询某时间范围的聚合结果
- 给定Metric Name，查询所有相关Time Series在某时间范围的统计信息

OpenTSDB – RowKey设计

OpenTSDB Version : 2.3



GeoMesa – 数据模型



GeoMesa is an Apache licensed open source suite of tools that enables large-scale geospatial analytics on cloud and distributed computing systems, letting you manage and analyze the huge spatio-temporal datasets that IoT, social media, tracking, and mobile phone applications seek to take advantage of today.

GeoMesa的一条数据，称之为一个SimpleFeature，SimpleFeature中主要包含如下数据：

1. 时间信息
2. 空间信息
3. 其它属性

GeoMesa – 典型查询场景

- 查询某一个地理区域在某个时间范围发生的关键事件
- 一个任意区域的流量信息
- 给出2015年受血吸虫病影响的区域
- 查找最近10分钟进入飓风区域的汽车？
- 查找某一个点附近所有的酒店
- 查找某嫌疑人在2018年9月的移动轨迹
- ……

时空查询：区域 + 时间区间

空间查询：区域信息

时序查询：轨迹查看

属性查询：主题属性信息

GeoMesa – RowKey设计

GeoMesa Version : 2.11

- 针对Point的时间+空间三维索引(Z3)
  ShardKey(1byte) + Epoch Week(2bytes) + Z3(x, y, t) (8bytes)+ FeatureID

- 针对Point的空间索引(Z2)
  ShardKey(1byte) + Z2(x, y) (2bytes) + FeatureID

- 针对复杂空间对象(如Polygon)的时间+空间三维索引(XZ3)
  ShardKey(1byte) + Epoch Week(2bytes) + XZ2(minX, minY, maxX, maxY)(8bytes)
  + FeatureID

- 针对复杂空间对象(如Polygon)的空间二维索引 (XZ2)
  ShardKey(1byte) + XZ2(minX, minY, maxX, maxY)(8bytes) + FeatureID

- Attribute索引
  IdxBytes(2bytes) + ShardKey(1byte) + AttrValue + SplitByte(1byte)
  + SecondaryIndex(Z3/XZ3/Z2/XZ2) + FeatureID

- ID索引
  FeatureID

“**知识点备注**：Z2/Z3的核心思想是基于Z-Order空间填充曲线将二维/三维信息映射成一维信息，在这个一维信息中能够很好的保持时空距离信息。而XZ2/XZ3的核心思想是XZ-Order空间填充曲线，XZ-Order是基于Z-Order做的扩展，从而能够支持将Polygon/Rectangle等复杂空间对象映射成一维信息。

CloudTable功能：OpenTSDB时序数据存储、查询和分析

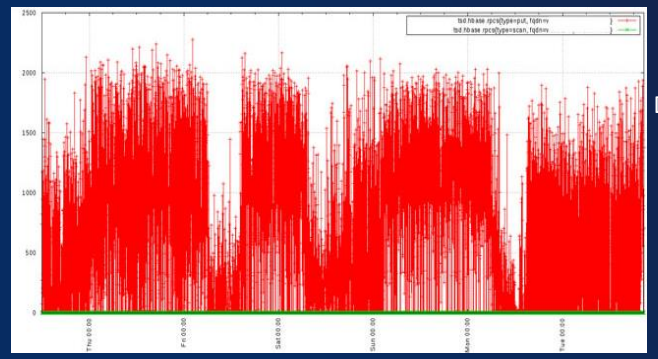
metrics	timestamp	tagk1	tagv1	tagk2	tagv2	value
sys.cpu.system	2017-01-01 01:00:00	dc	dal	host	web01	3
sys.cpu.system	2017-01-01 01:00:00	dc	dal	host	web02	2
sys.cpu.system	2017-01-01 01:00:00	dc	dal	host	web03	10
sys.cpu.system	2017-01-01 01:00:00	host	web01			1
sys.cpu.system	2017-01-01 01:00:00	host	web01	owner	joe	4
sys.cpu.system	2017-01-01 01:00:00	dc	lax	host	web01	8
sys.cpu.system	2017-01-01 01:00:00	dc	lax	host	web02	4

CloudTable (OpenTSDB)

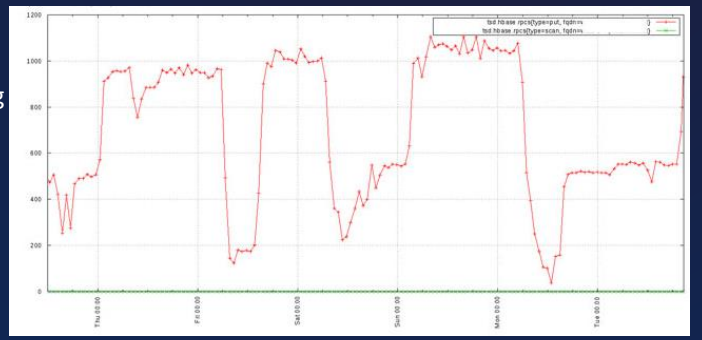
rowkey	+0	+1	+2	+3
{0,0,1},1465920000[{0,0,1}={0,0,1},{0,0,2}={0,0,3}]	50	50	50	50
{0,0,1},1465920000[{0,0,1}={0,0,1},{0,0,2}={0,0,4}]	50	50	50	50
{0,0,1},1465920000[{0,0,1}={0,0,2},{0,0,2}={0,0,4}]	50	50	50	50

Compaction

rowkey	v
{0,0,1},1465920000[{0,0,1}={0,0,1},{0,0,2}={0,0,3}]	(+0,50),(+1,50),(+2,50),(+3,50)
{0,0,1},1465920000[{0,0,1}={0,0,1},{0,0,2}={0,0,4}]	(+0,50),(+1,50),(+2,50),(+3,50)
{0,0,1},1465920000[{0,0,1}={0,0,2},{0,0,2}={0,0,4}]	(+0,50),(+1,50),(+2,50),(+3,50)



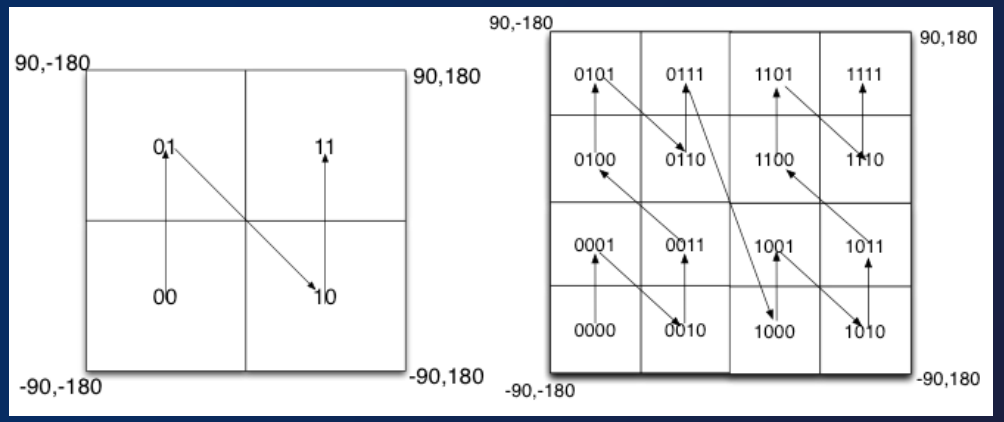
Downsampling



时序数据库OpenTSDB关键能力：

- ✓ 低成本
  - 时间戳采用delta编码进行压缩，数据值采用XOR进行压缩，单个数据点平均2个字节，压缩比约为10:1。
  - 存储计算解耦，为IoT场景海量数据、动态热点的数据特征量身打造，方便按照并发度和存储量按需独立扩容
- ✓ 企业级
  - 百万级数据点1s写入，百万数据点读取，5s返回
  - 分布式架构，横向水平扩展
  - 高压缩率算法，节约成本的同时，提升查询速度
- ✓ 兼容性
  - 兼容OpenTSDB社区最新版本
  - 兼容OpenTSDB原生接口，业务迁移应用“0”改动
- ✓ 时序数据计算
  - 插值，缺失的数据点，支持线性插值数据补全
  - 降精度，支持预降精度和实时降精度计算，满足高效查询需求
  - 空间聚合，支持按照不同的Tag进行空间聚合和分组计算
  - 丰富的聚合函数，提供AVG,SUM,MAX,MIN等聚合函数

CloudTable功能： GeoMesa地理大数据存储、查询和分析



GeoMesa基于Geohash编码以及空间填充曲线Z-Order理论基础，做到了将二维经纬度转换成一维字符串，将三维时空（经纬度和时间）转换成一维字符串，为高性能查询打下了基础。

GeoHash编码步骤：

1) 根据经纬度计算GeoHash二进制编码，分别得到经度和纬度的二进制编码
2) 偶数位放经度，奇数位放纬度，组合生成新串
3) 最后使用0-9、b-z（去掉a, i, l, o）这32个字母进行base32编码

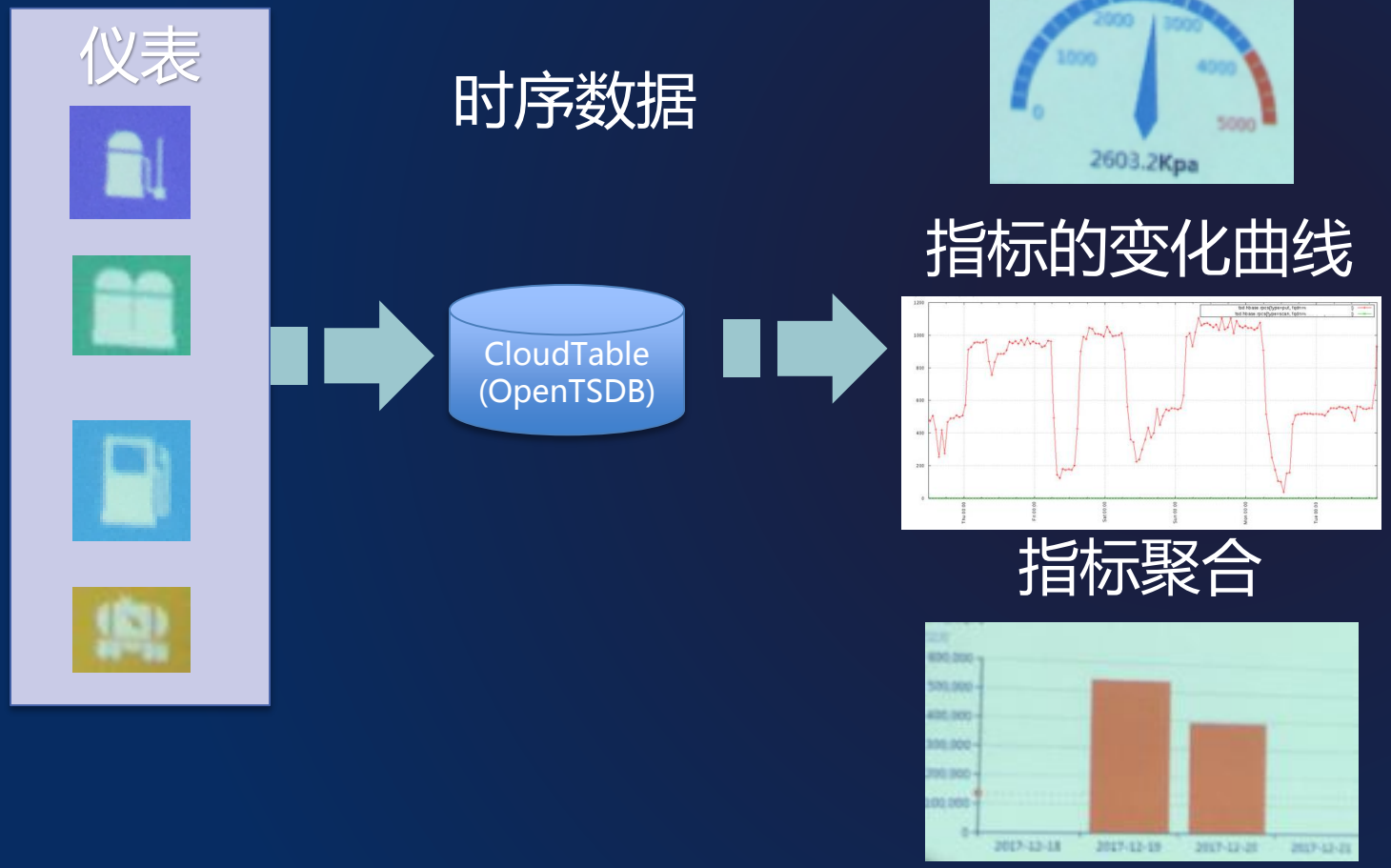
HBase中索引设计

1) Id index: Feature ID的索引
2) Z2 index：空间索引（point）
3) Z3 index：时空索引表（point）
4) XZ2 index：空间索引（非point）
5) XZ3 index：时空索引（非 point）
6) Attribute Index: 属性索引表

时空数据库GeoMesa关键能力：

1) 位置查询：查询某Feature的位置
2) 轨迹查询：查询某Feature的轨迹
3) 时空范围内位置查询：查询时间+空间范围内的位置
4) 时空范围内的轨迹查询：查询时间+空间范围内的轨迹
5) OD 分析：起终点间的交通出行量

CloudTable(OpenTSDB)典型应用案例：**燃气集团燃气管网监控



业务简介：

燃气管网监控：管道天然气是由国家干线下气，通过省干线输给各地市，由各地市燃气公司负责输送，部分大型工商企业、工厂等单独输送。

1）在城市母站和各分支站，以及大型工商企业、工厂管线通过流量计来监控当前管道内天然气的压力、流量、温度、泄露指标，通过DTU（Data Transfer unit）上报，并通过监控大屏展示。

2）通过监控平台识别异常，需要给DTU下发指令，关闭异常流量所在的管道。

3）当前监控点需要调压（增压、减压）需要人工操作；希望通过历史数据，训练模型，自动调控压力。

客户价值：

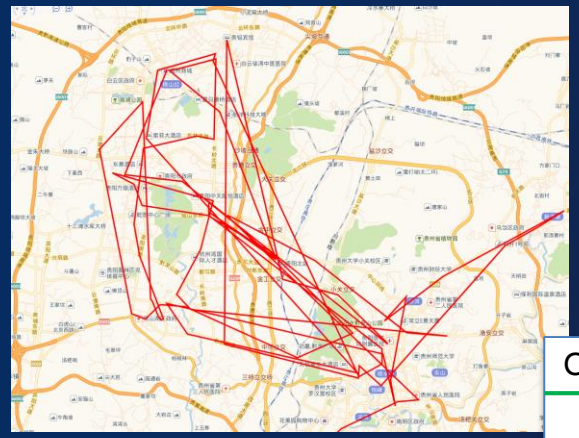
1）30+万条数据需要达到秒级返回

2）对数据上报有严格的时序要求

3）在补数插值时需要保持时序不乱

4）展示时支持在时间维度上降低精度显示

CloudTable(GeoMesa)典型应用案例：深圳交警联创项目



OD 分析：起终点间的交通出行量

- 交通疏导，交通线路规划；
- 出行强度，出行规律分布
- 聚集分析，职住分析

深圳交警

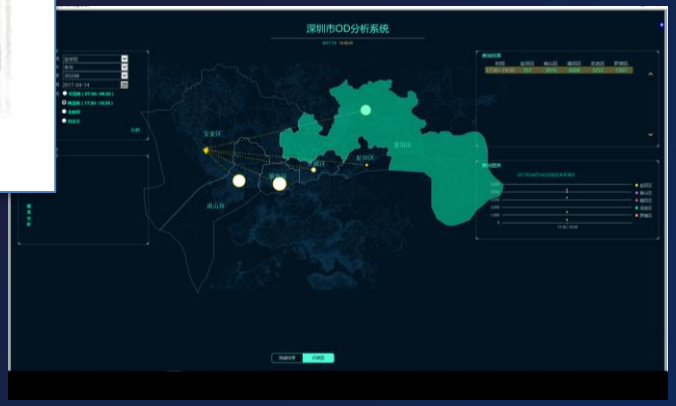
数据量：
近4亿数据

效果：
查询秒级返回

移动对象分析：
实时检索三维轨迹、实时计算移动规律、预测移动状态、

OD分析-利用车数据进行出行分析

区域分析：
实时计算行政区间车辆来源、去向、分析拥堵原因疏导路线



城市通勤分析：
计算城市出行规律、规划站点、枢纽区域

华为云CloudTable(HBase) vs. 自建HBase集群

易运维

针对内核深度优化。版本保持和社区特性、优化和问题的同步，对HBase的参数进行调优，软硬件垂直优化，保证高性能

低成本

按需付费，减少积压成本
计算与存储分离，
按使用计费

易扩容

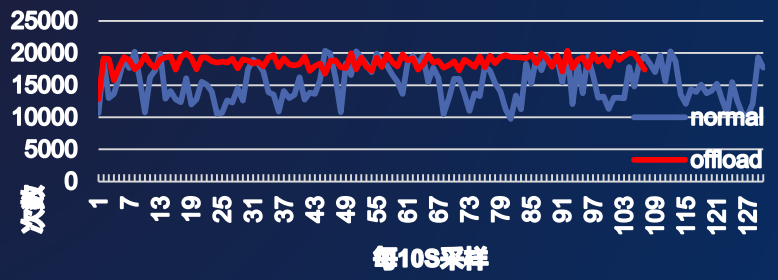
支持按照CU在线简单扩容减容，标准化的CU设计，减少基于业务规格设计机器规格的复杂度，便捷扩容减容

易上手

华为云上全托管服务，简单几步申请集群，不需要熟悉集群安装流程，技术门槛低，专注于你的应用

CloudTable云服务优势

- **应用0改动：**
  - 兼容HBase、OpenTSDB、GeoMesa原生接口，毫秒级NoSQL数据库，提供时序数据库和时空数据库生态，业务迁移应用“0”改动
- **使用门槛低：**
  - 全托管服务，简单几步申请集群，减少集群安装。
  - 针对内核深度优化，版本保持和社区特性、优化和问题的同步，对HBase的参数进行调优，节省平台运维开销。
  - 针对时序数据库OpenTSDB进行深度优化，
- **企业级：**
  - 架构高可用。Master为两个节点，主备模式，HA实时检测；计算单元的故障，region秒级转移，保证业务的高可用。
  - 标准化规格CU设计，简单横向扩展应对不同应用规格
  - 软硬件垂直优化实现平稳的TPS
  - HBase平均时延15ms，P99小于80ms
  - OpenTSDB写并发单CU位80000 Datapoint(阿里5W)



- **低成本：**
  - 海量存储按需付费
  - 计算与存储解耦，适用于海量数据、动态热点的特征数据，计算分离，共享存储，方便按照并发度和存储量按需独立扩容
  - **时序数据**高压缩，压缩比约为10:1。

API 简介

CloudTable提供的外部API主要有：

- **HBase接口**

  兼容Apache HBase原生的Java API接口。请参考 http://hbase.apache.org/1.2/apidocs/index.html

- **OpenTSDB HTTP Restful API**

  请求方式是通过向资源对应的路径发送标准的HTTP请求，请求包含GET、POST方法。

  与开源OpenTSDB接口保持一致，请参考http://opentsdb.net/docs/build/html/api_http/index.html

- **GeoMesa Java API**

  兼容原生的Java API接口，具体请参考GeoTools接口：http://docs.geotools.org/stable/userguide/

常见问题

Q：有API和SDK吗？
A：CloudTable兼容现有的HBase1.3.1版本客户端，同时提供SDK。OpenTSDB不提供SDK，提供Restful接口。GeoMesa提供CQL接口，也可以通过SparkSQL进行访问。针对集群操作API目前还没有开放，所有集群的操作目前是通过Web页面进行

Q：创建CloudTable集群要准备什么？
A：需要提前规划好需要使用的多个服务的网络，包括VPC、安全组等，保证CloudTable的客户端的ECS或者服务与CloudTable服务处于一个VPC中，安全组保证能互通。

Q：CloudTable的高可用有哪些措施？
A：计算层HMaster进程HA部署，Zookeeper进程3节点部署，region秒级转移；存储层使用华为高端存储，坏盘不影响数据读写。

Q：客户自建的HBase/OpenTSDB集群 可以直接迁移到我们的上边吗？
A：只要HBase1.3.1以下的版本，应用代码不需要任何改动。如果客户希望自建HBase上的数据迁移到CloudTable上，可以通过CDM服务进行数据批量迁移，也可以通过写数据接口写入CloudTable。

Q：时空数据库能支持哪些场景？
A：主要适用于包含位置（X\Y坐标）和时间维度的数据存储和查询，可以帮助物联网存储和分析海量时空(spatio-temporal)数据，提供轨迹查询、区域分布统计、区域查询、密度分析、聚合、OD分析等功能。

Q：时序数据库能支持哪些场景？
A：主要适用于物联网存储和分析时间序列数据，提供指标数据的聚合、插值、降精度等查询和分析功能。

JOIN US IN

BUILDING A BETTER CONNECTED WORLD

THANK YOU

HUAWEI