网易大数据
Netease Big Data

开放 · 生态 · 赋能

网易分布式时序数据库揭秘

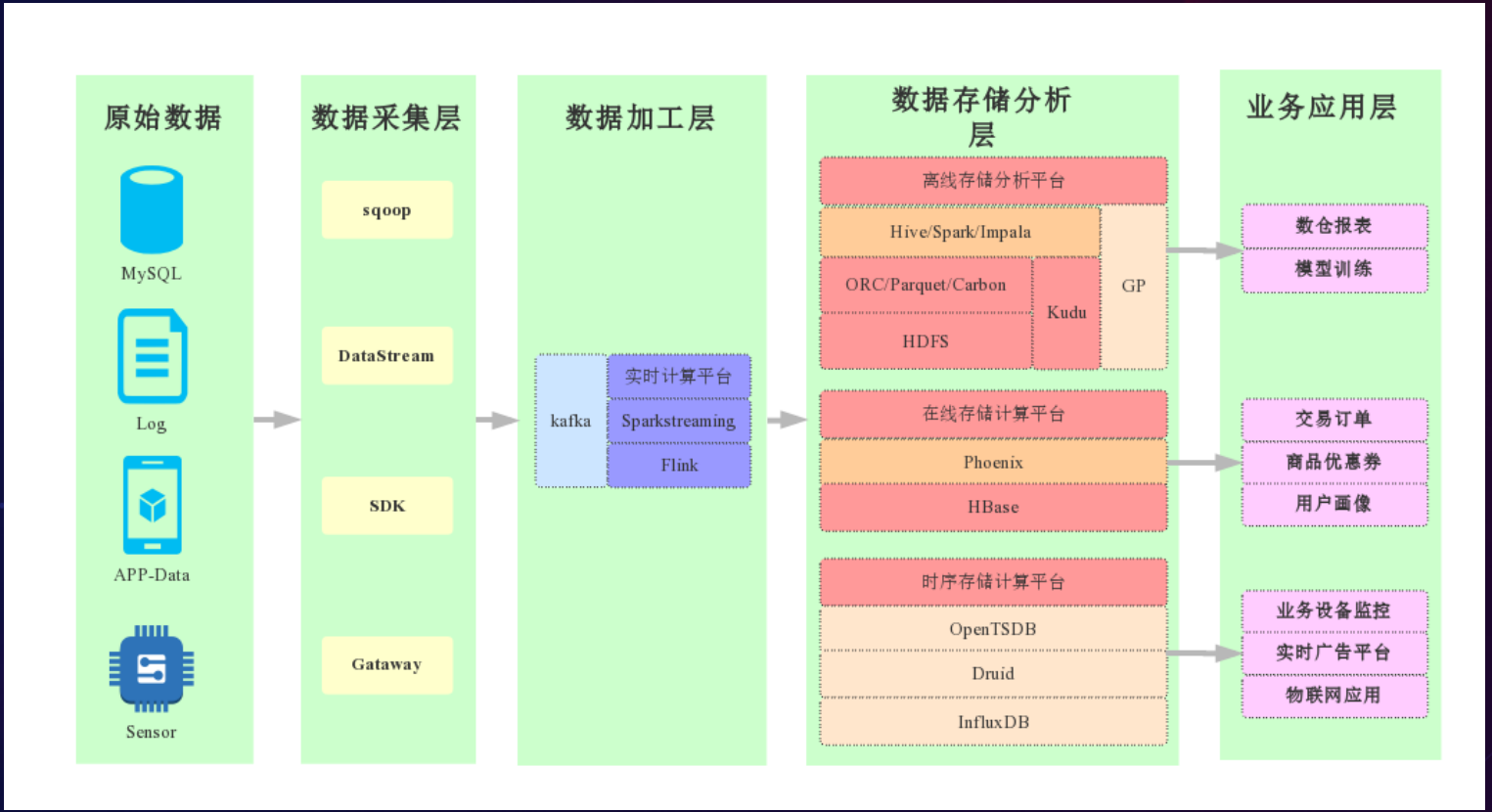
范欣欣

网易大数据
Netease Big Data

PART 01 时序数据库应用场景

时序数据平台主要业务场景

网易大数据
Netease Big Data



时序数据平台主要业务场景

网易大数据
Netease Big Data

电商
- 业务大盘：查看单量，金额，发货等业务指标
- 异常大盘：查看超卖，库存校准耗时，商品回调耗时，各种类型下单错误等异常指标

广告
- 广告曝光点击消耗实时统计
- 流量地域分配

- 调用链全息排查
- 全局调用拓扑
- 链路依赖项分析梳理

- 应用调用次数，错误占比，页面加载延迟统计、地域统计分析
- 慢加载追踪，慢SQL
- 异常会话追踪

- 查看指定hadoop任务耗用内存、CPU、IO利用率等
- 查看集群消耗资源TopN任务、节点等
- 统计集群任务执行耗时

- 物理机、云主机、容器：CPU、内存、IO等
- 组件服务：数据库集群、Kafka集群、HBase集群

时间序列平台场景核心特征

网易大数据
Netease Big Data

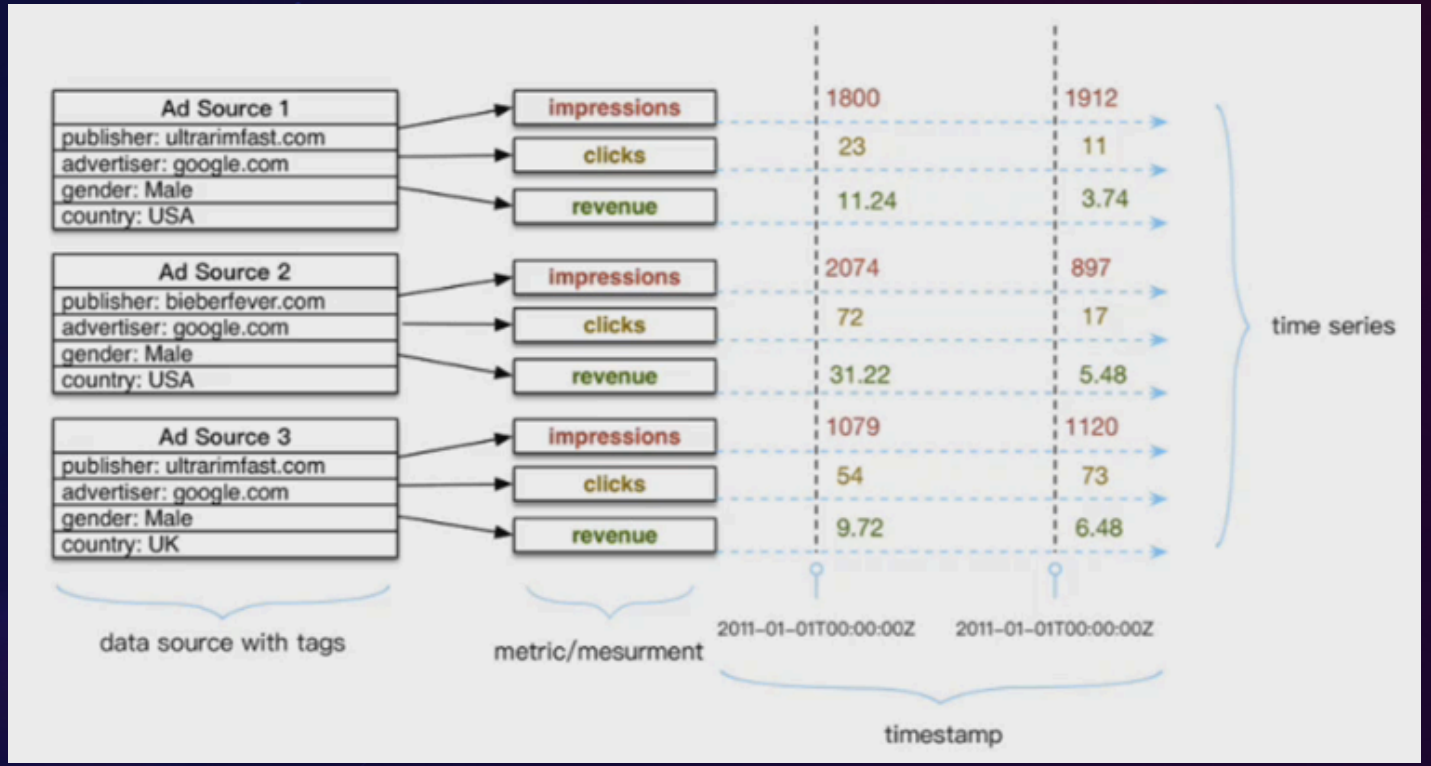
- 时间区间查询。最近时间区间查询频率远大于历史数据查询
- 多维条件查询。多维字段随机组合查询
- 支持TTL机制。数据可以自动过期
- 支持高压缩率。数据压缩比要达到10以上
- 支持高效聚合
- 支持集群可扩展，服务高可用，数据高可靠

网易大数据
Netease Big Data

PART 01 时间序列数据模型

时间序列数据模型

网易大数据
Netease Big Data



时间序列数据模型

网易大数据
Netease Big Data

Time	Publisher	Advertiser	Gender	Country	Impression	Clicks	revenue
2011-01-01T00:00:00	Ultra	Google	Male	USA	1800	23	11.24
2011-01-01T00:00:00	Bieber	Google	Male	USA	2074	72	31.22
2011-01-01T00:00:00	Ultra	Google	Male	UK	1079	54	9.72
...	...	...	...	...	...	...	...
2011-01-01T00:00:01	Ultra	Google	Male	USA	1912	11	3.74
2011-01-01T00:00:01	Bieber	Google	Male	USA	897	17	5.48
2011-01-01T00:00:01	Ultra	Google	Male	UK	1120	73	6.48

行式存储

列式存储

时间序列数据模型

网易大数据
Netease Big Data

Time	Publisher	Advertiser	Gender	Country	Impression	Clicks	revenue
2011-01-01T00:00:00	Ultra	Google	Male	USA	1800	23	11.24
2011-01-01T00:00:01	Ultra	Google	Male	USA	1912	11	3.74
2011-01-01T00:00:00	Bieber	Google	Male	USA	2074	72	31.22
2011-01-01T00:00:01	Bieber	Google	Male	USA	897	17	5.48
2011-01-01T00:00:00	Ultra	Google	Male	UK	1079	54	9.72
2011-01-01T00:00:01	Ultra	Google	Male	UK	1120	73	6.48
...	...	...	...	...	..	...	...

时间列　　维度列-倒排索引　　数值列

时间序列数据模型

网易大数据 Netease Big Data

Time	Publisher	Advertiser	Gender	Country	Impression	Clicks	revenue
2011-01-01T00:00:00	Ultra	Google	Male	USA	1800	23	11.24
2011-01-01T00:00:01	Ultra	Google	Male	USA	1912	11	3.74
2011-01-01T00:00:00	Bieber	Google	Male	USA	2074	72	31.22
2011-01-01T00:00:01	Bieber	Google	Male	USA	897	17	5.48
2011-01-01T00:00:00	Ultra	Google	Male	UK	1079	54	9.72
2011-01-01T00:00:01	Ultra	Google	Male	UK	1120	73	6.48
...	...	...	...	...	..	...	...

数据源 - SeriesKey

时间列

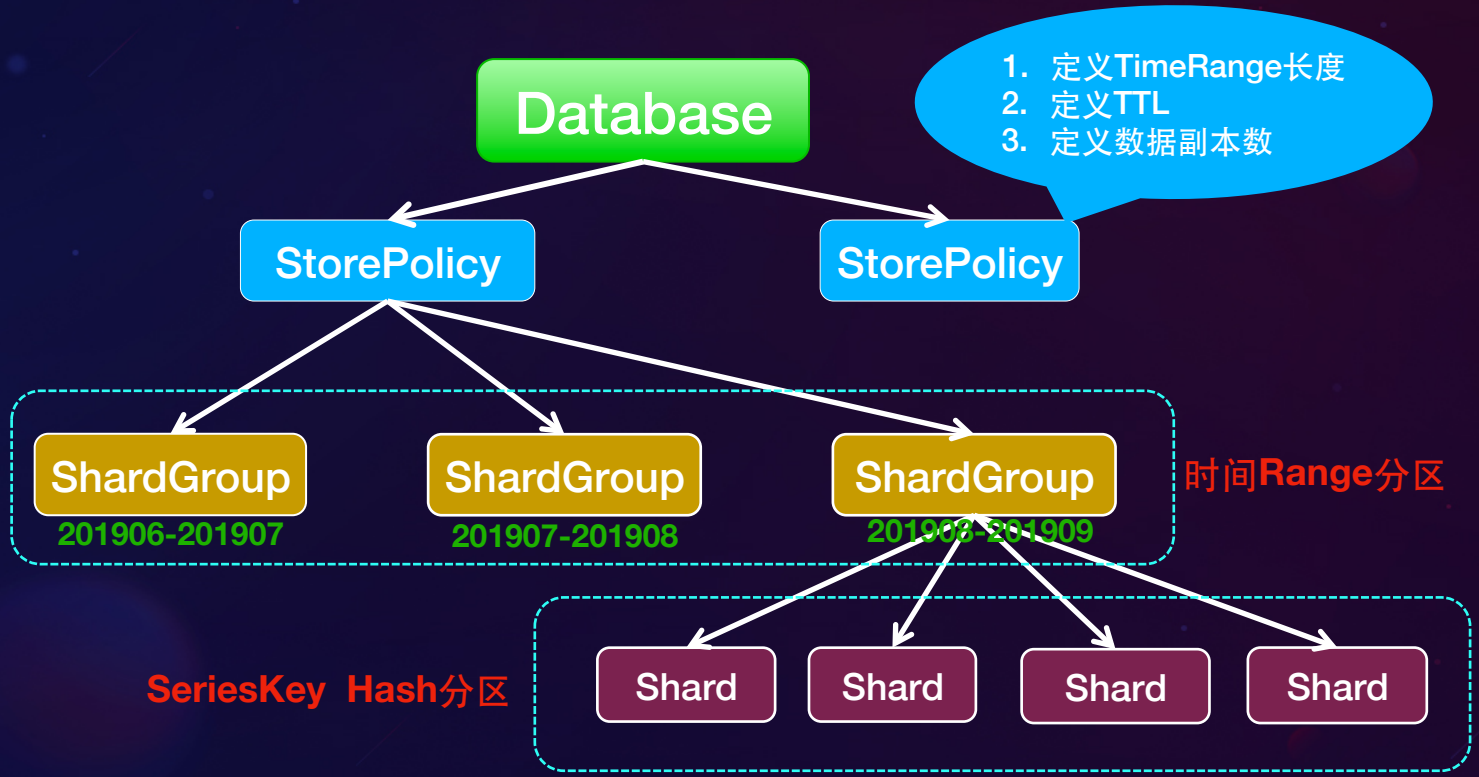
数值列

网易大数据
Netease Big Data

PART 03 NTSDB分布式架构体系

NTSDB - 分片策略

网易大数据
Netease Big Data



NTSDB分布式系统架构

网易大数据
Netease Big Data

NTSDB - MPP计算架构

网易大数据
Netease Big Data

NTSDB - Master

网易大数据
Netease Big Data

NTSDB - 多副本特性

网易大数据
Netease Big Data

- 副本策略：用户可以自定义副本数，副本数不大于节点数。
- 副本放置策略
  - 随机放置策略
  - 分组放置策略：可以将某些表的副本集中放置于某些节点。业务之间隔离。

NTSDB - 多副本读写一致性

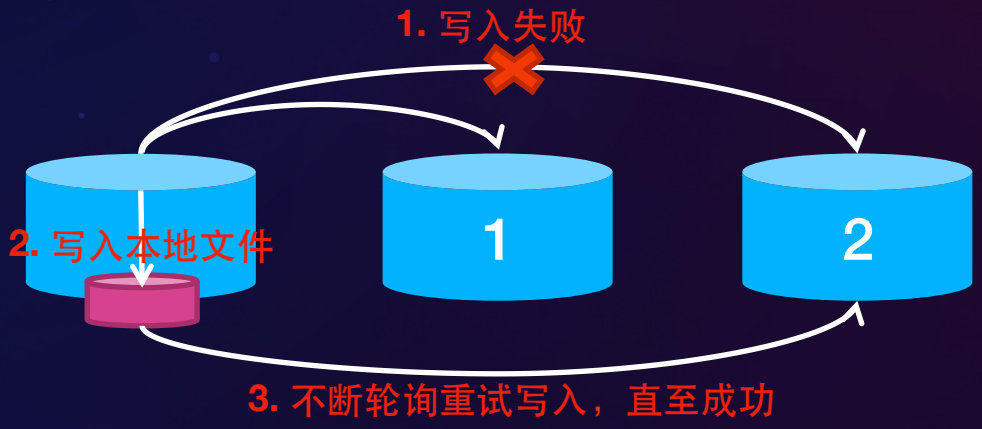
网易大数据
Netease Big Data

- 多副本读写一致性保证：最终一致性
- **Level : one/quorum/all**

NTSDB - 多副本读写一致性

网易大数据
Netease Big Data

- Hinted handoff



- Anti-entropy Node Repair

NTSDB – 故障恢复

网易大数据 Netease Big Data

- 故障检测：ShardServer -> Master心跳检测

- 故障处理
  - unresponsibility : truncate shard
  - failed : shard replica rebuild

NTSDB – 集群扩容

网易大数据
Netease Big Data

- cold shard 数据拷贝迁移
- hot shard下一个数据分片自动分配到扩容机器

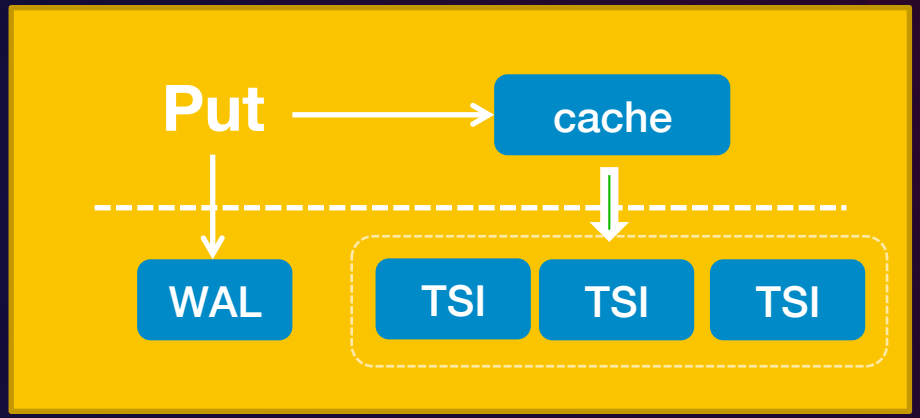
网易大数据
Netease Big Data

PART 04 NTSDB内核核心实现

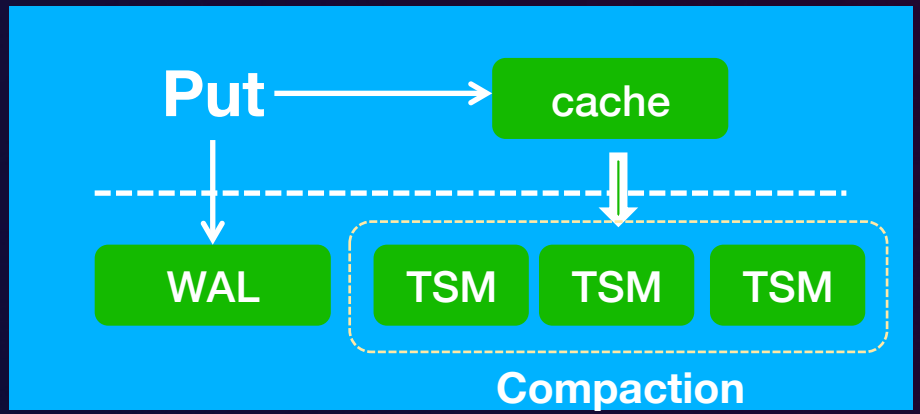
NTSDB - 存储引擎

网易大数据
Netease Big Data

倒排索引引擎



时序存储引擎



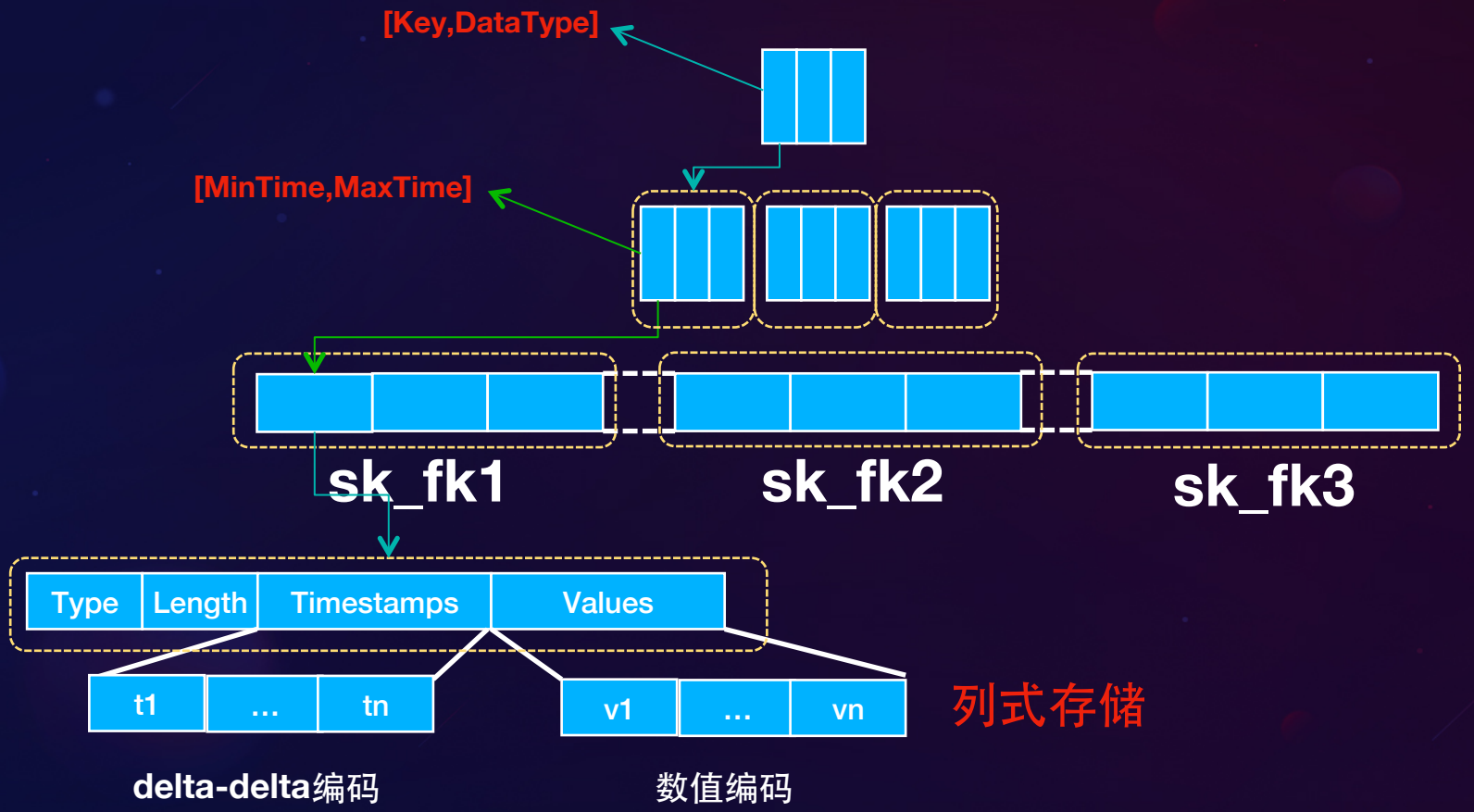
NTSDB - 时序存储引擎

网易大数据
Netease Big Data

select clicks from table where sk in (‘ad1’,’ad2’) and time in [t1,tn]

NTSDB - 时序存储引擎

网易大数据
Netease Big Data



时间序列数据模型

网易大数据 Netease Big Data

Time	Publisher	Advertiser	Gender	Country	Impression	Clicks	revenue
2011-01-01T00:00:00	Ultra	Google	Male	USA	1800	23	11.24
2011-01-01T00:00:01	Ultra	Google	Male	USA	1912	11	3.74
2011-01-01T00:00:00	Bieber	Google	Male	USA	2074	72	31.22
2011-01-01T00:00:01	Bieber	Google	Male	USA	897	17	5.48
2011-01-01T00:00:00	Ultra	Google	Male	UK	1079	54	9.72
2011-01-01T00:00:01	Ultra	Google	Male	UK	1120	73	6.48
...	...	...	...	...	..	...	...

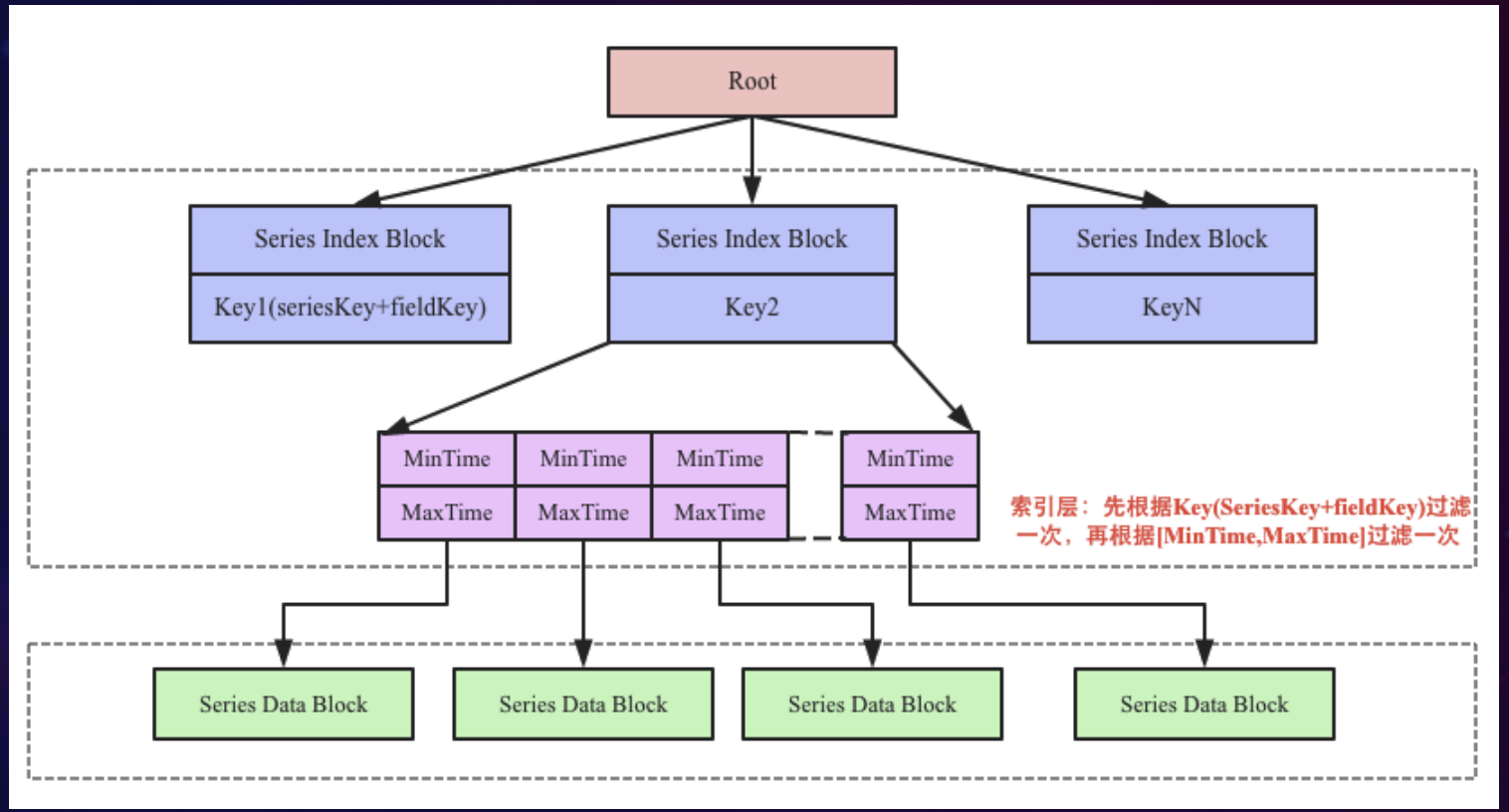
数据源 - SeriesKey

时间列

数值列

NTSDB - 时序存储引擎

网易大数据
Netease Big Data

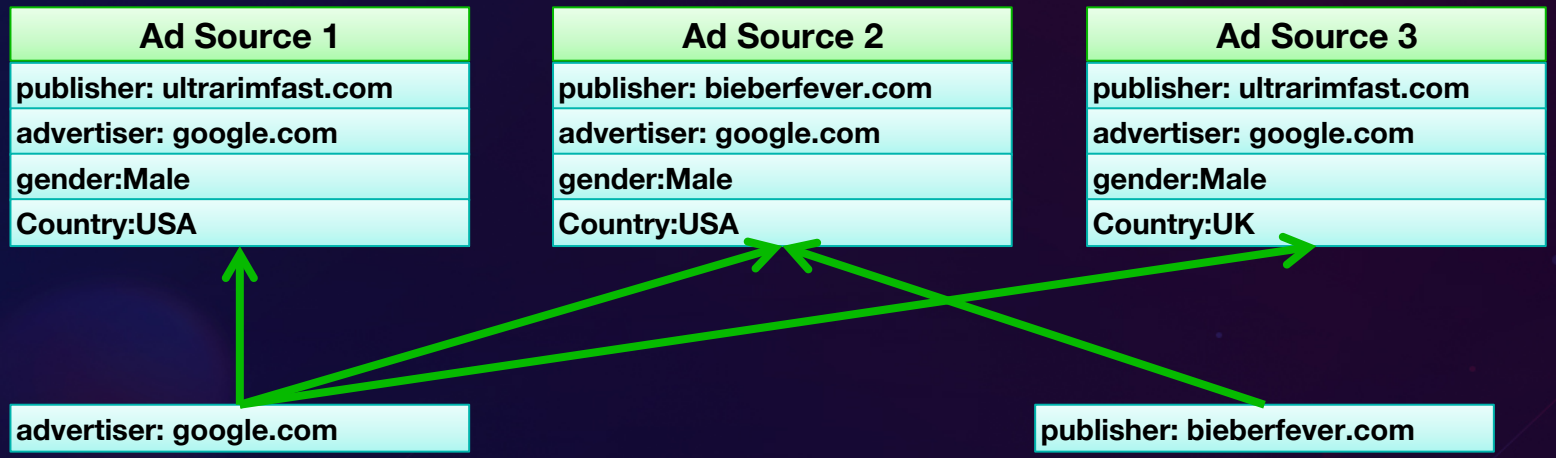


NTSDB – 倒排索引引擎

网易大数据
Netease Big Data

Shard（倒排索引）

select sum(clicks) from ad where time > now() – 1h and advertiser = ‘google’ and publisher = ‘bieberfeer.com’



NTSDB – 倒排索引引擎

网易大数据
Netease Big Data

基于Map序列化实现

基于Bitmap索引实现

NTSDB – 倒排索引引擎

网易大数据
Netease Big Data

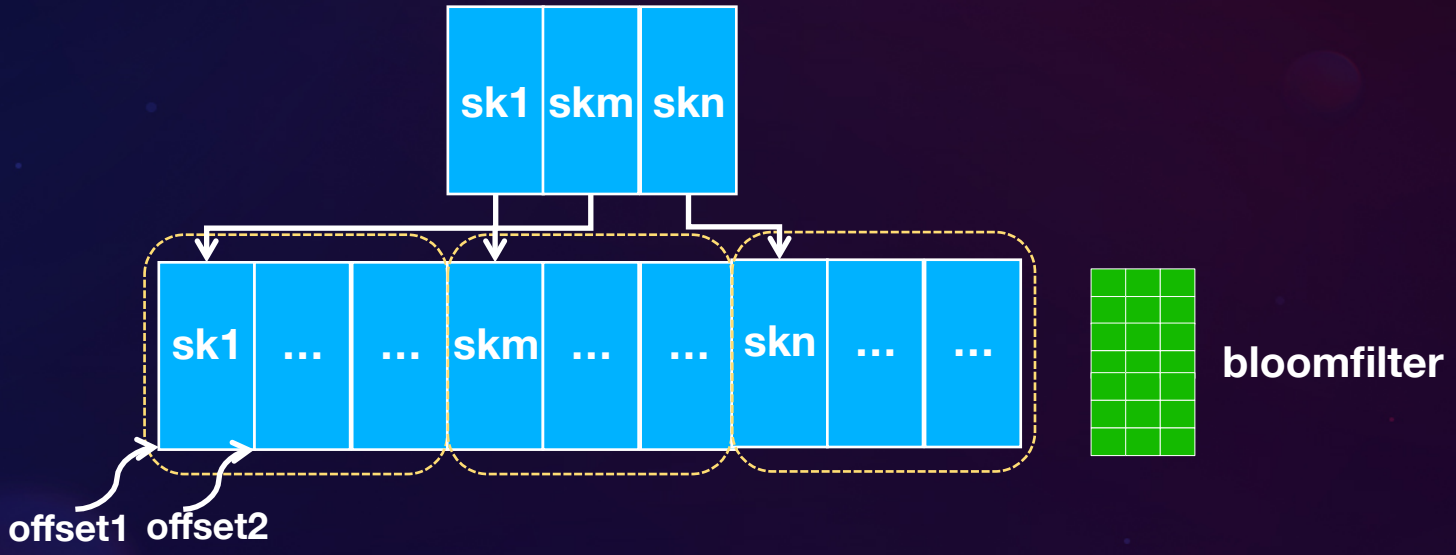
TreeMap<tagkey, TreeMap<tagvalue, LinkedList<SeriesKey>>

- tagkey: 维度列列名，比如advertiser
- tagvalue: 指定维度列可选维度值，比如google.com

NTSDB – 倒排索引引擎

网易大数据 Netease Big Data

TreeMap<tagkey, TreeMap<tagvalue, LinkedList<SeriesKey>>>



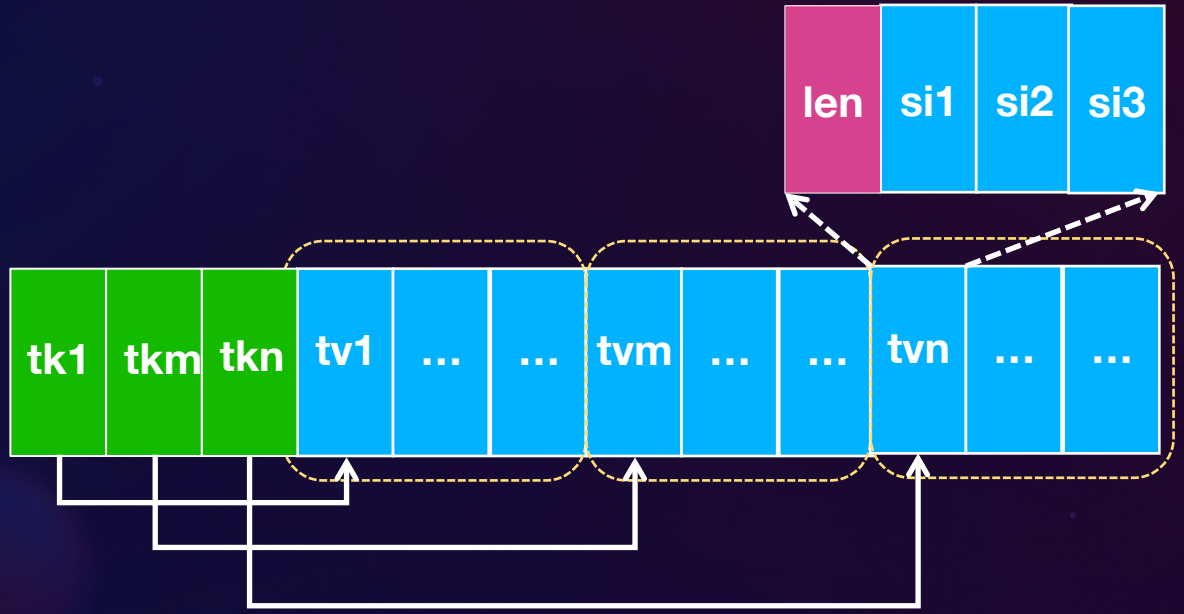
TreeMap<tagkey, TreeMap<tagvalue, LinkedList<offset>>

TreeMap<tagkey, TreeMap<tagvalue, LinkedList<SeriesId>>

NTSDB – 倒排索引引擎

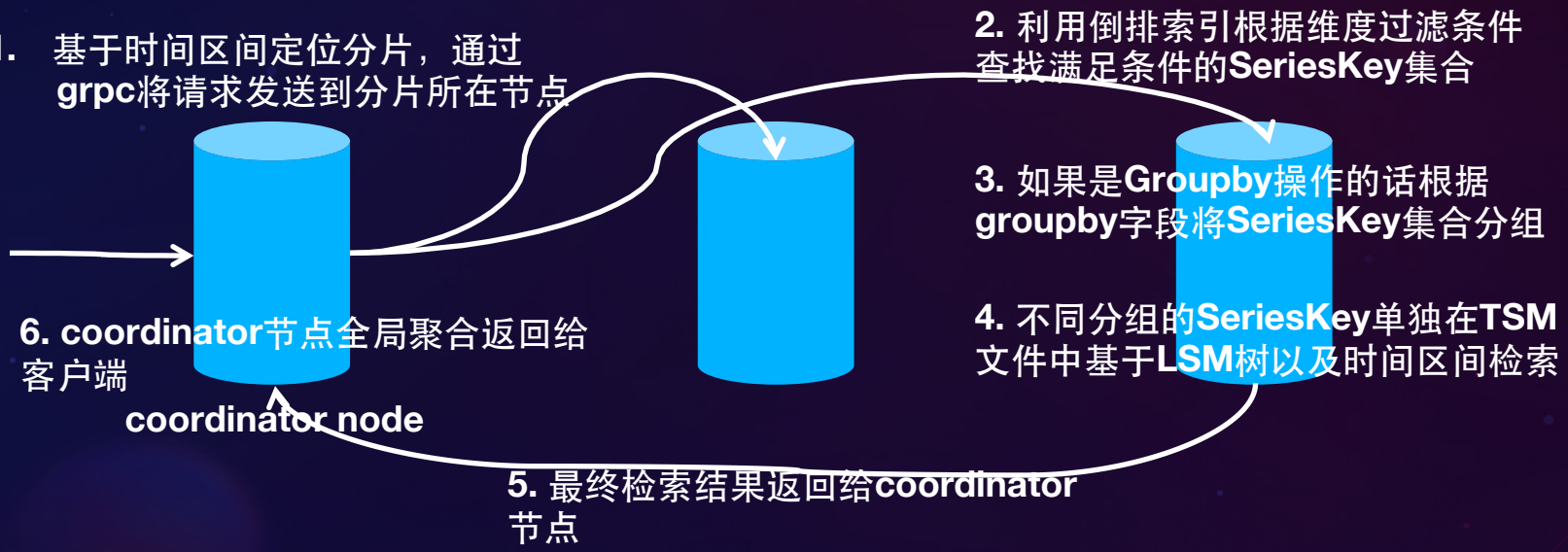
网易大数据
Netease Big Data

TreeMap<tagkey, TreeMap<tagvalue, LinkedList<SeriesId>>



NTSDB – 存储引擎

网易大数据
Netease Big Data



NTSDB - 优化启动效率

网易大数据
Netease Big Data

多级存储优化

- hot shard全量索引加载到内存，cold shard只加载root索引
- hot shard数据放在ssd，cold shard数据放在hdd

NTSDB – 其他内核工作

网易大数据
Netease Big Data

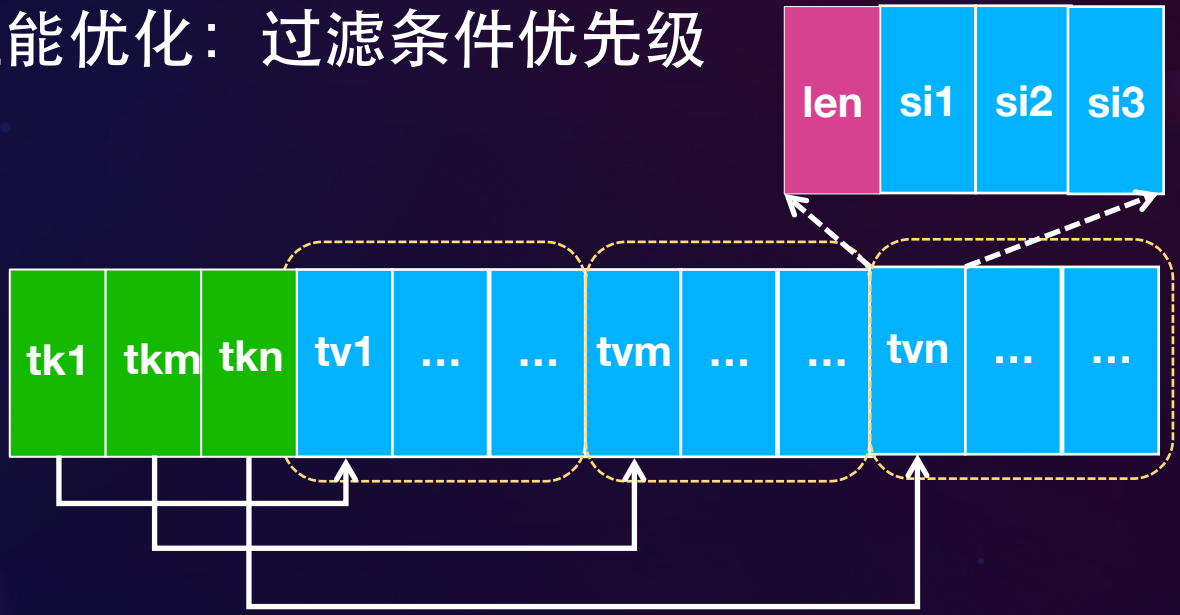
■ 查询性能优化：过滤条件优先级

select mean(cpu) from metrics where host = ‘ysd001’ and cluster = ‘cluster11’

NTSDB – 其他内核工作

网易大数据
Netease Big Data

- 查询性能优化：过滤条件优先级



判断满足各自过滤条件的数据源基数，如果存在基数很小的值，忽略其他过滤条件

NTSDB – 其他内核工作

网易大数据
Netease Big Data

- SQL内存执行限制
- Explain analyze开发与增强

```
|   |   └── create_iterator
|   |       ├── labels
|   |       |   ├── cond: application::tag = 'kaola-s2b-oms-compose'
|   |       |   ├── measurement: aggregation_invoke_statistic
|   |       |   └── shard_id: 12457
|   |       ├── seriesKey_fetch_time: 123.999µs
|   |       ├── dataBlock_location_time: 200.663µs
|   |       ├── cursors_ref: 3
|   |       ├── cursors_aux: 0
|   |       ├── cursors_cond: 0
|   |       ├── float_blocks_decoded: 2
|   |       ├── float_blocks_size_bytes: 68
|   |       ├── integer_blocks_decoded: 0
|   |       ├── integer_blocks_size_bytes: 0
|   |       ├── unsigned_blocks_decoded: 0
|   |       ├── unsigned_blocks_size_bytes: 0
|   |       ├── string_blocks_decoded: 0
|   |       ├── string_blocks_size_bytes: 0
|   |       ├── boolean_blocks_decoded: 0
|   |       ├── boolean_blocks_size_bytes: 0
|   |       └── planning_time: 7.624169ms
```

NTSDB – 其他...

网易大数据
Netease Big Data

- NTSDB上到网易云做成Paas服务
- NTSDB支持K8S部署，对接prometheus

网易大数据
Netease Big Data