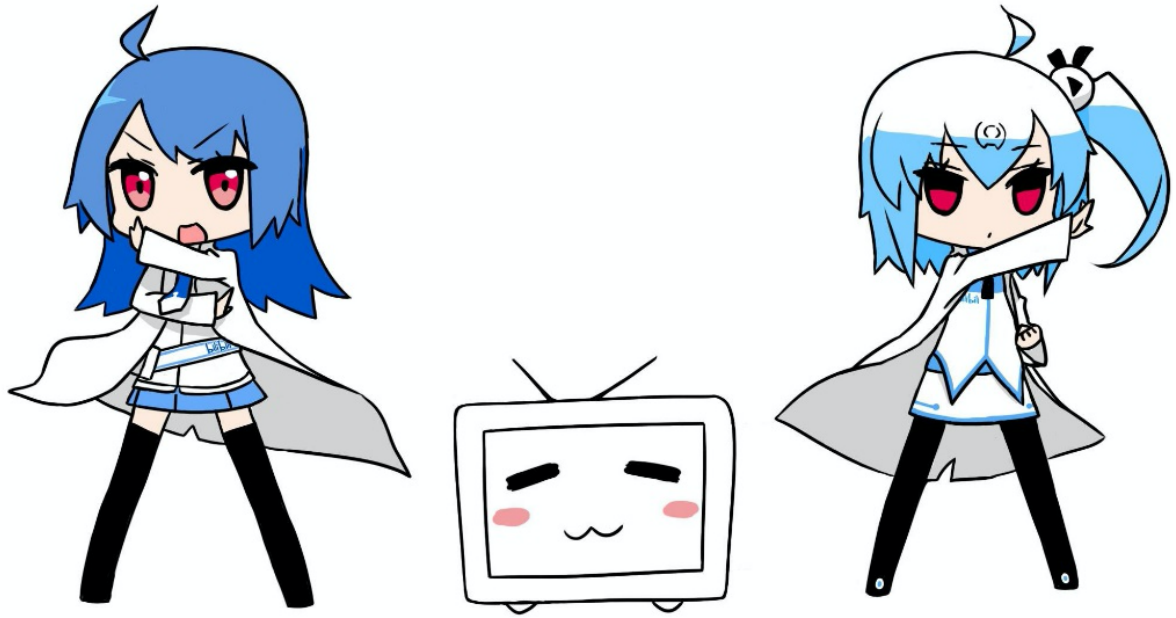
ClickHouse在B站用户行为分析的实践

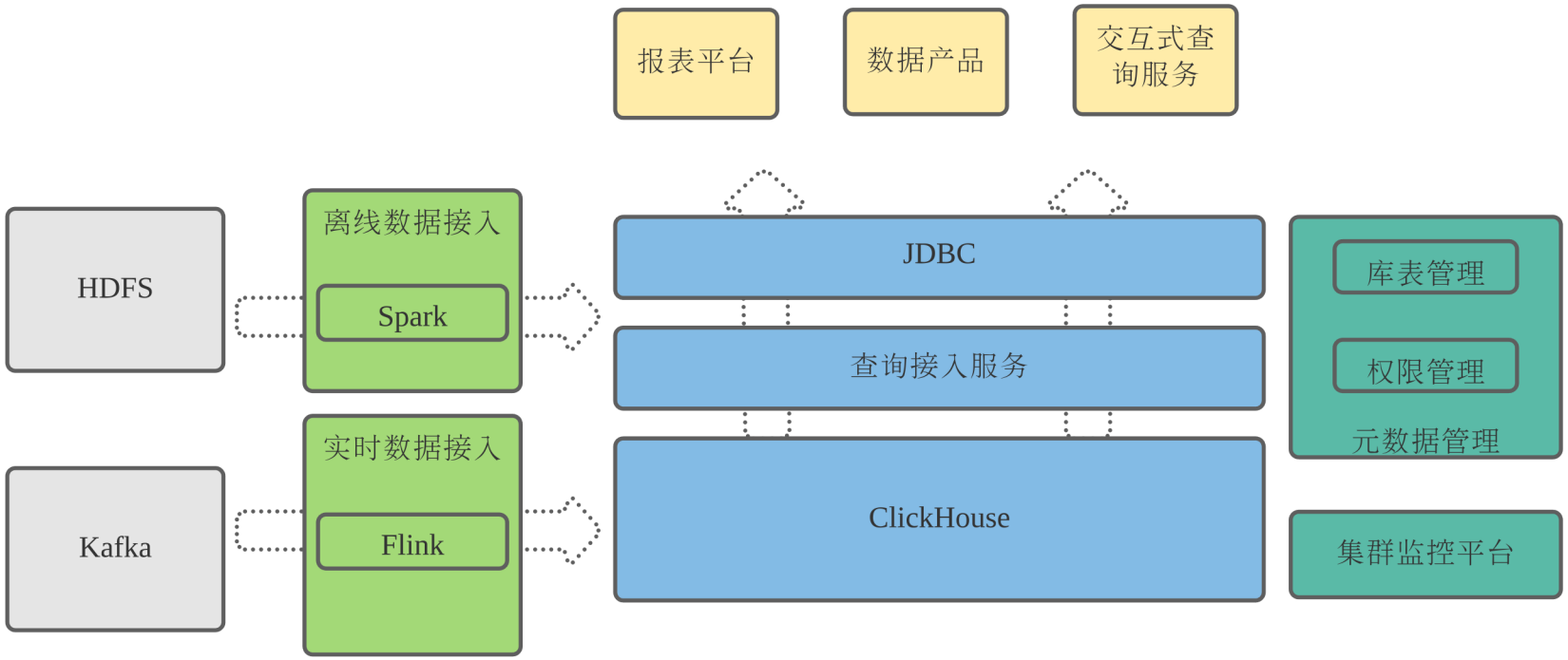
李呈祥



Bilibili OLAP平台现状

- 上百台节点，SSD+HDD存储。
- 每天上千亿级数据摄入。
- 引擎从Kylin/Druid统一收敛到ClickHouse。
- 主要场景包括用户行为分析，标签圈人，监控数据分析等等。

OLAP平台服务


报表平台
数据产品
交互式查询服务
HDFS
离线数据接入
Spark
JDBC
查询接入服务
库表管理
权限管理
元数据管理
Kafka
实时数据接入
Flink
ClickHouse
集群监控平台

用户行为分析需求

数据

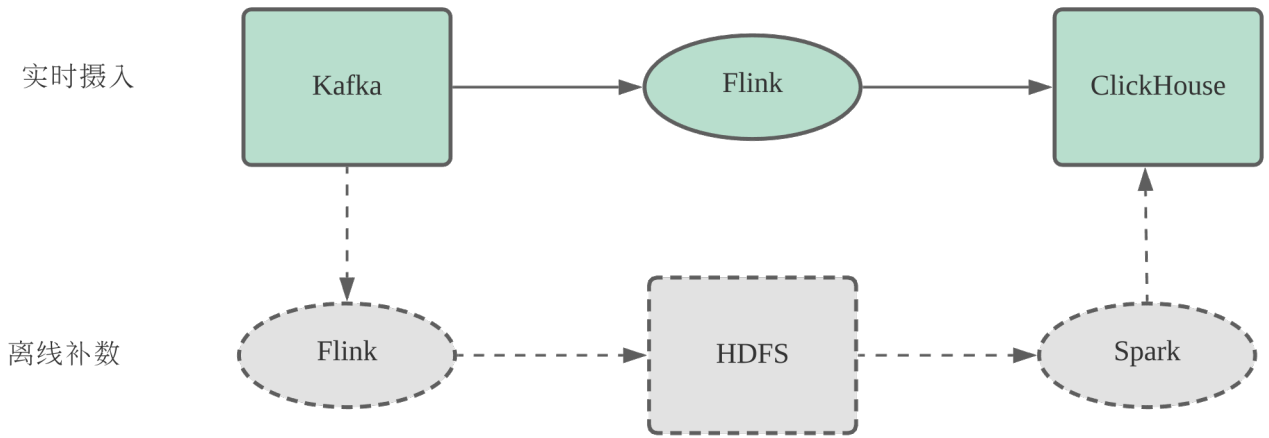
- 每天千亿级数据
- 几千个事件类型
- 实时接入

查询

- 每天上千个查询
- 任意维度
- 任意事件

如何交互式响应超大规模数据的用户行为分析？

数据摄入



Flink实时数据摄入：

- BatchSize 500000，20S内写入。
- ConnectionTimeout/SocketTimeout

表的设计一

```
CREATE TABLE event_analysis_table (
  ......
  `buvid` String CODEC(ZSTD(15)),
  ......
  `brand` LowCardinality(String),
  ......
  `page_type` Int32 CODEC(Delta(4), LZ4HC(6)),
  ......
) ENGINE = ReplicatedMergeTree(...)
PARTITION BY (log_date, app_id)
ORDER BY (...)
TTL ...
SETTINGS storage_policy = 'hot_and_cold',
  use_minimalistic_part_header_in_zookeeper = 1
```

1. ZSTD压缩效率较高，也耗费更多CPU。

2. 对于连续整型数据，Delta压缩非常高效。

3. 对于低基数字段，使用

StringWithDictionary可以大大减少存储开销。

表的设计二

```
CREATE TABLE event_analysis_table (
  ......
  `buvid` String CODEC(ZSTD(15)),
  ......
  `brand` LowCardinality(String),
  ......
  `page_type` Int32 CODEC(Delta(4), LZ4HC(6)),
  ......
) ENGINE = ReplicatedMergeTree(...)
PARTITION BY (log_date, app_id)
ORDER BY (...)
TTL ...
SETTINGS storage_policy = 'hot_and_cold',
  use_minimalistic_part_header_in_zookeeper = 1
```

1. 固定过滤的低基数字段加到分区中。

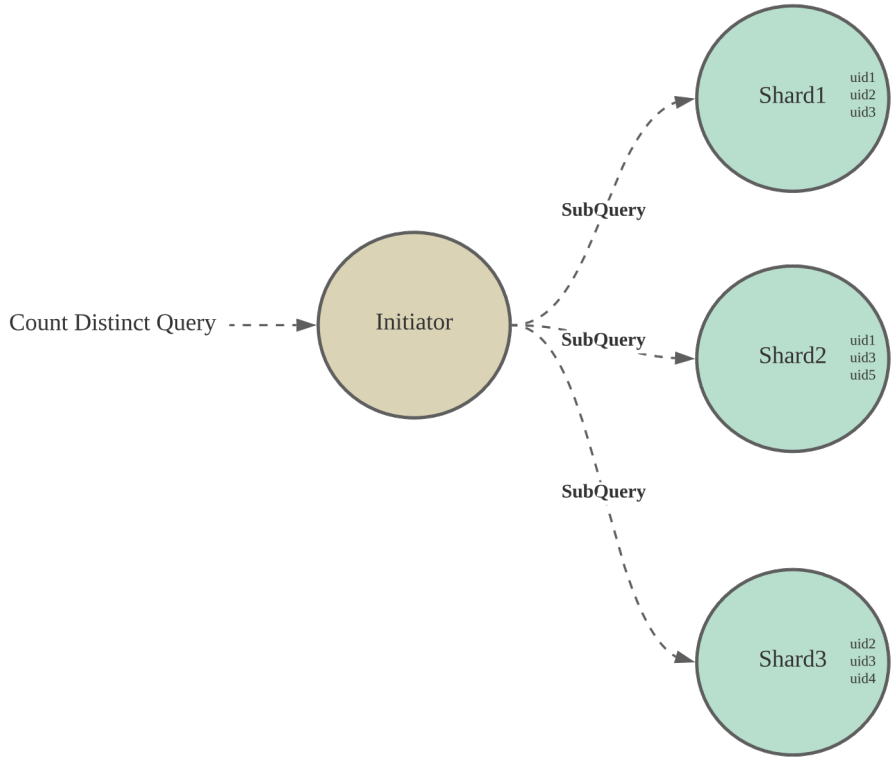
2. 配置存储策略，使用SSD存储新摄入数据。

3. 减少zk存储压力。

```
<policies>
    <hot_and_cold>
      <volumes>
        <hot_volume>
          <disk>ssd1</disk>
          ......
        <hot_volume>
        <cold_volume>
          <disk>storage01</disk>
          ......
        </cold_volume>
      </volumes>
      <move_factor>0.2</move_factor>
    </hot_and_cold>
</policies>
```

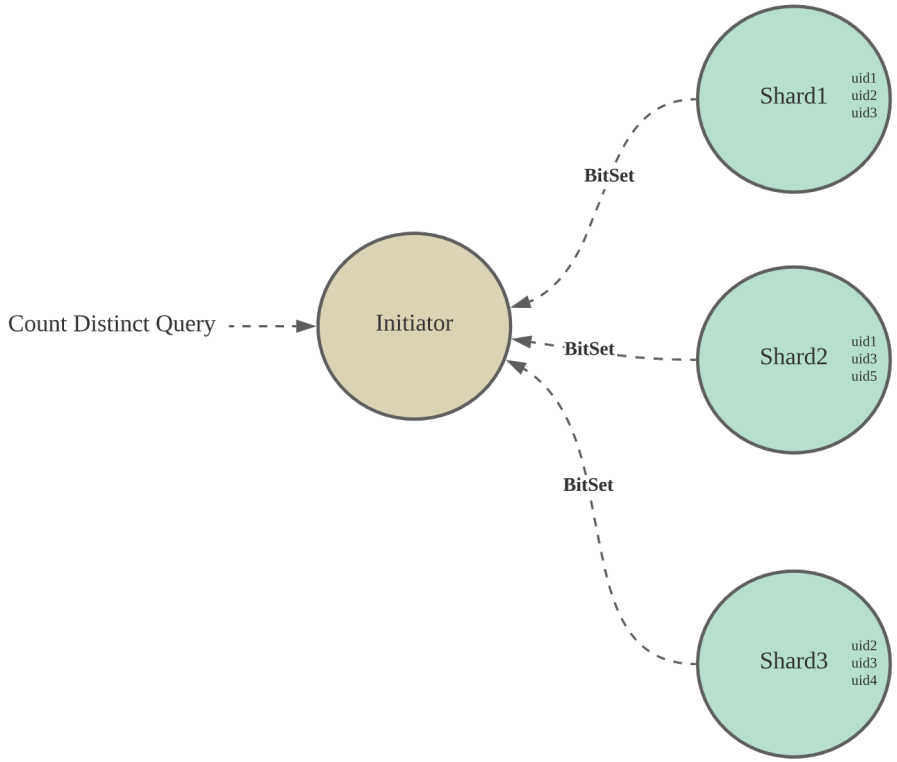
优化实践—Buvid分shard存储

```
SELECT log_date, city, uniqExact(buvid)

FROM event_analysis_table

WHERE log_date >=    '20210129'

      AND log_date <=   '20210206'

      AND event_id='pgc.pgc-video-detail.0.0.pv'

      AND app_id=1

GROUP BY log_date, city
```

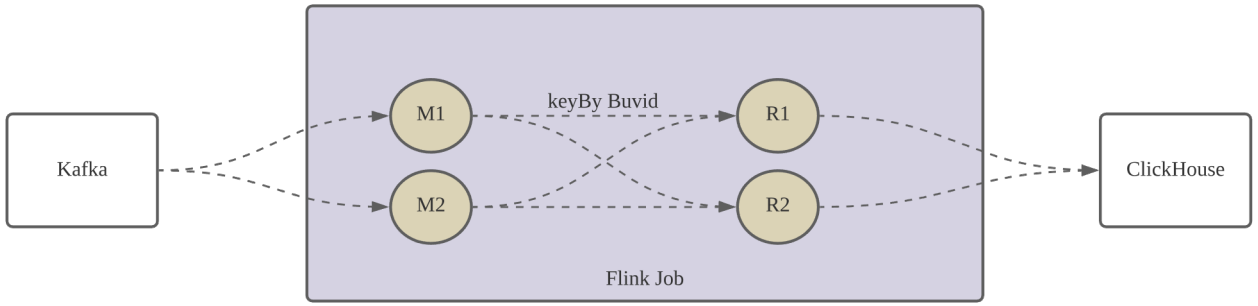


优化实践—Buvid分shard存储

```
SELECT log_date, city, uniqExact(buvid)

FROM event_analysis_table

WHERE log_date >=    '20210129'

      AND log_date <=   '20210206'

      AND event_id='pgc.pgc-video-detail.0.0.pv'

      AND app_id=1

GROUP BY log_date, city
```



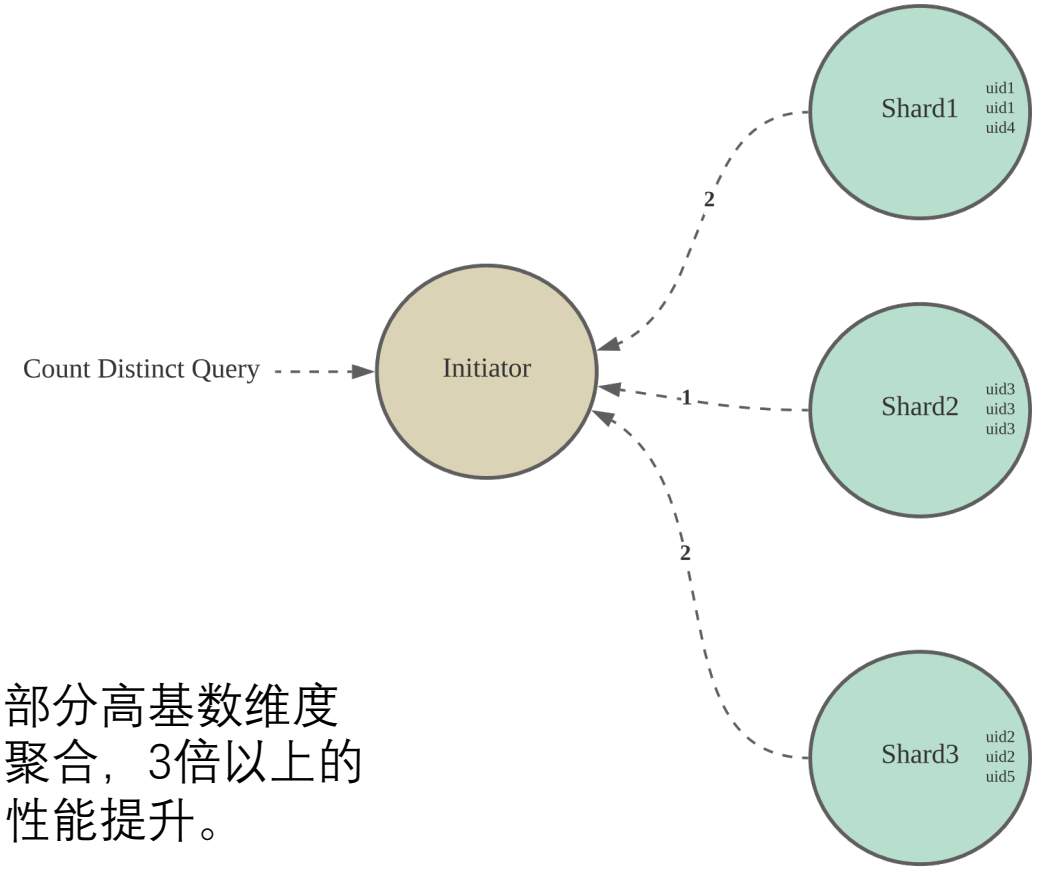
优化实践—Buvid分shard存储



1. Flink任务根据buvid Shuffle。
2. Reduce Task直连Local ClickHouse Table。
3. 通过distributed_group_by_no_merge下推子查询。

优化实践—Buvid分shard存储

```
SELECT log_date, event_id, city, SUM(indicator) AS indicator
FROM
  (SELECT log_date, event_id, city, uniqExact(buvid) AS indicator
  FROM event_analysis_table
  WHERE log_date >= '20201022'
        AND log_date <='20201025'
        AND event_id='pgc.pgc-video-detail.0.0.pv'
        AND app_id=1
  GROUP BY log_date, city
  SETTINGS distributed_group_by_no_merge=1 ) t
GROUP BY logDate, city
```



部分高基数维度聚合，3倍以上的性能提升。

优化实践—Buvid分shard存储

通过参数控制自动在Initiator节点合并结果，对用户透明。

```
SELECT log_date, event_id, city, SUM(indicator) AS indicator
FROM
  (SELECT log_date, event_id, city, uniqExact(buvid) AS indicator
  FROM event_analysis_table
  WHERE log_date >= '20201022'
        AND log_date <='20201025'
        AND event_id='pgc.pgc-video-detail.0.0.pv'
        AND app_id=1
  GROUP BY log_date, city
  SETTINGS distributed_group_by_no_merge=1 ) t
GROUP BY logDate, city
```

```
SELECT log_date, city, uniqExact(buvid)
FROM event_analysis_table
WHERE log_date >=  '20210129'
      AND log_date <=  '20210206'
      AND event_id='pgc.pgc-video-detail.0.0.pv'
      AND app_id=1
GROUP BY log_date, city
SETTINGS distributed_group_by_merge_finalized=1
```

优化实践—物化视图

好处：

1. 预计算加速查询速度。
2. 查询时可能更少的磁盘IO。

UniqExact => Croaring BitMap

Uniq => HyperLogLog

代价：

1. 没有查询计划自动重写，业务需重构SQL。
2. 额外的数据冗余，写入代价变大。
3. 引入高基数维度会导致存储代价太大，查询效率降低。

1. 区分高基数维度和低基数维度。
2. 在写入速度和查询性能之间tradeoff

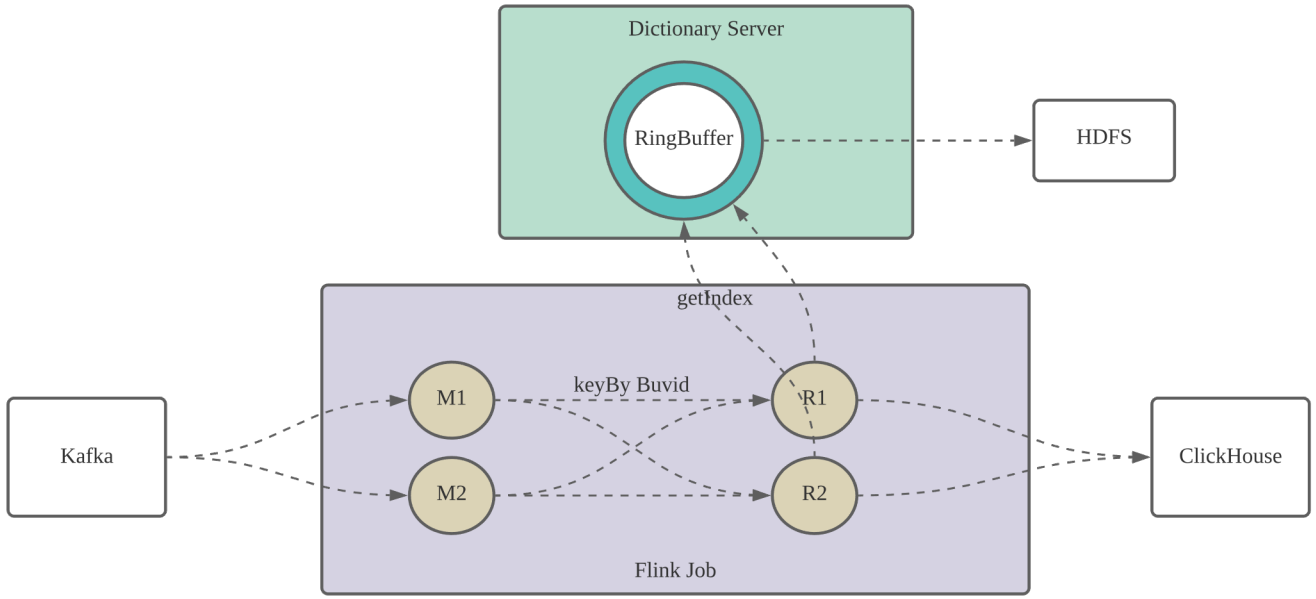
优化实践—Buvid字典映射

相比于groupBitmap，UniqExact计算需要：

1. 更多的磁盘IO。
2. 更复杂的计算。
3. 更多的内存占用。

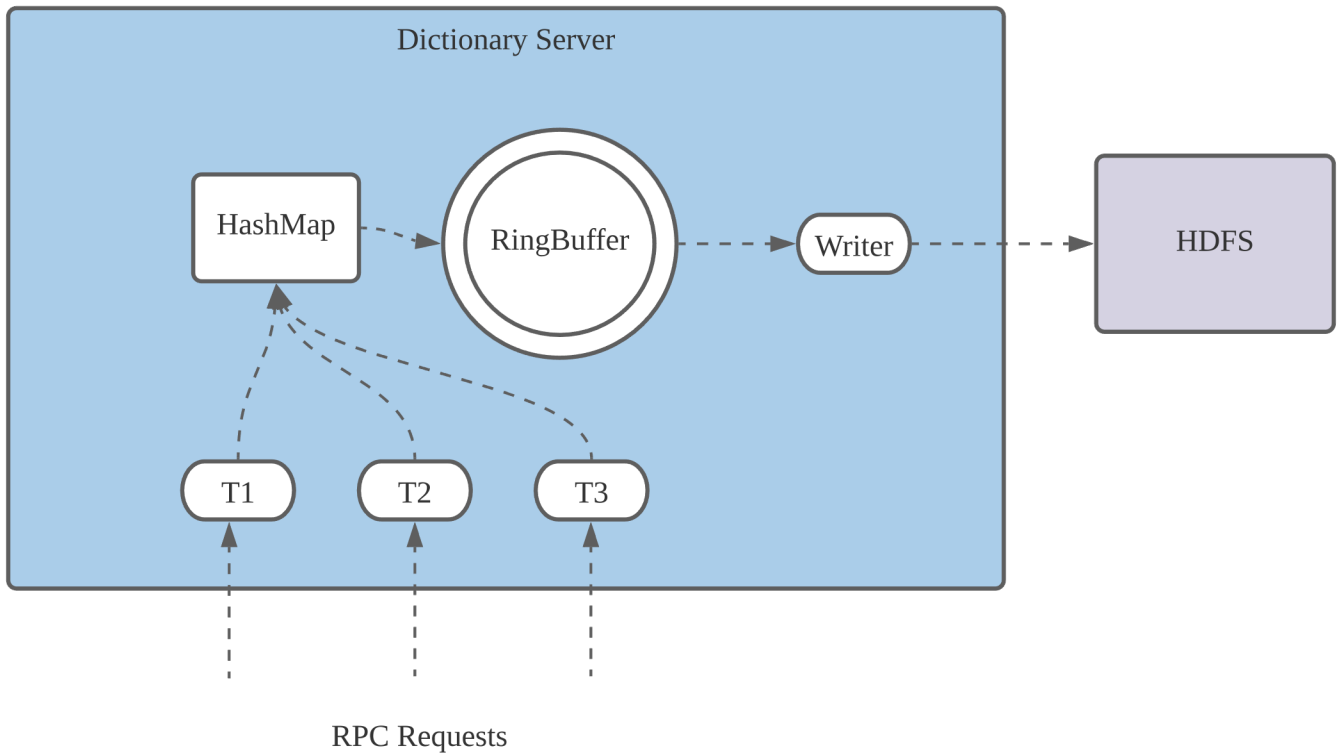
大概2-10倍的性能提升

优化实践—Buvid字典映射



1. 外部Dictionary Server，字典映射值插入ClickHouse。
2. Shuffle by Buvid，在Flink task中缓存字典映射。
3. 离线补数关联Hive字典映射表。

优化实践—Buvid字典映射



1. 单台128G内存可以支持10亿Buvid字典映射。
2. 多生产者单消费者，异步线程持久化数据到HDFS。

查询性能

平均响应时间：
3s

P90响应时间：
5s

每天查询次数：上千次

查询数据时间范围：1个月

优化实践—更多优化计划

1. Array类型查询优化。
2. GroupingSet支持。
3. 漏斗分析定制优化。

We Are Hiring!