ArchSummit
全球架构师峰会 2017

主办方 Geekbang 极客邦科技 InfoQ

FreeWheel 在OLAP上的实践

Bing Jiang
FreeWheel Principal Engineer

极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新

扫一扫下载极客时间App

主办方 Geekbang. 极客邦科技 InfoQ

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14　北京国际会议中心

扫描关注大会官网

SPEAKER INTRODUCE

Bing Jiang（姜冰）

FreeWheel Principal Engineer

姜冰是大数据、分布式系统和系统性能的资深专家。他毕业于中科院计算所，现任FreeWheel数据平台首席工程师，主管数据平台的研发工作。

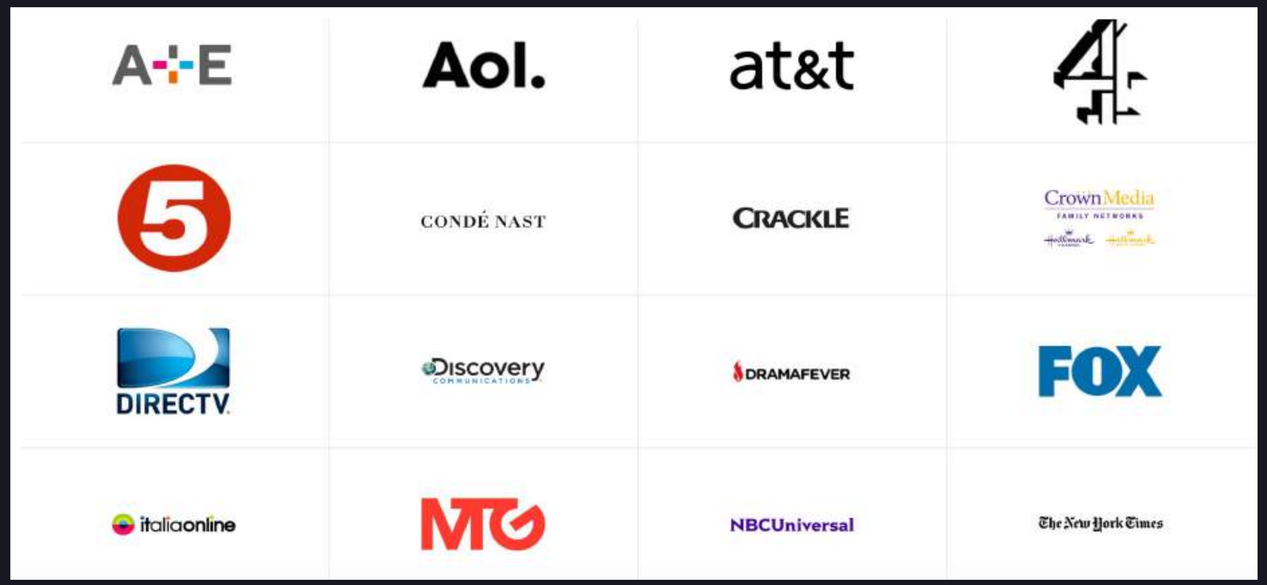
曾供职于Yahoo Hadoop Team，拥有超过8年的大数据系统的研究和实战经验，技术涉猎广泛，对分布式系统、大数据存储、消息队列等领域有深入的理解和实战经验，并擅长排查解决分布式系统的疑难问题。

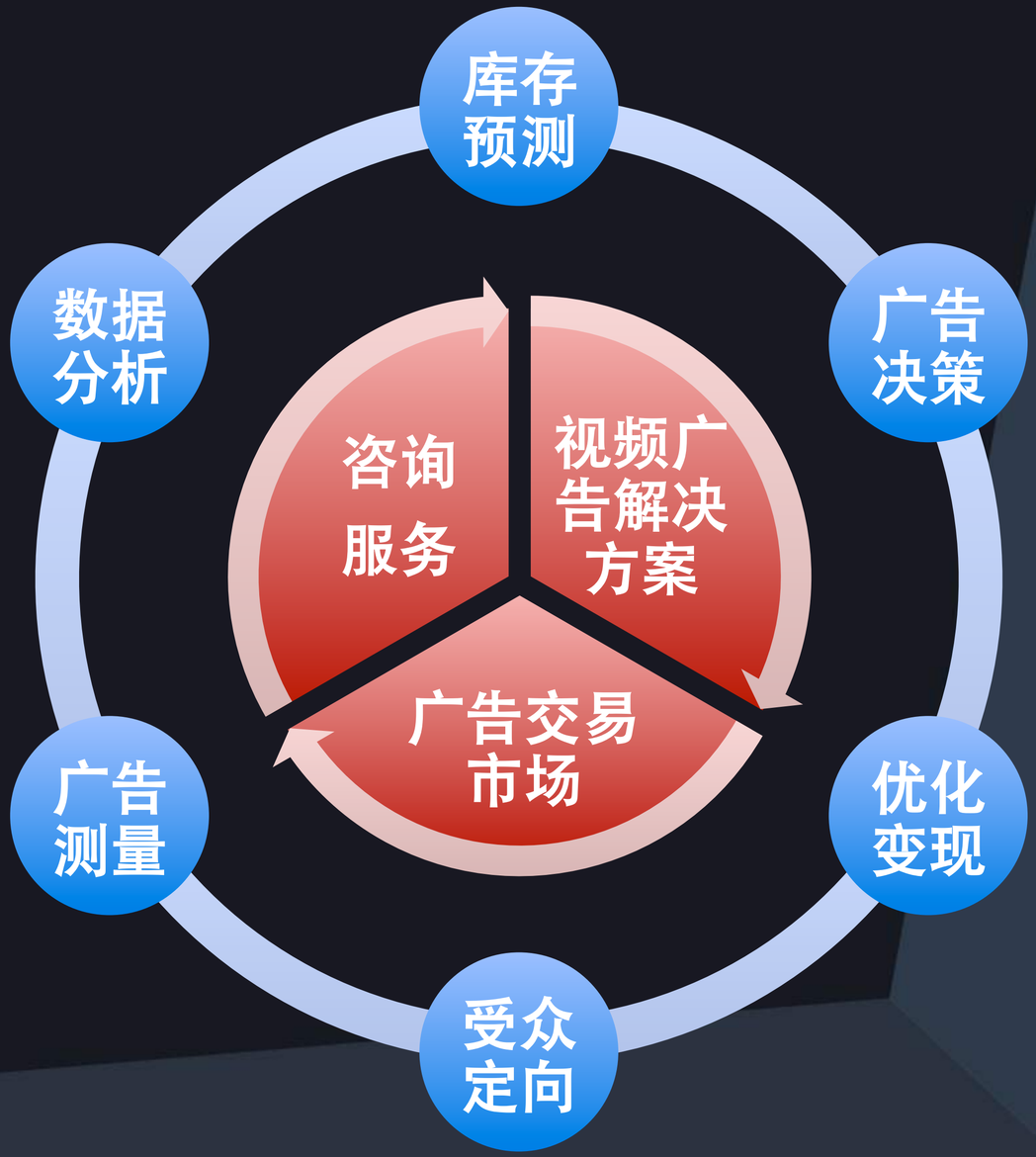
TABLE OF

CONTENTS 大纲

- FreeWheel数据平台简介
- Metadata Service
- CacheLayer Service
- AWS Cloud部署

FreeWheel介绍

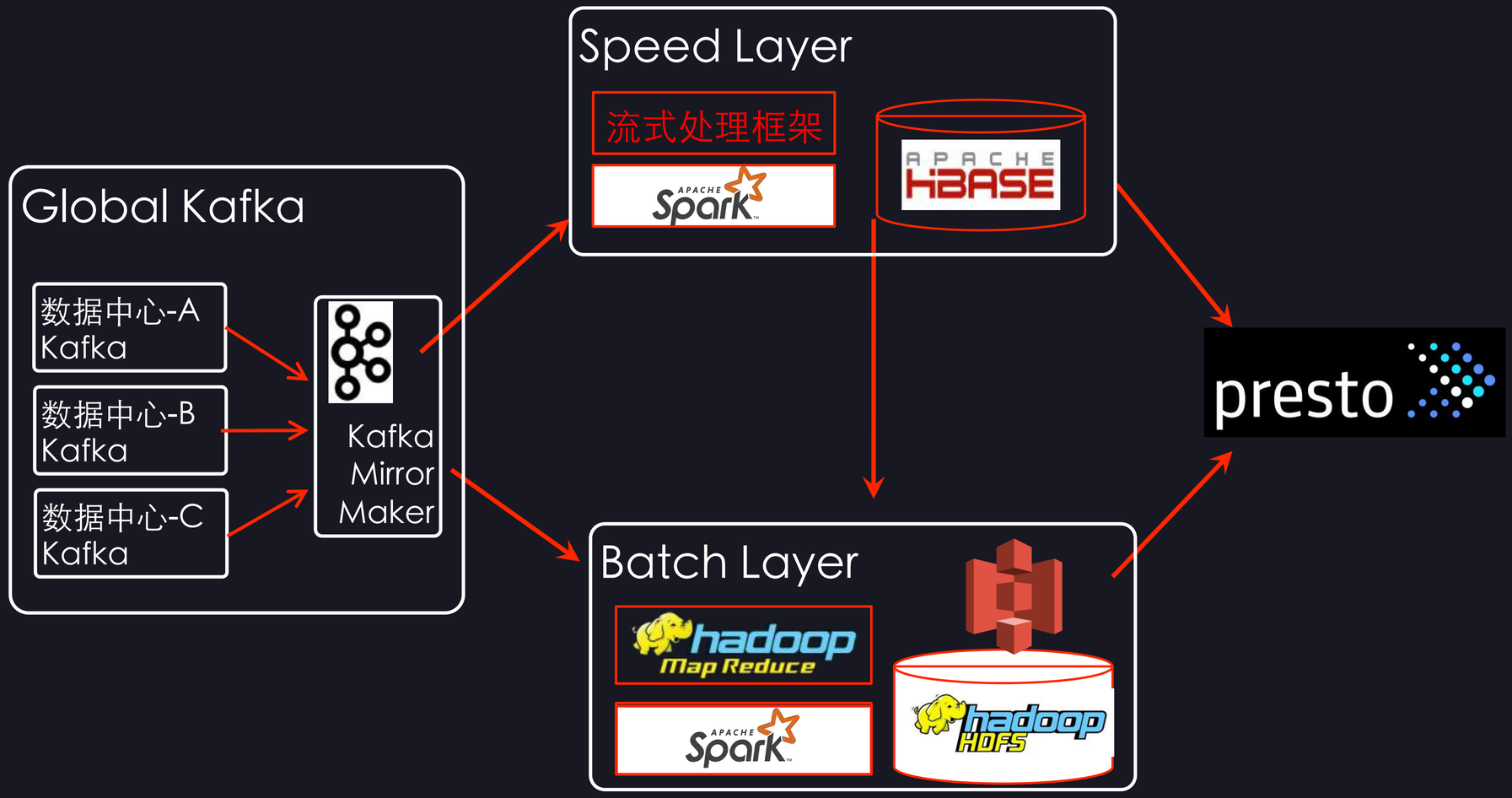




FreeWheel数据特点

- 数据规模
  - >10亿记录/天
  - >3TB日志/天
  - 跨DC
- 数据类型
  - 会话性数据
  - 嵌套结构日志数据
  - 结构化属性数据
- 数据应用类型
  - **报表统计**
  - **业务流量追踪**
  - **实时分析查询**
  - **预测反馈**

数据平台介绍



- Ingesting pipeline
  - 匹配业务日志，写入HBase
  - 周期性将HBase数据写入S3
  - S3文件存储使用Parquet
- Presto
  - 定制化Presto Connector
  - 多个Presto集群
  - 15K+/天的查询量

Presto遇到的问题(1)

- 边界的管理
- 数据可见的原子性

Presto遇到的问题(2)

- 查询性能问题
  - Log Record: 200多列，protobuf -> parquet-avro (较强的业务逻辑侵入)
  - Parquet文件索引信息不足
  - 文件多，600+/h，获取split性能不好

```
{
        "namespace": "tv.freewheel.reporting.avro.schema",
        "name": "Request",
        "type": "record",
        "fields": [
                { "name": "timestamp", "type": "long" },
                { "name": "flags", "type": "long" },

                { "name": "context", "type": "Context" },
                { "name": "visitor", "type": "Visitor" },
                { "name": "slots", "type": { "type": "array", "items": "Slot" } },
                { "name": "advertisements", "type": { "type": "array", "items": "Advertisement" } },

                { "name": "magnifier", "type": ["int", "null"] },
                { "name": "cbp", "type": ["CBP", "null"] },
                { "name": "scores", "type": { "type": "array", "items": "Score" } },
                { "name": "candidates", "type": { "type": "array", "items": "long" } },
                { "name": "private_data_accessible_networks", "type": { "type": "array", "items": "long" } },
                { "name": "external_candidate_ad", "type": ["null", { "type": "array", "items": "External_Candidate"}] },
                { "name": "server_id", "type": ["string", "null"] },
                { "name": "vod_session_id", "type": ["string", "null"] },
                { "name": "is_first_request", "type": ["boolean", "null"] },
                { "name": "identifier", "type": ["Identifier", "null"] },
                { "name": "process_timestamp", "type": ["long", "null"] },
                { "name": "errors", "type": ["null", { "type": "array", "items": "Error" }]},
                { "name": "time_record", "type": ["Time_Record", "null"] },
                { "name": "inventory_group", "type": ["null", { "type": "array", "items": "Inventory_Group"}]},
                { "name": "soft_guaranteed_ad", "type": ["null", { "type": "array", "items": "Soft_Guaranteed_Ad"}] },
                { "name": "audience_item", "type": ["null", { "type": "array", "items": "Audience_Item"}] },
                { "name": "phantom_candidate", "type": ["null", { "type": "array", "items": "Phantom_Candidate"}] },
        { "name": "callback_counters_for_wasted_inventory", "type": ["null", "string"] },
```

Presto遇到的问题 (3)

- 数据变化带来的维护负担
  - 底层文件组织形式的变化
  - 底层存储系统变化 (AWS S3, HDFS)
  - 优化对文件类型有依赖

```
/case-1/2017/12/01/00/file1
/case-2/2017/12/01/file2
```

HDFS	S3
/data/2017/12/01/file1	/b121/2017/12/01/file1
/data/2017/12/02/file2	/ac84/2017/12/02/file2
/data/2017/12/02/file3	/ac84/2017/12/02/file3

- Parquet/ORC/CSV/Text/...

Metadata服务

SQL

```
SELECT
    event_date,
    sum(impression) as imp
FROM
    transaction
WHERE
event_date >= timestmp ‘2017-12-01’
and event_date < timestamp
‘2017-12-04’
and network_id = 263548
GROUP BY 1
```

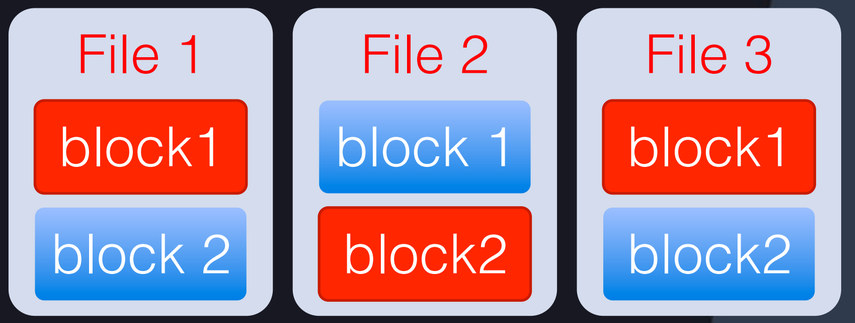
Presto Coordinator

```
event_date: 2017-12-01
network_id: 263548
```

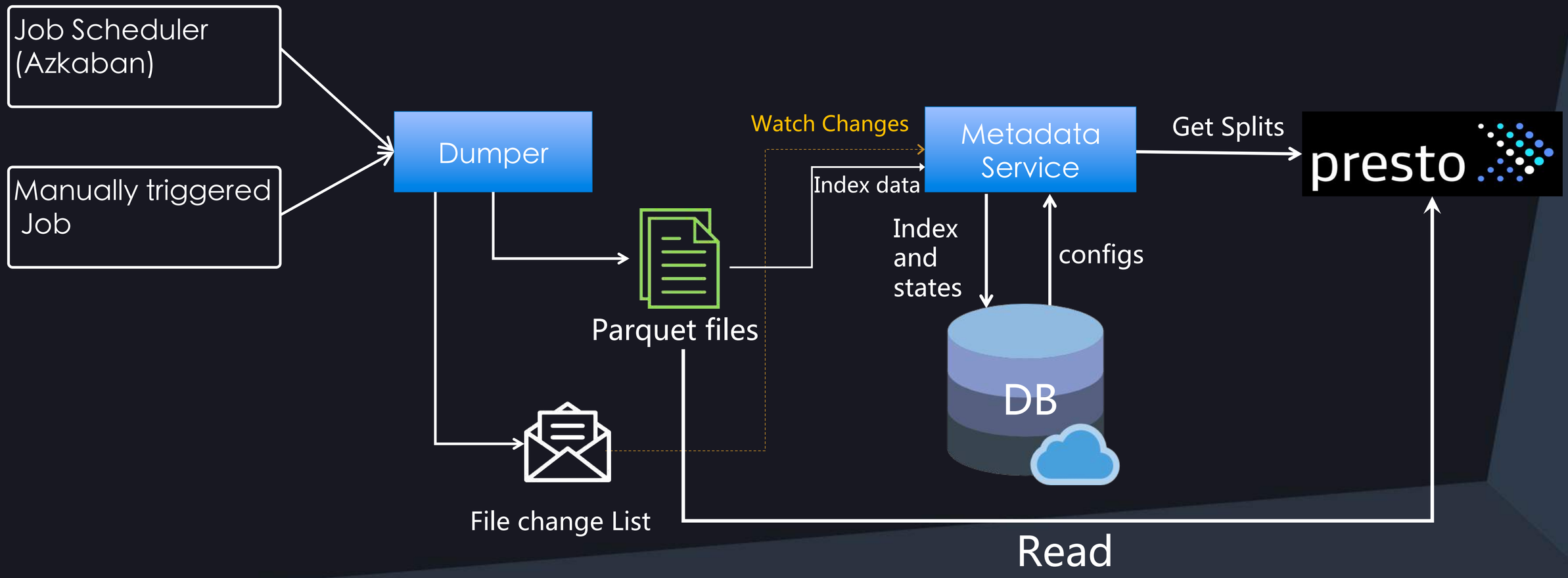
```
event_date: 2017-12-02
network_id: 263548
```

```
event_date: 2017-12-03
network_id: 263548
```

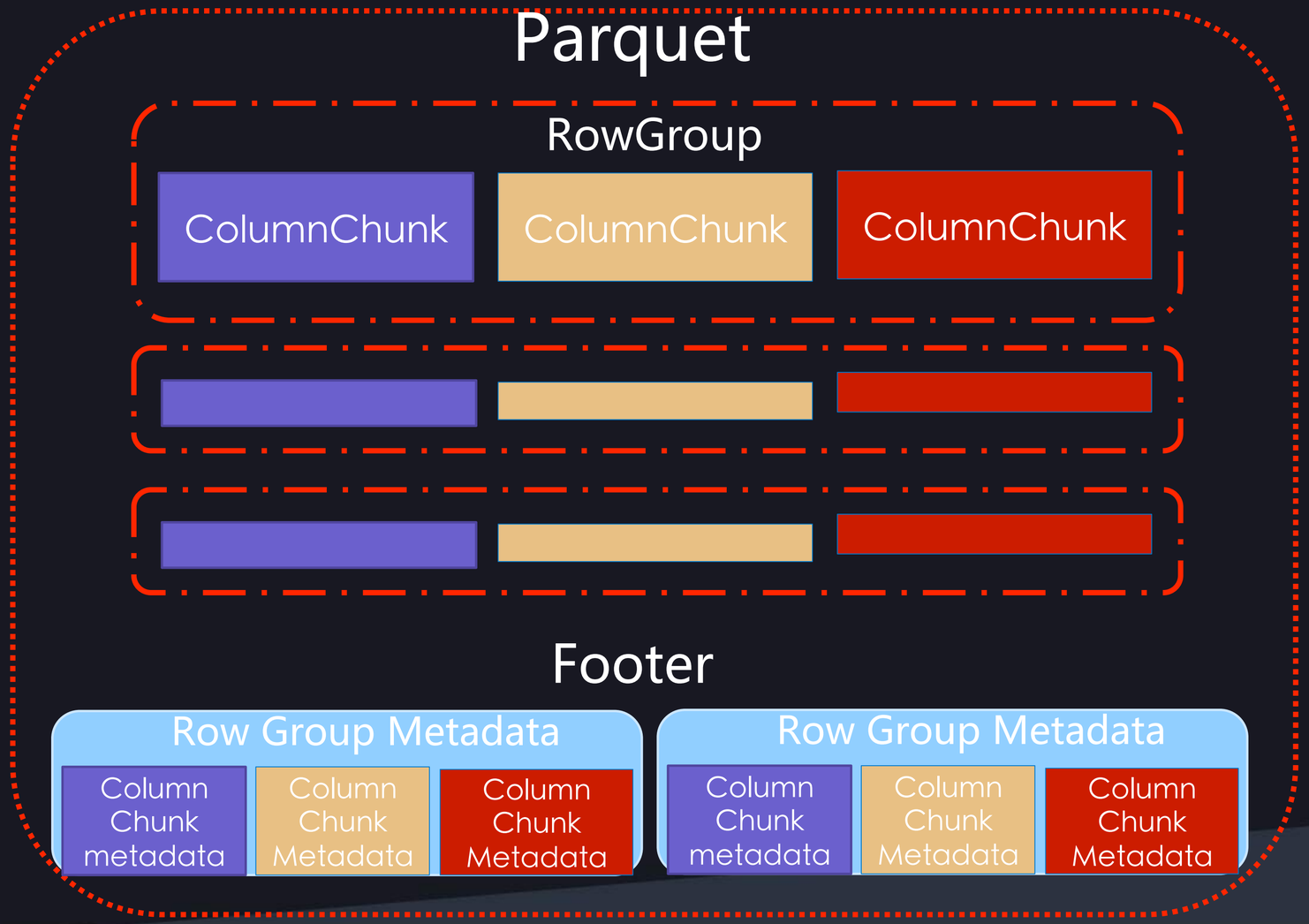
Metadata Service



Metadata服务

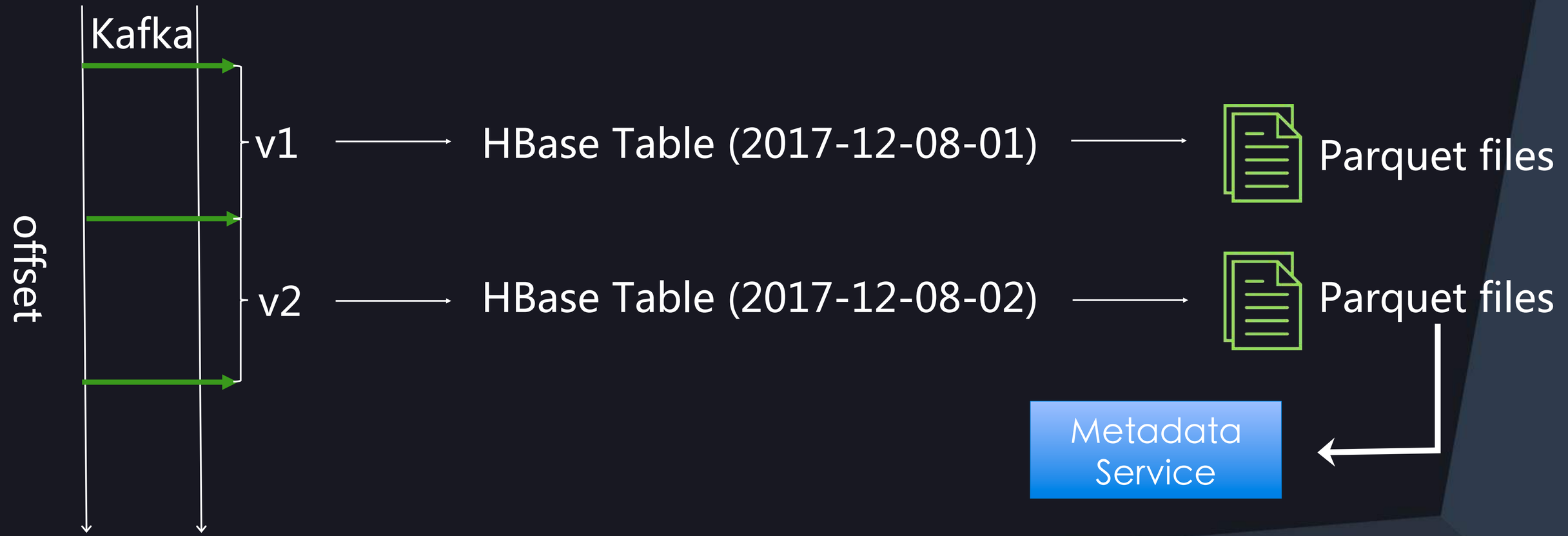

Job Scheduler (Azkaban)
Manually triggered Job
Dumper
Watch Changes
Metadata Service
Get Splits
presto
Index data
Index and states
configs
Parquet files
DB
File change List
Read

与内建索引的区别

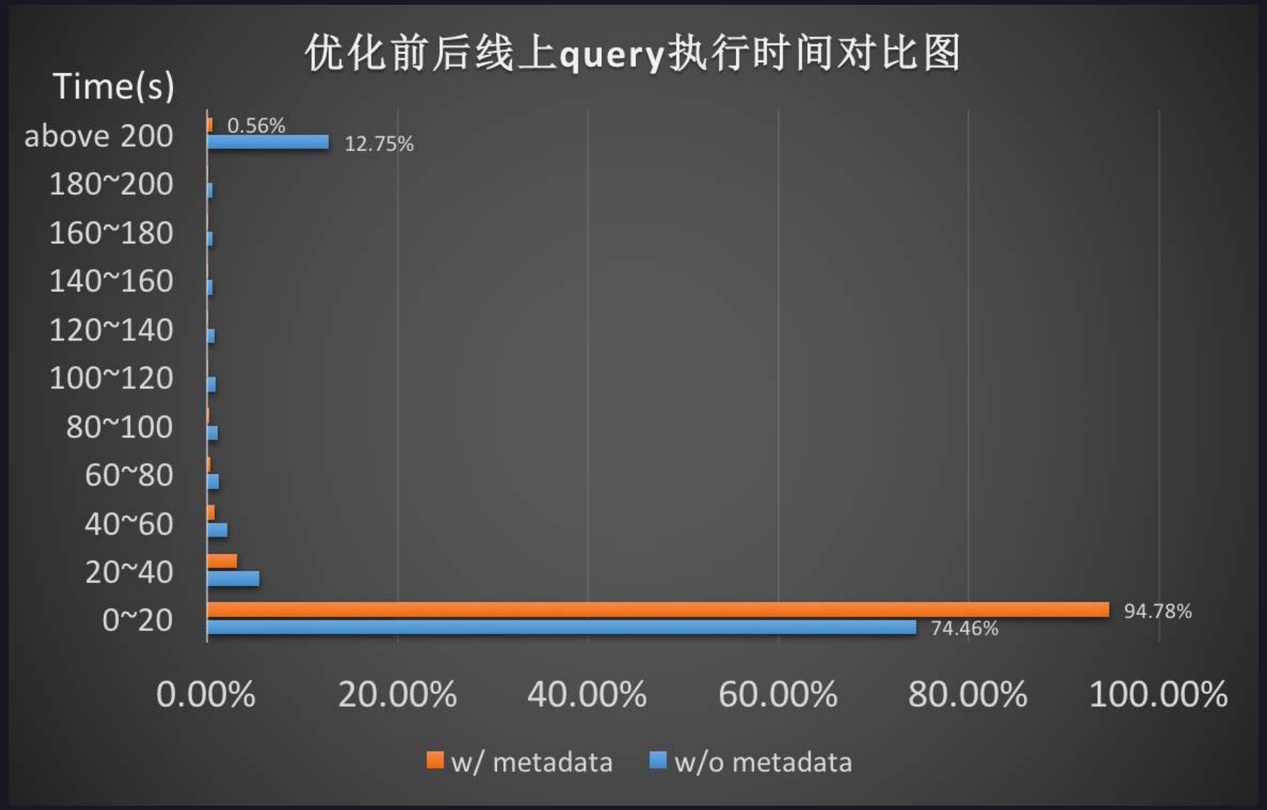


- Parquet: Max/Min
- Metadata Service: Max/Min, Dict info, and bloom filter.
- Parquet文件内建元数据不完整。
- MapReduce/Online indexing
- 内置索引至少要读一次Footer。

数据发布


Kafka
offset
v1
v2
HBase Table (2017-12-08-01)
HBase Table (2017-12-08-02)
Parquet files
Parquet files
Metadata Service

Metadata服务效果



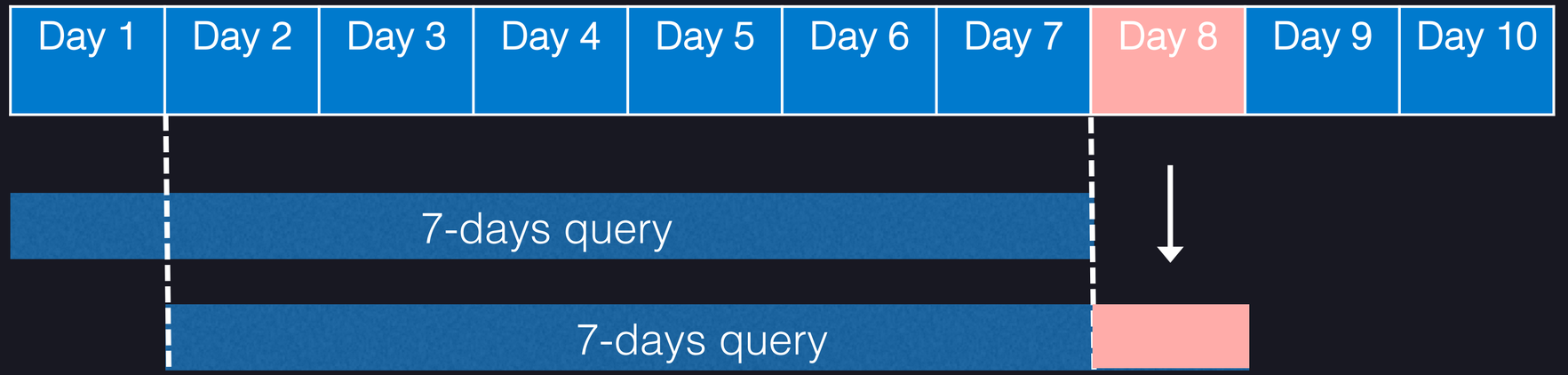
采集Metadata service上线前后各一周的query执行状况。

percentile	W/O	W/
95%	42.01	21.36
99%	269.55	128.69
99.9%	2077.28	819.84

- 上线Metadata Service之后，超过200秒query的比例明显下降。
- 小于20秒的query的比例达到近95%的比例。
- 从95%，99%，99.9%Percentile的执行时间对比，性能提升了近一倍。

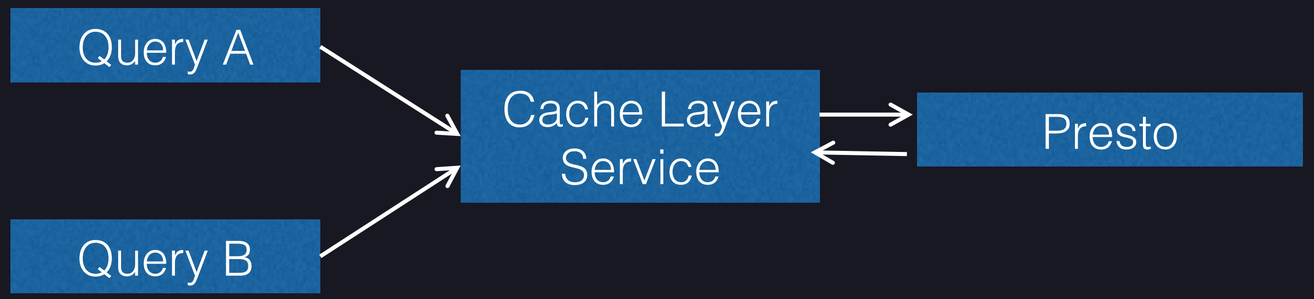
新的挑战

- 周期性的周报、月报查询



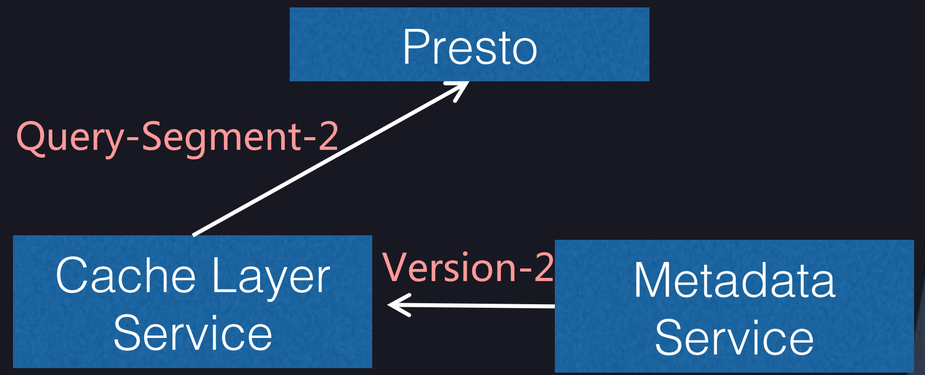
如何避免重复计算？

Cache Layer服务



Query-A	Values	TTL
Query-B	Values	TTL
Query-B-Segment-1	Value	TTL
Query-B-Segment-2	Value	TTL

Child-Query更新



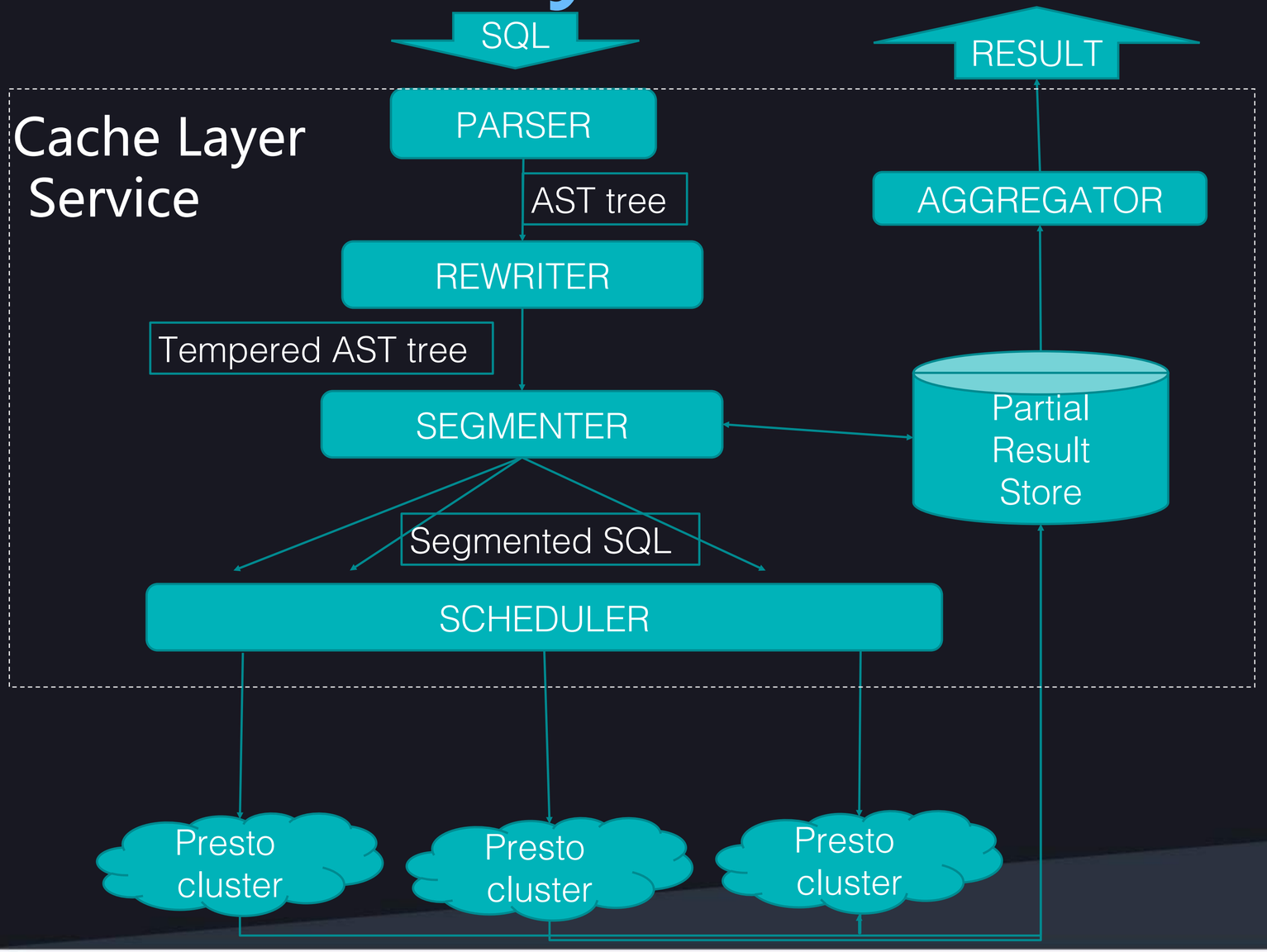
Query-Segment-1	Values	Version-1
Query-Segment-2	Values	Version-2

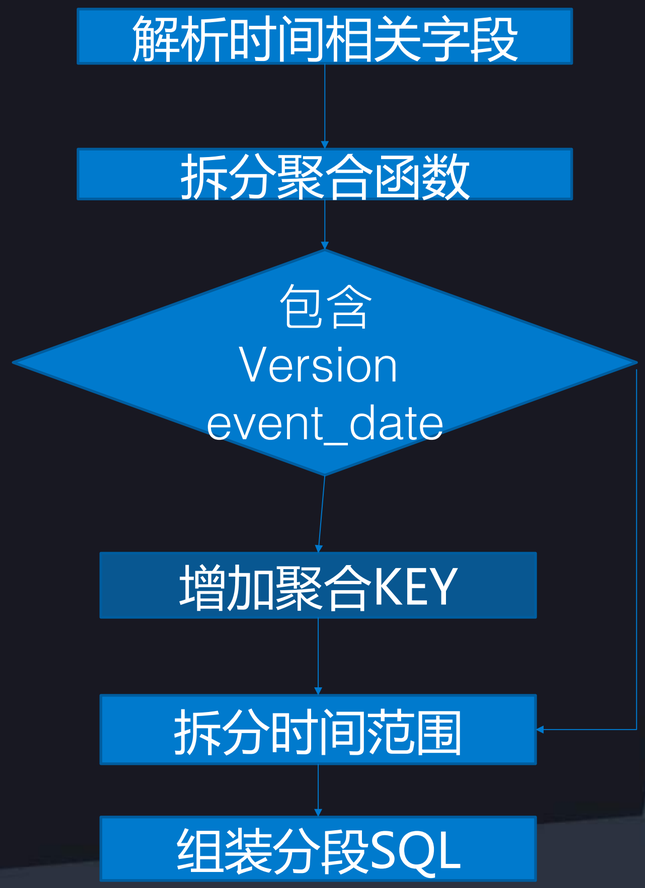
Cache Layer改写查询

```
SELECT
    sum(???) as imp
FROM
    transaction
WHERE
event_date >= timestmp
‘2017-12-01’ and
event_date < timestamp
‘2017-12-08’
and network_id = 263548
```

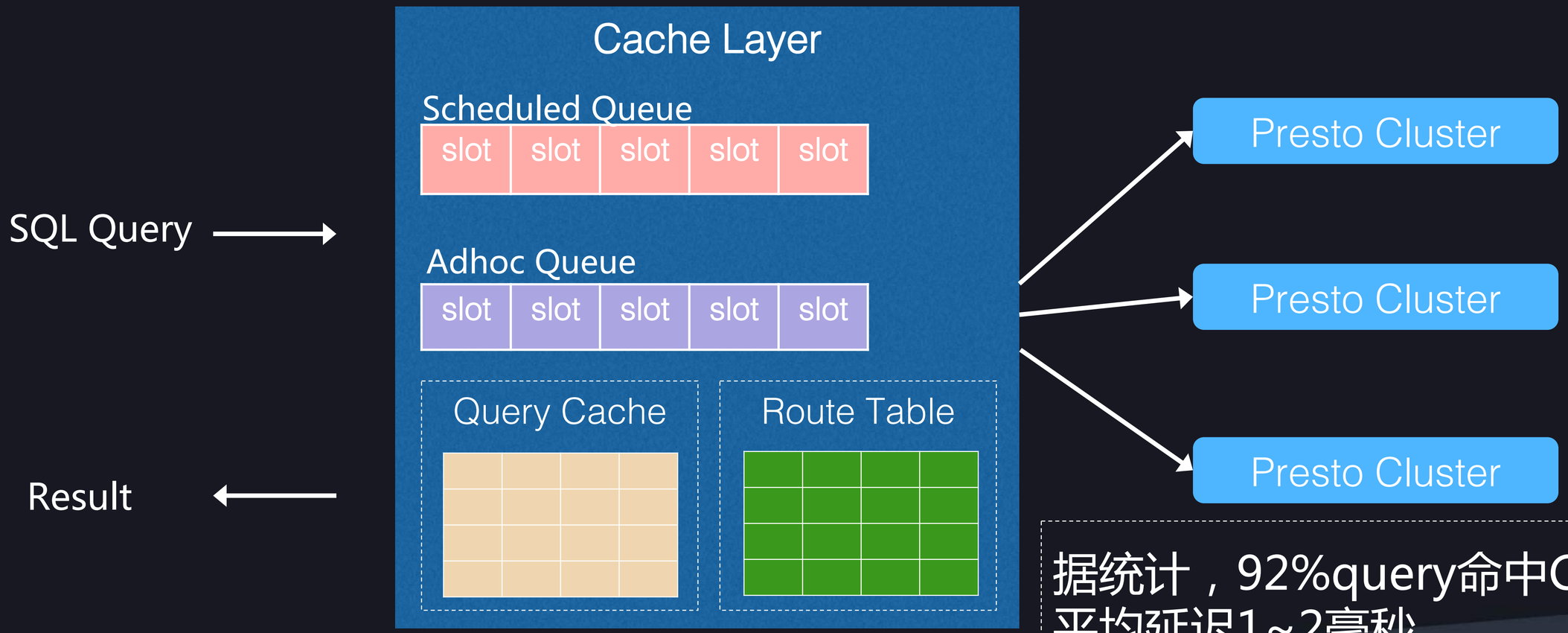
```
SELECT
    event_date,
    data_version,
    sum(???) as imp_n,
FROM
    transaction
WHERE
event_date >= timestmp ‘2017-12-07’
and event_date < timestamp
‘2017-12-08’
  and network_id = 263548
group by 1, 2
```

Cache Layer Service 执行流程图





Cache 调度/Routing

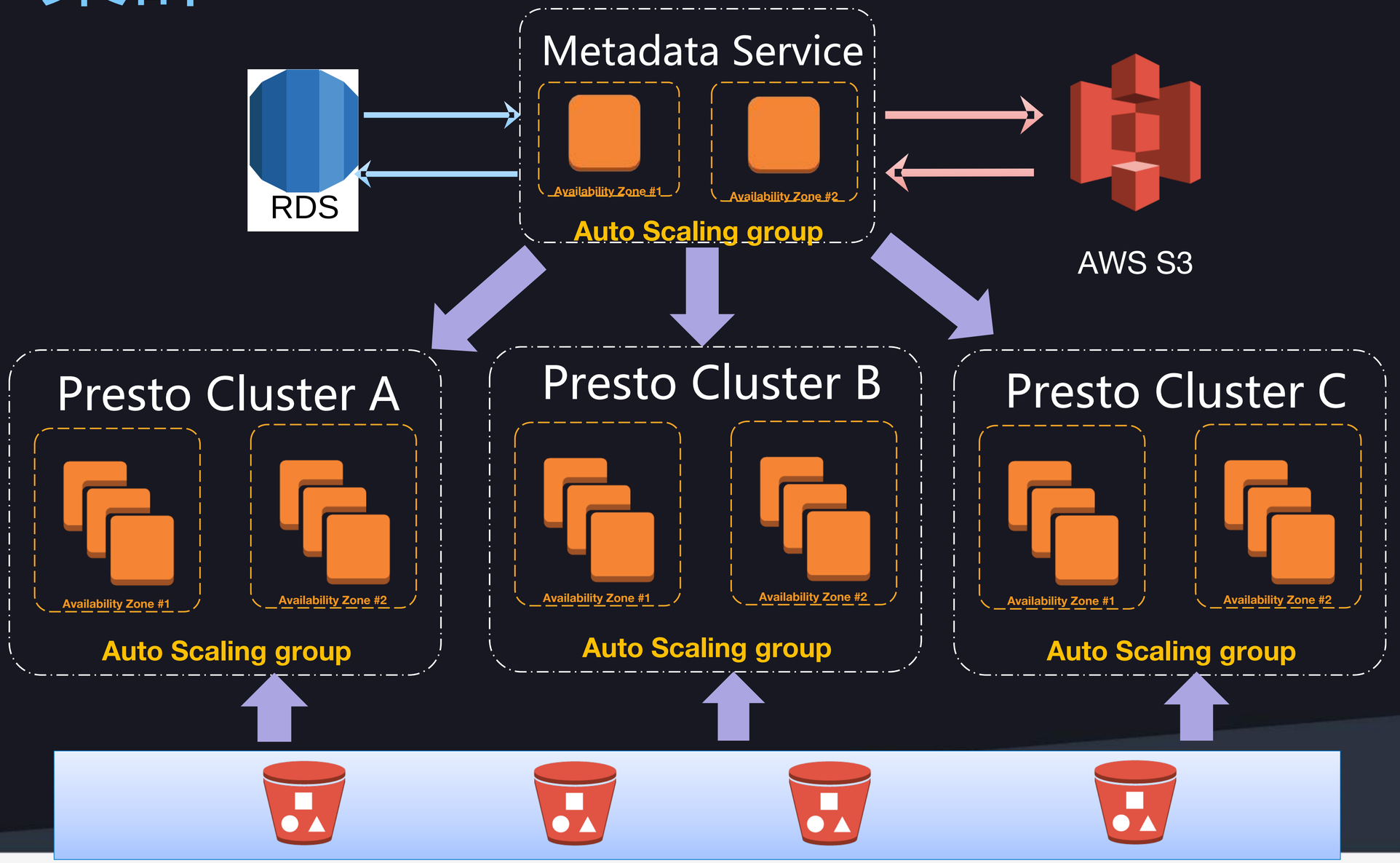


据统计，92%query命中Cache Layer，平均延迟1~2毫秒。

新的问题

我们需要一个资源可伸缩的OLAP。

Presto集群 on AWS



Presto集群伸缩

Admin

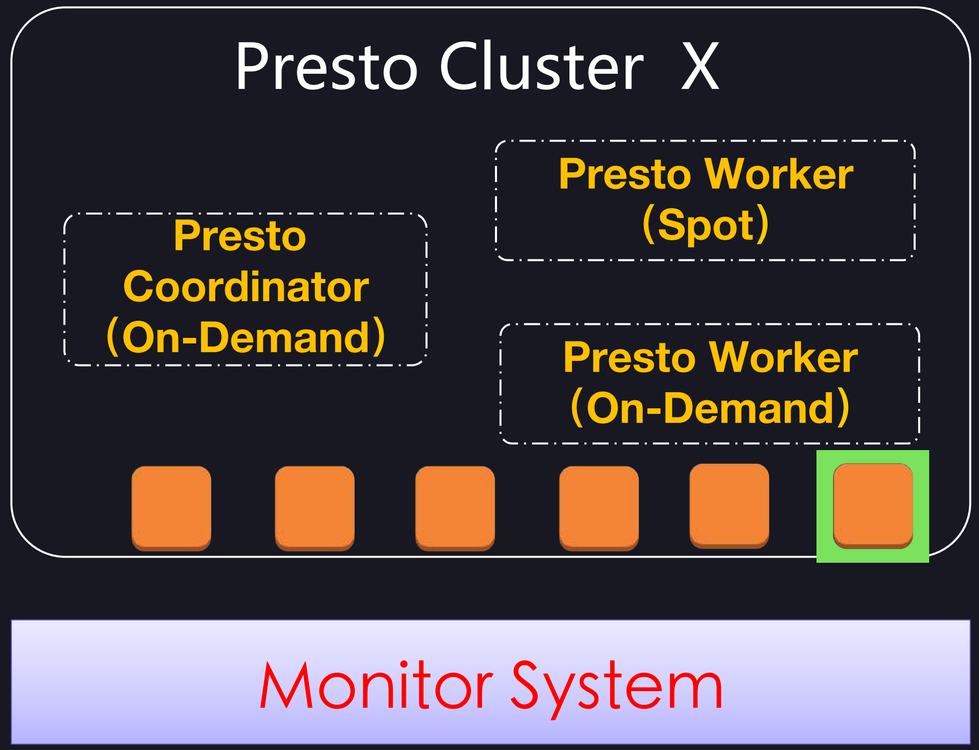
Presto Cluster Launcher

- 集群状态信息（节点/ASG/版本）
- Scale-In: 发送SHUTTING_DOWN信号给 Presto worker(s).
- Scale-Out: 指定ASG加入节点。

- 使用场景：

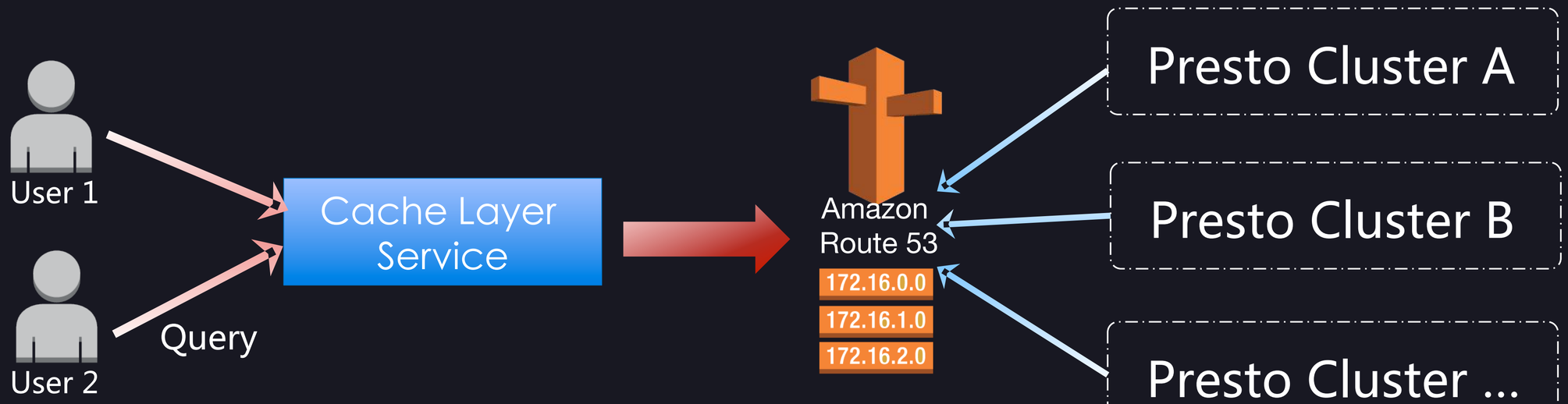
✓ Scheduled Query → Scale-out----running----trigger scale-in →资源回收

✓ Presto InsufficientResourcesFailures → trigger scale-out → 发出告警



- 系统升级维护
- 资源使用率
- 使用Spot Instance需要容忍机器被回收 => 支持On-Demand,Spot-Instance混合部署

Cache Layer on AWS



用户	Primary Cluster	Secondary Cluster
User 1	Presto Cluster A	Presto Cluster C
User 2	Presto Cluster B	Presto Cluster C

- 注册Presto Coordinator地址到AWS Route 53 (DNS服务)。
- Cache Layer根据定义的路由规则和集群容错顺序提交query。

总结

Metadata Service

- 查询效率优化与执行框架解耦
- 数据发布原子性
- 持续提升查询性能

Cache Layer Service

- 降低重复计算
- 缓存query结果，提升对外服务能力
- 设置查询路由表，高可用建设。

Presto on AWS

- 跨AZ高可用架构
- 多Presto集群管理
- Scale-in/Scale-out

以Presto为核心的OLAP服务支撑FreeWheel
快速增长的数据产品和业务需求

FreeWheel

- 姜冰
- Principal Software Engineer
- bjiang@freewheel.tv

THANK YOU

如有需求，欢迎至 [讲师交流会议室] 与我们的讲师进一步交流

ArchSummit
全球架构师峰会 2017