DTCC 2010-2022

第十三届中国数据库技术大会

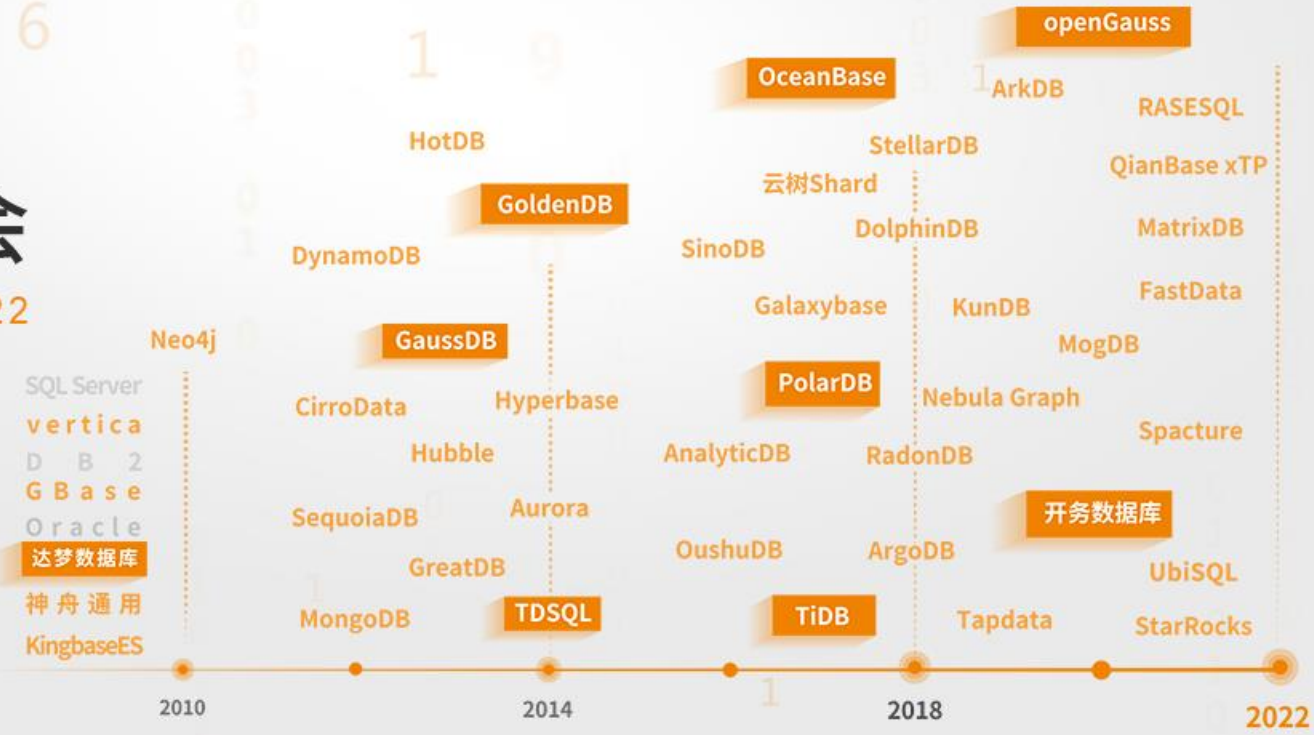
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



Presto在B站性能优化

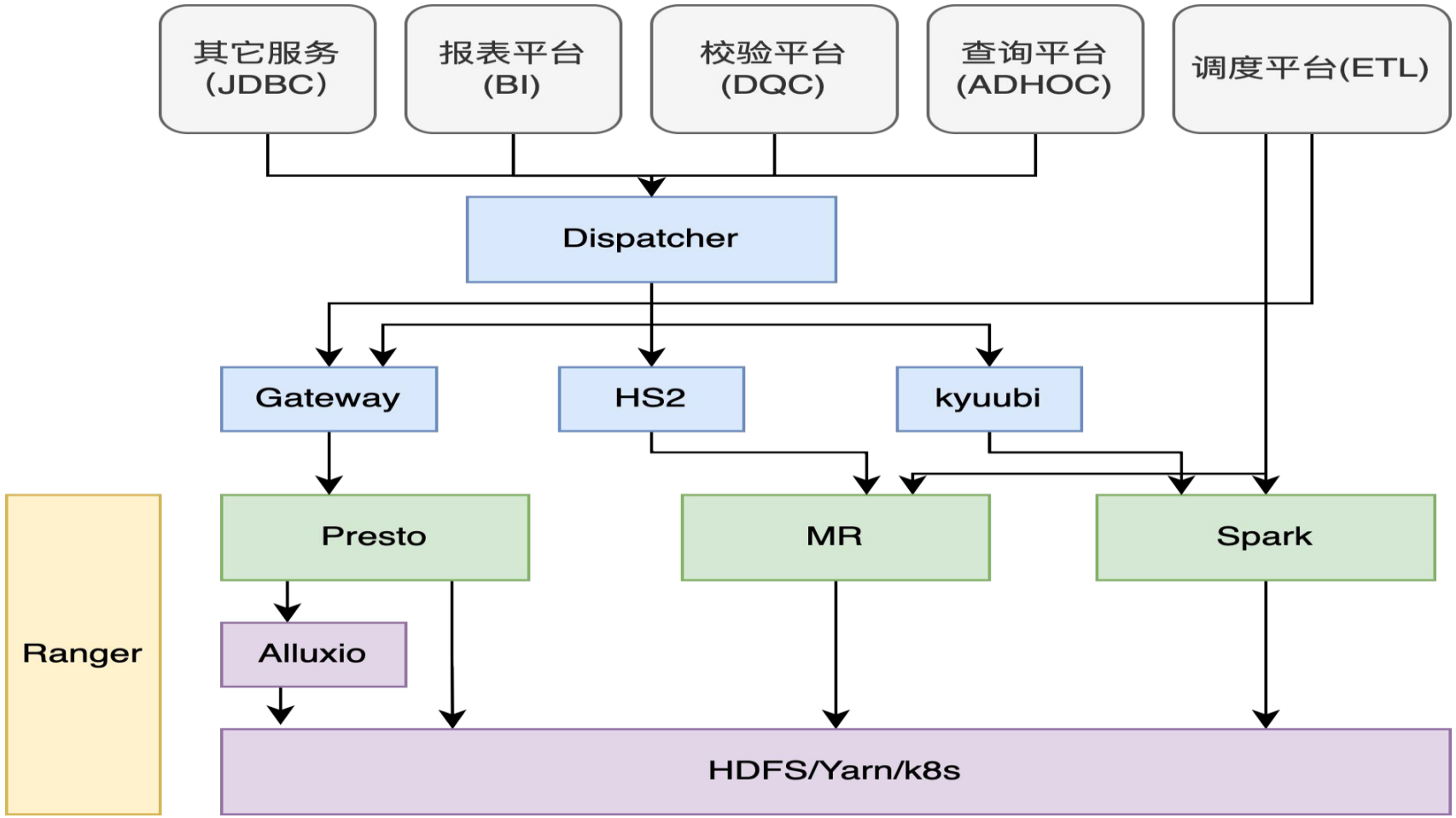
郭建华+b站+研发工程师

个人简介

- 2016～2020 携程大数据离线平台
- 2020～至今 B站大数据离线平台
- 负责Presto计算引擎

目录

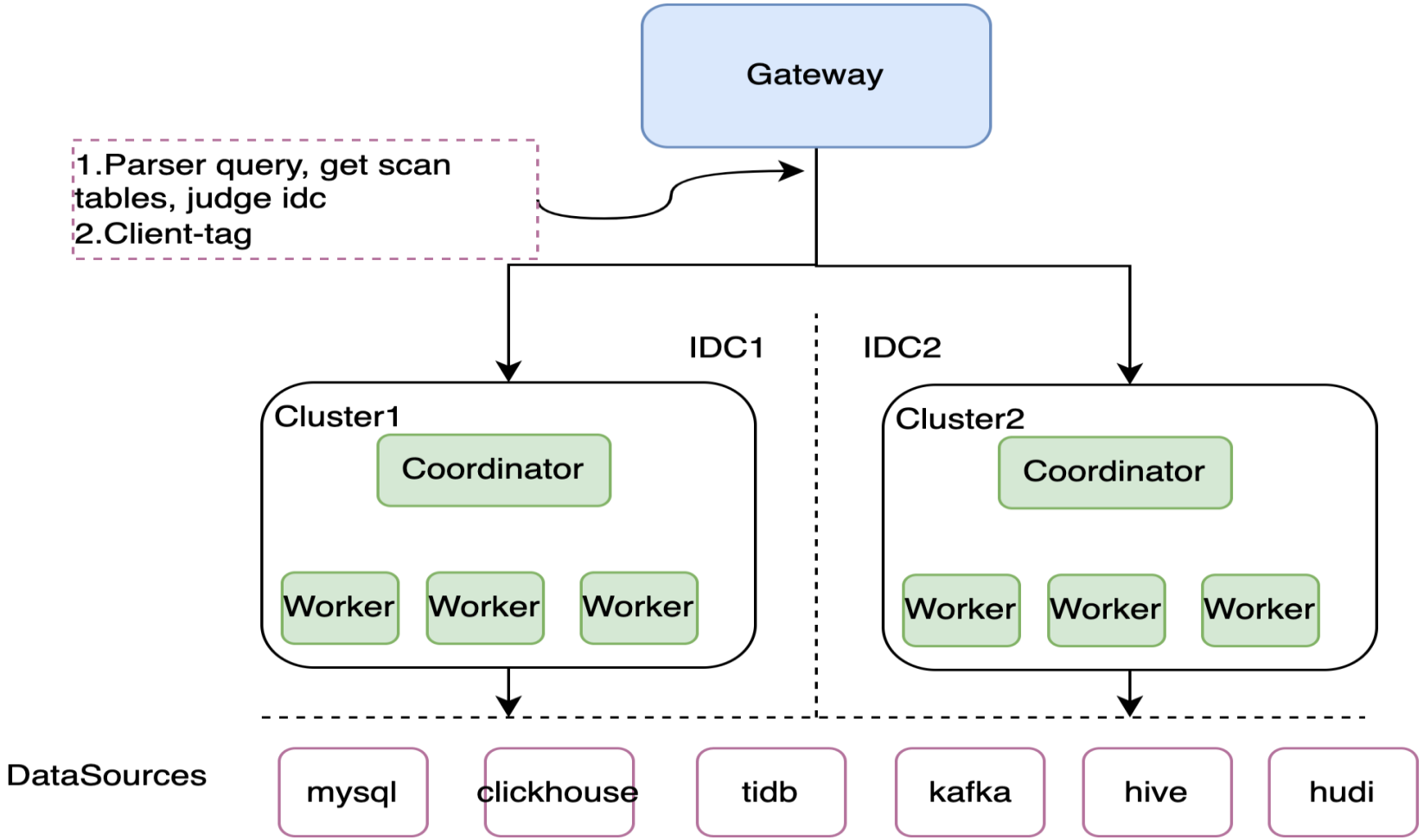
架构


其它服务
(JDBC)
报表平台
(BI)
校验平台
(DQC)
查询平台
(ADHOC)
调度平台(ETL)
Dispatcher
Gateway
HS2
kyuubi
Ranger
Presto
MR
Spark
Alluxio
HDFS/Yarn/k8s

Presto集群现状

- 30W/天
- 20PB/天
- 1200+Worker
- 2 IDC/6 cluster
- Presto-330

Presto集群现状


Gateway
1.Parser query, get scan tables, judge idc
2.Client-tag
IDC1
IDC2
Cluster1
Coordinator
Worker
Worker
Worker
Cluster2
Coordinator
Worker
Worker
Worker
DataSources
mysql
clickhouse
tidb
kafka
hive
hudi

Presto Local Cache--背景

1. 容易受到慢DN影响，导致查询不稳定
2. 少量热表反复被查询
3. 减少网络传输，增加locality提升查询性能

Presto Local Cache--热度统计

```
select
  items
from
  tmp_db.tmp_table
where
  (log_date between '20220209' and '20220210')
  and (log_hour between '21' and '22') limit 10;
```


```
{
    "queryId": "20220216_085627_00000_inqkv",
    "querystr": "select items... from ai.tablexxx where log_date||log_hour between '20220209'||'21' and '20220210'||'22' limit 10;"
    "lineageInfo": "{\"inputs\":[{\"catalogName\":\"hive\",\"schema\":\"ai\",\"table\":\"xxxx\",\"columns\":[\"key\",\"log_hour\",\"value#features\",\"log
    Info\":{\"partitionIds\":[\"log_date=20220209/log_hour=21\",\"log_date=20220209/log_hour=22\",\"log_date=20220209/log_hour=23\",\"log_date=20220210/
    log_hour=00\",\"log_date=20220210/log_hour=01\",\"log_date=20220210/log_hour=02\",\"log_date=20220210/log_hour=03\",\"log_date=20220210/log_hour=04\",
    log_hour=05\",\"log_date=20220210/log_hour=06\",\"log_date=20220210/log_hour=07\",\"log_date=20220210/log_hour=08\",\"log_date=20220210/log_hour=09\",
    log_hour=10\",\"log_date=20220210/log_hour=11\",\"log_date=20220210/log_hour=12\",\"log_date=20220210/log_hour=13\",\"log_date=20220210/log_hour=14\",
    log_hour=15\",\"log_date=20220210/log_hour=16\",\"log_date=20220210/log_hour=17\",\"log_date=20220210/log_hour=18\",\"log_date=20220210/log_hour=19\",
    log_hour=20\",\"log_date=20220210/log_hour=21\",\"log_date=20220210/log_hour=22\"],\"truncated\":false}}],\"output\":null,
    "inputTables": "[ai.xxx]",
    "inputCols": "[ai.xxx.key, ai.xxx.log_date, ai.xxx.log_hour, ai.xxx.value#features]"}"
}
```

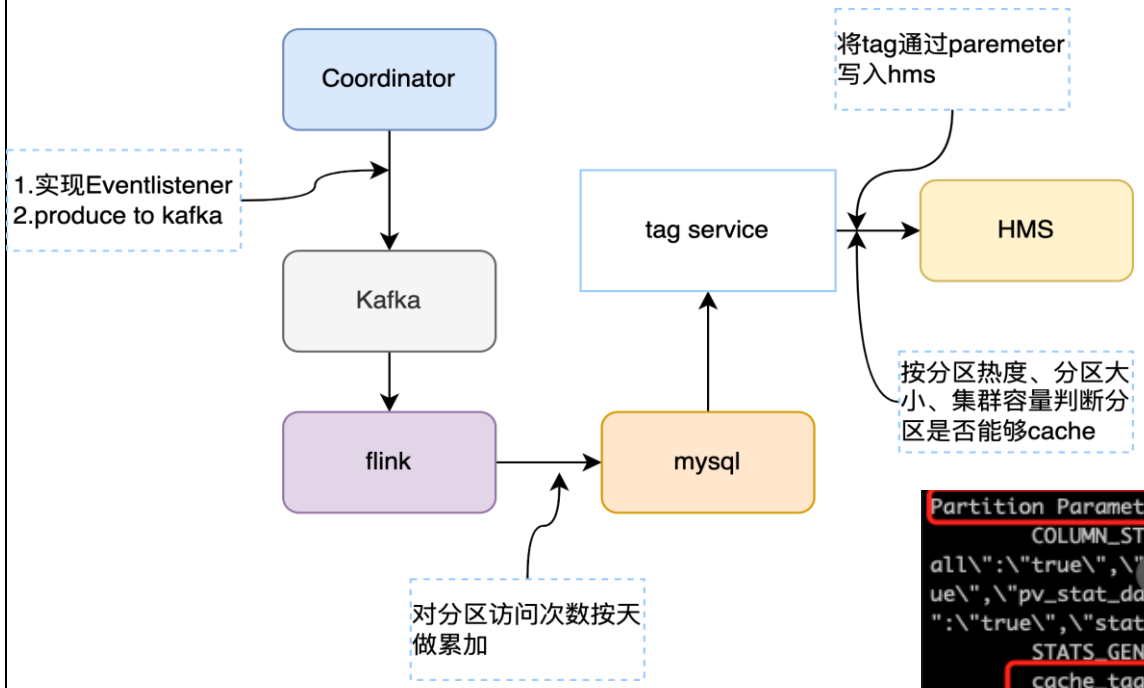
Presto Local Cache--热度统计

cluster: jscs-ai-offl dbName: tableName: tag: 1 查询

cluster	dbName	tableName	partitionTable	location	sameLocation	addTime	updateTime	tableHeat	tag	ttl	action
jscs-ai-offline	ai	dalao_crow	1	viewfs://jssz-big-cluster-par tmen wa reho dal ao_c l_c ard	1	2021-11-11 T08:31:54.000+00:00	2022-11-02 T09:52:31.000+00:00	2097	1	8	消除tag 详情
jscs-ai-offline	ai	search_senj	1	viewfs://jssz-bi a-cl uste par tme /wa reh /se arch nju_ int ce	1	2021-11-16 T14:04:14.000+00:00	2022-11-02 T09:54:39.000+00:00	619	1	96	消除tag 详情
jscs-ai-offline	ai	video_unde	1	viewfs://jssz-big-cluster-par tmer /CV _NL deo _un tan d	1	2022-02-14 T08:51:29.000+00:00	2022-11-02 T09:29:17.000+00:00	495	1	45	消除tag 详情

Presto Local Cache--如何只缓存热表

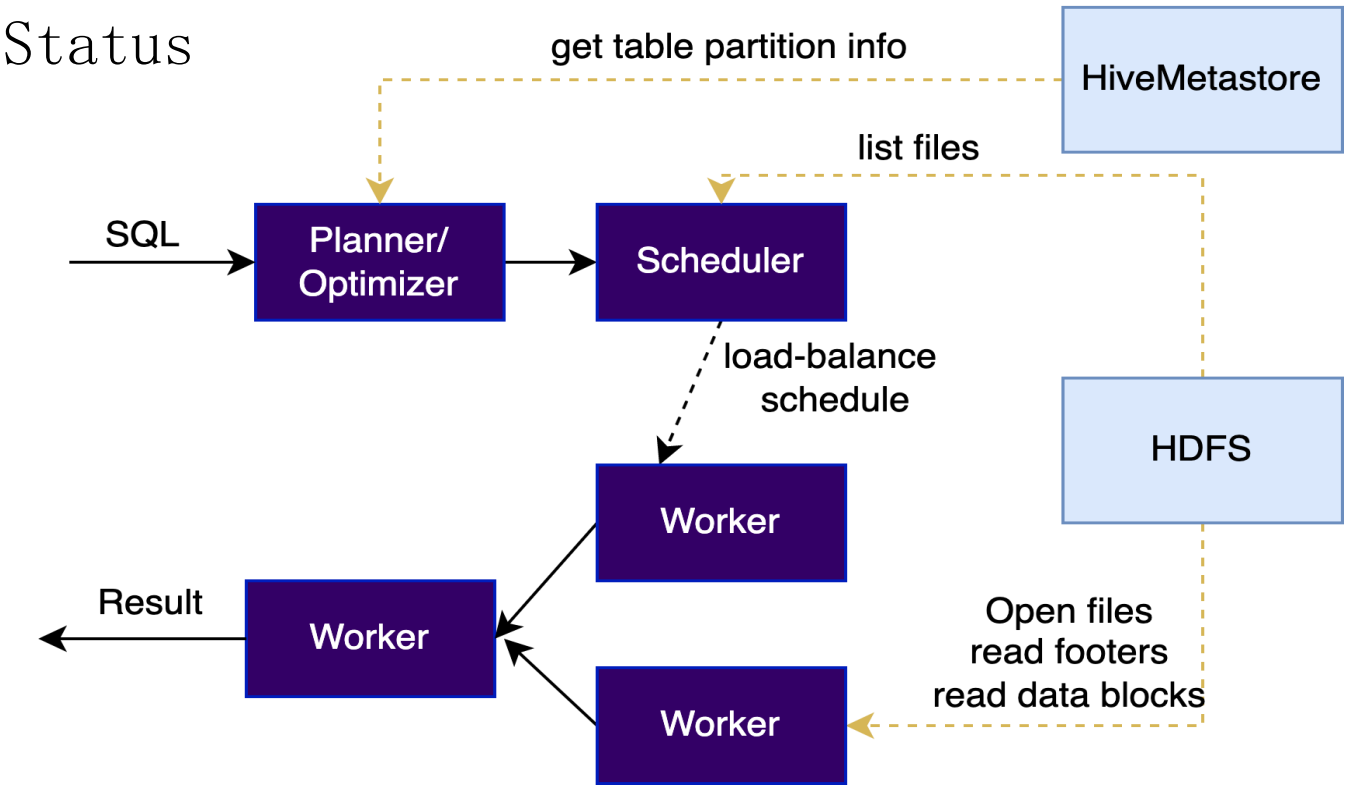
1. 开发一个标记服务，将适合缓存的分区进行标记
2. Presto在构建split时，load partition即可识别




```
Partition Parameters:
        COLUMN_STATS_ACCURATE    {\"COLUMN_STATS\":{                ":\"true\",\"build\":\"true\",\"buv
all\":\"true\",\"                  \":\"true\",\"ip\":\"true\",\ match_count_list\":\"true\",\"mid\":\"true
ue\",\"pv_stat_data . . true\",\"qtime\":\"true\",\"query\":\"true\",'                  \":\"true\",\"requ
":\"true\",\"state\":\"true\",\"trackid\":\"true\",\"url\":\"true\"}}
        STATS_GENERATED_VIA_STATS_TASK  workaround for potential lack of HIVE-12730
        cache_tag               jscs-ai-offline_tag,ai-offline_tag,ogv_tag,jscs-ai-online_tag
        numFiles                80
```

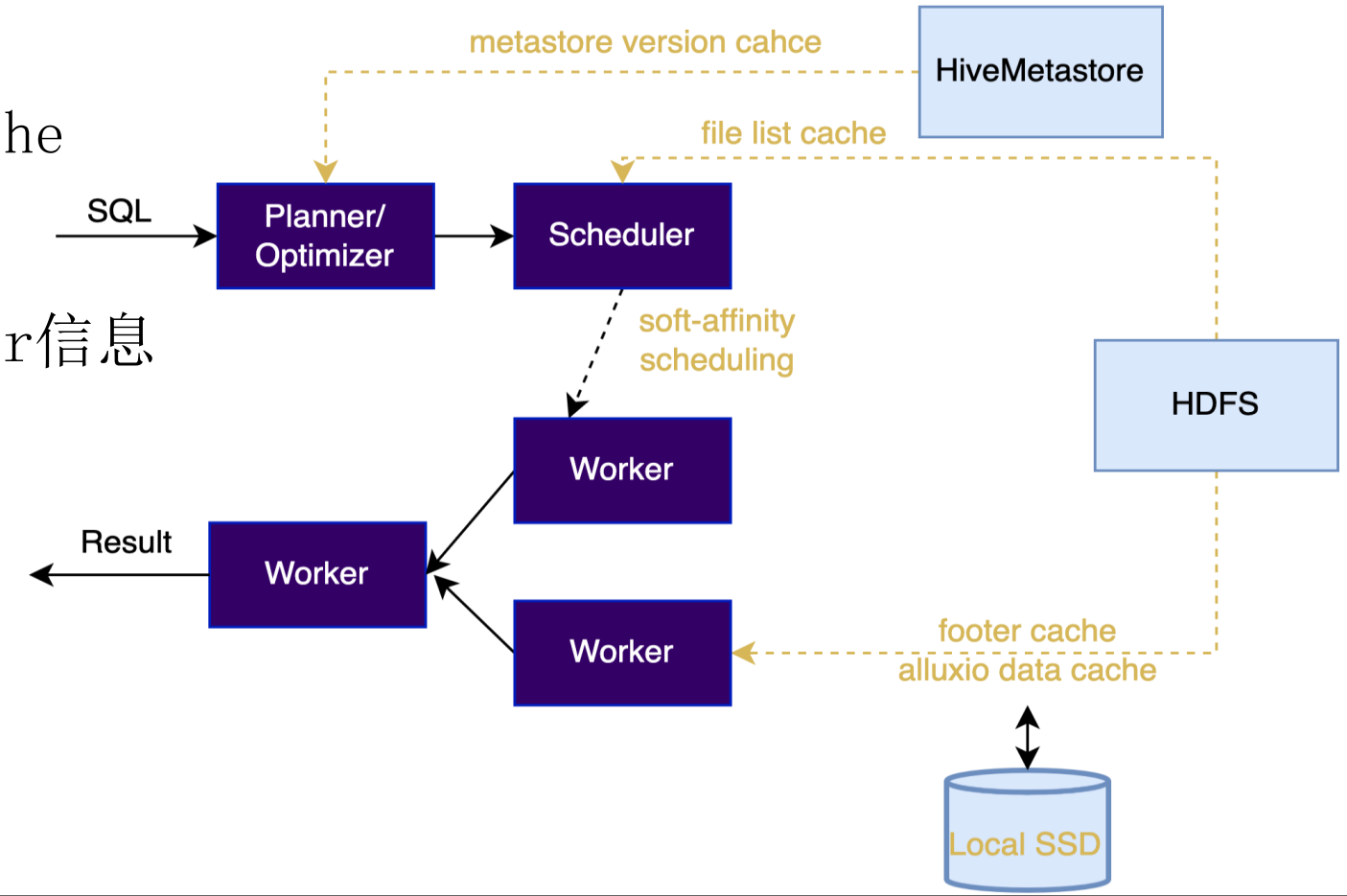
Presto Local Cache--架构

1. 强依赖HMS，且频繁请求
2. 构建split，访问NN获取FileStatus
3. 读数据，访问DN读block



Presto Local Cache--架构

1. 基于版本的元数据缓存
2. 主节点支持开启FileStatus cache
3. HDFS数据缓存到worker的ssd中
4. Worker缓存orc/parquet的footer信息
5. 主节点软亲和调度



Presto Local Cache--Soft-Affinity调度

热点问题：

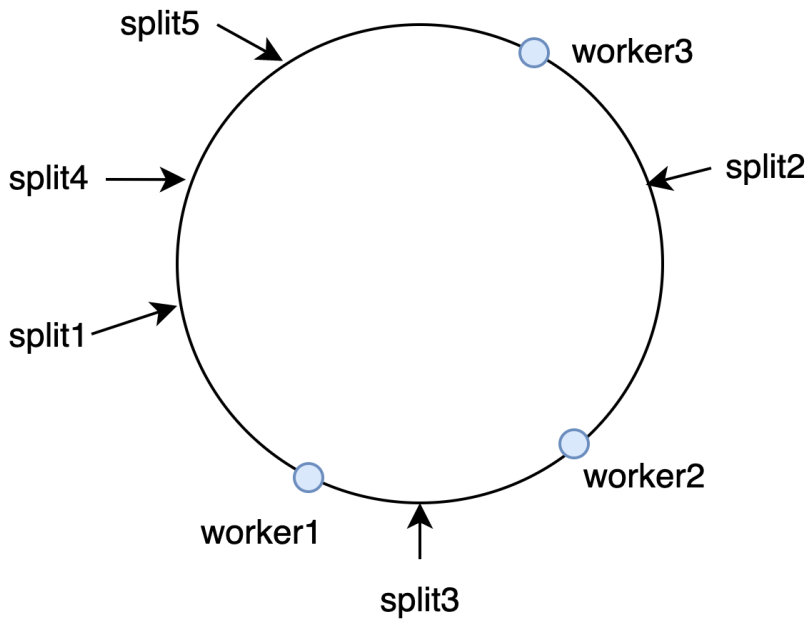
1.WorkerID1 =Hash(splitID) % workerCount
2.WorkerID2 =Hash(splitID) % workerCount + 1

Task集中问题：

Hive、iceberg、hudi：

```
public String getSoftAffinityFilePath()
{
   return path + start;
}
```

节点缩扩容问题：



Presto Local Cache--Soft-Affinity调度

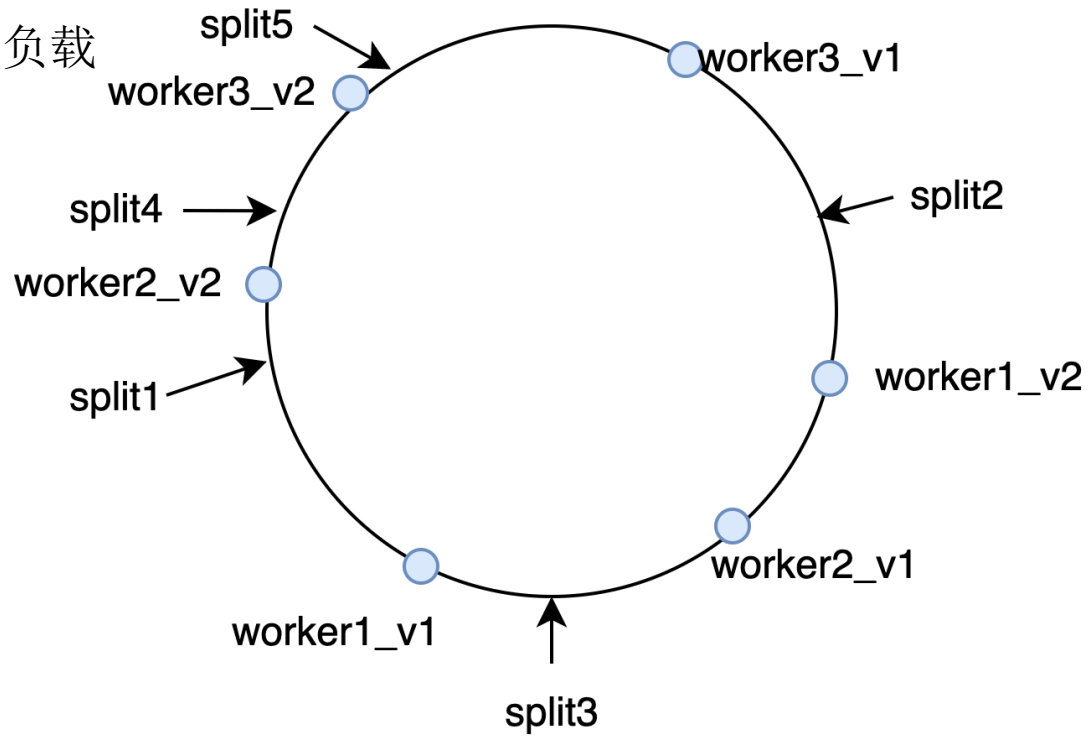
分布不均问题：

掉线一台节点，相邻节点承载掉线节点全部负载

解决：

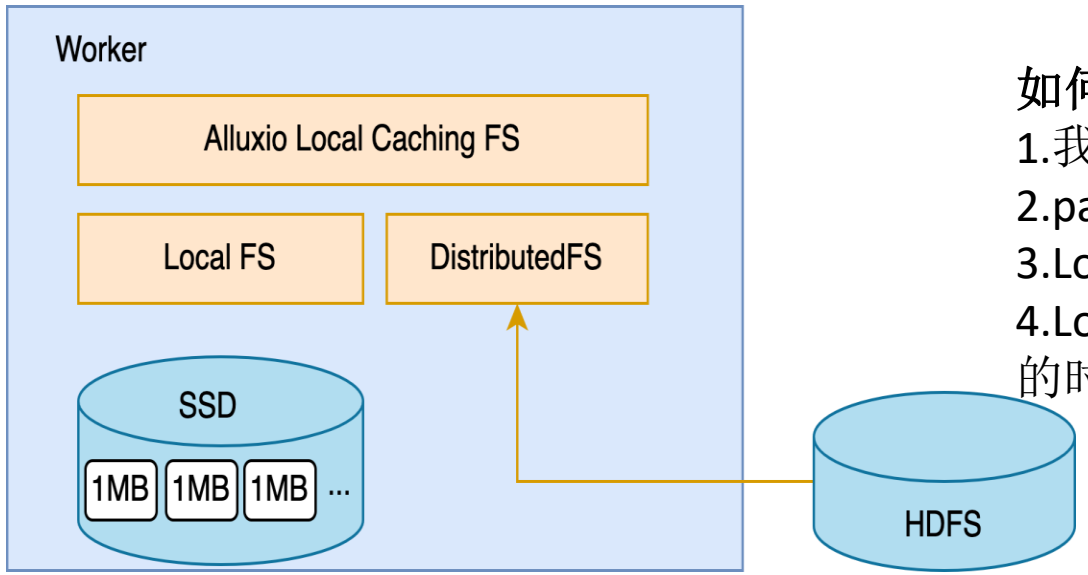
引入虚拟节点的概念

1. 更好的负载分布
2. 掉线节点负载均摊给各节点



Presto Local Cache--本地磁盘管理

1. Presto worker通过集成了AlluxioCachingFileSystem来进行热数据缓存
2. 读过来的数据被分成1MB为一个单位进行存储管理
3. 基于LRU来清理缓存块



如何解决缓存失效问题？

1.我们将split的FileModifiedTime 传递到FileContext
2.pageSource再将该时间传递给LocalFS
3.LocalFS缓存Page的时候，将time写入元 数据
4.LocalFS读文件的时候将该时间和存储的FileMeta中的时间进行对比

Presto Local Cache--改进

1. bugs

一、FileSystem提前close

```
Caused by: java.io.IOException: Filesystem closed
                at org.apache.hadoop.hdfs.DFSClient.checkOpen(DFSClient.java:482)
                at org.apache.hadoop.hdfs.DFSInputStream.close(DFSInputStream.java:723)
                at java.base/java.io.FilterInputStream.close(FilterInputStream.java:180)
                at org.apache.hadoop.util.LineReader.close(LineReader.java:164)
                at org.apache.hadoop.mapred.LineRecordReader.close(LineRecordReader.java:291)
```

FileSystemCache涉及到对象回收的一个bug

prestodb-17356

二、disabled Filesystem cache情况下，viewfs 存在bug

prestodb-17366

三、insert语句存在跨namespace问题

prestodb-17389

Presto Local Cache--改进

问题2：

CacheManager只支持单路径

社区：

通过hash&mod的方式存入多磁盘

```
private Path getRoot(PageId pageId) {
  int index = pageId.hashCode() % mRoots.size();
  index = index < 0 ? index + mRoots.size() : index;
  return mRoots.get(index);
}
```

改造：

基于AvailableSpace来做磁盘选择（借鉴HDFS）

```
基于可用空间的策略
举例：假设有五个盘，容量分别为1g、50g、25g、5g、30g，
现在需要基于该策略往某个盘写数据。
1）校验5个盘是否处于balanced
最大容量-最小容量<平衡态的阈值（默认10g）
若平衡的话，直接RoundRobin进行选择
2）划分为2列：highAvail与lowAvail
划分标准：判断是否大于（最小容量+平衡态阈值）
highAvail：50g、25g、30g
lowAvail：1g、5g
3）根据概率，选择某列进行RoundRobin
若数据大小超过lowAvail列最大值，则选择highAvail进行轮询
平衡概率值默认为0.75，
0.75选择highAvail轮询
0.25选择lowAvail轮询
```

Presto Local Cache--改进

问题3：

HMS不支持获取带版本的Partition和table

开启：
hive.partition-versioning-enabled=true

异常：
UnsupportedOperationException

改造：

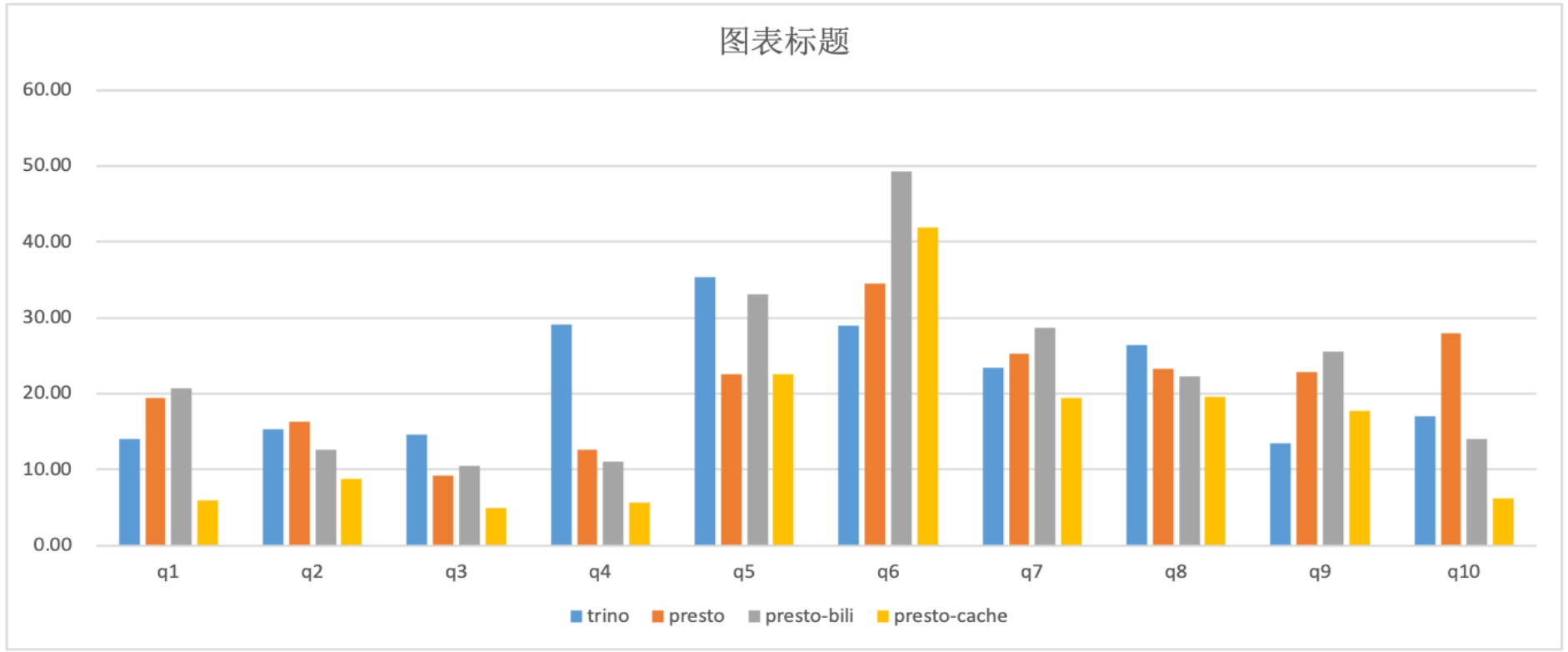
HMS基于分区的lastModifyTime新增版本API

```
@Override
public Map<String, Long> getPartitionsVersions(String databaseName, String tableName, List<String> partitionValues)
        throws TException
{
    return client.get_partitions_version(databaseName, tableName, partitionValues);
}
```

```
@Override
public long getTableVersion(String databaseName, String tableName)
        throws TException
{
    return client.get_table_version(databaseName, tableName);
}
```

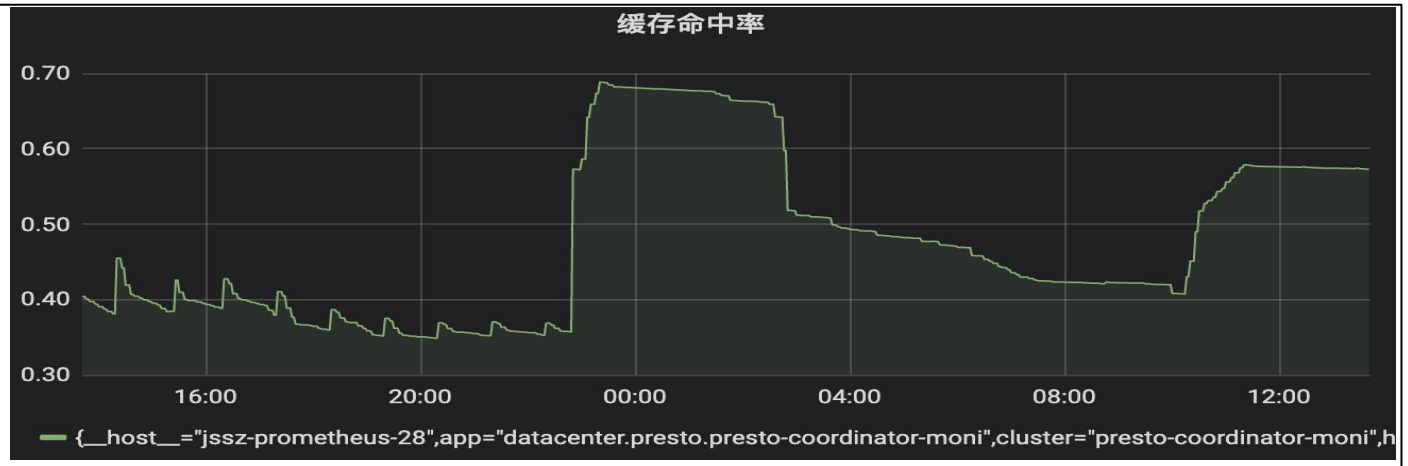
Presto Local Cache--性能对比

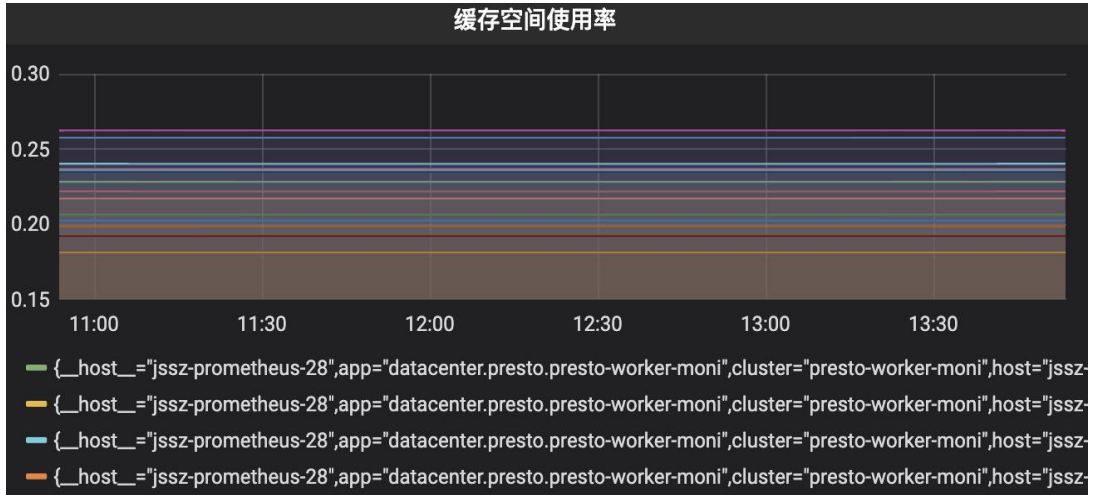
TPCH:30%～数倍性能提升

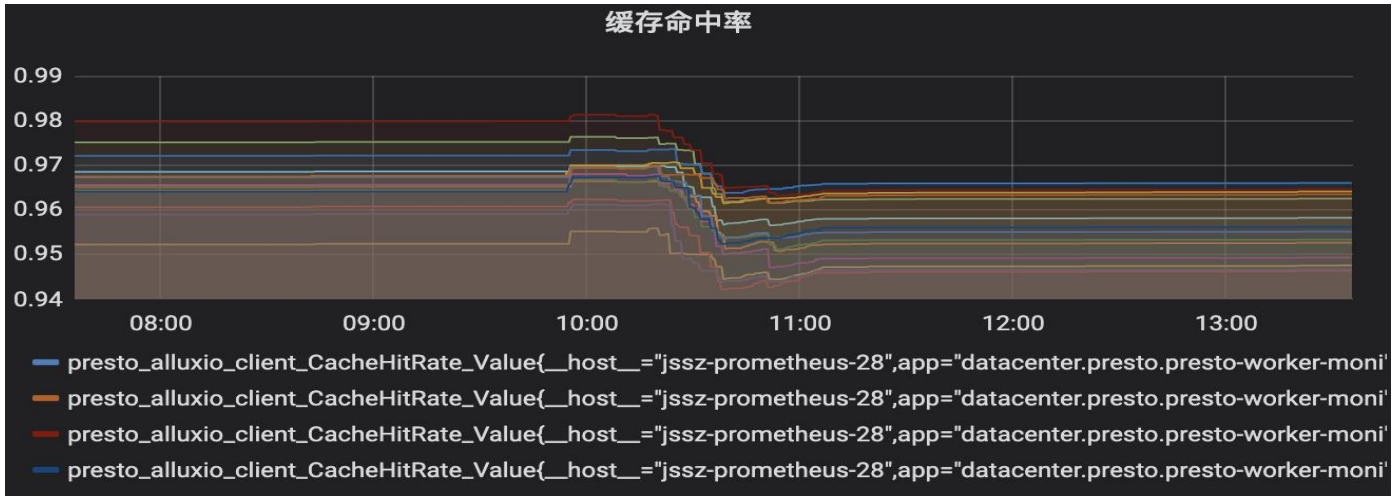


Presto Local Cache--线上情况

- 上线6个集群
- 分区走Cache命中率40%







Presto Index--背景

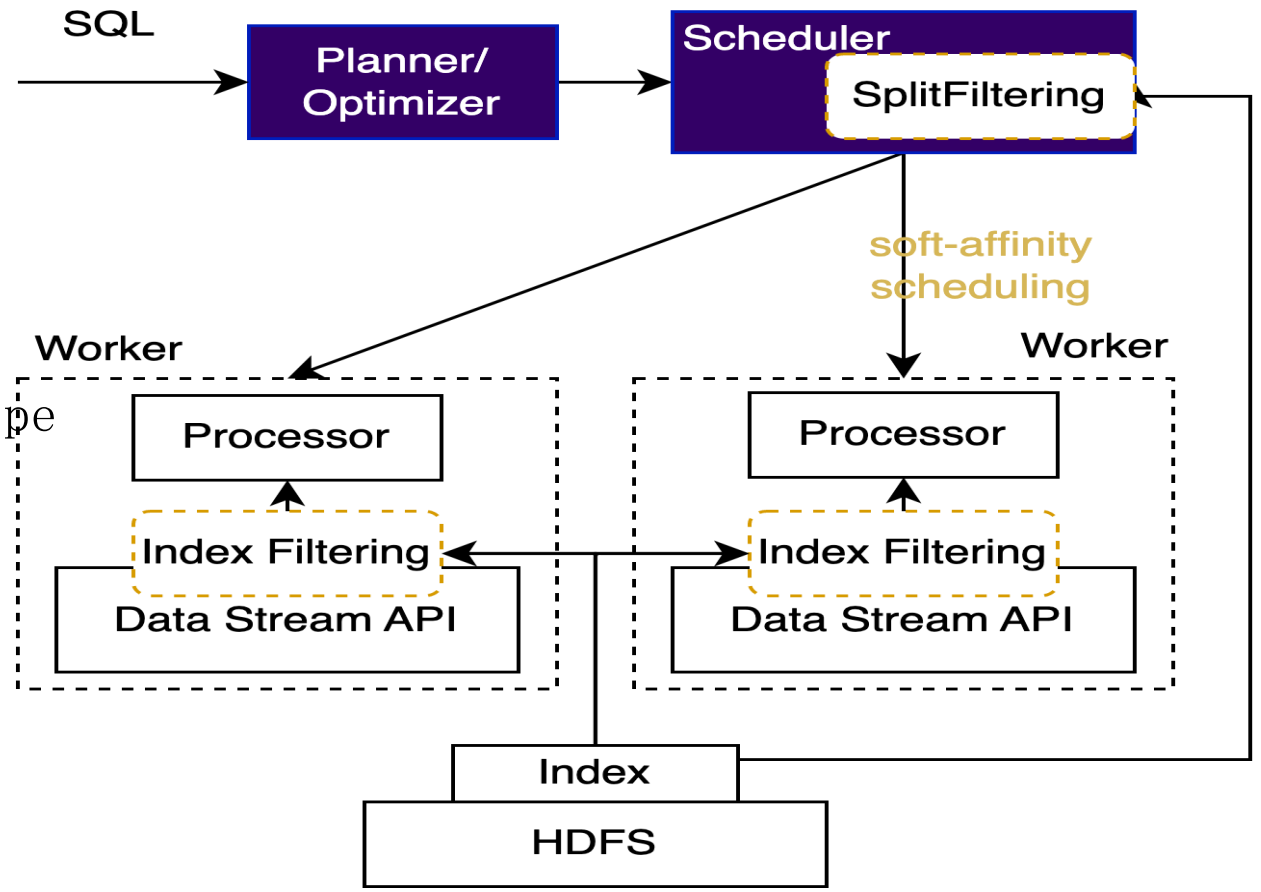
1. ORC/parquet 内嵌index少
2. 文件中的index只能在worker端读的时候进行过滤
3. 合适的index，能过显著提升查询性能

Presto Index--架构

1.实现index语法，构建index数据，并写入到Hdfs中进行持久化。

2.Coordinator侧source split调度利用index进行过滤。

3.Worker侧在读orc文件时，根据index过滤stripe或者具体的行数据。



Presto Index--indexes

1. BitMap Index

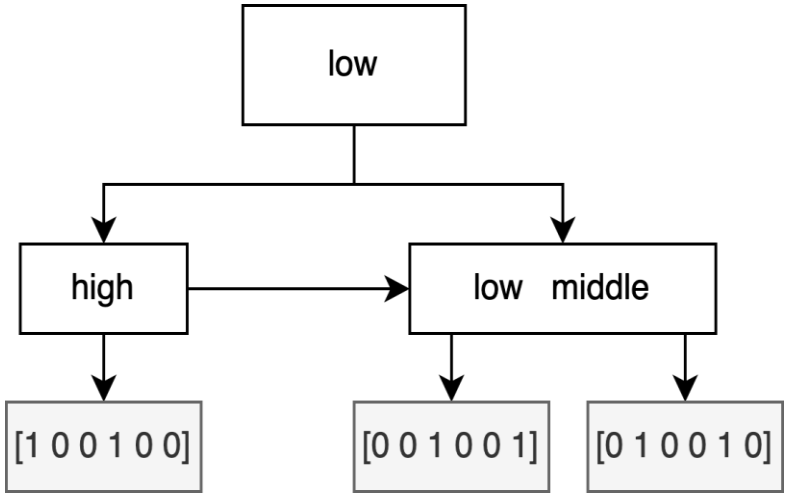
为索引字段的每个stripe构建bitmap+btree

特点：
1. 数据量大，低基数列
2. 直接读取相应行数数据
3. 支持区间查询和点查

缺点：
1. 只能在worker端过滤行
2. 不适合大基数列

/hive/db/tbl/000.orc	
name	height
bob	high
alex	middle
sam	low
peter	high
kael	middle
aux	low

	1	2	3	4	5	6
high	1	0	0	1	0	0
middle	0	1	0	0	1	0
low	0	0	1	0	0	1



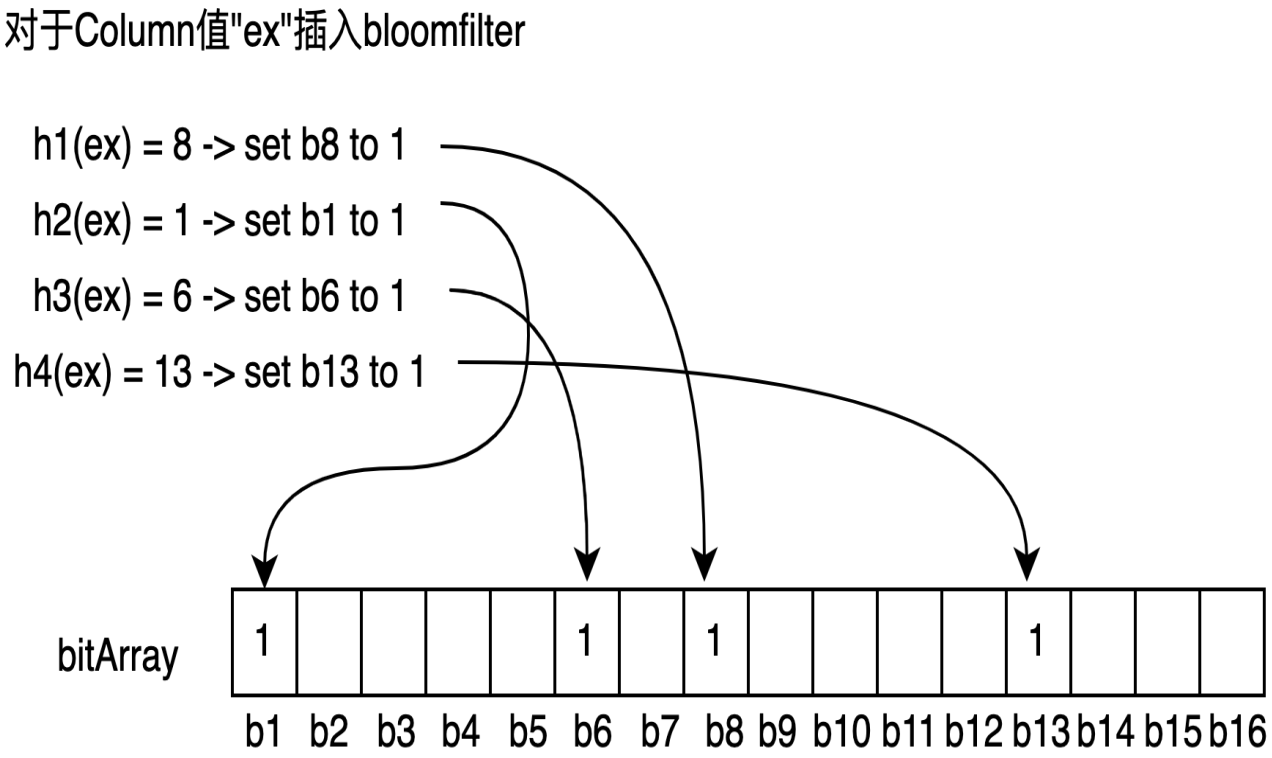
Presto Index--indexes

2.Bloomfilter Index

(1)对索引列的每个stripe维护一个Bloomfilter对象，直接将列值读入去重后插入BloomFilter。

特点：
1.相对BitMap更节省存储
2.能在coordinator过滤split
3.能够适用高基数列

缺点：
1. 只能应用等值表达式“=” “in”



Presto Index--indexes

3. Min-Max Index

记录了每个Stripe的min-max值

特点：
1. 占用空间小，
2. 适合coordinator过滤split

缺点：
1. 需要对column进行排序

Presto Index--column热度收集

获取colunn热度的时候，需要将能过下推到TableScan的表达式来统计column热度

思路:

Optimizer: PushPredicateIntoTableScan

1. 将Filter的expression 尝试将Predicate下推到TableScan

2. 如果能过下推的Predicate, 表达式存在HiveTableHandle 的effectivePredicate中

提取下推column及表达式：

使用Visitor遍历Logical Plan
在遍历到FileterNode，并且起Source为TableScan，则可以很方便的提取Column以及下推表达式

```
select * from tmp_bdp.tmp_test_index where b = 3
and a = 'bili_live' and get_json_object(c, '$.input') = 'c'
```

Res:
"effectivePredicate" :
"{\"hive:tmp_bdp:tmp_test_index.a\":{\"=\":1},\"hive:tmp_bdp:tmp_test_index.b\":{\"=\":1}}",

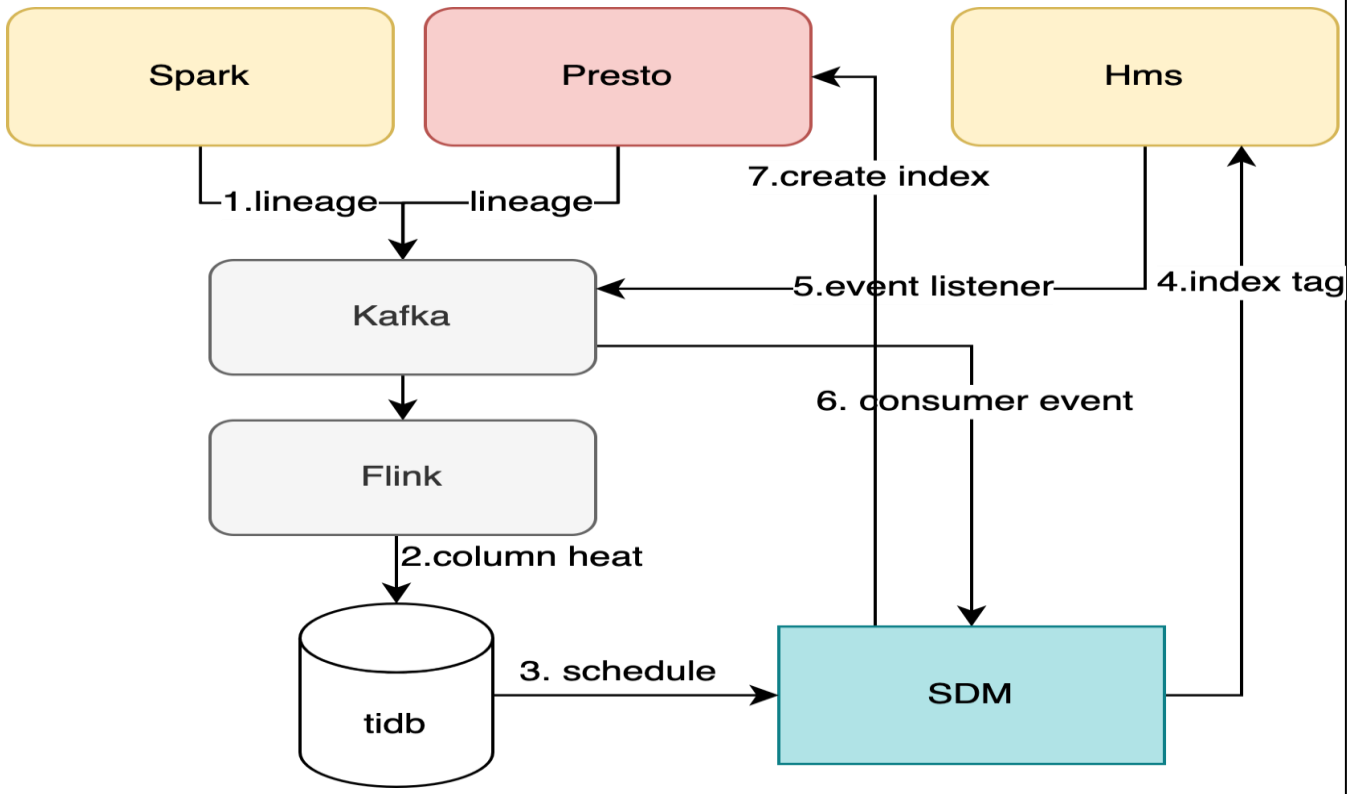
Presto Index--column热度收集

列热度统计

col_name	使用次数	<	<=	>	>=	=	!=	in	join	group by	order by
hive:r_csc:rods_s_maincenter_aegis_log01_l_min.ctime	96931	0	0	0	4895	5170	0	855	14744	73538	3352
hive:bili_dim:examine_operator_detail_new.nickname	24461	0	0	0	0	64	0	0	37	23994	62
hive:bili_dim:examine_operator_detail_new.entrytime	23657	0	3	0	4	4	0	1	0	23637	12
hive:bili_ogv:dim_season_hour_d.season_id	20035	92	0	25	0	1610	0	3036	3220	13662	0
hive:b_ods:ods_db485_work_todo_a_d.name	15715	0	0	0	0	15	0	15	1947	13747	6
hive:b_ods:ods_db485_work_todo_a_d.business_id	15484	0	0	0	0	0	0	0	1932	13552	0
hive:bili_dim:examine_operator_detail_new.area	13741	0	12	0	0	173	0	19	16	13234	318
hive:ai:recsys_avbasic_dist.avid	11089	252	49	216	333	310	0	130	5948	4230	0
hive:bili_dwd:dwd_tfc_ott_app_ubt_d.extended_fields	9444	0	0	0	0	787	0	0	787	7868	0
hive:ai:dalao_crowd_cards.id	8505	0	0	0	0	726	0	0	1787	5218	774
hive:bili_ogv:dim_season_full_d.title	6524	283	25	74	283	487	0	80	544	4344	914

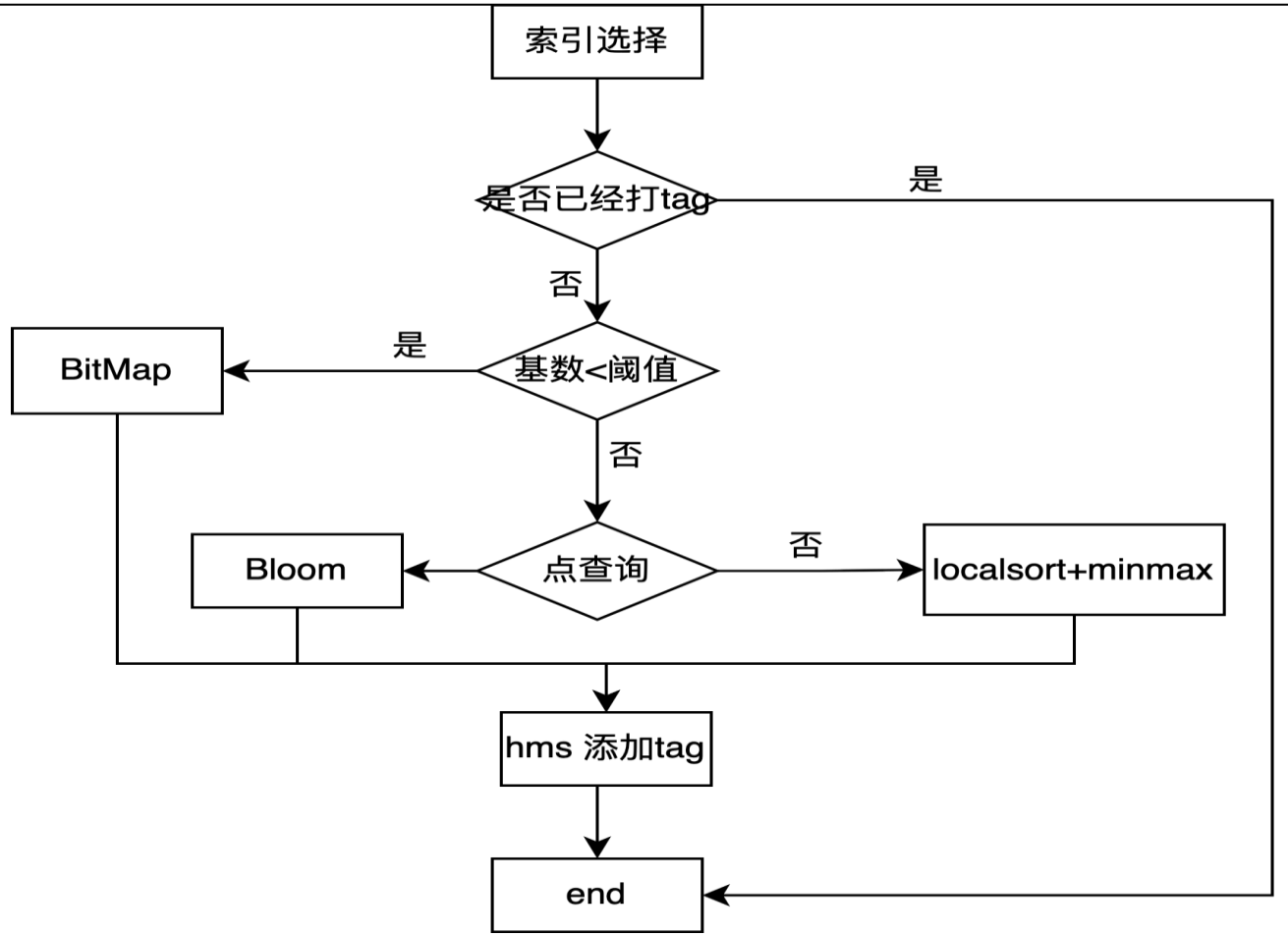
Presto Index--自动构建index

1. 根据上述column 热度信息进行排序
2. 过滤掉不合法的column信息(分区字段、非orc、复杂类型)
3. 根据决策树，在hms中响应表中打上index tag
4. 消费hms event 事件，自动构建相应index



Presto Index--如何选择index

1.低基数选择BitMap
2.点查选择Bloom
3.否则min-max,并且标记进行localsort



Presto Index--语法

创建：

```
CREATE INDEX index_name USING bloom ON hive.schema.table (column1)  WHERE p = part1;
```

删除：

```
DROP INDEX index_name (WHERE predicate);
```

更新：

```
REBUILD INDEX index_name;
```

查询：

```
SHOW INDEX (index_name);
```

Presto Index--性能对比

```
create index bili_infra_lineitem using bloom on bili_bdp.lineitem (partkey);

select sum(partkey), count(partkey) from bili_infra.lineitem  where partkey = 2724820;
```

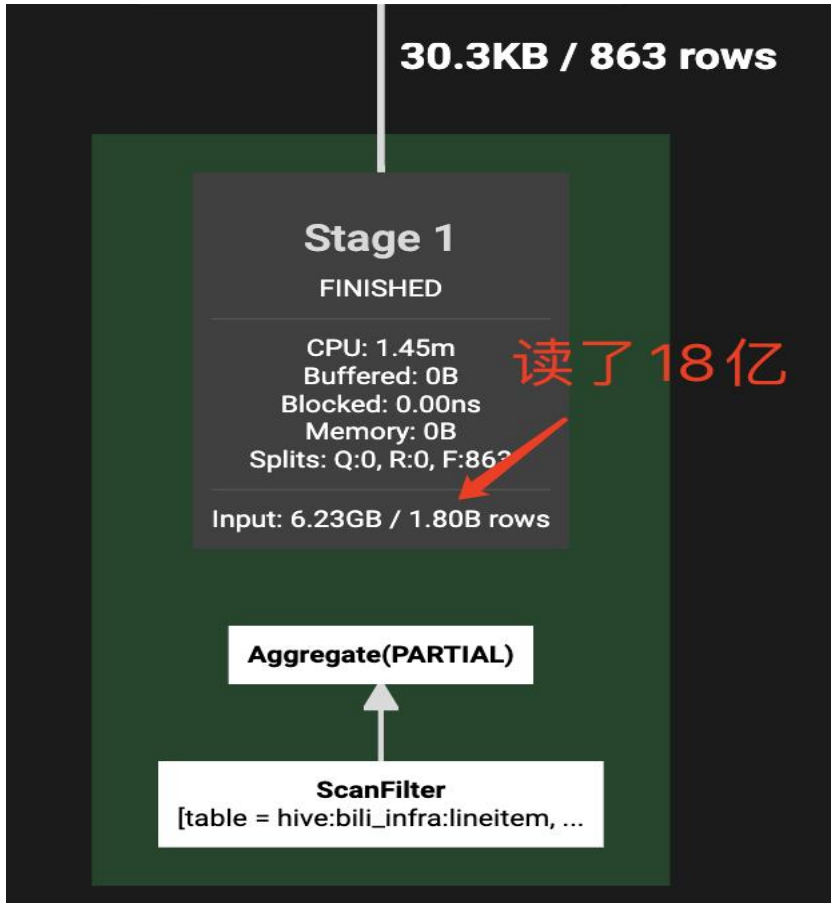
```
presto> select sum(partkey), count(partkey) from bili_infra.lineitem  where partkey = 2724820;
  _col0   | _col1
----------+-------
 21798560 |     8
(1 row)

Query 20220914_033811_00005_sxiir, FINISHED, 1 node
http://jssz-bigdata-test-05.host.bilibili.co:8282/ui/query.html?20220914_033811_00005_sxiir
Splits: 46 total, 46 done (100.00%)
CPU Time: 2.3s total, 8.62M rows/s, 30.6MB/s, 69% active
Per Node: 4.4 parallelism, 37.8M rows/s,  134MB/s
Parallelism: 4.4
Peak Memory: 0B
0:01 [20.2M rows, 71.7MB] [37.8M rows/s, 134MB/s]
```

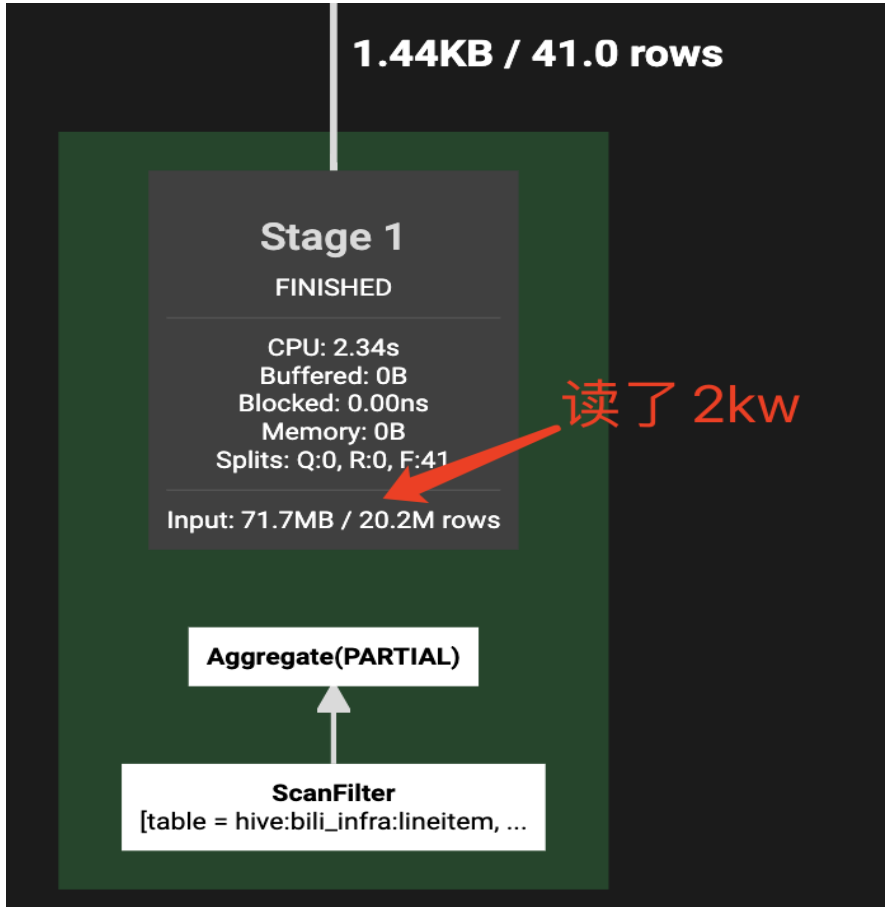
```
presto> set session index_filter_enabled=false;
SET SESSION
presto> select sum(partkey), count(partkey) from bili_infra.lineitem  where partkey = 2724820;
  _col0   | _col1
----------+-------
 21798560 |     8
(1 row)

Query 20220914_034148_00007_sxiir, FINISHED, 1 node
http://jssz-bigdata-test-05.host.bilibili.co:8282/ui/query.html?20220914_034148_00007_sxiir
Splits: 868 total, 868 done (100.00%)
CPU Time: 87.1s total, 20.7M rows/s, 73.3MB/s, 32% active
Per Node: 5.6 parallelism,  116M rows/s,  413MB/s
Parallelism: 5.6
Peak Memory: 64B
0:15 [1.8B rows, 6.23GB] [116M rows/s, 413MB/s]
```

Presto Index--性能对比



VS



未来计划

- 支持物化列的读写
- 支持物化视图提升查询性能
- Native engine探索

哔哩哔哩技术

微信扫描二维码，关注我的公众号


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

THANKS