
prestocon
DAY

prestocon DAY

Presto Pinot Data Lake Segment Reader

Mingjia Hang
Uber

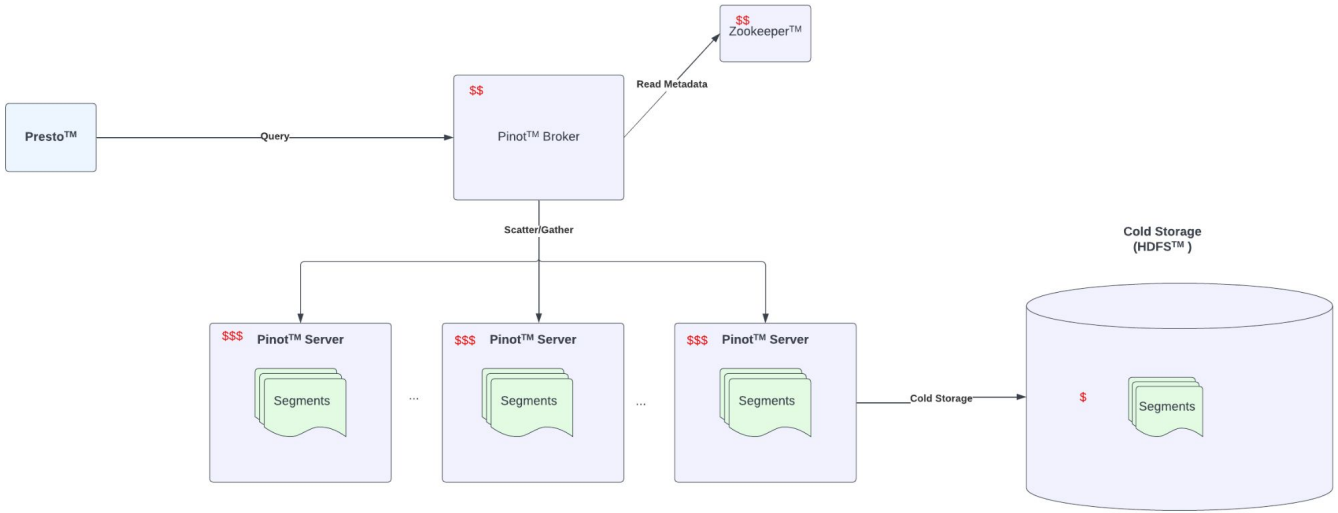
Background

Some use cases onboarded to Apache Pinot™:

- Over **6K** tables
- **7** days data retention (median)
- **~800** Hosts
- **~2 PB** data on Hosts
- **~2 PB** on HDFS (Deep Store Copies)

Customer Ask

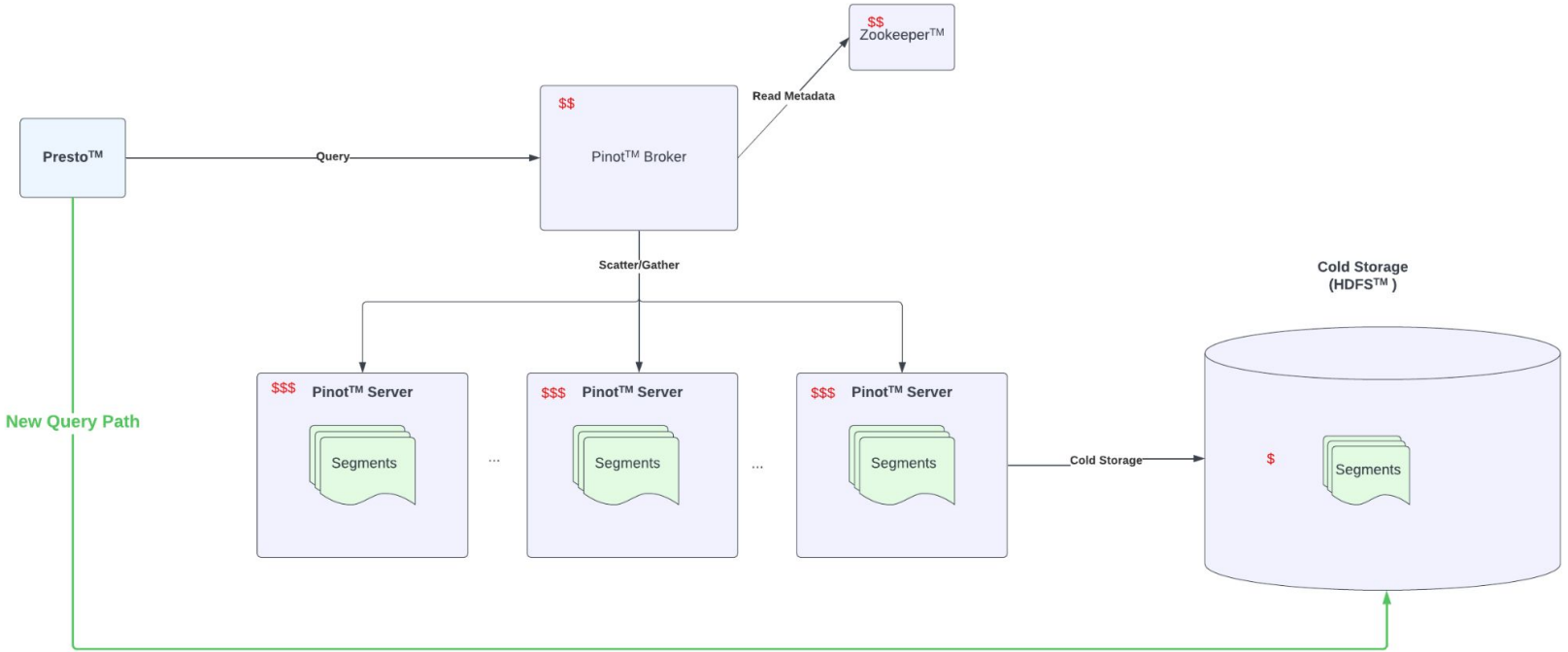
- Q1: Can we increase the retention to 30 days?
- Q2: What is the cost?



Cost Analysis

Use Case	Serving directly using Pinot™ Servers
Log search (30 days)	$30x

Alternative Path (for historical queries)



Cost Analysis

Use Case	Serving directly using Pinot™ Servers
Log search (30 days)	$30x

Benefits

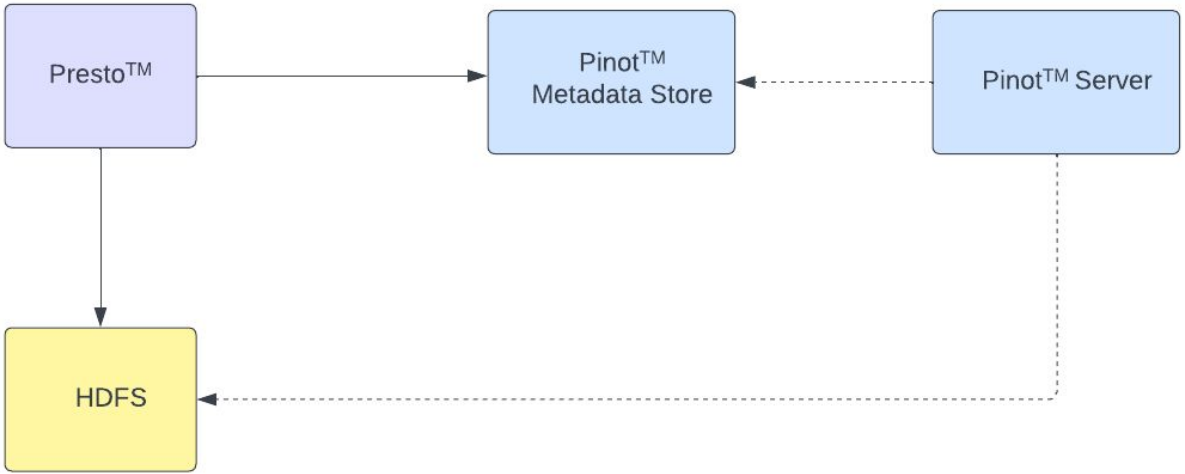
Cost Saving

Support complex queries (join, udf)

No duplicate copied needed

Overall Design Diagram

Catalog:
pinotdatalake



Metadata Store

- Use Pinot™ Controller as the metadata store
- Endpoints
  - /tables/{tableName}/shema
    - Return table schema

```
{
 “schemaName”: “tableName”,
 “dimensionFieldSpecs”: [
  {
   “name”: "example Column",
   "dataType": "STRING"
  }
 ]
}
```

Metadata Store

- **/segments/{tableName}**
  - Return all segments for specific table

```
[{
   "REALTIME": [
     "segmentFile1",
     "segmentFile2"]
 }]
```

- **/segments/{tableName}/{segmentName}/metadata**
  - Return deep store location for this table's segment

```
{
  "segment.start.time":"170425133058",
  ...
  "segment.download.url": "hdfs://path/tableName/fileName"
}
```

Alternative Metadata Store

Alternative Solutions	Pros	Cons
Hive™ Metadata Store(HMS)	1. Assured Scalability	1. Unnecessary Dependency on HMS 2. Compatibility Evaluation with Pinot Segments 3. Duplicate Metadata Store
Redis™*	1. Seamless Integration with Presto 2. Scalability Assurance 3. Redis Versatility	1. Data Structure Design Challenge
Internal Tool	1. Already stores table and cluster level metadata 2. Easy integration	1. Segment level data not available (need to access Pinot controller)

Segment Read

Download segment file to Presto disk

Decompress the file

Read the file row by row

Delete file

Support Time Range Filtering

Segment Size - one medium size table, 6K+ segments, 800MB+ for each segment(7 days) ~~ 20TB for 30 days data

Segment Name Example: tablename__0__16__20240506T0028Z

Segment Metadata Example: _ingestionEpochMs under timeFieldSpec

```
"timeFieldSpec": {
  "incomingGranularitySpec": {
    "name": "_ingestionEpochMs",
    "dataType": "LONG",
    "timeType": "MILLISECONDS"
  }
}
```

Support Time Range Filtering

Force user to add time range filtering: *select runtime_env,level from pinotdatalake.tableName where _ingestionEpochMs between TIMESTAMP '2024-05-15 23:50:00' and TIMESTAMP '2024-05-15 23:59:59'*

ErrorType is USER_ERROR

ErrorName is MISSING_DATE_FILTERING

ErrorCode is 84213762

Exception is com.facebook.presto.spi.PrestoException

Message is (host: phx5-wua) Query is running without required filter for table 'logging_scribe_composer_for_kibana'. Please include _ingestionEpochMs as filter in WHERE clause.

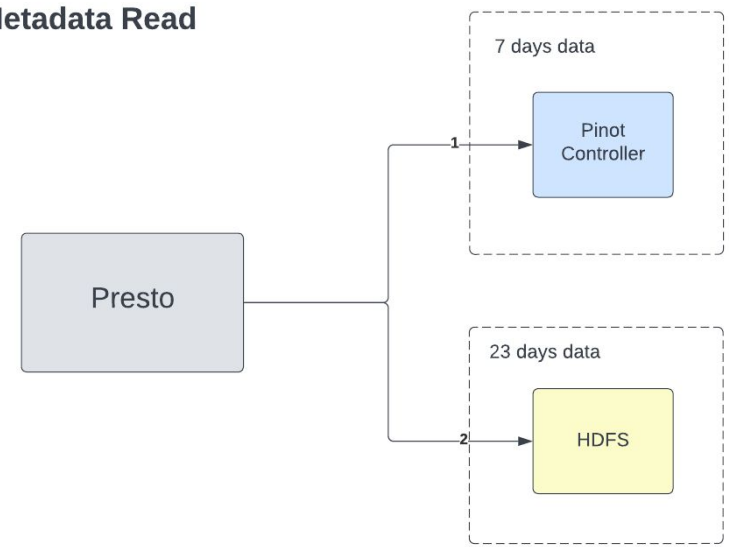
Support 30 days data reading

```
{
 "REALTIME": {
  "tableName": "tableName",
  "segmentsConfig": {
   "deletedSegmentsRetentionPeriod" : "30d",
   "retentionTimeUnit": "DAYS",
   "retentionTimeValue" : "7"
   "schemaName": "schemaName"
  }
 }
}
```

Support 30 days data reading

Segment Name Change:
'tableName__1__937__20240514T2304Z' to
'**tableName__1__937__20240514T2304Z__**RETENTION_UNTIL__202405220120**'

Location Change:
'hdfs://path/databaseName**/tableName/**' →
'hdfs://path/databaseName**/Deleted_Segments/tableName/**'.



Next Phase

Caching - Alluxio (segment is immutable)

Optimization - Min/Max column value

Ingestion - convert segment to Parquet format

Pinot Team

Ting Chen, Sr Staff engineer

Caner Balci, Sr engineer

prestocon DAY

Q&A

prestocon DAY

Thank you!

Mingjia Hang
Uber