

Presto Iceberg Connector

Chunxu Tang, Twitter

March 2021



Roadmap

1

Presto at Twitter

2

Presto Iceberg Connector





Presto at Twitter

A long path pursuing for scalability and availability

Presto at Twitter

9+ Clusters (on-prem/on-GCP)

3000+ Presto Workers

16+ Million Queries



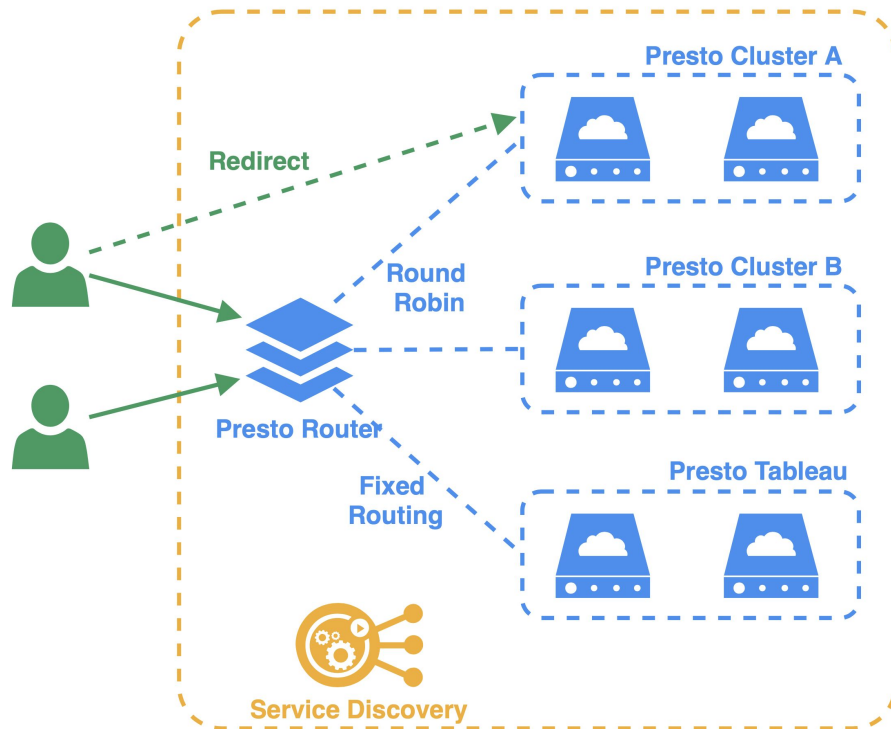
Presto Federation System

Scalability

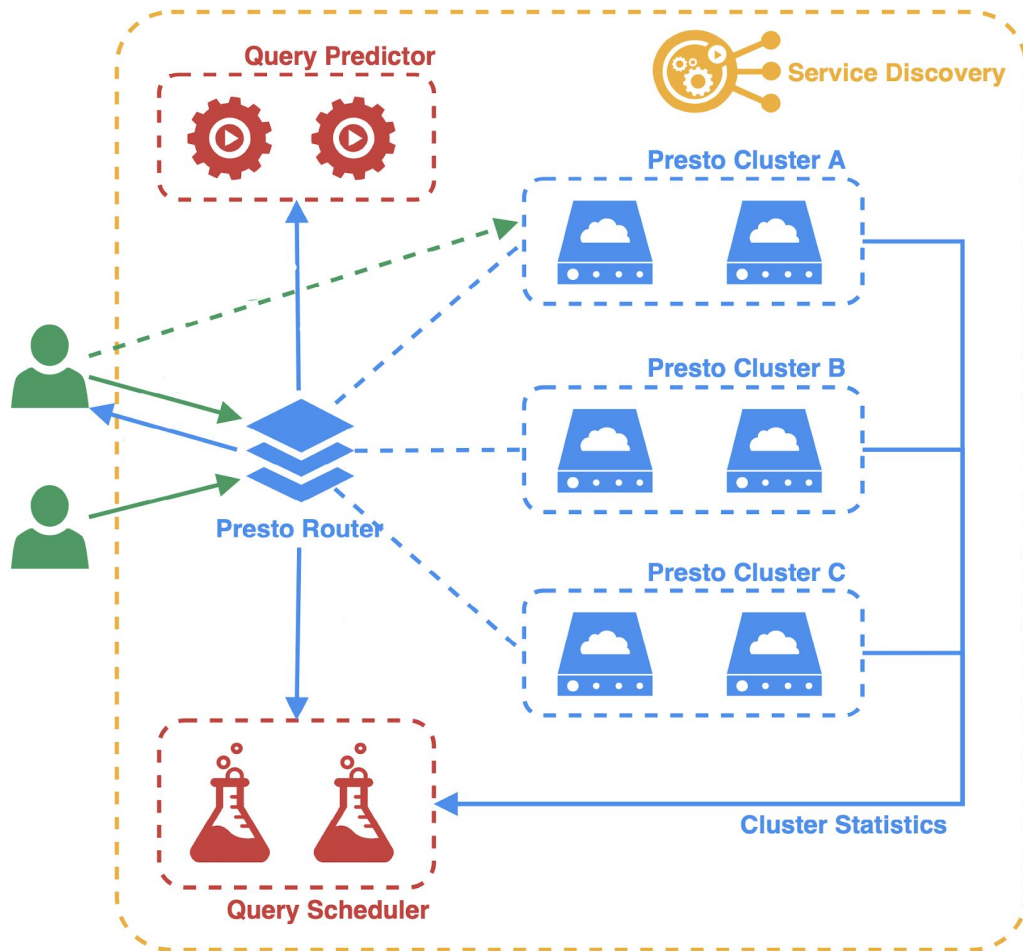
- ❑ Easily scaled
- ❑ Waiting queue splitted to sub-clusters
- ❑ Aggregated UI and APIs

Availability

- ❑ Rolling deployment

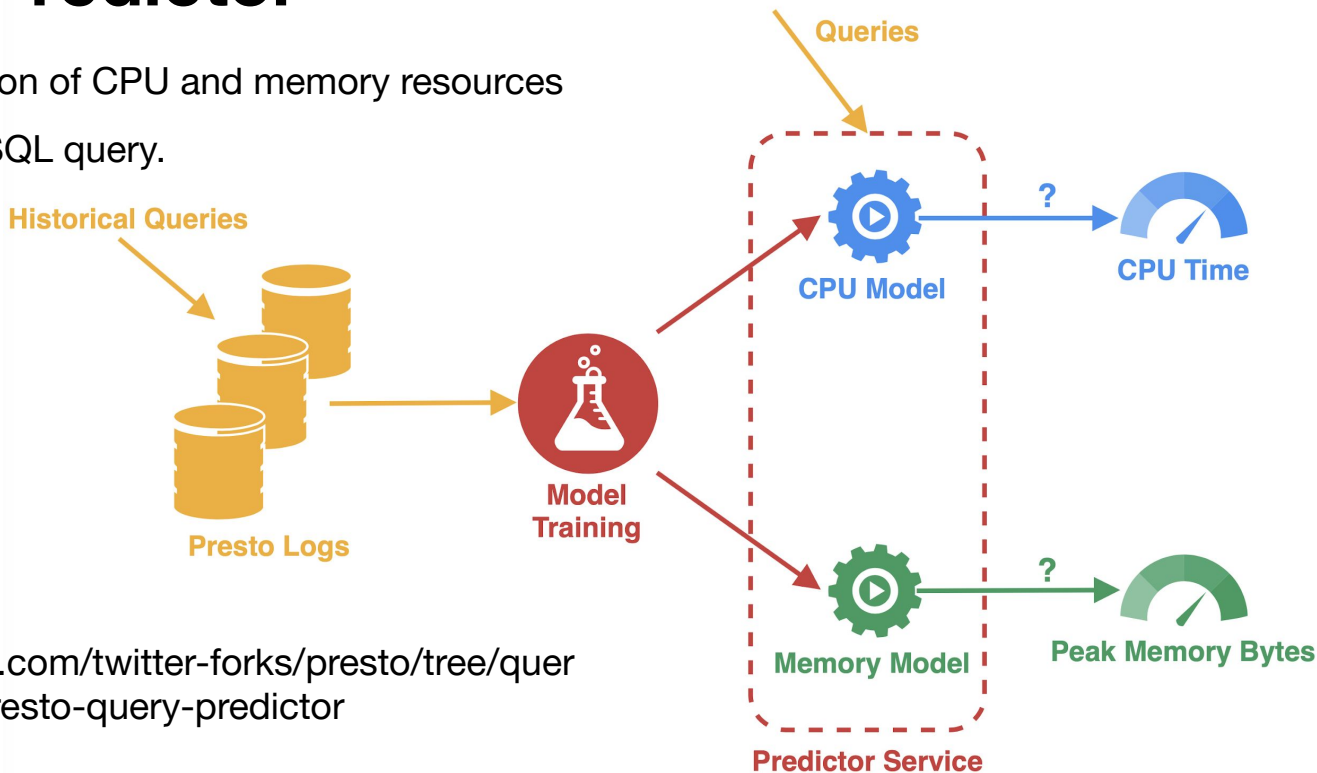


Intelligent Presto Cluster Management



Query Predictor

Quick estimation of CPU and memory resources needed for a SQL query.



<https://github.com/twitter-forks/presto/tree/query-predictor/presto-query-predictor>



Query Predictor Web App

Try a SQL statement below and predict it!

```
SELECT
  l_returnflag,
  l_linestatus,
  sum(l_quantity) as sum_qty,
  sum(l_extendedprice) as sum_base_price,
  sum(l_extendedprice * (1 - l_discount)) as sum_disc_price,
  sum(l_extendedprice * (1 - l_discount) * (1 + l_tax)) as sum_charge,
  avg(l_quantity) as avg_qty,
  avg(l_extendedprice) as avg_price,
  avg(l_discount) as avg_disc,
```

An example query from
TPC-H benchmark

Predict

Predicted CPU time: 30s - 1h

Predicted peak memory bytes: 1MB - 1TB





Presto Iceberg Connector

Empowering SQL queries on the open table format

Apache Iceberg

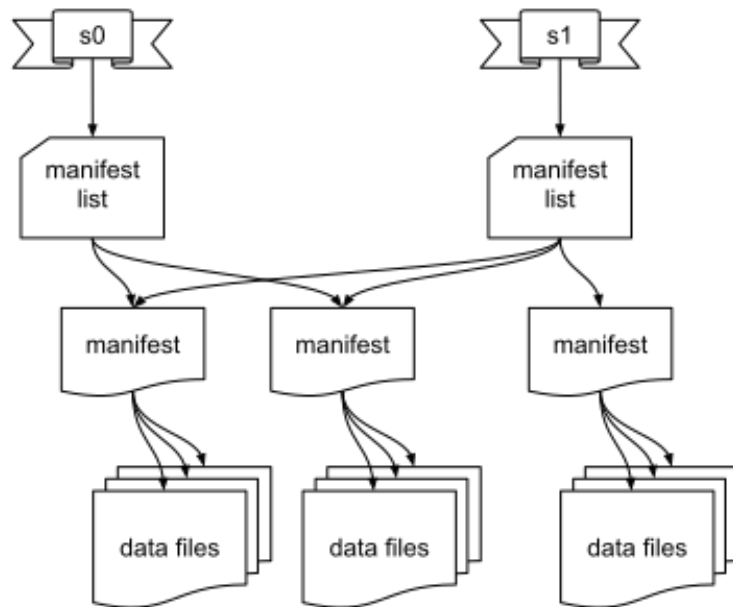
An open table format for huge analytic datasets

- ❑ Schema evolution
- ❑ Hidden partitioning
- ❑ Partition layout evolution
- ❑ Time travel
- ❑ Version rollback



Format

- ❑ Snapshot
- ❑ Manifest list
- ❑ Manifest file
- ❑ Partition spec
- ❑ Partition tuple
- ❑ Snapshot log



Example - TPCB

```
{
  "manifest_path" : "/var/folders/wg/f6prq0_
  "manifest_length" : 4746,
  "partition_spec_id" : 0,
  "added_snapshot_id" : {
    "long" : 2525750933897995572
  },
  "added_data_files_count" : {
    "int" : 1
  },
  "existing_data_files_count" : {
    "int" : 0
  },
  "deleted_data_files_count" : {
    "int" : 0
  },
  "partitions" : {
    "array" : [ ]
  },
  "added_rows_count" : {
    "long" : 5
  },
}
```

Manifest list (Avro)

```
{
  "status" : 1,
  "snapshot_id" : {
    "long" : 2525750933897995572
  },
  "data_file" : {
    "file_path" : "/var/folders/wg/f6prq0_
    "file_format" : "PARQUET",
    "partition" : { },
    "record_count" : 5,
    "file_size_in_bytes" : 805,
    "block_size_in_bytes" : 67108864,
    "column_sizes" : {
      "array" : [ {
        "key" : 1,
        "value" : 52
      }, {
        "key" : 2,
        "value" : 101
      }, {
        "key" : 3,
        "value" : 254
      } ]
    }
  },
}
```

Manifest (Avro)

```
{
  "format-version" : 1,
  "table-uuid" : "0a258e15-b687-42f8-9b32-b444
  "location" : "/var/folders/wg/f6prq0_s0w7cdh
  "last-updated-ms" : 1615779313737,
  "last-column-id" : 3,
  "schema" : {
    "type" : "struct",
    "fields" : [ {
      "id" : 1,
      "name" : "regionkey",
      "required" : false,
      "type" : "long"
    }, {
      "id" : 2,
      "name" : "name",
      "required" : false,
      "type" : "string"
    }, {
      "id" : 3,
      "name" : "comment",
      "required" : false,
      "type" : "string"
    } ]
  },
  "partition-spec" : [ ],
  "default-spec-id" : 0,
  "partition-specs" : [ {
    "spec-id" : 0,
    "fields" : [ ]
  } ],
}
```

Table metadata (JSON)



Example - TPCCH

```
{
  "manifest_path" : "/var/folders/wg/f6prq0",
  "manifest_length" : 4746,
  "partition_spec_id" : 0,
  "added_snapshot_id" : {
    "long" : 2525750933897995572
  },
  "added_data_files_count" : {
    "int" : 1
  },
  "existing_data_files_count" : {
    "int" : 0
  },
  "deleted_data_files_count" : {
    "int" : 0
  },
  "partitions" : {
    "array" : [ ]
  },
  "added_rows_count" : {
    "long" : 5
  },
}
```

Manifest list (Avro)

```
{
  "status" : 1,
  "snapshot_id" : {
    "long" : 2525750933897995572
  },
  "data_file" : {
    "file_path" : "/var/folders/wg/f6prq0",
    "file_format" : "PARQUET",
    "partition" : { },
    "record_count" : 5,
    "file_size_in_bytes" : 805,
    "block_size_in_bytes" : 67108864,
    "column_sizes" : {
      "array" : [ {
        "key" : 1,
        "value" : 52
      }, {
        "key" : 2,
        "value" : 101
      }, {
        "key" : 3,
        "value" : 254
      } ]
    }
  },
}
```

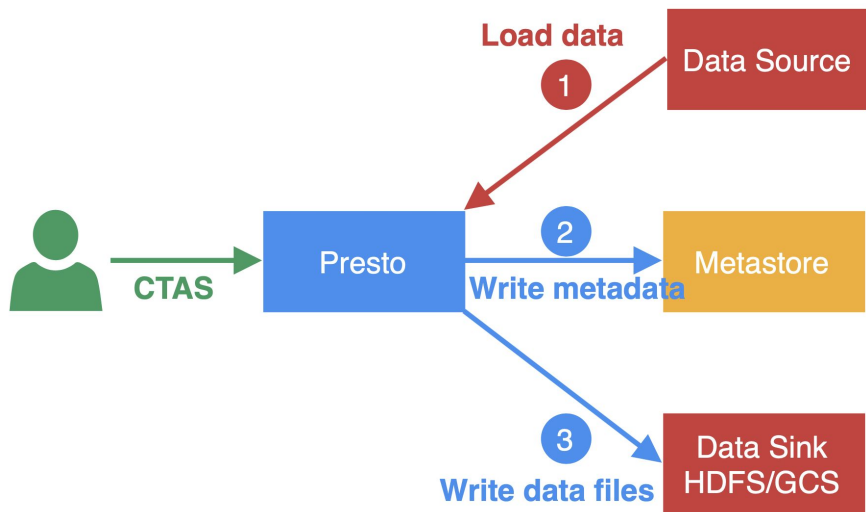
Manifest (Avro)

```
{
  "format-version" : 1,
  "table-uuid" : "0a258e15-b687-42f8-9b32-b444",
  "location" : "/var/folders/wg/f6prq0_s0w7cdh",
  "last-updated-ms" : 1615779313737,
  "last-column-id" : 3,
  "schema" : {
    "type" : "struct",
    "fields" : [ {
      "id" : 1,
      "name" : "regionkey",
      "required" : false,
      "type" : "long"
    }, {
      "id" : 2,
      "name" : "name",
      "required" : false,
      "type" : "string"
    }, {
      "id" : 3,
      "name" : "comment",
      "required" : false,
      "type" : "string"
    } ]
  },
  "partition-spec" : [ ],
  "default-spec-id" : 0,
  "partition-specs" : [ {
    "spec-id" : 0,
    "fields" : [ ]
  } ],
}
```

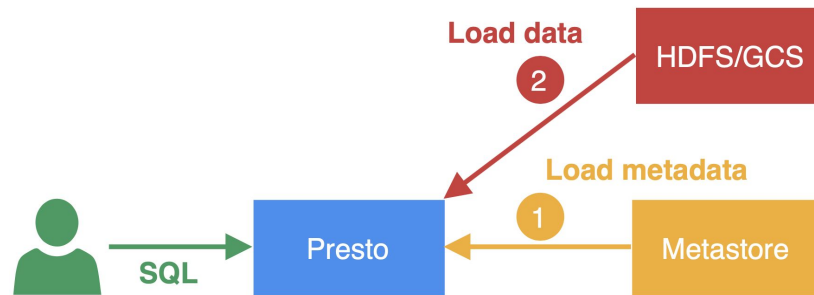
Table metadata (JSON)



Design



Write path



Read path



 Head

default ▼

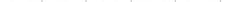
FINISHED     

```
create table if not exists ctas_nation as (select * from nation);
create table if not exists ctas_customer as (select * from customer);

select ctas_customer.name, ctas_nation.name
from ctas_customer
inner join ctas_nation on ctas_customer.nationkey=ctas_nation.nationkey
limit 20;
```

Query executed successfully. Affected rows : 25

Query executed successfully. Affected rows : 1500


[settings](#)

name	name
Customer#000001460	SAUDI ARABIA
Customer#000001489	INDONESIA
Customer#000001490	SAUDI ARABIA
Customer#000001495	IRAN
Customer#000001028	IRAN
Customer#000001035	IRAN
Customer#000001039	IRAN
Customer#000001180	JORDAN

1 to 20 of 20 |< < Page 1 of 1 > >|

Future Work

- ❑ ORC support
- ❑ Snapshot support
- ❑ Schema evolution
- ❑ Time travel
- ❑ Version rollback



Ready to help

Slack @Chunxu

