

Extracting SQL Query Logs from Ingestion

This provides a mechanism for

1. Having ingestion dump SQL query logs to a sqlite file
2. Extracting the sqlite file into a human-readable JSON file (whose format is the same as our SQL queries source)

This is useful for debugging ingestion sources, testing the SQL parser, and analyzing the query logs.

Limitations

- Requires that the ingestion source uses `SqlParsingAggregator`
- The query dumps can be large - in one instance, ~12k queries produced a 128MB sqlite db file.

Process

In the ingestion config, add extra environment variables to the config (`TMPDIR` variable can be adjusted, but needs to be defined explicitly):

Python

```
{"DATAHUB_SQL_AGG_QUERY_LOG": "STORE_ALL", "TMPDIR": "/tmp"}
```

Extra Environment Variables

Advanced: Set extra environment variables to an ingestion execution

```
{"DATAHUB_SQL_AGG_QUERY_LOG": "STORE_ALL", "TMPDIR": "/tmp"}
```

When the ingestion run is over, the report will contain a path to a file in the `sql_aggregator.query_log_path` field

```
"stateful_usage_ingestion_enabled": true,  
"num_unresolved_temp_columns": 0,  
"sql_aggregator": {  
  "query_log_path": "/tmp/tmp3edjnib4/query_log.db",  
  "num_observed_queries": 13927,  
  "num_observed_queries_failed": 325,  
  "num_observed_queries_column_timeout": 0,  
  "num_observed_queries_column_failed": 12,
```

Copy that file to your local machine. See this useful [stackoverflow answer](#)

Option 1: From ECS to local

Python

```
aws ecs execute-command \  
  --region <your-region> \  
  --cluster <your-cluster> \  
  --task <your-task-id> \  
  --container <your-container> \  
  --command "cat <query-log-path>" \  
  --interactive >> <desired-local-path>
```

Option 2: From ECS to S3 to local (possibly better for larger files)

Python

copy from ecs to s3

```
aws ecs execute-command \  
  --region <your-region> \  
  --cluster <your-cluster> \  
  --task <your-task-id> \  
  --container <your-container> \  
  --command "aws s3 cp <query-log-path>  
s3://your-bucket-name/desired-s3-path" \ --interactive
```

copy to local

```
aws s3 cp s3://your-bucket-name/desired-s3-path  
~/desired-local-path
```

Pass the above file to your Acryl representative.

=====

The below is unnecessary unless you want to do your own analysis

Using the datahub CLI, extract a readable JSON file from the query dump

Python

```
datahub check extract-sql-agg-log query_log.db --output  
queries.json
```

The final queries.json file will look something like this

None

```
[  
  {  
    "query": "insert into mytable ...",  
    "session_id": "1073856819",  
    "timestamp": "2024-03-12 08:57:53.612133",  
    "user": null,  
    "default_db": "dev",  
    "default_schema": "public"  
  },  
  {  
    "query": "..."  
    ...  
  }  
]
```