

# SQL Query Log Analysis using Python

Anish Kumar

CSCI-29 Final Project Fall 2020



# Background

- Millions of Query in Database Logging System (Teradata)
- DBAs are busy in fixing performance issue
- Existing day to day issue with SQL joins etc.
- New features (including AI/ML) are available in database
- Performance concern of using these feature (if not done properly)
- Hard to do analysis of Queries Logs by SQL select etc.
- Whether Customer using OR not using some advance feature/function of database



# Project Goal

- Extract SQL Logs from Database to Dev environment
- Define search keywords (SQL keyword or advance database functions)
- Perform Text Analytics of SQL Texts from the logs
- Utilize Word2Vec framework and find matches/nearest neighbor in SQL Texts export
- Find top 3 match and create a summary file (csv)
- Upload the summary file to database for daily review through dashboard/reports by DBA

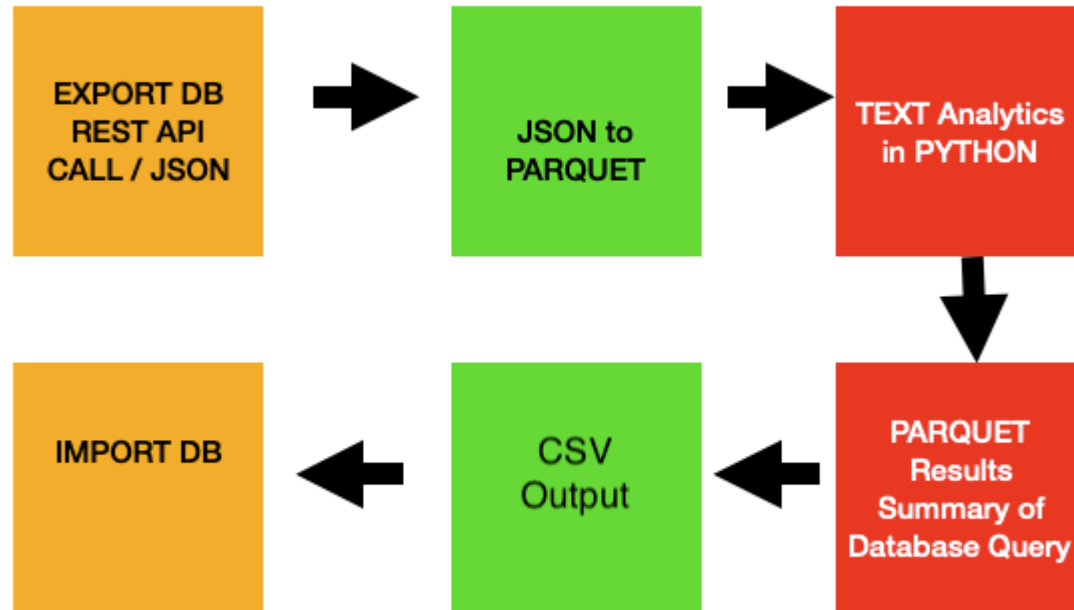


# Python Framework Utilized

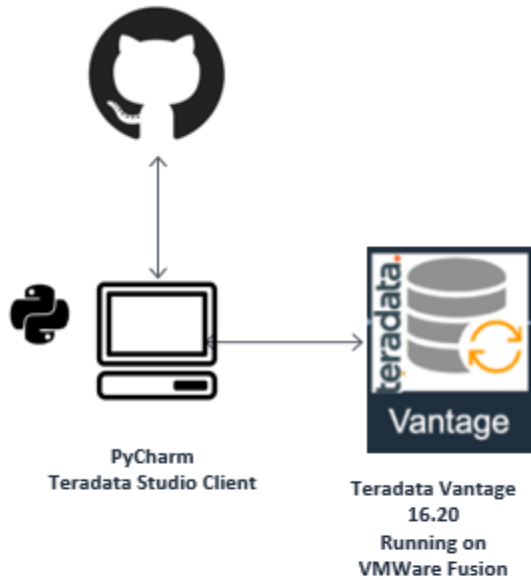
- Python is chosen since it contains many library to perform text analytics
- **fastapi, pydantic, uvicorn, teradatasql** : Utilized for exporting database SQL Log data in JSON from database using SQL Query
- **luigi, requests**: Utilized to build a wrapper around fast api / rest and produce JSON file
- **json2parquet**: To convert JSON to Parquet format
- **pandas, numpy, atomic\_write, word2vec, awscli, csv**: find matches/nearest top 3 neighbor in SQL Texts based on match criteria in env file and generate csv result
- **sqlalchemy**: connecting to database and loading summary csv file



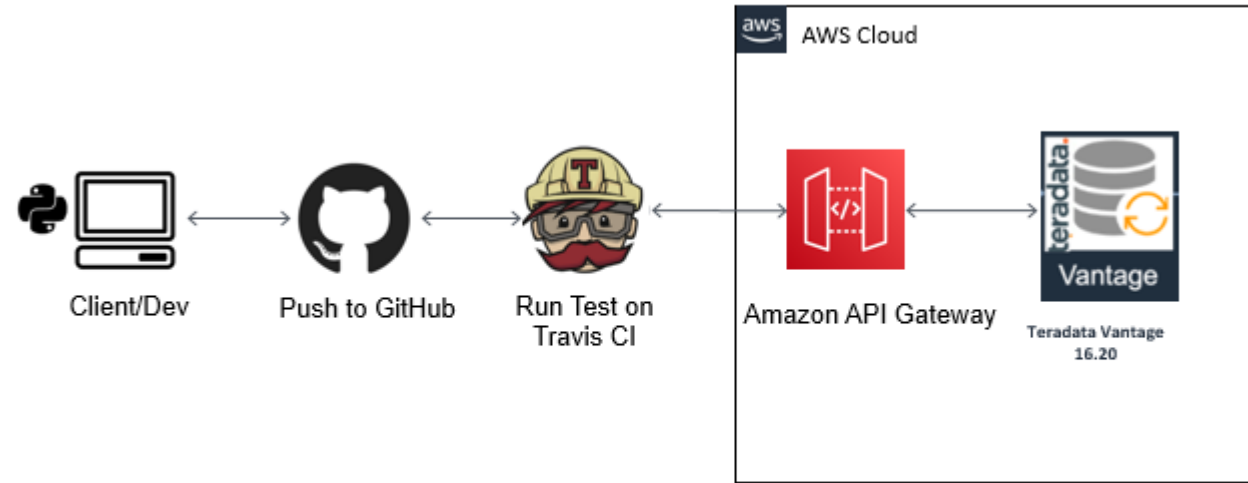
# Flow Diagram



# Architecture Diagram



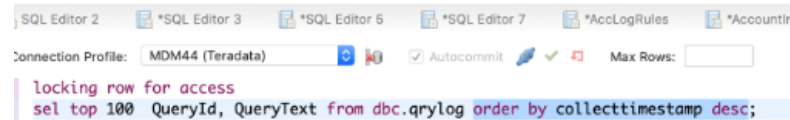
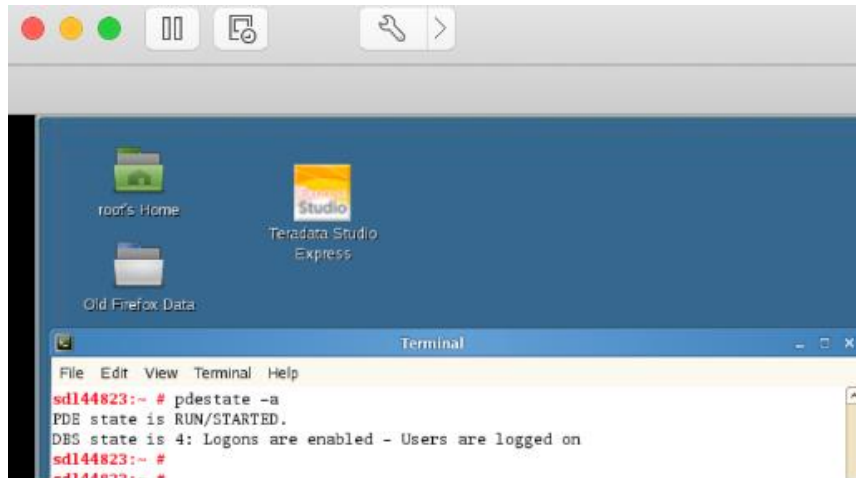
Dev Environment



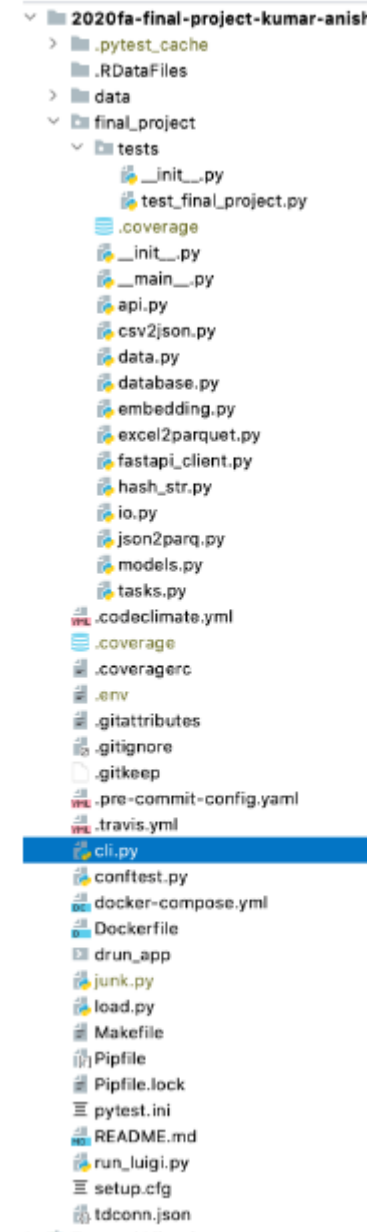
Prod Environment



# Dev Env Details / SQL Logs



Teradata Result Set Viewer			
Result Set - SQL Edi	Result Set - CL_CAS_	Result Set - SQL Edi	Result Set - SQL Edi
QueryID			
12	307190438469261069	SELECT TableName, TRIM (TableKind) AS TableKind, TRIM (CommentString) AS CommentStrin	
13	307190438469261061	SELECT FunctionName, TRIM(Fcns.FunctionType) AS FunctionType, TRIM(Fcns.SrcFileLangua	
14	307190438469261064	LOCK DBC.TableSizeV FOR ACCESS SELECT TableName, SUM(CurrentPerm) AS CurrentPerm,	
15	307190438469261060	SELECT FunctionName, TRIM(Fcns.FunctionType) AS FunctionType, TRIM(Fcns.SrcFileLangua	
16	307190438469261056	SELECT TableName AS ProcedureName, TRIM (CommentString) AS CommentString , ChildCo	
17	307190438469261062	SELECT TableName AS MacroName, RequestText, TRIM (CommentString) AS CommentString	
18	307190438469261066	LOCK DBC.JournalsV FOR ACCESS SELECT TableName, Journals_DB, JournalName FROM DB	
19	307190438469261084	locking row for access sel top 100 * QueryId, QueryText from dbc.qrylog order by collecttimestamp desc;	
20	307190438469261059	SELECT TableName AS FunctionName, TRIM(RequestText) AS RequestText, TRIM (CommentSt	
21	307190438469261030	locking row for access sel * from dbc.qrylog;	



# Design

- **api.py** (FastAPI/ uvicorn implementation) EXPORT FROM DB USING REST API CALL /API Output: JSON HTTP GET REQUEST is /querylog/ which execute query to fetch top 100 recent database query logs Below is the Query details
- **fastapi\_client.py** Use for testing the rest service
- **tasks.py** Luigi task which calls rest api and gives output of JSON file. /data/query\_logs.json
- **json2parq.py** Converts JSON to PARQUET
- **excel2parq.py** Converts Excel to PARQUET
- **data.py** loads numpy array and words.txt along with SQL Query Text parquet file (Code Reuse)
- **cli.py** SQL LOG TEXT Analytics (Word2Vec) in PYTHON Input numpy array, Words.txt & SQL Log / Output CSV summary (Code Modified)
- **embedding.py** (Code reuse by cli.py)
- **database.py** sqlalchemy database session management ( using env variable DB\_CONN)
- **models.py** sqlalchemy model for database table insert
- **load.py** Write CSV result summary to Teradata database (QUERYRESULT table)





# Project in Action – FastAPI, uvicorn

```
(2020fa-final-project-kumar-anish) ka@ak-imac 2020fa-final-project-kumar-anish % python final_project/api.py
```

```
INFO:      Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)
INFO:      Started reloader process [41437] using statreload
INFO:      Started server process [41443]
INFO:      Waiting for application startup.
INFO:      Application startup complete.
```



## Welcome to Final Project

This is a sample Python application that uses Teradata database to fetch query logs to users via REST API



## QueryLog

0.1.0 OAS3

/openapi.json

Sample REST API Application

### default

GET / Root

GET /querylog/ Querylog List



```
{
  "QueryID": "307190438469282088",
  "QueryText": "locking row for access\\rsel top 100 QueryId, QueryText from dbc.qrylog order by collecttimestamp desc;",
  "initialised": true,
  "QueryID": "307190438469261357",
  "QueryText": "locking row for access\\rsel top 100 QueryId, QueryText from dbc.qrylog order by collecttimestamp desc;",
  "initialised": true,
  "QueryID": "307190438469261270",
  "QueryText": "locking row for access\\rsel top 100 QueryId, QueryText from dbc.qrylog order by collecttimestamp desc;",
  "initialised": true,
  "QueryID": "307190438469261212",
  "QueryText": "locking row for access\\rsel top 100 QueryId, QueryText from dbc.qrylog order by collecttimestamp desc;",
  "initialised": true,
  "QueryID": "307190438469261165",
  "QueryText": "locking row for access\\rsel top 100 QueryId, QueryText from dbc.qrylog order by collecttimestamp desc;",
  "initialised": true,
  "QueryID": "307190438469261124",
  "QueryText": "locking row for access\\rsel top 100 QueryId, QueryText from dbc.qrylog order by collecttimestamp desc;",
  "initialised": true,
  "QueryID": "307190438469261065",
  "QueryText": "LOCK DBC.TableSizeV WHERE DataBaseName = ? GROUP BY TableName ORDER BY TableName",
  "initialised": true,
  "QueryID": "307190438469261057",
  "QueryText": "SELECT TableName AS ProcedureName, TRIM (CommentString) AS CommentString WHERE TableKind = 'P' AND DataBaseName = ?",
  "initialised": true,
  "QueryID": "307190438469261063",
  "QueryText": "RequestText, TRIM (CommentString) AS CommentString FROM DBC.TablesV WHERE DataBaseName=? AND TableKind = 'M' OF MacroName",
  "initialised": true,
  "QueryID": "307190438469261064",
  "QueryText": "LOCK DBC.TableSizeV FOR ACCESS SELECT SUM(CurrentPerm, SUM(PeakPerm) AS PeakPerm, (100 - (AVG(CurrentPerm)/NULLIFZERO(MAX(CurrentPerm))*100)) AS SkewFactor DataBaseName = ? GROUP BY TableName ORDER BY TableName",
  "initialised": true,
  "QueryID": "307190438469261058",
  "QueryText": "FunctionName, TRIM(RequestText), TRIM (CommentString) AS CommentString FROM DBC.TablesV WHERE DataBaseName=? AND TableKind = 'M' OF MacroName",
  "initialised": true,
  "QueryID": "307190438469261066",
  "QueryText": "LOCK DBC.JournalsV FOR ACCESS SELECT Te FROM DBC.JournalsV WHERE Tables_DB=?",
  "initialised": true,
  "QueryID": "307190438469261062",
  "QueryText": "RequestText, TRIM (CommentString) AS CommentString FROM DBC.TablesV WHERE DataBaseName=? AND TableKind = 'M' OF MacroName",
  "initialised": true,
  "QueryID": "307190438469261061",
  "QueryText": "SELECT FunctionName, TRIM(Fcns.Fcns.ExtFileReference) AS SrcFileLanguage, TRIM(Fcns.DeterministicOpt) AS DeterministicOpt, TRIM(Fcns.Externs.Fcns.ExtFileReference, TRIM(Fcns.NullCall) AS NullCall, TRIM(Fcns.SpecificName) AS SpecificName, TRIM(Fcns.Param (Tbls.CommentString) AS CommentString FROM DBC.FunctionsV Fcns .DBC.TablesV Tbls WHERE Fcns.DatabaseName = ? AND
```

```
INFO:      Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)
INFO:      Started reloader process [45437] using statreload
INFO:      Started server process [45439]
INFO:      Waiting for application startup.
INFO:      Application startup complete.
INFO: 127.0.0.1:61295 - "GET / HTTP/1.1" 200 OK
INFO: 127.0.0.1:61295 - "GET /favicon.ico HTTP/1.1" 404 Not Found
INFO: 127.0.0.1:61311 - "GET /querylog HTTP/1.1" 307 Temporary Redirect
SELECT TOP 100 QueryId
      , QueryText
FROM dbc.qrylog
ORDER BY collecttimestamp desc
INFO: 127.0.0.1:61311 - "GET /querylog/ HTTP/1.1" 200 OK
```



# Luigi, Text Analytics, Load & Results

```
(2020fa-final-project-kumar-anish) ka@ak-imac 2020fa-final-project-kumar-anish % python run_luigi.py
DEBUG: Checking if SQLLogsToJSON() is complete
INFO: Informed scheduler that task SQLLogsToJSON__99914b932b has status PENDING
INFO: Done scheduling tasks
INFO: Running Worker with 1 processes
DEBUG: Asking scheduler for work...
DEBUG: Pending tasks: 1
INFO: [pid 41519] Worker Worker(salt=332335116, workers=1, host=ak-imac.local, username=ka, pid=41519)
INFO: [pid 41519] Worker Worker(salt=332335116, workers=1, host=ak-imac.local, username=ka, pid=41519)
DEBUG: 1 running tasks, waiting for next task to finish
INFO: Informed scheduler that task SQLLogsToJSON__99914b932b has status DONE
DEBUG: Asking scheduler for work...
DEBUG: Done
DEBUG: There are no more tasks to run at this time
INFO: Worker Worker(salt=332335116, workers=1, host=ak-imac.local, username=ka, pid=41519) was stopped
INFO:
===== Luigi Execution Summary =====

Scheduled 1 tasks of which:
* 1 ran successfully:
  - 1 SQLLogsToJSON()

This progress looks :) because there were no failed tasks or missing dependencies

===== Luigi Execution Summary =====
```

```
(2020fa-final-project-kumar-anish) ka@ak-imac 2020fa-final-project-kumar-anish % python load.py
loading csv results to database - start...
loading csv results to database - done...
(2020fa-final-project-kumar-anish) ka@ak-imac 2020fa-final-project-kumar-anish %
```

```
(2020fa-final-project-kumar-anish) ka@ak-imac 2020fa-final-project-kumar-anish % python cli.py
my_matched_sql_keywords_text:
NULLIFZERO SUM AVG TRIM FunctionName
```

top 3 matched with high scores / shorted distance sql texts:

```
id          : 17
distance    : 0.42183470726013184
QueryText text:
SELECT FunctionName, TRIM(Fcns.FunctionType) AS FunctionType, TRIM(Fcns.SrcFileLanguage) AS Src
s.ExtFileReference, TRIM(Fcns.NullCall) AS NullCall, TRIM(Fcns.SpecificName) AS SpecificName, T
Fcns ,DBC.TablesV Tbls WHERE Fcns.DatabaseName = ? AND Fcns.RoutineKind = 'R' AND Fcns.Specifi
```

```
id          : 16
distance    : 0.42183470726013184
QueryText text:
SELECT FunctionName, TRIM(Fcns.FunctionType) AS FunctionType, TRIM(Fcns.SrcFileLanguage) AS Src
s.ExtFileReference, TRIM(Fcns.NullCall) AS NullCall, TRIM(Fcns.SpecificName) AS SpecificName, T
Fcns ,DBC.TablesV Tbls WHERE Fcns.DatabaseName = ? AND Fcns.RoutineKind = 'R' AND Fcns.Specifi
```

```
id          : 31
distance    : 0.4315459132194519
QueryText text:
LOCK DBC.TableSizeV FOR ACCESS SELECT TableName, SUM(CurrentPerm) AS CurrentPerm, SUM(PeakPerm)
WHERE DataBaseName = ? GROUP BY TableName ORDER BY TableName
```

```
(2020fa-final-project-kumar-anish) ka@ak-imac 2020fa-final-project-kumar-anish %
```

```
sel * from QUERYRESULT;
```

Teradata Result Set Viewer		
Result Set - SQL Editor (1)		Result Set - SQL Editor 1 (1)
	QUERYID	MATCHSCORE
1	307190438469261061	0.42183470726013184
2	307190438469260990	0.4315459132194519
3	307190438469261060	0.42183470726013184



# Reference

GitHub Repo:

<https://github.com/csci-e-29/2020fa-final-project-kumar-anish>

Fast API:

<https://fastapi.tiangolo.com/>

json2parquet:

<https://pypi.org/project/json2parquet/>

