FINAL REPORT – ETL PROJECT

ROADWORKS ACROSS PERTH METROPOLITAN AREA

TEAM QUOKKA - MARCH 2021

1 INTRODUCTION

This report presents a final summary of the ETL project for all the roadworks currently undertaken by Mainroads across the Perth Metropolitan Area, within the scope requested by the 'Client', whose main proposal aims to assess the disruption of roadworks on their commercial activities and community events in each suburb. The project's guidelines are located at WAUS-PERTH-DATA-PT-12-2020-U master repository.

In our preliminary team discussion, the scope management/responsibilities were structured as following:

Name	***Responsabilities***
M. Venables/ T. Zhao	Data research and selection
T. Zhao	ETL – Dataset1– GEOJSON format
M. Venables	ETL – Dataset2 – CSV format
M. Venables/ T. Zhao	Final Report

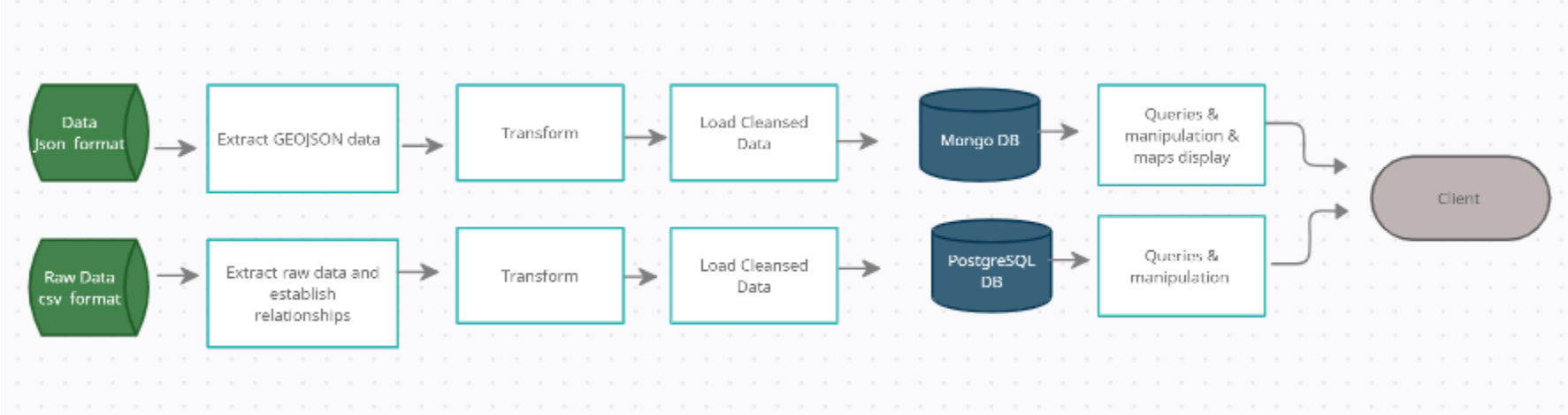
Our proposed workflow, in the section below, aimed to extend the current search options allowing our client to integrate both existing qualitative/quantitative datasets, and geographical search tools in their final analysis.

Regarding database selection, we decided the following for our Backend Architecture:

- MongoDB database used as a presenting area for the ETL process of JSON files due to its efficiency, support of unstructured data, and query capabilities. This was the preferred option as loading raw data into a SQL database requires the raw data to be error-free, and relationships between entities to be consistent, which was not achievable due to the timeline of this project.
- PostgreSQL database that contains operational data as our client has a highly structured data model.

1.1 WORKFLOW

After establishing our data sources and dependences, the project workflow was defined as following:



2 ETL – DATASET 1

This section was implemented using pandas, sqlalchemy, geopandas, folium, pymongo, numpy.

2.1 COLLECTION

General roadworks fields for the whole Metro and Regional areas projects:

```
#build general info field and insert into MongoDB
rwork = {}
gov_link = "https://catalogue.data.wa.gov.au/dataset/mrwa-roadworks"
geojson_link = "https://portal-mainroads.opendata.arcgis.com/datasets/f8df3952b47a4a73aac2df74a84a5b65_1.geojson"
csv_link = "https://portal-mainroads.opendata.arcgis.com/datasets/f8df3952b47a4a73aac2df74a84a5b65_1.csv"
general_map = "Perth_metro_roadwok.html"
thanks = "Thank you for visit, enjoy, best regards from team Quokka"
rwork["source"] = gov_link
rwork["gjson"] = geojson_link
rwork["csv"] = csv_link
rwork["general_map"] = general_map
rwork["Thanks"] = thanks

db.rworks.insert_one(rwork)
```

Detailed roadworks fields for the only Metro areas projects:

```
#Load the whole data into MongoDB
#However mongoDB doesn't accept folium map, I give it up for the maps' inserting.

for i in range(len(job_list)):
    bag = {}
    bag['Id'] = int(rmetro_df.iloc[i, 0])
    bag['StartDate'] = rmetro_df.iloc[i, 1]
    bag['FinishDate'] = rmetro_df.iloc[i, 2]
    bag['WorkType'] = rmetro_df.iloc[i, 3]
    bag['Description'] = rmetro_df.iloc[i, 4]
    bag['TrafficImpact'] = rmetro_df.iloc[i, 5]
    #bag['geometry'] = rmetro_df.iloc[i, 6]
    #bag['map'] = rmetro_df.iloc[i, 7]
```

2.2 EXTRACT

The raw dataset_2 (“roadworks. geojson”) was obtained from [mrwa-roadworks] website as a csv file and saved under Resources folder in our ETL project main GitHub repository. After download, the data was extracted into DataFrames using pandas as following:

```
<class 'geopandas.geodataframe.GeoDataFrame'>
RangeIndex: 136 entries, 0 to 135
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   OBJECTID                136 non-null    int64
 1   Id                      136 non-null    int64
 2   DateStarted             136 non-null    object
 3   EstimatedCompletionDate 136 non-null    object
 4   WorkType                136 non-null    object
 5   Description             136 non-null    object
 6   Suburb                  136 non-null    object
 7   Road                    136 non-null    object
 8   TrafficImpact           136 non-null    object
 9   Region                  136 non-null    object
 10  EntryDate               136 non-null    object
 11  geometry                136 non-null    geometry
dtypes: geometry(1), int64(2), object(9)
memory usage: 12.9+ KB
```

2.3 TRANSFORM

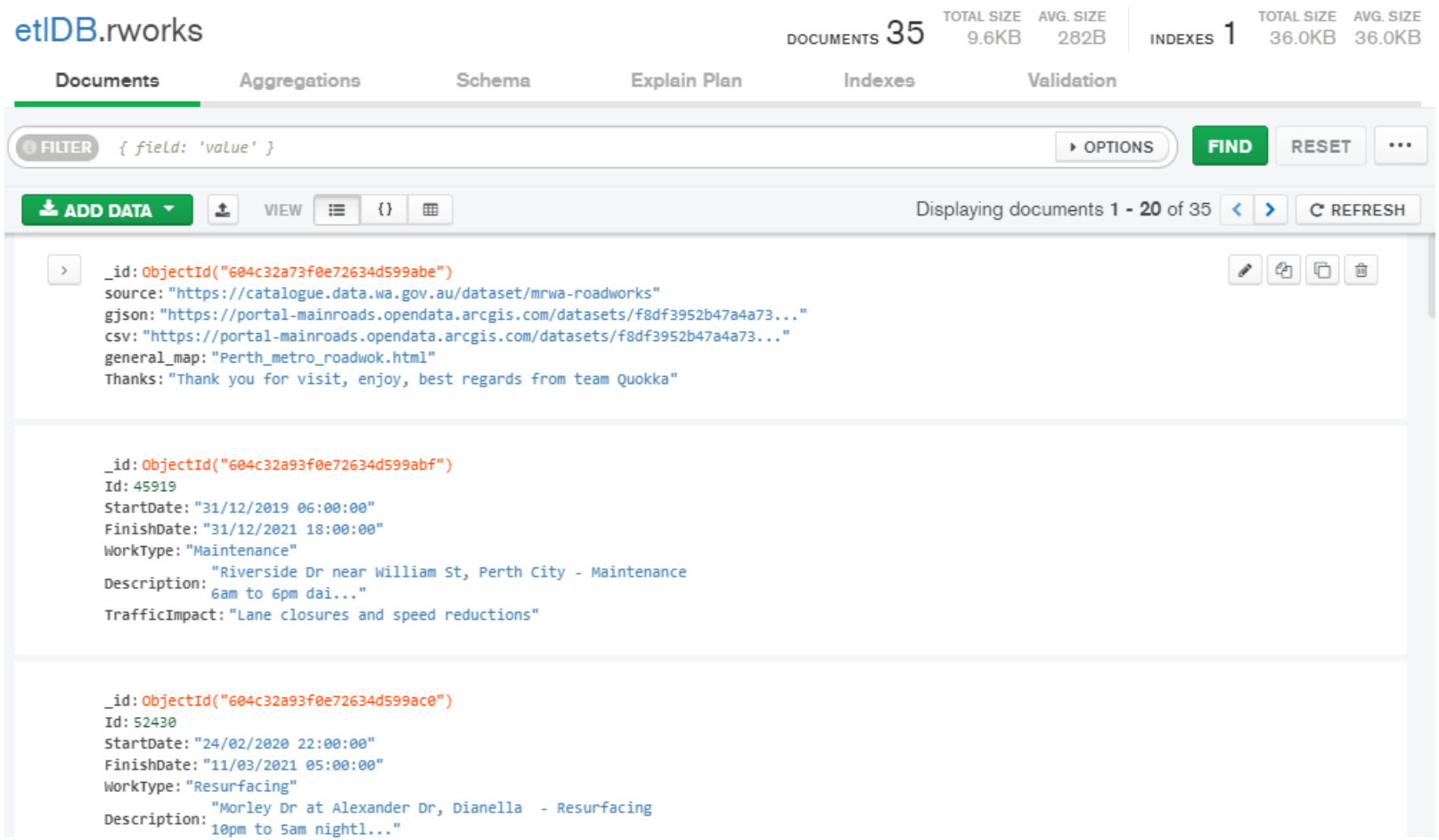
For the transformation step, the following was pursued:

- Create a filtered dataframe from specific columns and rename columns as per schema.
- Create a filtered dataframe from specific location that includes only roadworks current undertaken at Perth Metropolitan Area (‘Metro’).
- Add a new empty column map to the dataframe for future use.

```
<class 'geopandas.geodataframe.GeoDataFrame'>
RangeIndex: 34 entries, 0 to 33
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             34 non-null     int64
 1   StartDate      34 non-null     object
 2   FinishDate     34 non-null     object
 3   WorkType       34 non-null     object
 4   Description    34 non-null     object
 5   TrafficImpact  34 non-null     object
 6   geometry       34 non-null     geometry
 7   map            34 non-null     object
dtypes: geometry(1), int64(1), object(6)
memory usage: 2.2+ KB
```

2.4 LOAD

The clean data was successfully loaded to the MongoDB database under collection ‘roadworks’ as per below:



2.5 QUERIES & DISPLAY

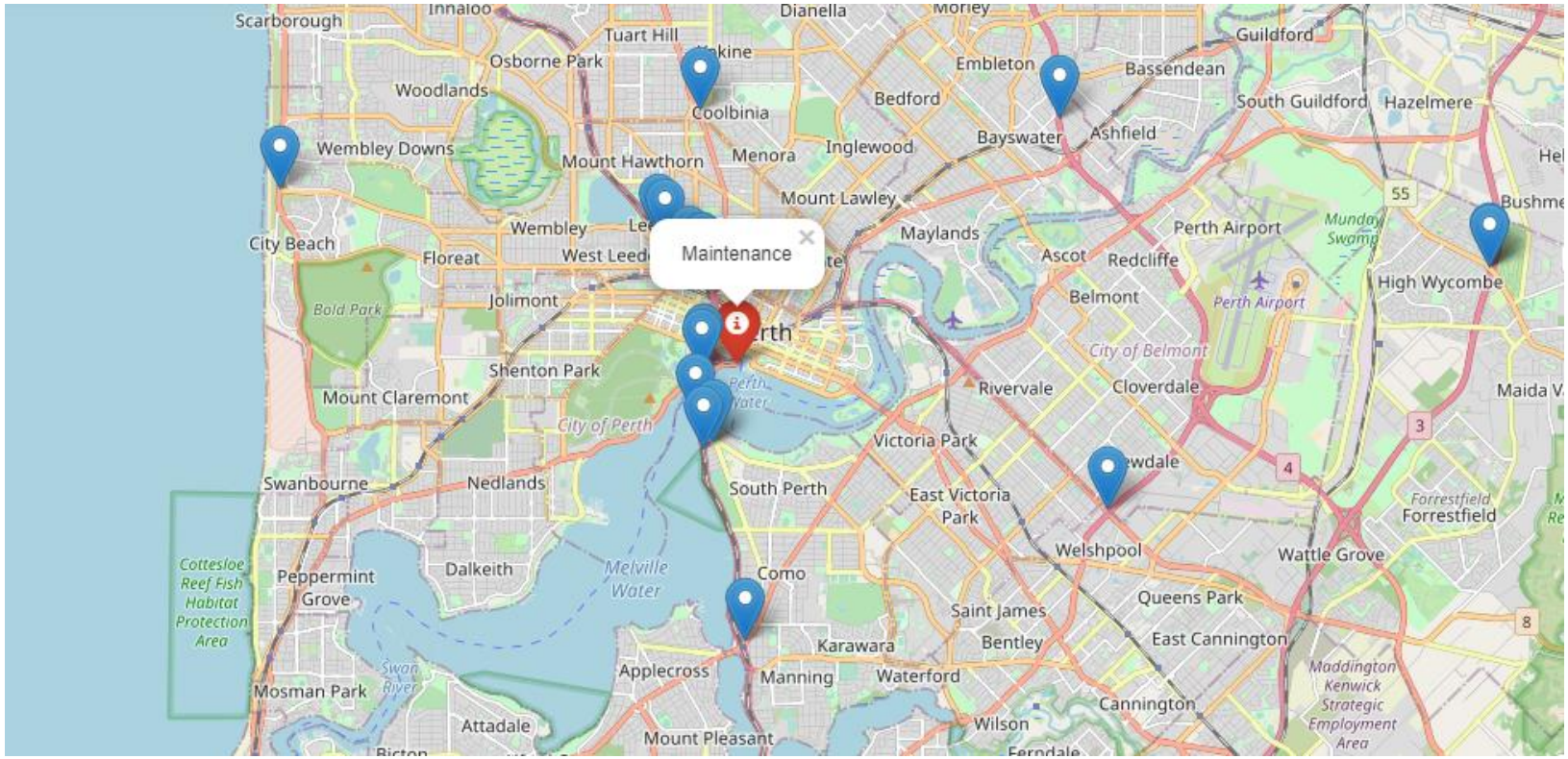
The queries for the project's location map as per below:

```
n = db.rworks.count_documents({})
print (n)
```

```
37
```

```
roadworks = db.rworks.find()
for r in roadworks:
    print(r)
```

```
{'_id': ObjectId('604acc6752696ec426c17dd0'), 'source': 'https://catalogue.data.wa.gov.au/dataset/mrwa-roadworks', 'gjson':
'https://portal-mainroads.opendata.arcgis.com/datasets/f8df3952b47a4a73aac2df74a84a5b65_1.geojson', 'csv': 'https://portal-ma
inroads.opendata.arcgis.com/datasets/f8df3952b47a4a73aac2df74a84a5b65_1.csv', 'Thanks': 'Thank you for visit, enjoy, best reg
ards from team Quokka'}
{'_id': ObjectId('604acddc52696ec426c17dd1'), 'source': 'https://catalogue.data.wa.gov.au/dataset/mrwa-roadworks', 'gjson':
'https://portal-mainroads.opendata.arcgis.com/datasets/f8df3952b47a4a73aac2df74a84a5b65_1.geojson', 'csv': 'https://portal-ma
inroads.opendata.arcgis.com/datasets/f8df3952b47a4a73aac2df74a84a5b65_1.csv', 'general_map': 'Perth_metro_roadwok.html', 'Tha
nks': 'Thank you for visit, enjoy, best regards from team Quokka'}
```



3 ETL – DATASET 2

This section was implemented using pandas, SQLAlchemy and Postgresql.

3.1 SCHEMA

```sql
-- Create Two Tables
CREATE TABLE roadworks_details(
  Id int PRIMARY KEY,
  StartDate Date,
  FinishDate Date,
  WorkType varchar(250),
  Suburb varchar(250),
  Road varchar(250),
  Region varchar(250),
  TrafficImpact varchar(250)
);
```

3.2 EXTRACT

The raw dataset_2 (**"**roadworks.csv**")** was obtained from [mrwa-roadworks] website as a csv file and saved under Resource's folder in our ETL project main GitHub repository. After download, the data was extracted into DataFrames using pandas as following:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 136 entries, 0 to 135
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   X                        136 non-null    float64
 1   Y                        136 non-null    float64
 2   OBJECTID                 136 non-null    int64
 3   Id                       136 non-null    int64
 4   DateStarted              136 non-null    object
 5   EstimatedCompletionDate  136 non-null    object
 6   WorkType                 136 non-null    object
 7   Description              136 non-null    object
 8   Suburb                   92 non-null     object
 9   Road                     136 non-null    object
 10  TrafficImpact            136 non-null    object
 11  Region                   136 non-null    object
 12  EntryDate                136 non-null    object
dtypes: float64(2), int64(2), object(9)
memory usage: 13.9+ KB
```

3.3 TRANSFORM

For the transformation step, the following was pursued:

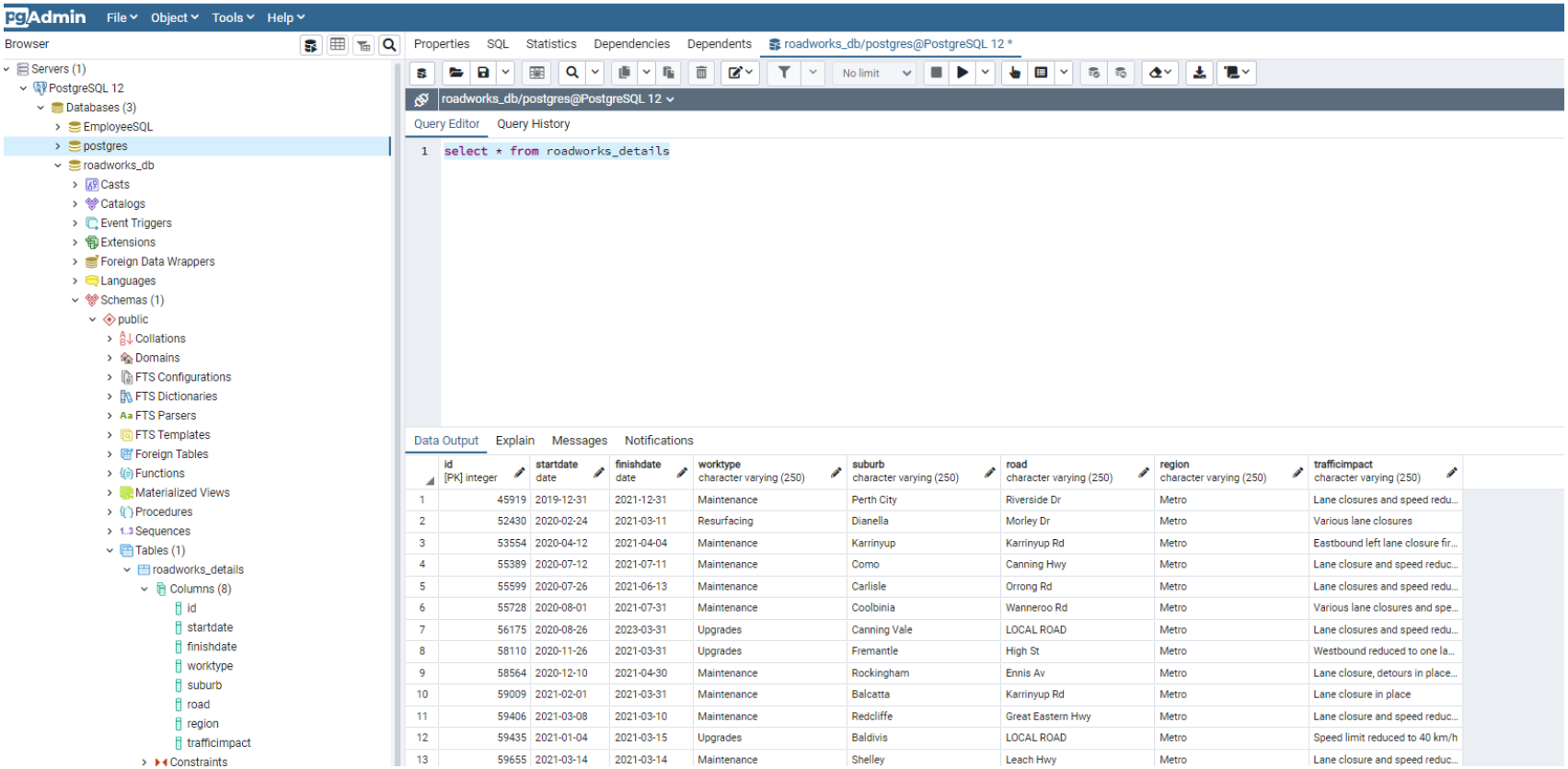
- Create a filtered dataframe from specific columns and rename columns as per schema.
- Create a filtered dataframe from a specific location that includes only roadworks currently undertaken in the Perth Metropolitan Area (‘Metro’).

The data after transformation was displayed as following:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17 entries, 0 to 71
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             17 non-null     int64
 1   startdate      17 non-null     object
 2   finishdate     17 non-null     object
 3   worktype       17 non-null     object
 4   suburb         17 non-null     object
 5   road           17 non-null     object
 6   region         17 non-null     object
 7   trafficimpact  17 non-null     object
dtypes: int64(1), object(7)
memory usage: 1.2+ KB
```

3.4 LOAD

We encountered a few issues to establish the initial connection with our database '*roadworks_db'*, however after removing all the connection errors the clean data was successfully loaded to the PostgresSQL database table '*roadworks_details'* as per below:



	id [PK] integer	startdate date	finishdate date	worktype character varying (250)	suburb character varying (250)	road character varying (250)	region character varying (250)	trafficimpact character varying (250)
1	45919	2019-12-31	2021-12-31	Maintenance	Perth City	Riverside Dr	Metro	Lane closures and speed redu...
2	52430	2020-02-24	2021-03-11	Resurfacing	Dianella	Morley Dr	Metro	Various lane closures
3	53554	2020-04-12	2021-04-04	Maintenance	Karrinyup	Karrinyup Rd	Metro	Eastbound left lane closure fir...
4	55389	2020-07-12	2021-07-11	Maintenance	Como	Canning Hwy	Metro	Lane closure and speed reduc...
5	55599	2020-07-26	2021-06-13	Maintenance	Carlisle	Orrong Rd	Metro	Lane closures and speed redu...
6	55728	2020-08-01	2021-07-31	Maintenance	Coolbinia	Wanneroo Rd	Metro	Various lane closures and spe...
7	56175	2020-08-26	2023-03-31	Upgrades	Canning Vale	LOCAL ROAD	Metro	Lane closures and speed redu...
8	58110	2020-11-26	2021-03-31	Upgrades	Fremantle	High St	Metro	Westbound reduced to one la...
9	58564	2020-12-10	2021-04-30	Maintenance	Rockingham	Ennis Av	Metro	Lane closure, detours in place...
10	59009	2021-02-01	2021-03-31	Maintenance	Balcatta	Karrinyup Rd	Metro	Lane closure in place
11	59406	2021-03-08	2021-03-10	Maintenance	Redcliffe	Great Eastern Hwy	Metro	Lane closure and speed reduc...
12	59435	2021-01-04	2021-03-15	Upgrades	Baldivis	LOCAL ROAD	Metro	Speed limit reduced to 40 km/h
13	59655	2021-03-14	2021-03-14	Maintenance	Shelley	Leach Hwy	Metro	Lane closure and speed reduc...

4 FINAL CONSIDERATIONS

The outcome of this project (***Extract – Transform – Load***) and the steps necessary to obtain these results was straightforward given the timeframe and raw data available for the chosen subject, the clean data was successfully loaded to both , Postgresql and Mongodb, databases as per client's request.

5 REFERENCES

- https://catalogue.data.wa.gov.au/dataset/mrwa-roadworks
- https://pandas.pydata.org/pandas-docs/version/0.25.3/