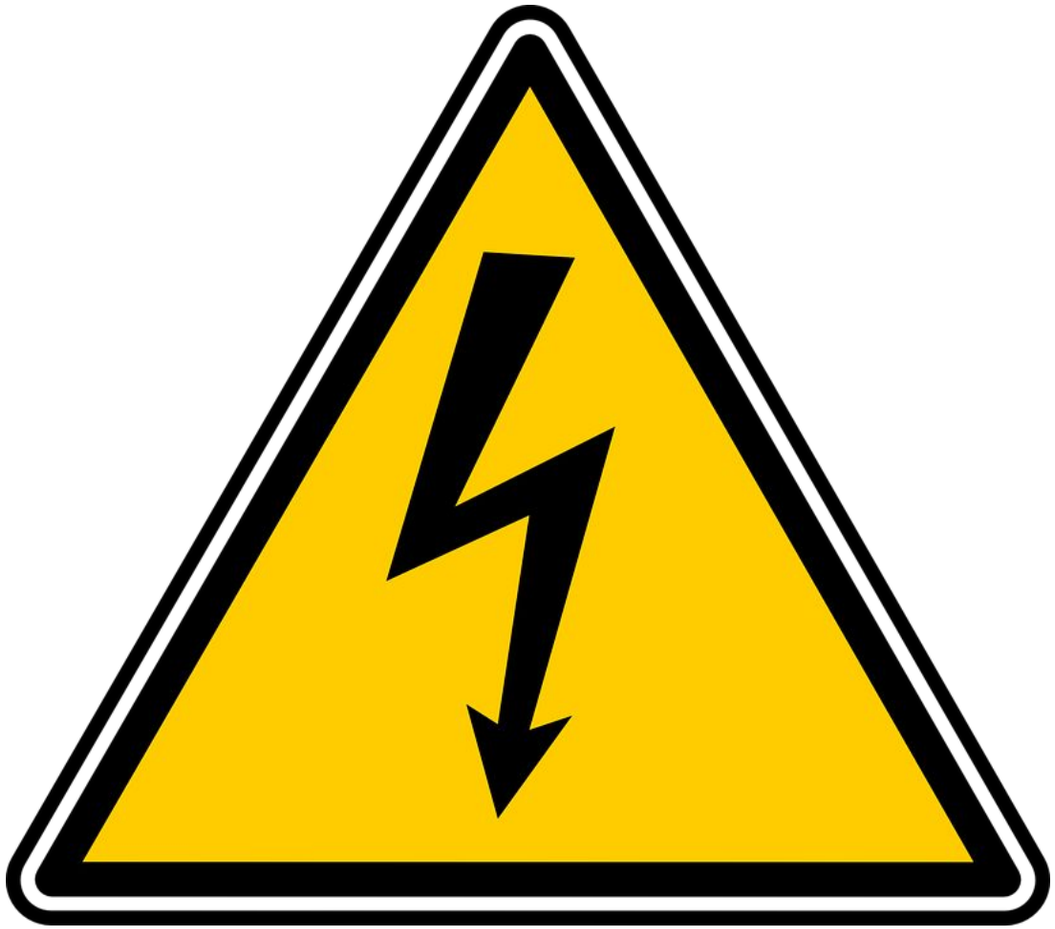
EXPLORE || DATA SCIENCE ACADEMY

Gather

Predict Instructions

Topics to be covered

1. **Problem Statement**
2. **ETL Pipelines**
3. **Importing Eskom data**
   - What data is available?
   - Restoring the database
4. **Exploring the data**
   - Tables in the database
   - Relationships in the database
5. **Creating/Editing ER diagrams**



Problem Statement

To whom it may concern,

Your tender application has been successful.

You have been tasked to create a new system that would assist in the estimation of future demand based on historical data. To achieve this we need a new and improved data infrastructure that is able to handle large amounts of data and is organised in a manner that allows for computationally efficient access.

Additionally, we would like to track customer satisfaction around our products and services. We require a data pipeline that can stream twitter data into an existing database, which will later on be analysed for sentiment.

We are committed to becoming a world class energy provider.
The power is in your hands.

Data Pipelines

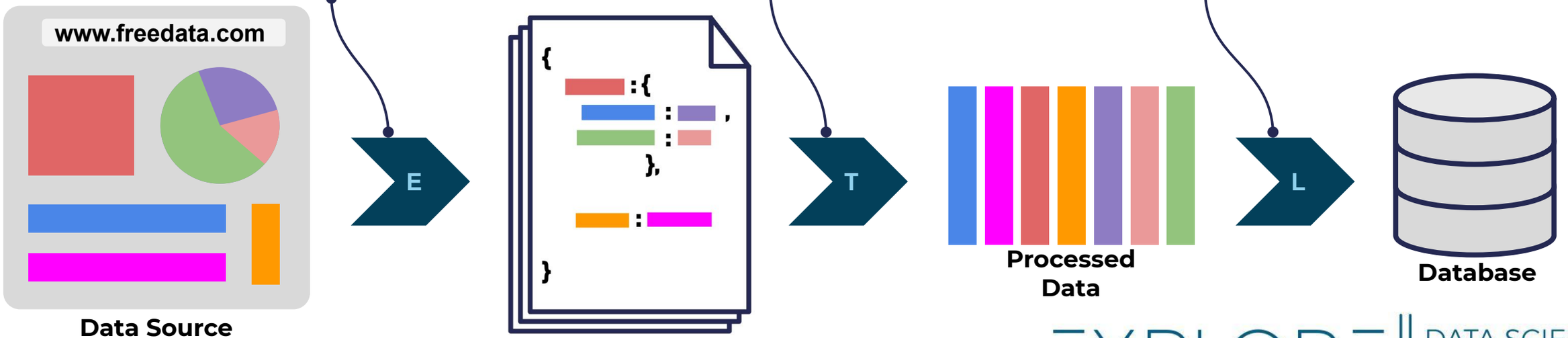
While working with CSV's in an ad-hoc manner can be sufficient for basic data science programs, larger data systems require that the data extraction and processing to be deterministic, reliable, and efficient. In most data products and applications this is achieved by setting up a **data pipeline**.

Data pipelines facilitate the flow of data from a data source to a destination, with the aim of making the data suitable for consumption by the destination system. They are typically made up of a sequence of processing nodes connected in series which apply operations to the data as it flows through them and are typically designed to be reusable and maintainable. Data pipelines can vary depending on the volume, frequency, or nature of data flowing through them.

Data Pipelines

A common example of a data pipeline is an **ETL pipeline**, where the data pipeline is broken down into data **E**xtraction, data **T**ransforming, and data **L**oading.

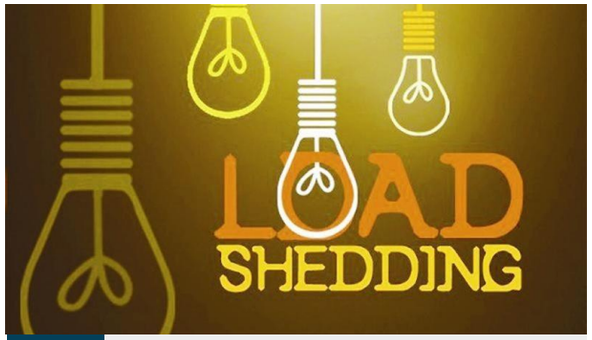
EXTRACT	TRANSFORM	LOAD
Collecting raw data from data source(s)*, for example through accessing API endpoints, crawling or scraping websites, running surveys, etc. * For larger systems, this data is often collected into a central location e.g. a **datalake** and is separated from an ETL pipeline, but for the purposes of this predict, each of the data sources will be treated as part of the ETL pipeline.	Since the collected raw data typically comes in different formats and requires that we transform the data, i.e. converting to other formats, mapping, handling incomplete or corrupted parts of the data, pivoting, etc.	Finally, the transformed data has to be loaded into a storage system i.e. databases, data warehouses, Amazon S3, etc. The loading process is informed by the volume of the data as well as the the requirements of the data product or application being developed.



EXPLORE | DATA SCIENCE ACADEMY

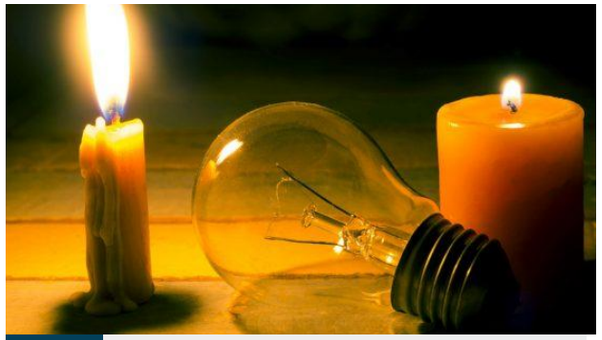
Gather | **Predict**

Project: Collect data from various sources and store it in a database.



DRAFT

- Design a problem statement
- Identify data sources



DO

- Collect data from sources i.e. via web scraping or API's
- Build SQL database of collected data
- Data clean up and DB normalisation

DELIVER

- ETL pipeline.
  i.e. Python functions for querying the database.
- .bak file of SQL Database. i.e.

DECOMPRESS

- Insights
- Reuse module in subsequent sprints
- Feedback: SQL Server Security Considerations

EXPLORE | DATA SCIENCE ACADEMY

What data is available?

The Eskom data that has been provided was collected from a number of online available resources, such as the city of StatsSA,twitter, and Eskom. It includes information about the following -

- Stations (Wind , Nuclear ,Gas ,Hydroelectric and Coal) data
- Twitter data
- Electricity consumption data.
- Province consumption statistics
- Yearly consumption statistics.
- Number of homes/Infrastructure electrified each year.

Exploring the Database

Province

- **Province_ID**
- **Province** names

province_electrification

- **Year columns:** (eg. 2010,2011,2012, etc)
- **province** Names

all_stations

- **station_name:** Name of power plant
- **_Total_installed_Capacity_MW :** Intended full-load sustained output of a facility
- **_Total_Nominal_Capacity_MW_:** Actual full-load sustained output of a facility
- **Station_type:** Source of electricity generated
- **Location:** Region/city/location at which the power plant is located

Population

- **Year** of recording
- **Population** size

Station_table

- **Station Name**
- **Station_ID**

twitter_table *

- **tweets :** actual posted tweet
- **Date :** date and time of the tweet

* Must be added to the database

Exploring the Database... Continued

electricity_generated

- **electricity_distributed:** Amount of electricity distributed to the grid (Gigawatt/hour)
- **Date**: month-year of recording

year_on_year_change

- **Months as rows (eg. JAN, FEB, etc)**
- **Years** as columns **(eg, 2009, 2010, 2011, etc)**
- **Matrix** showing the percentage change of the electricity distributed in each year and month

Different station tables (coal, hydroelectric, gas turbine, wind, nuclear)

- **Station_Name:** All stations names.
- **Province :** Province the station is located.
- **Date commissioned (planned) :** Dates of commission.
- **Capacity (MW) planned :** Power produced in MegaWatts.
- **Status :** Status whether is operational or not operational.
- **Coordinates :** Coordinates of the location of the station.
- **Operator :** The owner or subcontracted operator.
- **Planned decommissioning dates:** Date planned to close or shutdown the station.
- **Notes :** Facts and additional information on the station.

Deliverables For Gather Predict:

1

A **single notebook** which encompasses the entire **ETL pipeline**. i.e. functions

- Extracting data from twitter, specifically from *Eskom_SA*.
- Transforming the data using the given python functions.
- Connecting to and querying the given SQL database.
- Loading transformed data to appropriate table/tables in the database, updating them.

2

A **backup SQL file** of a cleaned up **database** containing Eskom data

- ER diagram describing the different connections in the DB in a 3rd normalized form.
- Database saved/exported as .bak file, which will be zipped and uploaded to Athena.

Team projects

Tasks for ETL Pipeline Notebook

1

Project instructions available under the ‘Predict’ tab on Athena. You will need to:

- **Build an ETL Pipeline** which :
  - **Scrapes Twitter** Data
  - **Cleans** the data using **one** of the text **functions** built in **Analyse**
  - **Updates** twitter table in **SQL** with the clean data

All to answer at the end of fundamentals:

- *What are the infrastructural requirements that Eskom needs to prioritise in order to meet current and future electricity demands?*
- *What are the factors driving the current sentiment around the state of Eskom?*

Team projects

Tasks for Cleaning the Database

Project instructions available under the ‘Predict’ tab on Athena. You will need to:

- **Restore** a SQL database containing some data
- **Collect** and add twitter data to relevant tables in the database
- **Explore & Clean** the database and make use of keys to define or correct relationships between tables
- Create and edit an Entity Relationship Diagram (**ERD**) describing those relationships
- Optionally, search for and include additional data to the database

All to answer at the end of fundamentals:

- *What are the infrastructural requirements that Eskom needs to prioritise in order to meet current and future electricity demands?*
- *What are the factors driving the current sentiment around the state of Eskom?*