

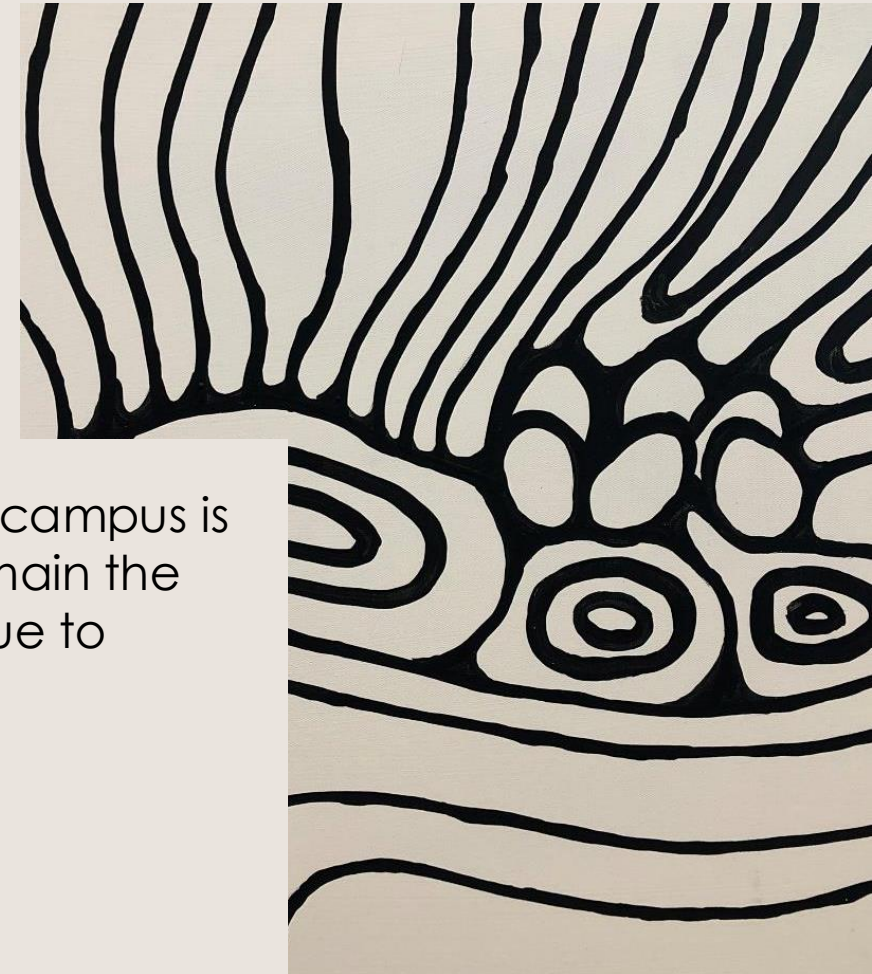


Topic Nine: Data Workflows

INMT5526: Business Intelligence

Acknowledgement of country

The University of Western Australia acknowledges that its campus is situated on Noongar land, and that Noongar people remain the spiritual and cultural custodians of their land, and continue to practise their values, languages, beliefs and knowledge.



Artist: Dr Richard Barry Walley OAM

What is a data workflow?

- A data workflow is a sequence of tasks undertaken to process data.
 - Depending on what we are trying to do, as well as the content of the dataset itself, the workflow will differ, but in a higher-level sense it will contain a subset of the same steps.
 - When considering the generation of a report (or eventually dashboard) in the Business Intelligence sense, there will be a common high-level workflow we will undertake.

Business Intelligence workflow

- Everyone will have a different opinion on the specifics of this style of '*data analysis workflow*', but the following provides a good summary:



- It is often said that the majority of time is spent in the first three steps of this process!

Problem identification

- Often, we forget why we do the things we do.
 - “Failure to plan is planning to fail” – Benjamin Franklin?
- Hence, our first step is to ensure we consider:
 - what the question is that we are trying to answer, and;
 - who will be utilising our dashboards and report.
- These answers may not be fixed; they may change during the project, or we may be making something for our own exploration. That is still perfectly OK!

Problem statement

- One way that we can frame this is in terms of a problem statement that defines the problem we are trying to solve and who we are solving it for:
 - This week's practical: I am interested in the **average daily percentage change** of the **stock prices of Microsoft, Apple and Amazon**.
- From this, we are able to determine the data we will need, how we will analyse it and it enables us to think about how we will present it.

Data acquisition

- This may sound silly, but we actually have to find fit-for-purpose data that can help answer our problem. Consider the following questions:
 - What considerations should we be making when we are acquiring data?
 - What does the 'who' tell us in our problem statement? Why is that important?

Acquisition – theoretical issues (I)

- Some questions we can ask of the dataset we are looking to use:
 - Is the source of the data trustworthy and authoritative?
 - Are we provided with metadata (description)?
 - Does the data conform to a recognised standard?
 - Does the data (help) answer the identified problem?
 - Are there any licensing issues? (cost, time, legal)

Acquisition – theoretical issues (II)

- Additional questions we can ask of the data:
 - Is the data current? If not, is that suitable?
 - How accurate is the data?
 - How is the data recorded? (e.g. time series frequency)
 - What data format is the data in? Does it work with our tools?
 - Is the data 'imported' or 'connected'?

Acquisition – practical issues

- This stage is concerned with the practical elements of acquiring the data and the issues this raises, rather than the theoretical issues previously described.
- This is then the process of getting the data from another system (where it is generated or summarised) into your system, as well as the issues raised:
 - How big is the data file?
 - Can it be sent over the internet?
 - How long will it take for the data to be generated?

Data preparation

- Data is (almost) never exactly how we want it.
 - Some data may be missing...
 - Some data may be in the wrong format...
 - Some data may be inconsistent...
 - Some data may need to be normalised...
 - Some data may need to be 'enriched'...
 - Some data may not need to be there...

ETL vs data preparation

- The acronym ETL (Extract / Elaborate Transform Load) is commonly used by many practitioners interchangeably with the term data preparation.
 - Components of data preparation include data transformation and data cleaning;
 - The concept of ETL dates back around 50 years and is linked with the concept of a data warehouse – but the two are certainly not the same, just work together!

ETL vs data preparation

- It is hard to find agreement on a true meaning of the differences, but it is safe to say that ETL is one method to cover data acquisition and preparation:
 - **Extract** the data from the source system;
 - **Transform** it to the format of the destination;
 - **Load** it into the destination system.
- As you can see, this is fundamentally an abstract process, but one that models the general process we use building dashboards and reports in Power BI Desktop.

Data analysis

- This is the part where we answer the problem we started with!
 - Although with Power BI, analysis is tightly integrated with visualisation – consider how we generated our R visuals last week from the analysis!
- To find answers to our questions, a wide variety of techniques can be used:
 - Often, we first undertake Exploratory Data Analysis with descriptive statistics and visualisation, to get a rough feel for what the data is like;
 - Then, we can move on to 'the real thing'...

Analysis activities (I)

- Basic analytic activities have been taxonomized (categorised) as follows; it's a great way to look at it, however, it's not the only way to look at it:
 - **Get Value**: what was the closing price today for BHP?
 - **Filter**: which products sold more than 500 units?
 - **Derive Value**: what is the average CTR of each ad?
 - **Find Extreme**: which salesperson sold the most? (value)
 - **Sort**: order the ASX200 by profit (largest to smallest)
 - **Find Range**: what was the minimum and maximum revenue for our franchisees?

Analysis activities (II)

- Continuing on from the last slide, there is also the following:
 - **Distribution:** what is the mean and standard deviation of the discounts on product X?
 - **Anomaly Detection:** which of our developers is an outlier? (by lines of code written)
 - **Clustering:** is there a cluster of the number of hours worked per week?
 - **Correlation:** is there a relationship between the productivity and remuneration of our workers (or two other variables)? *Think back to our R visuals last week!*
 - **Contextualisation:** which bus routes can get me to my meeting in Joondalup?

MECE principle

- We must ensure that we can answer our question with data, which means...
 - That the data we have is of suitable quality and is complete in nature;
 - The observations (a.k.a. measurements) and their attributes (a.k.a. dimensions) are relevant to and can answer the question (a.k.a. problem)!
- Many years ago, McKinsey developed a method to break down these problems into smaller ones, considering the above and named the MECE principle.
 - It simply refers to the property of being **mutually exclusive** and **collectively exhaustive**.
 - We'll explain this problem in context with an example on the next slide.

MECE example

- Consider a business that sells 10 different products – with profit and stock levels:
 - The businesses' total profit is the **collectively exhaustive sum** of the **mutually exclusive profit** of each of the ten items – that is, we add together each items' profit;
 - The total stock on hand is the **collectively exhaustive sum** of the **mutually exclusive stock level** of each product – that is, we add together each items' stock level.
- Through this method, we can **drill up** and **down** data and break problems down into smaller ones and conversely summarise the information at a higher level.

Other things we can do with data

- In the field of business intelligence, we generally output our analysis results as visualisations. However, we can do other things with our data:
 - Generate derived data sets – for further analysis by others or for other uses that we have;
 - Apply mathematical models (formulas/algorithms) – consider this ‘advanced analysis’;
- However, the overall process that has just been described is somewhat iterative.
 - We start out with simple, make analysis more complex, then add more data sources.
 - You will learn (complex) modelling in other units – there is plenty of interesting things to do.

Photo: [Unsplash](#)



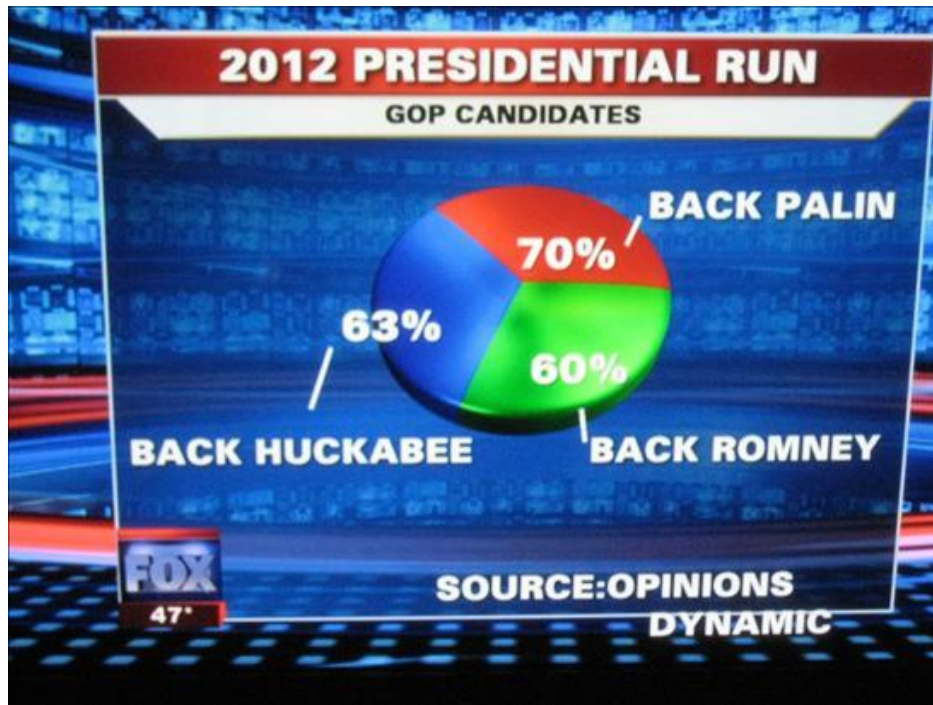
Topic Nine: Data Visualisation

INMT5526: Business Intelligence

Data visualisation

- The last step in our methodology is to present the results to the target audience in an understandable way. There are however a couple of considerations:
 - Confusing fact for opinion, ensuring it is the data that tells the story;
 - Misrepresenting data through selective representation of data or the use of misleading visualisations (e.g. wrong type, unusual axes, poorly labelled).

Examples of bad presentation!



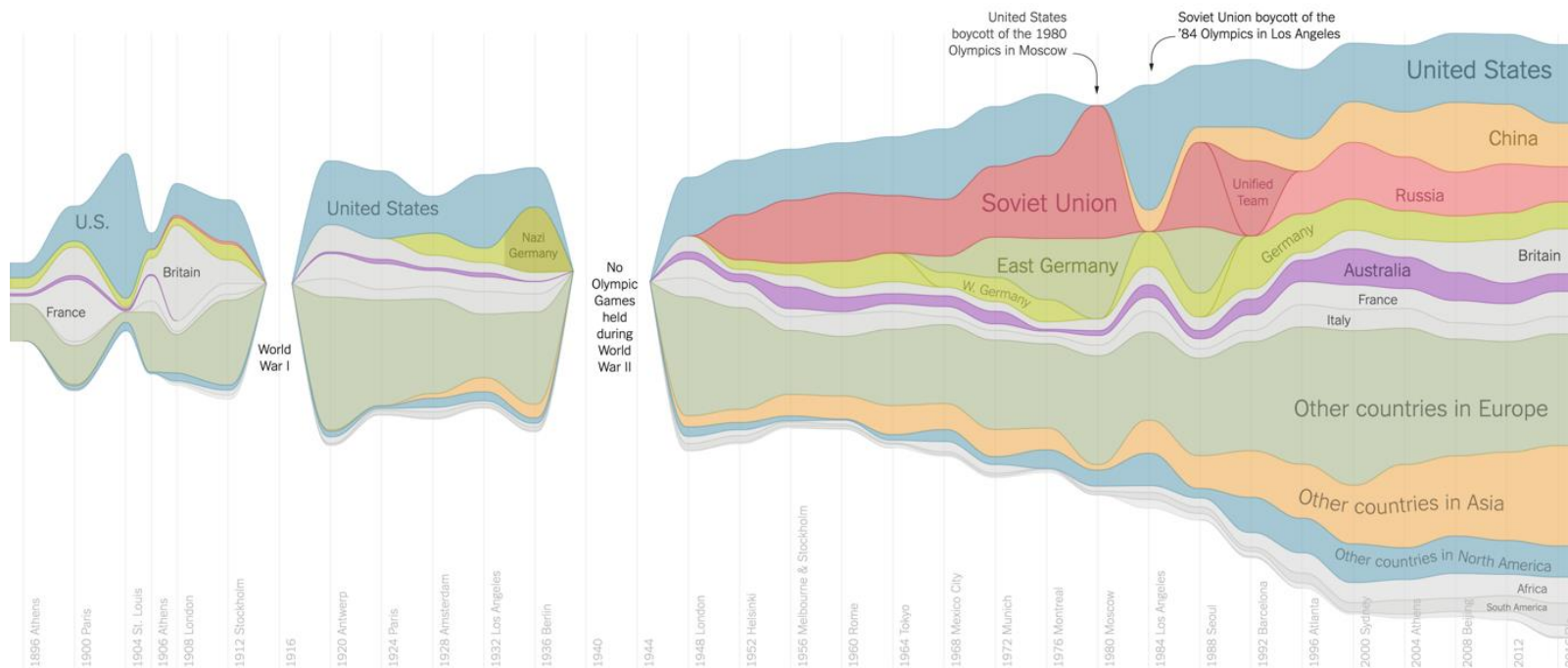
Source: FOX News



Source: Waterman Broadcasting of Florida (NBC2)

An excuse for a good graphic!

- Found during research, I thought this was relevant...



Source: New York Times

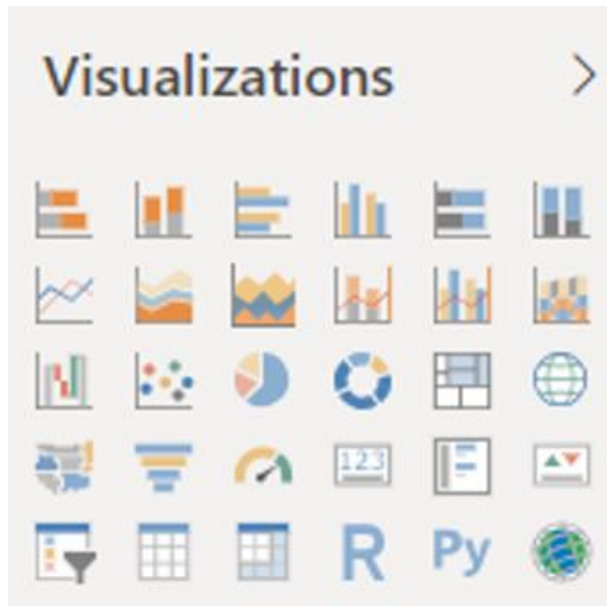
Data visualisation

- What can we show in our visualisations? Another categorisation:
 - **Time series:** how a quantity changed over time – e.g. GDP per quarter since 1972;
 - **Order/Ranking:** largest to smallest, A to Z – e.g. products by total sales volume;
 - **Partition:** such as percentages of a whole – e.g. percentage of sales by country, Olympic (or Commonwealth Games) medals, as we saw on the previous slide;
 - **Deviation:** planned versus actual – e.g. expenditure within a business unit;

Data visualisation (II)

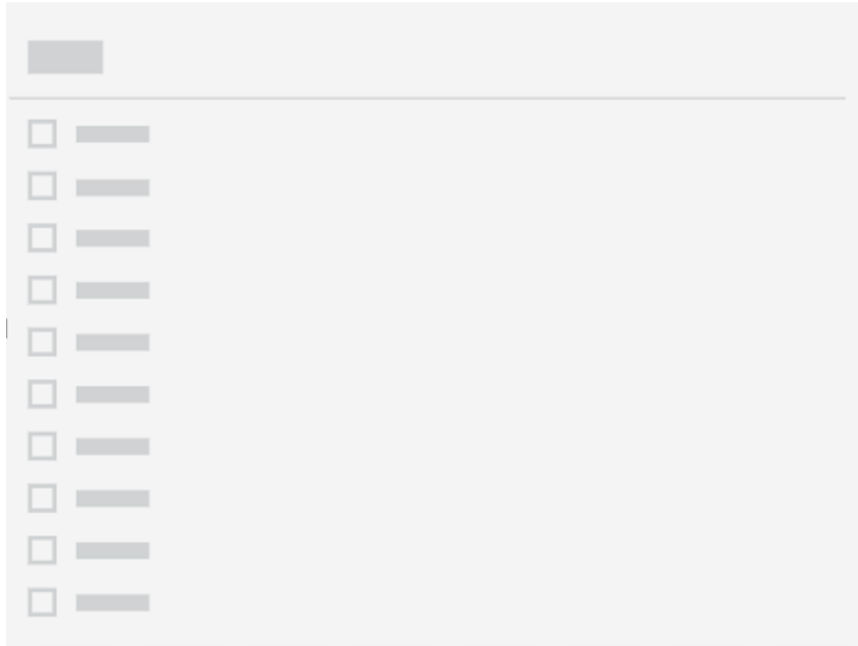
- Additionally, we can (or could) also show the following:
 - **Frequency:** how often a quantity occurred – e.g. visitors to a retail store per hour;
 - **Correlation:** this is two-dimensional; with one variable compared to another – e.g. extending from above, determining relationship between total hours worked vs GDP;
 - **Comparison:** comparing categorical data – e.g. total sales per sales channel;
 - **Geospatial:** map or plan diagram – e.g. employees per physical level of a building;

Types of visualisations



- We can see the common types in **Power BI** such as bar, column, line, scatter, map and so forth.
- Types used depend on what we are visualising.
- Consider when it is appropriate to use a particular type of visualisation.

Drilling up and down



- Not all are strictly visualisations; consider our slicer from last week.
- We must ensure that our data is mutually exclusive for our slicer to function in a practical way.
- We looked at other ‘not quite’ visualisations last week!

Photo: [Unsplash](#)



Topic Nine: BI Project of the Week

INMT5526: Business Intelligence

MySQL and Power BI

- This week, in the practical we will create a dashboard showing some of the data within our MySQL database that we worked on earlier in semester.
 - Create a new project (a.k.a. dashboard/file) in Power BI Desktop.
 - Choose to “Get Data” (from another source) from a “MySQL database” (then Connect).
 - The “Server” is **db.tris.id.au** as it has been this semester ☺
 - The “Database” is **<YourStudentID>_db** as it was earlier in semester too – replacing the part at the start with your student ID, as we had before!
- There always seems to be issues with the MySQL connector, so do not worry if this does not work – it is not the main focus of the practical.

This week's project

- This is a very complex and long project!
 - We bring a lot of what we have done this semester together!
- Do not worry if you don't finish it during the practical.
 - You'll have a bit of time next week to look at it as well.
- Read ahead as you go through...
 - The problem you have may just not have been solved just yet!

The brief (for the Practical)

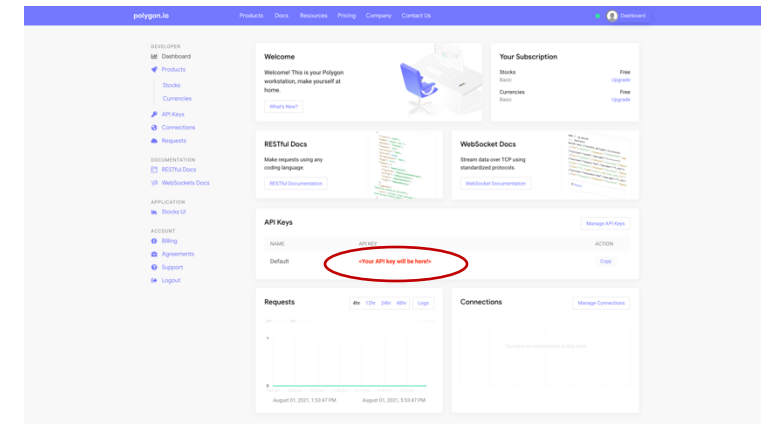
- The FAANG group of companies are considered the main players in the industry of 'big tech' that influences much of business and leisure.
- Respectively, the letters in the abbreviation refer to the following companies with the following stock symbols (among others) traded on the NASDAQ exchange:
 - **F**acebook's parent company Meta (NASDAQ:META);
 - **A**mazon (NASDAQ:AMZN);
 - **A**pple (NASDAQ:AAPL);
 - **N**etflix (NASDAQ:NFLX) and;
 - **G**oogle's parent company, Alphabet (NASDAQ:GOOG).

Polygon API key

Next, we will use try to use the Polygon.io REST API to gather connected data regarding stock prices and related information from the Web.

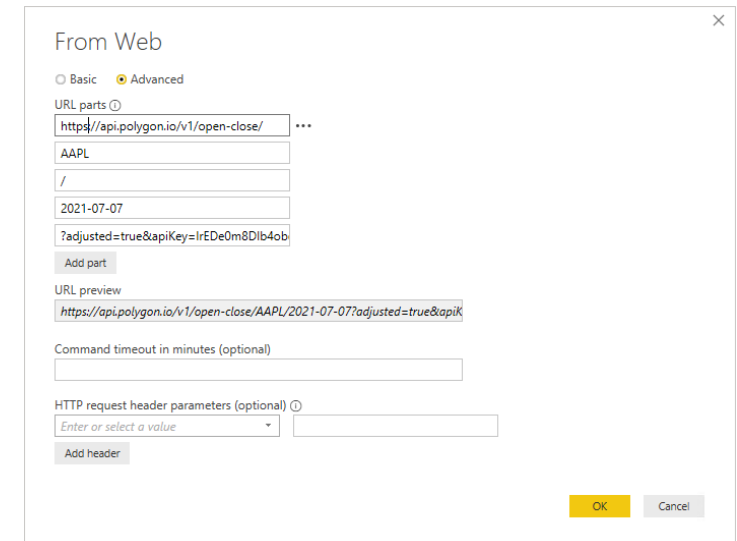
Visit the Polygon.io website link to sign up (<https://polygon.io/dashboard/signup>) which will give you an API key which is required to connect to the data source from within our Power BI document so that we can use it.

Once you sign up, you will be presented with a dashboard that shows clearly, your API key.



Connecting the data

- Now that we have an API key, we can connect to (not import) our data within Power BI Desktop, to proceed towards the visualisation and analysis of it.
- On the Power BI Desktop splash screen, select the 'Get Data' option to connect to our data source, similar to as we have done before.
- We will connect to the data as a 'Web' source (this will be changed for us in time).



From Web

☐ Basic ☒ Advanced

URL parts ⓘ

https://api.polygon.io/v1/open-close/ ...

AAPL

/

2021-07-07

?adjusted=true&apiKey=IrEDe0m8D1b4ob

Add part

URL preview

https://api.polygon.io/v1/open-close/AAPL/2021-07-07?adjusted=true&apiK

Command timeout in minutes (optional)

HTTP request header parameters (optional) ⓘ

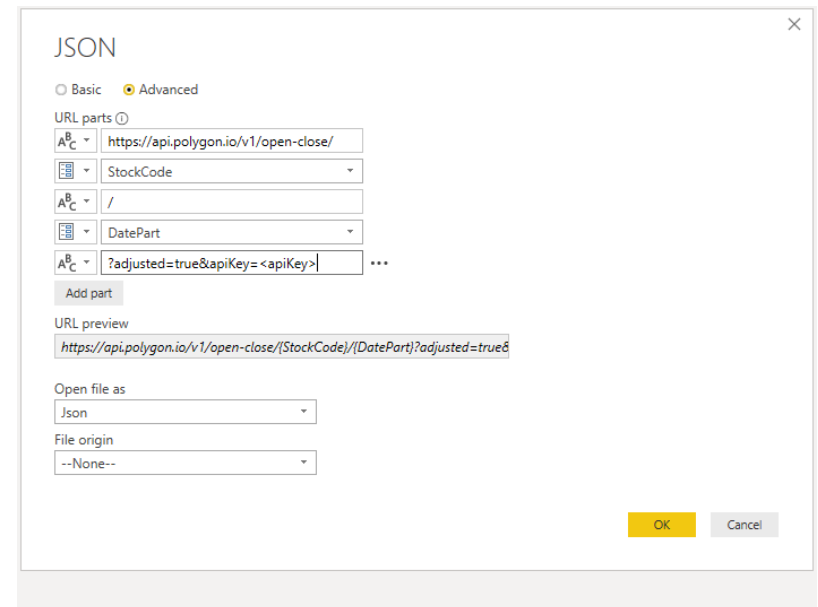
Enter or select a value

Add header

OK Cancel

Generalising the data

- Instead of going through and doing each request manually, from the 'Home' tab, we will click 'Manage Parameters' and then continue on...
- This will bring up a window where we can configure the parameter values.
- Click 'New' and name the parameter '**StockCode**'.
- It will be of type 'Text' with suggested values of the format 'List of values'.
- Enter the five stock codes listed above in the box:
META, AMZN, AAPL, NFLX and **GOOG...**



The screenshot shows a 'JSON' configuration window with a close button (X) in the top right corner. It has two tabs: 'Basic' and 'Advanced', with 'Advanced' selected. Under 'URL parts', there are four rows, each with a dropdown menu and a text input field. The first row has a dropdown with 'A^BC' and a text field containing 'https://api.polygon.io/v1/open-close/'. The second row has a dropdown with a grid icon and a text field containing 'StockCode'. The third row has a dropdown with 'A^BC' and a text field containing '/'. The fourth row has a dropdown with a grid icon and a text field containing 'DatePart'. Below these is a fifth row with a dropdown with 'A^BC' and a text field containing '?adjusted=true&apiKey=<apiKey>' followed by three dots. There is an 'Add part' button below the fifth row. Under 'URL preview', the text 'https://api.polygon.io/v1/open-close/{StockCode}/{DatePart}?adjusted=true&' is shown. Below that is an 'Open file as' section with a dropdown menu showing 'Json'. Below that is a 'File origin' section with a dropdown menu showing '--None--'. At the bottom right are 'OK' and 'Cancel' buttons.

Creating functions

- We then need to extract only the information that we need and transform it into a useful format for analysis and visualisations.
 - First, we use the 'Convert' tab to convert each of these queries to tables.
 - Next, we use the 'Transpose' option from under 'Transform' and then 'Use First Row as Headers'. This will start to look more familiar.
- If your data looks like the example in the screenshot (note the shape and headers) seen on the slide below, then Power BI has done it automatically for us already.

Close & Apply

New Source

Recent Sources

Enter Data

Data source settings

Manage Parameters

Refresh Preview

Manage

Choose Columns

Remove Columns

Keep Rows

Remove Rows

Sort

Split Column

Group By

Replace Values

Combine Files

Azure Machine Learning

Close

New Query

Data Sources

Parameters

Query

Manage Columns

Reduce Rows

Transform

Combine

AI Insights

Queries [1]

CompanyInfo

fx

Table.TransformColumnTypes(#"Converted to Table",{{"logo", type text}, {"listdate", type date}, {"cik", Int64.Type}, {"bloomberg", type text}, {"figi", type any}, {"lei", type text}, {"sic", type text}, {"country", type text}, {"industry", type text}})

	logo	listdate	cik	bloomberg	figi	lei	sic	country	industry	
1	https://s3.polygon.io/logos/aapl/logo.png	2/01/1990	320193	EQ0010169500001000		null	HWUPKROMPOU8FGXBT394	3571	usa	Computer Harc

Query Settings

PROPERTIES

Name

CompanyInfo

All Properties

APPLIED STEPS

Source

Converted to Table

Changed Type

Creating tables

- We must then use the 'Enter Data' tool under 'Home' to create two new tables.
 - One named '**Stocks**' which contains the five stock codes in one column and;
 - Another named '**Dates**' which contains the date codes in one column, ensuring again that you enter them in the correct format (such as **2022-09-30**).
- Make sure these columns are in 'Text' format.
 - Use the 'Transform' bar if needed or 'step back' in the 'Query Settings'.
- Now comes the hard part, where we invoke various functions to build our data set.

Real-world problems

- There will probably be some errors as we are requesting too much data.
 - Well, we are actually requesting data too quickly – after all, it is a free services!
 - However, if we expand this new column, we will see that in the background the request has been retried and the data is now likely there for us to peruse.
 - We can now rename the expanded columns and remove some of the columns.
 - Then, we can rename our query / table – to what we actually need it to be.

Creating the report

- Once your data model has been created and populated, it is now time to create the reports and/or dashboards as suggested above in the brief.
 - Consider the appropriate number of each of these and the appropriate visuals to use within each of them.
- Once you have created the dashboards and reports, reflect upon them.
 - Could you improve how the data is presented?
 - Is the data misleading (due to the way it has been presented) in any way?
 - You can now answer the questions in the brief.

The End: Thank You

Any Questions? Ask via email (tristan.reed@uwa.edu.au)