# Extract – Transform – Load

Project Proposal

## Introduction

The National Disability Insurance Scheme provides supports to over half a million individuals with disabilities. The scheme has an annual budget of over 15 billion dollars and services are provided by almost ten thousand service providers across Australia.

## Data Sources

The NDIS publishes de-identified data on participants, providers, market conditions and utilisation of plan funding on a quarterly basis.

This data is available on their [website](https://data.ndis.gov.au/explore-data).

This data can be filtered on the NDIS website and/or downloaded in a CSV format. There is also an [archive](https://data.ndis.gov.au/data-downloads) of previous datasets in a CSV format. There is also documentation on the data rules applied to the datasets.

This project will use all four main datasets (participants, providers, market conditions and utilisation) filtered to Western Australia only.

## Project Aim

The aim for this project will be to build a relational database in SQL of these four datasets using all available longitudinal data.

The secondary aim will be to develop a Jupiter notebook that can expediate the process of adding future NDIS data releases to this SQL database.

The expectation is that this database could serve as a useful data source for business intelligence for a disability service organisation.

## ETL Process

The data extraction process will in the first instance involve downloading the available CSVs. Where possible the data will be filtered prior to downloading the CSV. (This appears only to be an option with current data, and not the archived data.)

The data transformation process will include the following steps using Python and Pandas:

* Filtering any unfiltered CSVs to Western Australia data only.
* Removing or addressing any null values in the datasets.
* Any further data cleaning as required.
* Adding a date column for longitudinal data, to enable the creating of a single dataframe for each category of data.
* Creating additional dataframe objects using groupby and joins to present data by service district, by disability group, by age and (where relevant) by support class.

The data loading process will involve converting and loading these dataframes into a SQL database using SQLAlchemy. The location of this database is yet to be confirmed. At this stage the preferred option is that it will be hosted on a local server only (with PostgreSQL as a front-end for example). However, there are other options that will be considered including hosting the database on an Azure SQL server (for demonstration purposes only, choosing the free and temporary option only).

## Documentation

As part of this project, the following documentation will be developed:

* A technical report detailing the ETL process.
* An ERD for the database created.