

Unlocking the Power of Data:

Streamlined ETL Solutions for the Digital Anthroposphere

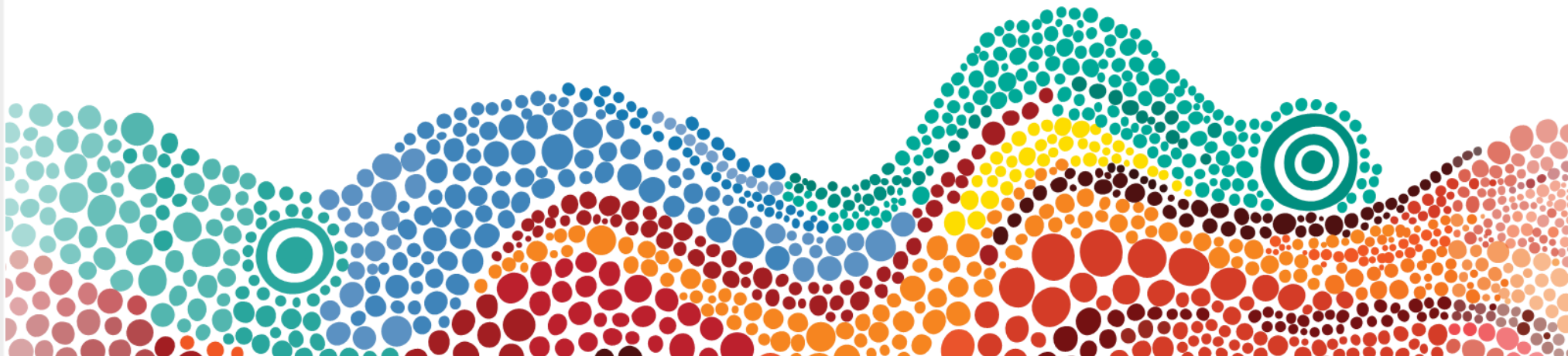
eResearch conference, Brisbane

Presenter: Dr. Ali Asghari



Acknowledgement of Country

We acknowledge the Traditional Owners of the land on which this event is taking place and pay respect to their Elders (past and present) and families.



AURIN: An Introduction

What is AURIN?



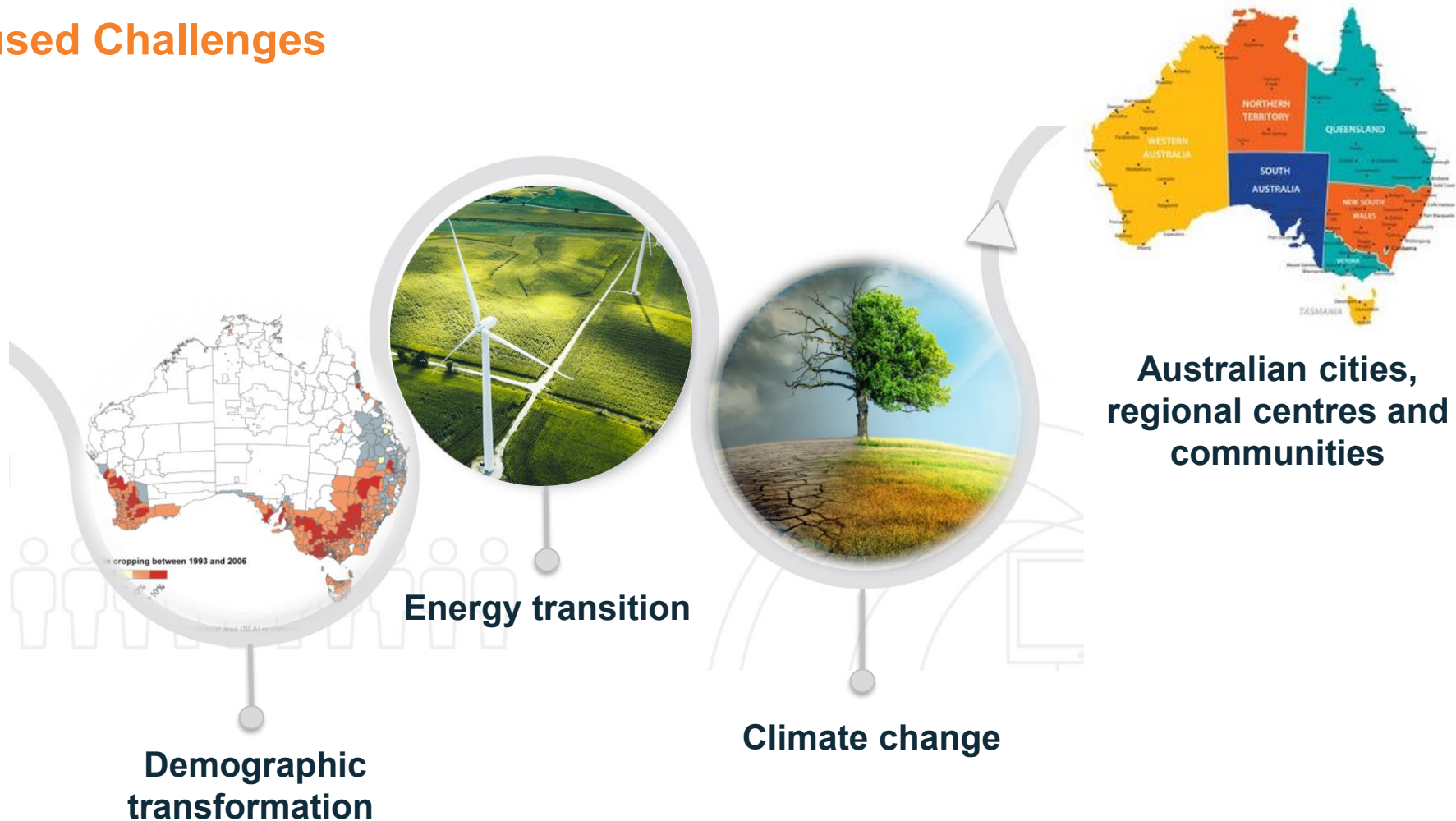
Australian Urban **R**esearch
Infrastructure **N**etwork



National Collaborative
Research **I**nfrastructure
Strategy

AURIN: Key Focused Challenges

Key Focused Challenges



AURIN: Key Strategic Areas

Urban
DaaS

Urban Data-as-a-Service

Unlocking, generating, curating, cataloguing, and sharing high value and hard-to-get data for urban and regional research and planning.

Urban
AaaS

Urban Analytics-as-a-Service

Designing, developing and maintaining a modular, cloud-first digital research infrastructure to support urban and regional analytics and modelling.

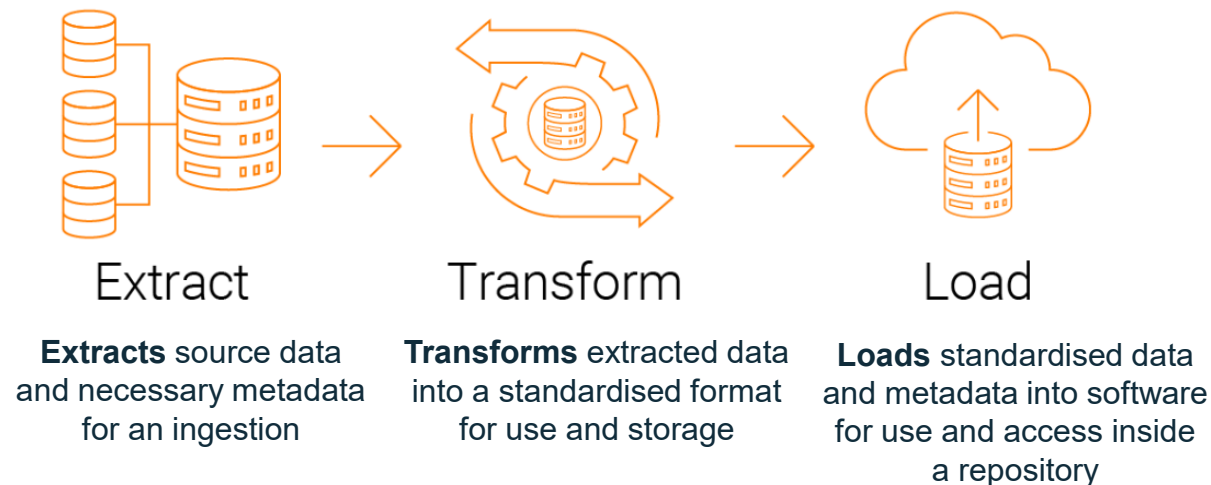
UDT

Foundations for Australian Urban Digital Twins

Understanding the needs of UDTs to inform decisions and positively impact lives. Applying, testing, prototyping and sharing best practice standards, semantics, workflows and cutting-edge technologies for use in UDTs.

How does AURIN Manage and publish data?

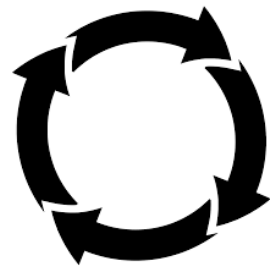
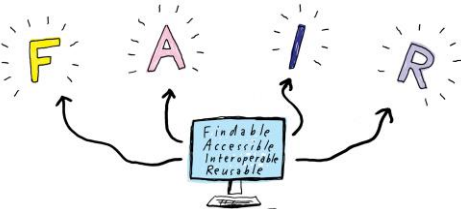
- AURIN Manages and publishes data using the AURIN ETL pipeline
- Allow to standardise highly valuable and hard-to-get data from a vast range of different data sources (organisations)
- Publish that data in a consistent and machine-readable format and manage it in a modernised way.



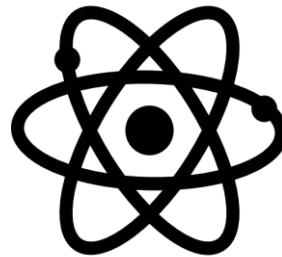
AURIN ETL Pipeline: Key Characteristics

Why is AURIN ETL Pipeline developed?

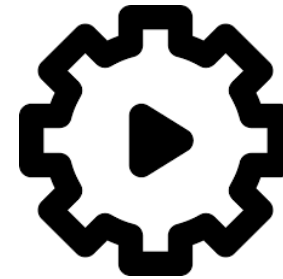
To enable the creation of **F A I R** data products in a reproducible, atomic, automated, and flexible manner



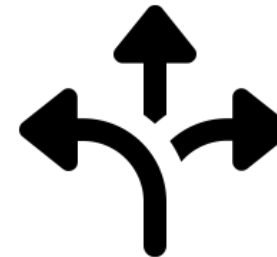
Reproducibility



Atomicity

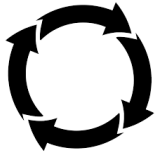


Automation



Flexibility

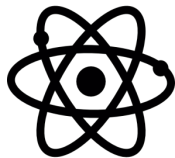
AURIN ETL Pipeline: Key Characteristics



Reproducibility

Definition: The ability to recreate any version of a data product at any time, involving documentation and preservation of the data's sources and generation processes.

Importance: Crucial for **data integrity and transparency**, enabling verification, validation, understanding of lineage, and independent result replication.

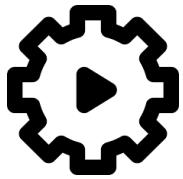


Atomicity

Definition: The concept of treating a sequence of data operations as a single, indivisible unit, where all operations succeed or fail together to prevent partial or inconsistent data states.

Importance: Vital for preserving **data integrity and consistency**, ensuring well-structured data for easier access, analysis, and maintenance.

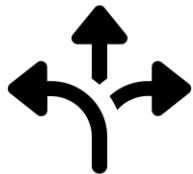
AURIN ETL Pipeline: Key Characteristics



Automation

Definition: Ability to conduct data ingestion processes with minimal manual intervention using scripts, automation tools, or workflows to automate data collection, transformation, and loading tasks.

Importance: Essential for improving **efficiency and scalability** while speeding up data processing and reducing errors.



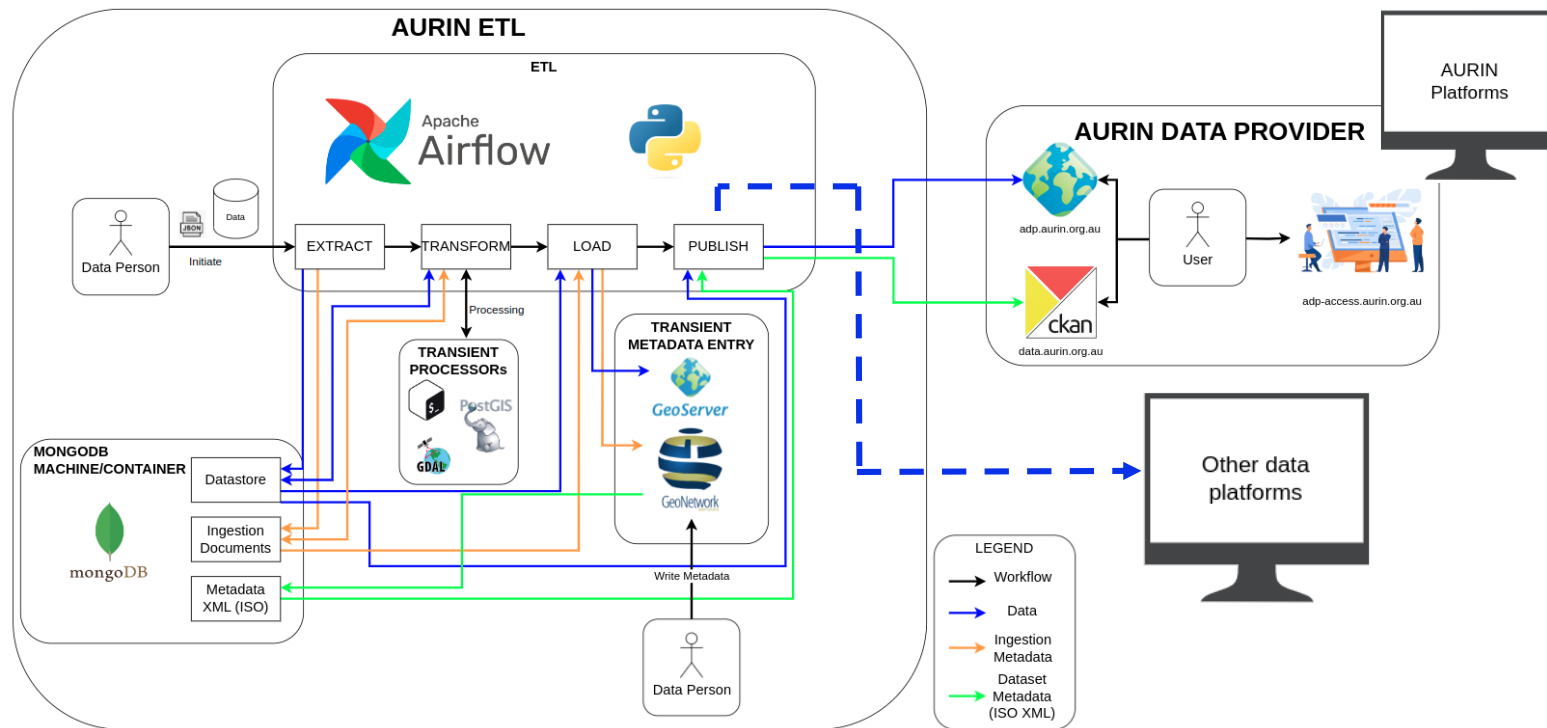
Flexibility

Definition: Ability to adapt to various data formats, schemas, and sources, producing a controlled and standardised output.

Importance: Ensures **interoperability and adaptability** to changing data sources over time.

AURIN ETL Pipeline: Architecture

AURIN ETL Architecture



AURIN ETL Pipeline: Applications and technologies

Infrastructure and DevOps



nectar



docker

Geospatial Data Server



GeoServer

Workflow Management



Apache
Airflow



python™

Database Management System



PostgreSQL



mongoDB.

Metadata Management



GeoNetwork
open source



Urban Data-as-a-Service: AURIN Data Provider (ADP)

AURIN Data Catalogue

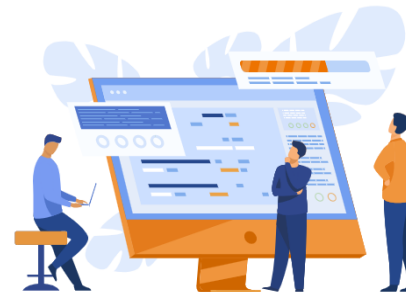
Search data

Try searching by:

- property
- urban
- human society
- health sciences
- parks
- built environment
- employment
- transport
- infrastructure
- income
- indigenous studies
- bicycle
- demographics
- migration

5081
datasets

106
organisations



AURIN Data Provider

AURIN - National Education Facilities - TAFEs (Point) 2018

This dataset presents the campus locations of technical and further education (TAFE) institutions in Australia. The National Education Dataset has been developed by AURIN. Data was collected and validated from September 2018 to November 2018. Educational institution names and addresses from the Department of Education and Training were extracted and converted into their respective spatial location using the AURIN Portal Geocoder Tool. The coordinates have been verified through comparison with other authoritative datasets.

For further information about the resources used in the creation of this dataset, please visit:

- [PSMA Geocoder](#).
- [Commonwealth of Australia - Department of Education and Training](#).

WFS Data URL

https://adp.aurin.org.au/geoserver/wfs?request=GetFeature&typename=datasource-AURIN-UoM_AURIN_DB:aurin_national_education_dataset_tafes_2018

Format

GeoJSON

Download

Login

If you're having issues logging in, click here

More information:

- [AURIN Training](#)

What are the future directions for AURIN?

- Focusing on establishing the foundations for Australia's Urban Digital Twins, to address climate change, energy transition, and changing demographics.
- Analysing the requirements of UDT applications and looking to fulfil them wherever possible with data, analytics, workflows, semantic solutions, etc.
- Providing diverse and reliable datasets for researchers will allow them to address more urban challenges and find practical solutions to those challenges.
- Improving the searchability of our AURIN Data Catalogue (metadata enrichments) will help researchers to better find their desired datasets
- Developing data validation methods to make sure the data quality
- Providing new data formats including raster, real-time and 3D datasets for researcher to broaden their horizon into new research ideas and address new challenges

THANK YOU!

COME SEE US AT THE AURIN BOOTH

Dr. Ali Asghari
Data Engineer
E: asgharia@unimelb.edu.au
www.aurin.org.au
 [@aurin_org_au](https://twitter.com/aurin_org_au)

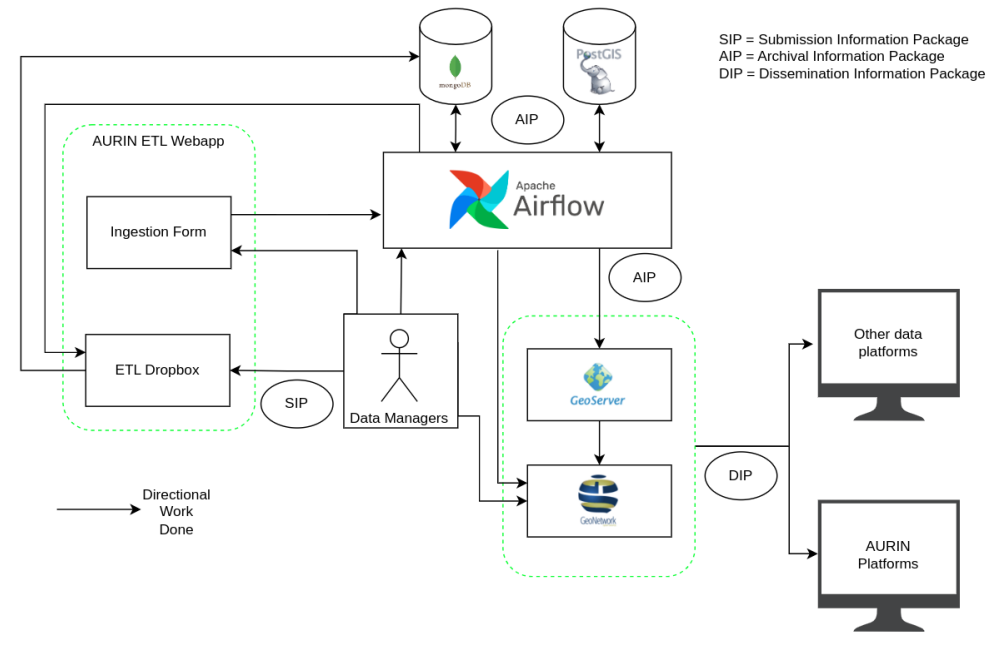


aurin.org.au

AURIN ETL Pipeline: Workflow

How it works – Workflow

Workflow for an AURIN ETL data ingestion



AURIN ETL Pipeline: Workflow

How it works – Develop the workflow

DAG Descriptor

```
File Edit Selection View Go Run Terminal Help
au_govt_dese-lm_salm.py - dag_creation - Visual Studio Code

EXPLORER
DAG_CREATION
├── all_mobility_australia.json
├── au_govt_abs_data_by_region.json
├── au_govt_abs_personal_income.json
├── au_govt_dese_lm_salm.json
├── etl_demo2.json
├── ingest_accesses.json
├── ingest_geometries.json
├── ingest_licences.json
├── ingest_organisations.json
├── legacy_ingestion.json
├── README.md
├── wfs_ingestion.json
├── dags
│   ├── __init__.py
│   ├── airflowignore
│   ├── access_ingestion.py
│   ├── all_mobility_australia.py
│   ├── apm-dag.py
│   ├── au_govt_abs_data_by_region.py
│   ├── au_govt_abs_personal_income.py
│   └── au_govt_dese_lm_salm.py
├── data_preload.py
├── data_publish.py
├── data_unpublish.py
├── geometry_ingestion.py
├── legacy_ingestion.py
├── licence_ingestion.py
├── metadata_down.py
├── metadata_up.py
├── organisation_ingestion.py
├── resource_up.py
├── upload_extract_files.py
├── wfs_ingestion.py
├── docs
├── geo
├── __pycache__
├── __init__.py
├── Geoserver.py
├── Style.py
└── ...

data-etl > dataetl > dags > au_govt_dese-lm_salm.py > Jupyter > au_govt_dese_lm_salm
1 from airflow.decorators import dag, task
2 from airflow.operators.python import get_current_context
3 from airflow.operators.trigger_dagrun import TriggerDagRunOperator
4 from airflow.utils.dates import days_ago
5 from dataetl import db, extract, load, transform
6
7 args = {'owner': 'ali', 'start_date': days_ago(1)}
8
9
10 @dag(dag_id='au_govt_dese_lm_salm',
11      default_args=args,
12      schedule_interval=None,
13      start_date=days_ago(1),
14      tags=['data-etl', 'DESE', 'au_govt_dese'])
15 def au_govt_dese_lm_salm():
16     """Transforms data in the AU Govt DESE - Labour Markets collections
17     Requires a MongoDB extract document ObjectId string as input.
18     Notes
19     -----
20     Trigger with the following config:
21     {
22     "extract_oid": "61ad4075707c22311e7be8a5"
23     }
24     """
25     @task(task_id="transform")
26     def transform_sheets():
27         context = get_current_context()
28         print(context['dag_run'].conf)
29         extract_oid = context['dag_run'].conf['extract_oid']
30
31         m = db.MongoDB() # Set up Mongo connection
32         # Choose the DAG document by name
33         m.dag.choose('au_govt_dese_lm_salm', by="name")
34         m.ingestion.choose(extract_oid, 'extract') # Choose the extract doc
35
36         # Get extract file IDs
37         extract_file_ids = m.ingestion.get_file_ids()
38
39         # Create a new transform doc for the ingestion
40         transform_id = m.ingestion.new('transform')
41
42         # Set up a Postgres connection.
```

DAG Document

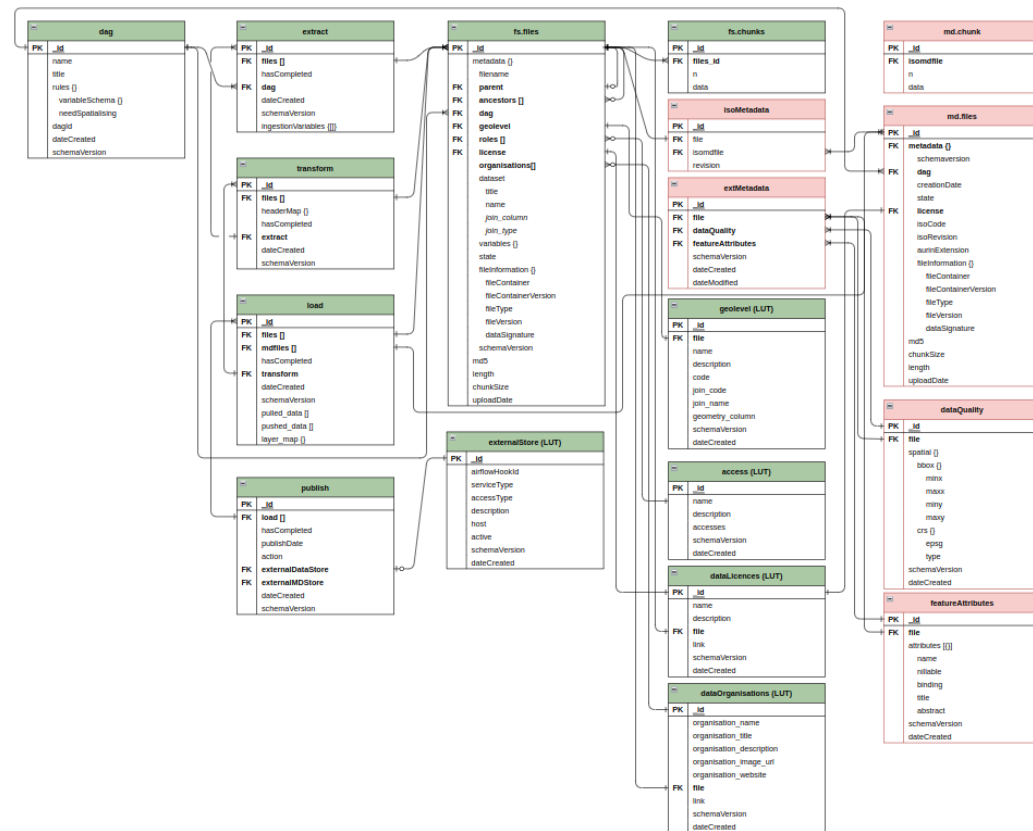
```
EXPLORER
DAG_CREATION
├── __init__.py
├── airflowignore
├── access_ingestion.py
├── all_mobility_australia.py
├── apm-dag.py
├── au_govt_abs_data_by_region.py
├── au_govt_abs_personal_income.py
├── au_govt_dese_lm_salm.json
├── etl_demo2.json
├── ingest_accesses.json
├── ingest_geometries.json
├── ingest_licences.json
├── ingest_organisations.json
├── legacy_ingestion.json
├── README.md
├── wfs_ingestion.json
├── dags
│   ├── __init__.py
│   ├── airflowignore
│   ├── access_ingestion.py
│   ├── all_mobility_australia.py
│   ├── apm-dag.py
│   ├── au_govt_abs_data_by_region.py
│   ├── au_govt_abs_personal_income.py
│   ├── au_govt_dese_lm_salm.py
│   ├── data_preload.py
│   ├── data_publish.py
│   ├── data_unpublish.py
│   ├── geometry_ingestion.py
│   ├── legacy_ingestion.py
│   ├── licence_ingestion.py
│   ├── metadata_down.py
│   ├── metadata_up.py
│   ├── organisation_ingestion.py
│   ├── resource_up.py
│   ├── upload_extract_files.py
│   ├── wfs_ingestion.py
│   ├── docs
│   ├── geo
│   ├── __pycache__
│   ├── __init__.py
│   ├── Geoserver.py
│   ├── Style.py
│   └── ...

data-etl > dataetl > dagdocs > au_govt_dese_lm_salm.json > {} rules: {} variableschema > {} items > {} properties > {} multihdr_start > {} type
1 {
2   "title": "AU Govt DESE - Labour Markets",
3   "name": "au_govt_dese_lm_salm",
4   "dagId": "au_govt_dese-lm_salm",
5   "rules": {
6     "variableschema": {
7       "title": "AU Govt DESE - Labour Markets",
8       "type": "array",
9       "description": "Metadata for each file in the AU Govt DESE - Labour Markets DAG",
10      "items": {
11        "type": "object",
12        "required": [
13          "filename",
14          "ags column",
15          "multihdr start",
16          "multihdr end",
17          "role",
18          "licence",
19          "organisation",
20          "sheets"
21        ],
22        "properties": {
23          "filename": {
24            "type": "string",
25            "description": "The name of the original file"
26          },
27          "ags column": {
28            "type": "string",
29            "description": "The name of the column defining the column with either the ASOS code or ASOS name",
30            "default": "Region Name"
31          },
32          "aggregation type": {
33            "type": "string",
34            "description": "It defines the type of aggregation whether it is based on code or name",
35            "default": "code"
36          },
37          "multihdr start": {
38            "type": "number",
39            "description": "The row number defining the start of the data header",
40            "default": 3
41          },
42          "multihdr end": {
43            "type": "number",
44            "description": "The row number defining the end of the data header",
45            "default": 3
46          }
47        }
48      }
49    }
50  }
```

AURIN ETL Pipeline: Workflow

How it works – Register the workflow

MongoDB Schema



AURIN ETL Pipeline: Workflow

How it works – Populate information required for the workflow

Ingestion Form

× au_govt_dese_lm_salm | 647e834cef032056cea8c0d7

http://localhost:1263/jsonserver/api/jsons/647e834cef032056cea8c0d7

Background:

1. Sheet Information Add Delete

* filename: SALM_Smoothed_SA2_Datafiles_ASGS_2016_-_September_quarter_2021.xlsx

* asgs_column: SA2 Code (2016 ASGS)

aggregation_type: code

* multihheader_start: 3

* multihheader_end: 3

* role: ROLE_ACCEPTED_TOU

* licence: CC-BY-4.0

* organisation: au_govt_dese

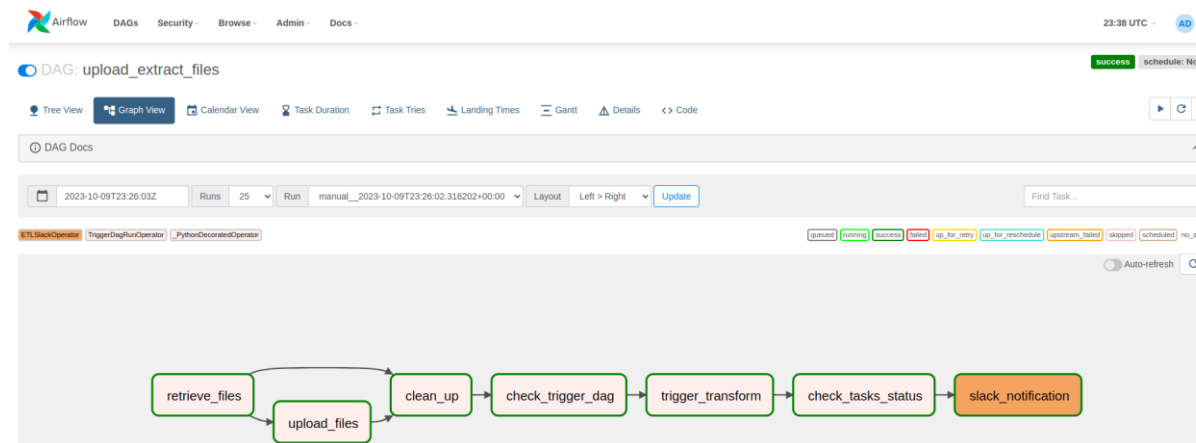
Array that sets out the variables per sheet within a file Add

* title:	DESE - SALM - Smoothed Labour Force (SA2) Q4 2010 - Q3 2021
* table_name:	dese_salm_sa2_asgs_2016_sep_qrt_2021_smhd_sa2_lbr_frc
* sheet_name:	Smoothed SA2 labour force
* geolevel:	sa2_2016_aust
* aggregations:	sa2
Delete	
* title:	DESE - SALM - Smoothed Unemployment (SA2) Q4 2010 - Q3 2021
* table_name:	dese_salm_sa2_asgs_2016_sep_qrt_2021_smhd_sa2_unemp

AURIN ETL Pipeline: Workflow

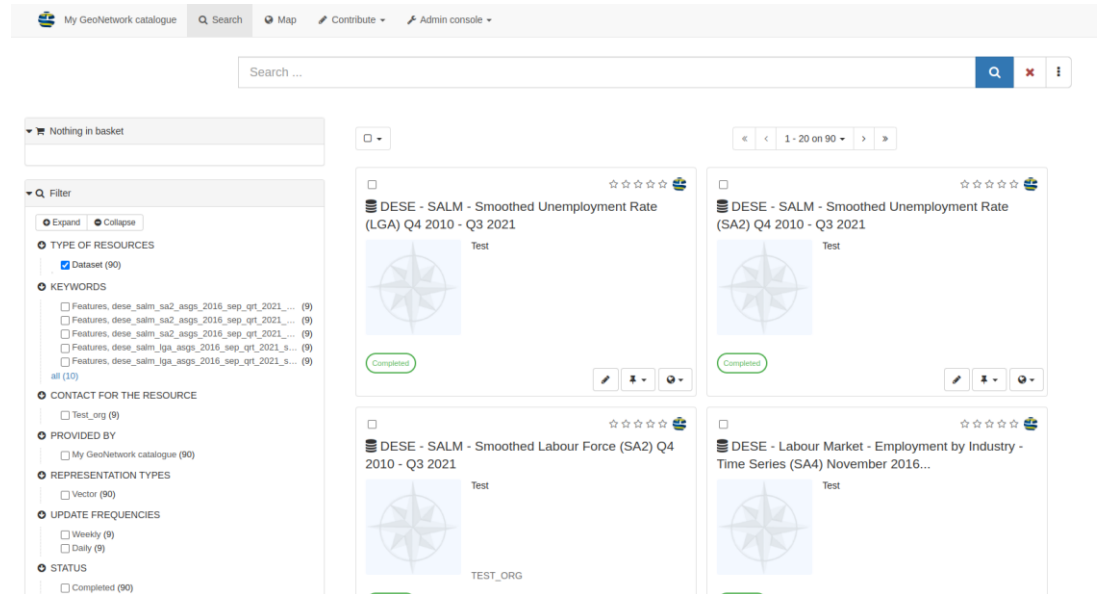
How it works – Run the workflow

Running DAGs



How it works – Populate and/or enrich the metadata

GeoNetwork standardised template




The screenshot displays the GeoNetwork catalogue interface. At the top, there is a navigation bar with 'My GeoNetwork catalogue', a search bar, and links for 'Map', 'Contribute', and 'Admin console'. Below the navigation bar is a search input field with the text 'Search ...'. On the left side, there is a filter panel with a 'Filter' section. The filter panel includes a 'Nothing in basket' notification and a 'Filter' section with 'Expand' and 'Collapse' buttons. The filter section is divided into several categories: 'TYPE OF RESOURCES' (with 'Dataset (90)' selected), 'KEYWORDS' (with several checkboxes for features), 'CONTACT FOR THE RESOURCE' (with 'Test_org (9)' selected), 'PROVIDED BY' (with 'My GeoNetwork catalogue (90)' selected), 'REPRESENTATION TYPES' (with 'Vector (90)' selected), 'UPDATE FREQUENCIES' (with 'Weekly (9)' and 'Daily (9)' selected), and 'STATUS' (with 'Completed (90)' selected). The main content area shows a grid of search results. The first two results are 'DESE - SALM - Smoothed Unemployment Rate (LGA) Q4 2010 - Q3 2021' and 'DESE - SALM - Smoothed Unemployment Rate (SA2) Q4 2010 - Q3 2021'. The third result is 'DESE - SALM - Smoothed Labour Force (SA2) Q4 2010 - Q3 2021'. The fourth result is 'DESE - Labour Market - Employment by Industry - Time Series (SA4) November 2016...'. Each result card includes a compass icon, a 'Test' label, a 'Completed' status indicator, and a set of action icons (edit, share, download).

AURIN ETL Pipeline: Workflow

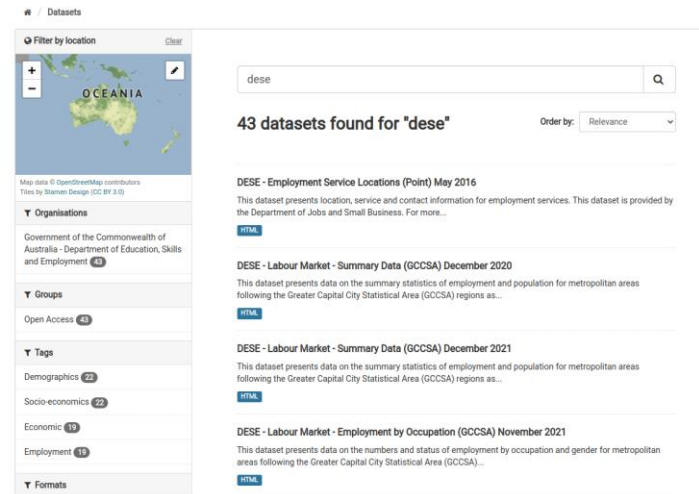
How it works – Publish the data

Publish Data into GeoServer



The screenshot shows the GeoServer web interface. On the left is a navigation menu with categories like About & Status, Data, Services, Settings, The Caching, Security, and Tools. The main content area is titled 'Layers' and contains a table of published layers. The table has columns for Type, Title, Name, Store, Enabled, and Native SRS. The layers listed include various geographic data such as 'World rectangle', 'Manhattan (NY) points of interest', 'Manhattan (NY) landmarks', 'Manhattan (NY) roads', 'A sample ArcGIS file', 'North America sample imagery', 'North America sample imagery', 'mosaic', 'USA Population', 'Tasmania cities', 'Tasmania roads', 'Tasmania state boundaries', 'Tasmania water bodies', 'Spearfish archaeological sites', 'Spearfish bug locations', 'Spearfish restricted areas', 'Spearfish roads', 'Spearfish elevation', 'Spearfish streams', 'NSWTrains_yearly_flow_2020', 'building_heights', 'crime_statistics_2020', 'crime_statistics_2021', 'lsm_dem_1arc_10m_offline_2010_11_2a', and 'new_public_housing_allocations_2012_2013'.

Publish Data into CKAN



The screenshot shows the CKAN web interface. At the top, there's a search bar with the text 'dese' and a search icon. Below the search bar, it says '43 datasets found for "dese"'. To the right of this text is a dropdown menu for 'Order by' set to 'Relevance'. Below the search results, there are sections for 'Organisations', 'Groups', 'Tags', and 'Formats'. The 'Organisations' section shows 'Government of the Commonwealth of Australia - Department of Education, Skills and Employment'. The 'Tags' section shows 'Demographics (22)', 'Socio-economics (22)', 'Economic (17)', and 'Employment (13)'. The 'Formats' section shows 'HTML'.