Australian Company Data Pipeline Documentation

Susanta Baidya

March 2, 2025

Abstract

This document describes the implementation of a data pipeline that integrates Australian company website data from Common Crawl with business information from the Australian Business Register (ABR). The system collects, cleans, and stores data in a MySQL database, enabling comprehensive business analysis.

1 Pipeline Architecture

The data processing workflow follows three key stages:

- **Data Ingestion**: Python scripts extract raw data from Common Crawl (200k+ URLs) and ABR
- **Transformation**: DBT models clean, normalize, and merge datasets
- **Storage**: Processed data stored in MySQL with optimized schema
- **Validation**: Automated data quality checks using DBT tests

2 Implementation Details

2.1 Data Collection

```python
# Common Crawl URL processing
def process_url(url):
    try:
        response = requests.get(url, timeout=10)
        soup = BeautifulSoup(response.text, 'html.parser')
        return extract_metadata(soup)
    except Exception as e:
        log_error(url, str(e))
        return None
```

2.2 Data Cleaning

DBT staging model for ABR data:

```sql
{{ config(materialized='view') }}

SELECT
    abn,
    TRIM(company_name) AS company_name,
    entity_type,
    state,
    postcode,
    registration_date
FROM {{ source('raw', 'abr_raw') }}
WHERE LENGTH(abn) = 11
```

3 Database Schema

Optimized MySQL table structure:

```sql
CREATE TABLE companies (
    company_id INT AUTO_INCREMENT PRIMARY KEY,
    url VARCHAR(512) NOT NULL UNIQUE,
    company_name VARCHAR(255) NOT NULL,
    industry VARCHAR(255),
    abn CHAR(11),
    entity_type VARCHAR(50),
    state CHAR(3),
    postcode CHAR(4),
    registration_date DATE
);
```

4 Data Validation

Implemented quality checks:

- **Uniqueness**: Enforced through database constraints
- **Completeness**: NULL checks for mandatory fields
- **Consistency**: ABN format validation (11 digits)
- **Referential Integrity**: Relationship between tables

5 Example Queries

Top industries analysis:

```sql
SELECT industry, COUNT(*) AS company_count
FROM companies
GROUP BY industry
ORDER BY company_count DESC
LIMIT 10;
```

Recent registrations:

```sql
SELECT company_name, registration_date
FROM companies
WHERE registration_date > '2023-01-01'
ORDER BY registration_date DESC;
```

References

- Common Crawl: `https://commoncrawl.org`
- ABR Dataset: `https://data.gov.au`
- DBT Documentation: `https://docs.getdbt.com`
- MySQL Documentation: `https://dev.mysql.com/doc`