Apache Airflow

Orchestrates pipeline task flow & schedules batch extractions

Scrapy Engine

AWS S3 Bucket

Apache Spark Cluster

Web Application

Relational Database

supabase

PostgreSQL

Frontend

NEXT.js

Users

Web Application

Database feeds visualization tools and user queries.

Supabase database and relational data schema is optimized for web application traffic.

Multi-threaded Requests

Uploads

< raw JSON >

< CSV Files >

Requests data batch

< raw JSON >

APACHE Spark

APACHE Spark

APACHE Spark

Queries

Data

< Structured CSV Files >

Data Warehouse

Snowflake

< Structured CSV Files >

Analytics

Warehouse data schema is optimized for extremely large, relatively infrequent queries.

Analysts can complete large queries simultaneously without affecting web application.

Computational power scales automatically.

Large Query

Large Query

Data Analyst

Data Scientist

Analysts can still access raw data as needed.

Data Source	Extraction	transform (minimal)	Load	Transform
House of Commons Website expenditure reports.	Scrapy runs multi-threaded scrape of website	CSV data is aggregated into JSON file	Raw data dumped into staging cloud S3 bucket.	Cluster applies calculations and transforms the unstructured data into data models & relationships.
	CSV files are downloaded and sent to transform pipeline.	Remove duplicates & invalid rows, sanitize data	Raw data remains available to analysts to decouple workflows.	CSV Files are uploaded to SnowFlake or Supabase.
	Metadata is generated and attached to file downloads.	No business / relationship logic applied to data		