End-to-End Job Data Analysis, Recommendation, and Resume Customization System

Data Scraping: Rowan

DevOps : Rowan, David, Sam

Gen AI: Kevin

ETL & Reporting: Jasper

00

Introduction and Project Overview

Goal:

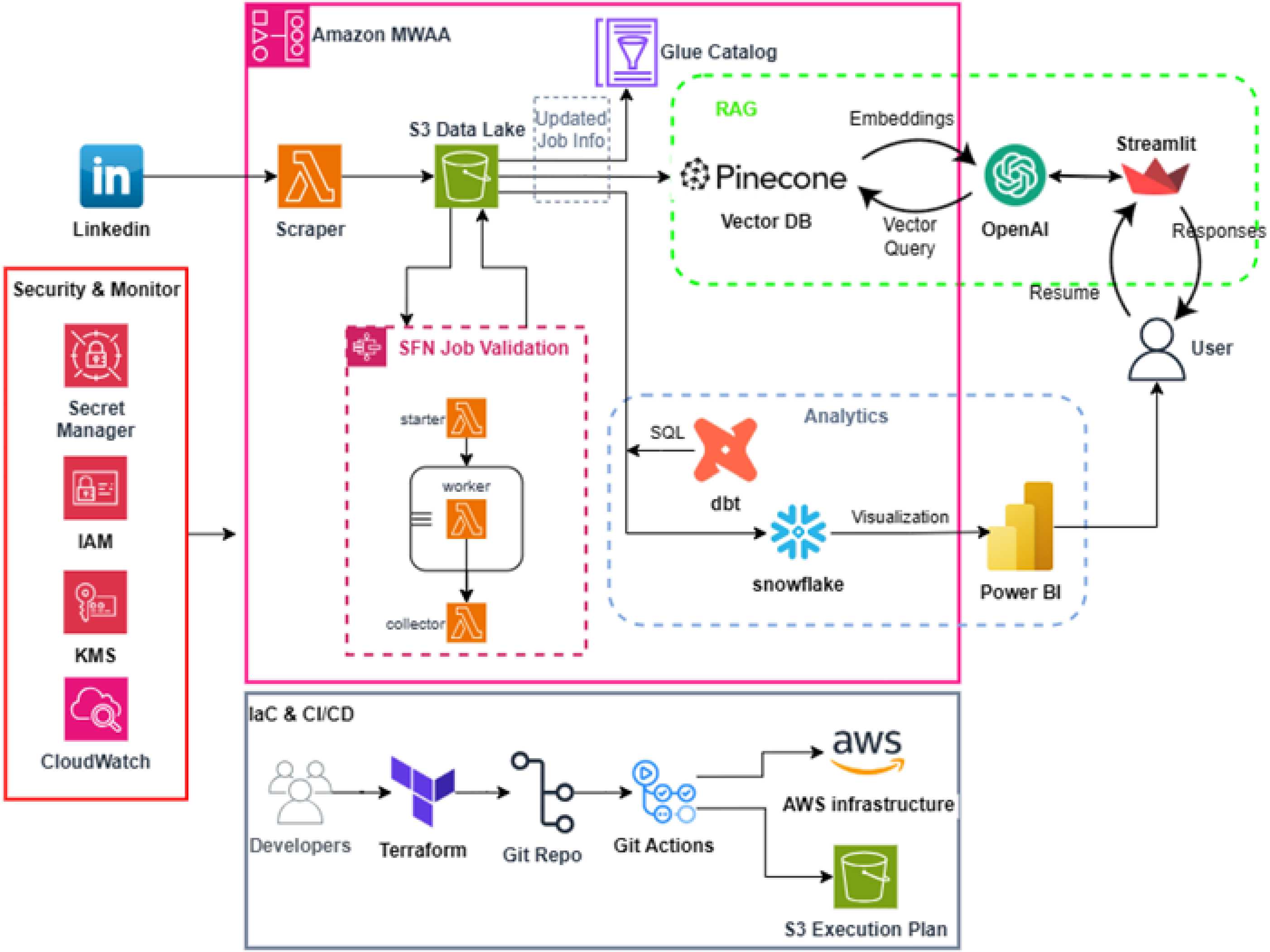
To develop an end-to-end automated system that empowers job seekers and analysts by leveraging advanced data processing, AI, and visualization technologies.

Data Lake + Gen AI + Analytics + Visualization

- Personalized Job Matching
- Customized Resumes and Cover Letters
- Job Market Analysis
- Seamless User Experience
- Interactive Visual Dashboards

Pipeline

Spark on kubernetes?



Contents

01

Data Scraping

02

Data Storage, Processing & Pipeline Orchestration

03

Gen AI

04

Data Transformation & Reporting

05

CI/CD & IaC

01

Data Scraping

LinkedIn Job Scraper Architecture

Technology Stack

- Python-based LinkedIn API
- Direct access to LinkedIn's Voyager endpoints
- AWS Lambda & S3 for deployment

Key Features

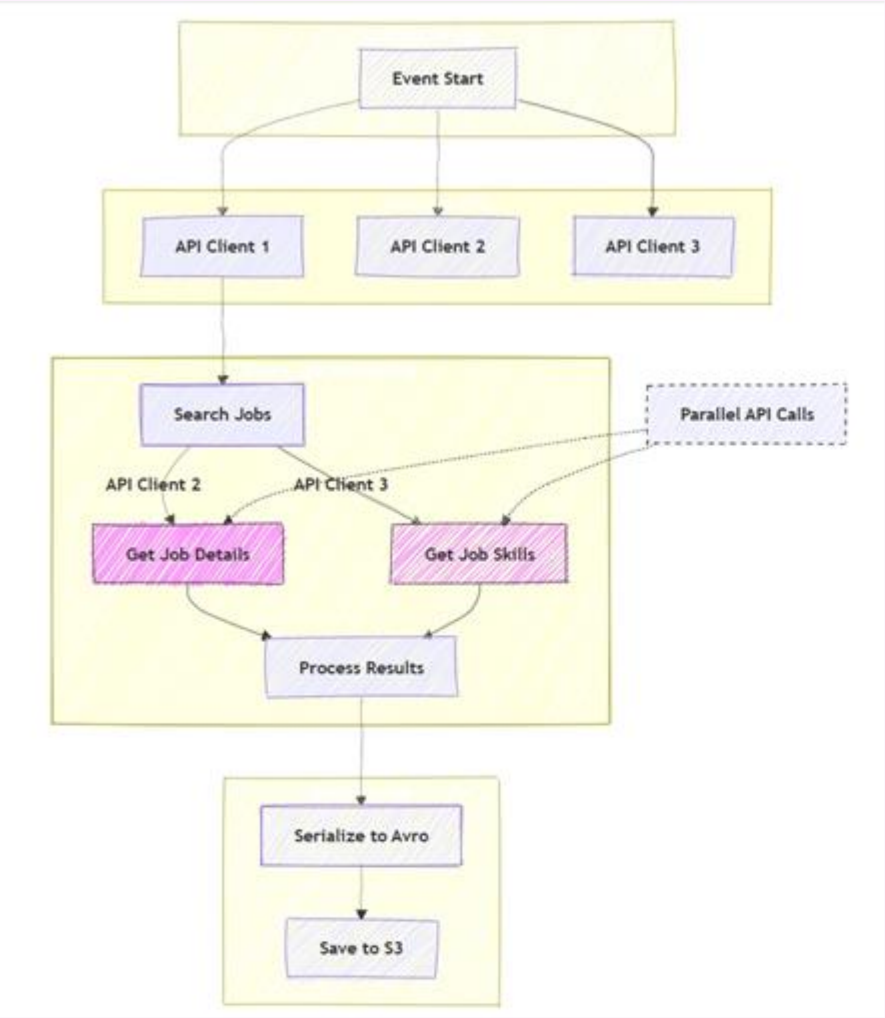
- Asynchronous processing using 3 accounts
- Rate limit handling & Data deduplication

Data Processing

- Job details and skills data extraction
- Storage in Avro format

Maintenance Note

- Cookie refresh required every 3 months
- Stored securely in S3



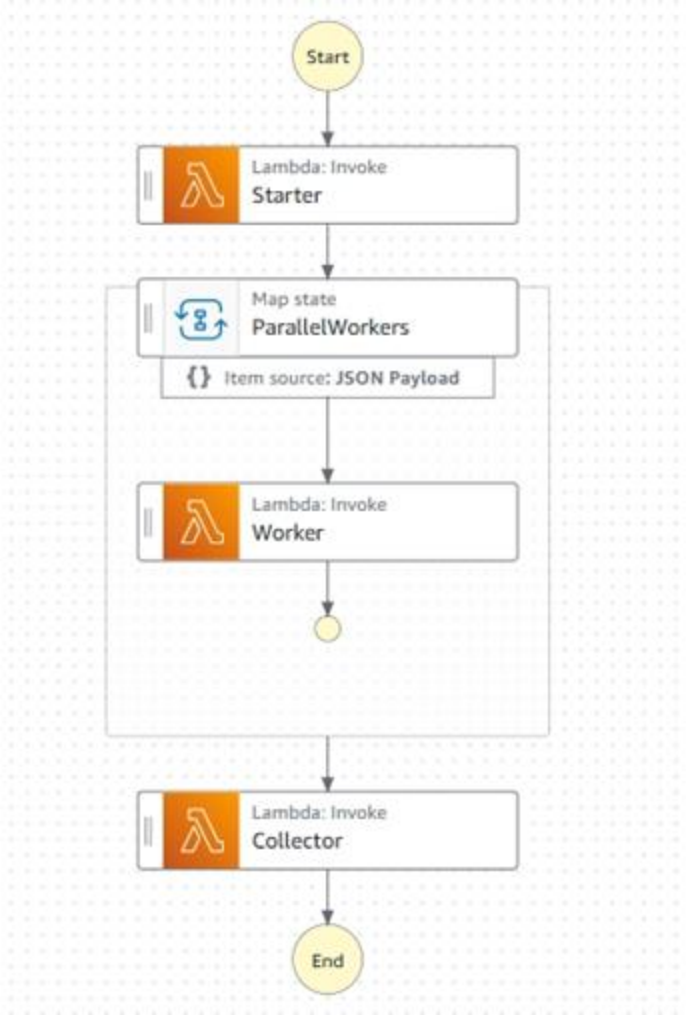
LinkedIn Job Validation

Architecture Overview

- Distributed Processing Architecture with AWS Step Functions
- Dynamic Worker Allocation
- Parallel Job Status Validation
- Automated Invalid Job Filtering

Key Benefits of Step Functions

- Native AWS Service Integration
- Automatic State Management
- Effortless to Build



02

Data Lake, Processing & Pipeline Orchestration

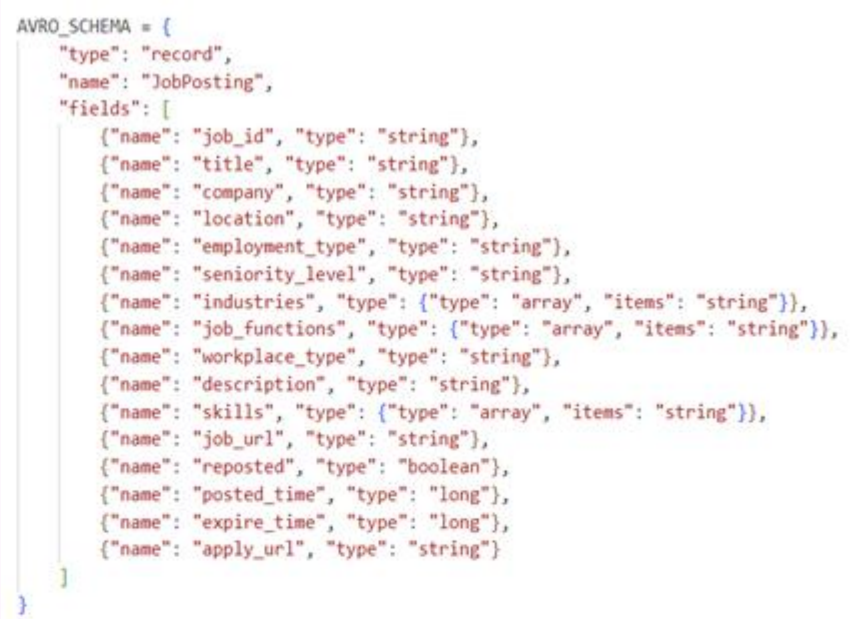
Storage

Storage Architecture

- S3 as scalable data lake
- Snowflake for data warehousing

Why Avro Format?

- Strong support for fields evolution
- Store schema with data
- Built-in schema validation
- Efficient binary compression



```
AVRO_SCHEMA = {
    "type": "record",
    "name": "JobPosting",
    "fields": [
        {"name": "job_id", "type": "string"},
        {"name": "title", "type": "string"},
        {"name": "company", "type": "string"},
        {"name": "location", "type": "string"},
        {"name": "employment_type", "type": "string"},
        {"name": "seniority_level", "type": "string"},
        {"name": "industries", "type": {"type": "array", "items": "string"}},
        {"name": "job_functions", "type": {"type": "array", "items": "string"}},
        {"name": "workplace_type", "type": "string"},
        {"name": "description", "type": "string"},
        {"name": "skills", "type": {"type": "array", "items": "string"}},
        {"name": "job_url", "type": "string"},
        {"name": "reposted", "type": "boolean"},
        {"name": "posted_time", "type": "long"},
        {"name": "expire_time", "type": "long"},
        {"name": "apply_url", "type": "string"}
    ]
}
```

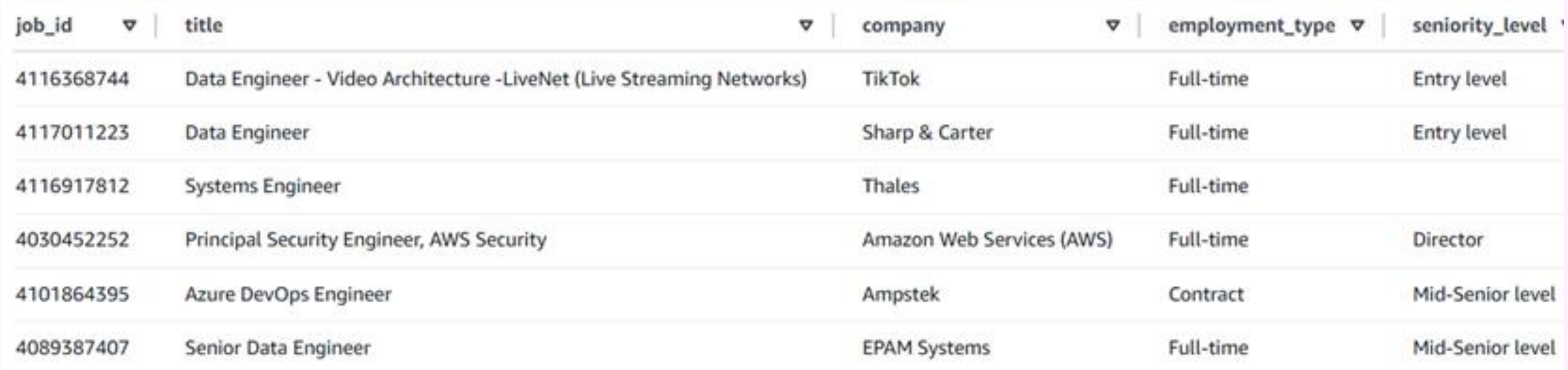
Glue Catalog

Table Configuration

- External table for Avro data
- Partitioned by year/month/day/keyword
- Managed through Terraform

Crawler Features

- Automated metadata updates
- S3 encryption enabled
- Schema change logging



job_id	title	company	employment_type	seniority_level
4116368744	Data Engineer - Video Architecture -LiveNet (Live Streaming Networks)	TikTok	Full-time	Entry level
4117011223	Data Engineer	Sharp & Carter	Full-time	Entry level
4116917812	Systems Engineer	Thales	Full-time	
4030452252	Principal Security Engineer, AWS Security	Amazon Web Services (AWS)	Full-time	Director
4101864395	Azure DevOps Engineer	Ampstek	Contract	Mid-Senior level
4089387407	Senior Data Engineer	EPAM Systems	Full-time	Mid-Senior level

Data Processing

AWS Lambda

- LinkedIn data extraction via Python
- Data storage in Avro format
- 15-minute timeout sufficient
- Automatic scaling

Benefits

Serverless • Cost-efficient • Scalable

Pipeline Orchestration

Orchestration – Automation of daily tasks

- pipeline is triggered everyday to scrap jobs posted in the last 24 hours

Airflow

- Comprehensive Workflow Management & Flexibility
- Robust Monitoring & Error Handling
- Integration with wide range of services

Step Function

- AWS native service for workflow orchestration
- Limited to AWS services primarily, less flexible for non-AWS tasks

Why Include Step Functions Within Airflow?

Problem statement: During the job validation step, we need to split the workload and invoke a dynamic number of Lambdas to call the LinkedIn API from different IPs, avoiding being blocked.

Challenges with Airflow:

- **Static Nature of DAGs**: Airflow DAGs are predefined and static. Adjusting workflows dynamically is possible but requires workarounds that are not elegant.
- **Limitations of `LambdaInvokeFunctionOperator`: Not Deferrable**
  - ○ **Synchronous Invocation**: Blocks Airflow workers while waiting (up to 15 minutes until Lambda timeout).
  - ○ **Asynchronous Invocation**: Cannot track Lambda execution status, leaving success/failure unknown.

Advantages of AWS Step Functions:

- **Dynamic Workflow Support:**
  - ○ The first Lambda outputs the number of subsequent Lambdas required.
  - ○ Step Functions orchestrate and dynamically invoke these Lambdas, managing execution statuses seamlessly.
- **Deferrable in Airflow:**
  - ○ Using the `StepFunctionStartExecutionOperator` in Airflow avoids occupying workers while waiting for Step Functions to complete. This enhances resource efficiency and scalability.

Airflow DAGs
Demonstration in MWAA UI

1d419daf-46dd-4d52-9e7d-c8fe4633b6f1.c7.ap-southeast-2.airflow.amazonaws.com

Security:

MWAA

- **Login**: Users authenticate to MWAA through AWS credentials, with optional Multi-Factor Authentication (MFA) for enhanced security.
- **User Roles**: Permissions and roles within Airflow are managed using AWS IAM policies.
- **Authorization**: Access to AWS services is managed through the MWAA execution role, which defines the actions Airflow is allowed to perform on AWS resources.

Within airflow

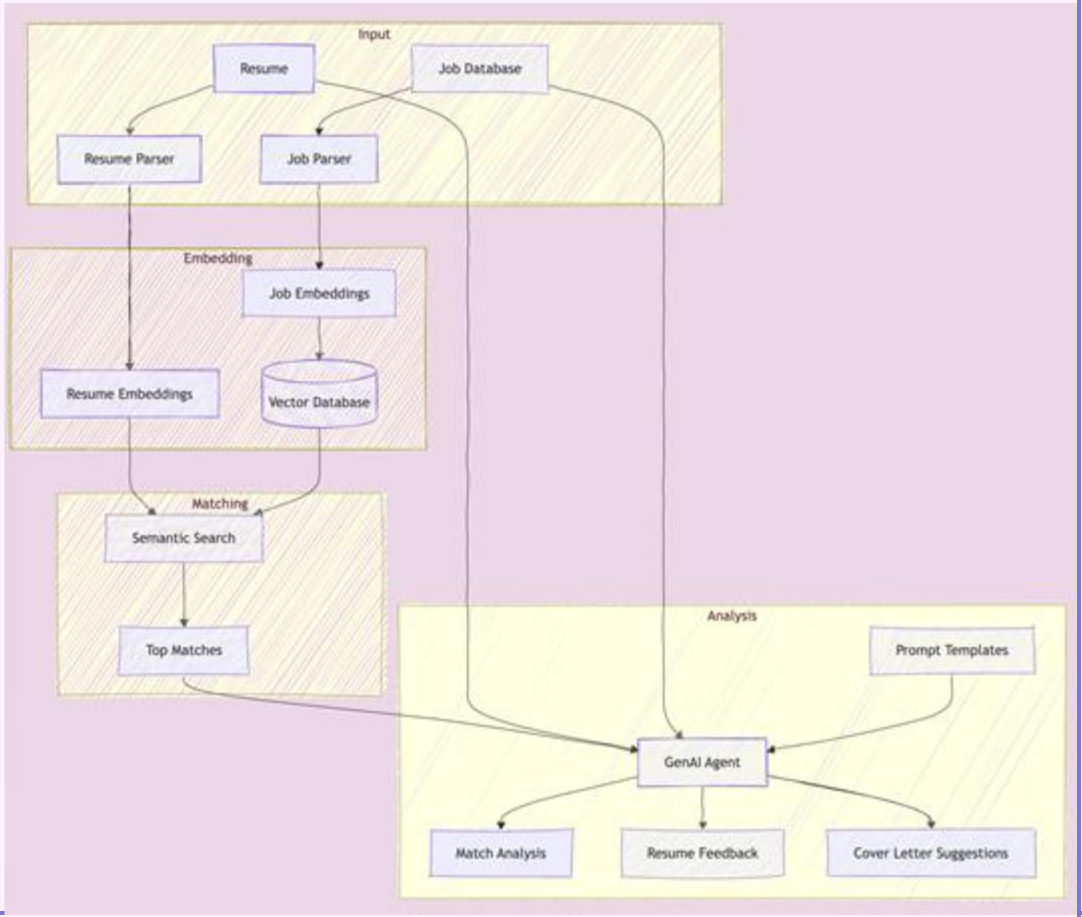
- **Data Encryption in Airflow**:
  - Configured AIRFLOW__CORE__FERNET_KEY to encrypt sensitive data such as variables, connections, and XComs within Airflow, ensuring confidentiality.
- **AWS Secrets Manager as Secret Backend**:
  - Integrated AWS Secrets Manager as the secret backend for Airflow to securely store and manage sensitive information like secrets, API keys, and connection details.
  - Enables Airflow to securely access third-party services, such as:
    - Gmail: For sending notification emails,
    - OpenAI and Pinecone: For injecting new jobs and removing invalid jobs from the vector database,
    - Snowflake: For extracting data and loading it into the data warehouse.

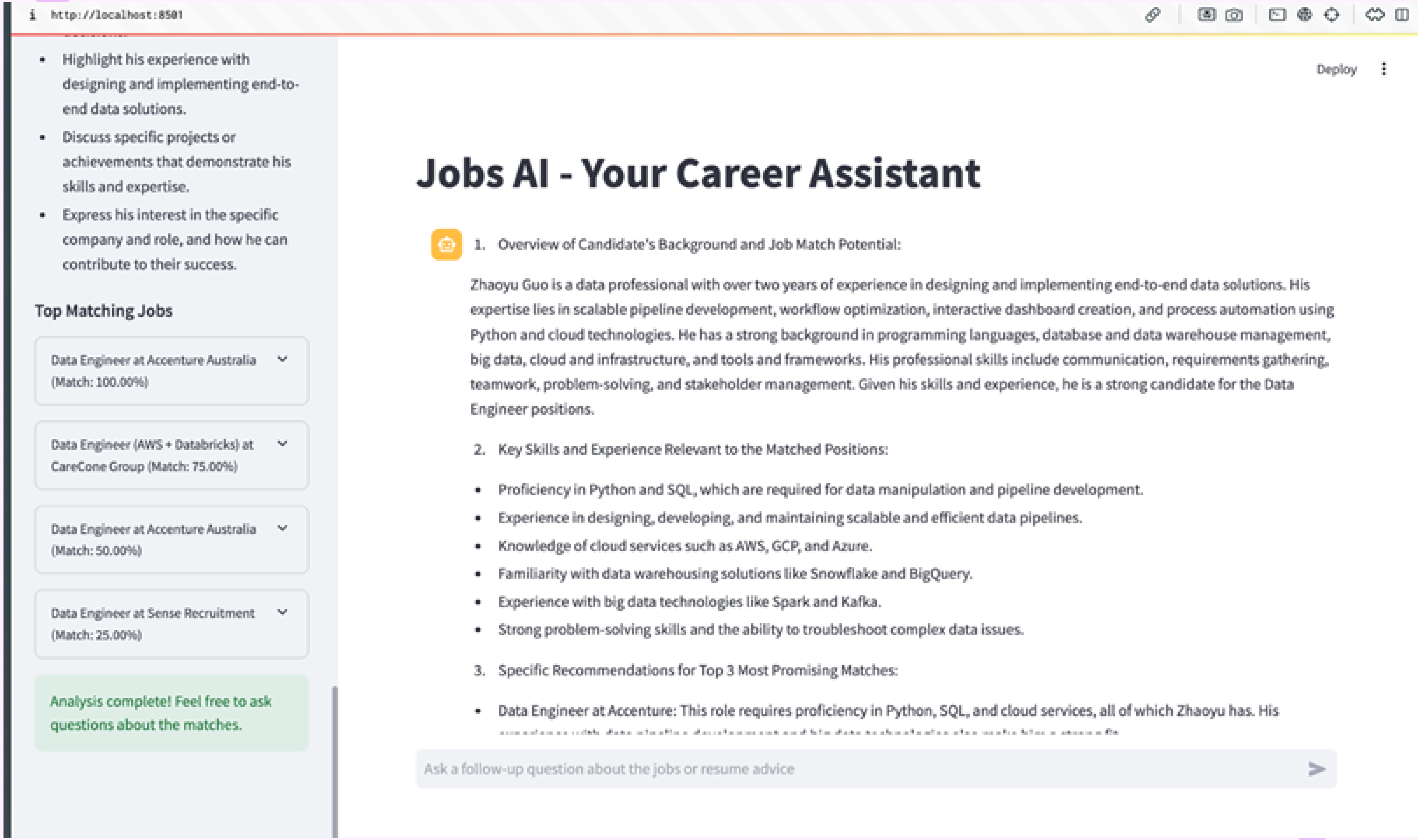
03

Gen AI

AI powered job hunting

- Key steps
  - Retrieval
  - Embedding
  - Generation
  - Analysis
  - Evaluation
- Key features
  - Intelligent job matching
  - Resume analysis and feedback
  - Cover letter generation & support





- Highlight his experience with designing and implementing end-to-end data solutions.
- Discuss specific projects or achievements that demonstrate his skills and expertise.
- Express his interest in the specific company and role, and how he can contribute to their success.

Top Matching Jobs

- Data Engineer at Accenture Australia (Match: 100.00%)
- Data Engineer (AWS + Databricks) at CareCone Group (Match: 75.00%)
- Data Engineer at Accenture Australia (Match: 50.00%)
- Data Engineer at Sense Recruitment (Match: 25.00%)

Analysis complete! Feel free to ask questions about the matches.

Jobs AI - Your Career Assistant

1. Overview of Candidate's Background and Job Match Potential:

Zhaoyu Guo is a data professional with over two years of experience in designing and implementing end-to-end data solutions. His expertise lies in scalable pipeline development, workflow optimization, interactive dashboard creation, and process automation using Python and cloud technologies. He has a strong background in programming languages, database and data warehouse management, big data, cloud and infrastructure, and tools and frameworks. His professional skills include communication, requirements gathering, teamwork, problem-solving, and stakeholder management. Given his skills and experience, he is a strong candidate for the Data Engineer positions.

2. Key Skills and Experience Relevant to the Matched Positions:

- Proficiency in Python and SQL, which are required for data manipulation and pipeline development.
- Experience in designing, developing, and maintaining scalable and efficient data pipelines.
- Knowledge of cloud services such as AWS, GCP, and Azure.
- Familiarity with data warehousing solutions like Snowflake and BigQuery.
- Experience with big data technologies like Spark and Kafka.
- Strong problem-solving skills and the ability to troubleshoot complex data issues.

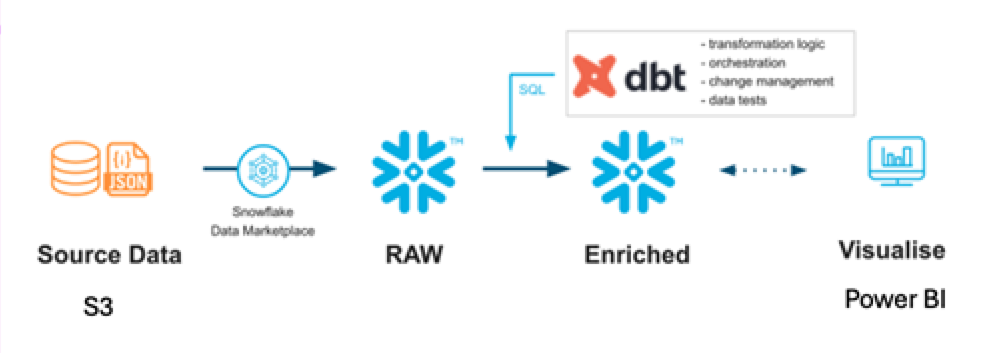
3. Specific Recommendations for Top 3 Most Promising Matches:

- Data Engineer at Accenture: This role requires proficiency in Python, SQL, and cloud services, all of which Zhaoyu has. His experience with data pipeline development and big data technologies also make him a strong fit.

04

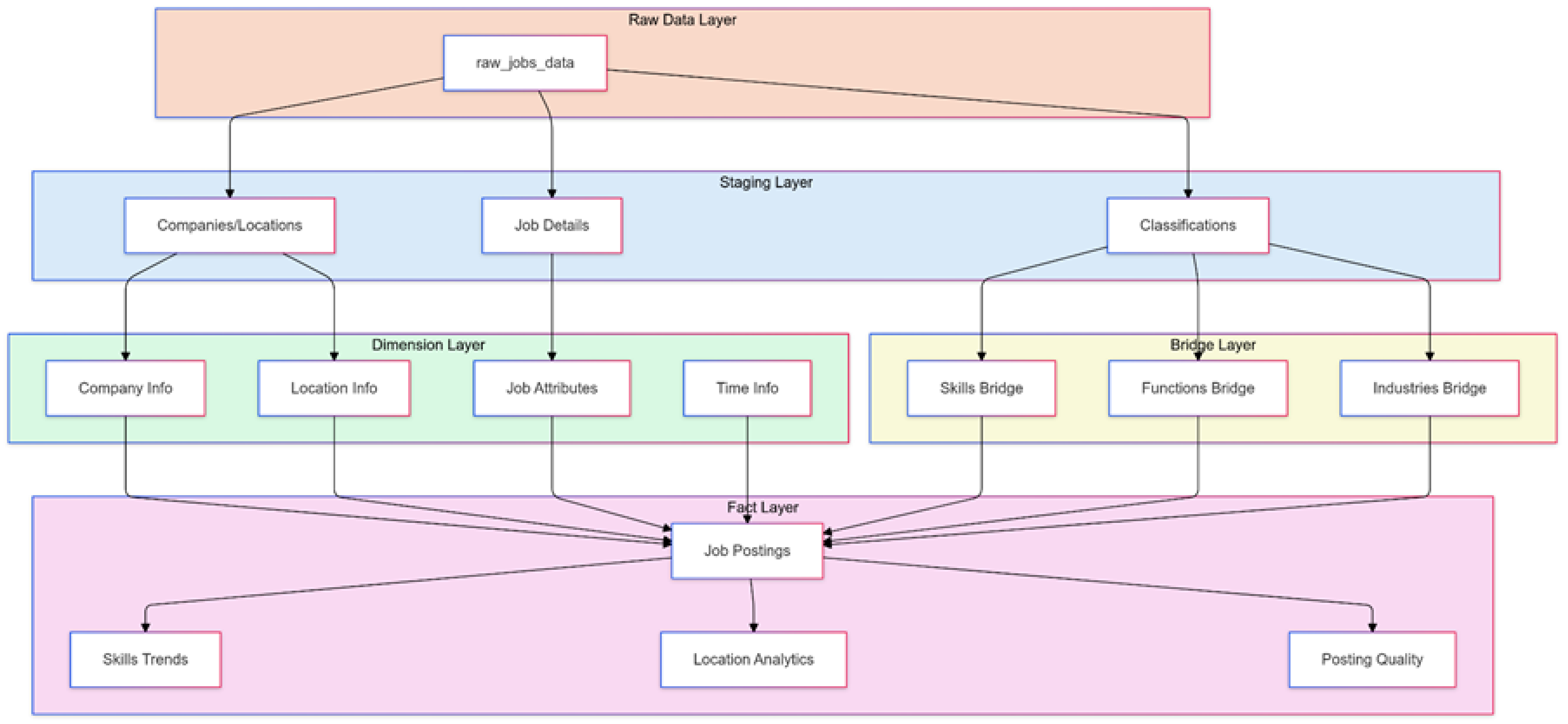
Data Transformation & Reporting

ELT pipeline



1. Snowflake handles **data storage** and compute:
   - Stores raw data from sources (S3, JSON, AVRO)
   - Manages data marketplace access
   - Provides scalable compute for transformations
2. DBT manages **transformation** layer:
   - Handles SQL transformations
   - Orchestrates data flows
   - Implements testing
   - Manages version control
   - Documents data lineage

Data Transform using DBT



Kimball’s dimensional modeling

Fact Tables:

1. Main fact_job_postings:
   - One row per unique job posting ID
   - Links to all dimensions via surrogate keys
2. Analytical fact tables:
   - skill_trends
   - location_analytics
   - posting_quality

Dimension Design:

- Company: Basic dimension with company metrics
- Location: Hierarchy (Country -> Region -> State -> City) + workplace type
- Employment: Type and seniority level standardization
- Time: Date dimension with year, month, quarter
- Bridge tables for many-to-many relationships:
  - job_skills
  - job_functions
  - job_industries

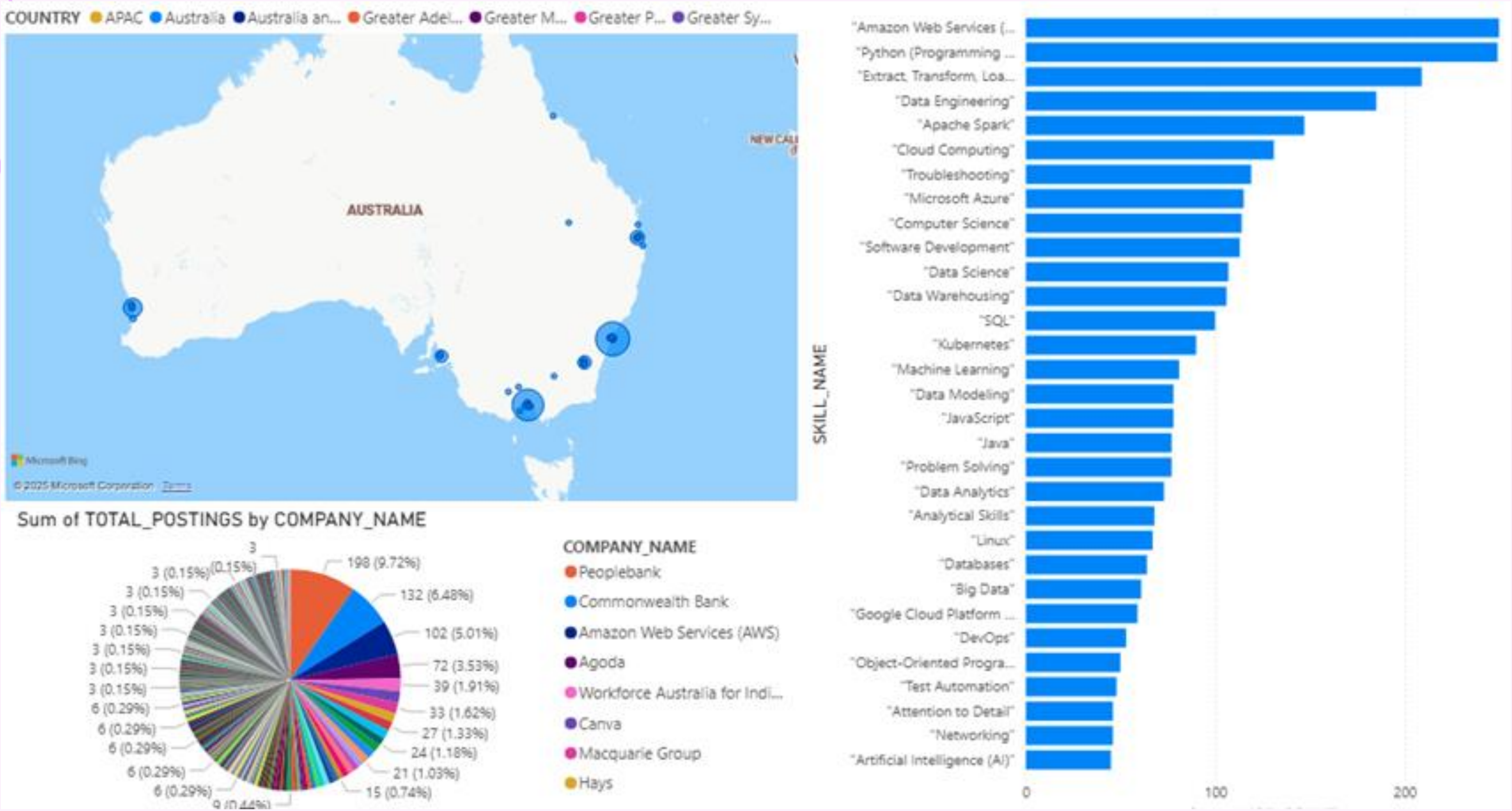
Visualization using Power BI

skill demand trends, hiring patterns by industry, location-based job insights, …

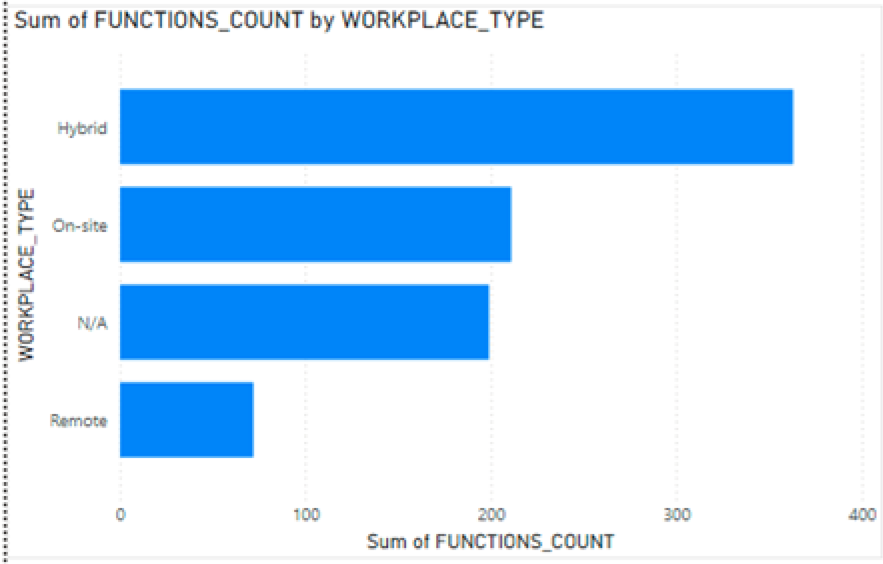
Power BI is preferred over Tableau

- Seamless integration with Excel, Azure and Microsoft 365 suite
- Direct connection to Excel / SharePoint data
- Data modelling capability and strong DAX language for complex calculations
- Extensive Power Query capabilities for data transformation
- Version control support (on Power BI Server)

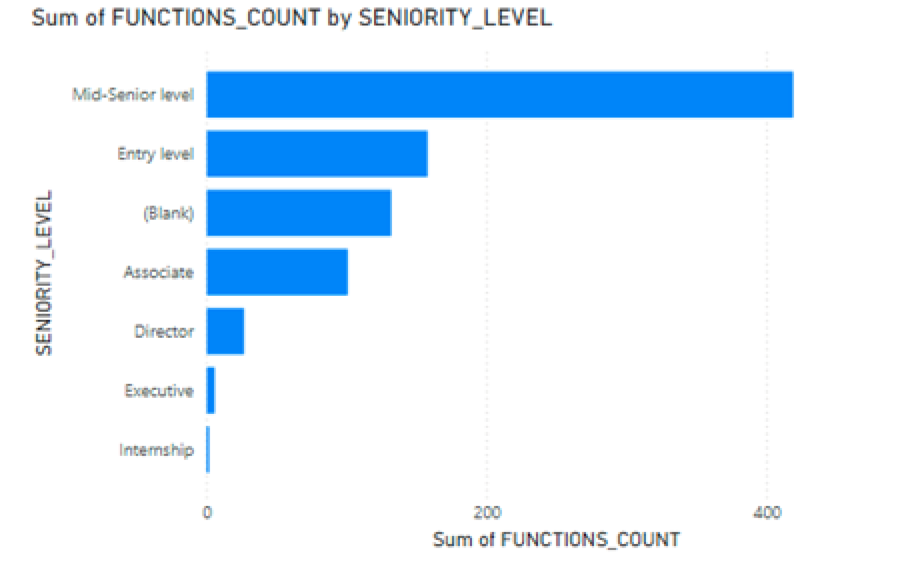
Visualization using Power BI



Visualization using Power BI



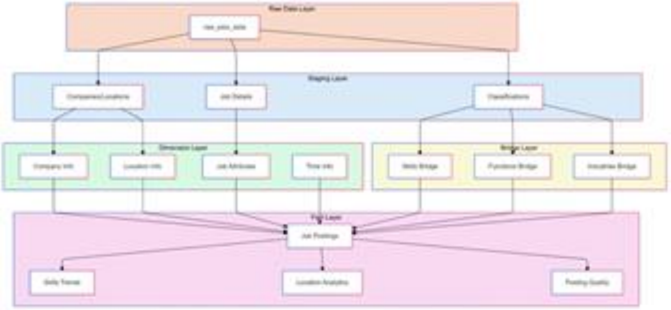
Sum of FUNCTIONS_COUNT by WORKPLACE_TYPE

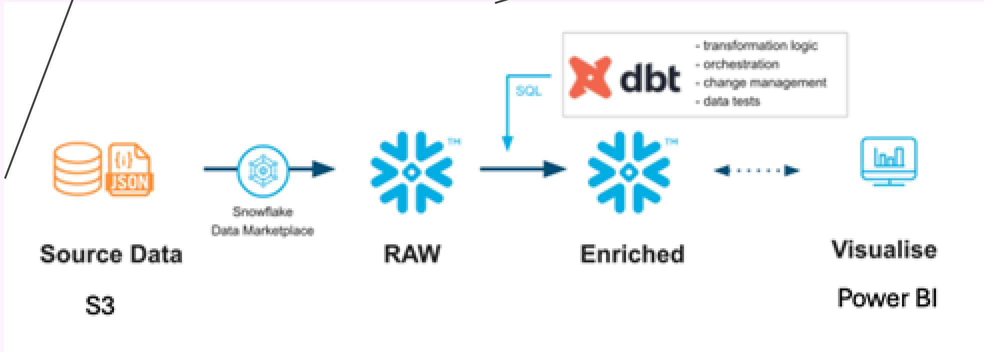


Sum of FUNCTIONS_COUNT by SENIORITY_LEVEL

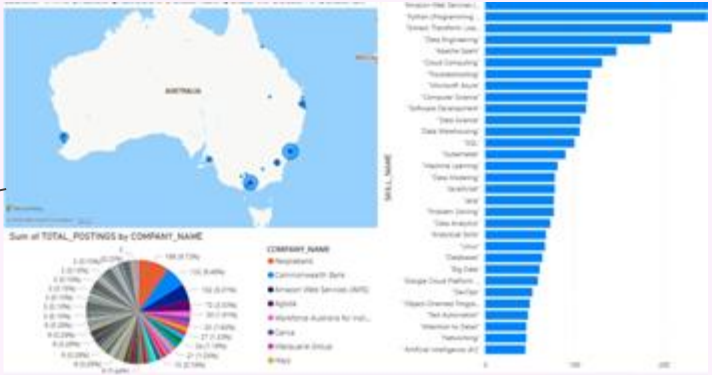
ELT pipeline

Amazon S3





- transformation logic
- orchestration
- change management
- data tests



05

DevOps & Project Management

DevOps

- Terraform for Infrastructure as Code
- GitHub for version control and GitHub Actions for CI/CD
- Airflow for workflow orchestration and monitoring
- CloudWatch for infrastructure and application monitoring

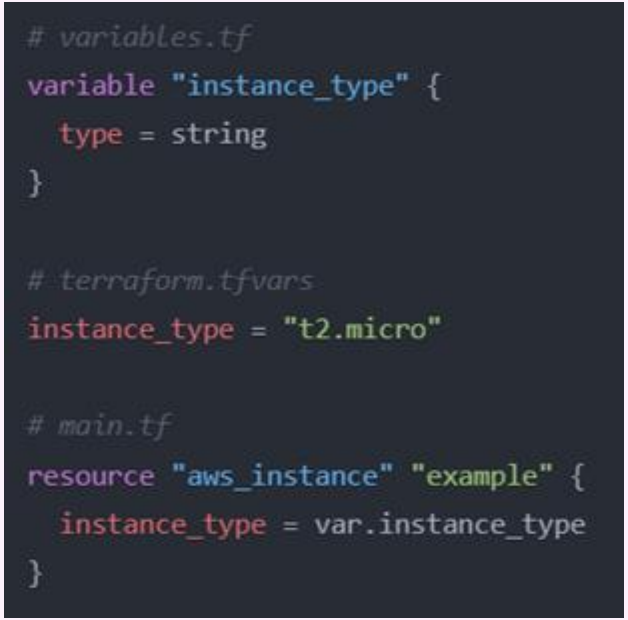
IaC

Terraform

- Infrastructure as Code (IaC) automation
- Multi-cloud support
- Version-controlled infrastructure
- Consistent environments
- Easy to code & Reduced human error

Advantages over CloudFormation

- Provider-agnostic (not AWS-only)
- HCL syntax (more readable than JSON/YAML)
- State management capabilities
- Faster deployment execution
- Better dependency handling



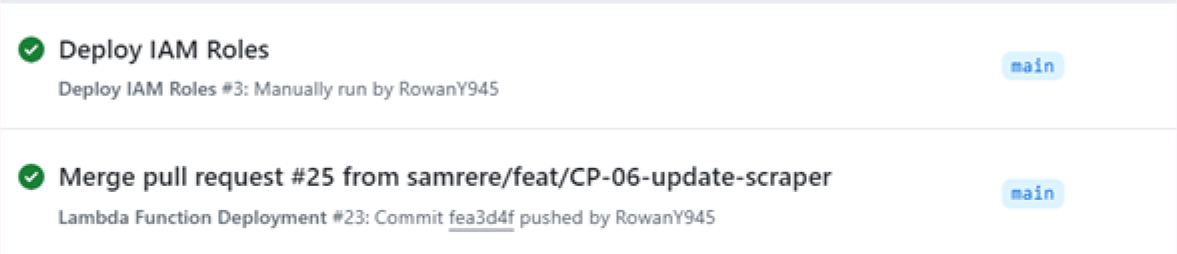
```
# variables.tf
variable "instance_type" {
  type = string
}

# terraform.tfvars
instance_type = "t2.micro"

# main.tf
resource "aws_instance" "example" {
  instance_type = var.instance_type
}
```

CI/CD: GitHub Action

Lambda Function Depolyment



Deploy IAM Roles

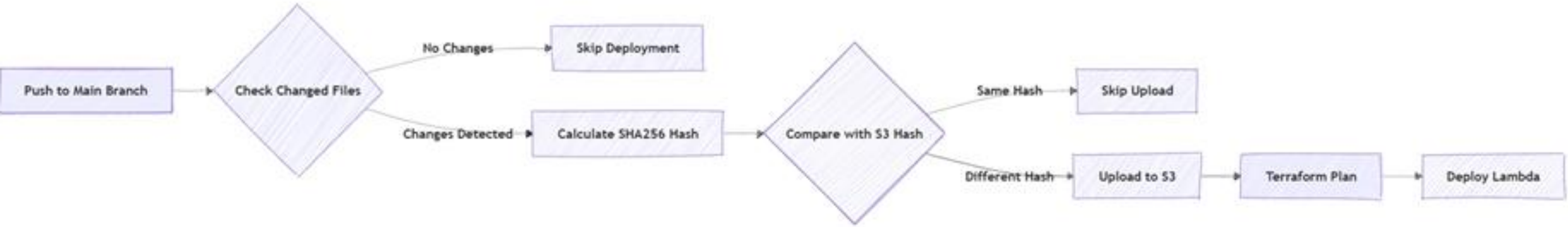
Deploy IAM Roles #3: Manually run by RowanY945

main

Merge pull request #25 from samrere/feat/CP-06-update-scraper

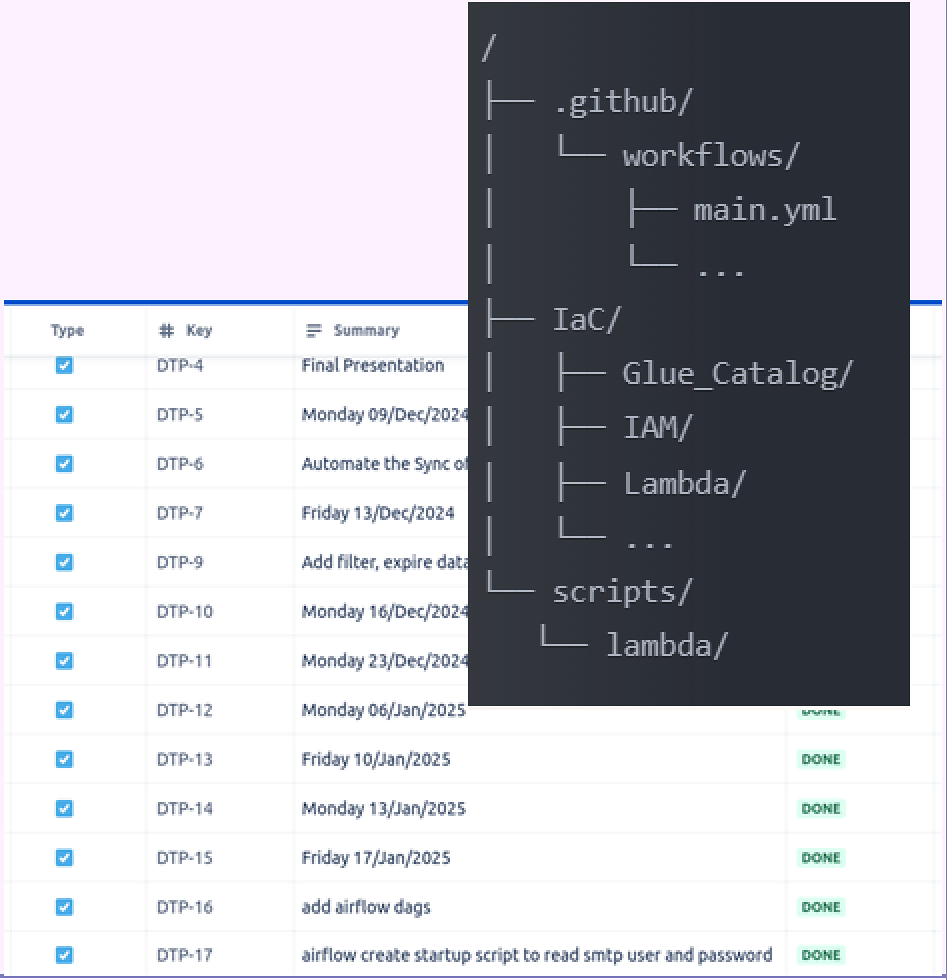
Lambda Function Deployment #23: Commit fea3d4f pushed by RowanY945

main



Code repo and Project Management

- Github for code repo and CI/CD
  - Scraper code and IaC
  - Data Modeling (dbt models)
  - RAG
  - Orchestration (Airflow dags)
- JIRA for Planning and Task Management
- Confluence for documentation, knowledge management, and collaboration



```
/
├── .github/
│   └── workflows/
│       ├── main.yml
│       └── ...
├── IaC/
│   ├── Glue_Catalog/
│   ├── IAM/
│   ├── Lambda/
│   └── ...
└── scripts/
    └── lambda/
```

Type	# Key	Summary	
☑	DTP-4	Final Presentation	
☑	DTP-5	Monday 09/Dec/2024	
☑	DTP-6	Automate the Sync of	
☑	DTP-7	Friday 13/Dec/2024	
☑	DTP-9	Add filter, expire data	
☑	DTP-10	Monday 16/Dec/2024	
☑	DTP-11	Monday 23/Dec/2024	
☑	DTP-12	Monday 06/Jan/2025	DONE
☑	DTP-13	Friday 10/Jan/2025	DONE
☑	DTP-14	Monday 13/Jan/2025	DONE
☑	DTP-15	Friday 17/Jan/2025	DONE
☑	DTP-16	add airflow dags	DONE
☑	DTP-17	airflow create startup script to read smtp user and password	DONE

Thanks!