

A cloud-based near real-time system for large scale water quantity and quality models

The challenges overcome and lessons learnt in developing an operational-ready system


Michael Pegios

Richard Laugesen

Urooj Khan

Alex Cornish

Introduction

- The **Bureau of Meteorology** is a 24/7 operational agency that cover weather, climate, ocean and space weather services for Australia
- The Bureau plays a critical role in number of sectors including emergency management, energy, resources, agriculture and water.
-  • An official weather app for access to hourly and 7-day forecasts, radar and warnings - wherever you are
- **My role:** Scientific Software Engineer
- **Team:** Water Models Team
- **Project:** eReefs





Agenda

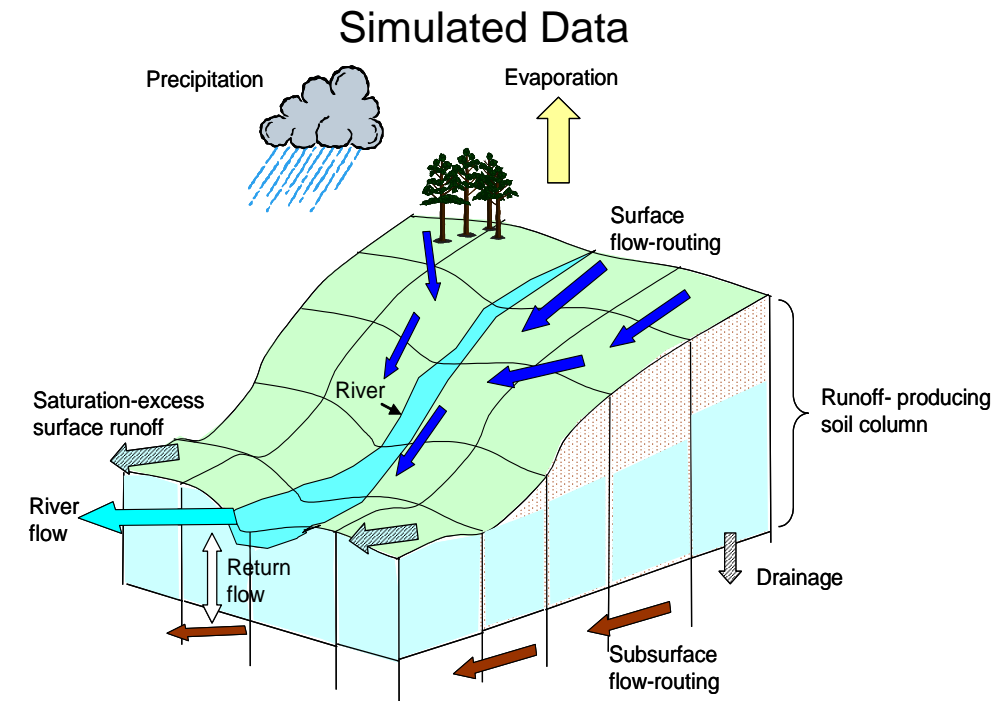
- **Objectives & Challenges**
- **Bureau Data Hub**
- **Terraform Infrastructure Set Up**
- **CI/CD Deployment Workflow**
- **AWS System Architecture**



Objectives and Challenges

Objectives

- **Develop a system to run the largest hydrological model in the southern hemisphere**
- **Deploy large-scale water quantity and quality models developed for the GBR catchments to AWS:**
 - Operationalisation of the high spatiotemporal resolution Grid-to-Grid (G2G) model
 - Operationalisation of water quality models for all gauged and ungauged rivers
 - Daily updates of real-time observation and modelled flow, and water quality data
- **Develop a workflow to generate and deliver daily data to customers in Near Real-Time**
- **Provide a portable solution where production generation and data modelling tasks can be run outside of the Bureau**

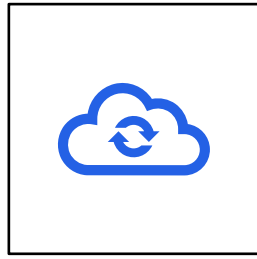


Moore et al., 2006

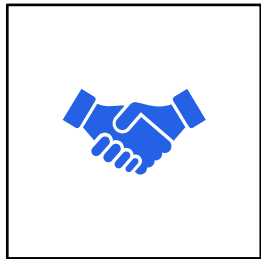
Challenges



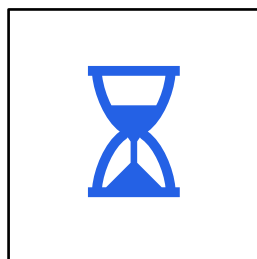
How to sync real-time data between internal Bureau systems and AWS daily?



How can AWS access and modify the same data?

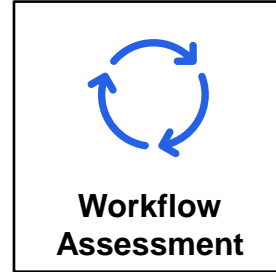


How to create a re-deployable system we can transfer ownership of to partner agencies

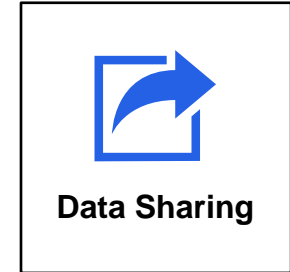


How to quickly develop a solution given time constraints

Reasons for Selecting AWS



Workflow Assessment



Data Sharing



Managed Workflow for Apache Airflow (MWAA)



Cloud First / Cloud Right



Open Source





Systems Overview

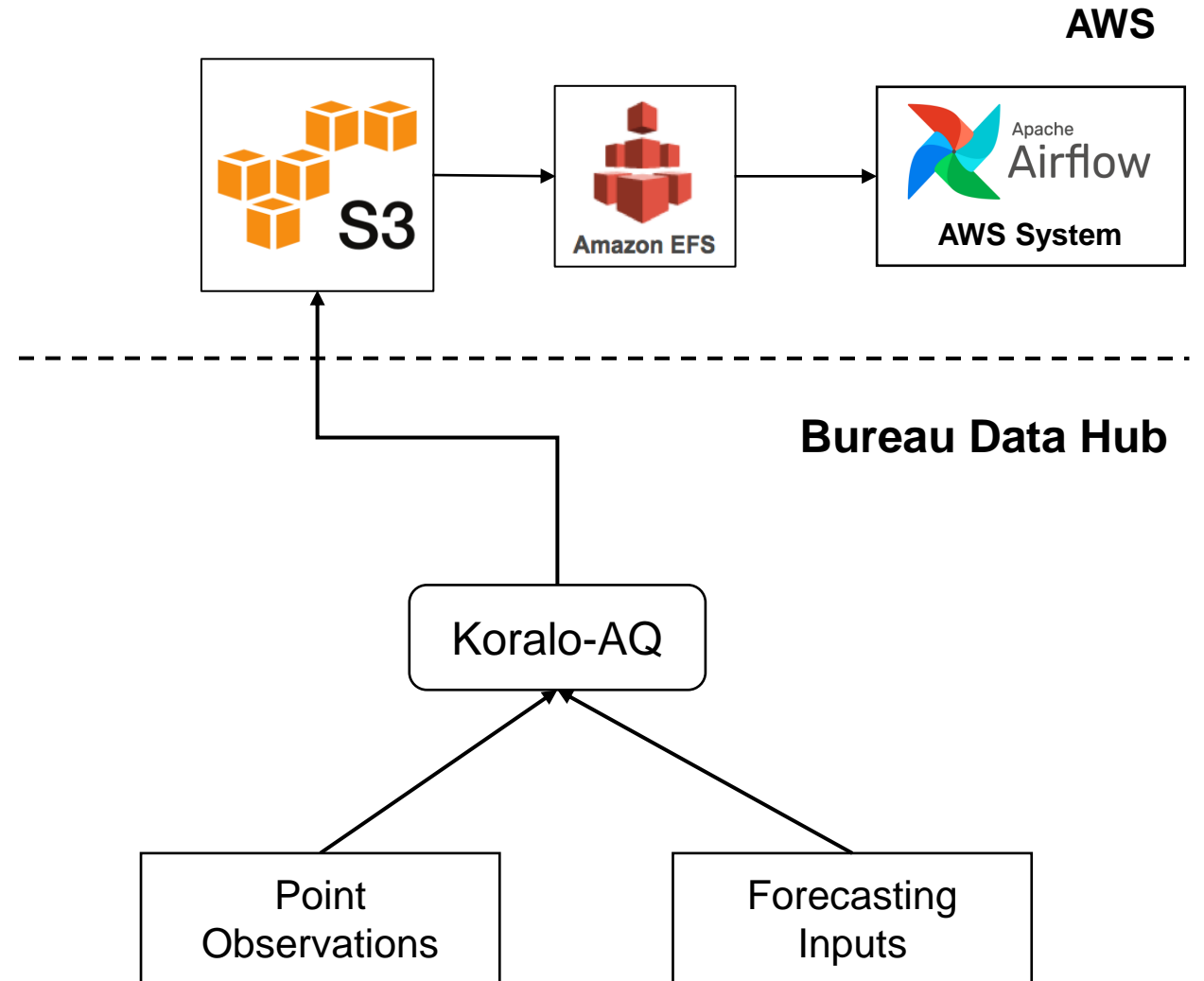
Bureau Data Hub

- **Point Observation Inputs**

- Hourly Rainfall Observation Data
- Hourly Streamflow Observation Data

- **Forecasting Inputs:**

- Daily Gridded Rainfall data
- Daily Potential Evapotranspiration (PET) data



Bureau Data Sources

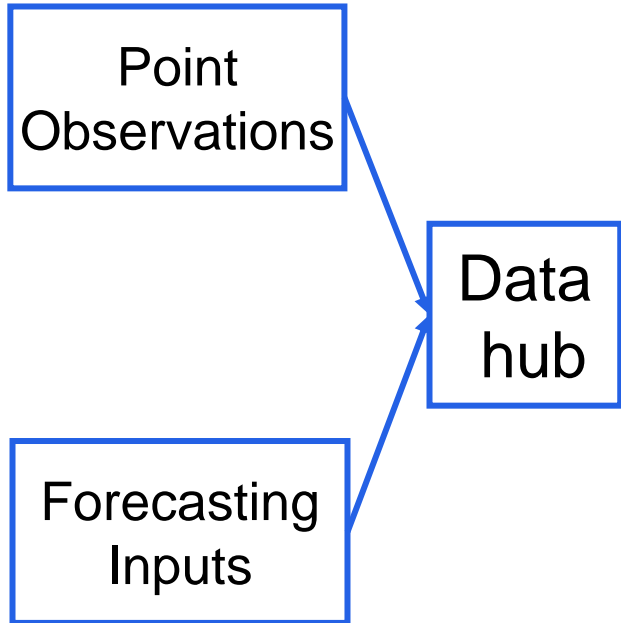


Systems solution



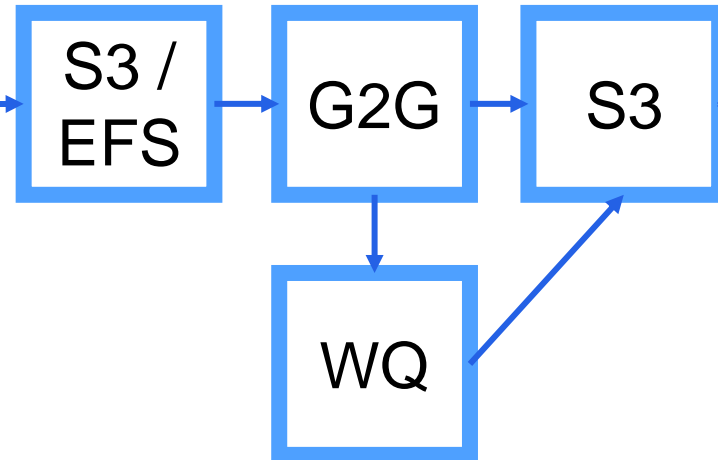
Bureau Systems

Ongoing data provider



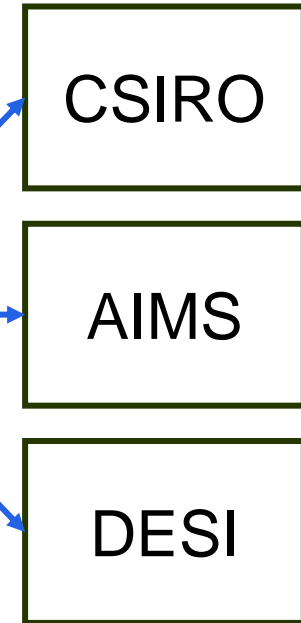
Amazon Cloud

Transfer to operational delivery entity

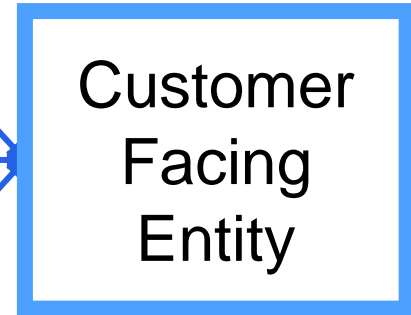


Near Real-Time

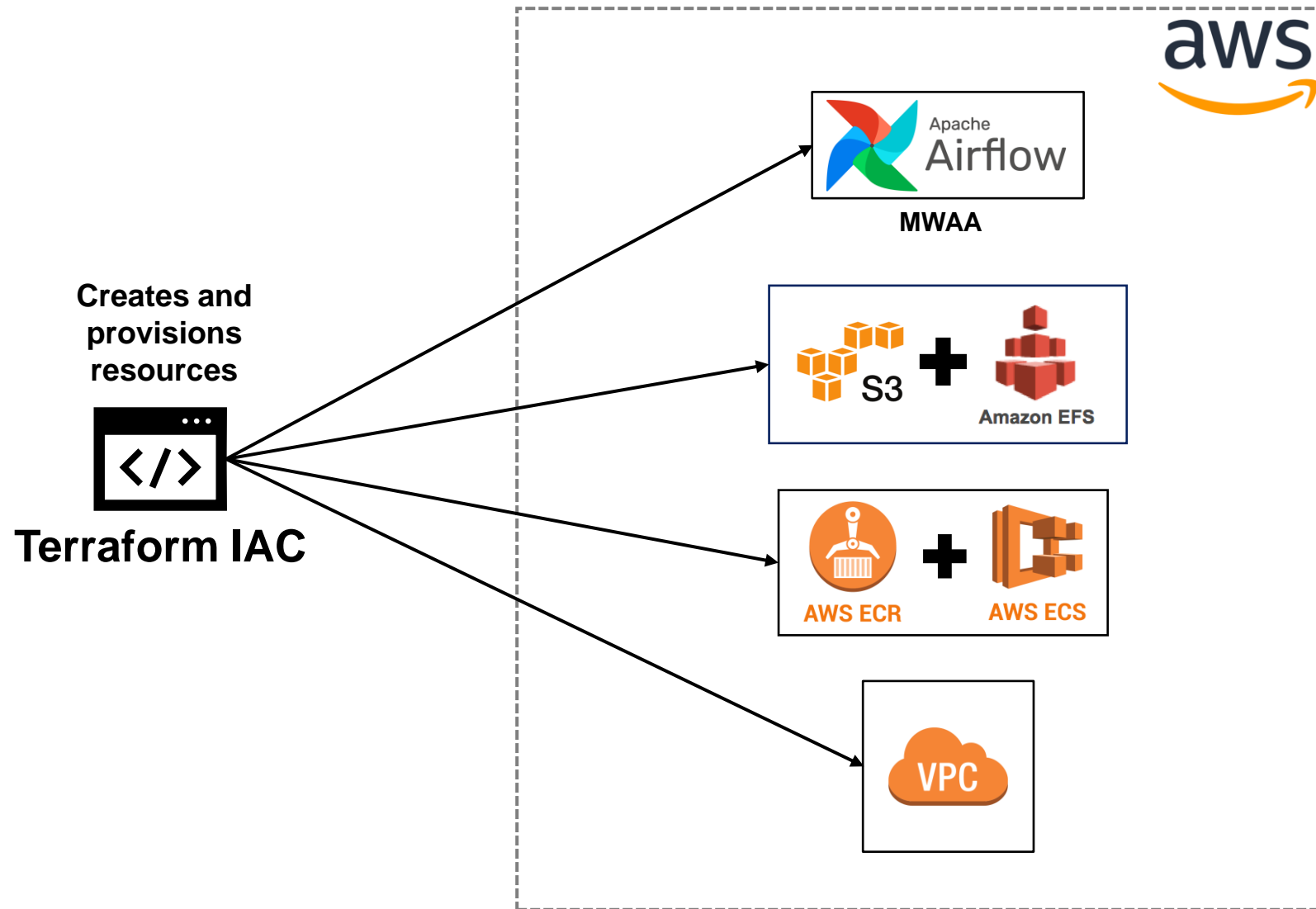
Partner Systems



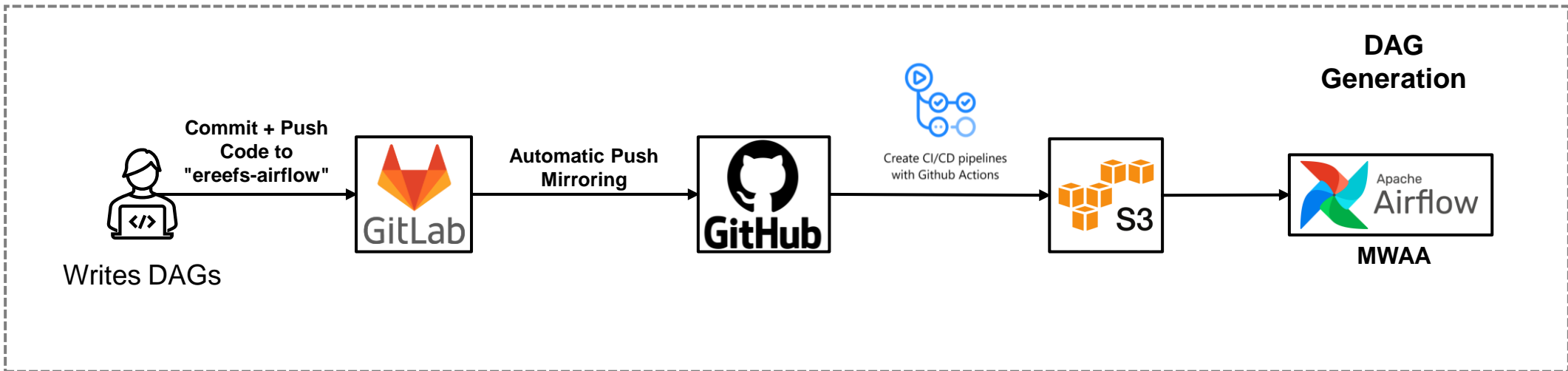
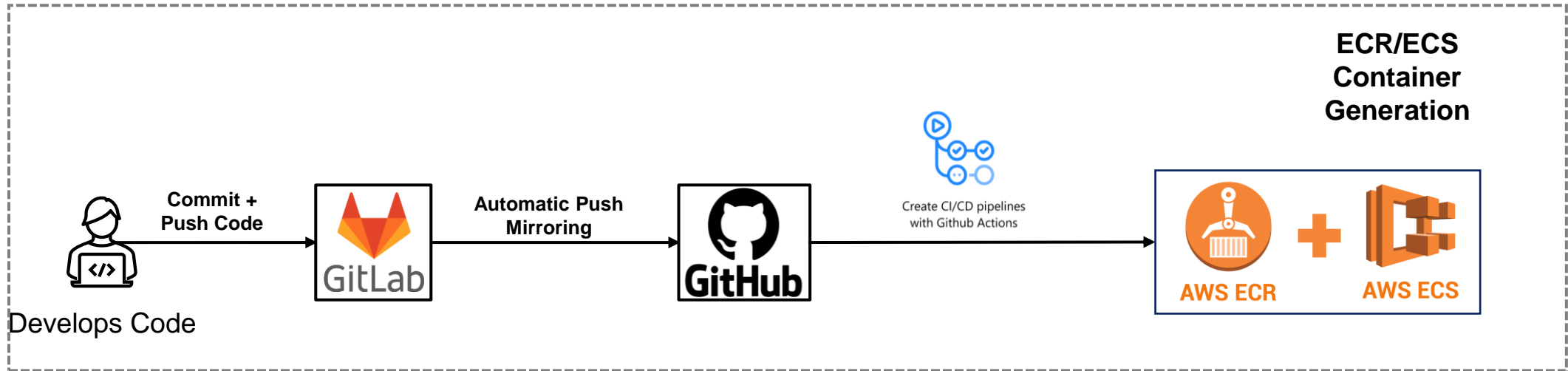
Operational delivery entity



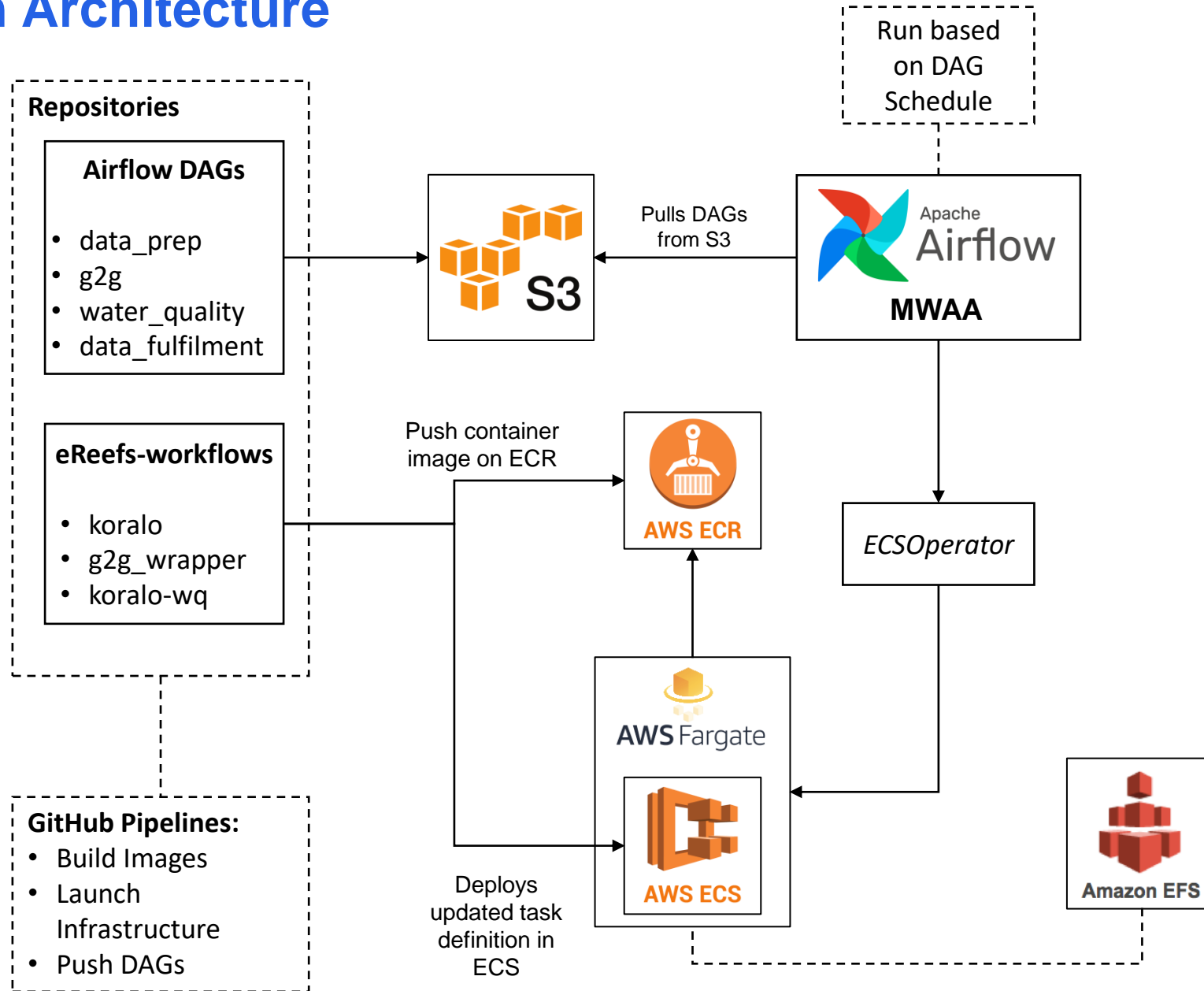
Terraform Infrastructure Set Up



CI/CD Deployment Workflow



AWS System Architecture



End-to-End System Workflow

Bureau
Data Hub

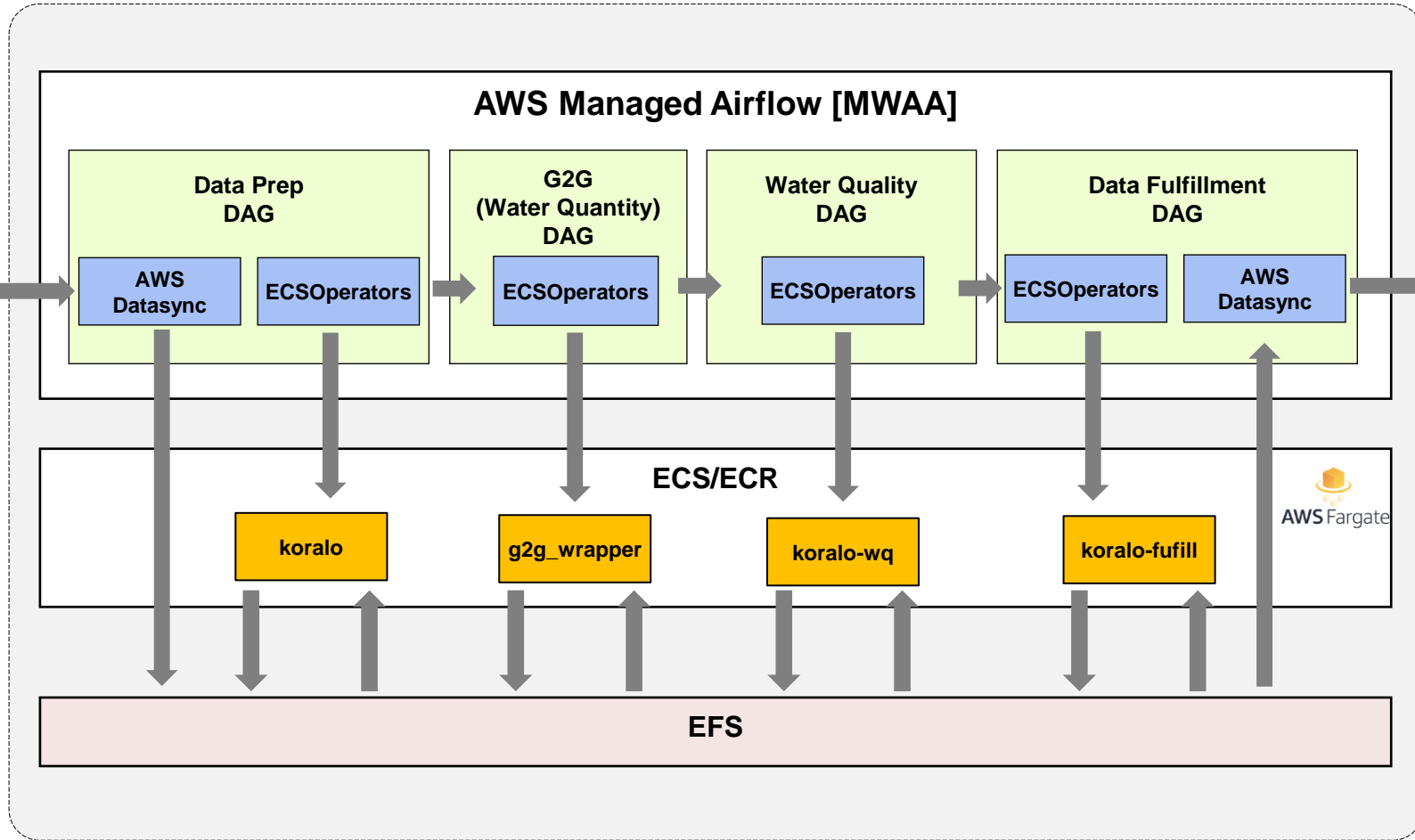
AWS System

Airflow

Data Hub
DAG

Python
Operators

Koralo-AQ



Deployed System

The screenshot displays the Airflow web interface for a DAG named 'g2g'. The interface includes a navigation menu, a header with the DAG name and schedule, and a main content area with a bar chart and a summary table.

Header: Airflow | DAGs | Datasets | Security | Browse | Admin | Docs | 02:38 UTC | User Profile

DAG Information: DAG: g2g BoM eReefs G2G workflow | Schedule: 30 1*** | Next Run: 2024-05-17, 01:30:00

Navigation: Grid | Graph | Calendar | Task Duration | Task Tries | Landing Times | Gantt | Details | Code | Audit Log

Filters: 17/05/2024 02:37:42 AM | 25 | All Run Types | All Run States | Clear Filters

Run States: deferred, failed, queued, removed, restarting, running, scheduled, shutdown, skipped, success, up_for_reschedule, up_for_retry, upstream_failed, no_status

Auto-refresh:

Bar Chart: Duration (00:00:00 to 00:05:56) | May 06, 01:30 | May 09, 03:10 | May 14, 23:44

Task List: validate_params, g2g_run, convert_g2g_output_run

DAG Details:

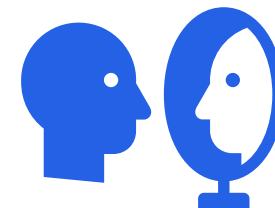
DAG Runs Summary	
Total Runs Displayed	25
Total success	25
Total failed	0
First Run Start	2024-05-06, 01:30:00 UTC
Last Run Start	2024-05-17, 01:30:00 UTC
Max Run Duration	00:05:56

Summary

- **AWS** and **MWAA** for the underlying infrastructure
- **Airflow** for orchestration and monitoring
- Utilisation of **S3/EFS** storage solutions
- **Containerisation** of data processing and modelling tasks
- **GitHub** Actions
- Portability achieved through **Terraform** IAC

Reflections

- How **Airflow** compared to other workflow tools
- Consideration of adopting **FSx for Lustre**
- Use of **Direct Connect**
- Leveraging **Terraform** earlier
- Separation of environments for **dev** and **test**





The Bureau
of Meteorology

Thank you

Michael Pegios – michael.pegios@bom.gov.au