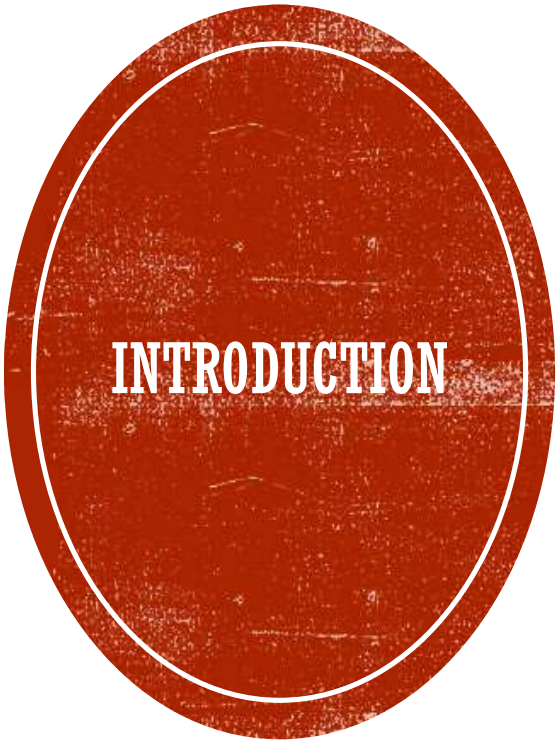

# AWS Data Lake Proof Of Concept




Project outline and business case

Legacy Technology Stack

AWS Proof of Concept

Future Roadmap

INCREASING BUSINESS DEMAND

COST

FUNCTIONALITY

PERFORMANCE

# PROJECT OUTLINE AND BUSINESS CASE

# LEGACY TECHNOLOGY STACK

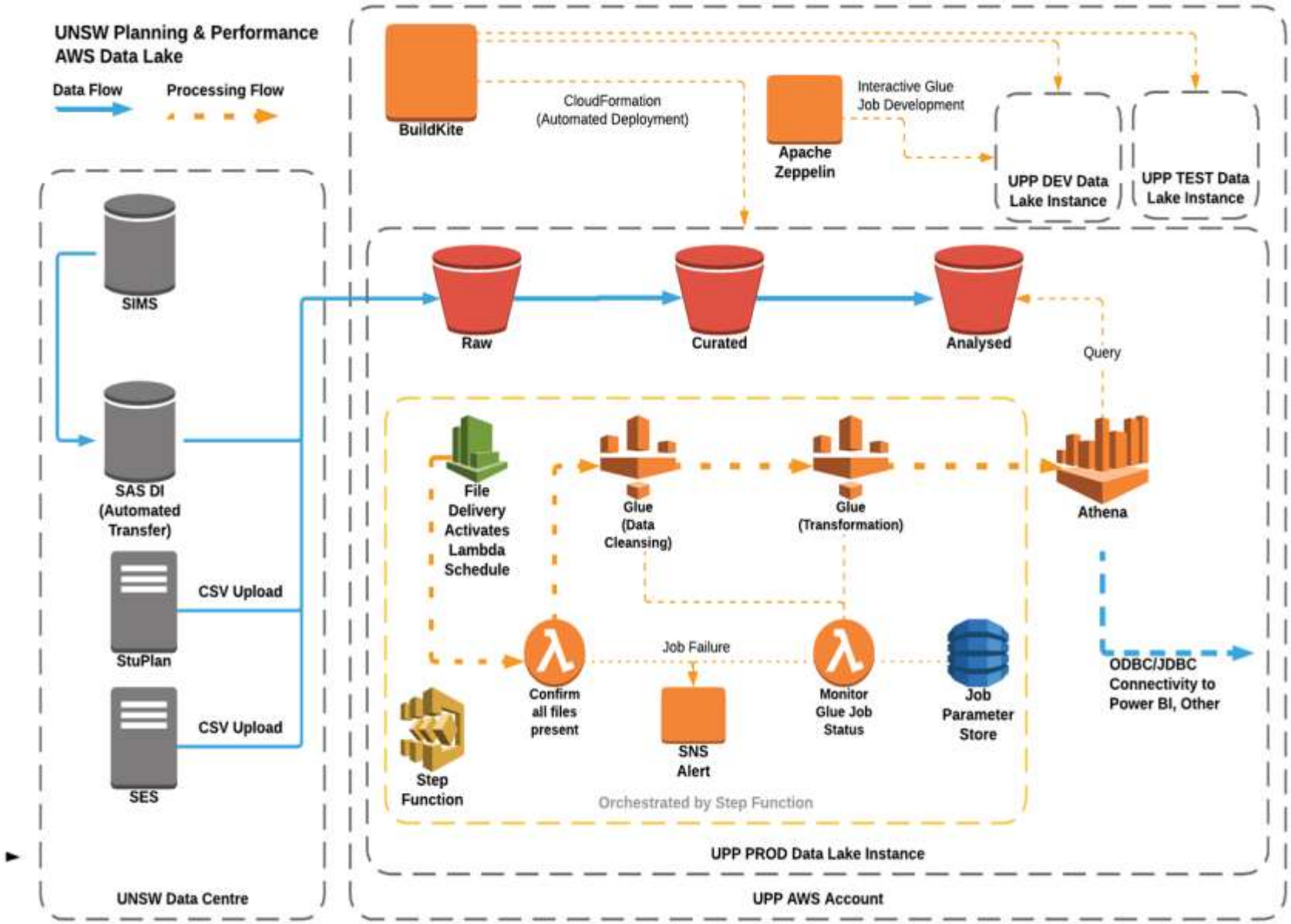
- Changing business requirements
- SAS Resourcing challenges
- Too costly to maintain and difficult to scale infrastructures
- Complex legacy business rules
- Lack of Documentation

# AWS PROOF OF CONCEPT

- Asked for 6 weeks POC using AWS but delivered within 4 weeks.
- Serverless data lake
- 5 new dashboards
- Full automation

# LESSONS LEARNED

- Find the right technology partner/consultants
- Make sure you have IT supports
- Discuss with Stakeholder early
- Need data SME (Subject Matter Expert)
- Document your success and share it


UNSW Planning & Performance
AWS Data Lake
Data Flow
Processing Flow
BuildKite
CloudFormation
(Automated Deployment)
Apache
Zeppelin
Interactive Glue
Job Development
UPP DEV Data
Lake Instance
UPP TEST Data
Lake Instance
SIMS
SAS DI
(Automated
Transfer)
StuPlan
CSV Upload
SES
CSV Upload
Raw
Curated
Analysed
Query
File
Delivery
Activates
Lambda
Schedule
Glue
(Data
Cleansing)
Glue
(Transformation)
Athena
Confirm
all files
present
Job Failure
SNS
Alert
Monitor
Glue Job
Status
Job
Parameter
Store
ODBC/JDBC
Connectivity to
Power BI, Other
Step
Function
Orchestrated by Step Function
UPP PROD Data Lake Instance
UNSW Data Centre
UPP AWS Account

# SOLUTION COMPONENT BREAKDOWN

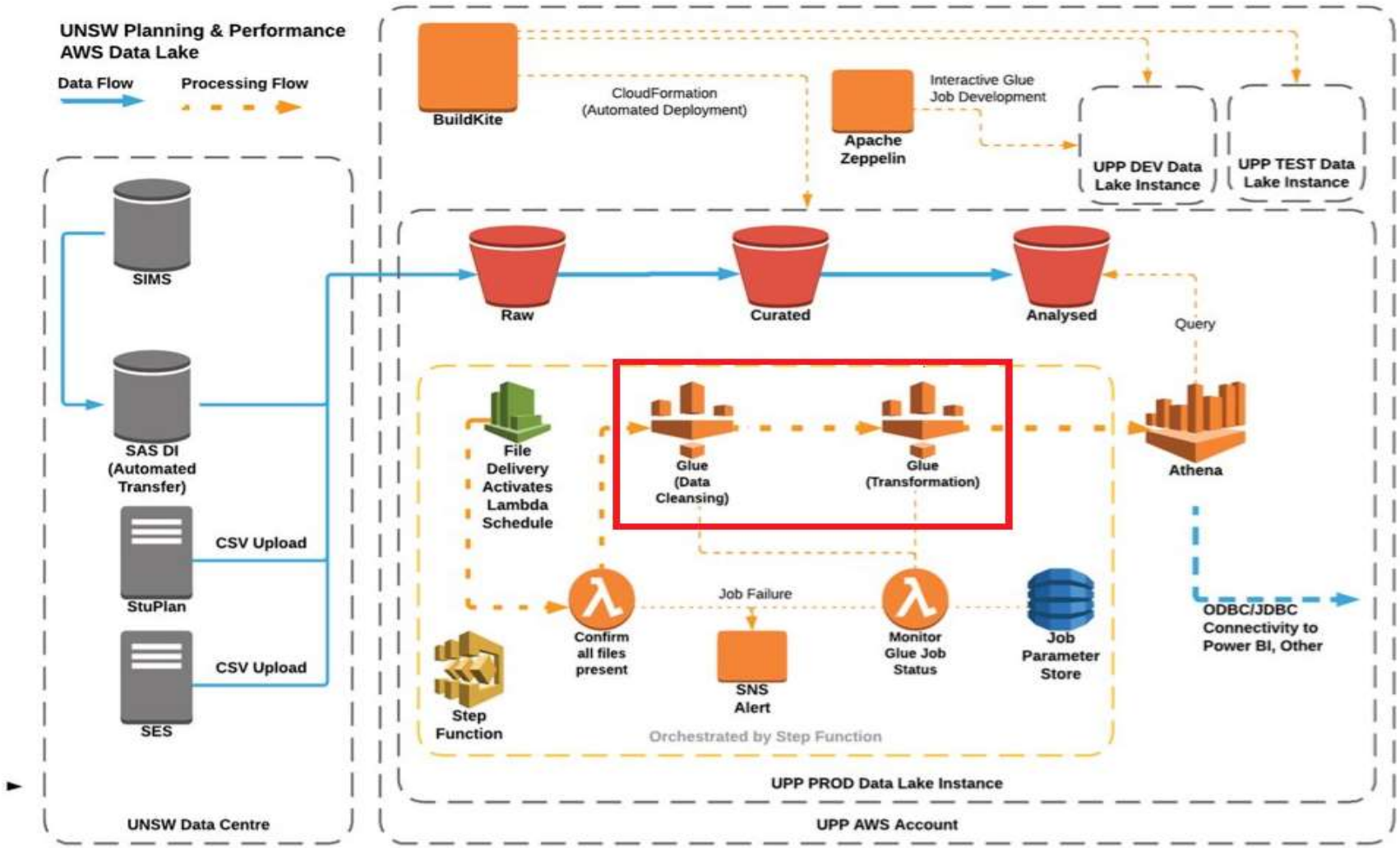
- **AWS Native**

  - Glue
  - Athena
  - S3
  - Lambda
  - Step Function
  - DynamoDB
  - SNS
  - CloudWatch
  - CloudFormation

- **3rd Party Tooling**

  - BitBucket
  - BuildKite
  - Zeppelin Notebook

# GLUE


UNSW Planning & Performance
AWS Data Lake
Data Flow
Processing Flow
BuildKite
CloudFormation
(Automated Deployment)
Apache
Zeppelin
Interactive Glue
Job Development
UPP DEV Data
Lake Instance
UPP TEST Data
Lake Instance
SIMS
SAS DI
(Automated
Transfer)
StuPlan
CSV Upload
SES
CSV Upload
Raw
Curated
Analysed
Query
File
Delivery
Activates
Lambda
Schedule
Glue
(Data
Cleansing)
Glue
(Transformation)
Athena
Confirm
all files
present
Job Failure
SNS
Alert
Monitor
Glue Job
Status
Job
Parameter
Store
ODBC/JDBC
Connectivity to
Power BI, Other
Step
Function
Orchestrated by Step Function
UPP PROD Data Lake Instance
UNSW Data Centre
UPP AWS Account

# GLUE

- Glue is a serverless ETL service for batch workload.

**Glue Catalog**

Replaced Hive Metastore, stores your data's metadata

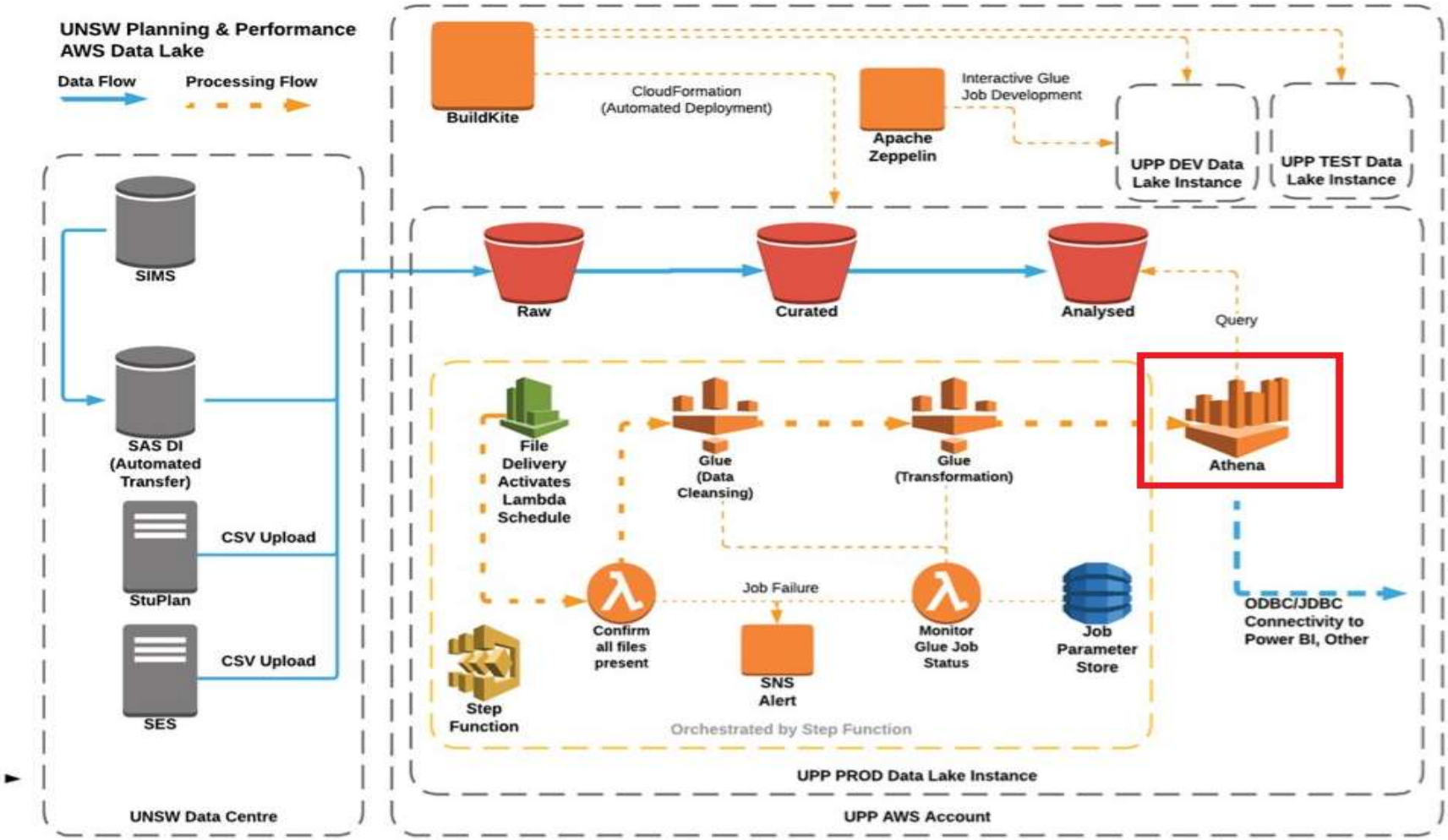
**Glue Crawler**

Crawls new file that comes in and creates table/partitions in the glue catalog

**Glue Job**

Runs Spark Code in either Scala or Python

# ATHENA


UNSW Planning & Performance
AWS Data Lake
Data Flow
Processing Flow
BuildKite
CloudFormation
(Automated Deployment)
Apache
Zeppelin
Interactive Glue
Job Development
UPP DEV Data
Lake Instance
UPP TEST Data
Lake Instance
SIMS
SAS DI
(Automated
Transfer)
CSV Upload
StuPlan
CSV Upload
SES
Raw
Curated
Analysed
Query
File
Delivery
Activates
Lambda
Schedule
Glue
(Data
Cleansing)
Glue
(Transformation)
Athena
Confirm
all files
present
Job Failure
SNS
Alert
Monitor
Glue Job
Status
Job
Parameter
Store
ODBC/JDBC
Connectivity to
Power BI, Other
Step
Function
Orchestrated by Step Function
UPP PROD Data Lake Instance
UNSW Data Centre
UPP AWS Account

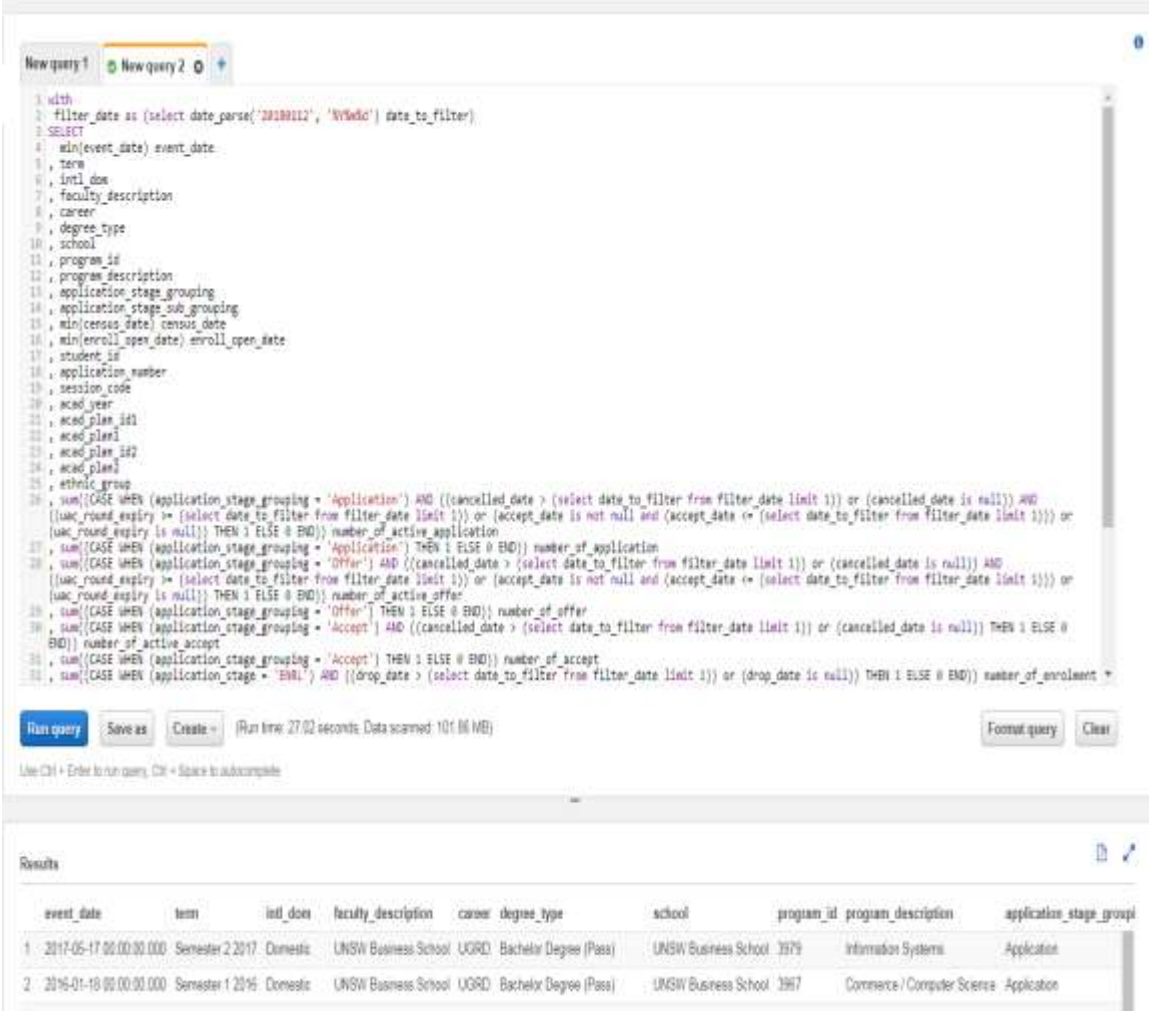
# ATHENA

Interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.

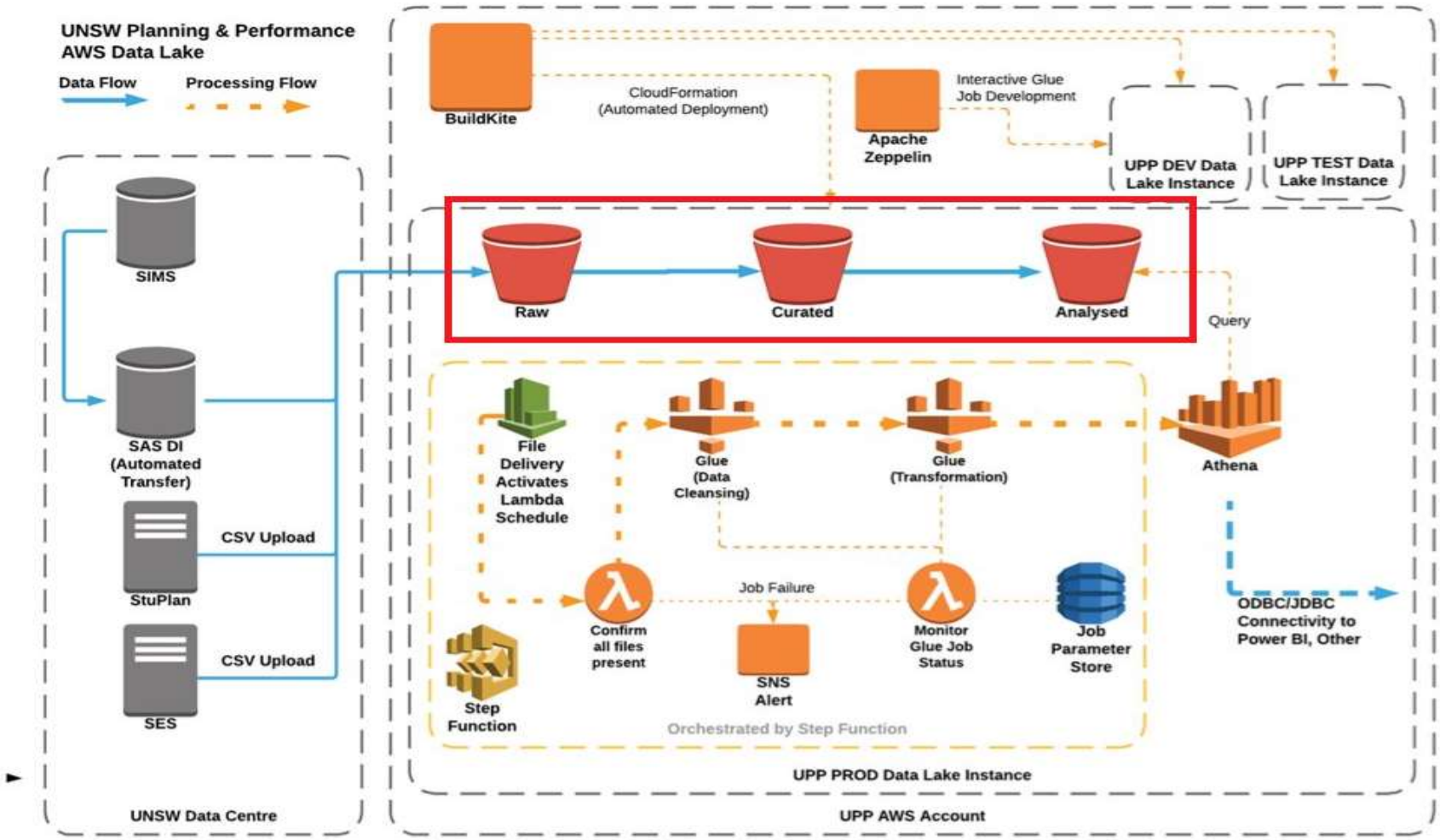
Query using AWS console or your favourite thick client using JDBC/ODBC connection

Build in integration with popular BI tools like Tableau/PowerBI

Serverless

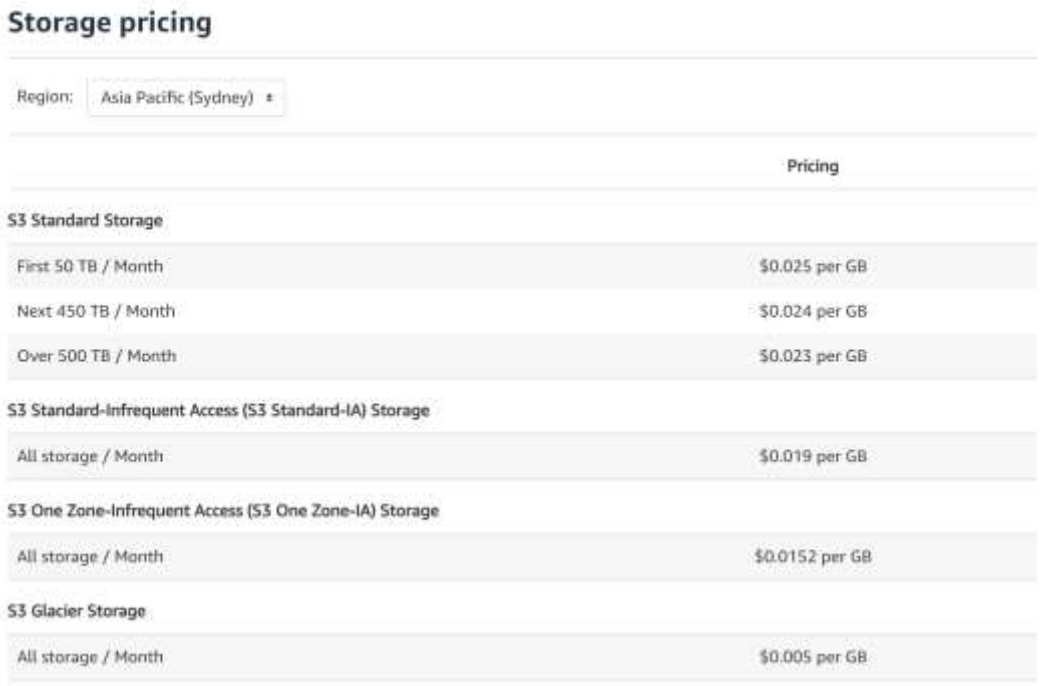


# SIMPLE STORAGE SERVICE (S3)



# SIMPLE STORAGE SERVICE (S3)

- Object Storage
- Unlimited on size
- Serverless
- Cheap
- Resilient (99.999999999% Durability of objects)



**Storage pricing**

Region: Asia Pacific (Sydney)

| | Pricing |
|---|---|
| **S3 Standard Storage** | |
| First 50 TB / Month | $0.025 per GB |
| Next 450 TB / Month | $0.024 per GB |
| Over 500 TB / Month | $0.023 per GB |
| **S3 Standard-Infrequent Access (S3 Standard-IA) Storage** | |
| All storage / Month | $0.019 per GB |
| **S3 One Zone-Infrequent Access (S3 One Zone-IA) Storage** | |
| All storage / Month | $0.0152 per GB |
| **S3 Glacier Storage** | |
| All storage / Month | $0.005 per GB |

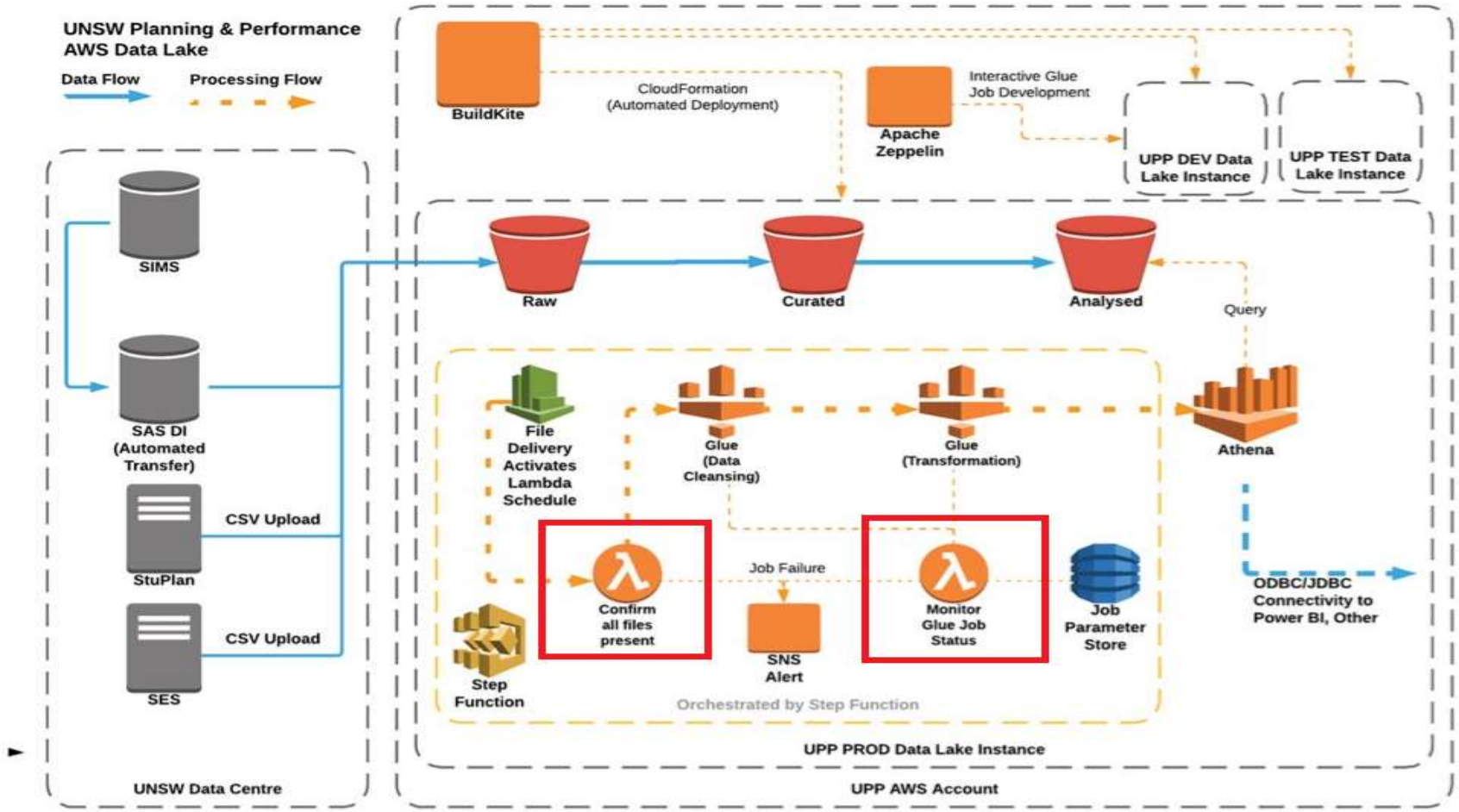
- Allows event driven workflow by S3 event.

# DYNAMODB

Serverless No-SQL DB, key-value storage
We use this for state storage for our serverless services

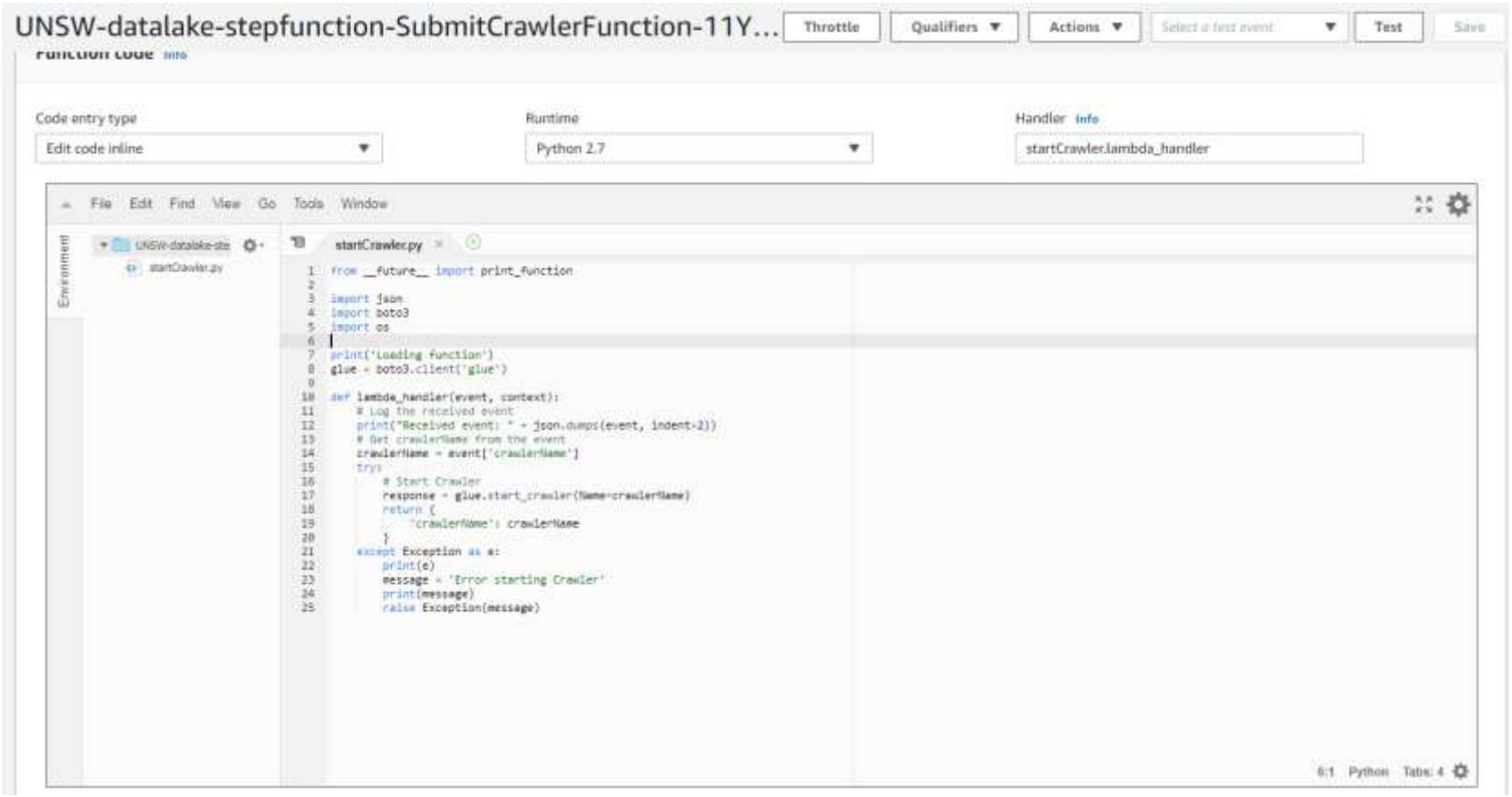
amazon
DynamoDB

# LAMBDA

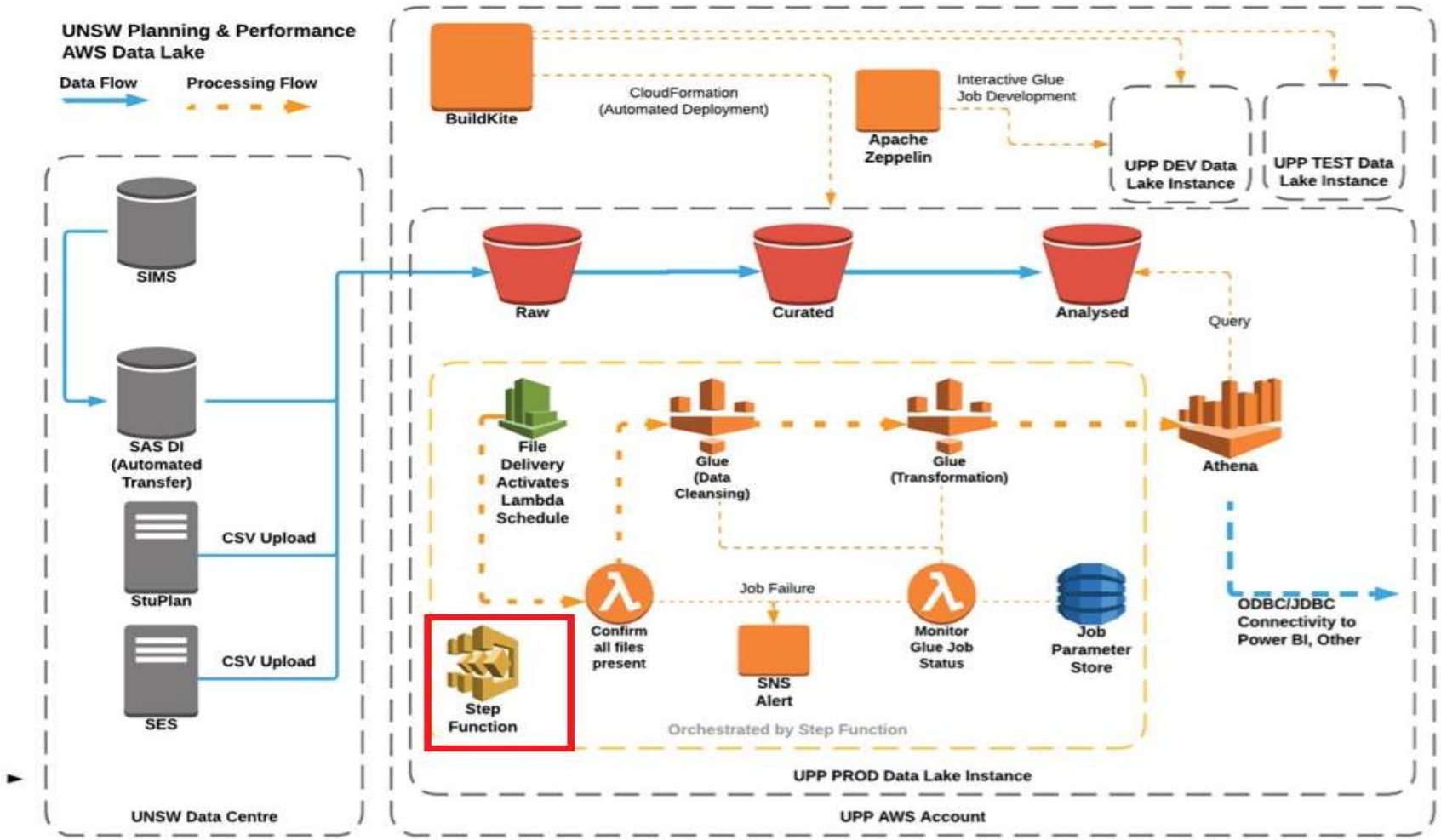

UNSW Planning & Performance
AWS Data Lake
Data Flow
Processing Flow
BuildKite
CloudFormation
(Automated Deployment)
Apache
Zeppelin
Interactive Glue
Job Development
UPP DEV Data
Lake Instance
UPP TEST Data
Lake Instance
SIMS
SAS DI
(Automated
Transfer)
StuPlan
CSV Upload
SES
CSV Upload
Raw
Curated
Analysed
Query
File
Delivery
Activates
Lambda
Schedule
Glue
(Data
Cleansing)
Glue
(Transformation)
Athena
Confirm
all files
present
Job Failure
Monitor
Glue Job
Status
Job
Parameter
Store
ODBC/JDBC
Connectivity to
Power BI, Other
Step
Function
SNS
Alert
Orchestrated by Step Function
UPP PROD Data Lake Instance
UNSW Data Centre
UPP AWS Account

# LAMBDA

Serverless function, per second billing. Can integrate with different AWS service e.g. S3, Step Function, SNS

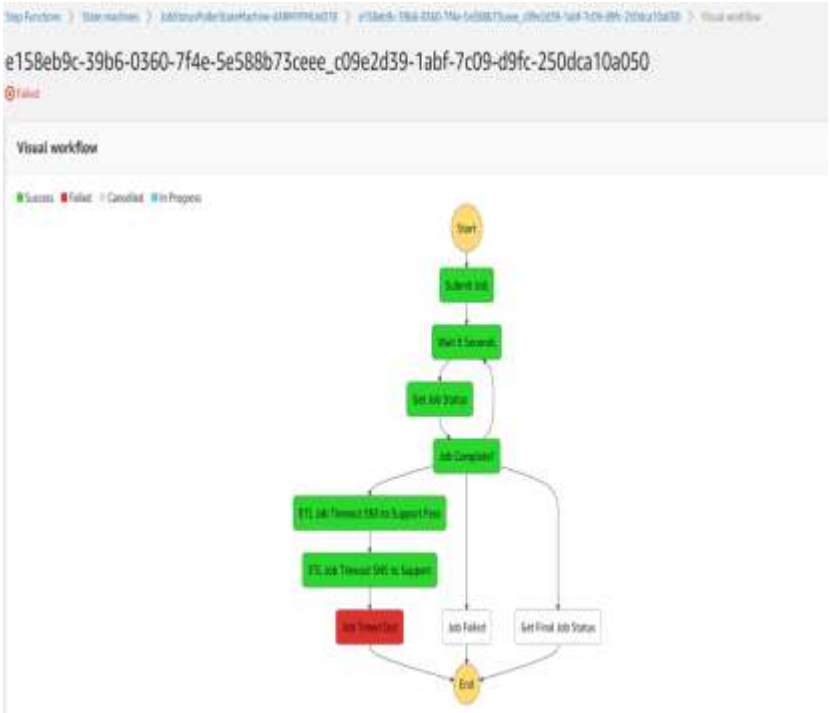


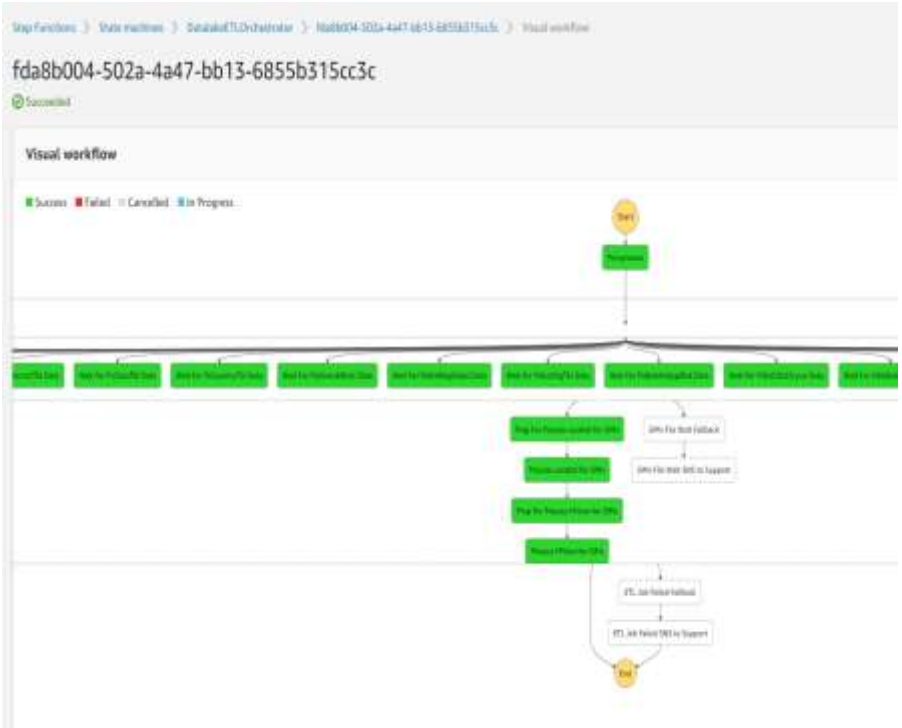
# STEP FUNCTION



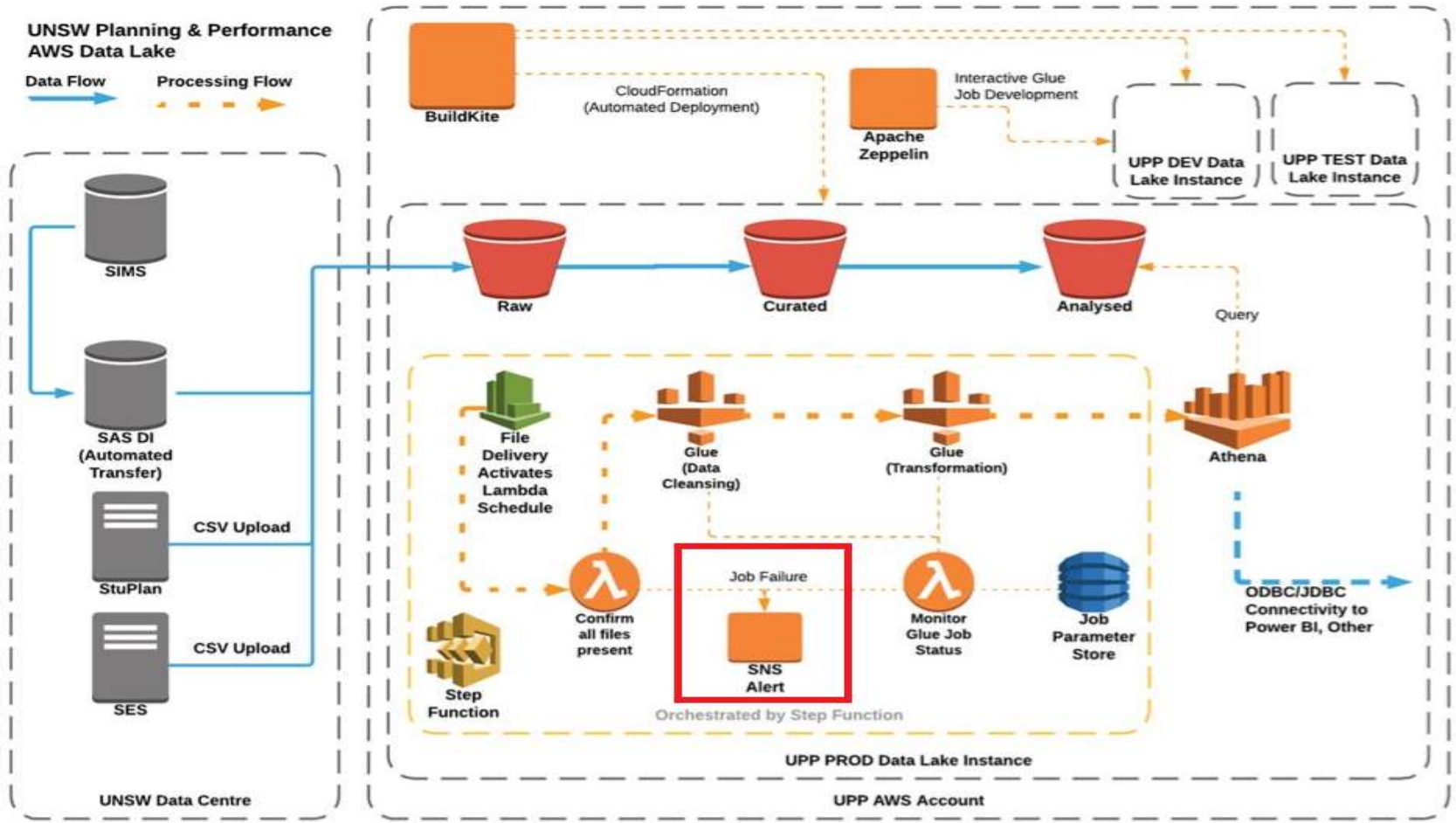
# STEP FUNCTION

AWS Step Functions lets you coordinate multiple AWS services into serverless workflows  We use step function in combination with S3 event to kick off and chain glue job, also with notification on time-out and failure



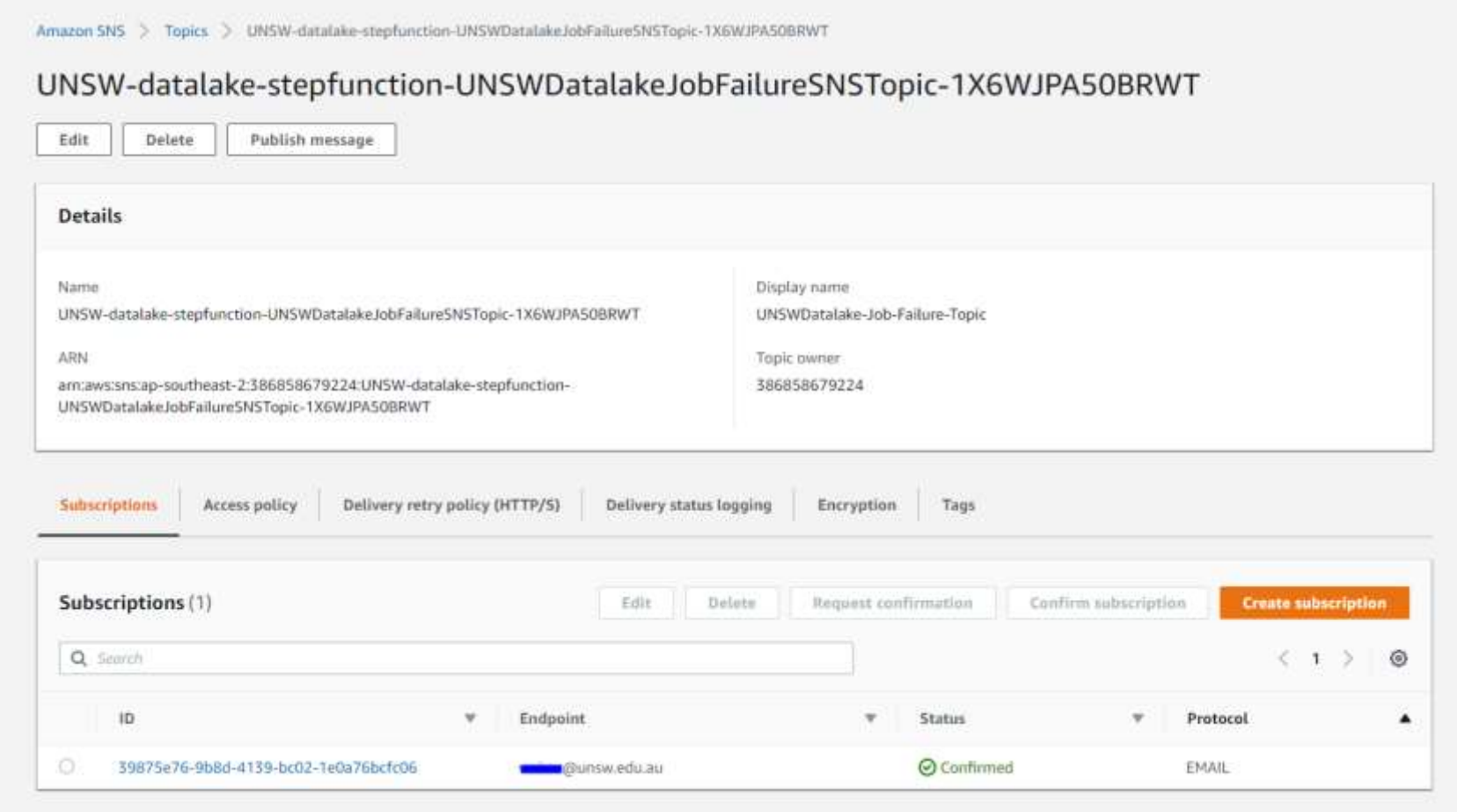



# SIMPLE NOTIFICATION SERVICE (SNS)



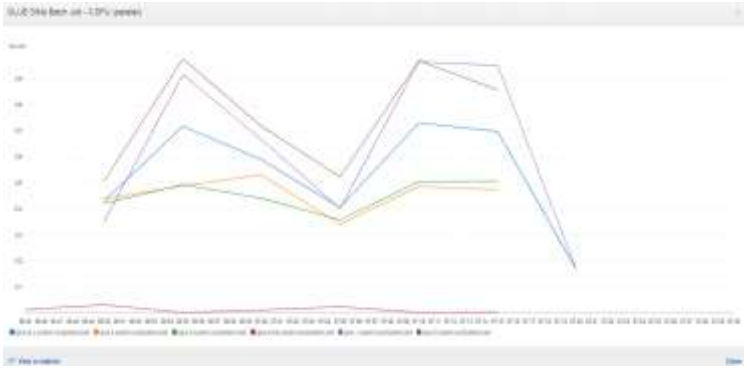
# SIMPLE NOTIFICATION SERVICE (SNS)

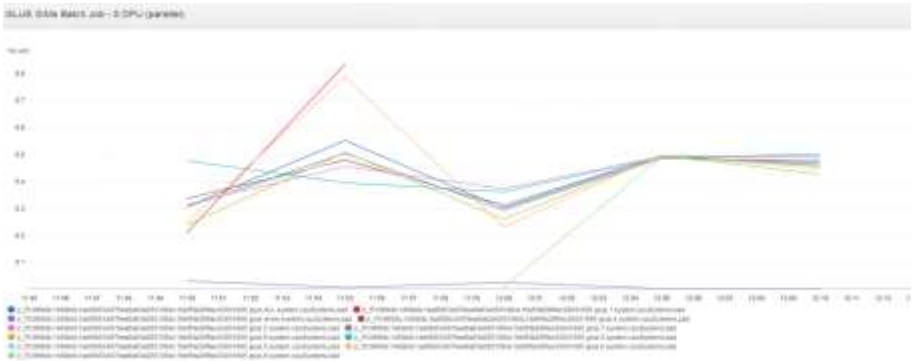
Failure Notifications



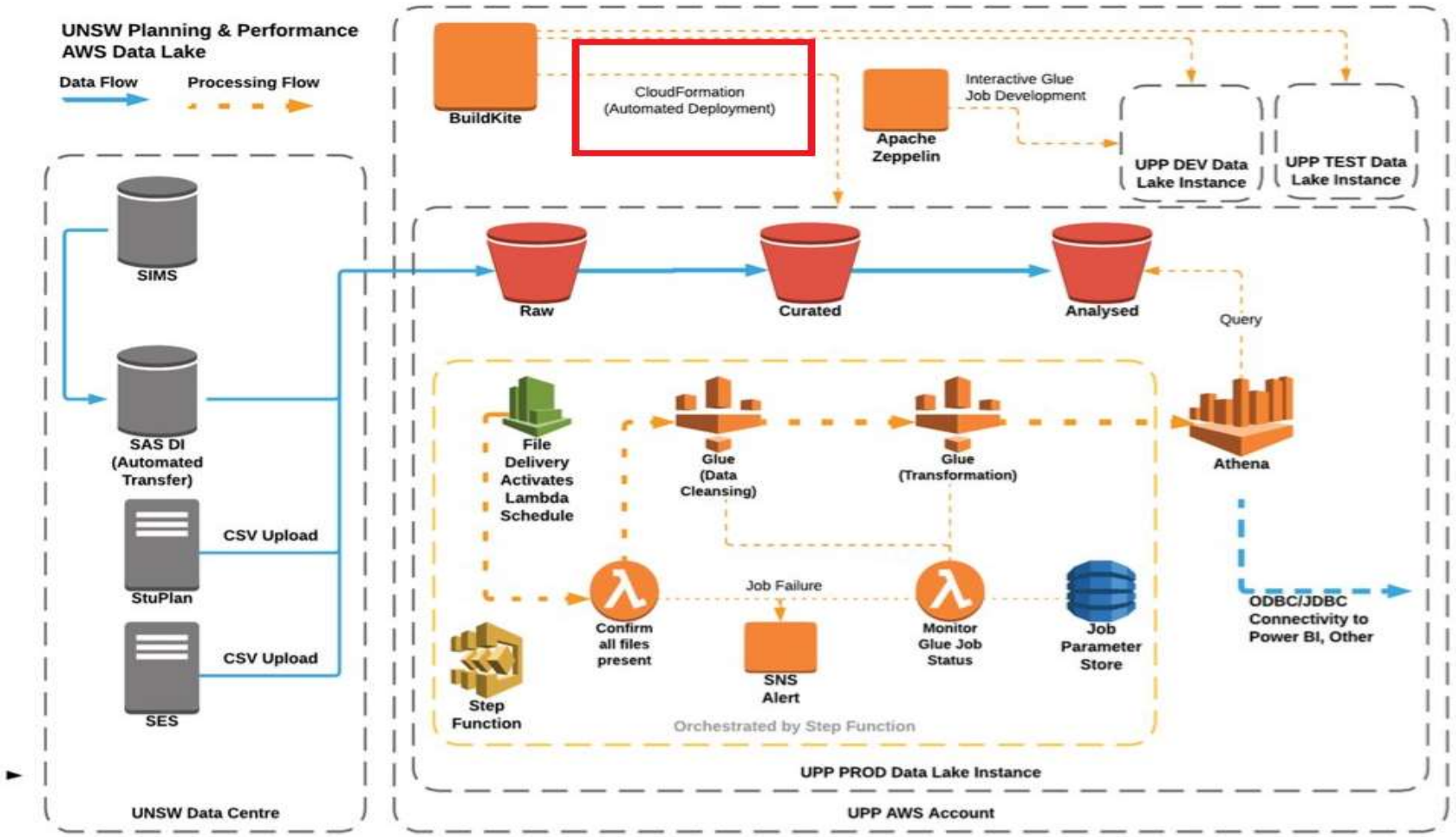
# CLOUDWATCH

Monitoring and log (Glue and Lambda)



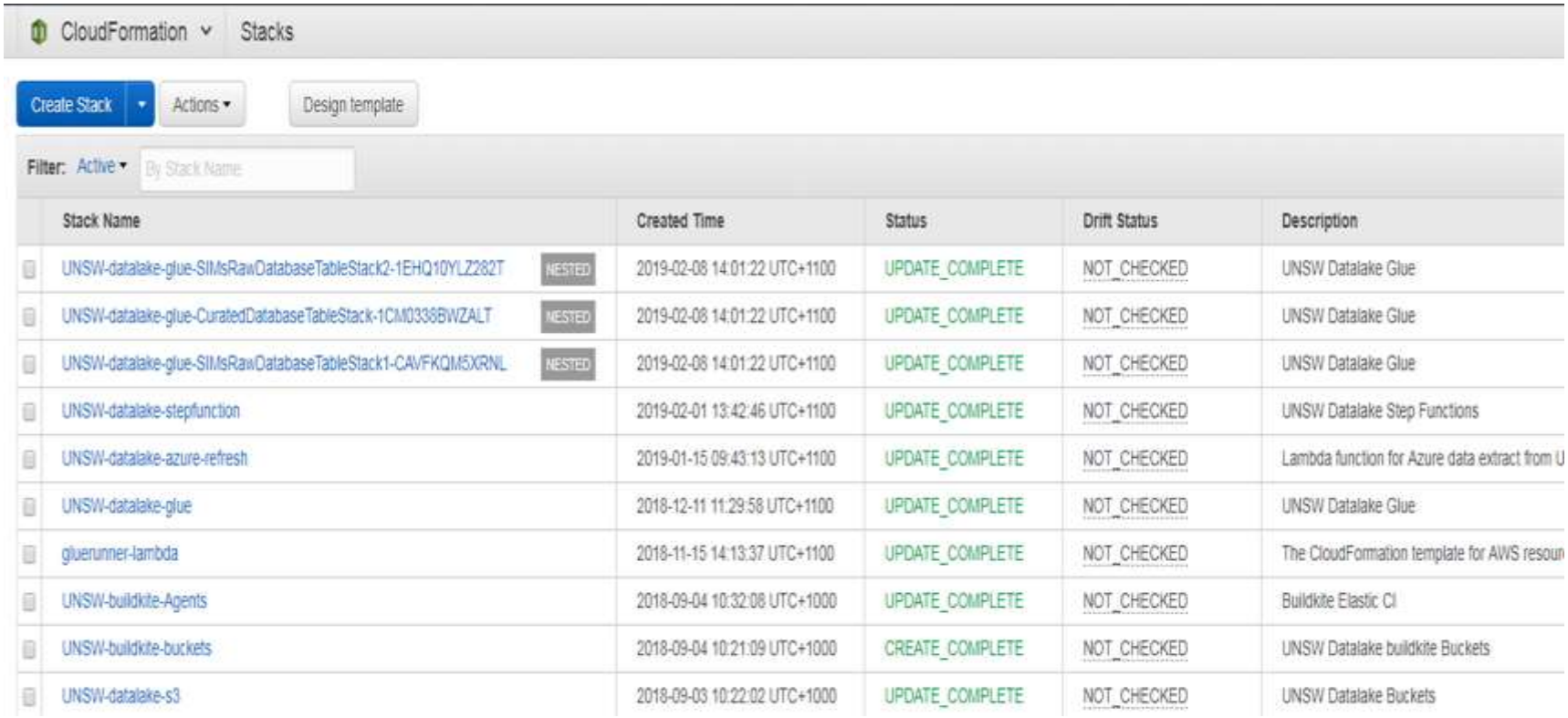


# CLOUDFORMATION



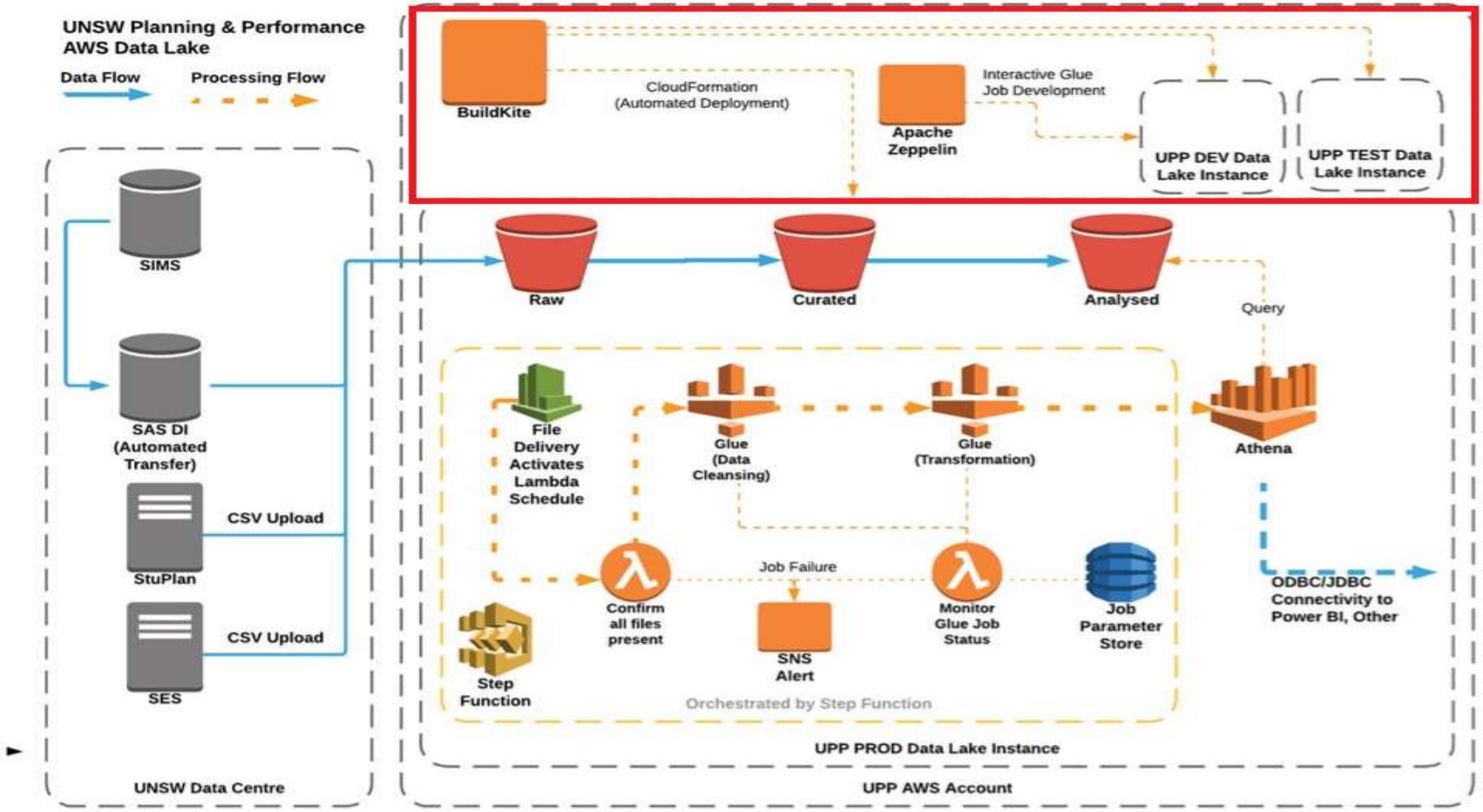
# CLOUDFORMATION

Code Templates for Deployment, Easily Control and Track Changes




| Stack Name | | Created Time | Status | Drift Status | Description |
|---|---|---|---|---|---|
| UNSW-datalake-glue-SIMsRawDatabaseTableStack2-1EHQ10YLZ282T | NESTED | 2019-02-08 14:01:22 UTC+1100 | UPDATE_COMPLETE | NOT_CHECKED | UNSW Datalake Glue |
| UNSW-datalake-glue-CuratedDatabaseTableStack-1CMI0336BWZALT | NESTED | 2019-02-08 14:01:22 UTC+1100 | UPDATE_COMPLETE | NOT_CHECKED | UNSW Datalake Glue |
| UNSW-datalake-glue-SIMsRawDatabaseTableStack1-CAVFKQM5XRNL | NESTED | 2019-02-08 14:01:22 UTC+1100 | UPDATE_COMPLETE | NOT_CHECKED | UNSW Datalake Glue |
| UNSW-datalake-stepfunction | | 2019-02-01 13:42:46 UTC+1100 | UPDATE_COMPLETE | NOT_CHECKED | UNSW Datalake Step Functions |
| UNSW-datalake-azure-refresh | | 2019-01-15 09:43:13 UTC+1100 | UPDATE_COMPLETE | NOT_CHECKED | Lambda function for Azure data extract from U |
| UNSW-datalake-glue | | 2018-12-11 11:29:58 UTC+1100 | UPDATE_COMPLETE | NOT_CHECKED | UNSW Datalake Glue |
| gluerunner-lambda | | 2018-11-15 14:13:37 UTC+1100 | UPDATE_COMPLETE | NOT_CHECKED | The CloudFormation template for AWS resour |
| UNSW-buildkite-Agents | | 2018-09-04 10:32:08 UTC+1000 | UPDATE_COMPLETE | NOT_CHECKED | Buildkite Elastic CI |
| UNSW-buildkite-buckets | | 2018-09-04 10:21:09 UTC+1000 | CREATE_COMPLETE | NOT_CHECKED | UNSW Datalake buildkite Buckets |
| UNSW-datalake-s3 | | 2018-09-03 10:22:02 UTC+1000 | UPDATE_COMPLETE | NOT_CHECKED | UNSW Datalake Buckets |

# OTHER SERVICES AND TOOLS



# OTHER SERVICES AND TOOLS

- BitBucket
- CI/CD
- BuildKite
- Zeppelin Notebook (Spark Script development and interact with Glue Dev Endpoint)

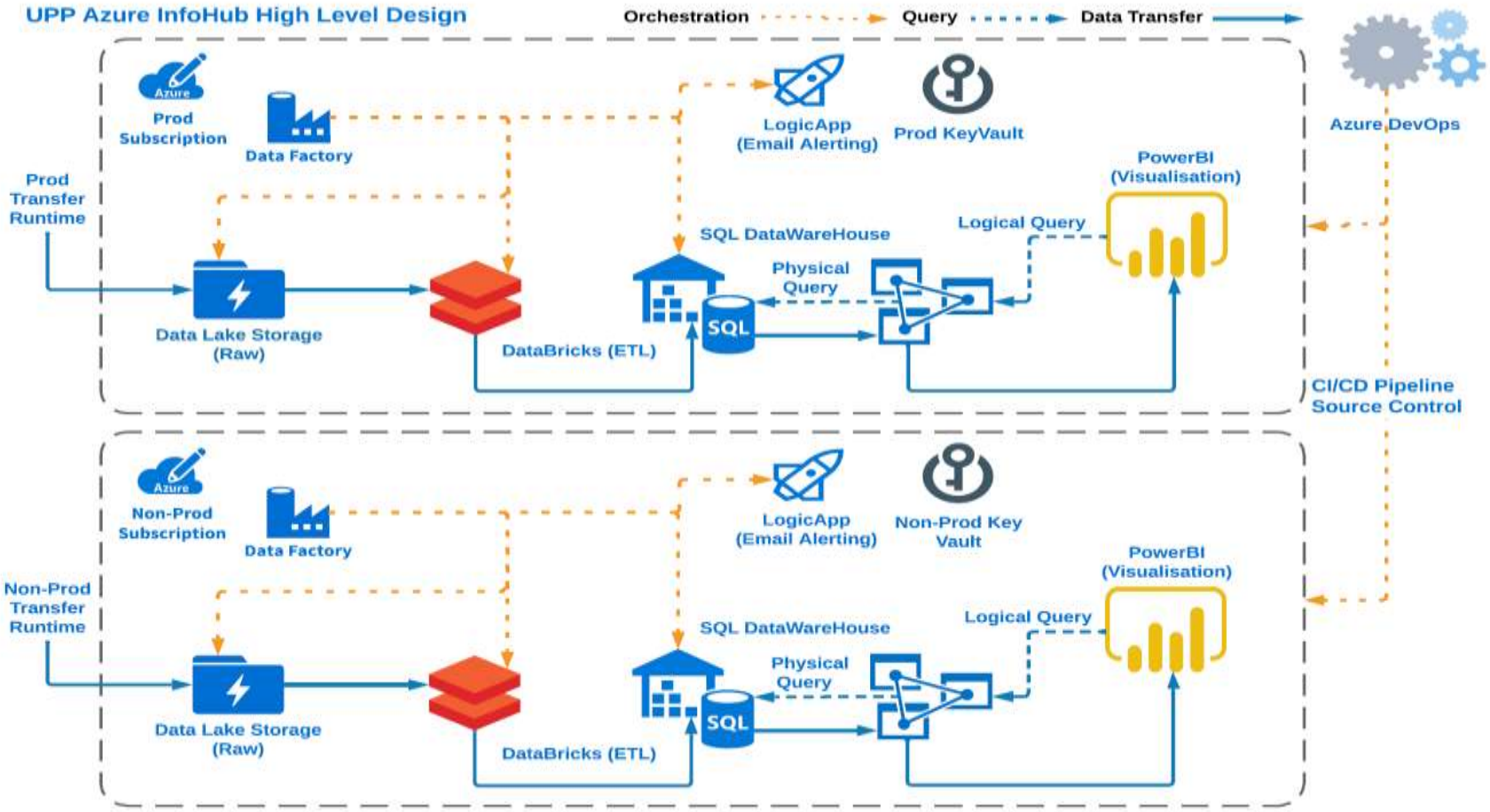
# ARCHITECTURAL CONSIDERATIONS

- **EMR vs Glue + Athena + Step Function + Lambda**
  - UNSW has minimal admin staff and does not want to maintain a hadoop cluster
  - Lower cost
- **Glue Job Batching**
  - Glue has minimum cost of 10 mins, some small job runs less than 1 min. Batched up to save cost.
- **Visualization Tool**
  - Quicksight - lacks of Important features such as
    - Code promotion
    - Object ownership
    - Visualisation Features

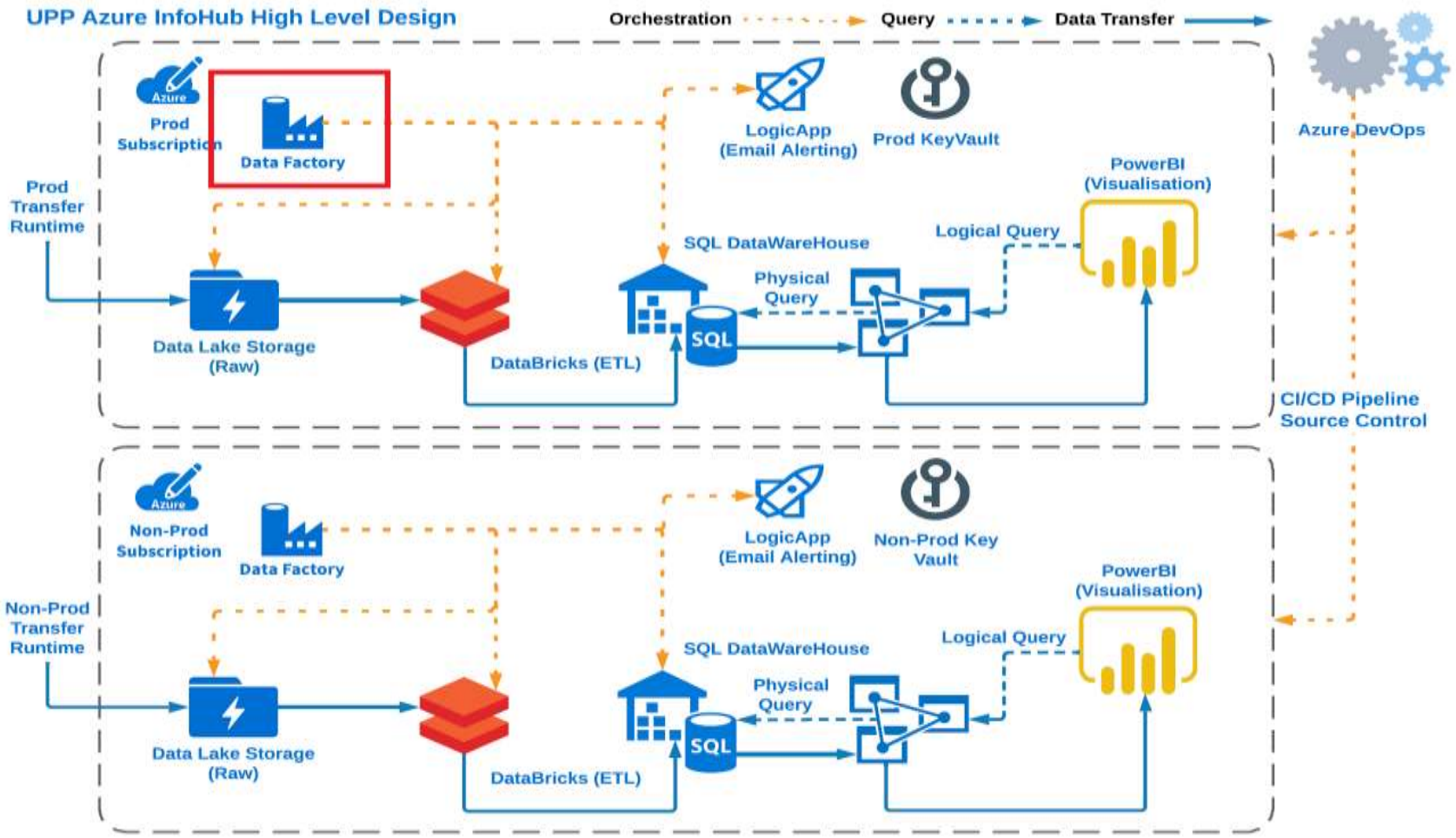
# KEY ACHIEVEMENTS

- Created the first **visual pipeline** of Students into the university, breaking down applicants by country of origin, age, degree and other data points through the Uni Application process as far back as the first student enrolled in the university in 1950 - a Mining Engineering Major.
- Established first true **data lake** for UPP storing raw, curated and analysed data from both source systems and legacy data marts in a central pipeline for easy consumption as a service.
- Purely **serverless** implementation of a data lake with all functions controlled through **event-driven orchestration**, ensuring lowest cost per workload.
- Implemented **CI/CD tooling** and **SCM** for the data lake with BuildKite, allowing one-click deployments
- Illustrated an example of **serverless cost-of-exit** - with a new DWH being built, data extracts could be provided in a matter of days via ODBC connection to Azure from Athena, facilitating transition.
- UPP team **uplifted skills** in Python and SQL etc.

# FUTURE ROADMAP - UPP AZURE INFOHUB

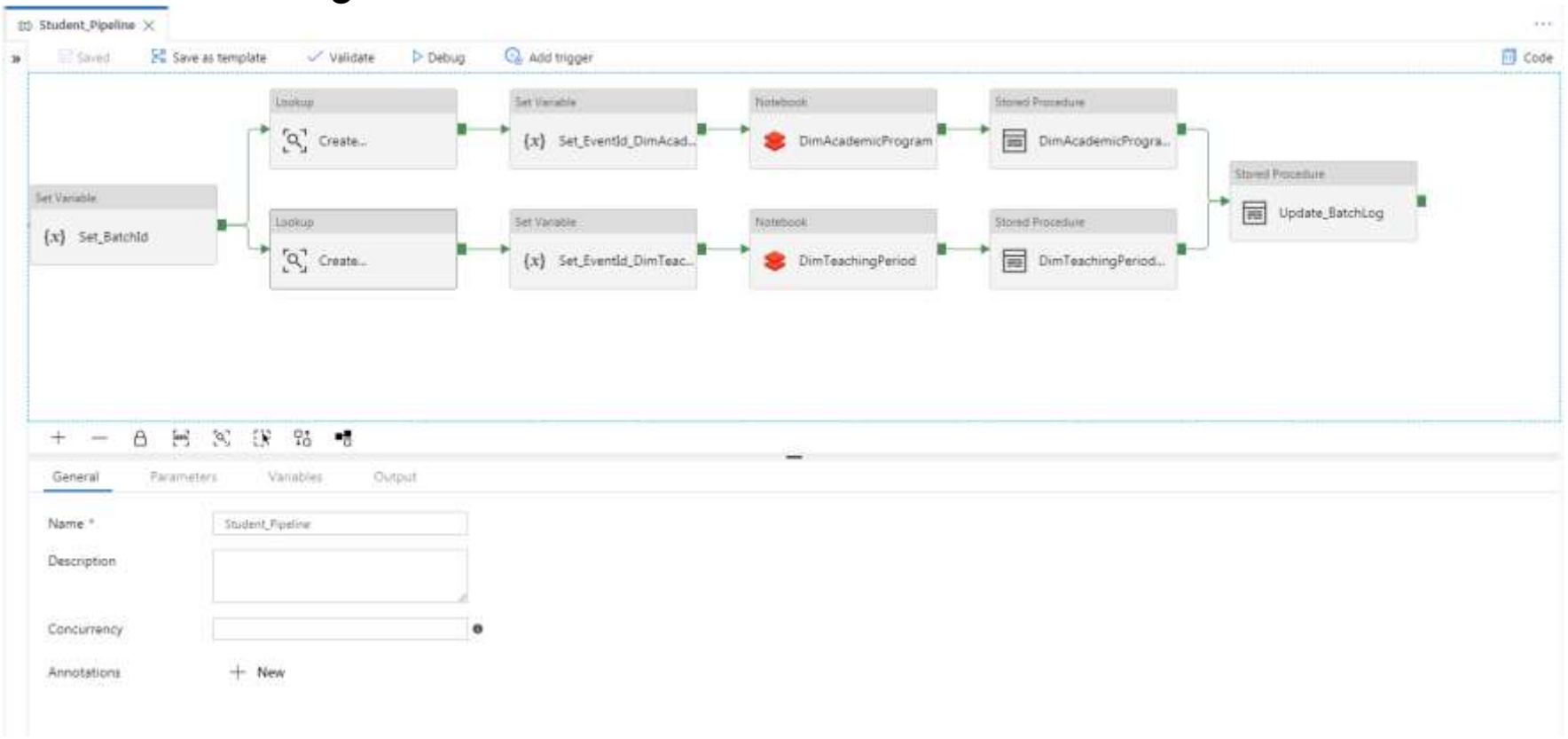


# AZURE DATA FACTORY

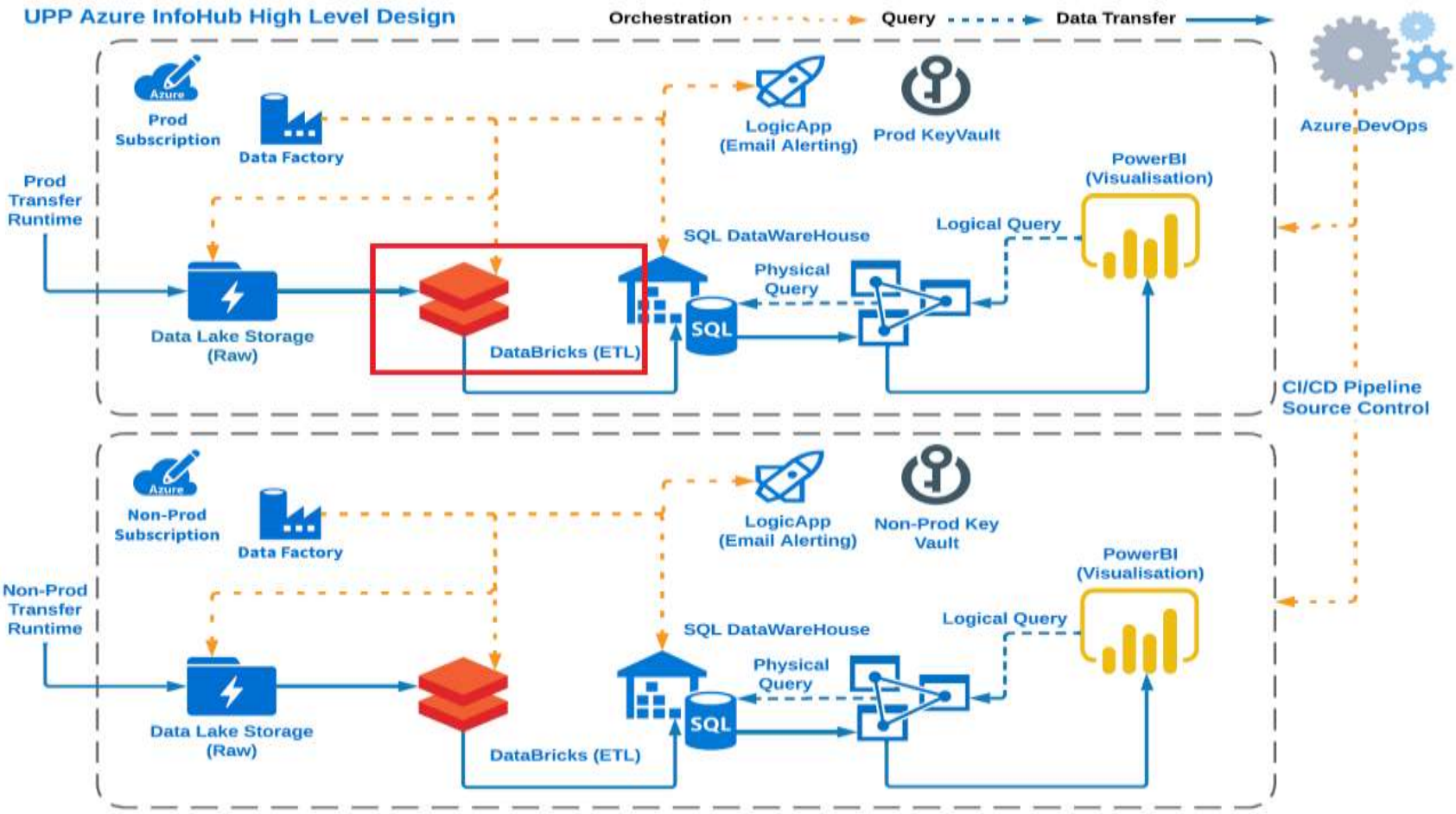


# AZURE DATA FACTORY

- ETL and ELT processes code-free or write your own code
- Natively built and maintenance-free connectors
- serverless integration service

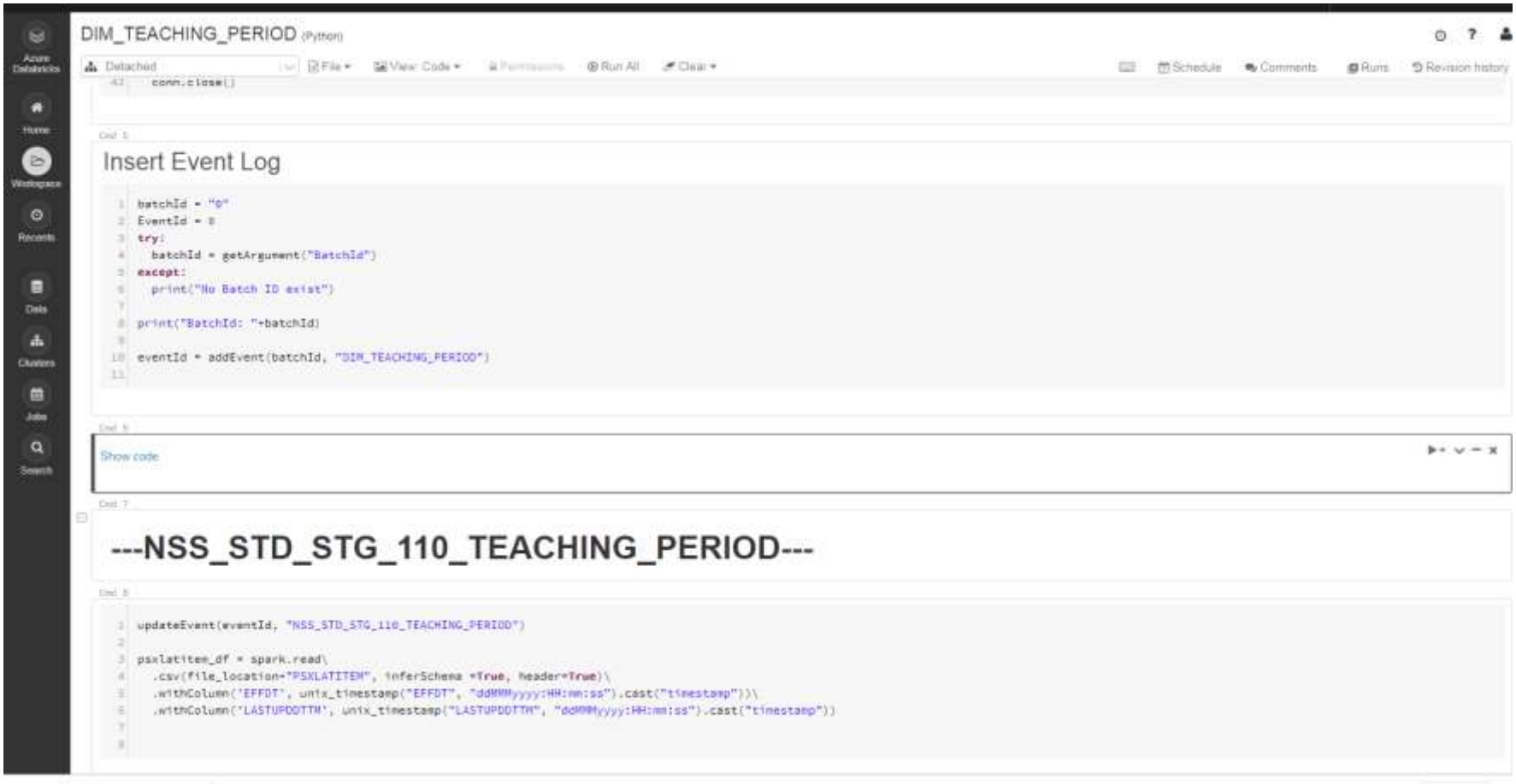


# AZURE DATABRICKS

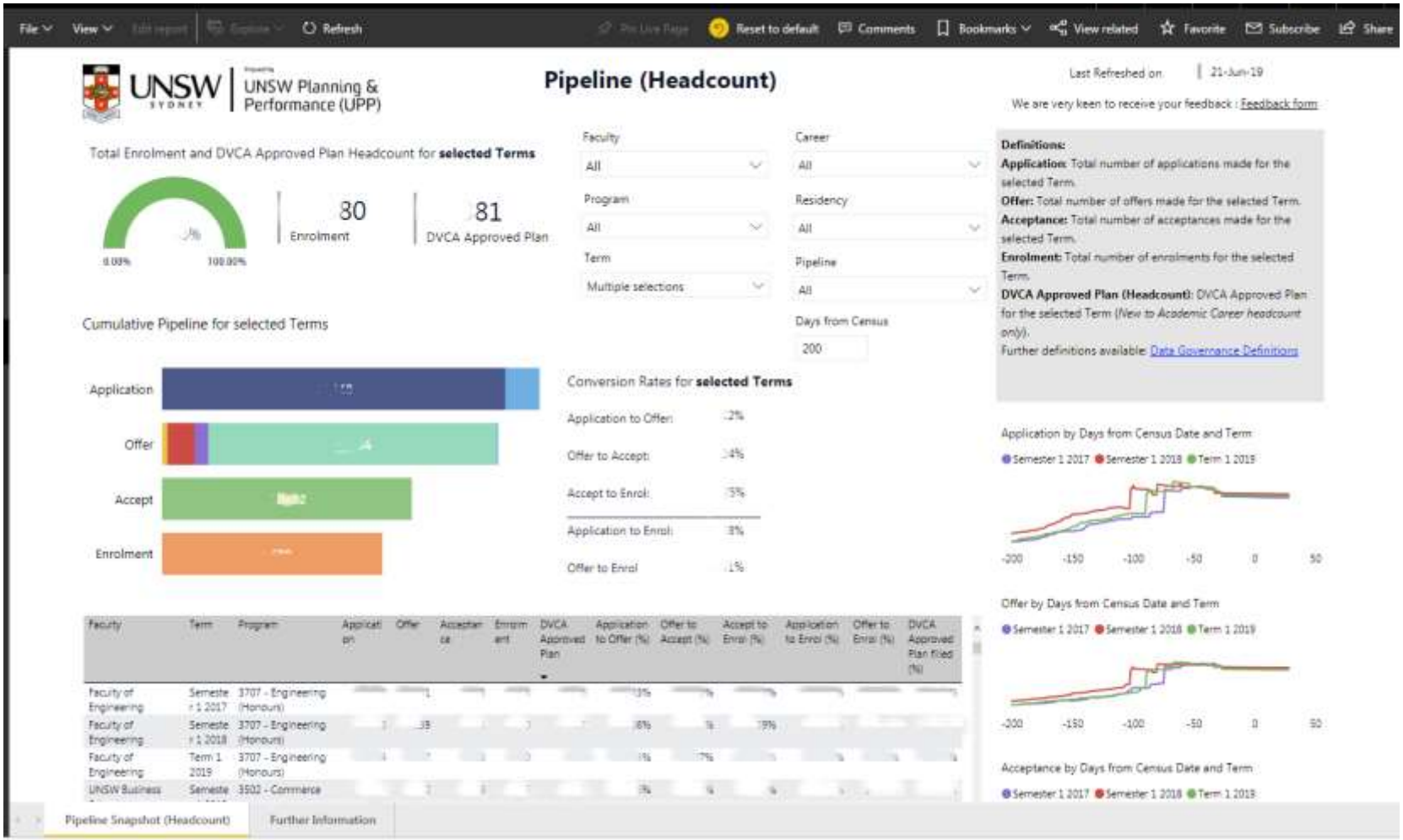


# AZURE DATABRICKS

Spark-based analytics service



# POWERBI


File
View
Refresh
Reset to default
Comments
Bookmarks
View related
Favorite
Subscribe
Share
UNSW SYDNEY
UNSW Planning & Performance (UPP)
Pipeline (Headcount)
Last Refreshed on 21-Jun-19
We are very keen to receive your feedback : Feedback form
Total Enrolment and DVCA Approved Plan Headcount for selected Terms
30
Enrolment
81
DVCA Approved Plan
Faculty
All
Career
All
Program
All
Residency
All
Term
Multiple selections
Pipeline
All
Days from Census
200
Definitions:
Application: Total number of applications made for the selected Term.
Offer: Total number of offers made for the selected Term.
Acceptance: Total number of acceptances made for the selected Term.
Enrolment: Total number of enrolments for the selected Term.
DVCA Approved Plan (Headcount): DVCA Approved Plan for the selected Term (New to Academic Career headcount only).
Further definitions available: Data Governance Definitions
Cumulative Pipeline for selected Terms
Application
Offer
Accept
Enrolment
Conversion Rates for selected Terms
Application to Offer:
Offer to Accept:
Accept to Enrol:
Application to Enrol:
Offer to Enrol
Application by Days from Census Date and Term
Semester 1 2017
Semester 1 2018
Term 1 2019
Offer by Days from Census Date and Term
Acceptance by Days from Census Date and Term
Faculty
Term
Program
Application
Offer
Acceptance
Enrolment
DVCA Approved Plan
Application to Offer (%)
Offer to Accept (%)
Accept to Enrol (%)
Application to Enrol (%)
Offer to Enrol (%)
DVCA Approved Plan filled (%)
Faculty of Engineering
Semester 1 2017
3707 - Engineering (Honours)
Faculty of Engineering
Semester 1 2018
3707 - Engineering (Honours)
Faculty of Engineering
Term 1 2019
3707 - Engineering (Honours)
UNSW Business
Semester
3502 - Commerce
Pipeline Snapshot (Headcount)
Further Information

# AWS VS AZURE

| AWS | Azure | Description |
|---|---|---|
| Step Function, AWS Glue | Data Factory | Move data between storage. Schedule and orchestrate |
| Amazon Athena | Azure data Lake Analytics | Serverless interactive query service |
| Glue | Azure Databricks | Spark-based analytics |
| SNS | LogicApp | Sent out alerting email |
| CloudWatch | Azure Monitor | Track performance and create logs |
| Parameter Store | Key Vault | Store parameters |
| Simple Storage Service(S3) | Azure Blob storage | Object storage |
| QuickSight | PowerBI | BI visualisation tools |


Q
A