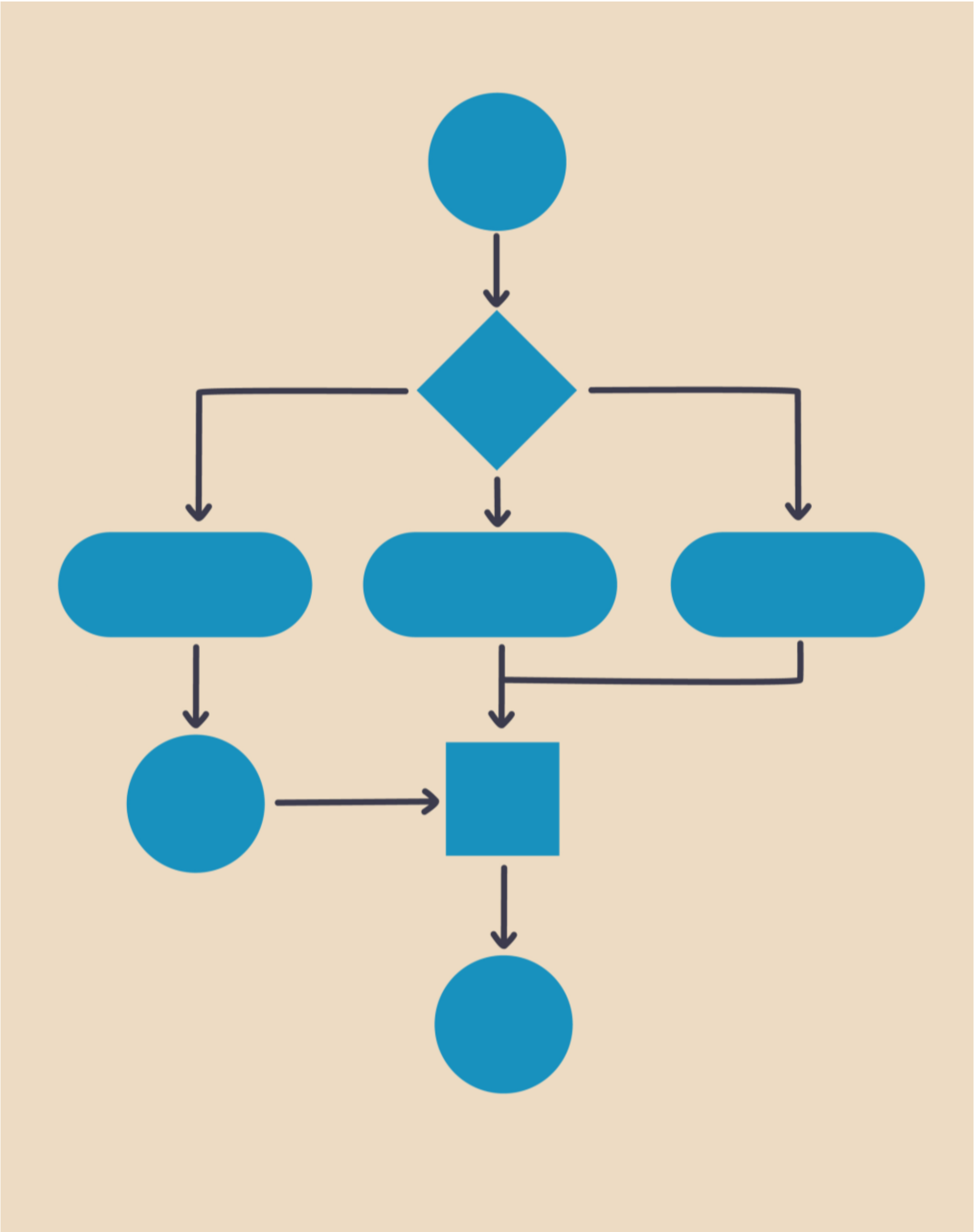
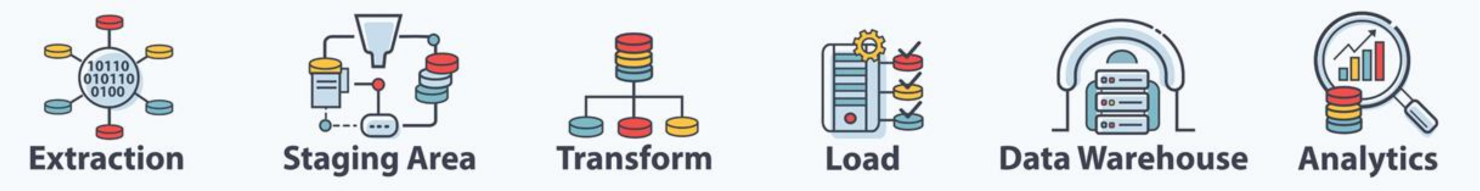
Orchestration of Data Pipeline

Orchestration of Data Pipeline



- Automation for data movement from multiple storage locations into a centralised repository.



- May include various tools, services, triggers, paths, sub-paths, schedules.
- May involve sequential, parallel processing steps.
- Should be a robust system that can handle failures, retries.

Orchestration Service in AWS

Orchestration Service in AWS

AWS Glue Workflows

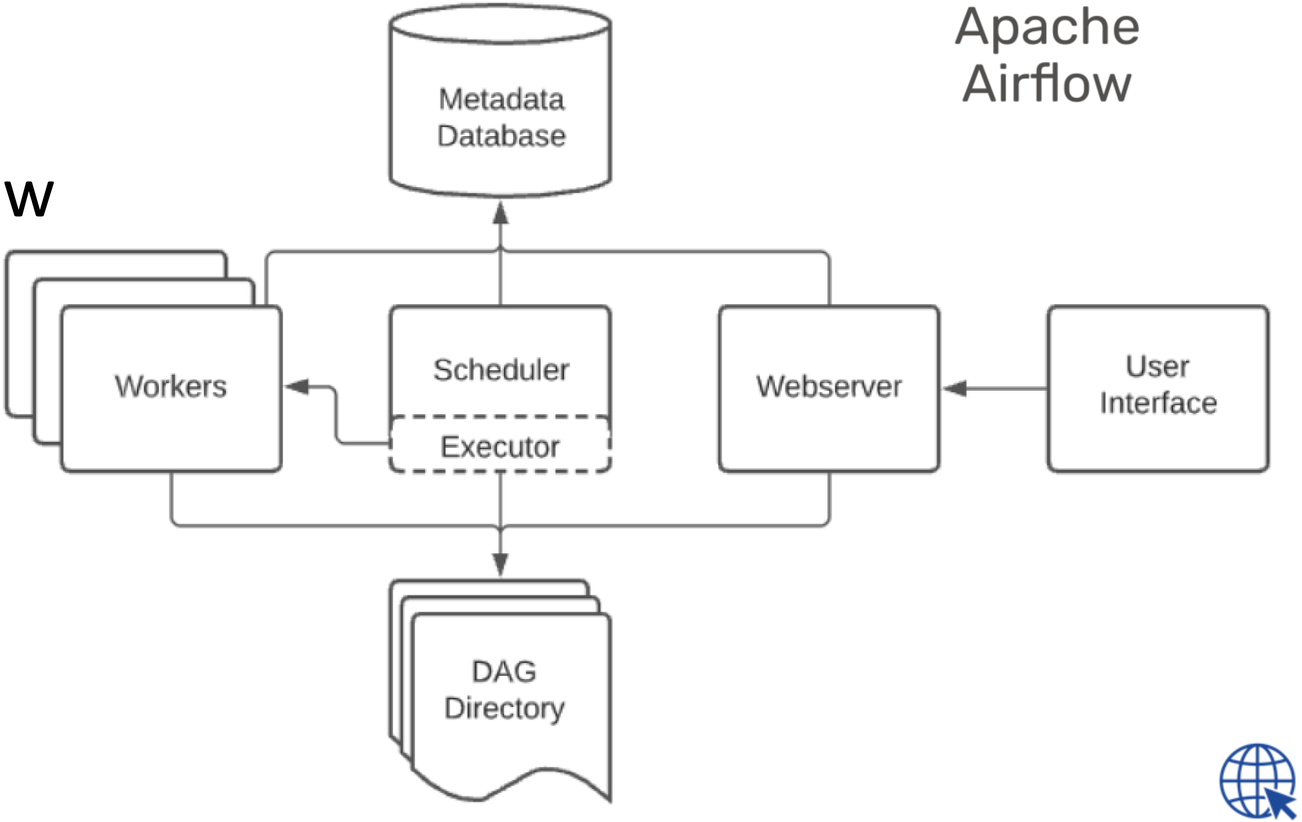
AWS Step Functions

Amazon Managed Workflows for Apache Airflow (Amazon MWAA)

Apache Airflow

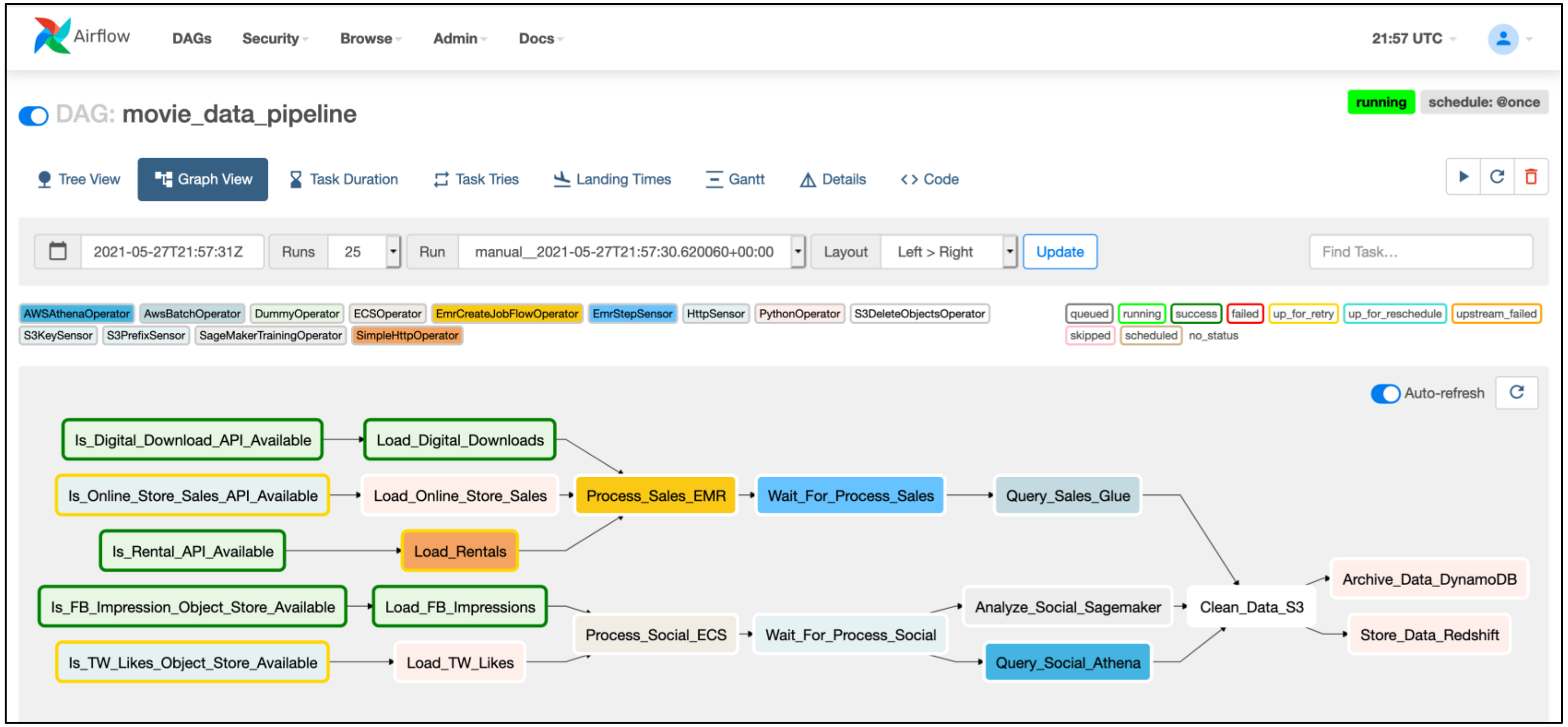
- Apache Airflow is an open-source platform for developing, scheduling, and monitoring batch-oriented workflows.
- Key Use Cases for Apache Airflow
  - ETL
  - AI/ML
  - DevOps
- Main Components of Apache Airflow
  - Meta Database
  - Scheduler
  - Worker
  - Web Server

Apache Airflow



Apache Airflow

- In Airflow, a DAG – or a Directed Acyclic Graph – is a workflow: a collection of all the tasks you want to run, organized to reflect their relationships and dependencies.
- A DAG is defined in a Python script, which represents the tasks and their dependencies as code.



DAG: movie_data_pipeline

running schedule: @once

2021-05-27T21:57:31Z Runs 25 Run manual__2021-05-27T21:57:30.620060+00:00 Layout Left > Right

Challenges with Self-Managed Apache Airflow

Setup

Scaling

Security

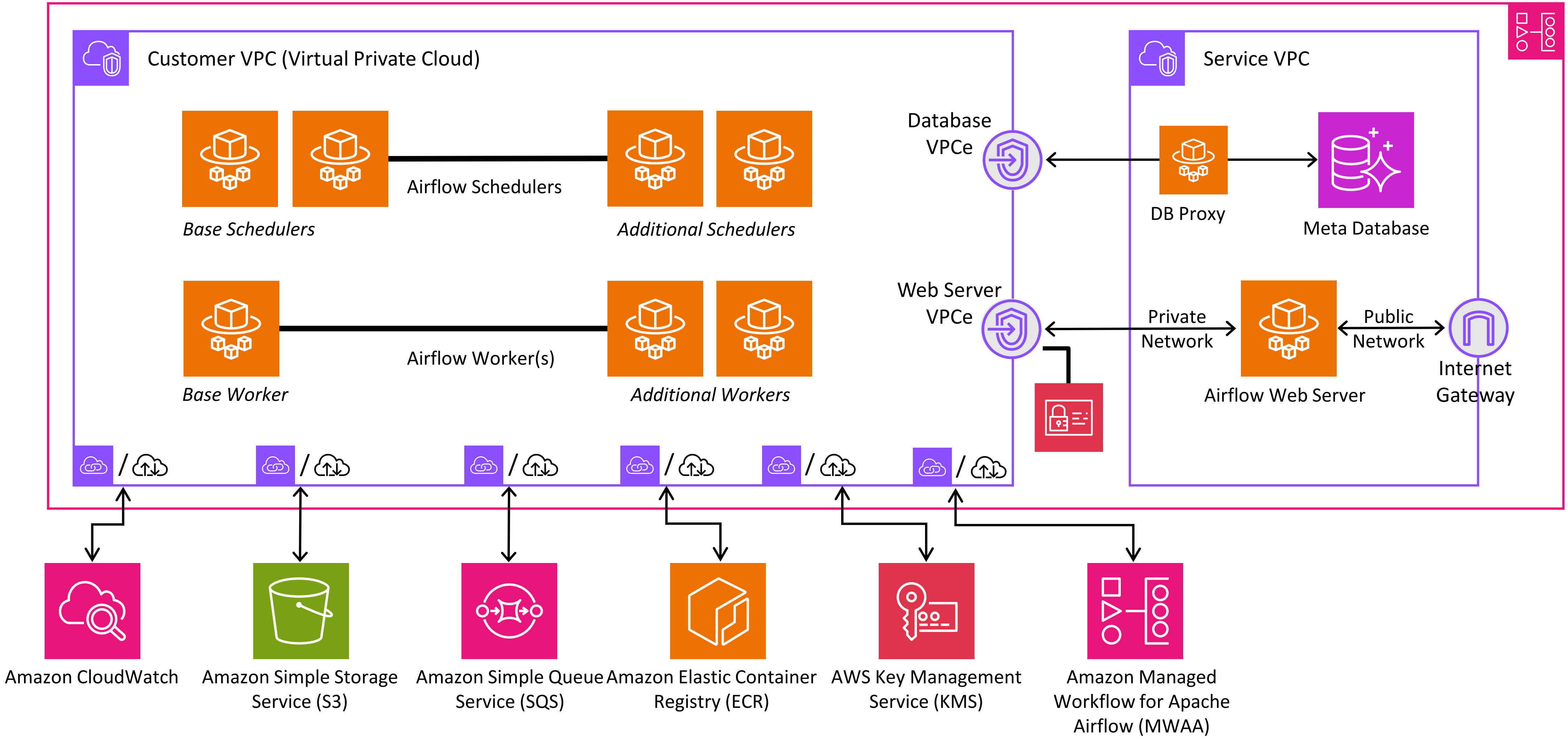
Upgrade

Maintenance

Amazon Managed Workflows for Apache Airflow (MWAA)

- A managed service for Apache Airflow, making it easy for data engineers and data scientists to invoke data processing workflows on AWS.
- MWAA is 100% Open Source. It is not a Fork or "branch" of Airflow.
- How Amazon MWAA Helps?
  - Deployments and Operations
    - Easy to Set Up and Maintain
  - Availability and Sizing
    - Multi-AZ for HA with Airflow on ECS Fargate
  - Scaling
    - Auto Scaling and Celery Executor
  - Security
    - IAM and VPC

Amazon MWAA Architecture



Comparison	AWS Glue Workflow	AWS Step Function	Amazon Managed Workflows for Apache Airflow (Amazon MWAA)
Terminology	Workflows	State Machine	Directed Acyclic Graph (DAG)
Deployment	Serverless	Serverless	Runs on managed servers
Authoring	Visual Designer SDK, CDK	Visual Designer SDK, CDK, State Language	Python
Type	Purpose built	General purpose	General purpose
High availability	Built-in	Built-in	Configurable across AZs
Integration	Primarily integrate with Glue services (crawlers, jobs)	Out-of-box integration with various AWS Services	Can be integrated with AWS Services and third-party tools as Apache Hadoop, Presto, Hive and Spark
Suitability	Glue specific orchestration	Integration with AWS Services	Migration from On-prem Airflow Prioritization of open source
Pricing	No charges for workflows. You pay of underlying Glue services.	Charged based on state transition (standard workflow) or duration (express workflow)	Infrastructure charges (compute and storage)

Amazon AppFlow

Amazon AppFlow

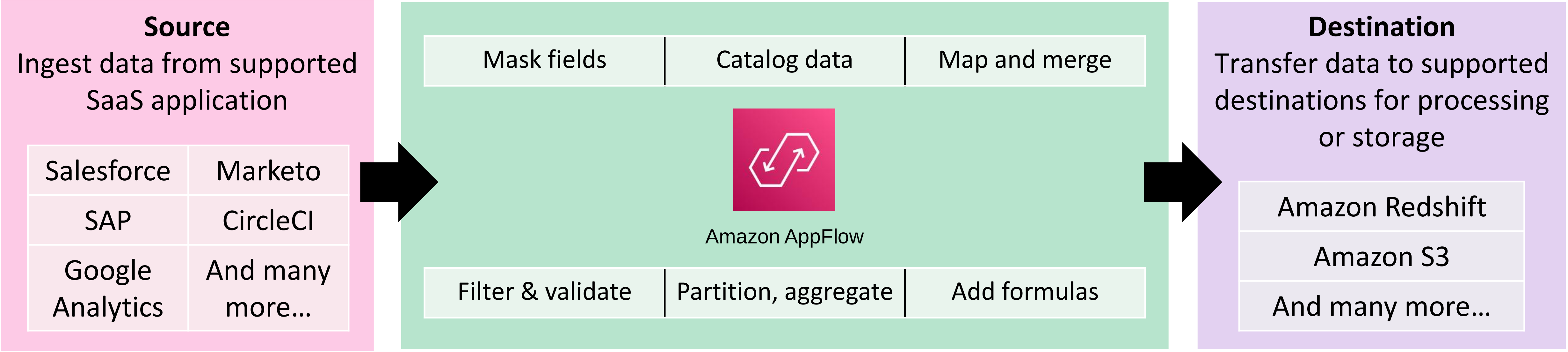
- A fully managed integration service that enables you to securely transfer data between SaaS applications like Salesforce, SAP, Zendesk, Slack, and ServiceNow, and AWS services like Amazon S3 and Amazon Redshift.

NO / LOW CODE

COST SAVINGS

SPEED & AGILITY

SECURE & SCALABLE



Source

Ingest data from supported SaaS application

Salesforce	Marketo
SAP	CircleCI
Google Analytics	And many more…

Mask fields	Catalog data	Map and merge

Filter & validate	Partition, aggregate	Add formulas

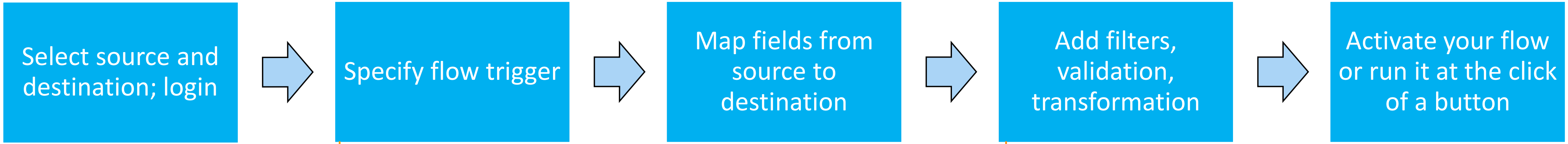
Destination

Transfer data to supported destinations for processing or storage

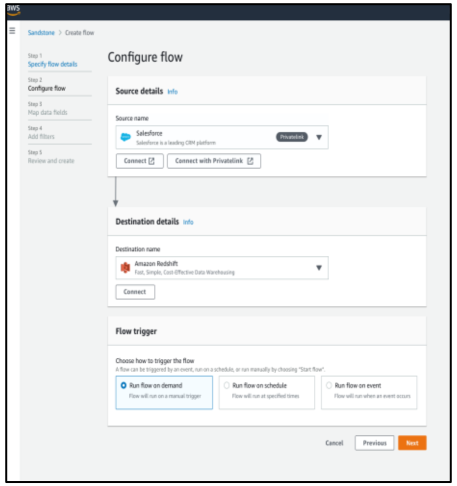
Amazon Redshift
Amazon S3
And many more…

How it works?

Creating a flow



Select source and destination; login → Specify flow trigger → Map fields from source to destination → Add filters, validation, transformation → Activate your flow or run it at the click of a button



Sandstone > Create flow

Step 1 Specify flow details
Step 2 Configure flow
Step 3 Map data fields
Step 4 Add filters
Step 5 Review and create

Configure flow

Source details Info

Source name: Salesforce — Salesforce is a leading CRM platform (Privatelink)

Connect | Connect with Privatelink

Destination details Info

Destination name: Amazon Redshift — Fast, Simple, Cost-Effective Data Warehousing

Connect

Flow trigger

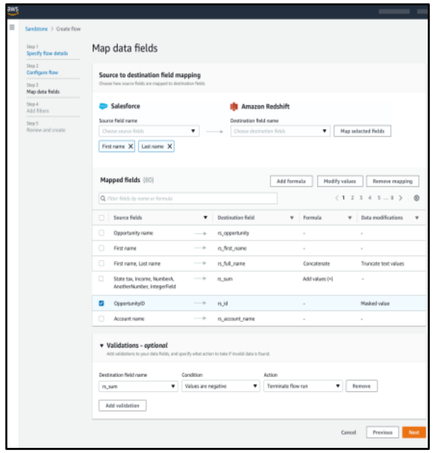
Choose how to trigger the flow
A flow can be triggered by an event, run on a schedule, or run manually by choosing "Start flow".

● Run flow on demand — Flow will run on a manual trigger
○ Run flow on schedule — Flow will run at specified times
○ Run flow on event — Flow will run when an event occurs

Cancel | Previous | Next

Specify flow trigger:

- Event Based
- Scheduled (1 minute minimum)
- On-demand (run flow immediately)



Sandstone > Create flow

Step 1 Specify flow details
Step 2 Configure flow
Step 3 Map data fields
Step 4 Add filters
Step 5 Review and create

Map data fields

Source to destination field mapping
Choose how source fields are mapped to destination fields

Salesforce → Amazon Redshift

Source field name: Choose source fields → Destination field name: Choose destination fields | Map selected fields

First name ✕ | Last name ✕

Mapped fields (80) — Add formula | Modify values | Remove mapping

Filter fields by name or formula — 1 2 3 4 5 ... 8

	Source fields		Destination field	Formula	Data modifications
□	Opportunity name	→	rs_opportunity	-	-
□	First name	→	rs_first_name	-	-
□	First name, Last name	→	rs_full_name	Concatenate	Truncate text values
□	State tax, Income, NumberA, AnotherNumber, IntegerField	→	rs_sum	Add values (+)	-
☑	OpportunityID	→	rs_id	-	Masked value
□	Account name	→	rs_account_name	-	-

Validations - optional
Add validations to your data fields, and specify what action to take if invalid data is found.

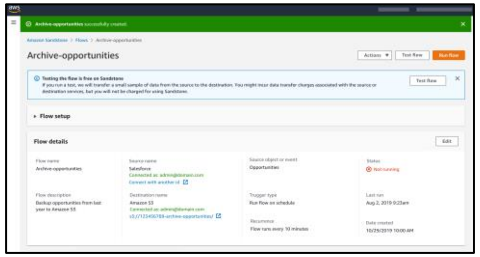
Destination field name: rs_sum | Condition: Values are negative | Action: Terminate flow run | Remove

Add validation

Cancel | Previous | Next

Add filters, validation, transformation:

- Filter records
- Mask sensitive fields
- Combine fields to create new ones
- Validate records
- Truncate fields



Archive-opportunities successfully created

Amazon Sandstone > Flows > Archive-opportunities

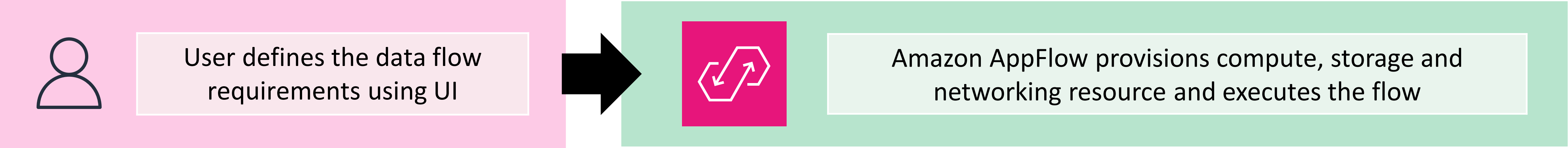
Archive-opportunities — Actions | Test flow | Run flow

Testing the flow is free on Sandstone
If you run a test, we will transfer a small sample of data from the source to the destination. You might incur data transfer charges associated with the source or destination services, but you will not be charged for using Sandstone. — Test flow

Flow setup

Flow details — Edit

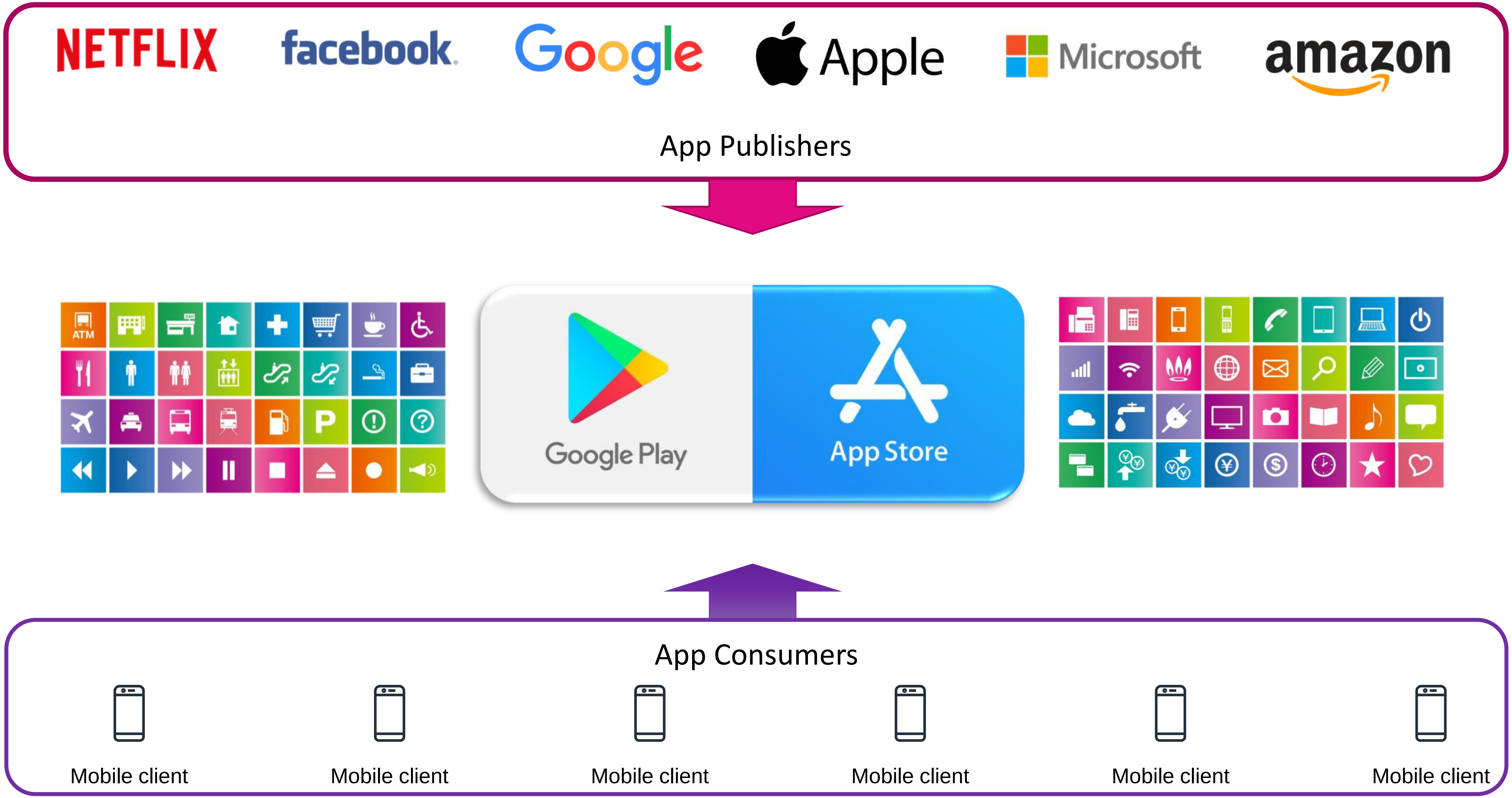
Flow name: Archive-opportunities
Source name: Salesforce — Connected as: admin@domain.com — Connect with another id
Source object or event: Opportunities
Status: Not running
Flow description: Backup opportunities from last year to Amazon S3
Destination name: Amazon S3 — Connected as: admin@domain.com — s3://123456789-archive-opportunities/
Trigger type: Run flow on schedule
Recurrence: Flow runs every 10 minutes
Last run: Aug 2, 2019 9:23am
Date created: 10/25/2019 10:00 AM



User defines the data flow requirements using UI → Amazon AppFlow provisions compute, storage and networking resource and executes the flow

AWS Data Exchange

How we use Mobile applications?



NETFLIX facebook Google Apple Microsoft amazon

App Publishers

Google Play

App Store

App Consumers

Mobile client

Mobile client

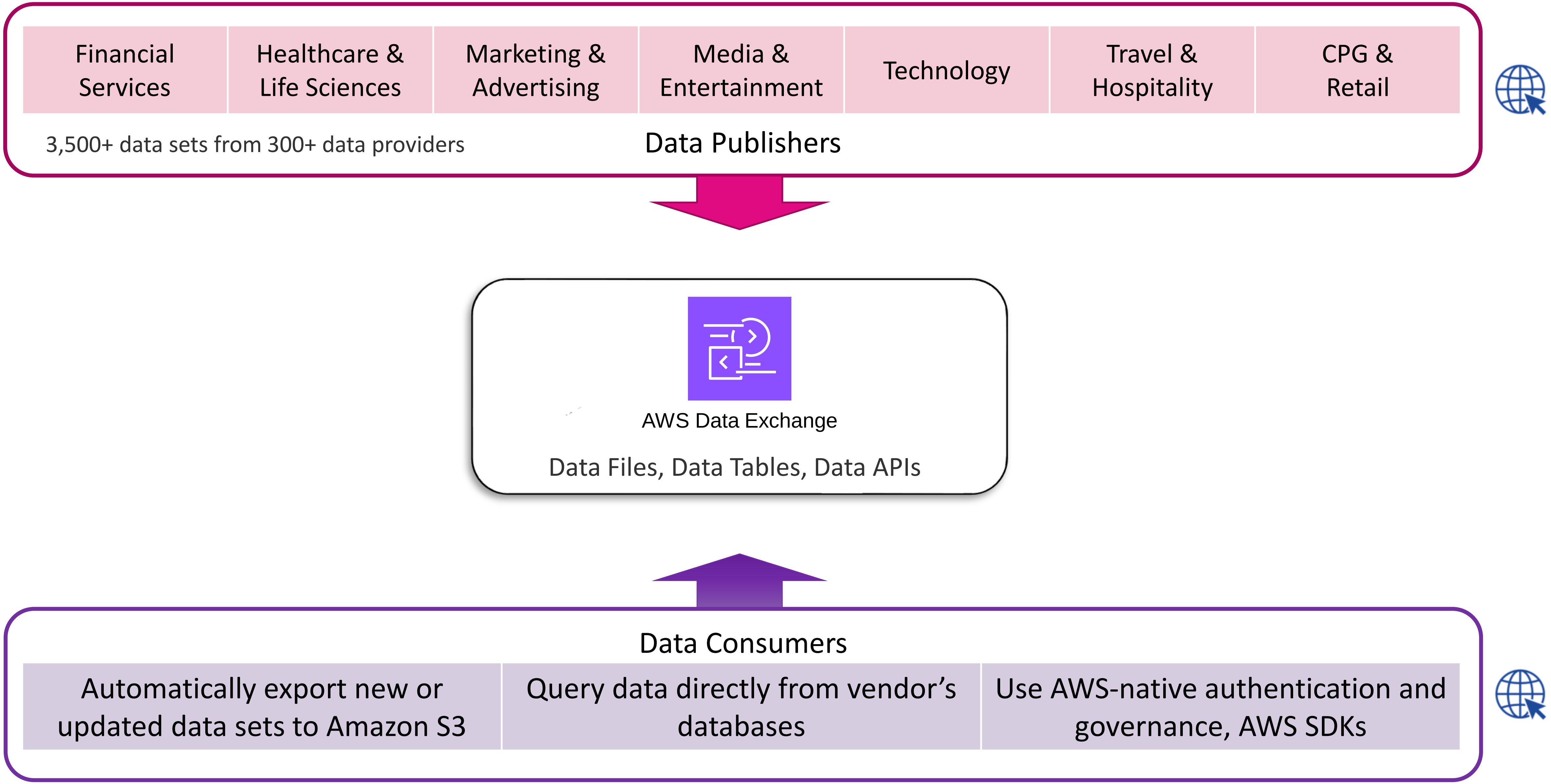
Mobile client

Mobile client

Mobile client

Mobile client

AWS Data Exchange



Financial Services

Healthcare & Life Sciences

Marketing & Advertising

Media & Entertainment

Technology

Travel & Hospitality

CPG & Retail

3,500+ data sets from 300+ data providers

Data Publishers

AWS Data Exchange

Data Files, Data Tables, Data APIs

Data Consumers

Automatically export new or updated data sets to Amazon S3

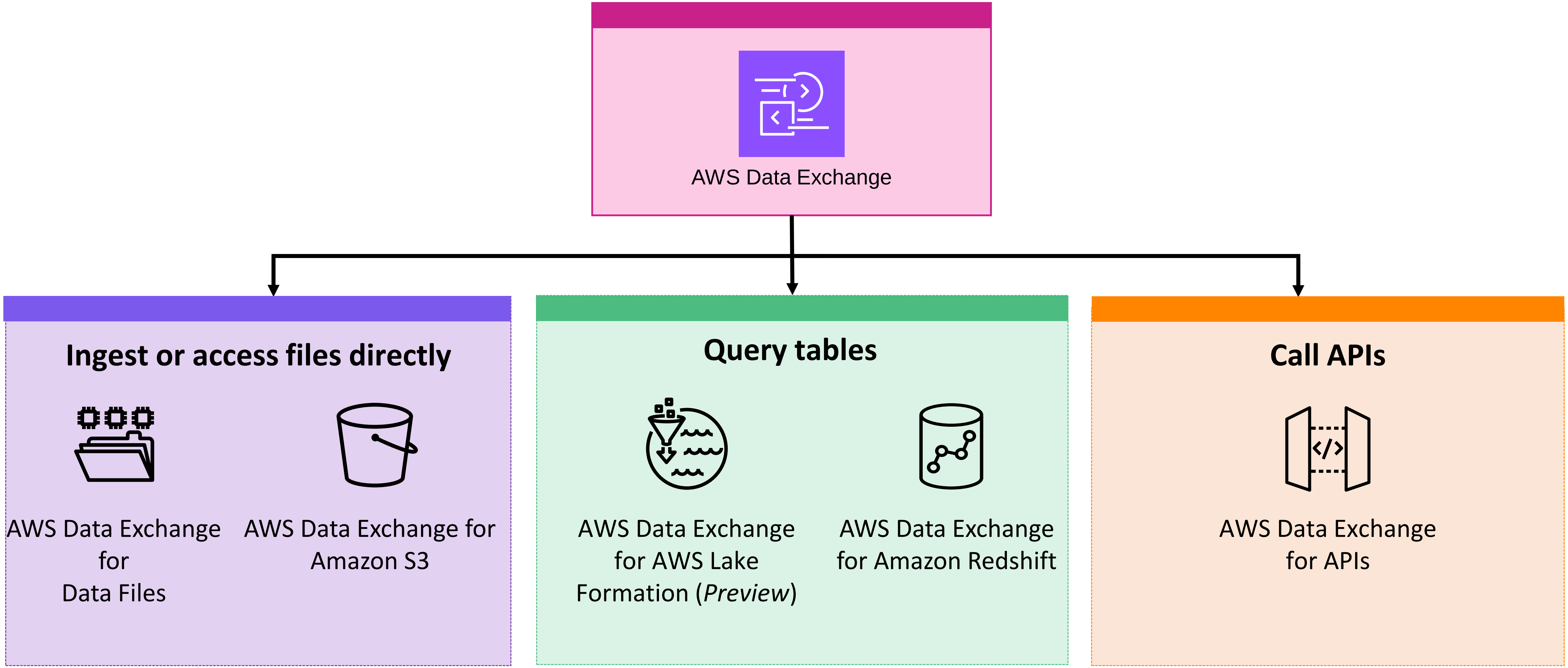
Query data directly from vendor's databases

Use AWS-native authentication and governance, AWS SDKs

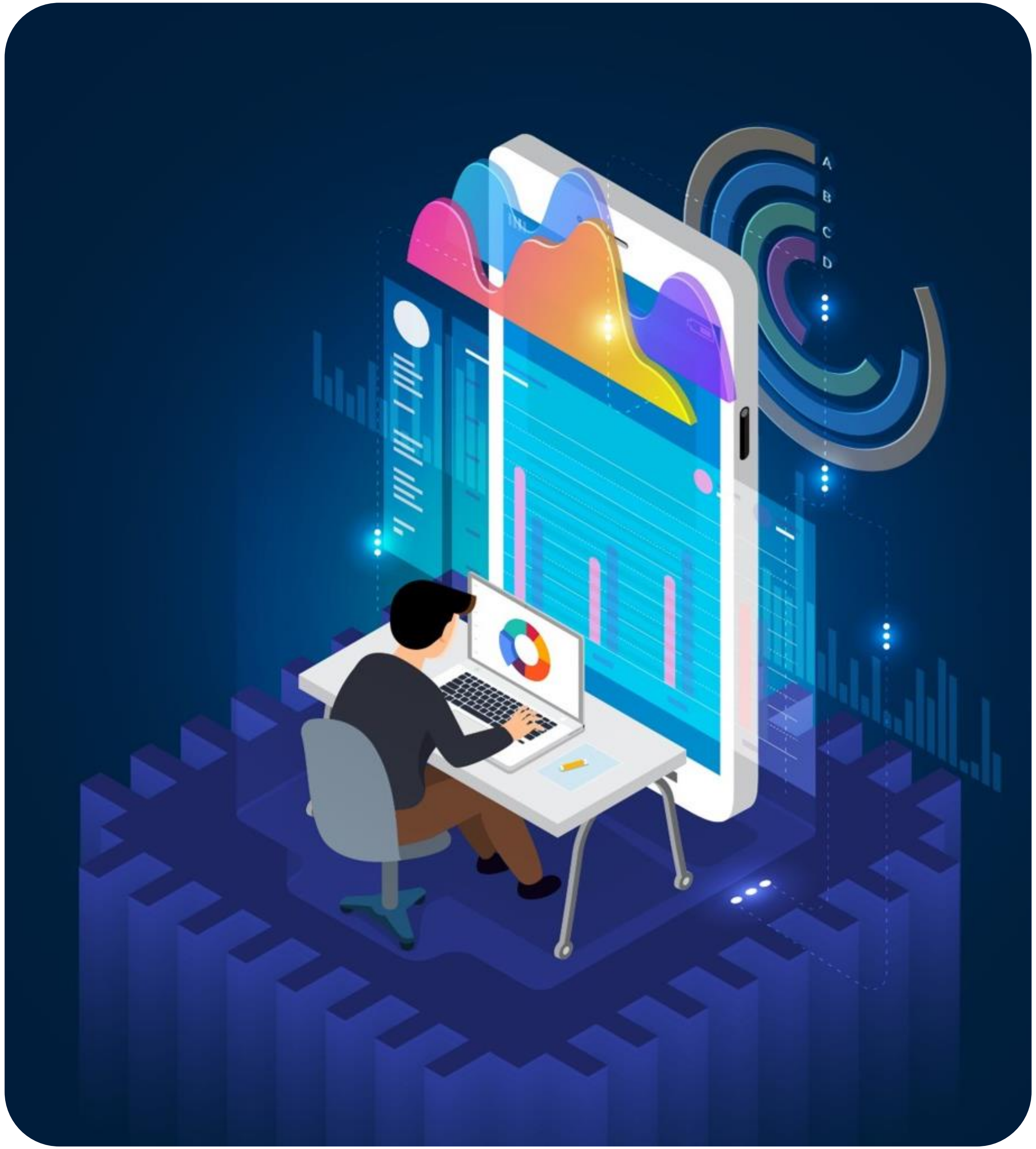
AWS Data Exchange

- AWS Data Exchange makes it easy to ingest third-party data into your data warehouses and data lakes
- It allows you to quickly and easily start using third-party data in your applications, analytics, and machine learning models.
- Data providers benefits
  - Package and publish data sets inside free or paid products for the price and terms that you control.
  - Data owners can curate data in S3, local files, Redshift datashares, APIs, or Lake Formation tables/databases.
  - Reach millions of AWS customers
- Data subscribers benefits
  - Quickly find diverse data in one place and analyze data as it’s published
  - Spend less time and effort to use data in production.

Ways to license third-party data faster and easier



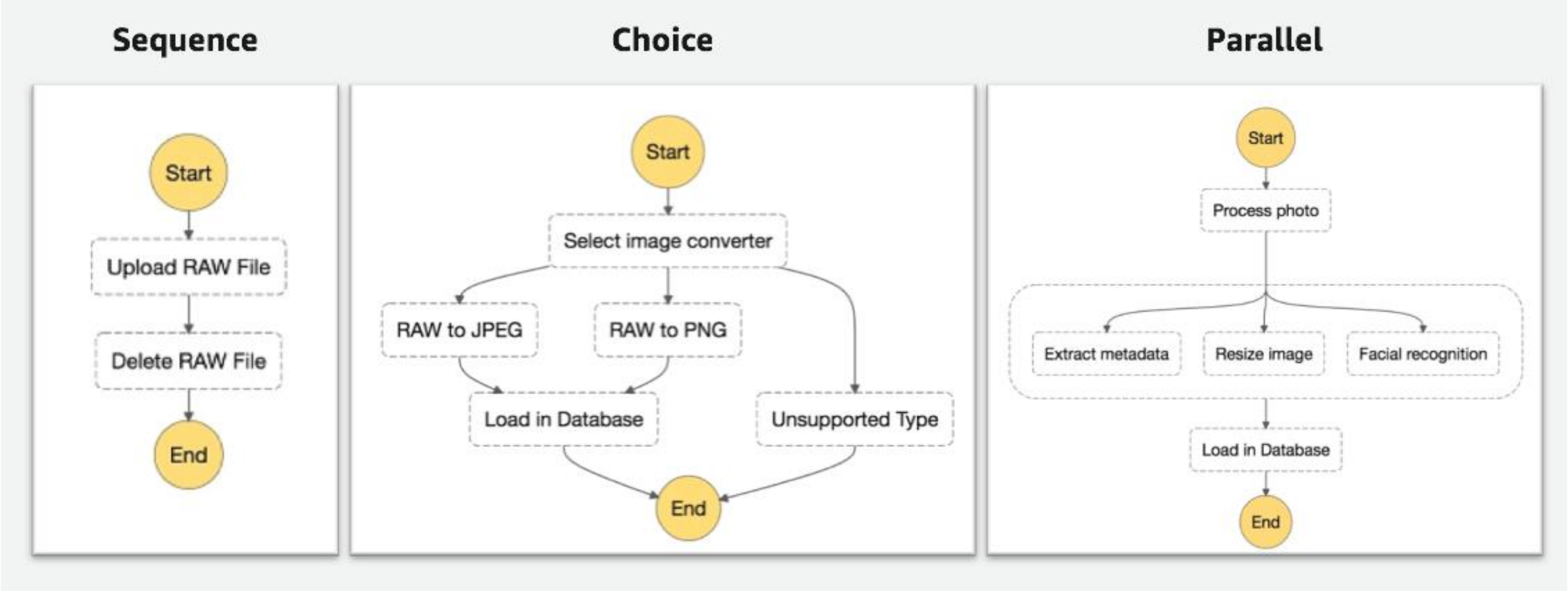
AWS Data Exchange
Ingest or access files directly
AWS Data Exchange for Data Files
AWS Data Exchange for Amazon S3
Query tables
AWS Data Exchange for AWS Lake Formation (*Preview*)
AWS Data Exchange for Amazon Redshift
Call APIs
AWS Data Exchange for APIs



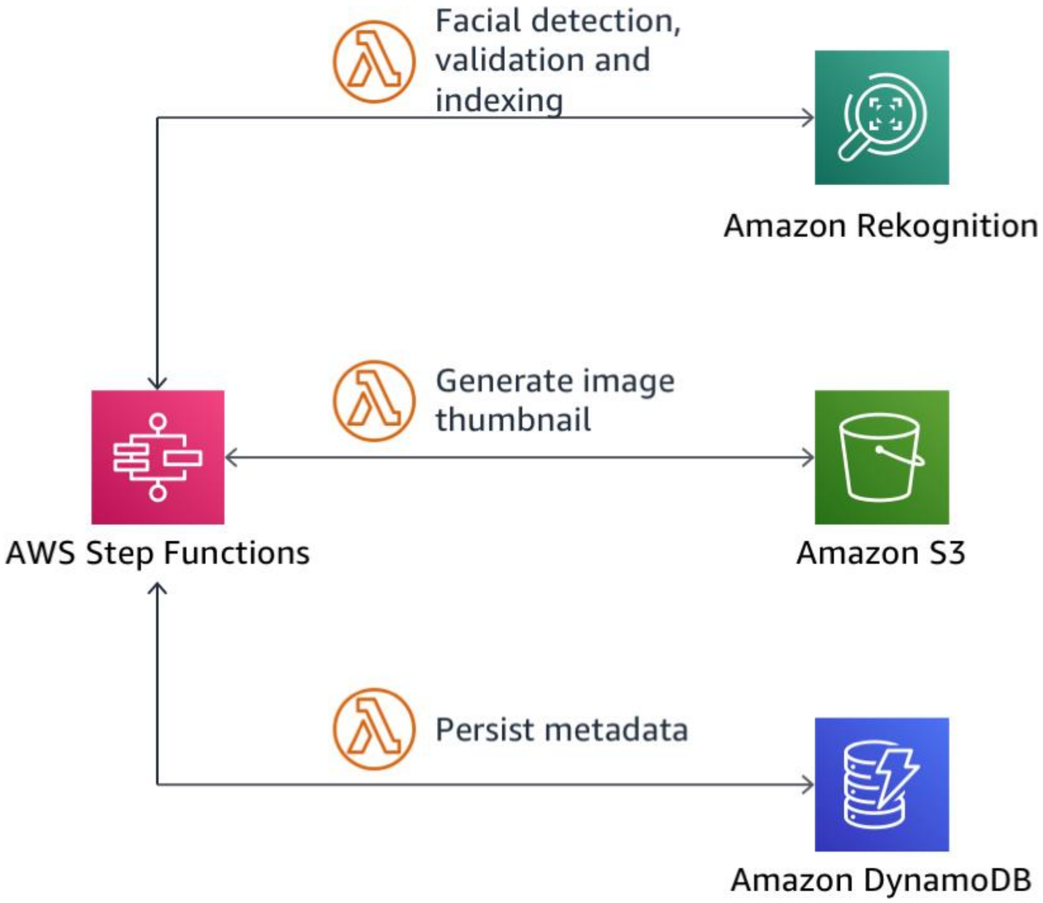
AWS Step Function

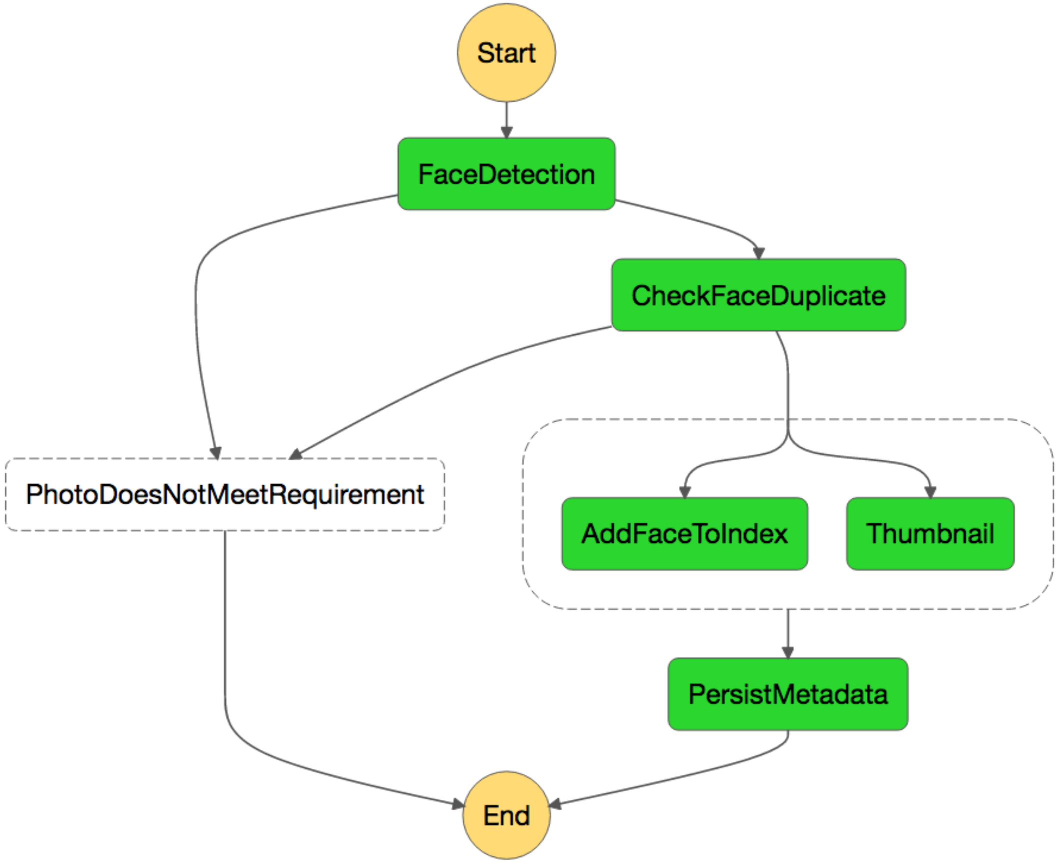
AWS Step Functions

- A workflow for Lambda Function
- Visual Interface to create and run workflows
- The output of one step acts as an input to the next.

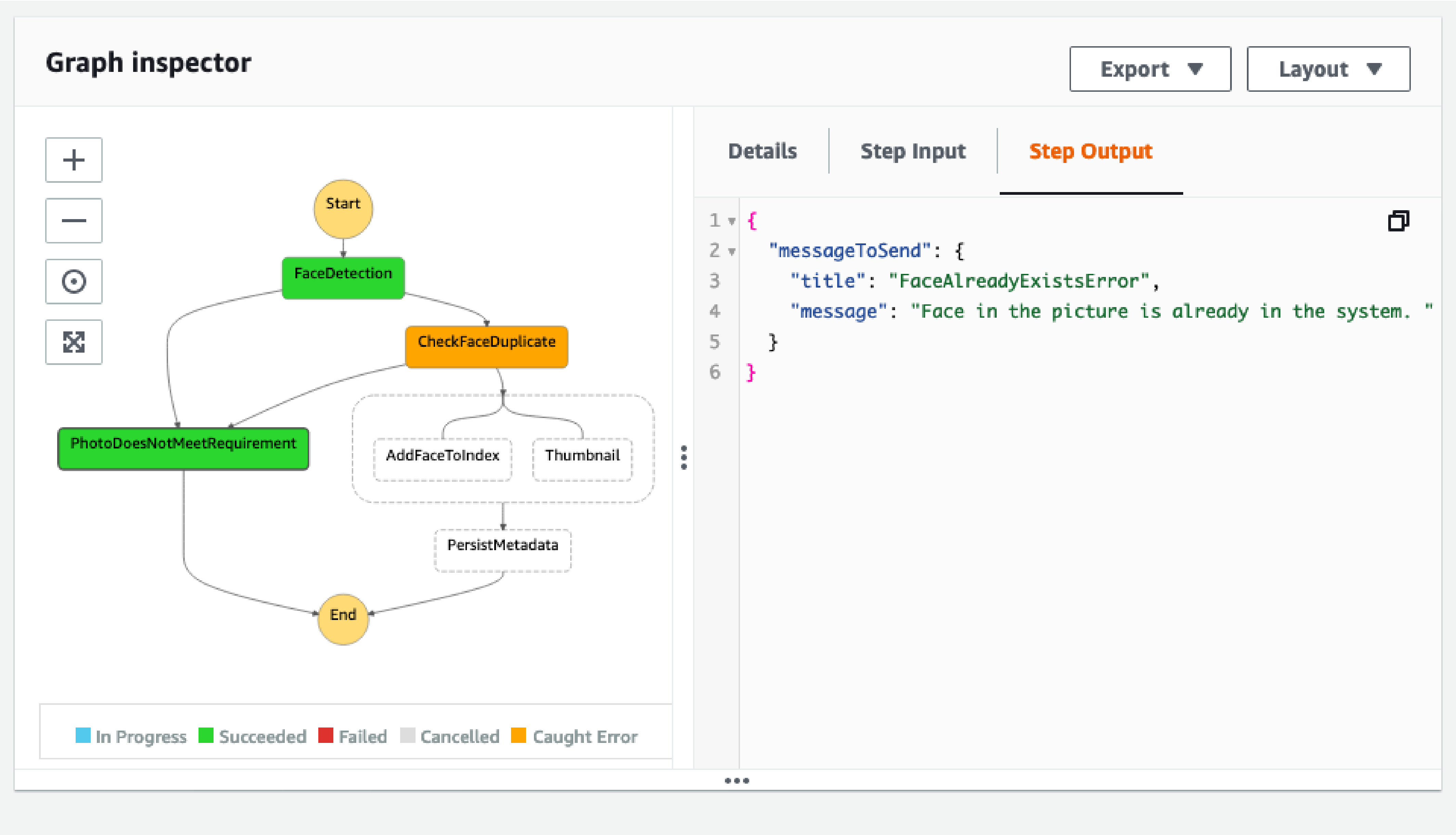


Serverless Image Processing Workshop



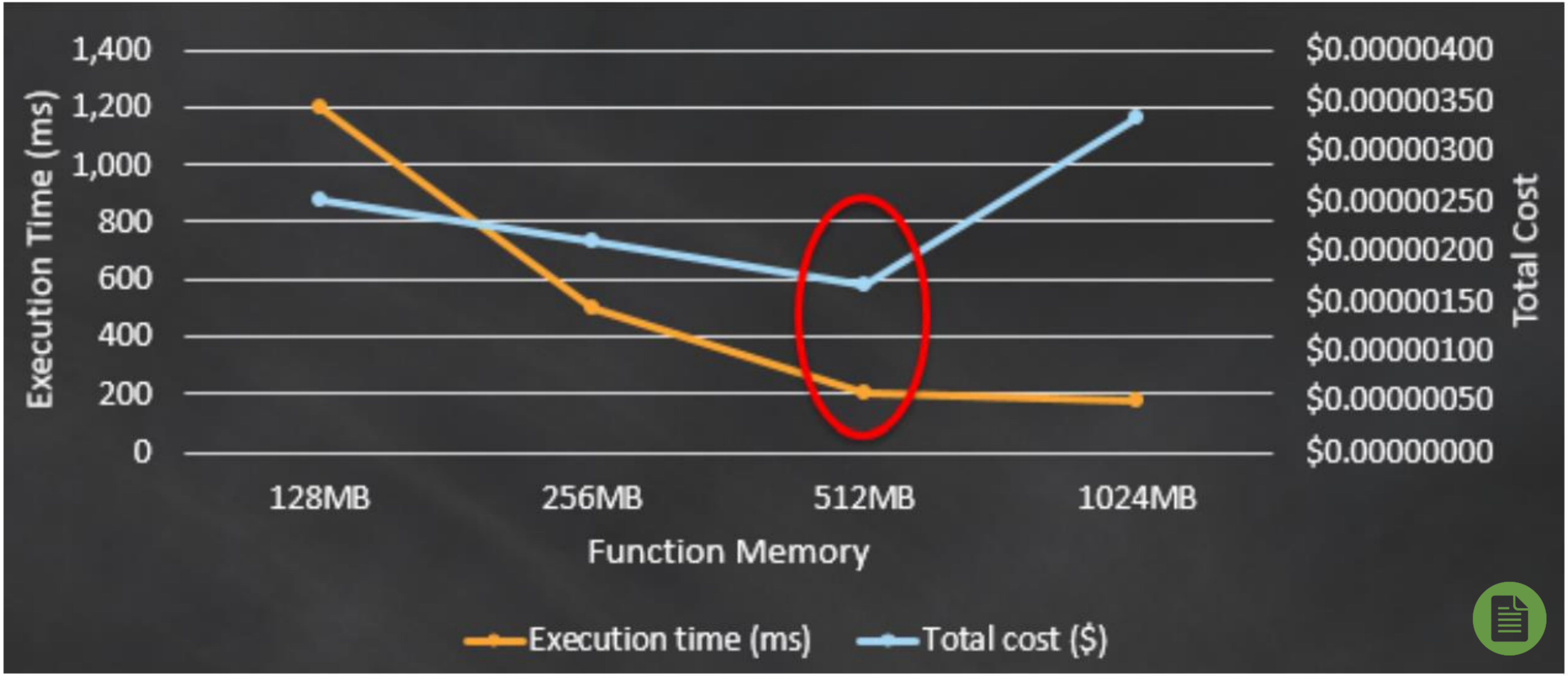


Serverless Image Processing Workshop

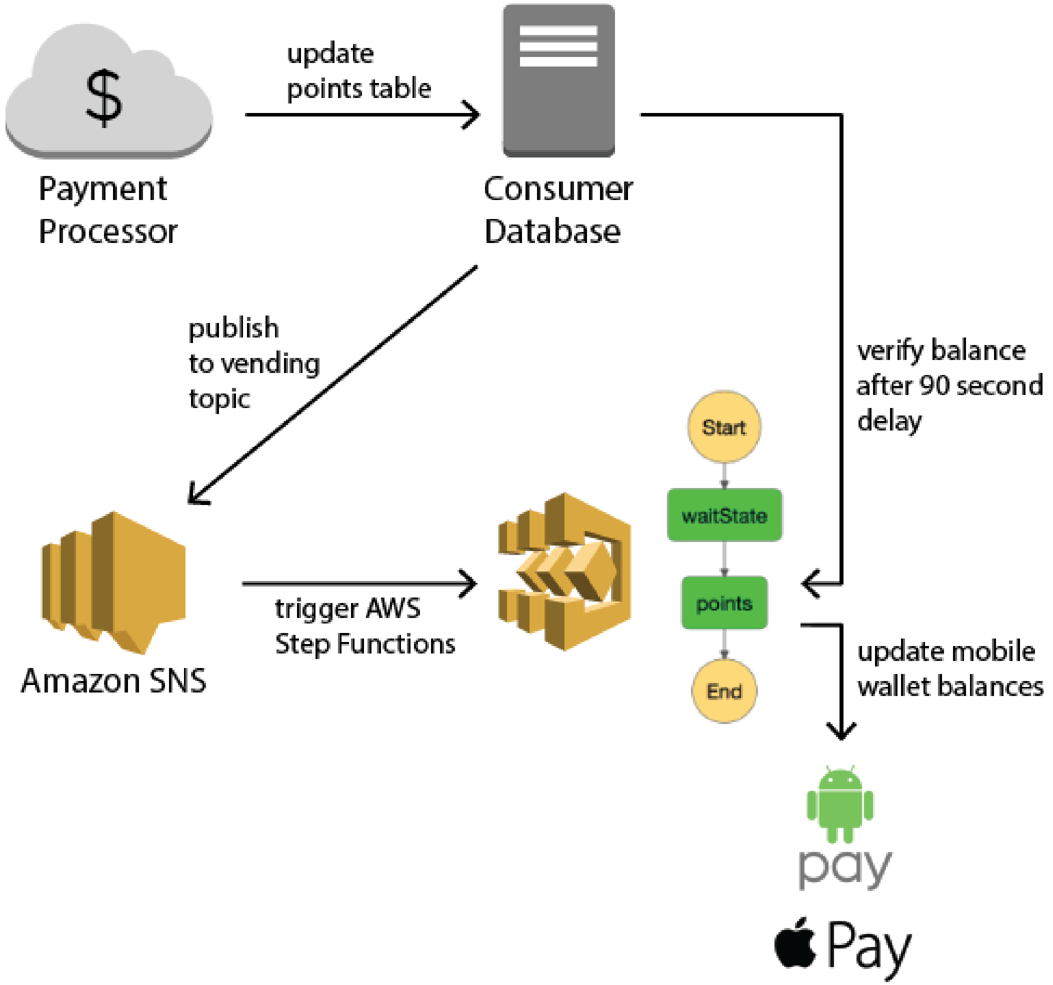


```
{
  "messageToSend": {
    "title": "FaceAlreadyExistsError",
    "message": "Face in the picture is already in the system. "
  }
}
```

AWS Lambda Power Tuning



Things Go Better With Step Functions



$
Payment Processor
update points table
Consumer Database
publish to vending topic
verify balance after 90 second delay
Start
waitState
points
End
Amazon SNS
trigger AWS Step Functions
update mobile wallet balances
pay
Pay