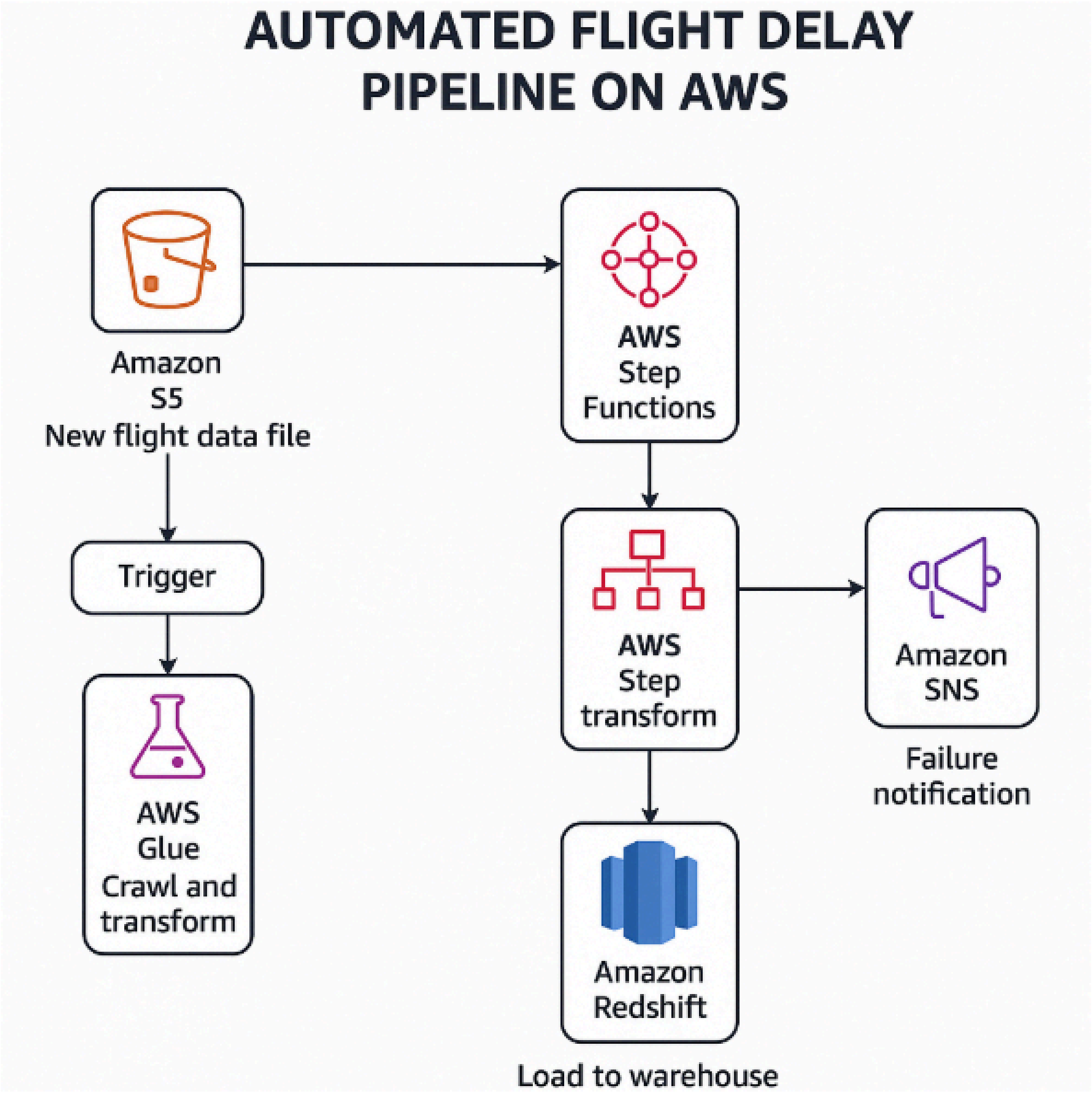
"Automated Flight Delay Processing Pipeline" project using **AWS Step Functions, EventBridge, SNS, S3, Glue, and Redshift**.



"Automated Flight Delay Processing Pipeline" project using **AWS Step Functions, EventBridge, SNS, S3, Glue, and Redshift**.

You can include this in your project README or technical documentation.

✈️ Automated Flight Delay Pipeline on AWS

📌 Project Overview

This project implements a **serverless, event-driven data pipeline** using AWS to automatically detect, transform, and analyze new flight data with a focus on **delays exceeding 60 minutes**. The solution ingests new data files added to an S3 bucket, processes and transforms them using Glue, and then loads the refined data into Amazon Redshift for analytics and reporting.

The pipeline also incorporates **Step Functions**, **EventBridge**, and **SNS** to automate the flow, monitor failures, and notify stakeholders in real time — **fully hands-free and scalable**.

🏗️ Architecture Summary

1. **Amazon S3** stores raw flight datasets.
2. **Amazon EventBridge** watches for new files added to the S3 bucket.
3. **AWS Step Functions** orchestrates the full data processing workflow.
4. **AWS Glue Crawler** scans and catalogs new data.
5. **AWS Glue Job** transforms the data to extract only the flights with delay times > 60 minutes.
6. **Amazon Redshift** stores the transformed data in a new analytics-ready table.
7. **Amazon SNS** sends failure notifications if any step fails, via email or other subscribers.

⚙️ Technologies Used

Service	Purpose
Amazon S3	Source bucket for raw flight data files
Amazon EventBridge	Detects new file uploads and triggers Step Functions
AWS Step Functions	Orchestrates data pipeline across Glue, Redshift, and SNS
AWS Glue Crawler	Automatically infers schema from new data
AWS Glue Job	Transforms the raw data (filters delay > 60 mins)
Amazon Redshift	Stores and visualizes the final results
Amazon SNS	Sends failure alerts (via email or SMS) on pipeline errors

🔁 Workflow Execution

1. **Trigger (EventBridge + S3):**

    - Whenever a new `.csv` or `.json` file is uploaded to the raw flight data S3 bucket, **EventBridge** detects this event.

    - It automatically **triggers the Step Function workflow** without any manual input.

2. **Crawling & Cataloging (Glue Crawler):**

    - A **Glue Crawler** scans the new dataset and updates the Data Catalog.

    - This ensures up-to-date schema metadata is available for transformation.

3. **Transformation (Glue Job):**

    - The **Glue Job** filters out flights with delay times greater than 60 minutes.

    - Data is cleaned, validated, and converted into a consistent format.

    - Output is stored in a separate S3 path or passed directly for loading into Redshift.

4. **Loading into Redshift:**

    - The filtered dataset is loaded into a **Redshift table called `flights_delayed_over_60`**.

    - This table supports further querying, dashboarding, or BI integrations.

5. **Error Notifications (SNS):**

    - If any stage fails (e.g., job error, Redshift load issue), the **Step Function triggers an SNS notification**.

    - A pre-configured email or SMS receives real-time alerts.

📊 Result

- A Redshift table named `flights_delayed_over_60` is populated with only those records where the delay exceeds 60 minutes.

- This enables quick downstream analytics to identify peak delay times, affected routes, or carrier patterns.

🚨 Error Handling & Monitoring

- **SNS Failures:** If Glue fails or Redshift load breaks, SNS sends a failure alert.

- **Step Function logs:** All executions (successful or failed) are logged in AWS Step Functions console.

- **CloudWatch Metrics:** Integrated monitoring available for each component, including Lambda logs (if used), job run time, crawler status, etc.

🔒 Security & Permissions

- IAM roles are tightly scoped for:
  - `StepFunctions-airline-data-pipeline-role` to run Glue, SNS, Redshift actions.
  - SNS topic is permissioned to only allow publishing from trusted roles.
  - S3 buckets have lifecycle rules and versioning enabled (optional).

🧪 Testing the Pipeline

To test the pipeline:

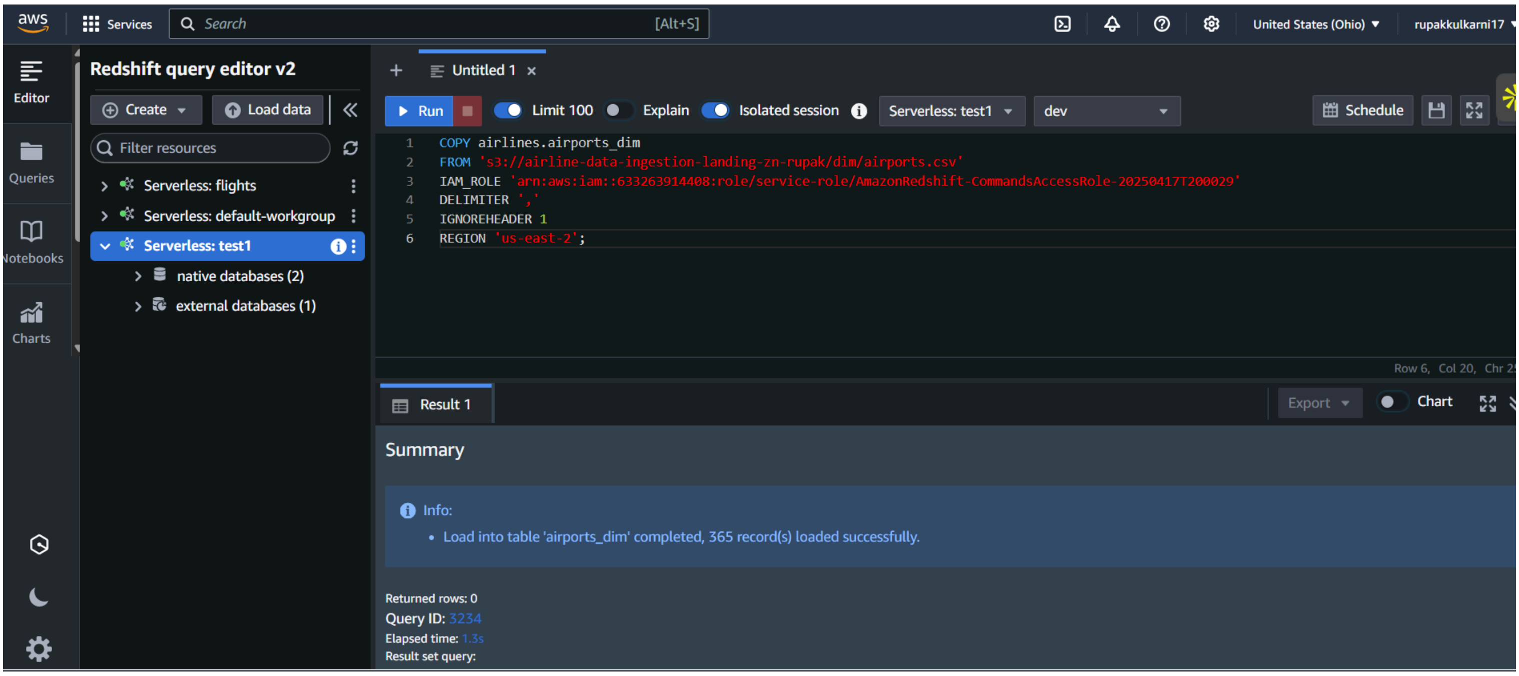
1. Upload a sample `.csv` or `.json` file with flight data to the S3 source bucket.
2. Watch the Step Function trigger and walk through each step.
3. Confirm the Redshift table is populated with correct delayed flights.
4. Introduce a failure (e.g., incorrect format) to verify SNS notification triggers.

🌟 Key Benefits

- **Fully automated**: No manual triggers required
- **Event-driven**: Responds in real-time to data arrival
- **Scalable**: Each component is serverless and cost-effective
- **Insight-ready**: Filtered data available in Redshift for instant BI consumption
- **Monitored**: End-to-end observability via logs, state tracing, and alerts

S3 bucket has been created. It has airports dim, date folder which will have updated data for everyday and a flights csv for loading data.

And going to create tables three tables for it one dimension and probably two facts When our trip need to create workgroups and name spaces as done in the previous projects and need to give access to the bucket



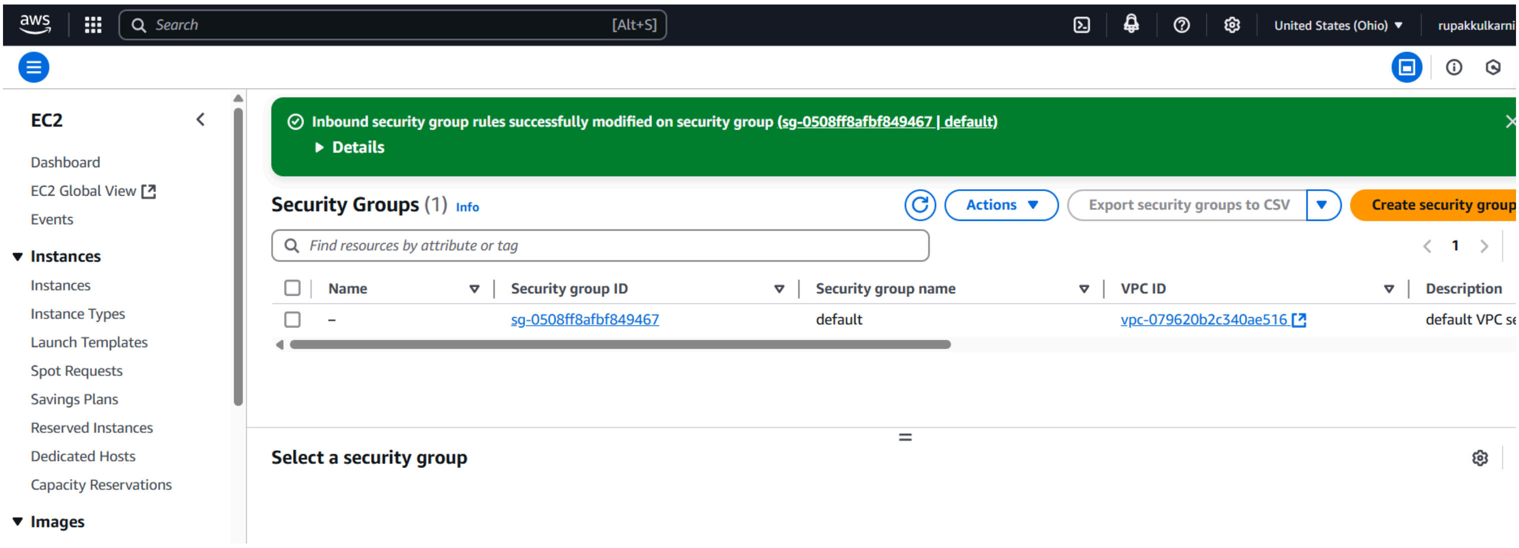
```
COPY airlines.airports_dim
FROM 's3://airline-data-ingestion-landing-zn-rupak/dim/airports.csv'
IAM_ROLE 'arn:aws:iam::633263914408:role/service-role/AmazonRedshift-CommandsAccessRole-20250417T200029'
DELIMITER ','
IGNOREHEADER 1
REGION 'us-east-2';
```

Summary

Info:
- Load into table 'airports_dim' completed, 365 record(s) loaded successfully.

Returned rows: 0
Query ID: 3234
Elapsed time: 1.3s
Result set query:

Opened inbound rules for this port 5439 used in redshift



Inbound security group rules successfully modified on security group (sg-0508ff8afbf849467 | default)

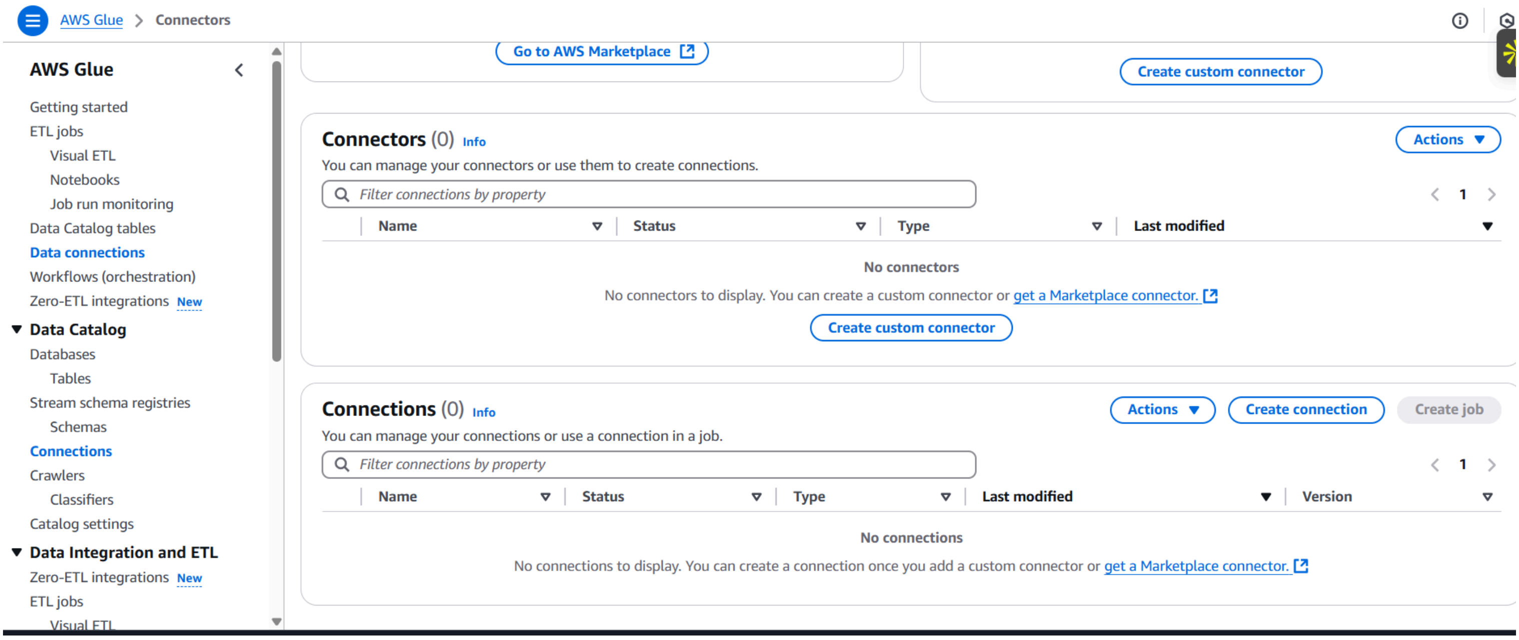
Details

Security Groups (1) Info

□	Name	Security group ID	Security group name	VPC ID	Description
□	–	sg-0508ff8afbf849467	default	vpc-079620b2c340ae516	default VPC se

Select a security group

Glue connections:



Connectors (0) Info

You can manage your connectors or use them to create connections.

Name	Status	Type	Last modified

No connectors

No connectors to display. You can create a custom connector or get a Marketplace connector.

Connections (0) Info

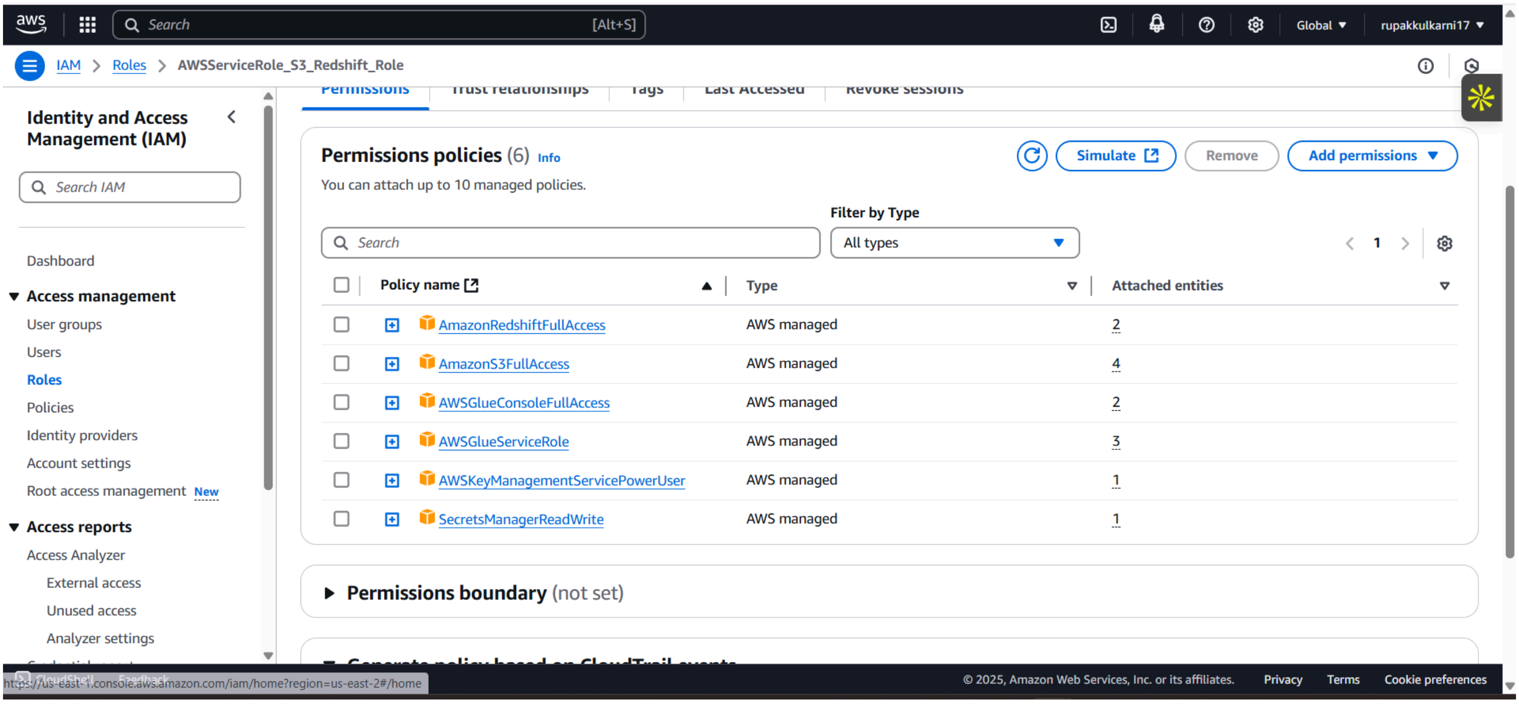
You can manage your connections or use a connection in a job.

Name	Status	Type	Last modified	Version

No connections

No connections to display. You can create a connection once you add a custom connector or get a Marketplace connector.

Connecting Glue with Redshift:



Permissions policies (6) Info

You can attach up to 10 managed policies.

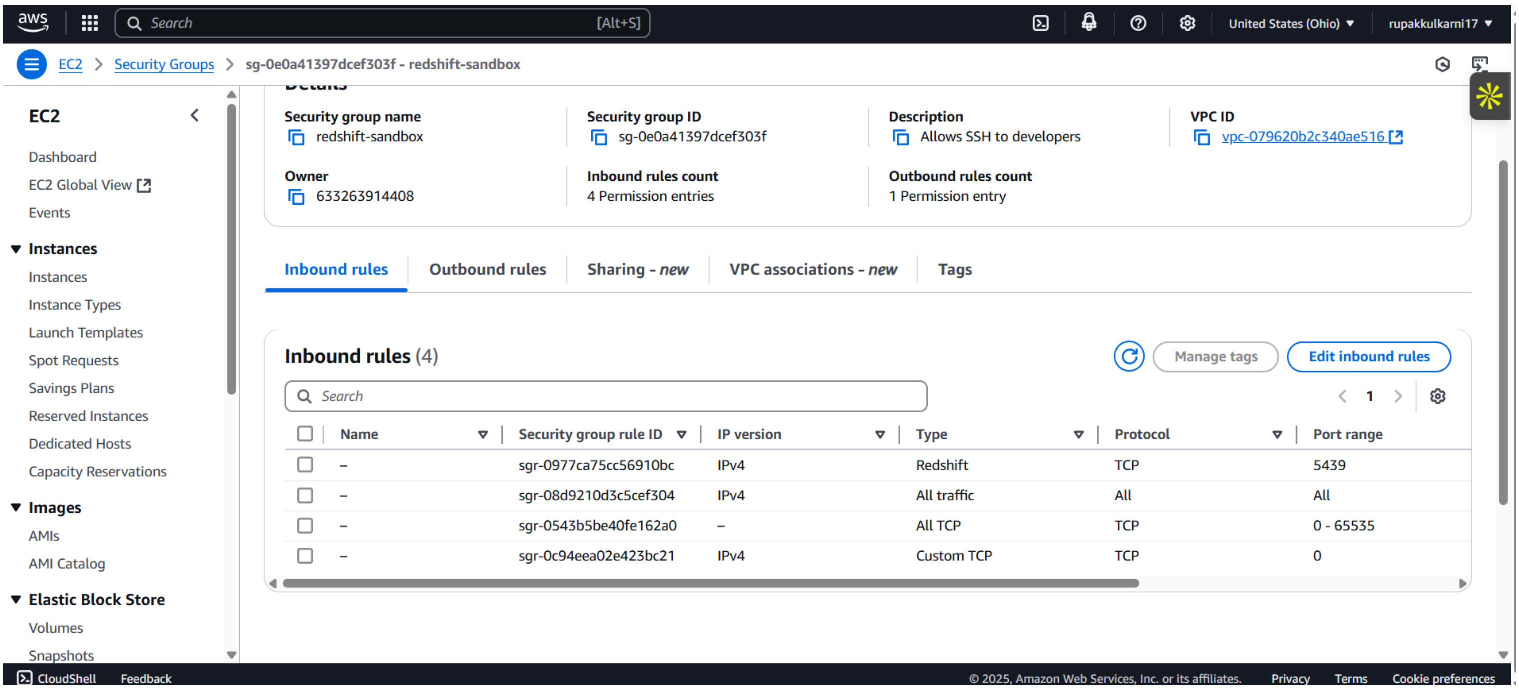
Filter by Type: All types

Policy name	Type	Attached entities
AmazonRedshiftFullAccess	AWS managed	2
AmazonS3FullAccess	AWS managed	4
AWSGlueConsoleFullAccess	AWS managed	2
AWSGlueServiceRole	AWS managed	3
AWSKeyManagementServicePowerUser	AWS managed	1
SecretsManagerReadWrite	AWS managed	1

Permissions boundary (not set)

Generate policy based on CloudTrail events

Inbound rules in redshift sandbox security group:



Details

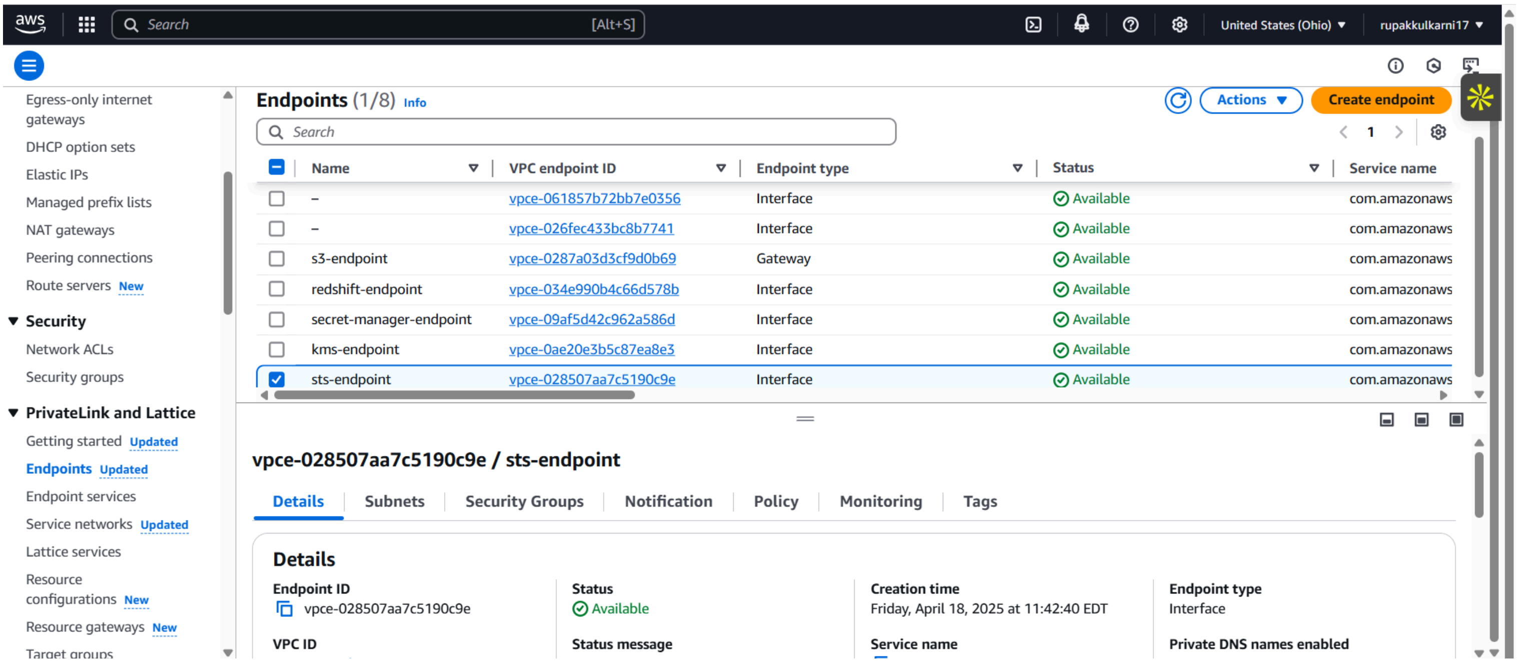
Security group name	Security group ID	Description	VPC ID
redshift-sandbox	sg-0e0a41397dcef303f	Allows SSH to developers	vpc-079620b2c340ae516

Owner	Inbound rules count	Outbound rules count
633263914408	4 Permission entries	1 Permission entry

Inbound rules (4)

Name	Security group rule ID	IP version	Type	Protocol	Port range
–	sgr-0977ca75cc56910bc	IPv4	Redshift	TCP	5439
–	sgr-08d9210d3c5cef304	IPv4	All traffic	All	All
–	sgr-0543b5be40fe162a0	–	All TCP	TCP	0 - 65535
–	sgr-0c94eea02e423bc21	IPv4	Custom TCP	TCP	0

Creation of endpoints:



Endpoints (1/8) Info

	Name	VPC endpoint ID	Endpoint type	Status	Service name
□	–	vpce-061857b72bb7e0356	Interface	Available	com.amazonaws
□	–	vpce-026fec433bc8b7741	Interface	Available	com.amazonaws
□	s3-endpoint	vpce-0287a03d3cf9d0b69	Gateway	Available	com.amazonaws
□	redshift-endpoint	vpce-034e990b4c66d578b	Interface	Available	com.amazonaws
□	secret-manager-endpoint	vpce-09af5d42c962a586d	Interface	Available	com.amazonaws
□	kms-endpoint	vpce-0ae20e3b5c87ea8e3	Interface	Available	com.amazonaws
☑	sts-endpoint	vpce-028507aa7c5190c9e	Interface	Available	com.amazonaws

vpce-028507aa7c5190c9e / sts-endpoint

Details

Endpoint ID: vpce-028507aa7c5190c9e

Status: Available

Creation time: Friday, April 18, 2025 at 11:42:40 EDT

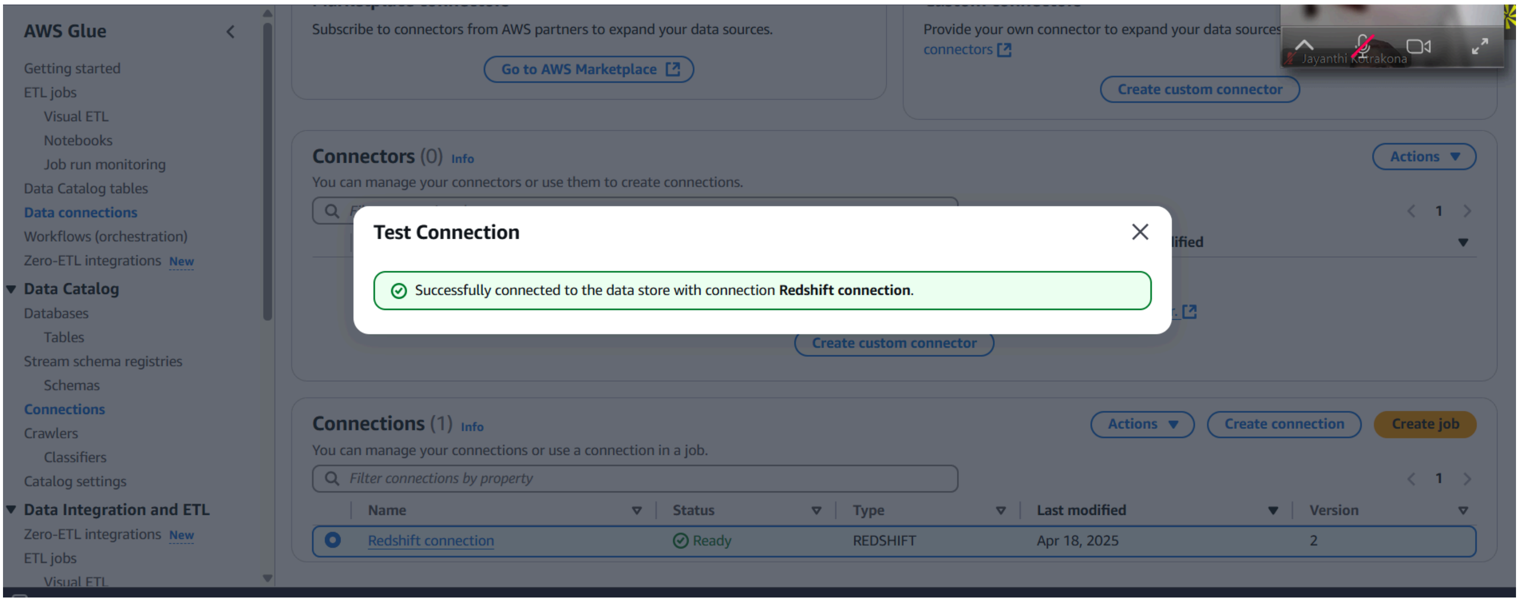
Endpoint type: Interface

VPC ID

Status message

Service name

Private DNS names enabled



Subscribe to connectors from AWS partners to expand your data sources.

Provide your own connector to expand your data sources

Connectors (0) Info

You can manage your connectors or use them to create connections.

Test Connection

Successfully connected to the data store with connection **Redshift connection**.

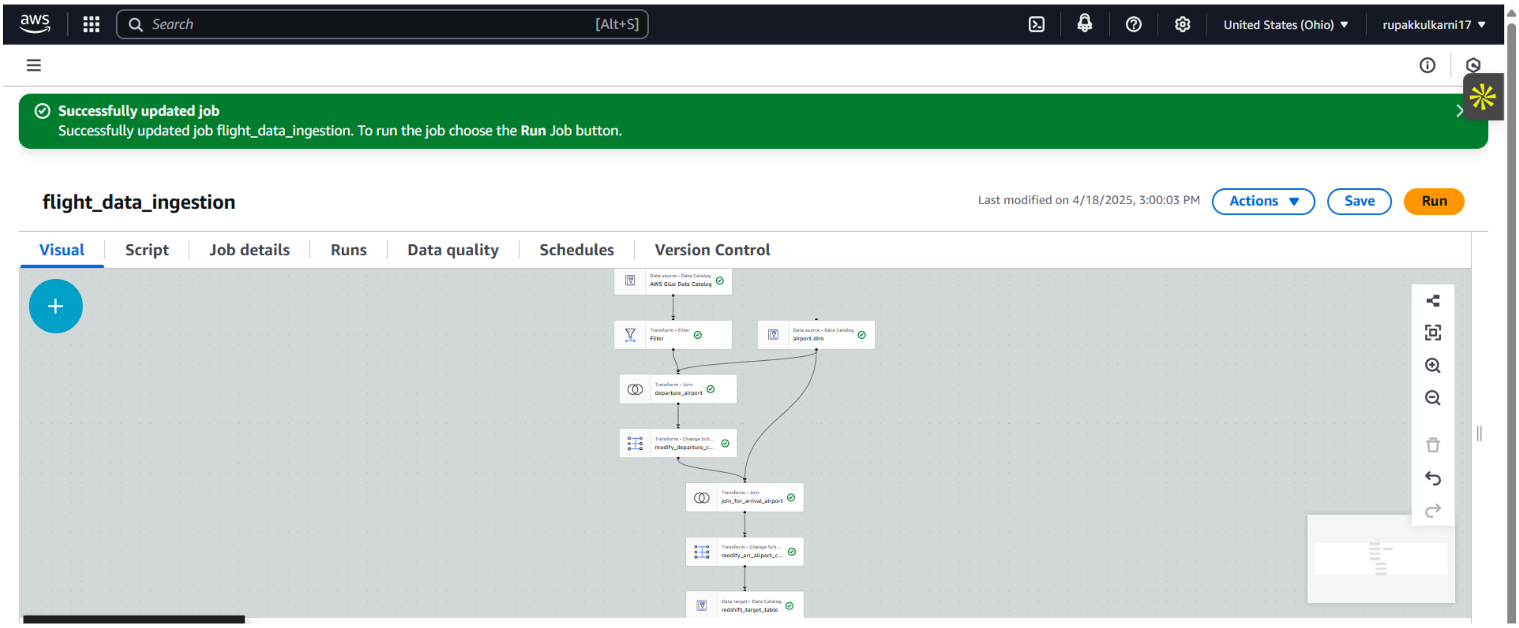
Connections (1) Info

You can manage your connections or use a connection in a job.

Filter connections by property

	Name	Status	Type	Last modified	Version
●	Redshift connection	Ready	REDSHIFT	Apr 18, 2025	2

Data Transformation to push data in daily folder



Successfully updated job

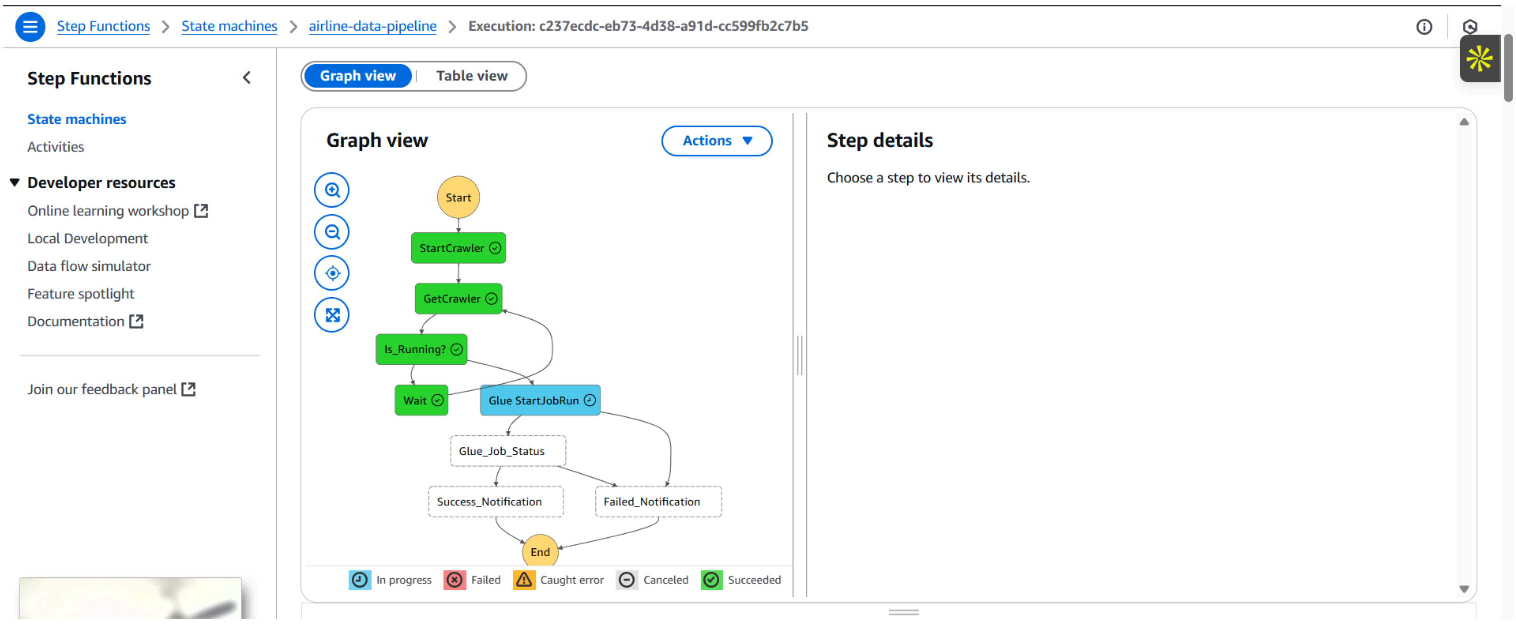
Successfully updated job flight_data_ingestion. To run the job choose the **Run** Job button.

flight_data_ingestion

Last modified on 4/18/2025, 3:00:03 PM

Turn on event bridge in the s3 bucket properties

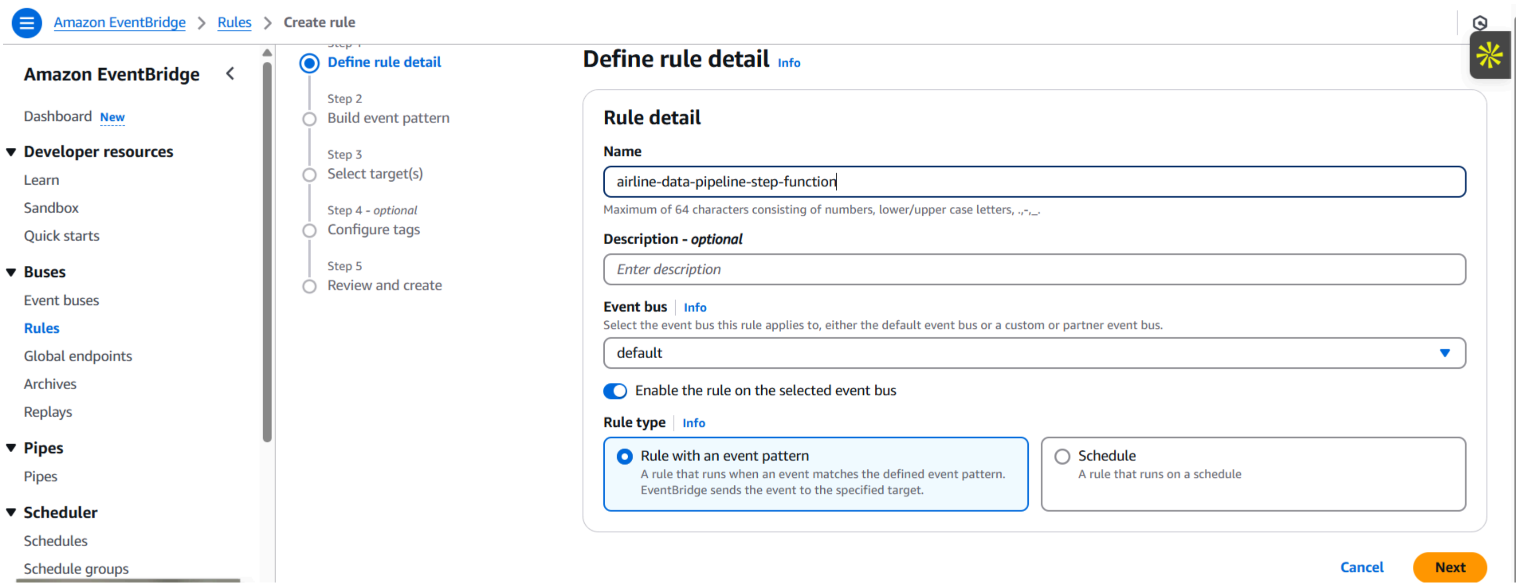
Prepare step function:



Step details

Choose a step to view its details.

Event bridge rule creation



Define rule detail Info

Rule detail

Name

airline-data-pipeline-step-function

Maximum of 64 characters consisting of numbers, lower/upper case letters, .,-,_.

Description - *optional*

Enter description

Event bus Info

Select the event bus this rule applies to, either the default event bus or a custom or partner event bus.

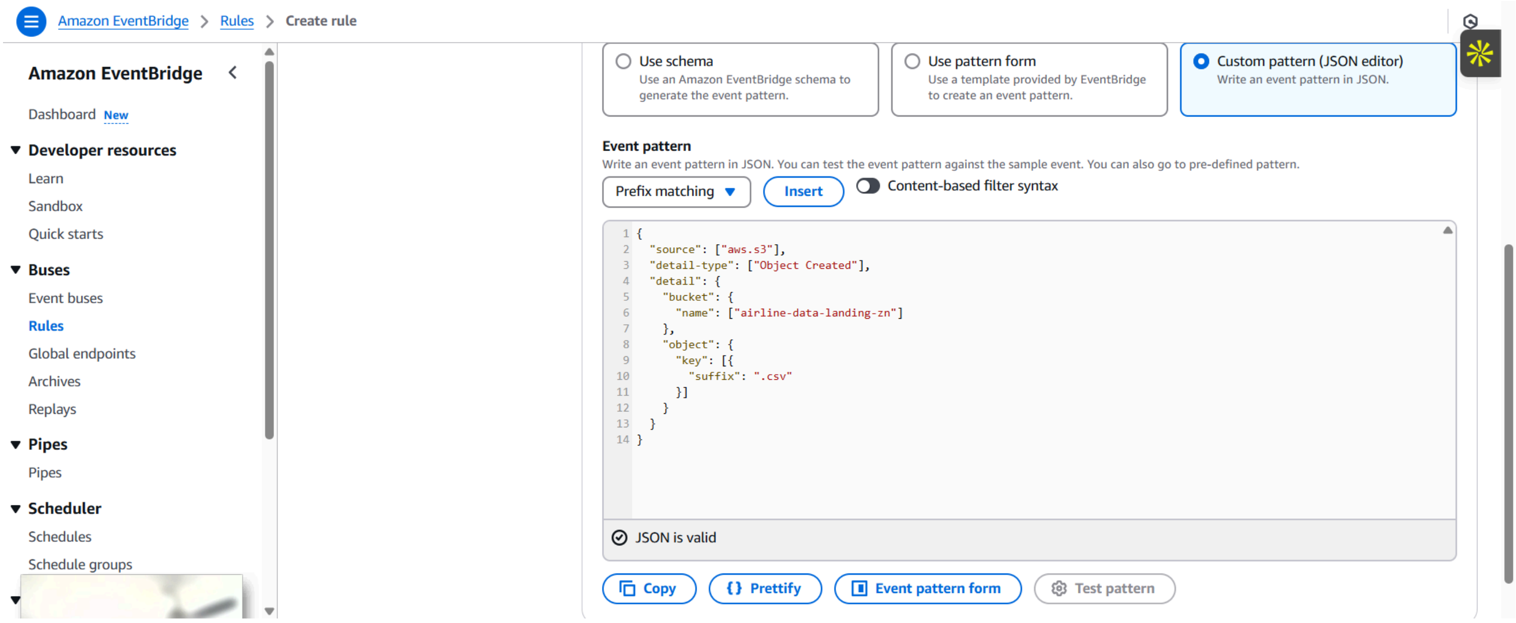
default

Enable the rule on the selected event bus

Rule type Info

● Rule with an event pattern
A rule that runs when an event matches the defined event pattern. EventBridge sends the event to the specified target.

○ Schedule
A rule that runs on a schedule



○ Use schema
Use an Amazon EventBridge schema to generate the event pattern.

○ Use pattern form
Use a template provided by EventBridge to create an event pattern.

● Custom pattern (JSON editor)
Write an event pattern in JSON.

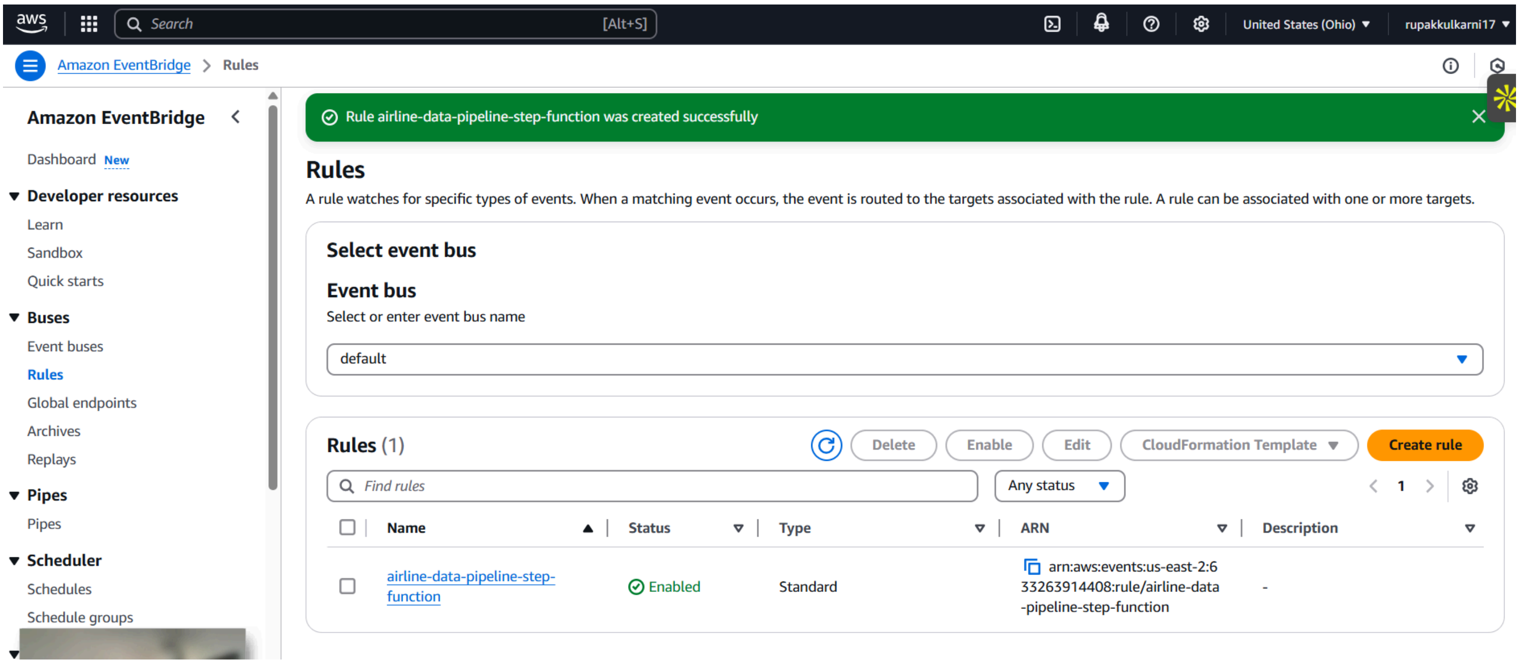
Event pattern

Write an event pattern in JSON. You can test the event pattern against the sample event. You can also go to pre-defined pattern.

Content-based filter syntax

```
{
  "source": ["aws.s3"],
  "detail-type": ["Object Created"],
  "detail": {
    "bucket": {
      "name": ["airline-data-landing-zn"]
    },
    "object": {
      "key": [{
        "suffix": ".csv"
      }]
    }
  }
}
```

JSON is valid



Rule airline-data-pipeline-step-function was created successfully

Rules

A rule watches for specific types of events. When a matching event occurs, the event is routed to the targets associated with the rule. A rule can be associated with one or more targets.

Select event bus

Event bus

Select or enter event bus name

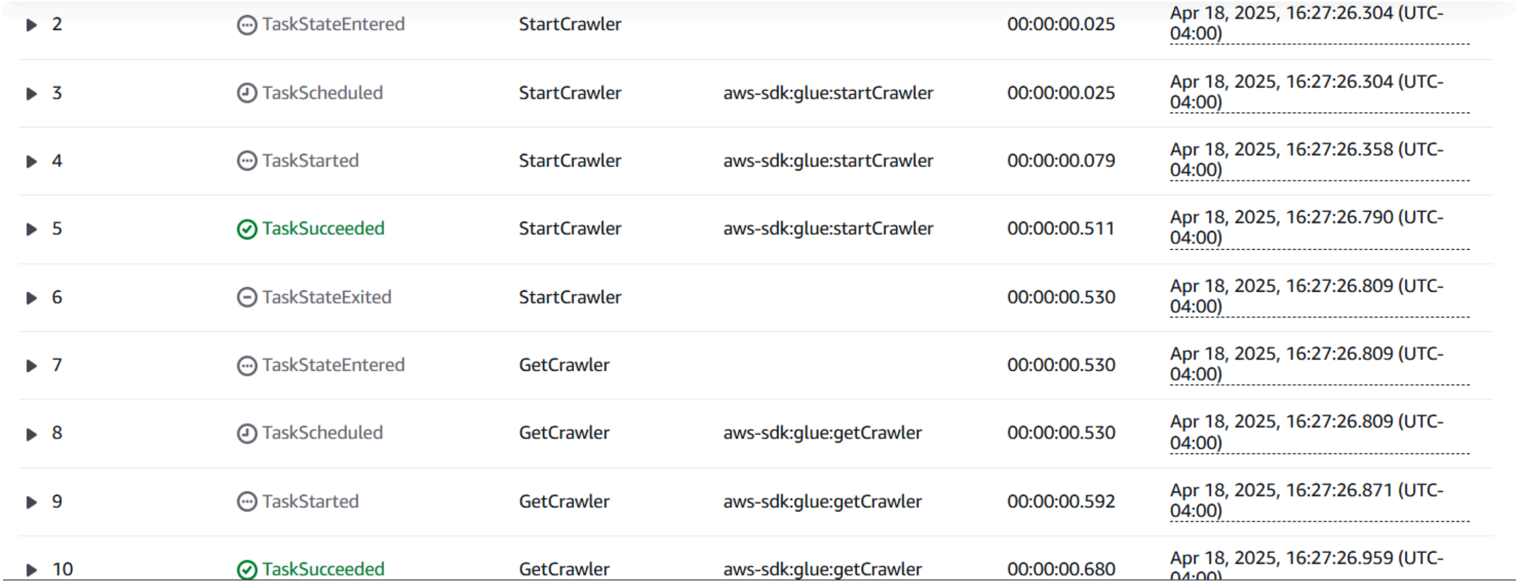
default

Rules (1)

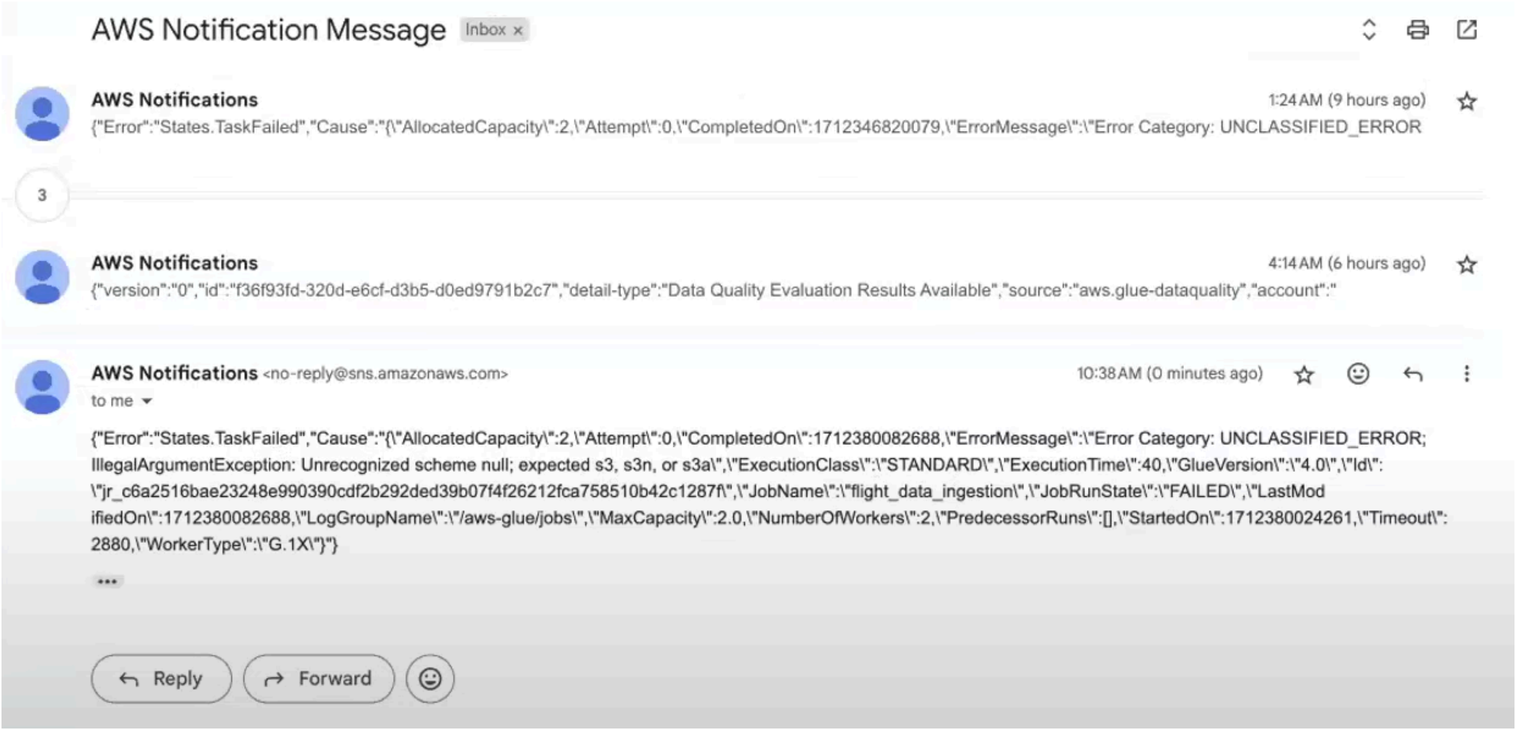
Name	Status	Type	ARN	Description
airline-data-pipeline-step-function	Enabled	Standard	arn:aws:events:us-east-2:633263914408:rule/airline-data-pipeline-step-function	-

Now delete flight.csv. And again add it.

– you will see state machine is started automatically due to event bridge.
-state machine starts the crawler —-- all state run



2	TaskStateEntered	StartCrawler		00:00:00.025	Apr 18, 2025, 16:27:26.304 (UTC-04:00)
3	TaskScheduled	StartCrawler	aws-sdk:glue:startCrawler	00:00:00.025	Apr 18, 2025, 16:27:26.304 (UTC-04:00)
4	TaskStarted	StartCrawler	aws-sdk:glue:startCrawler	00:00:00.079	Apr 18, 2025, 16:27:26.358 (UTC-04:00)
5	TaskSucceeded	StartCrawler	aws-sdk:glue:startCrawler	00:00:00.511	Apr 18, 2025, 16:27:26.790 (UTC-04:00)
6	TaskStateExited	StartCrawler		00:00:00.530	Apr 18, 2025, 16:27:26.809 (UTC-04:00)
7	TaskStateEntered	GetCrawler		00:00:00.530	Apr 18, 2025, 16:27:26.809 (UTC-04:00)
8	TaskScheduled	GetCrawler	aws-sdk:glue:getCrawler	00:00:00.530	Apr 18, 2025, 16:27:26.809 (UTC-04:00)
9	TaskStarted	GetCrawler	aws-sdk:glue:getCrawler	00:00:00.592	Apr 18, 2025, 16:27:26.871 (UTC-04:00)
10	TaskSucceeded	GetCrawler	aws-sdk:glue:getCrawler	00:00:00.680	Apr 18, 2025, 16:27:26.959 (UTC-04:00)



AWS Notification Message

AWS Notifications 1:24 AM (9 hours ago)

```
{"Error":"States.TaskFailed","Cause":"{\"AllocatedCapacity\":2,\"Attempt\":0,\"CompletedOn\":1712346820079,\"ErrorMessage\":\"Error Category: UNCLASSIFIED_ERROR
```

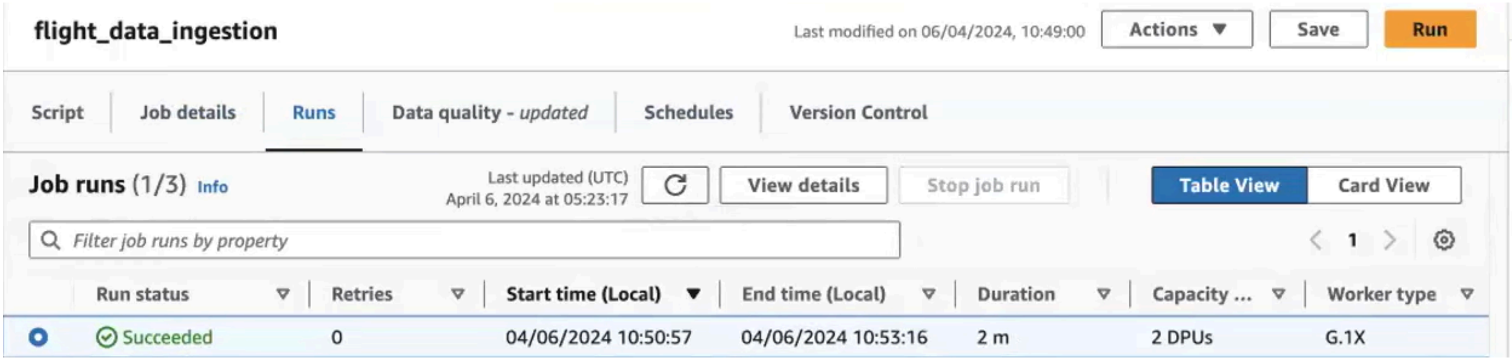
AWS Notifications 4:14 AM (6 hours ago)

```
{"version":"0","id":"f36f93fd-320d-e6cf-d3b5-d0ed9791b2c7","detail-type":"Data Quality Evaluation Results Available","source":"aws.glue-dataquality","account":"
```

AWS Notifications <no-reply@sns.amazonaws.com> 10:38 AM (0 minutes ago)

to me

```
{"Error":"States.TaskFailed","Cause":"{\"AllocatedCapacity\":2,\"Attempt\":0,\"CompletedOn\":1712380082688,\"ErrorMessage\":\"Error Category: UNCLASSIFIED_ERROR;
IllegalArgumentException: Unrecognized scheme null; expected s3, s3n, or s3a\",\"ExecutionClass\":\"STANDARD\",\"ExecutionTime\":40,\"GlueVersion\":\"4.0\",\"Id\":
\"jr_c6a2516bae23248e990390cdf2b292ded39b07f4f26212fca758510b42c1287f\",\"JobName\":\"flight_data_ingestion\",\"JobRunState\":\"FAILED\",\"LastMod
ifiedOn\":1712380082688,\"LogGroupName\":\"/aws-glue/jobs\",\"MaxCapacity\":2.0,\"NumberOfWorkers\":2,\"PredecessorRuns\":[],\"StartedOn\":1712380024261,\"Timeout\":
2880,\"WorkerType\":\"G.1X\"}"}
```



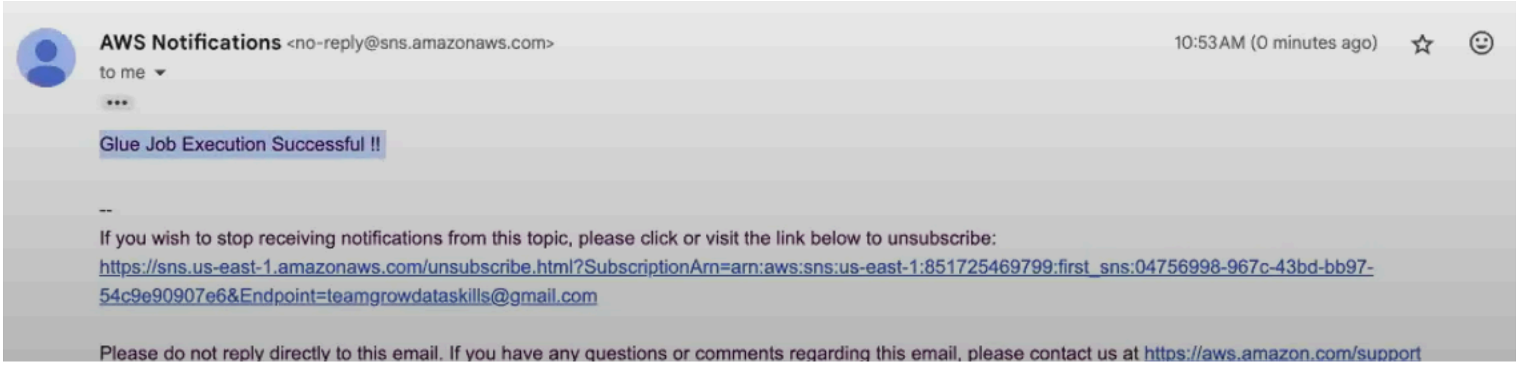
flight_data_ingestion

Last modified on 06/04/2024, 10:49:00

Job runs (1/3)

Last updated (UTC)
April 6, 2024 at 05:23:17

	Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity ...	Worker type
●	✓ Succeeded	0	04/06/2024 10:50:57	04/06/2024 10:53:16	2 m	2 DPUs	G.1X



AWS Notifications <no-reply@sns.amazonaws.com> 10:53 AM (0 minutes ago)

to me

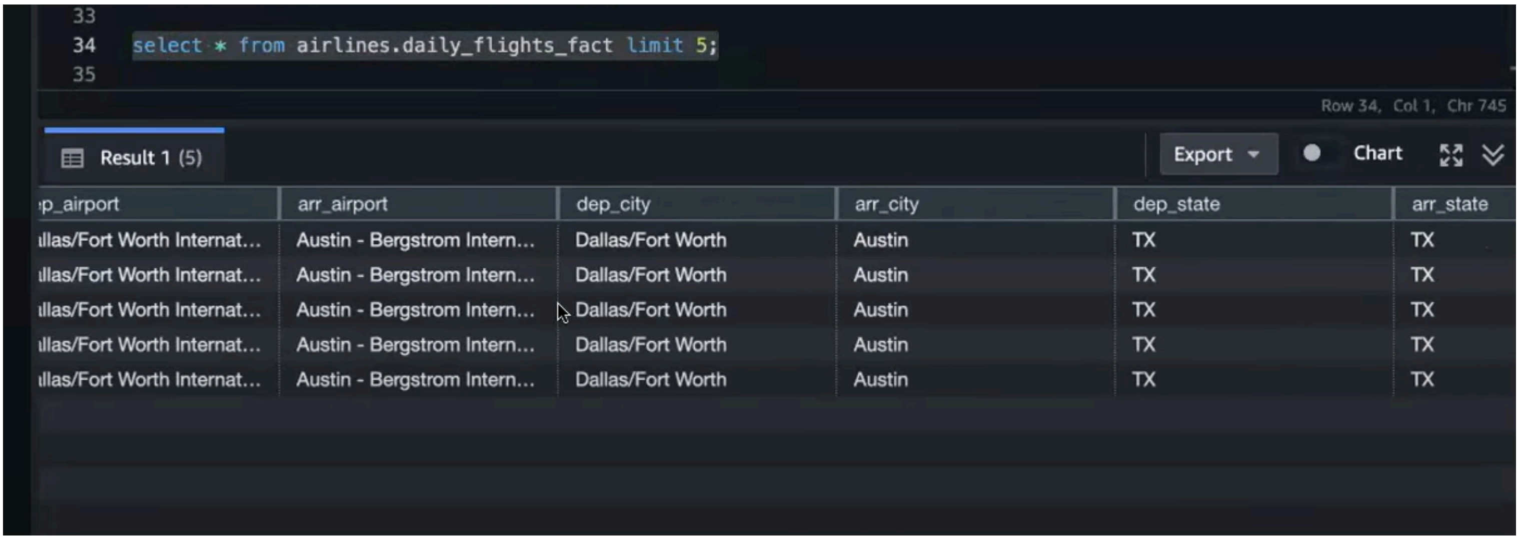
Glue Job Execution Successful !!

--

If you wish to stop receiving notifications from this topic, please click or visit the link below to unsubscribe:
https://sns.us-east-1.amazonaws.com/unsubscribe.html?SubscriptionArn=arn:aws:sns:us-east-1:851725469799:first_sns:04756998-967c-43bd-bb97-54c9e90907e6&Endpoint=teamgrowdataskills@gmail.com

Please do not reply directly to this email. If you have any questions or comments regarding this email, please contact us at https://aws.amazon.com/support

Departure delay greater than 60 minutes is seen in redshift cluster



```
select * from airlines.daily_flights_fact limit 5;
```

Result 1 (5)

p_airport	arr_airport	dep_city	arr_city	dep_state	arr_state
llas/Fort Worth Internat...	Austin - Bergstrom Intern...	Dallas/Fort Worth	Austin	TX	TX
llas/Fort Worth Internat...	Austin - Bergstrom Intern...	Dallas/Fort Worth	Austin	TX	TX
llas/Fort Worth Internat...	Austin - Bergstrom Intern...	Dallas/Fort Worth	Austin	TX	TX
llas/Fort Worth Internat...	Austin - Bergstrom Intern...	Dallas/Fort Worth	Austin	TX	TX
llas/Fort Worth Internat...	Austin - Bergstrom Intern...	Dallas/Fort Worth	Austin	TX	TX