

→ Joining with Secondary Tables

Create an Accelerometer Trusted Zone

If we have the sensitive data in the **accelerometer landing zone**, we can write a glue job that filters the data and moves compliant records into an **accelerometer trusted zone** for later analysis. Let's go to Glue Studio, and create a new Glue Job.

Search for Glue Studio in the AWS Console search bar




Q glue studio

X

Search results for 'glue studio'

Services


See all 5 results ▶



AWS Glue

☆


AWS Glue is a fully managed ETL (extract, transform, and load) service



Nimble Studio

☆


Accelerate building a cloud-based content creation studio



AWS Glue DataBrew

☆

Visual data preparation tool to clean and normalize data for analytics and machine learning




AWS Lake Formation

☆

AWS Lake Formation makes it easy to set up a secure data lake

Features



AWS Glue Studio

AWS Glue feature

Services (5)

Features (3)

Blogs (776)

Documentation (756)

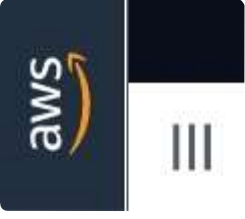
Knowledge Articles (30)

Tutorials (6)

Events (22)

Marketplace (5)

Click the three bars (hamburger menu) on the upper left corner of the AWS Console



Click the Hamburger menu

Click **Jobs**



Jobs menu

Accept the Visual with a source and target, then click Create

Create job [Info](#)

Create

☒ Visual with a source and target

Start with a source, ApplyMapping transform, and target.

☐ Python Shell script editor

Write or upload your own Python shell script.

Source



Amazon S3
JSON, CSV, or Parquet files stored in S3.



Target



Amazon S3
S3 bucket by specifying a bucket path as the data target.



☐ Visual with a blank canvas

Author using an interactive visual interface.

☐ Jupyter Notebook

Write your own code in a Jupyter Notebook for interactive development.

☐ Spark script editor

Write or upload your own Spark code.

You should see the **default** visual data flow



Source ▾



Transform ▾



Target ▾



Undo



Redo



Remove






Data source - S3 bucket
S3 bucket





Transform - ApplyMapping
ApplyMapping





Data target - S3 bucket

Configure the Job

Define Names for:

- Job (Accelerometer Landing to Trusted)
- Accelerometer Landing Node
- Join Customer Node
- Accelerometer Trusted Node

Define:

- IAM Role
- Job name
- Data Source Data Catalog table (accelerometer_landing)

The Visual Graph should look similar to this (**we're not finished!**):

- Data source type is **Data Catalog** (the accelerometer_landing table we created earlier)
- Transform type defaulted to **ApplyMapping** but the actual transformation will be a **Join**
- Data target defaulted to **S3** (we will change this to Data Catalog later)

Accelerometer Landing to Trusted

Visual

Script

Job details

Runs

Schedules

Source


Transform


Target


Undo

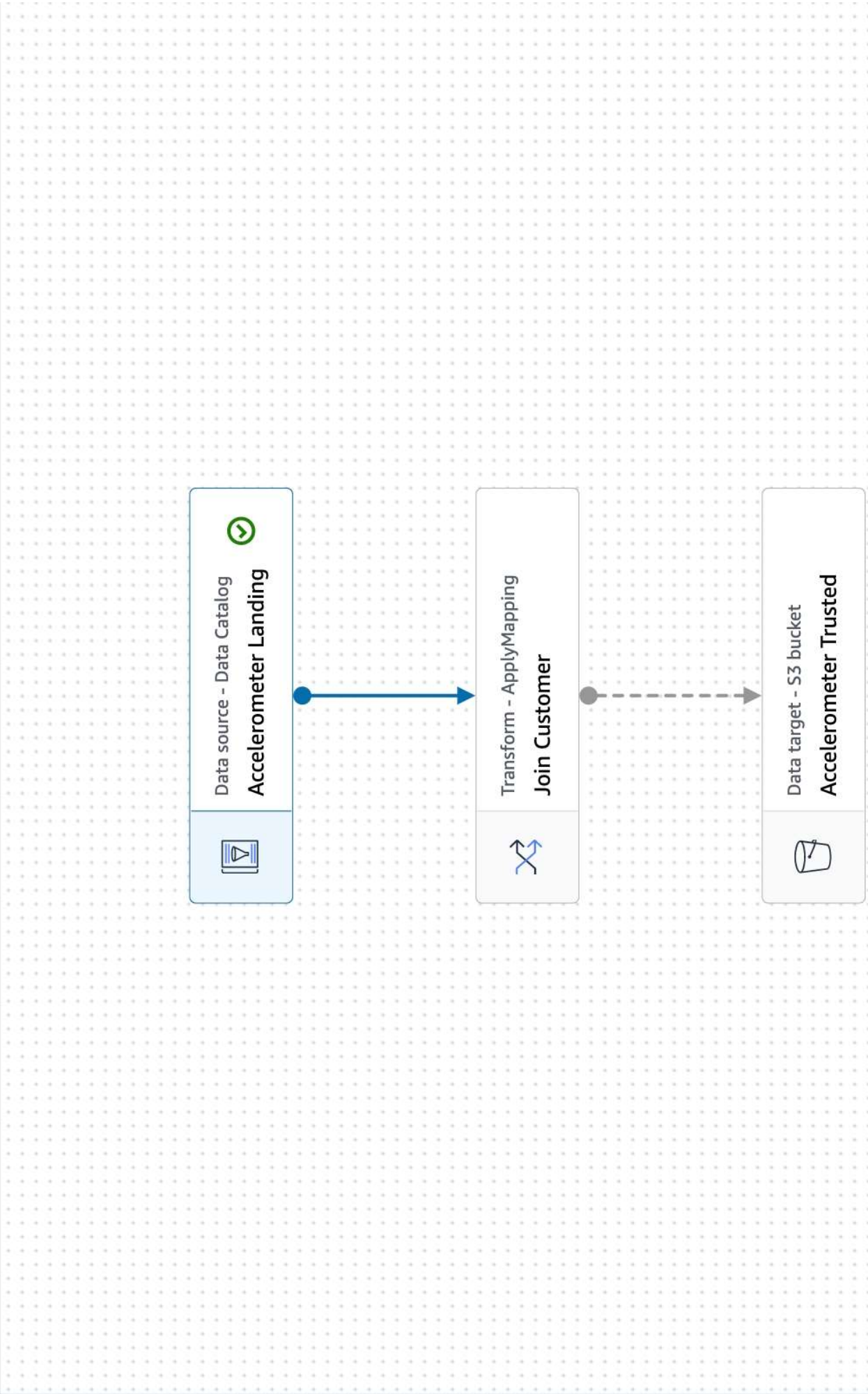
Redo

Remove









We need the **Customer Trusted Zone S3 datasource**. Click the Source dropdown, click Amazon S3:



- Configure the Amazon S3 datasource to point to the **Customer Trusted Zone S3** location
- Name the Node (**Customer Trusted Zone**)
- Click the **Infer schema** button

 Infer schema

Click the **Output schema** tab and you should see the inferred schema:

Node properties	Data source properties - S3	Output schema	Data preview	
Schema Info				<div>Edit</div>
Key		Data type	Partition	
serialNumber		string	-	
shareWithPublicAsOfDate		long	-	
birthDay		string	-	
registrationDate		long	-	
shareWithResearchAsOfDate		long	-	
customerName		string	-	
email		string	-	
lastUpdateDate		long	-	
phone		string	-	
shareWithFriendsAsOfDate		long	-	

As promised, now we change the Transform from **ApplyMapping** to **Join**

Node properties

Transform

Output schema

Data preview

Name

Join Customer

Node type

Choose which type of node to add to the job.

Apply Mapping

Map fields to new names and types of your choice.

Q |

Apply Mapping

Map fields to new names and types of your choice.

Select Fields

Choose which fields you want from your data.

Drop Fields

Remove selected fields from your data.

Drop Null Fields

Remove empty/nulls columns from your data.

Drop Duplicates

Drop duplicate rows in a data set.

Rename Field

Rename a single data field from your data set.

Spigot

Write sample data from a DynamicFrame.

Join

Join two sources into one output using a column header.

Check out our newly created join!



Node properties

Transform

Output schema

Data preview

Join type

Select the type of join to perform.



Inner join

Select all rows from both datasets that meet the join condition.



Join conditions

Select a field from each parent node for the join condition.



Insufficient source nodes

The Join transform requires two parent source nodes with selected tables.

Connect the **Customer Trusted Zone** node



Name

Join Customer

Node type

Choose which type of node to add to the job.

Join

Join two sources into one output using a column header.

Node parents

Choose which nodes will provide inputs for this one.

Select parents



Data sources

☒ Accelerometer Landing
Catalog - DataSource

☐ Customer Trusted Zone
S3 - DataSource

Transforms

Unclassified nodes

Click Add Condition

Join conditions

Select a field from each parent node for the join condition.

Add condition

All fields must have options selected

Add Condition

Choose the join fields you identified earlier to join accelerometer and customer

Node properties

Transform 1

Output schema

Data preview

Join type

Select the type of join to perform.

Inner join

Select all rows from both datasets that meet the join condition.

Join conditions

Select a field from each parent node for the join condition.

Accelerometer Landing

Customer Trusted Zone

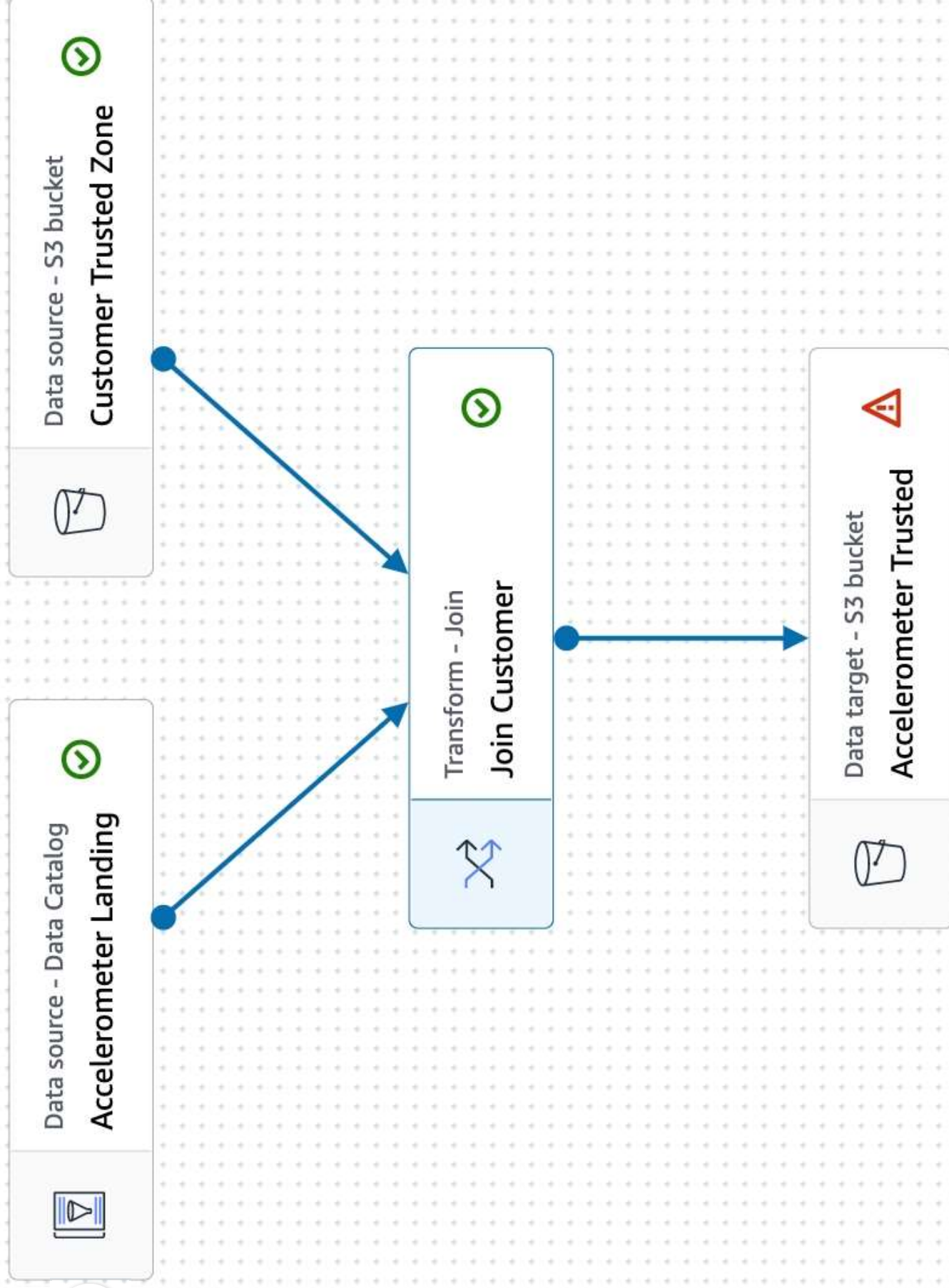
=

Add condition

All fields must have options selected

Configure the Join

Congratulations! You have created a join that will automatically drop Accelerometer rows unless they can be joined to a customer record in the **Trusted Zone**:



Accelerometer Trusted Zone

Click the **Accelerometer Trusted** Node

- Click Data target properties tab
- Add the new S3 path to the Accelerometer Trusted Zone: be sure it ends with a /
- **Do not update the Data Catalog**
- Choose the **JSON** format
- Leave Compression Type **None**



Format

JSON



Compression Type

None



S3 Target Location

Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).



s3://seans-stedi-lakehouse/accelerometer/trusted/



View



Browse S3

Data Catalog update options [Info](#)

Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

- ☒ Do not update the Data Catalog
- ☐ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
- ☐ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Partition keys - *optional*

Add partition keys.

Add a partition key

Glue Table

In the previous exercises, we used the Glue Console and the Athena Query editor to create a **Glue Table**. Choose one of those methods to create an **Accelerometer Trusted zone** table.

Click output schema to see the generated schema, notice the fields from **both** tables appear

Node properties	Data target properties - S3	Output schema	Data preview	
Schema				
Key			Data type	Partition
user			string	-
timestamp			long	-
x			float	-
y			float	-
z			float	-