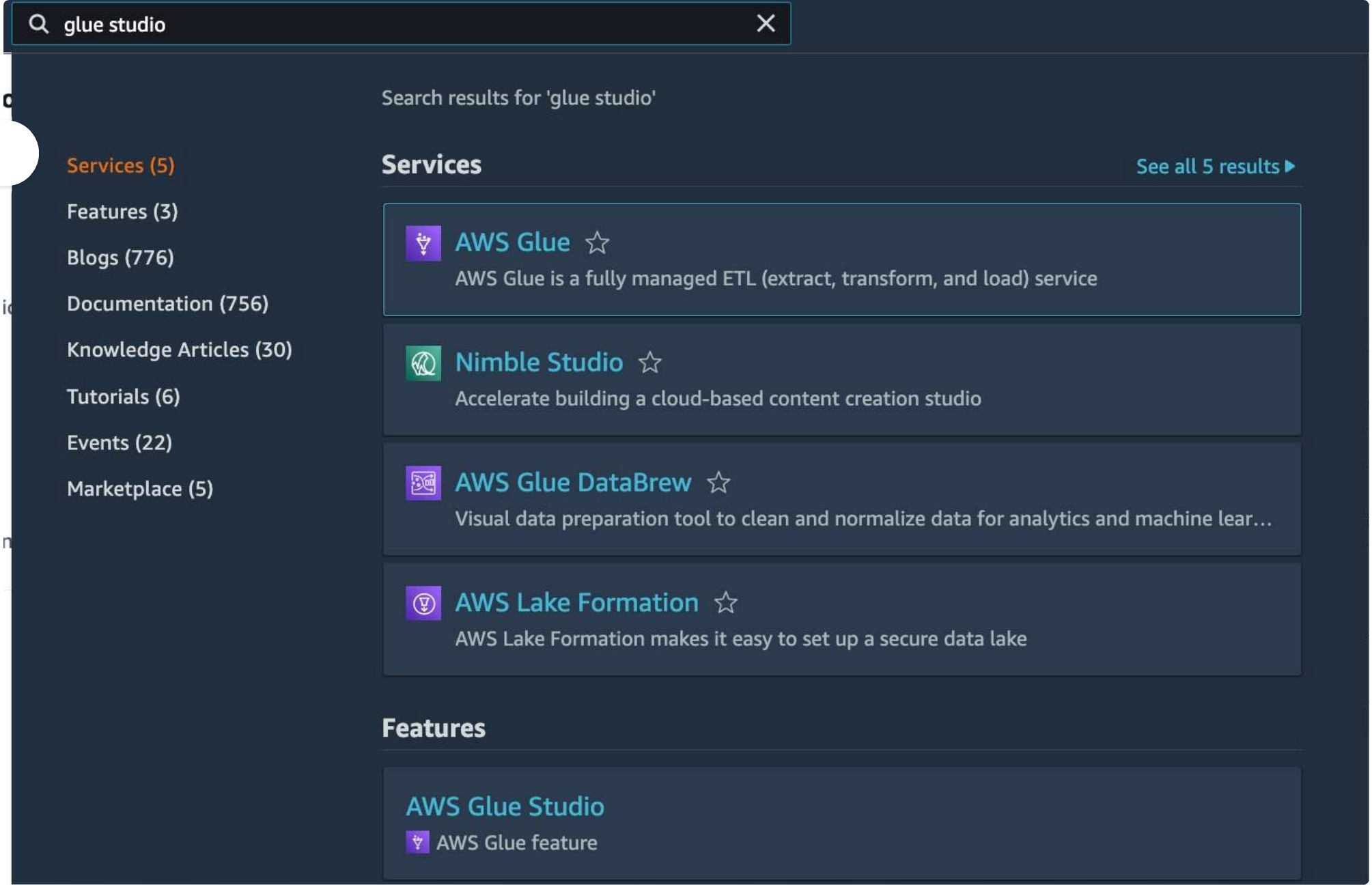
Joining with Secondary Tables

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.

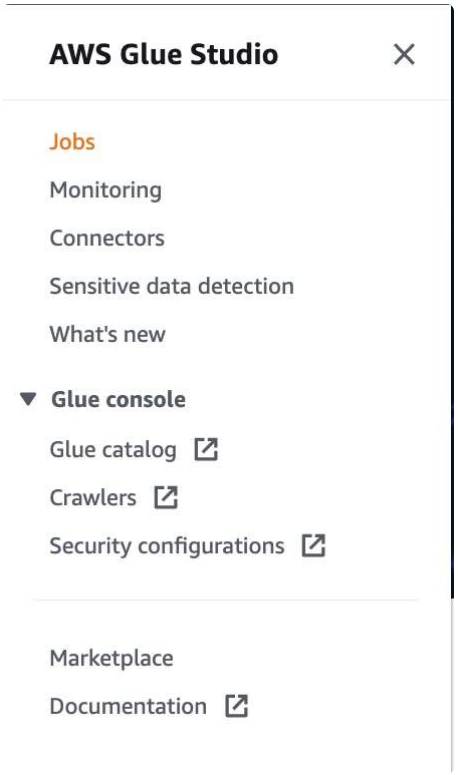
Search for Glue Studio in the AWS Console search bar



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

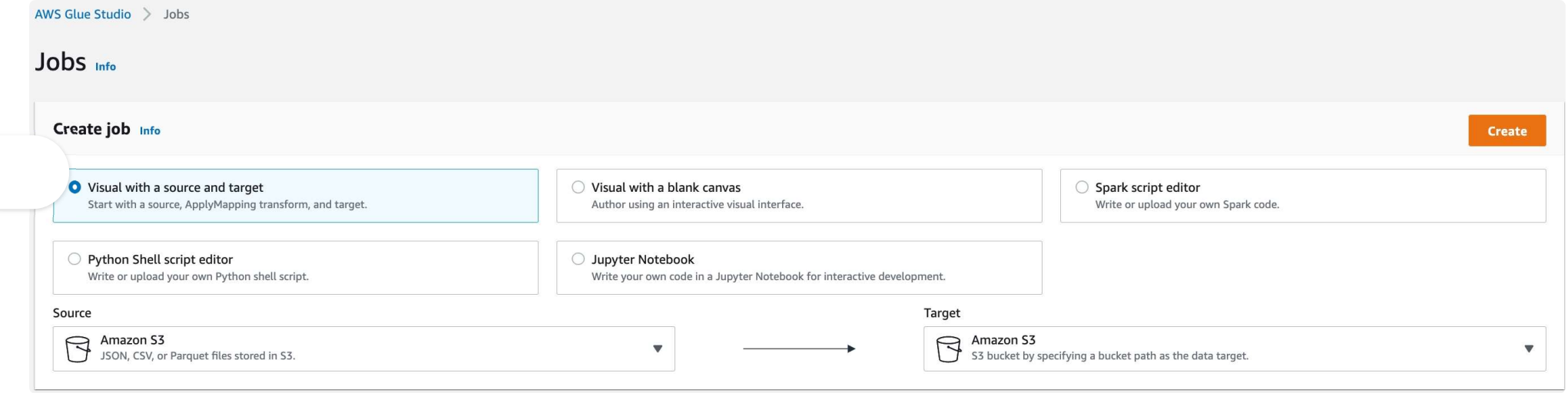
Click the Hamburger menu

Click **Jobs**



Jobs menu

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

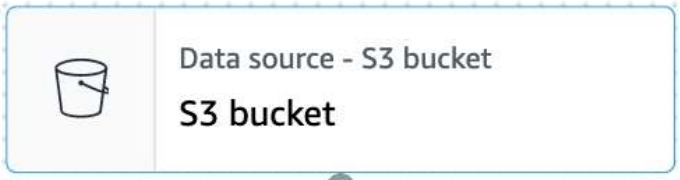


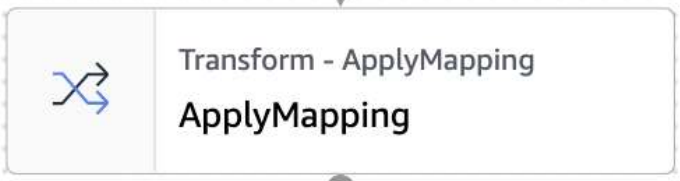
You should see the **default** visual data flow

Untitled job

Visual | Script | Job details | Runs | Schedules

Source | Transform | Target | Undo | Redo | Remove





Visual View of New Job

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)

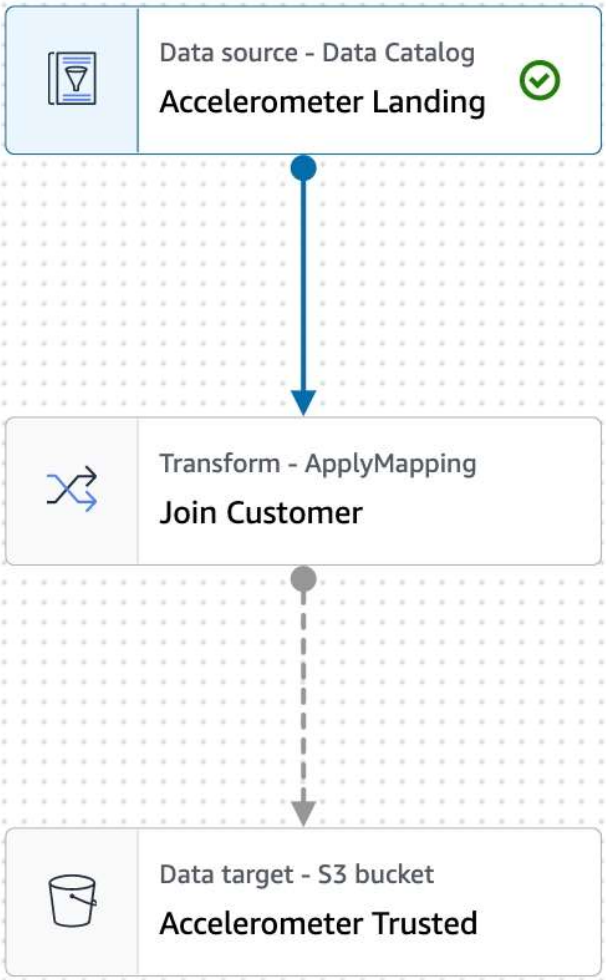
The Visual Graph should look similar to this (**we're not finished!**):

- 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

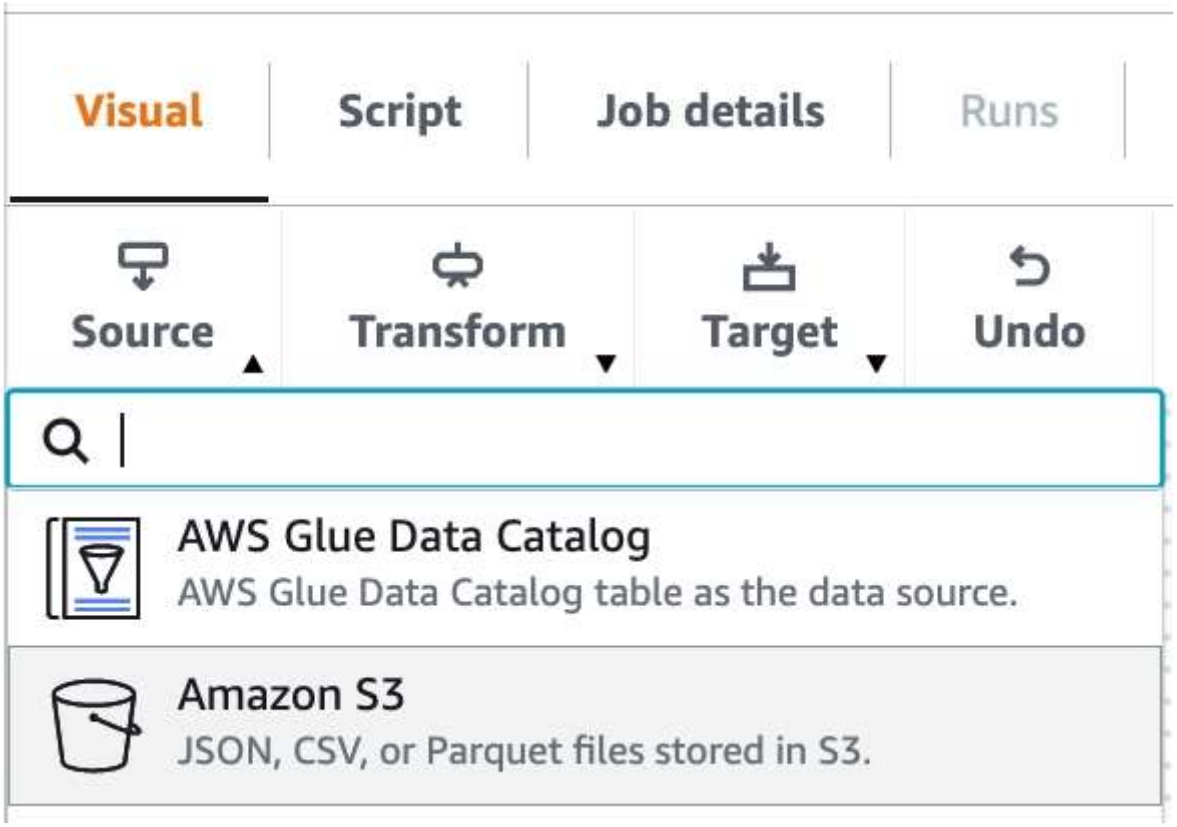
Visual | Script | Job details | Runs | Schedules

Source | Transform | Target | Undo | Redo | Remove



Configure the Node

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



AW

- 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

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

Node properties | Data source properties - S3 | Output schema | Data preview

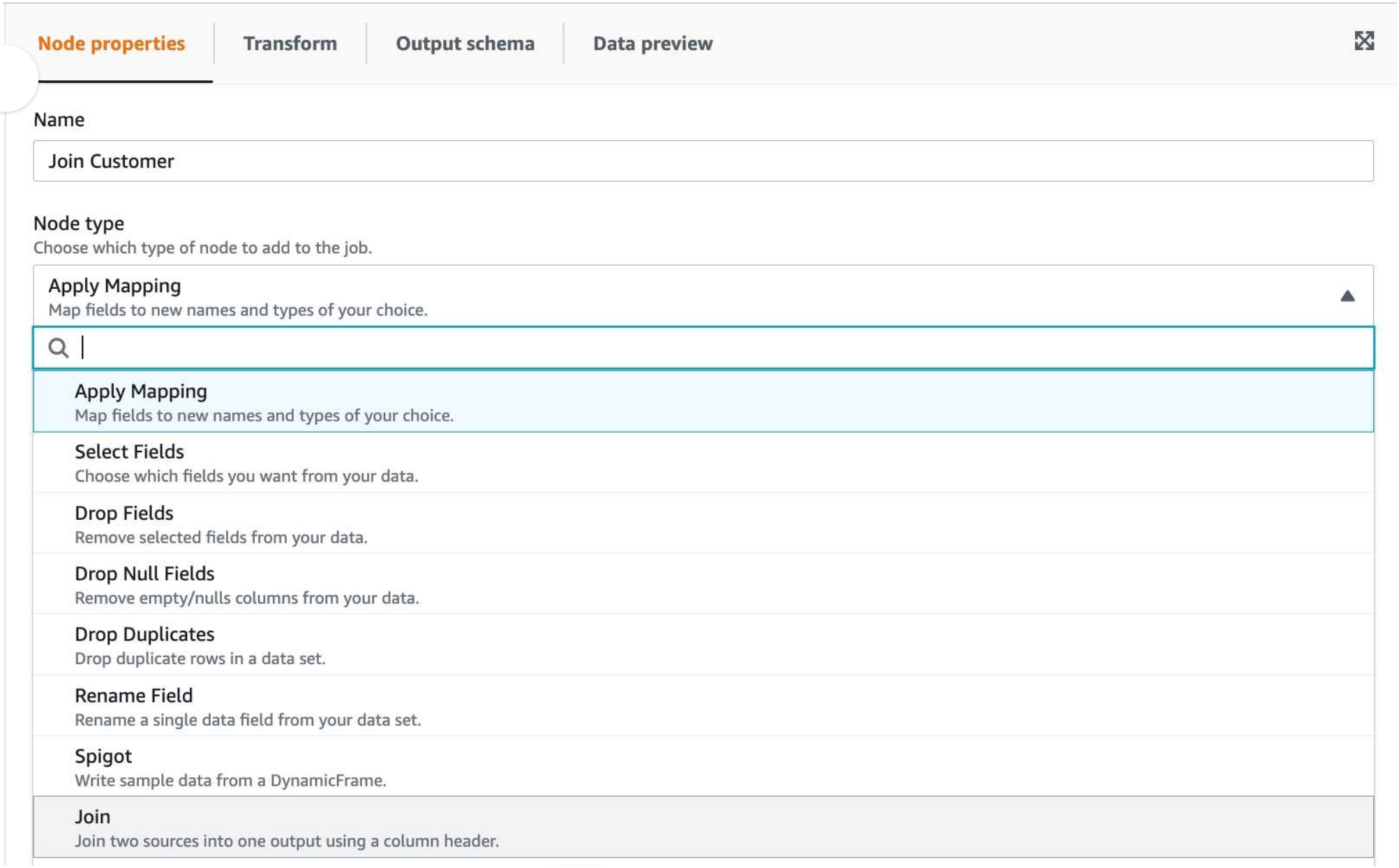
Schema Info

Edit

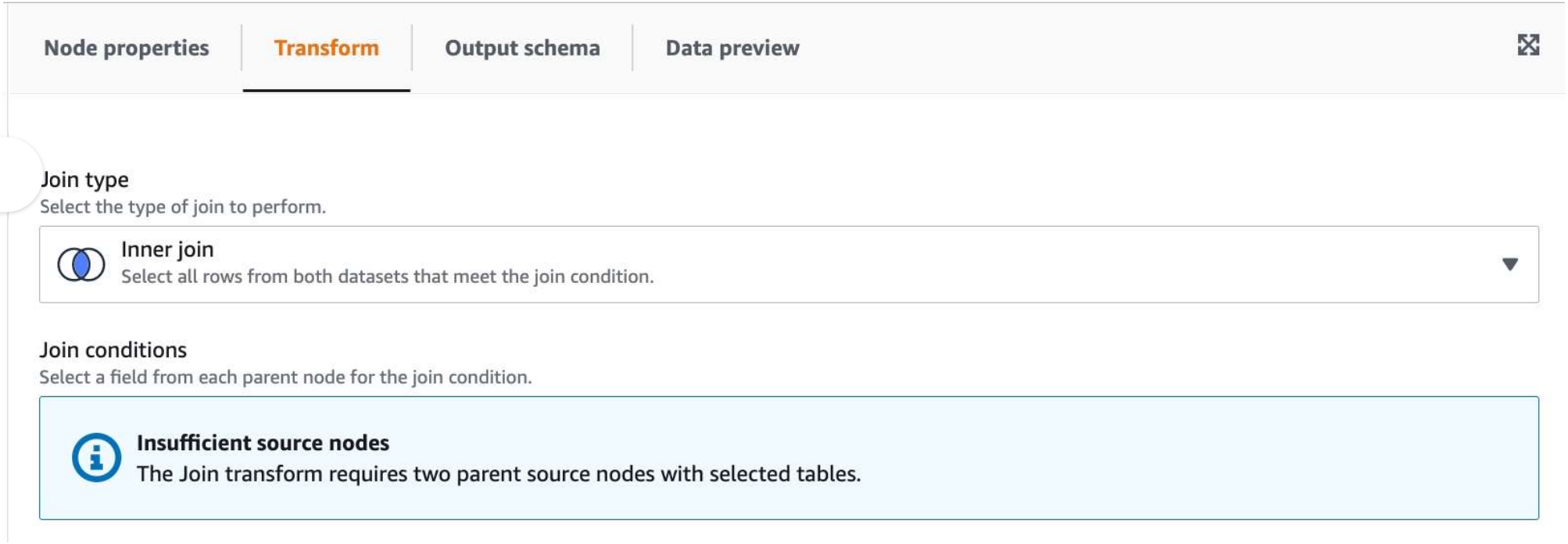
Key	Data type	Partition
serialNumber	string	-
shareWithPublicAsOfDate	long	-
birthDay	string	-
registrationDate	long	-
shareWithResearchAsOfDate	long	-
customerName	string	-
email	string	-
lastUpdateDate	long	-
phone	string	-
shareWithFriendsAsOfDate	long	-

Customer Trusted Zone Data Source Schema

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



Check out our newly created join!



Connect the **Customer Trusted Zone** node

Node properties | Transform | Output schema | Data preview

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

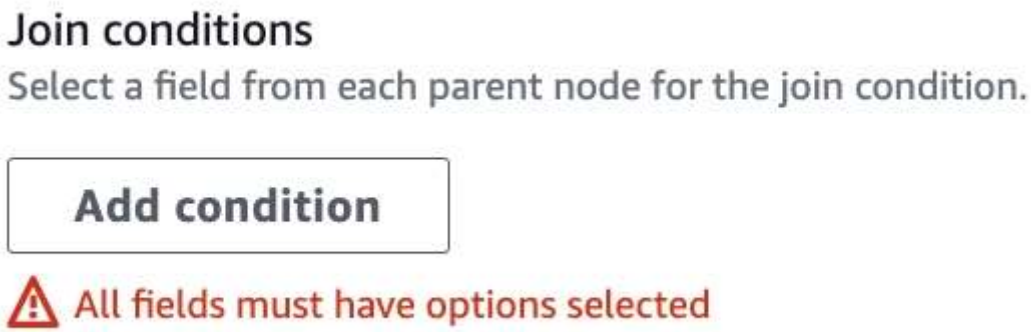
Data sources

- [x] Accelerometer Landing
  Catalog - DataSource
- [ ] Customer Trusted Zone
  S3 - DataSource

Transforms

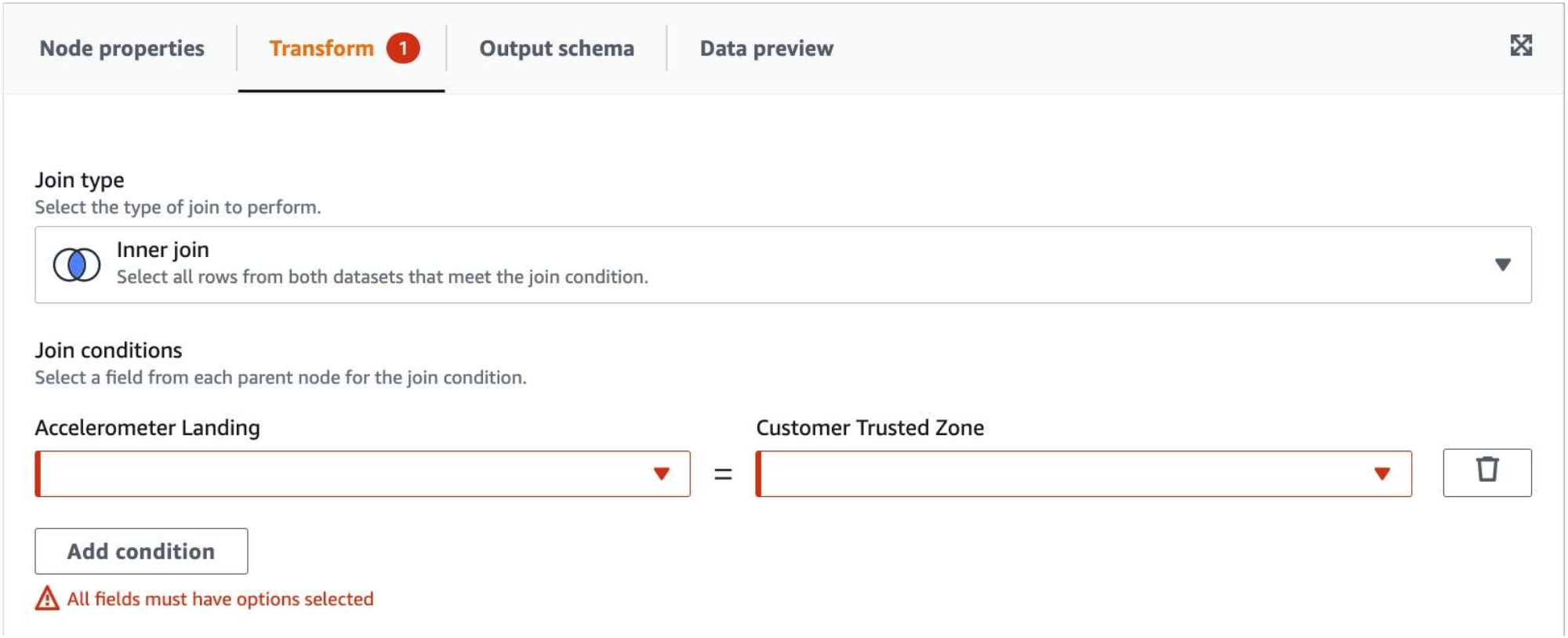
Unclassified nodes

Click Add Condition



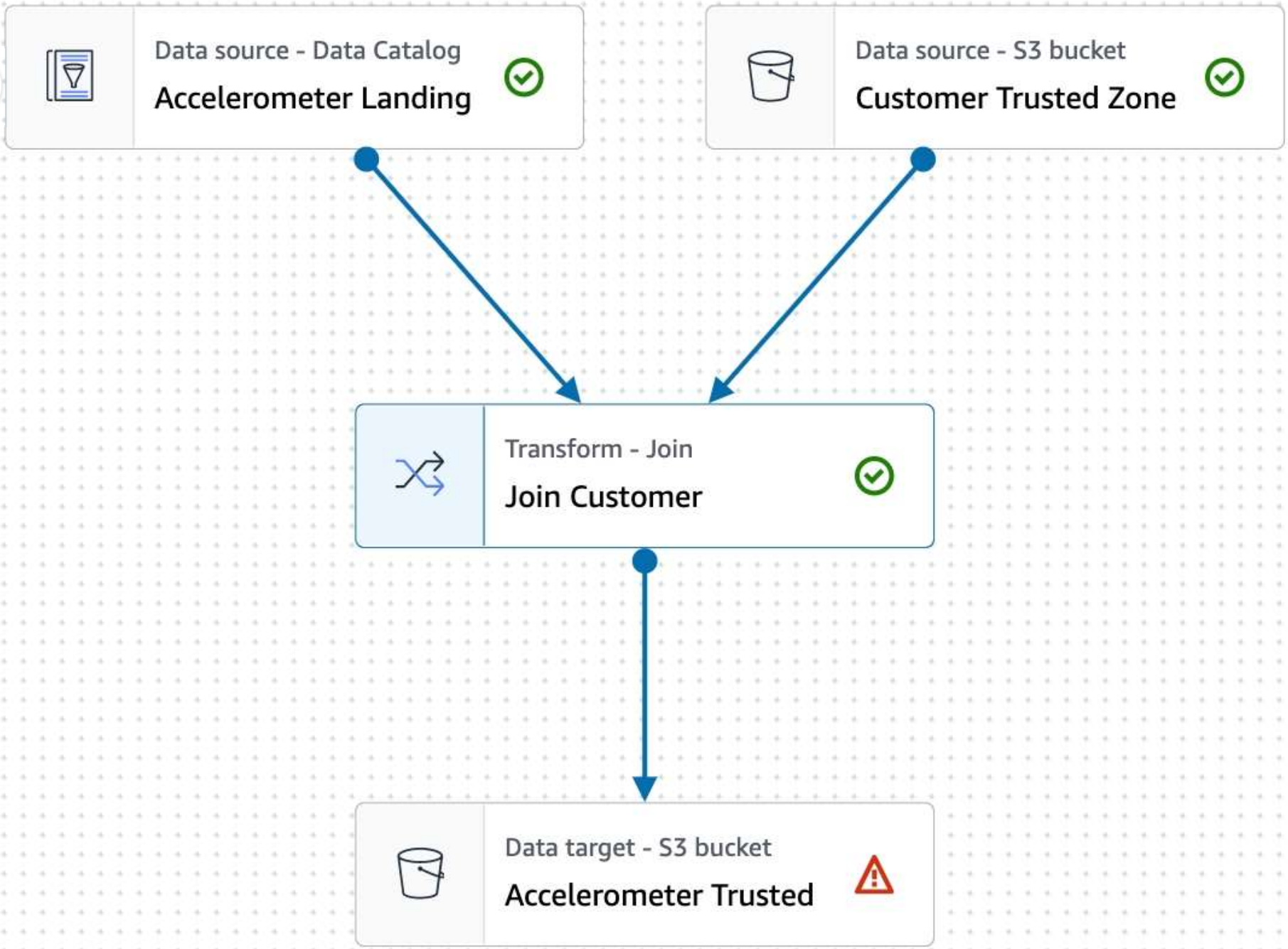
Add Condition

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



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:**


Data source - Data Catalog
Accelerometer Landing
Data source - S3 bucket
Customer Trusted Zone
Transform - Join
Join Customer
Data target - S3 bucket
Accelerometer Trusted

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**

Node properties | Data target properties - S3 | Output schema | Data preview

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 define 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	-