

Final project

- You have decided you want to get a feel for the data you are dealing with in a semi-structured format, so you decide to create **three Glue tables** for the three landing zones. Share your customer_landing.sql, accelerometer_landing.sql, and step_trainer_landing.sql scripts in git.

Customer_Landing:

```
CREATE EXTERNAL TABLE IF NOT EXISTS `finaldatabase`.`customer_landing` (  
  `customerName` string,  
  `email` string,  
  `phone` string,  
  `birthDay` string,  
  `serialNumber` string,  
  `registrationDate` bigint,  
  `lastUpdateDate` bigint,  
  `shareWithResearchAsOfDate` bigint,  
  `shareWithPublicAsOfDate` bigint,  
  `shareWithFriendsAsOfDate` bigint  
)  
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'  
WITH SERDEPROPERTIES (  
  'ignore.malformed.json' = 'FALSE',  
  'dots.in.keys' = 'FALSE',  
  'case.insensitive' = 'TRUE',  
  'mapping' = 'TRUE'  
)  
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT  
'org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat'  
LOCATION 's3://finalprojectsaleem/customer/landing/'  
TBLPROPERTIES ('classification' = 'json');
```

customer_landing_sq | X | accelerometer_landi... | X | steptrainer_sql | X | Query 15 | X | Query 16 | X | > | + | ▼

```

1 CREATE EXTERNAL TABLE IF NOT EXISTS `finaldatabase`.`customer_landing` (
2   `customerName` string,
3   `email` string,
4   `phone` string,
5   `birthDay` string,
6   `serialNumber` string,
7   `registrationDate` bigint,
8   `lastUpdateDate` bigint,
9   `shareWithResearchAsOfDate` bigint,
10  `shareWithPublicAsOfDate` bigint,
11  `shareWithFriendsAsOfDate` bigint
12 )
13 ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
14 WITH SERDEPROPERTIES (
15   'ignore.malformed.json' = 'FALSE',

```

SQL Ln 22, Col 43

Run again Explain Cancel Clear Create ▼

Reuse query results up to 60 minutes ago

Query results Query status

Completed Time in queue: 45 ms Run time: 411 ms Data scanned: -

Query successful.

Accelerometer_landing:

```

CREATE EXTERNAL TABLE IF NOT EXISTS `finaldatabase`.`accelerometer_landing` (
  `user` string,
  `timestamp` bigint,
  `x` float,
  `y` float,
  `z` float
)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
WITH SERDEPROPERTIES (
  'ignore.malformed.json' = 'FALSE',
  'dots.in.keys' = 'FALSE',
  'case.insensitive' = 'TRUE',
  'mapping' = 'TRUE'
)
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://finalprojectsaleem/accelerometer/landing/'
TBLPROPERTIES ('classification' = 'json');

```

customer_landing_sql
accelerometer_landi...
steptrainer_sql
Query 15
Query 16

```

1 CREATE EXTERNAL TABLE IF NOT EXISTS `finaldatabase`.`accelerometer_landing` (
2   `user` string,
3   `timestamp` bigint,
4   `x` float,
5   `y` float,
6   `z` float
7 )
8 ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
9 WITH SERDEPROPERTIES (
10   'ignore.malformed.json' = 'FALSE',
11   'dots.in.keys' = 'FALSE',
12   'case.insensitive' = 'TRUE',
13   'mapping' = 'TRUE'
14 )
15 STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io

```

SQL Ln 8, Col 30

Run again Explain Cancel Clear Create

☐ Reuse query results up to 60 minutes ago

Query results Query status

Completed
Time in queue: 44 ms Run time: 397 ms Data scanned: -

Query successful.

Steptrainer_landing:

```

CREATE EXTERNAL TABLE IF NOT EXISTS `finaldatabase`.`steptrainer_landing` (
  `sensorReadingTime` bigint,
  `serialNumber` string,
  `distanceFromObject` int
)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
WITH SERDEPROPERTIES (
  'ignore.malformed.json' = 'FALSE',
  'dots.in.keys' = 'FALSE',
  'case.insensitive' = 'TRUE',
  'mapping' = 'TRUE'
)
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://finalprojectsaleem/step_trainer/landing/'
TBLPROPERTIES ('classification' = 'json');

```

The screenshot displays the AWS Athena console interface. At the top, there are tabs for various queries, including 'customer_landing_sql', 'accelerometer_landi...', 'steptrainer_sql', 'Query 15', and 'Query 16'. The active query is 'steptrainer_sql', which contains the following SQL code:

```
1 CREATE EXTERNAL TABLE IF NOT EXISTS `finaldatabase`.`steptrainer_landing` (  
2   `sensorReadingTime` bigint,  
3   `serialNumber` string,  
4   `distanceFromObject` int  
5 )  
6 ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'  
7 WITH SERDEPROPERTIES (  
8   'ignore.malformed.json' = 'FALSE',  
9   'dots.in.keys' = 'FALSE',  
10  'case.insensitive' = 'TRUE',  
11  'mapping' = 'TRUE'  
12 )  
13 STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io  
   .HiveIgnoreKeyTextOutputFormat'  
14 LOCATION 's3://finalprojectsaleem/step_trainer/landing/'
```

Below the SQL editor, the status bar indicates 'SQL Ln 11, Col 21'. Action buttons include 'Run again' (orange), 'Explain' (blue), 'Cancel' (grey), 'Clear' (blue), and 'Create' (grey). A toggle for 'Reuse query results' is set to 'on', with a note 'up to 60 minutes ago'. The 'Query results' tab is active, showing a green bar with a checkmark and the text 'Completed'. Performance metrics are displayed: 'Time in queue: 131 ms', 'Run time: 538 ms', and 'Data scanned: -'. The final status is 'Query successful.'

- Query those tables using Athena, and take a screenshot of each one showing the resulting data. Name the screenshots customer_landing(.png,.jpeg, etc.), accelerometer_landing(.png,.jpeg, etc.), step_trainer_landing (.png, .jpeg, etc.).

Customer_landing:

accelerometer_landi...steptrainer_sqlQuery 15Query 16Query 17

1select * from customer_landing;

SQLLn 1, Col 32

Run again

Explain

Cancel

Clear

Create

Reuse query results

up to 60 minutes ago

Query results

Query status

Completed

Time in queue: 69 ms

Run time: 653 ms

Data scanned: 286.89 KB

Results (956)

Copy

Download results CSV

Search rows

< 1 ... >

Query results

Query status

Completed

Time in queue: 69 ms

Run time: 653 ms

Data scanned: 286.89 KB

Results (956)

Copy

Download results CSV

Search rows

< 1 ... >

#	customername	email	phone	birthday	serialnumber	registrat
1	Santosh Clayton	Santosh.Clayton@test.com	8015551212	1900-01-01	50f7b4f3-7af5-4b07-a421-7b902c8d2b7c	1655564
2	Ben Khatib	Ben.Khatib@test.com	8015551212	1399-01-01	20400202-c1da-4328-8e94-5bacc7d61ecb	1655564
3	Frank Hansen	Frank.Hansen@test.com	8015551212	1323-01-01	454a7430-d4ff-47cf-8666-7428f8a9894d	1655564
4	David Jones	David.Jones@test.com	8015551212	1111-01-01	fe784255-4faf-47ab-96f7-ea79c6e87f0c	1655564
5	Edward Clayton	Edward.Clayton@test.com	8015551212	1754-01-01	3d798cef-bc66-4128-85af-2c9d36c251b9	1655564
6	Edward Staples	Edward.Staples@test.com	8015551212	1631-01-01	9f14e906-5e19-4a70-9016-6e58f146d3f0	1655564
7	Jane Gonzalez	Jane.Gonzalez@test.com	8015551212	1141-01-01	2b6fe805-9059-4e1b-b7ed-720f5b4bf10c	1655564
8	Jaya Jackson	Jaya.Jackson@test.com	8015551212	1098-01-01	220437fd-77f5-463a-8d93-f59c944e60d4	1655564
9	Jerry Jefferson	Jerry.Jefferson@test.com	8015551212	1964-01-01	f1ad9449-f80b-4dcb-8fab-66352e929322	1655564
10	Jacob Habschied	Jacob.Habschied@test.com	8015551212	1934-01-01	276e1713-8675-495b-a4b3-c5015f0ca3a2	1655564

Accelerometer_landing:

accelerometer_landi...steptrainer_sqlQuery 15Query 16Query 17

1select * from accelerometer_landing;

SQLLn 1, Col 36

Run again

Explain

Cancel

Clear

Create

Reuse query results
up to 60 minutes ago

Query results

Query status

Completed

Time in queue: 104 ms

Run time: 1.336 sec

Data scanned: 6.55 MB

Results (81,273)

Copy

Download results CSV

Completed

Time in queue: 104 ms

Run time: 1.336 sec

Data scanned: 6.55 MB

Results (81,273)

Copy

Download results CSV

Search rows

< 1 ... >

#	user	timestamp	x	y	z
1	Jane.Olson@test.com	1655564439144	1.0	0.0	-1.0
2	Jane.Olson@test.com	1655564439144	1.0	-1.0	-1.0
3	Jane.Olson@test.com	1655564430945	-1.0	1.0	-1.0
4	Jane.Olson@test.com	1655564430945	-1.0	-1.0	-1.0
5	Jane.Olson@test.com	1655564417280	-1.0	-1.0	0.0
6	Jane.Olson@test.com	1655564414547	-1.0	0.0	0.0
7	Jane.Olson@test.com	1655564414547	0.0	-1.0	-1.0
8	Jane.Olson@test.com	1655564411814	0.0	-1.0	0.0
9	Jane.Olson@test.com	1655564411814	-1.0	-1.0	0.0
10	Danny.Hansen@test.com	1655564133234	1.0	-1.0	0.0
11	Danny.Hansen@test.com	1655564119902	0.0	0.0	-1.0
12	Dannv.Hansen@test.com	1655564103237	-1.0	-1.0	-1.0

Steptrainer_landing:

accelerometer_landi...
steptrainer_sql
Query 15
Query 16
Query 17

```
1 select * from steptrainer_landing;
```

SQL Ln 1, Col 34

Run again Explain Cancel Clear Create

☐ Reuse query results up to 60 minutes ago

Query results Query status

Completed
Time in queue: 107 ms Run time: 929 ms Data scanned: 3.15 MB

Results (28,680)
Copy Download results CSV

Completed
Time in queue: 107 ms Run time: 929 ms Data scanned: 3.15 MB

Results (28,680)
Copy Download results CSV

Search rows

#	sensorreadingtime	serialnumber	distancefromobject
1	1655564407947	81624e88-1469-4f54-977f-3dcb161ef79e	281
2	1655564365047	81624e88-1469-4f54-977f-3dcb161ef79e	272
3	1655564361147	81624e88-1469-4f54-977f-3dcb161ef79e	257
4	1655564353347	81624e88-1469-4f54-977f-3dcb161ef79e	260
5	1655564494769	058af4ef-68ed-4af7-b51d-b6223e5300ff	293
6	1655564488103	058af4ef-68ed-4af7-b51d-b6223e5300ff	278
7	1655564478104	058af4ef-68ed-4af7-b51d-b6223e5300ff	266
8	1655564471438	058af4ef-68ed-4af7-b51d-b6223e5300ff	269
9	1655564451440	058af4ef-68ed-4af7-b51d-b6223e5300ff	240
10	1655564441441	058af4ef-68ed-4af7-b51d-b6223e5300ff	268
11	1655564418110	058af4ef-68ed-4af7-b51d-b6223e5300ff	272
12	1655564411444	058af4ef-68ed-4af7-b51d-b6223e5300ff	249

The Data Science team has done some preliminary data analysis and determined that the Accelerometer Records each match one of the Customer Records. They would like you to create 2 AWS Glue Jobs that do the following:

- Sanitize the Customer data from the Website (Landing Zone) and only store the Customer Records who agreed to share their data for research purposes (Trusted Zone) - creating a Glue Table called `customer_trusted`.

- Customer_trusted:**

Customer Landing to Trusted
Last modified on 22/04/2025, 11:21:58
Actions Save Run

Visual
Script
Job details
Runs
Data quality
Schedules
Version Control

```

graph TD
    A[Data source - Data Catalog  
AWS Glue Data Catalog] --> B[Transform - SQL Query  
SQL Query]
    B --> C[Data target - S3 bucket  
Amazon S3]
        
```

Associate an alias with each input source Info

Edit the aliases used for the inputs to this node.

Input sources	SQL aliases
AWS Glue Data Catalog	customer_landing

SQL query

Enter a SQL statement to add to your job.

```

1 select * from customer_landing
2 where sharewithresearchasofdate is not null;
        
```

Data preview
Output schema

Data preview (200) Info READY ⓘ ⌂ End session Previewing 10 of 10 fields

Filter sample dataset

customerName	email	phone	birthDay
--------------	-------	-------	----------

accelerometer_landi...

steptrainer_sql

Query 15

Query 16

Query 17

1 select * from customer_trusted;

SQLLn 1, Col 32

Run again

Explain

Cancel

Clear

Create

Reuse query results
up to 60 minutes ago

Query results

Query status

Completed

Time in queue: 102 ms

Run time: 578 ms

Data scanned: 161.06 KB

Results (482)

Copy

Download results CSV

Completed

Time in queue: 102 ms

Run time: 578 ms

Data scanned: 161.06 KB

Results (482)

Copy

Download results CSV

Search rows

< 1 ... >

#	customername	email	phone	birthday	serialnumber	registrator
1	Santosh Clayton	Santosh.Clayton@test.com	8015551212	1900-01-01	50f7b4f3-7af5-4b07-a421-7b902c8d2b7c	165556437
2	Ben Khatib	Ben.Khatib@test.com	8015551212	1399-01-01	20400202-c1da-4328-8e94-5bacc7d61ecb	165556441
3	Jane Gonzalez	Jane.Gonzalez@test.com	8015551212	1141-01-01	2b6fe805-9059-4e1b-b7ed-720f5b4bf10c	165556443
4	Jaya Jackson	Jaya.Jackson@test.com	8015551212	1098-01-01	220437fd-77f5-463a-8d93-f59c944e60d4	165556443
5	Jerry Jefferson	Jerry.Jefferson@test.com	8015551212	1964-01-01	f1ad9449-f80b-4dcb-8fab-66352e929322	165556408
6	Senthil Smith	Senthil.Smith@test.com	8015551212	1587-01-01	dbfde3b7-69bf-4c56-a9e8-dfd48af6df89	165556440
7	Santosh Davis	Santosh.Davis@test.com	8015551212	1745-01-01	b5f6b144-0260-4c4f-83e1-469150f9055b	165556438
8	Sarah Doshi	Sarah.Doshi@test.com	8015551212	1126-01-01	466bb1b9-5dc3-4607-8f76-e0bcdcf6b9c7	165556414
9	Ben Phillips	Ben.Phillips@test.com	8015551212	1404-01-01	314ed64d-643f-4cd3-b164-f4a8ae02fc29	165556412
10	David Mitra	David.Mitra@test.com	8015551212	1201-01-01	ba582015-fbb3-4842-887a-e291f773df84	165556442
11	Christina Mitra	Christina.Mitra@test.com	8015551212	1160-01-01	7414a106-ee97-4824-ab25-70bedeae1136	165556443

Schema

Partitions

Indexes

Column statistics - new

Schema (10)

Edit schema as JSON

Edit schema

View and manage the table schema.

Q Filter schemas

< 1 > ⚙

#	Column name	Data type	Partition key	Comment
1	customername	string	-	-
2	email	string	-	-
3	phone	string	-	-
4	birthday	string	-	-
5	serialnumber	string	-	-
6	registrationdate	bigint	-	-
7	lastupdatedate	bigint	-	-
8	sharewithresearchasofdate	bigint	-	-
9	sharewithpublicasofdate	bigint	-	-
10	sharewithfriendsasofdate	bigint	-	-

Accelerometer_trusted:

accelerometer_trusted

Last modified on 22/04/2025, 11:29:21

Actions

Save

Run

Visual

Script

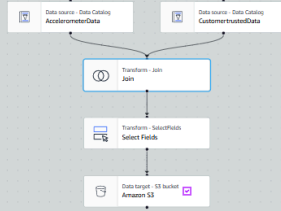
Job details

Runs

Data quality

Schedules

Version Control



Join

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

AccelerometerData X
Catalog - DataSource

CustomertrustedData X
Catalog - DataSource

Join type

Select the type of join to perform.

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

Join conditions

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

AccelerometerData
user

CustomertrustedData
a
email

Add condition

Data preview

Output schema

Data preview

Info

READY



End session

Previewing 0 of 0 fields

accelerometer_landi... steptrainer_sql Query 15 Query 16 Query 17

1 select * from accelerometer_trusted;

SQL Ln 1, Col 37

Run again

Explain

Cancel

Clear

Create

Reuse query results
up to 60 minutes ago

Query results

Query status

Completed

Time in queue: 99 ms

Run time: 761 ms

Data scanned: 3.30 MB

Results (40,981)

Copy

Download results CSV

Search rows

1 ...

Completed

Time in queue: 99 ms

Run time: 761 ms

Data scanned: 3.30 MB

Results (40,981)

Copy

Download results CSV

Search rows

< 1 ... > ⚙

#	z	user	y	x	timestamp
1	-1.0	Senthil.Huey@test.com	1.0	0.0	1655564499136
2	-1.0	Senthil.Huey@test.com	0.0	0.0	1655564499136
3	0.0	Senthil.Huey@test.com	0.0	1.0	1655564481340
4	-1.0	Senthil.Huey@test.com	-1.0	1.0	1655564475408
5	-1.0	Senthil.Huey@test.com	1.0	0.0	1655564475408
6	0.0	Senthil.Huey@test.com	1.0	0.0	1655564457612
7	0.0	Senthil.Huey@test.com	0.0	0.0	1655564436850
8	-1.0	Senthil.Huey@test.com	1.0	-1.0	1655564430918
9	0.0	Senthil.Huey@test.com	0.0	1.0	1655564424986
10	0.0	Senthil.Huey@test.com	0.0	1.0	1655564502102
11	-1.0	Senthil.Huey@test.com	-1.0	-1.0	1655564496170

Schema

Partitions

Indexes

Column statistics - new

Schema (5)

View and manage the table schema.

Edit schema as JSON

Edit schema

Filter schemas

< 1 > ⚙

#	Column name	Data type	Partition key	Comment
1	z	double	-	-
2	user	string	-	-
3	y	double	-	-
4	x	double	-	-
5	timestamp	bigint	-	-