

## Usecase 1: Dataset Data Integration using ML

A use case to match records across different datasets as well as identify and remove duplicate records, with little to no human intervention.

## Input

- Neat Dataset1 with some records (~2600).
- Dataset2 with some records (~64,000). Messy, duplicated.
- Matching records (labels) (~130)

## Steps

Reformat files → Create IAM role → upload to S3 bucket → Athena  
datasource → data glue ML → data glue job → get output file from  
S3

1. Change dataset format to JSON file with encoding **'utf-8'** and added one column called source to each dataset file using python script.

```
Open  dataset1.json
Data_FS/media/basma/Data_FS/AWS/usecases/PDSA_UseCase1_Resources/Students resources Save

1{"id": "conf_vldb_RusinkiewiczKTWM95", "title": "Towards a Cooperative Transaction Model - The Cooperative Activity Model", "author": "M
Rusinkiewicz, W Klas, T Tesch, J W\u00e4\u00bfg\u00bfsch, P Muth", "venue": "VLDB", "year": "1995", "source": "dataset1"}
2{"id": "journals_sigmod_EisenbergM02", "title": "SQL/XML is Making Good Progress", "author": "A Eisenberg, J Melton", "venue": "SIGMOD
Record", "year": "2002", "source": "dataset1"}
3{"id": "conf_vldb_AmmannJR95", "title": "Using Formal Methods to Reason about Semantics-Based Decompositions of Transactions", "author":
"P Ammann, S Jajodia, I Ray", "venue": "VLDB", "year": "1995", "source": "dataset1"}
4{"id": "journals_sigmod_Liu02", "title": "Editor's Notes", "author": "L Liu", "venue": "SIGMOD Record", "year": "2002", "source":
"dataset1"}
5{"id": "journals_sigmod_Hammer02", "title": "Report on the ACM Fourth International Workshop on Data Warehousing and OLAP (DOLAP 2001)",
"author": "N/A", "venue": "N/A", "year": "2002", "source": "dataset1"}
6{"id": "conf_vldb_FerrandinaMZFM95", "title": "Schema and Database Evolution in the O2 Object Database System", "author": "F Ferrandina,
T Meyer, R Zicari, G Ferran, J Madec", "venue": "VLDB", "year": "1995", "source": "dataset1"}
7{"id": "conf_vldb_SubietaKL95", "title": "Procedures in Object-Oriented Query Languages", "author": "K Subieta, Y Kambayashi, J
Leszczylowski", "venue": "VLDB", "year": "1995", "source": "dataset1"}
8{"id": "journals_sigmod_Bargal02", "title": "Phoenix Project: Fault-Tolerant Applications", "author": "R Barga, D Lomet", "venue":
```

2. Created S3 bucket and uploaded 2 dataset files and label file in CSV format.

Amazon S3 > usecase-1-bn

## usecase-1-bn [Info](#)

**Objects** | Properties | Permissions | Metrics | Management | Access Points

**Objects (5)**

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	dataset_json/	Folder	-	-	-
<input type="checkbox"/>	datasets/	Folder	-	-	-
<input type="checkbox"/>	label/	Folder	-	-	-
<input type="checkbox"/>	migrated_file.json	json	January 17, 2022, 11:54:39 (UTC+02:00)	13.7 MB	Standard
<input type="checkbox"/>	output/	Folder	-	-	-

Amazon S3 > usecase-1-bn > dataset\_json/

## dataset\_json/ [Copy S3 URI](#)

**Objects** | Properties

**Objects (2)**

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	dataset1.json	json	January 17, 2022, 12:52:50 (UTC+02:00)	533.4 KB	Standard
<input type="checkbox"/>	dataset2.json	json	January 17, 2022, 12:52:48 (UTC+02:00)	92.8 MB	Standard

### 3. created new IAM role for S3 and data glue

The screenshot shows the AWS IAM console interface. On the left is a navigation menu with options like Dashboard, Access management, Roles, Policies, and Access reports. The main content area is titled 'Roles > glueRole' and shows the 'Summary' tab. Key details include: Role ARN (arn:aws:iam::966008732841:role/glueRole), Role description (Allows Glue to call AWS services on your behalf), Instance Profile ARNs, Path (/), Creation time (2022-01-16 18:39 UTC+0200), Last activity (2022-01-17 22:31 UTC+0200 (Yesterday)), and Maximum session duration (1 hour). Below the summary, the 'Permissions' tab is active, showing 'Permissions policies (2 policies applied)'. A table lists the attached policies:

Policy name	Policy type	
AmazonS3FullAccess	AWS managed policy	✕
AWSGlueServiceRole	AWS managed policy	✕

### 4. Created new database using Athena to merge 2 datasets in one table.

The screenshot shows the Amazon Athena Query Editor interface. The 'Editor' tab is active, and the 'Data' panel on the left shows the 'Data Source' as 'AwsDataCatalog' and the 'Database' as 'usecase1-athena2'. The 'Tables and views' section shows a table named 'usecase1-athena2'. The main query editor displays a SQL query: `SELECT * FROM "usecase1-athena2"."usecase1-athena3" limit 10;`. The query is labeled 'Query 14'.

5. From data Glue, created a new FindMatches ML transform with 100% accuracy and 0.9 precision.

### Add transform

○ Transform properties

○ Data source

○ Primary key

○ Tune

○ Review

#### Configure transform properties

**Transform name**

transform-deduplicate

**Description (optional)**

Enter description

**IAM role** ⓘ

glueRole

Ensure that this role has permission to your Amazon S3 sources, temporary directory, and labeling files. [Create IAM role](#)

▸ Security configuration (optional)

### Choose a primary key

Choose a primary key column. This column typically contains a unique identifier for every record in the data source.

Filter by attributes or search by keyword

Showing: 1 - 6 of 6

Column name	Data type
<input checked="" type="radio"/> Id	string
<input type="radio"/> title	string
<input type="radio"/> authors	string
<input type="radio"/> venue	string
<input type="radio"/> year	double
<input type="radio"/> source	string

### Tune transform

Select a tuning option or adjust the sliders to tune the transform. Choosing a midpoint optimizes the transform to find duplicates in a reasonable time.

Recall01Precision

☐ **Balanced** (0.5) Even tradeoff between recall and precision

☐ **Favor recall** (0.2) Find more matches even if some are incorrectly matched

☒ **Favor precision** (0.9) Find matches with fewer incorrect matches

☐ Custom

Lower cost01Accuracy

☐ **Balanced** (0.5) Even tradeoff between accuracy and cost

☐ **Favor lower cost** (0.2) Use less resources to find matches

☐ **Favor accuracy** (0.9) Use more resources but potentially find more matches

☒ Custom

## Machine learning transforms Clean your data using machine learning transforms.

[Add transform](#) Action ▾  Showing: 1

Transform name ▾	Transform ID	Type ▾	Label count	Status ▾	Date created ▾	Last modified ▾	Descripti
<input type="radio"/> <a href="#">transform-deduplicate</a>	tfm-3d0c162dc4741bfd32dc...	Find matching records	129	Ready for use	17 January 2022 12:57 P...	17 January 2022 9:57 PM...	
<input type="radio"/> <a href="#">usecase1-deduplicate</a>	tfm-3646f79f29b2b185516c6...	Find matching records	0	Needs training	17 January 2022 12:02 P...	17 January 2022 12:36 P...	

### 6. Generated sample label file from glue by selecting “I don’t have labels” option.

Then added values to label column and added some matching labels from the shared label file. Then uploaded the new label file to S3.

### 7. Teach FindMatches by providing labeling records from S3.

A	B	C	D	E	F	G	H
labeling set id	label id	title	author	venue	year	source	
f570eef9-5ba6-35bd-a393-436e80d1beb6	0eBnT7lhV2LwJ	Aurora: A Data Stream	D Abadi, D Carney, U Cetin	Proceedings of the 2003 ACM SIGMOD Conference on		dataset2	
f570eef9-5ba6-35bd-a393-436e80d1beb6	0conf_sigmod_Aba	Aurora: A Data Stream	D Abadi, D Carney, U Cetin	SIGMOD Conference		2003 dataset1	
f570eef9-5ba6-35bd-a393-436e80d1beb6	0gBVNSFeS4P8J	Aurora: a new model an	DJ Abadi, D Carney, U A?	The VLDB Journal The International Journa		2003 dataset2	
f570eef9-5ba6-35bd-a393-436e80d1beb6	0VuY9Y49GqXgJ	Aurora: A Data Stream	DJ Abadi, D Carney, U Cetin	Intel, M Cherniack, C		dataset2	
f570eef9-5ba6-35bd-a393-436e80d1beb6	114XNgQag2nQJ	Improving Memory in O	B Levy	JOURNAL OF PERSONALITY AND SOCIA		1996 dataset2	
f570eef9-5ba6-35bd-a393-436e80d1beb6	20_Lktn3dJgAJ	Exploring Windows NT	R Duncan	PC		dataset2	
f570eef9-5ba6-35bd-a393-436e80d1beb6	3d9u9G1dulXgJ	Word for Word	D Haskins	PC		dataset2	
f570eef9-5ba6-35bd-a393-436e80d1beb6	4yEIHjCL5UykJ	Evolution and Change i	ACMS Anthology	SIGMOD Record,		2000 dataset2	
f570eef9-5ba6-35bd-a393-436e80d1beb6	5owCJmT0Mv2IJ	Impulse Detectors for	M Lukac	RADIOENGINEERING-PRAGUE-		2001 dataset2	
f570eef9-5ba6-35bd-a393-436e80d1beb6	6oFmPKSqVCEUJ	The common capability	B Levy	BT Technology Journal,		2005 dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	0AxpQwgyRyLgJ	Active XML Document	S Abiteboul, A Bonifati, G	ACM SIGMOD		dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	0Rjb06zlxblIJ	Dynamic XML document	S Abiteboul, A Bonifati, G	SIGMOD Conference,		2003 dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	0conf_sigmod_Abit	Dynamic XML document	S Abiteboul, A Bonifati, G	SIGMOD Conference		2003 dataset1	
e4cc1af1-1345-3d5d-8551-32997f7582db	2uri:http://www.csa	Modal analysis of aere	M Gennaretti, A Corbelli, P	European Rotorcraft Forum, 25 th, Rome, It		1999 dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	3uri:http://portal.ac	Content-Based Routing	P Bizaro, S Babu, D Dev	Proceedings of the 31st international confer		2005 dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	4uri:http://portal.ac	Caching with A??Good	H Guo, P?? Larson, R Re	Proceedings of the 31st international confer		2005 dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	5uri:http://www.csa	A parametric study of t	S Shaw, N Qin	European Rotorcraft Forum, 25 th, Rome, It		1999 dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	1WWaxLMlptTMJ	Self-tuning Histograms	A Aboulnaga, S Chaudhuri	SIGMOD Conference,		1999 dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	1xnDzelm2t1tQJ	Self-tuning Histograms	AA AC, S Chaudhuri	Proceedings of the ACM SIGMOD International Confe		dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	1conf_sigmod_Abo	Self-tuning Histograms	A Aboulnaga, S Chaudhuri	SIGMOD Conference		1999 dataset1	
e4cc1af1-1345-3d5d-8551-32997f7582db	0SFmnhhoQzgJ	The Lyric Language: Q	A Brodsky, Y Kornatzky			dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	06t_N5axEm7cJ	The Lyric Language: C	A Brodsky, Y Kornatzky	&hellip; SIGMOD International Conference on Manag		dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	0conf_sigmod_Bro	The Lyric Language: Q	A Brodsky, Y Kornatzky	SIGMOD Conference		1995 dataset1	
e4cc1af1-1345-3d5d-8551-32997f7582db	17mNckZ7D-R8J	Copy Detection Mecha	S Bin, J Davis, H Garcia-P	SIGMOD Conference,		1995 dataset2	
e4cc1af1-1345-3d5d-8551-32997f7582db	1JsKYXyM6laEJ	SCAM: A Copy Detecti	N Shiyakumar, H Garcia-Molina			dataset2	

Teach transform

Generate labeling file

Label data

Upload labels

Estimate quality

Upload labels

I do not have labels

I have labels

Upload labels from S3

The completed labeling file must be in the correct format and in Amazon S3.

Upload labeling file from S3

Back

Next

Teach transform

Generate labeling file

Label data

Upload labels

Estimate quality

new estimates can be viewed in the **Quality metrics** pane of the transform.

Estimate transform quality

Quality metric	Definition	Result	Last modified
Area under the Precision-Recall curve	Single number summarizing the performance of the transform	92.6984%	01/17/22 10:03 PM
Precision	When your transform predicts a match, how often is it correct?	85.7143%	01/17/22 10:03 PM
Recall upper limit	For an actual match, how often does your transform predict a match?	100%	01/17/22 10:03 PM
F1	Indicates transform's accuracy. Harmonic mean of Precision and Recall.	92.3077%	01/17/22 10:03 PM

\* Metrics shown are from the last quality estimation run.

\*\* End-to-End recall will tend to be closer to the upper limit as the cost-accuracy slider favors accuracy. See documentation for additional information about End-to-End recall.

Want to improve your results? [Generate a new labeling file](#), label it, and upload the labels to append to our existing labels.



## Teach transform

☒ Generate labeling file

☒ Label data

☒ Upload labels

☐ Estimate quality

\*\* End-to-End recall will tend to be closer to the upper limit as the cost-accuracy slider favors accuracy. See documentation for additional information about End-to-End recall.

Want to improve your results? [Generate a new labeling file](#), label it, and upload the labels to append to our existing labels.

### Column importance

Estimate your transform's ability to find matches. Estimates are calculated by comparing the transform match predictions using a subset of your labeled data against the labels you have provided. These estimates are approximate. To improve your transform quality, provide more labels.

Column name	Importance
title	0.555
author	0.421
venue	0.014
year	0.005
source	0.004

Back

Finish

Add transformAction

Filter by tags and attributes

Showing: 1 - 2

Transform name	Transform ID	Type	Label count	Status	Date created	Last modified	Description
<input checked="" type="radio"/> transform-deduplicate	tfm-3d0c162dc4741bfd32dc...	Find matching records	129	Ready for use	17 January 2022 12:57 P...	17 January 2022 9:57 PM...	
<input type="radio"/> usecase1-deduplicate	tfm-3646f79f29b2b185516c6...	Find matching records	0	Needs training	17 January 2022 12:02 P...	17 January 2022 12:36 P...	

Transform name

transform-deduplicate

Transform ID

tfm-3d0c162dc4741bfd32dc64b51d2b0652b7b1c64d

Source database

usecase1-athena2

Source data table

usecase1-athena2

Type

Find matching records

Spark Version

2.2

IAM Role

glueRole

Status

Ready for use

Date created

17 January 2022 12:57 PM UTC+2

Last modified

17 January 2022 9:57 PM UTC+2

Description

-

Label count

129

Precision-recall tradeoff

0.9

Accuracy-cost tradeoff

1

Force output to match labels

true

Worker type

G.2X

Number of workers

10

Tags

-

Security configuration

-

8. Created an AWS Glue ETL job that uses FindMatches ML transform.

configuration is specified.

**Python library path**

**Dependent jars path**

**Referenced files path**

**Worker type** ⓘ  
G.2X (Recommended for jobs with ML transforms)

**Number of workers**

The maximum number of workers you can define are 299 for G.1X, and 149 for G.2X.

**Max concurrency** ⓘ

[Save](#)

Job: deduplicate [Action](#) [Save](#) [Run job](#) [Generate diagram](#) ⓘ

Insert template at cursor ⓘ [Source](#) [Target](#) [Target Location](#) [Transform](#) [Spigot](#) ⓘ

**Database Name** usecase1-athena2  
**Table Name** usecase1-athena2

**Transform Name** ResolveChoice

**Transform Name** FindMatches

**Path** s3://usecase-1-bn/output

```
1 import sys
2 from awsglue.transforms import *
3 from awsglue.utils import getResolvedOptions
4 from pyspark.context import SparkContext
5 from awsglue.context import GlueContext
6 from awsglue.job import Job
7 from awsglue.transforms import FindMatches
8
9 ## @params: [JOB_NAME]
10 args = getResolvedOptions(sys.argv, ['JOB_NAME'])
11
12 sc = SparkContext()
13 glueContext = GlueContext(sc)
14 spark = glueContext.spark_session
15 job = Job(glueContext)
16 job.init(args['JOB_NAME'], args)
17
18 ## @type: DataSource
19 ## @return: DataSource
20
21 ## @inputs: []
22 datasource0 = glueContext.create_dynamic_frame.from_catalog(database = "usecase1-athena2", table_name = "usecase1-athena2", transformation_ctx = "datasource0")
23
24 ## @type: ResolveChoice
25
26
27
28
29
30
31
32
33
34
```

[Logs](#) [Schema](#)



9. Downloaded all output files from S3 and merged to one CSV file.

```
aws s3 cp --recursive s3://usecase-1-bn/usecase_1_output/ .  
awk '(NR == 1) || (FNR > 1)' * > output.csv
```

usecase\_1\_output/

Copy S3 URI

Objects

Properties

To enable sorting in the table below, use the search to reduce the size of the list to 999 objects or fewer.

Objects (999+)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Refresh

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

< 1 ... 7 8 9 10 11 12 13 >

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	run-1642434789836-part-r-00028	-	January 17, 2022, 21:04:03 (UTC+02:00)	435.5 KB	Standard
<input type="checkbox"/>	run-1642434789836-part-r-00029	-	January 17, 2022, 21:04:03 (UTC+02:00)	418.0 KB	Standard
<input type="checkbox"/>	run-1642434789836-part-r-00030	-	January 17, 2022, 21:04:03 (UTC+02:00)	432.1 KB	Standard
<input type="checkbox"/>	run-1642434789836-part-r-00031	-	January 17, 2022, 21:04:03 (UTC+02:00)	428.0 KB	Standard
<input type="checkbox"/>	run-1642434789836-part-r-00032	-	January 17, 2022, 21:04:03 (UTC+02:00)	431.0 KB	Standard
<input type="checkbox"/>	run-1642434789836-part-r-00033	-	January 17, 2022, 21:04:03 (UTC+02:00)	428.1 KB	Standard
<input type="checkbox"/>	run-1642434789836-part-r-00034	-	January 17, 2022, 21:04:03 (UTC+02:00)	437.3 KB	Standard

© 2022, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie prefer](#)

id	title	author	venue	year	source	match
conf_vldb_RusinkiewiczKTWM95	Towards a Cooperative Transaction I	M Rusinkiewicz, W Klas, T Tes	VLDB	1995	dataset1	0
wgK6p4mDSIMJ	SQL/XML is Making Good Progress	A Eisenberg, J Melton	SIGMOD Record,	2002	dataset2	1
journals_sigmod_EisenbergM02	SQL/XML is Making Good Progress	A Eisenberg, J Melton	SIGMOD Record	2002	dataset1	1
wgK6p4mDSIMJ	SQL/XML is Making Good Progress	A Eisenberg, J Melton	SIGMOD Record,	2002	dataset2	1
conf_vldb_AmmannJR95	Using Formal Methods to Reason ab	P Ammann, S Jajodia, I Ray	VLDB	1995	dataset1	2
2e-NAgqt-joJ	Applying Formal Methods to Seman	P Ammann, S Jajodia, I Ray	ACM Transactions on I	1997	dataset2	2
journals_tods_AmmannJR97	Applying Formal Methods to Seman	P Ammann, S Jajodia, I Ray	ACM Trans. Database	1997	dataset1	2
journals_sigmod_Liu02	Editor's Notes	L Liu	SIGMOD Record	2002	dataset1	3
journals_sigmod_Hammer02	Report on the ACM Fourth Internati	N/A	N/A	2002	dataset1	4
conf_vldb_FerrandinaMZFM95	Schema and Database Evolution in t	F Ferrandina, T Meyer, R Zica	VLDB	1995	dataset1	5
9rofzgQ6HtcJ	Schema and Database Evolution in t	F Ferrandina, T Meyer, R Zica	Proc. of the 21st IntÃ	1995	dataset2	5
conf_vldb_SubietaKL95	Procedures in Object-Oriented Que	K Subieta, Y Kambayashi, J Le	VLDB	1995	dataset1	6
conf_vldb_SubietaKL95	Procedures in Object-Oriented Que	K Subieta, Y Kambayashi, J Le	VLDB	1995	dataset1	6
gEFY87Ma0XUJ	Procedures in Object-Oriented Que	K Subieta, Y Kambayashi, J Le	PROCEEDINGS OF THE	1995	dataset2	6
hf84fEpX5agJ	Phoenix: Making Applications Robus	R Barga, D Lomet	Philadelphia, PA (June,		dataset2	7
journals_sigmod_BargaL02	Phoenix Project: Fault-Tolerant App	R Barga, D Lomet	SIGMOD Record	2002	dataset1	7
fY3kkkzBLw8J	Phoenix Project: Fault Tolerant App	D Lomet, R Barga	SIGMOD Record,		dataset2	7
conf_sigmod_BargaL99	Phoenix: Making Applications Robus	R Barga, D Lomet	SIGMOD Conference	1999	dataset1	7
QGxyK7bJQOMJ	Phoenix: Making Applications Robus	R Barga, DB Lomet			dataset2	7
QTzV3iNq2O8J	Phoenix Project: Fault-Tolerant App	R Barga, D Lomet	SIGMOD Record,	2002	dataset2	7