MCIT-AWS- PDSA- Intake2

Date : 11 – 04 -2022

Data Deduplication – Fuzzy Matching- Use Case 1

Presented by : Abeer Nada

Problem Description:

Use Case Description Input: 1- Neat Dataset1 with some records (~2600). 2- Dataset2 with some records (~64,000). Messy, duplicated. 3- Matching records (labels) (~350).

Task: Using ML, perform de-duplication for the datasets

Keywords:

Data Deduplication

Semantic matching

Record linkage

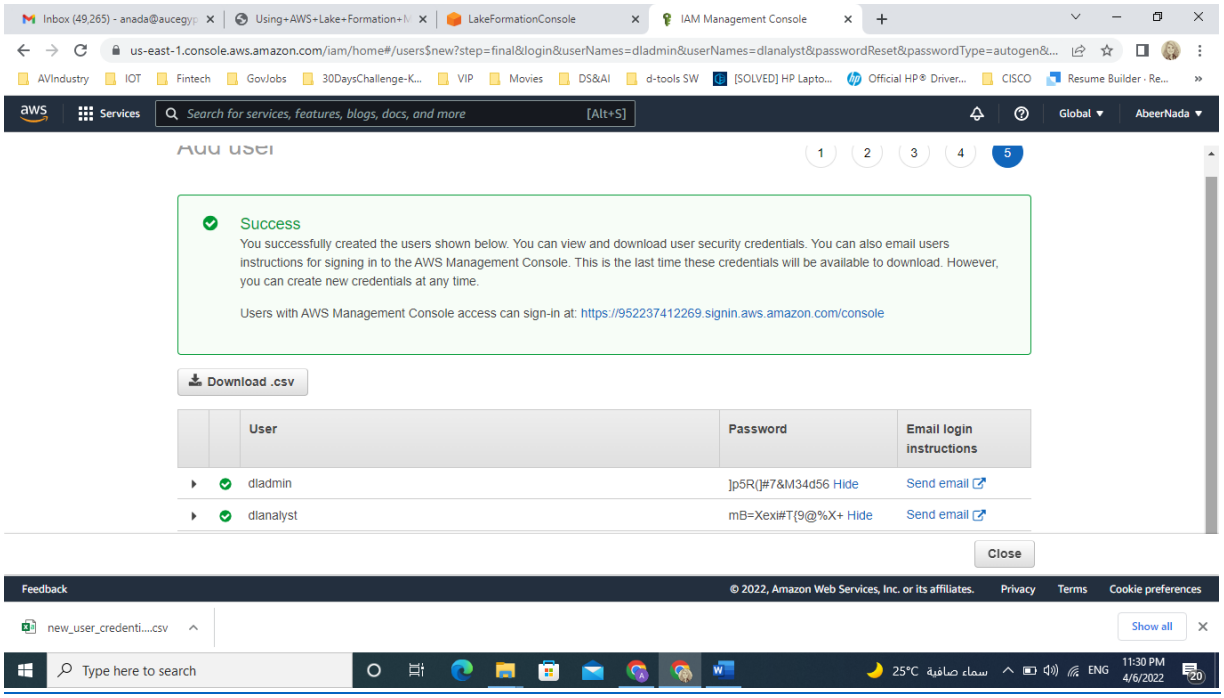
Entity resolution

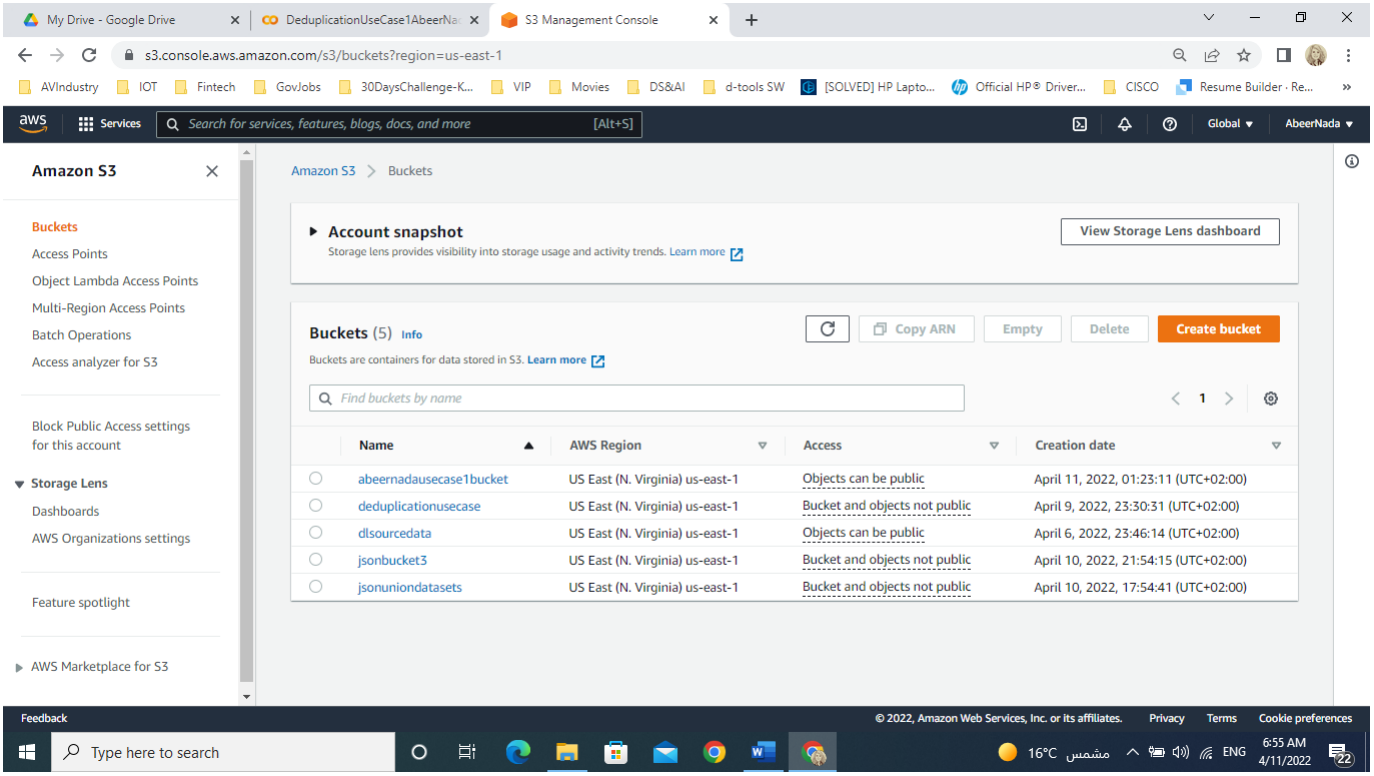
Approved Solving Methods:

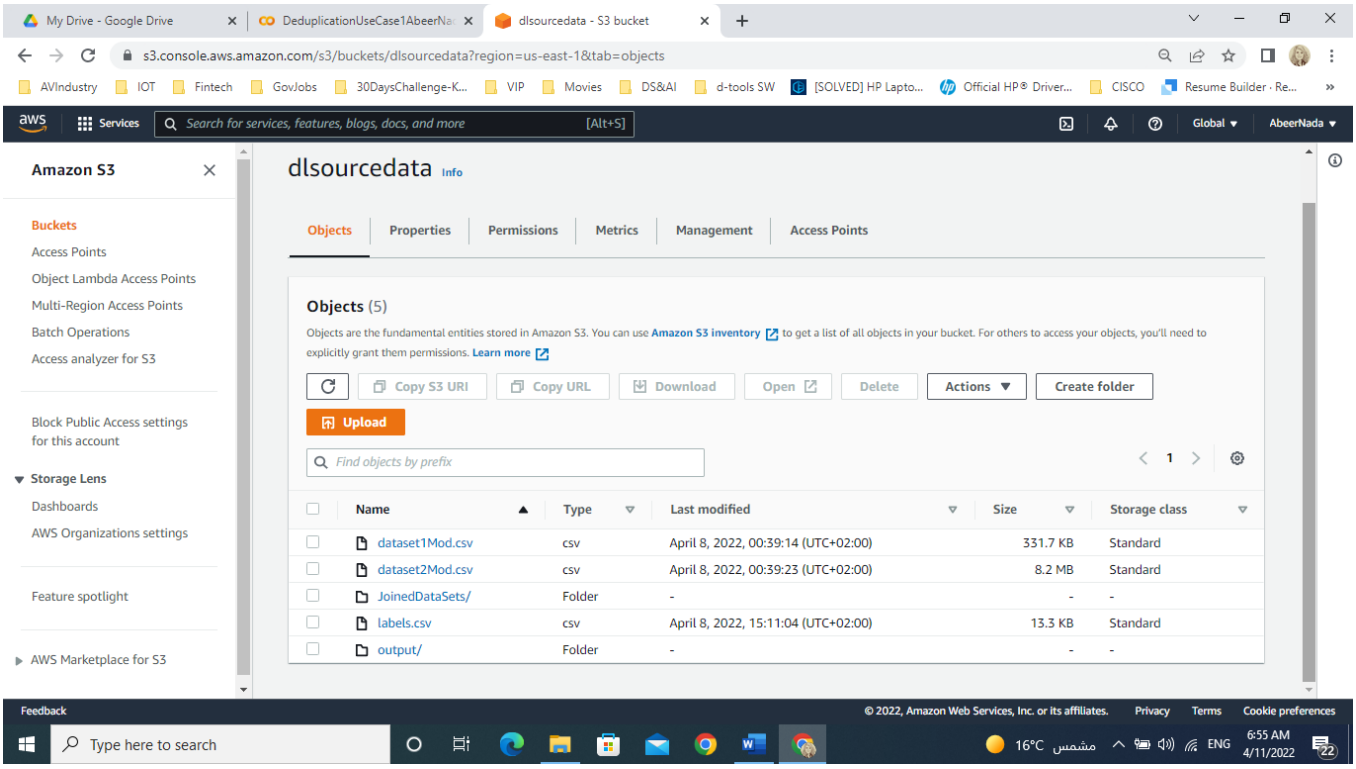
1- Build your own model – using supervised ML Algorithm for clustering
2- Fuzzy Score ML Model
3- Python Libraries with built-in ML Algorithms
4- Using AWS services

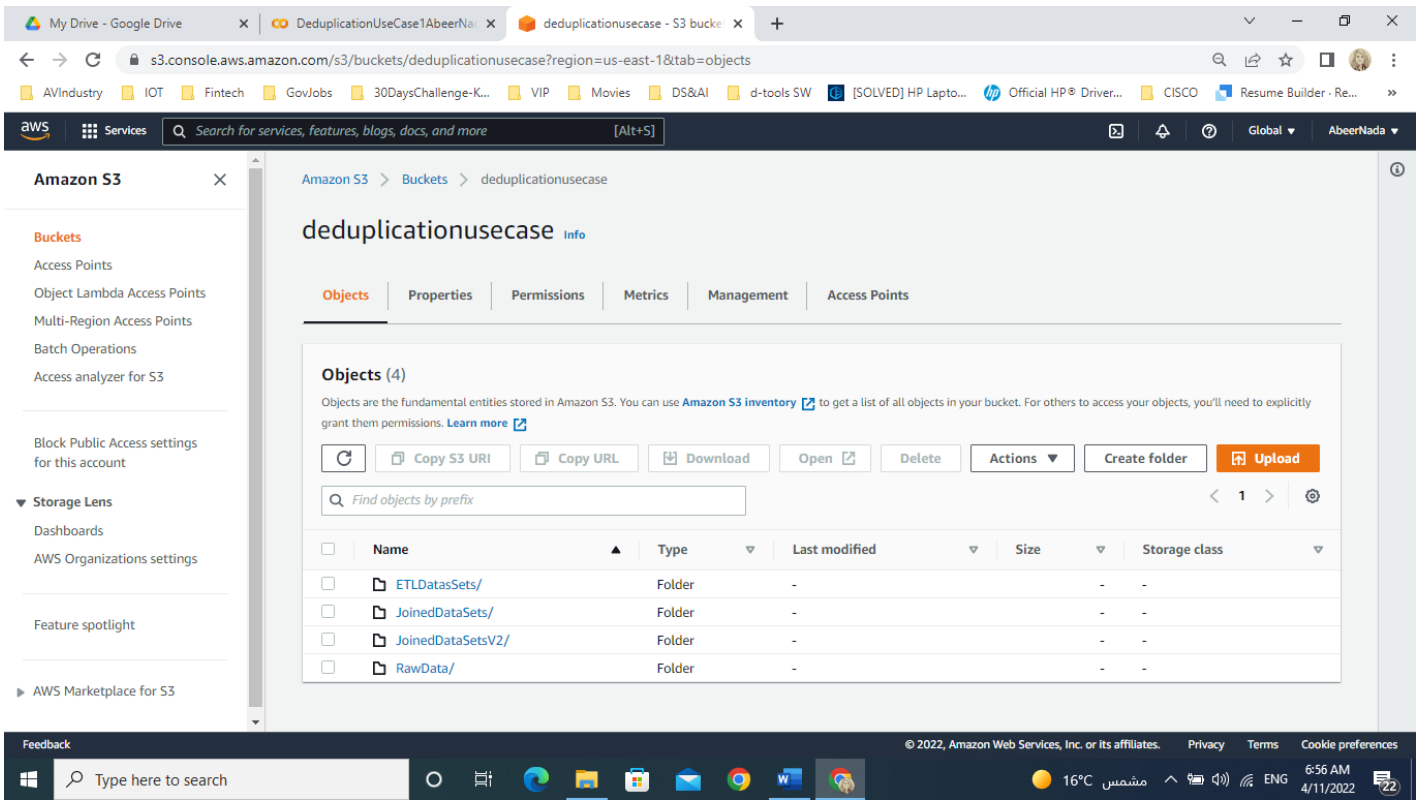
Chosen methods for the solution:

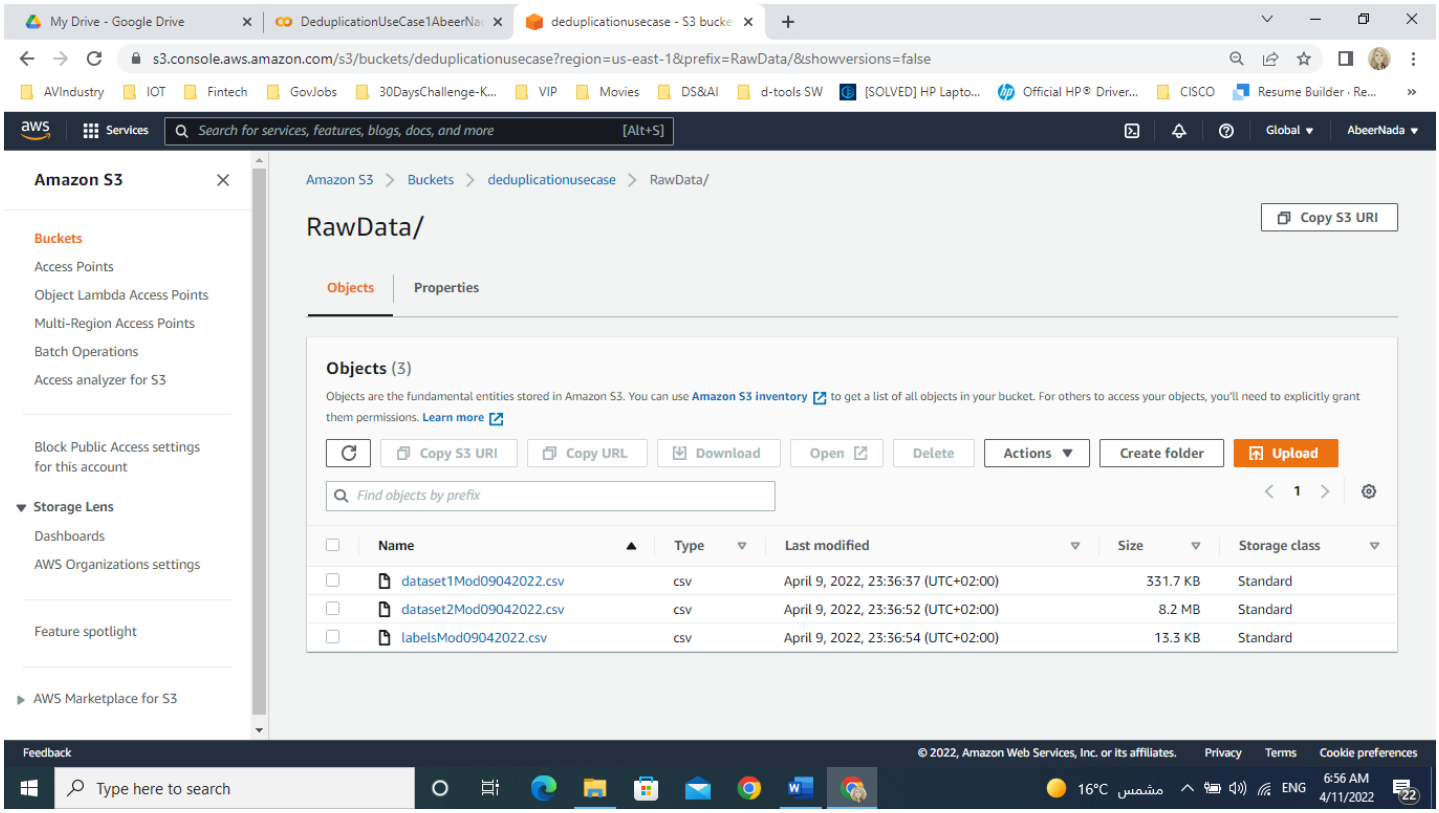
1- AWS for cleaning and transforming datasets using AWS S3, AWS Glue databrew. (A problem happened with the AWS crawler that prevent from completing the tasks on AWS
See blow screenshots:

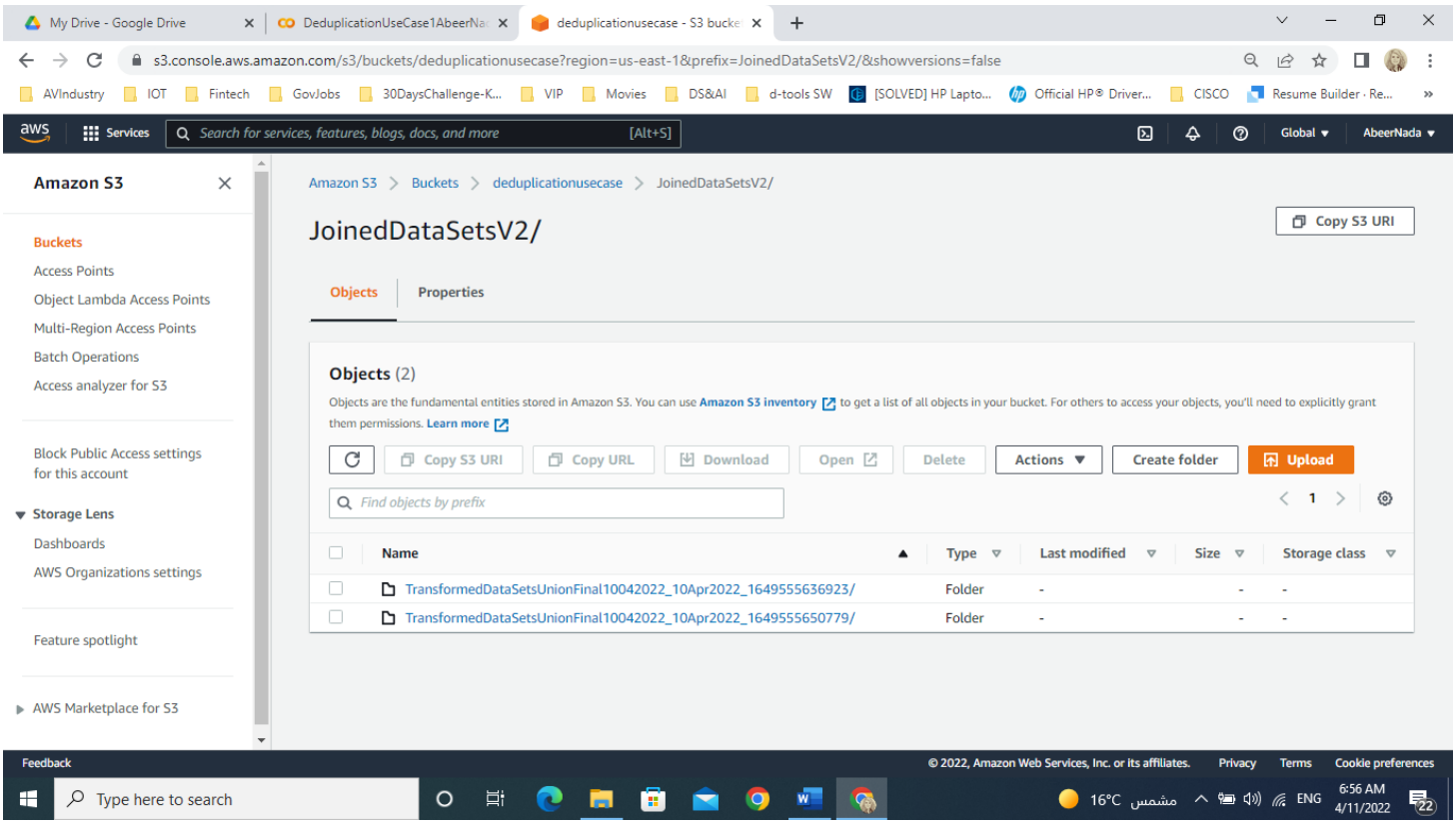

Add user
1 2 3 4 5
Success
You successfully created the users shown below. You can view and download user security credentials. You can also email users instructions for signing in to the AWS Management Console. This is the last time these credentials will be available to download. However, you can create new credentials at any time.
Users with AWS Management Console access can sign-in at: https://952237412269.signin.aws.amazon.com/console
Download .csv
User
Password
Email login instructions
dladmin
]p5R(]#7&M34d56 Hide
Send email
dlanalyst
mB=Xexi#T{9@%X+ Hide
Send email
Close
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
new_user_credenti....csv
Show all
Type here to search
25°C
ENG
11:30 PM
4/6/2022

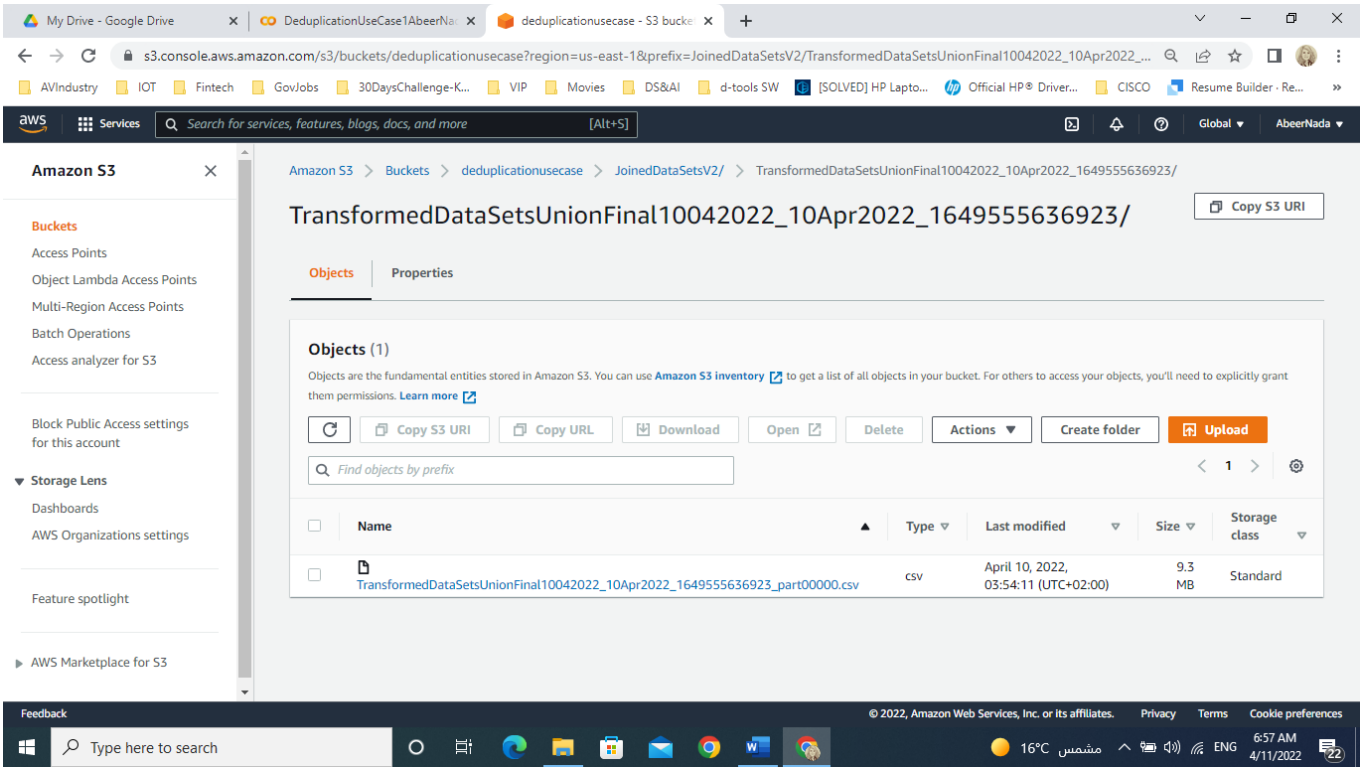

Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets
Account snapshot
Storage lens provides visibility into storage usage and activity trends. Learn more
View Storage Lens dashboard
Buckets (5) Info
Buckets are containers for data stored in S3. Learn more
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
Name
AWS Region
Access
Creation date
abeernadausecase1bucket
US East (N. Virginia) us-east-1
Objects can be public
April 11, 2022, 01:23:11 (UTC+02:00)
deduplicationusecase
US East (N. Virginia) us-east-1
Bucket and objects not public
April 9, 2022, 23:30:31 (UTC+02:00)
dlsourcedata
US East (N. Virginia) us-east-1
Objects can be public
April 6, 2022, 23:46:14 (UTC+02:00)
jsonbucket3
US East (N. Virginia) us-east-1
Bucket and objects not public
April 10, 2022, 21:54:15 (UTC+02:00)
jsonuniondatasets
US East (N. Virginia) us-east-1
Bucket and objects not public
April 10, 2022, 17:54:41 (UTC+02:00)
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
16°C
ENG
6:55 AM
4/11/2022

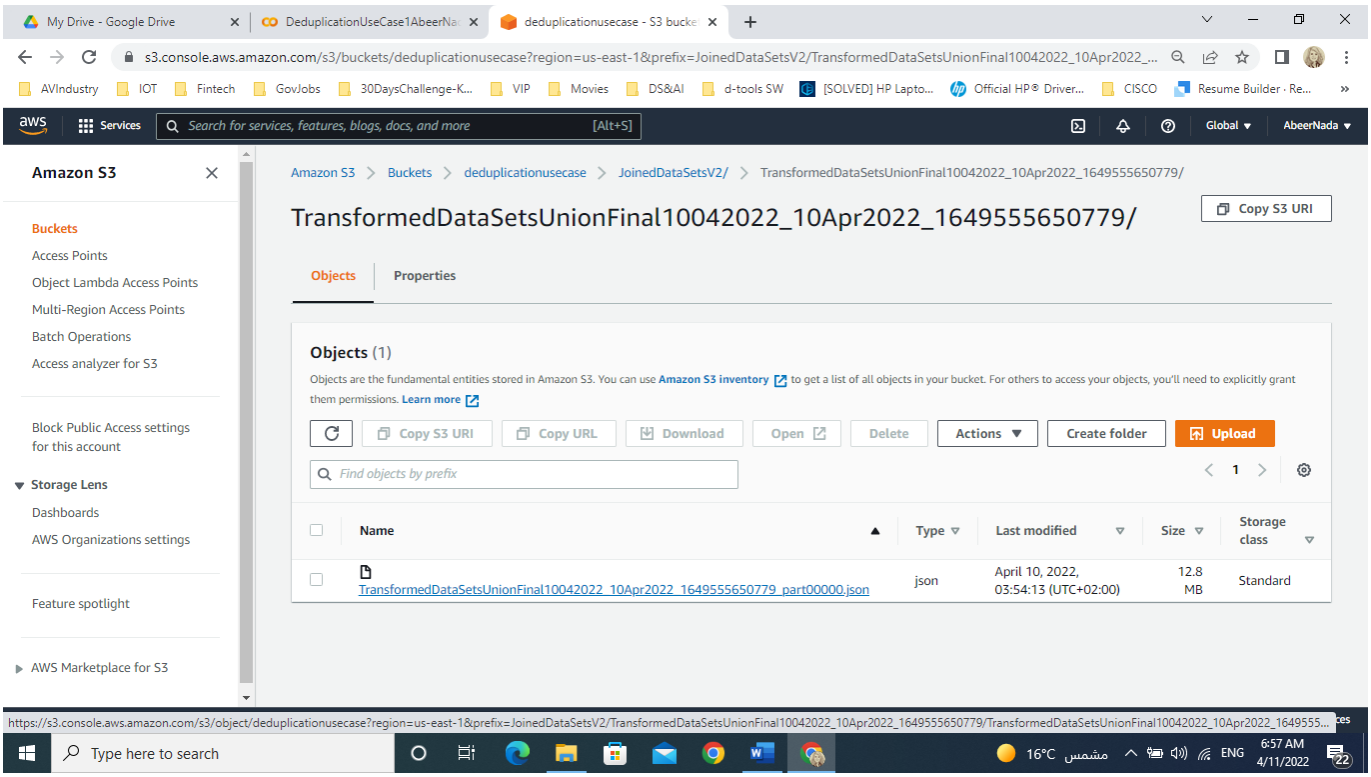

My Drive - Google Drive
DeduplicationUseCase1AbeerNa
dlsourcedata - S3 bucket
s3.console.aws.amazon.com/s3/buckets/dlsourcedata?region=us-east-1&tab=objects
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
dlsourcedata 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
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
dataset1Mod.csv csv April 8, 2022, 00:39:14 (UTC+02:00) 331.7 KB Standard
dataset2Mod.csv csv April 8, 2022, 00:39:23 (UTC+02:00) 8.2 MB Standard
JoinedDataSets/ Folder - - -
labels.csv csv April 8, 2022, 15:11:04 (UTC+02:00) 13.3 KB Standard
output/ Folder - - -
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
16°C
ENG
6:55 AM
4/11/2022


My Drive - Google Drive
DeduplicationUseCase1AbeerNa
deduplicationusecase - S3 bucke
s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&tab=objects
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets > deduplicationusecase
deduplicationusecase Info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (4)
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
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
ETLDatasSets/ Folder - - -
JoinedDataSets/ Folder - - -
JoinedDataSetsV2/ Folder - - -
RawData/ Folder - - -
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
16°C
ENG
6:56 AM
4/11/2022

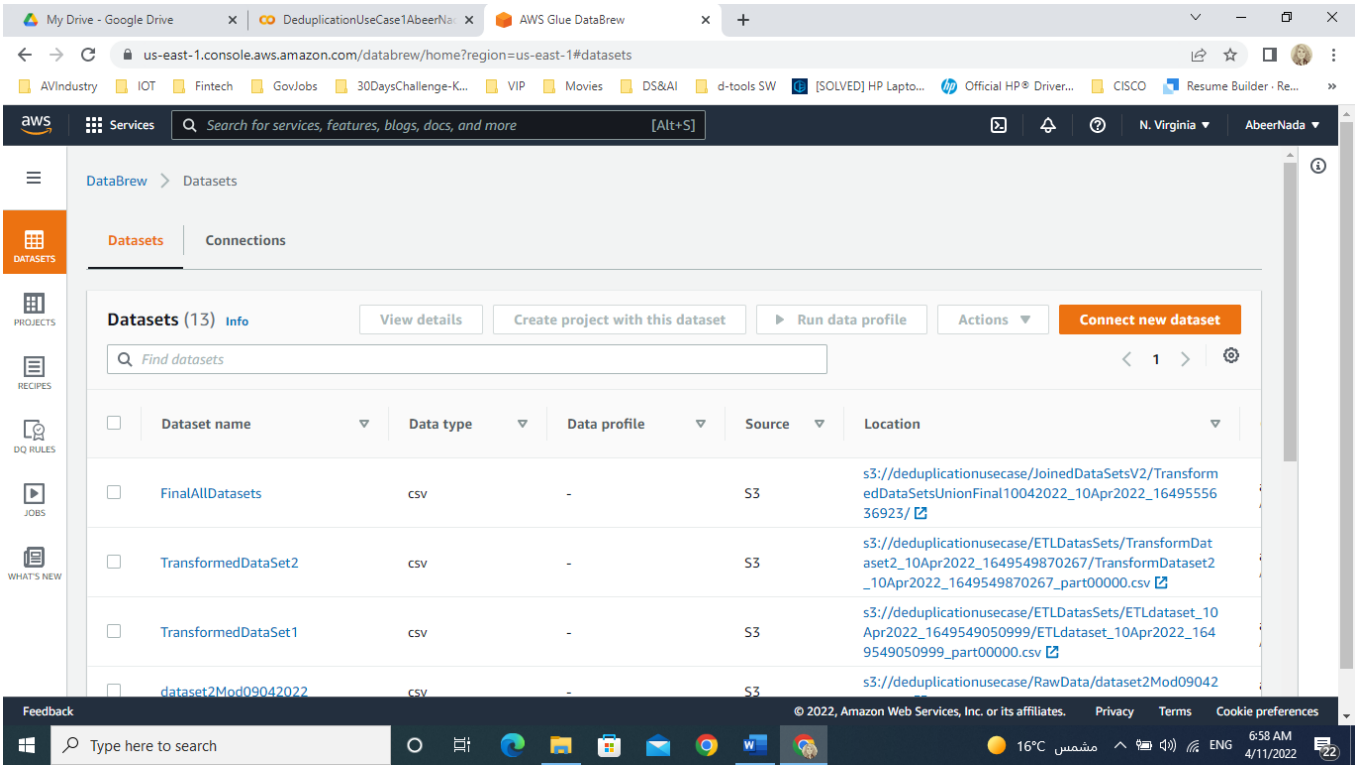

My Drive - Google Drive
DeduplicationUseCase1AbeerNa
deduplicationusecase - S3 bucke
s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&prefix=RawData/&showversions=false
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
AbeerNada
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets > deduplicationusecase > RawData/
RawData/
Copy S3 URI
Objects
Properties
Objects (3)
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
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
dataset1Mod09042022.csv
csv
April 9, 2022, 23:36:37 (UTC+02:00)
331.7 KB
Standard
dataset2Mod09042022.csv
csv
April 9, 2022, 23:36:52 (UTC+02:00)
8.2 MB
Standard
labelsMod09042022.csv
csv
April 9, 2022, 23:36:54 (UTC+02:00)
13.3 KB
Standard
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
16°C
ENG
6:56 AM
4/11/2022

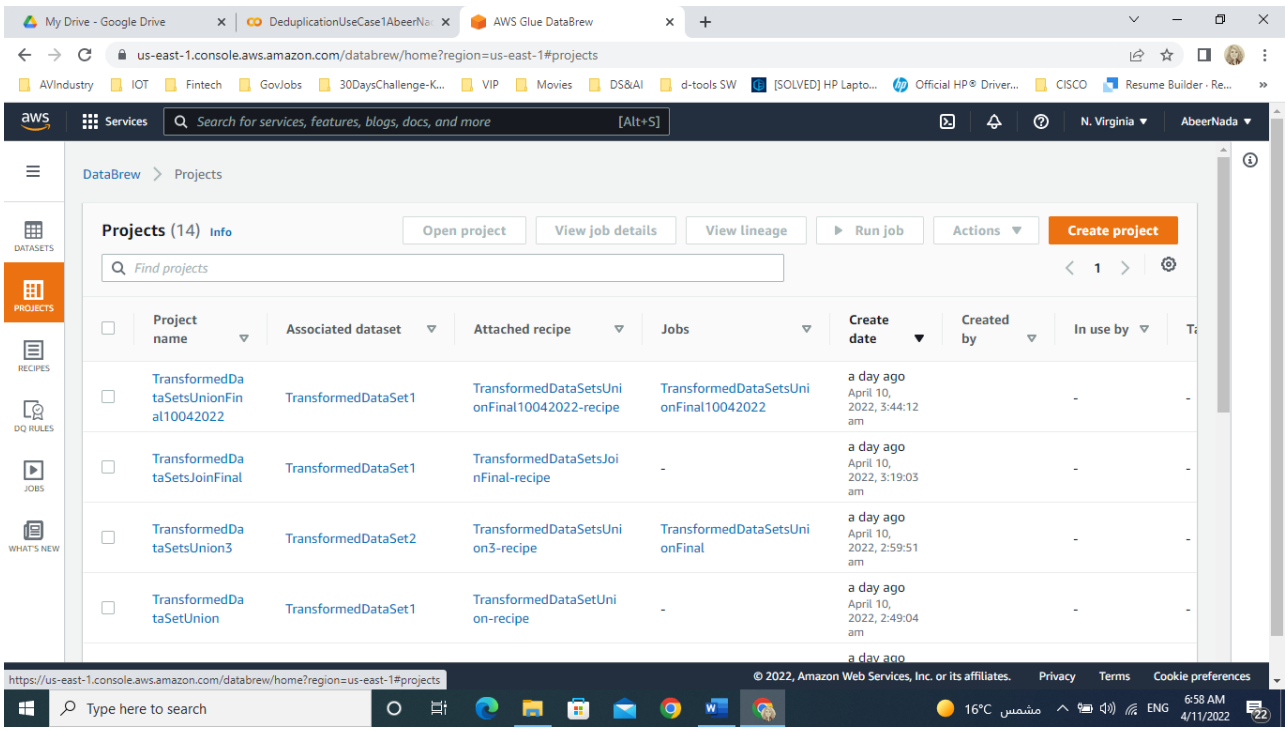

My Drive - Google Drive
DeduplicationUseCase1AbeerNa
deduplicationusecase - S3 bucke
s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&prefix=JoinedDataSetsV2/&showversions=false
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
AbeerNada
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets > deduplicationusecase > JoinedDataSetsV2/
JoinedDataSetsV2/
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
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923/
Folder
-
-
-
TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/
Folder
-
-
-
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
16°C
ENG
6:56 AM
4/11/2022

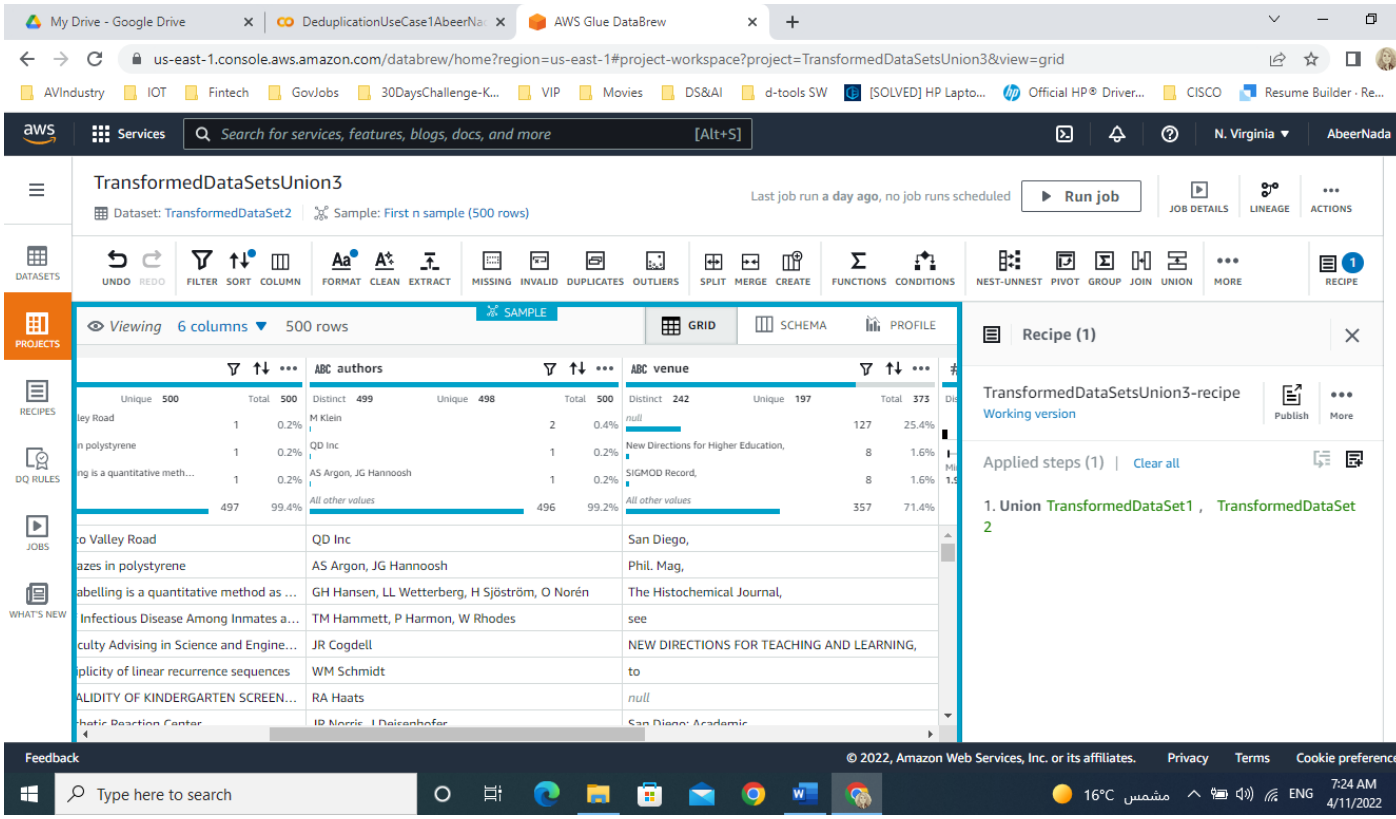

My Drive - Google Drive
DeduplicationUseCase1AbeerNa
deduplicationusecase - S3 bucke
s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&prefix=JoinedDataSetsV2/TransformedDataSetsUnionFinal10042022_10Apr2022_...
AVIndustry
IOT
Fintech
GovJobs
30DaysChallenge-K...
VIP
Movies
DS&AI
d-tools SW
[SOLVED] HP Lapto...
Official HP® Driver...
CISCO
Resume Builder - Re...
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
AbeerNada
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets > deduplicationusecase > JoinedDataSetsV2/ > TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923/
TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923/
Copy S3 URI
Objects
Properties
Objects (1)
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
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923_part00000.csv
csv
April 10, 2022, 03:54:11 (UTC+02:00)
9.3 MB
Standard
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
16°C
ENG
6:57 AM
4/11/2022

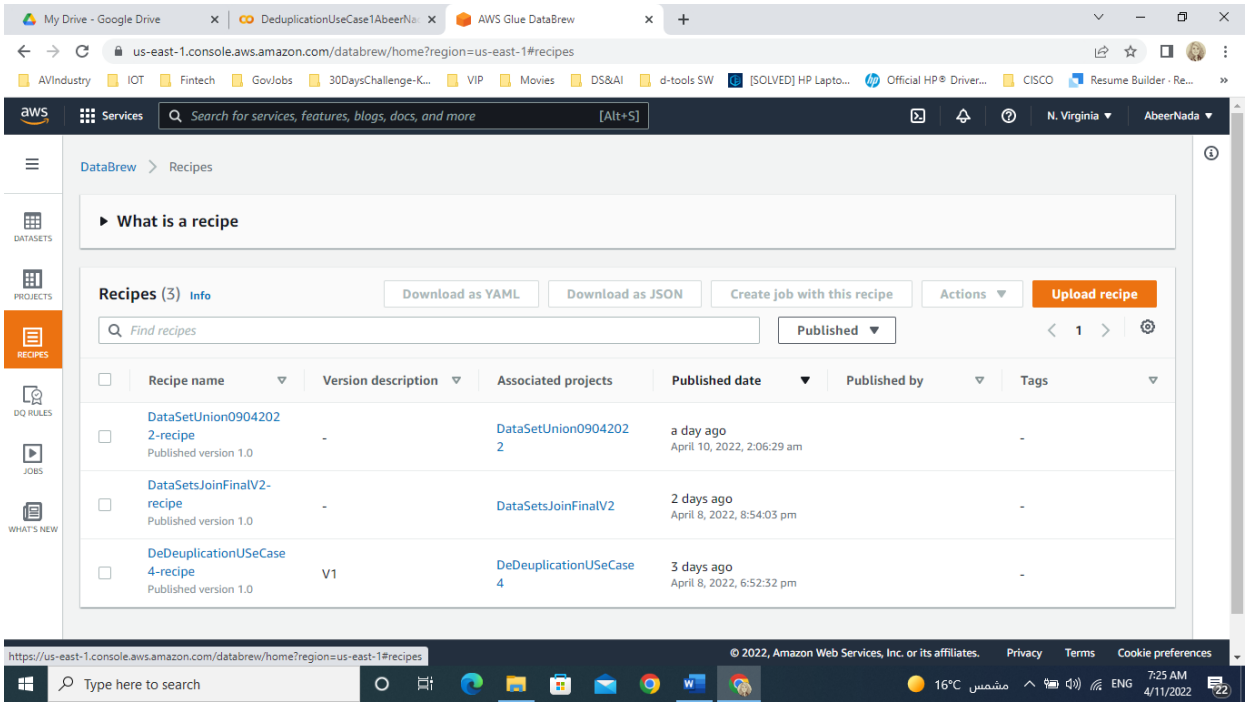

My Drive - Google Drive
DeduplicationUseCase1AbeerNa
deduplicationusecase - S3 bucke
s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&prefix=JoinedDataSetsV2/TransformedDataSetsUnionFinal10042022_10Apr2022_...
AVIndustry
IOT
Fintech
GovJobs
30DaysChallenge-K...
VIP
Movies
DS&AI
d-tools SW
[SOLVED] HP Lapto...
Official HP® Driver...
CISCO
Resume Builder - Re...
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
AbeerNada
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets > deduplicationusecase > JoinedDataSetsV2/ > TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/
TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/
Copy S3 URI
Objects
Properties
Objects (1)
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
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779_part00000.json
json
April 10, 2022, 03:54:13 (UTC+02:00)
12.8 MB
Standard
https://s3.console.aws.amazon.com/s3/object/deduplicationusecase?region=us-east-1&prefix=JoinedDataSetsV2/TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/TransformedDataSetsUnionFinal10042022_10Apr2022_1649555...
Type here to search
16°C
ENG
6:57 AM
4/11/2022

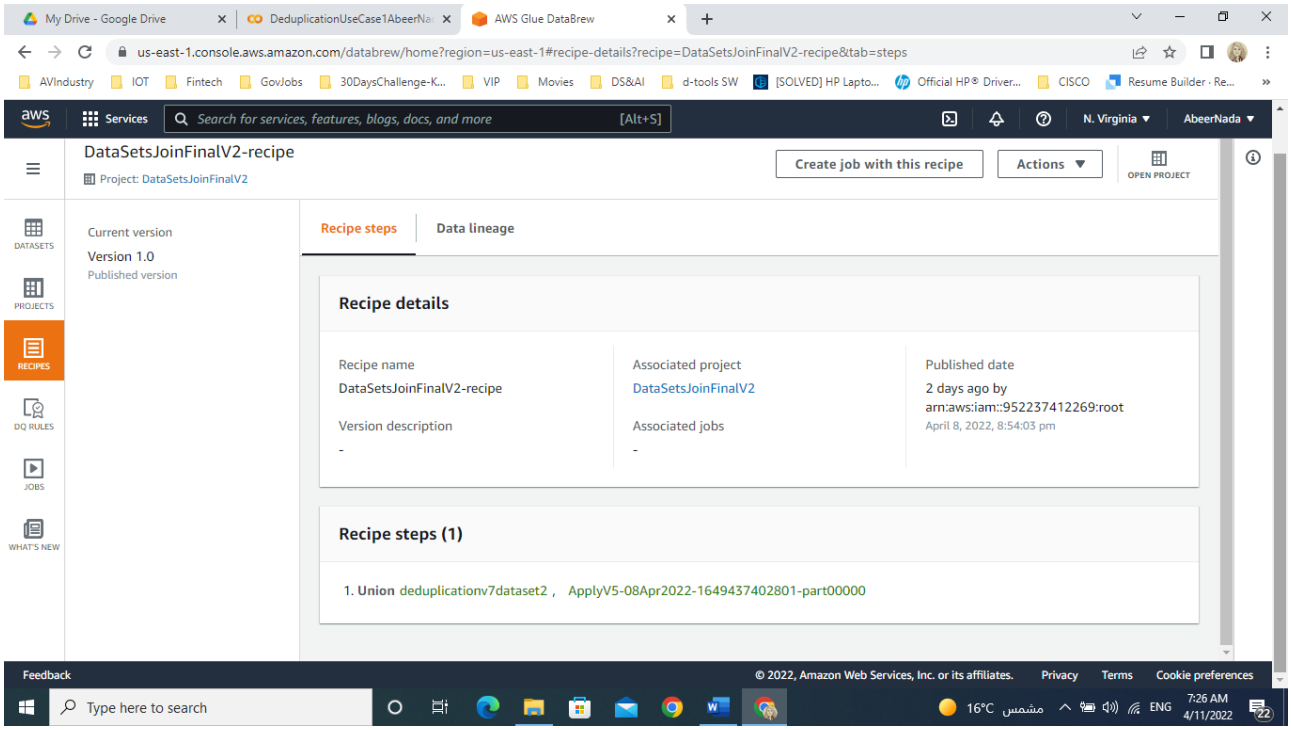
Glue Databrew:

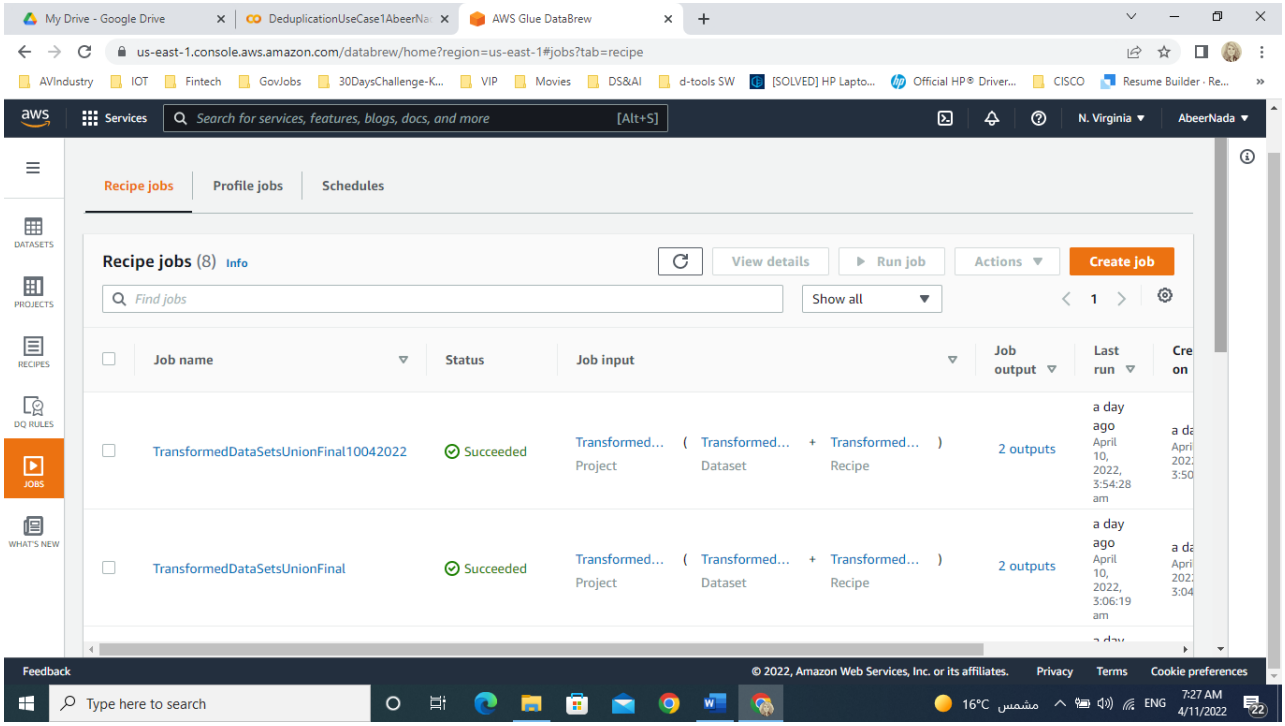





My Drive - Google Drive
DeduplicationUseCase1AbeerNa
AWS Glue DataBrew
us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#project-workspace?project=TransformedDataSetsUnion3&view=grid
Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia
AbeerNada
TransformedDataSetsUnion3
Dataset: TransformedDataSet2
Sample: First n sample (500 rows)
Last job run a day ago, no job runs scheduled
Run job
JOB DETAILS
LINEAGE
ACTIONS
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
UNDO
REDO
FILTER
SORT
COLUMN
FORMAT
CLEAN
EXTRACT
MISSING
INVALID
DUPLICATES
OUTLIERS
SPLIT
MERGE
CREATE
FUNCTIONS
CONDITIONS
NEST-UNNEST
PIVOT
GROUP
JOIN
UNION
MORE
RECIPE
Viewing 6 columns
500 rows
SAMPLE
GRID
SCHEMA
PROFILE
ABC authors
ABC venue
Recipe (1)
TransformedDataSetsUnion3-recipe
Working version
Publish
More
Applied steps (1)
Clear all
1. Union TransformedDataSet1 , TransformedDataSet 2
QD Inc
San Diego,
AS Argon, JG Hannoosh
Phil. Mag,
GH Hansen, LL Wetterberg, H Sjöström, O Norén
The Histochemical Journal,
TM Hammett, P Harmon, W Rhodes
see
JR Cogdell
NEW DIRECTIONS FOR TEACHING AND LEARNING,
WM Schmidt
to
RA Haats
null
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preference
Type here to search
16°C
ENG
7:24 AM
4/11/2022

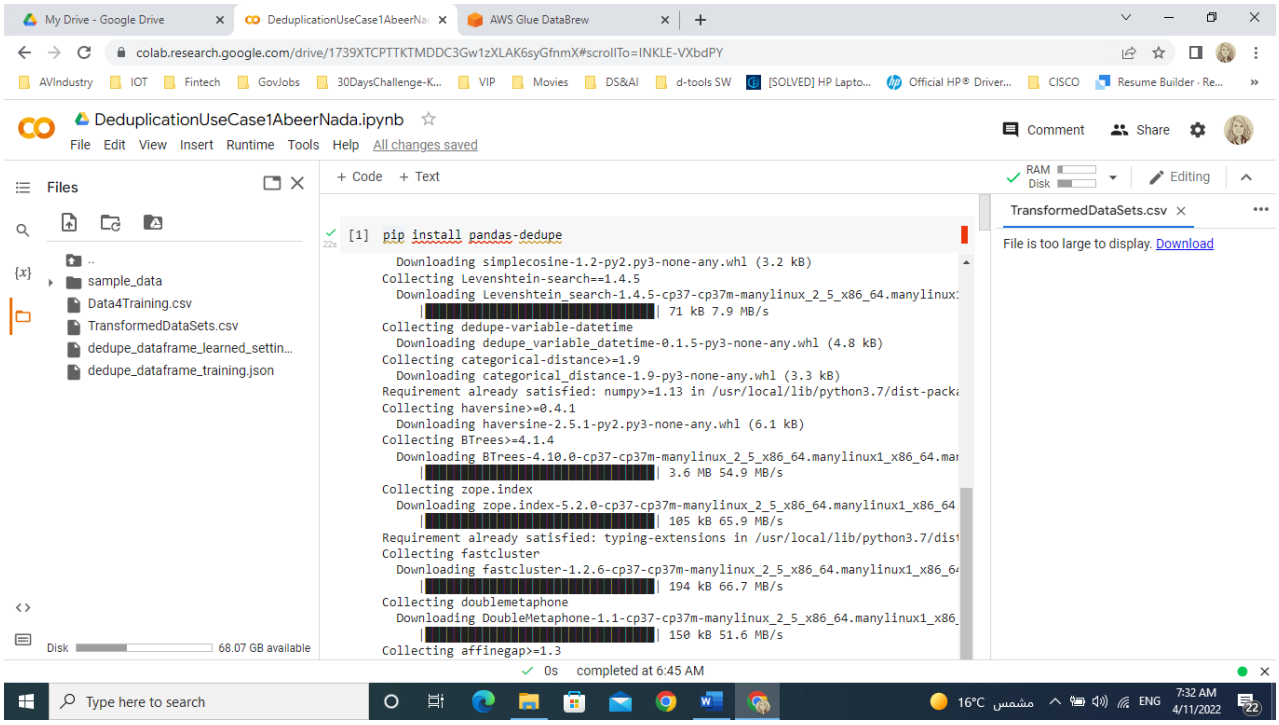

My Drive - Google Drive
DeduplicationUseCase1AbeerNa
AWS Glue DataBrew
us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#recipes
Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia
AbeerNada
DataBrew
Recipes
What is a recipe
Recipes (3) Info
Download as YAML
Download as JSON
Create job with this recipe
Actions
Upload recipe
Find recipes
Published
Recipe name
Version description
Associated projects
Published date
Published by
Tags
DataSetUnion09042022-recipe
Published version 1.0
DataSetUnion09042022
a day ago
April 10, 2022, 2:06:29 am
DataSetsJoinFinalV2-recipe
Published version 1.0
DataSetsJoinFinalV2
2 days ago
April 8, 2022, 8:54:03 pm
DeDeuplicationUSeCase4-recipe
Published version 1.0
V1
DeDeuplicationUSeCase4
3 days ago
April 8, 2022, 6:52:32 pm
https://us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#recipes
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
16°C
ENG
7:25 AM
4/11/2022





2- I generated one transformed csv and JSON file from the union of the two datasets (dataset1& dataset2) – please check attached files.

3- Using Dedupe Library in a python notebook to apply ML on the output dataset from the previous phase, please check attached notebook and file for the trained data.



References:

Using the Dedupe Machine Learning Library for Cleaning and Matching Data

https://www.azavea.com/blog/2019/08/30/using-the-dedupe-machine-learning-library-for-cleaning-and-matching-data/

Solving Data- Duplication problem using Machine Learning Algorithm

https://www.beyondkey.com/blog/solving-data-duplication-problem-using-machine-learning-algorithm/

https://stackoverflow.com/questions/16381133/using-machine-learning-to-de-duplicate-data

Why Machine Learning Is the Smart Way to Dedupe Salesforce

https://www.salesfix.com.au/blog/why-machine-learning-is-the-smart-way-to-dedupe-salesforce/

Dedupe - How it works

https://dedupe.io/documentation/how-it-works.html

Using Machine Learning to Detect Dupes: Some Real-Life Examples (Part II)

https://dzone.com/articles/using-machine-learning-to-detect-dupes-some-real-l