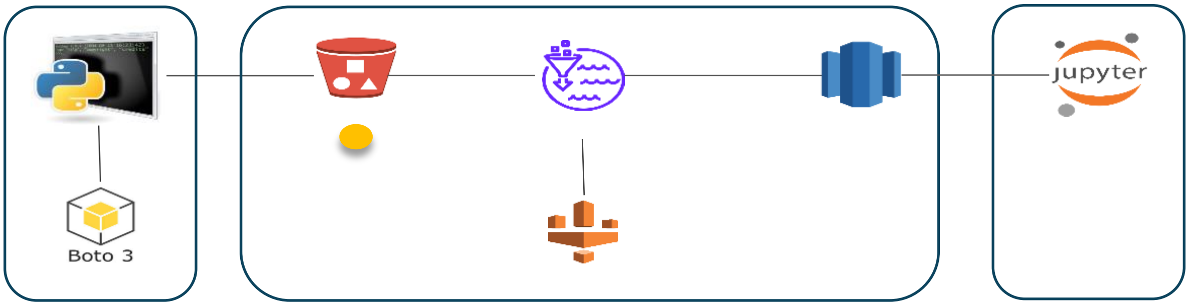
Table of Contents

Introduction

This project will be an analysis of Amazon Kindle book sales in 2023. The dataset is from Kaggle and can be found here. The scope of the analysis has 3 parts being Identifying the 10 most published authors based on their number of books published. Identifying the 10 most expensive books and the 10 most popular genres of book. Not in scope include total monetary value of sales and the number of copies a book has sold. While there is plenty that could be included in scope it will not be included to keep the length of the analysis manageable.

Tech Stack

The tech stack can be represented in the picture below first a python script and boto 3 are created to have a file be read into S3 storage. Then from S3 storage AWS Glue is used to transfer the data to a data lake formation. Using the data lake formation it is then transferred into a Redshift data warehouse. Finally, using the data warehouse it is then read into a Jupyter notebook. Where analysis using pandas can begin.

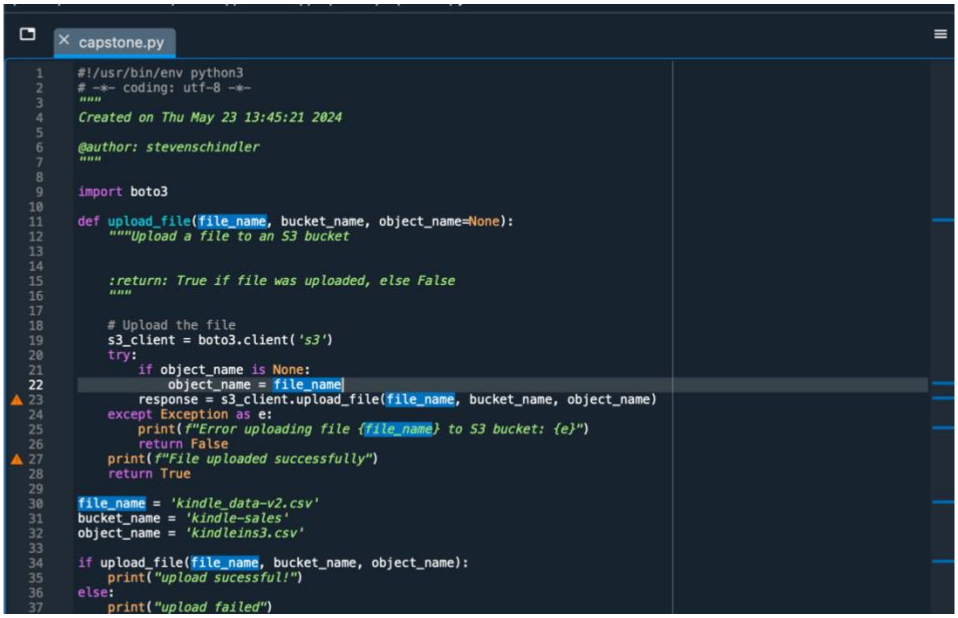


Part 1: Boto3

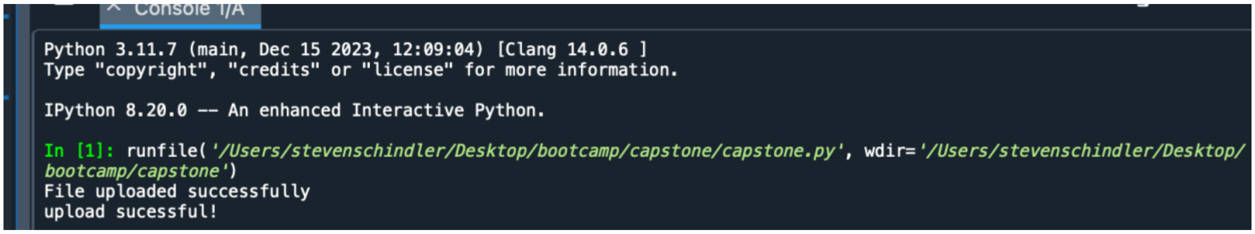
The first step is to write a Python script using boto3 to upload the file to the S3 storage bucket. Below is that script:

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Thu May 23 13:45:21 2024

@author: stevenschindler
"""


import boto3

def upload_file(file_name, bucket_name, object_name=None):
    """Upload a file to an S3 bucket



    :return: True if file was uploaded, else False
    """


    # Upload the file
    s3_client = boto3.client('s3')
    try:
        if object_name is None:
            object_name = file_name
        response = s3_client.upload_file(file_name, bucket_name, object_name)
    except Exception as e:
        print(f"Error uploading file {file_name} to S3 bucket: {e}")
        return False
    print(f"File uploaded successfully")
    return True

file_name = 'kindle_data-v2.csv'
bucket_name = 'kindlesales'
object_name = 'kindleins3.csv'

if upload_file(file_name, bucket_name, object_name):
    print("upload sucessful!")
else:
    print("upload failed")
```

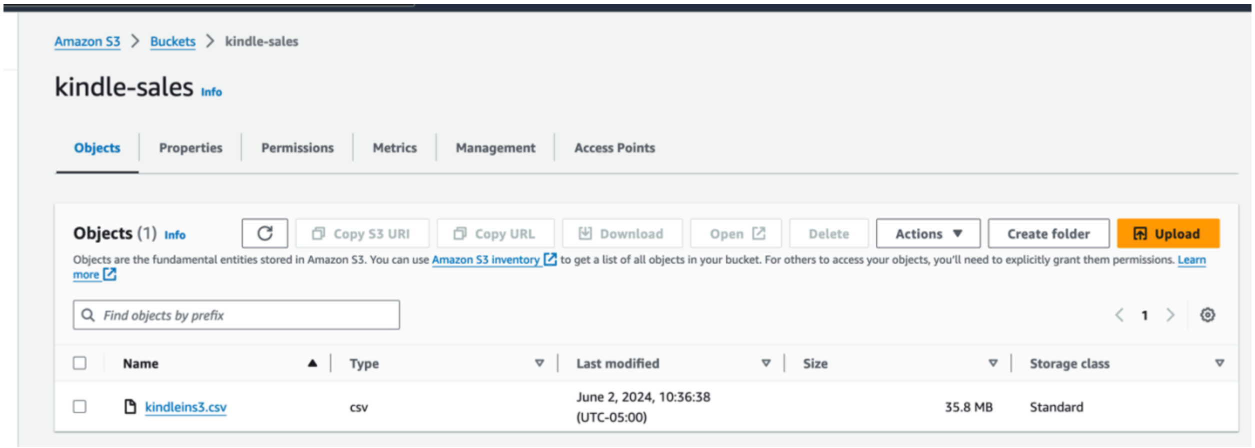
Next are screenshots showing the code and that it ran and uploaded successfully to an s3 bucket:



```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Thu May 23 13:45:21 2024

@author: stevenschindler
"""

import boto3

def upload_file(file_name, bucket_name, object_name=None):
    """Upload a file to an S3 bucket


    :return: True if file was uploaded, else False
    """

    # Upload the file
    s3_client = boto3.client('s3')
    try:
        if object_name is None:
            object_name = file_name
        response = s3_client.upload_file(file_name, bucket_name, object_name)
    except Exception as e:
        print(f"Error uploading file {file_name} to S3 bucket: {e}")
        return False
    print(f"File uploaded successfully")
    return True

file_name = 'kindle_data-v2.csv'
bucket_name = 'kindle-sales'
object_name = 'kindleins3.csv'

if upload_file(file_name, bucket_name, object_name):
    print("upload sucessful!")
else:
    print("upload failed")
```



```
Python 3.11.7 (main, Dec 15 2023, 12:09:04) [Clang 14.0.6 ]
Type "copyright", "credits" or "license" for more information.

IPython 8.20.0 -- An enhanced Interactive Python.

In [1]: runfile('/Users/stevenschindler/Desktop/bootcamp/capstone/capstone.py', wdir='/Users/stevenschindler/Desktop/
bootcamp/capstone')
File uploaded successfully
upload sucessful!
```

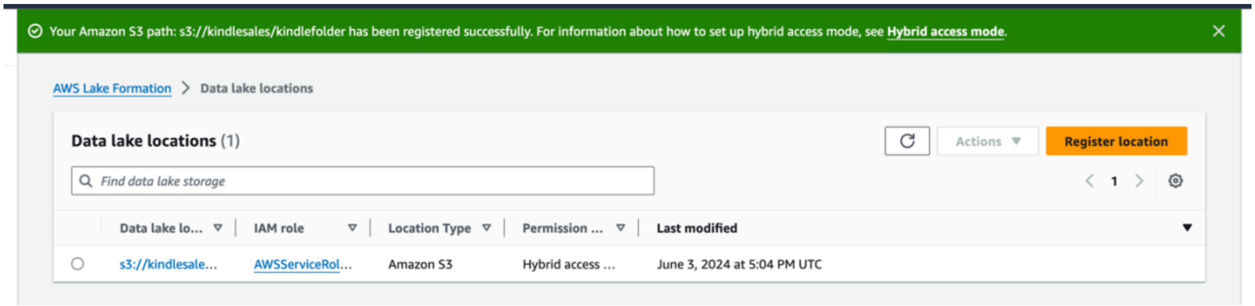


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	kindleins3.csv	csv	June 2, 2024, 10:36:38 (UTC-05:00)	35.8 MB	Standard

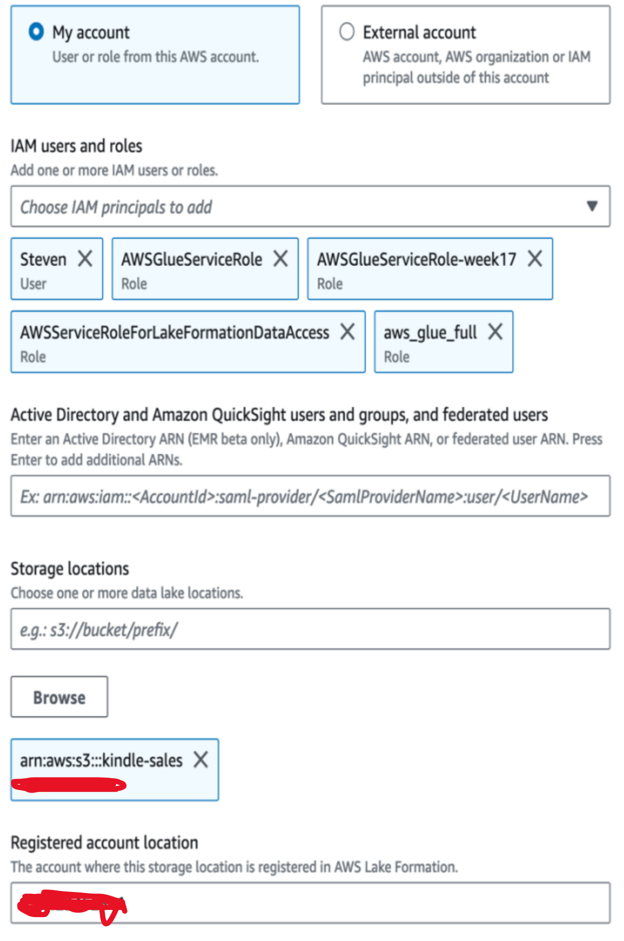
Part 2: AWS

Next is to create AWS Glue Crawler to transfer from the S3 bucket to the Data Lake. First permissions for the data lake need to be given as well as a location regristered.



Your Amazon S3 path: s3://kindlesales/kindlefolder has been registered successfully. For information about how to set up hybrid access mode, see Hybrid access mode.

	Data lake lo...	IAM role	Location Type	Permission ...	Last modified
○	s3://kindlesale...	AWSServiceRol...	Amazon S3	Hybrid access ...	June 3, 2024 at 5:04 PM UTC



● My account
User or role from this AWS account.

○ External account
AWS account, AWS organization or IAM principal outside of this account

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add ▼

Steven X User | AWSGlueServiceRole X Role | AWSGlueServiceRole-week17 X Role | AWSServiceRoleForLakeFormationDataAccess X Role | aws_glue_full X Role

Active Directory and Amazon QuickSight users and groups, and federated users
Enter an Active Directory ARN (EMR beta only), Amazon QuickSight ARN, or federated user ARN. Press Enter to add additional ARNs.

Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderName>:user/<UserName>

Storage locations
Choose one or more data lake locations.

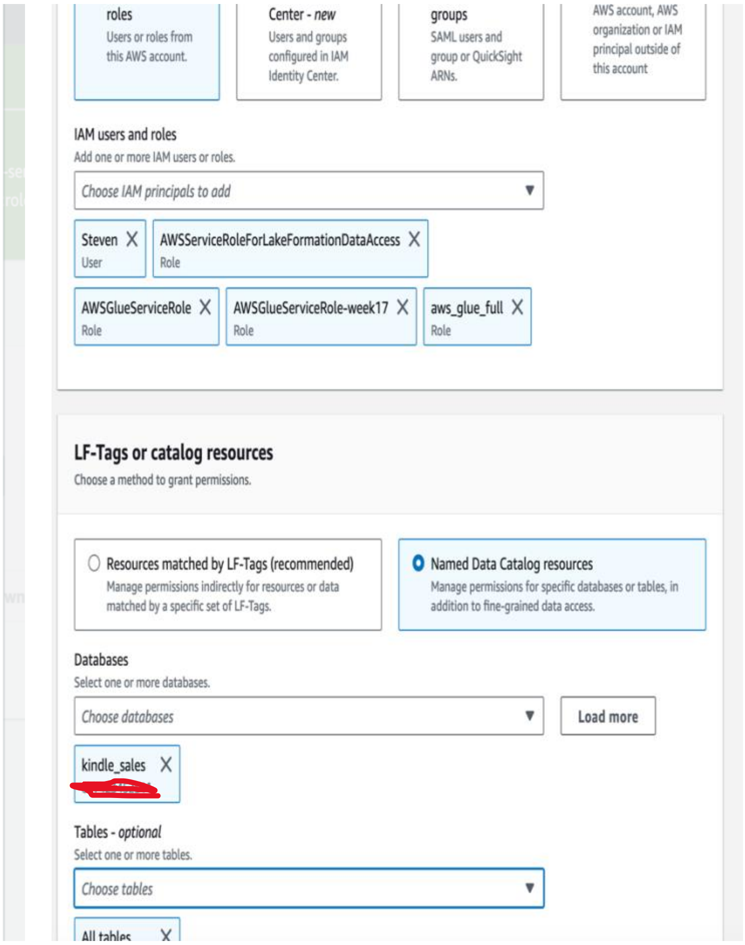
e.g.: s3://bucket/prefix/

Browse

arn:aws:s3:::kindle-sales X
[redacted]

Registered account location
The account where this storage location is registered in AWS Lake Formation.

[redacted]



roles
Users or roles from this AWS account.

Center - *new*
Users and groups configured in IAM Identity Center.

groups
SAML users and group or QuickSight ARNs.

AWS account, AWS organization or IAM principal outside of this account

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add ▼

Steven X User | AWSServiceRoleForLakeFormationDataAccess X Role | AWSGlueServiceRole X Role | AWSGlueServiceRole-week17 X Role | aws_glue_full X Role

LF-Tags or catalog resources
Choose a method to grant permissions.

○ Resources matched by LF-Tags (recommended)
Manage permissions indirectly for resources or data matched by a specific set of LF-Tags.

● Named Data Catalog resources
Manage permissions for specific databases or tables, in addition to fine-grained data access.

Databases
Select one or more databases.

Choose databases ▼ Load more

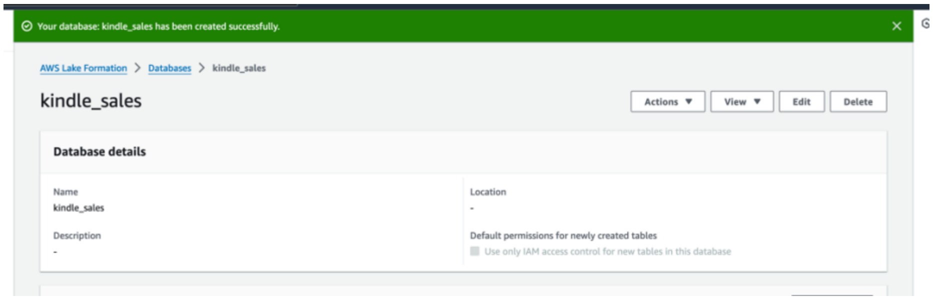
kindle_sales X
[redacted]

Tables - *optional*
Select one or more tables.

Choose tables ▼

All tables X

A database was also needed which was created before giving permissions on the right-hand side picture above.



Your database: kindle_sales has been created successfully.

AWS Lake Formation > Databases > kindle_sales

kindle_sales

Database details

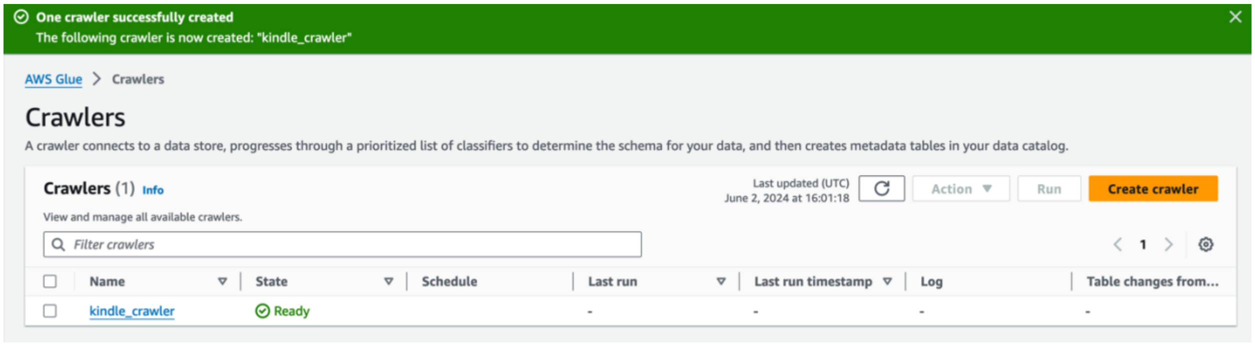
Name
kindle_sales

Description
-

Location
-

Default permissions for newly created tables
□ Use only IAM access control for new tables in this database

Now a Glue Crawler needs to be created and run to transfer data from the S3 bucket to the database kindle_sales.



One crawler successfully created
The following crawler is now created: "kindle_crawler"

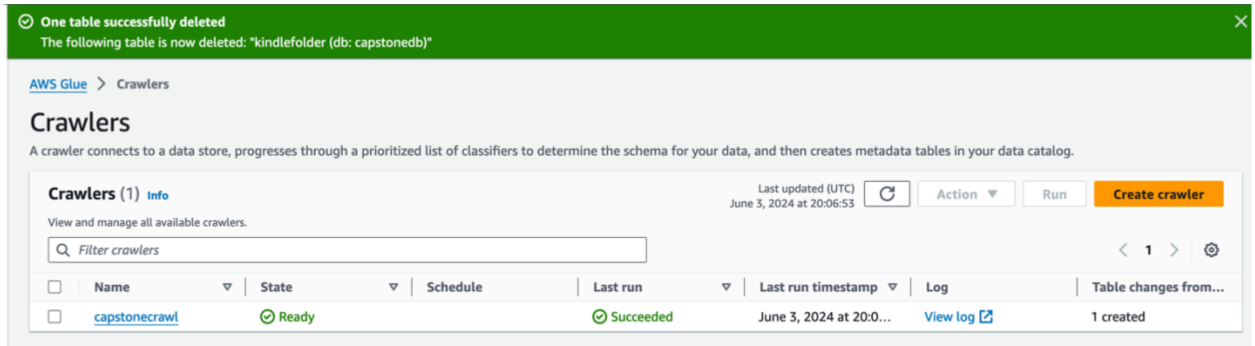
AWS Glue > Crawlers

Crawlers
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1) Info
View and manage all available crawlers.

Last updated (UTC)
June 2, 2024 at 16:01:18

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from...
kindle_crawler	Ready		-	-	-	-



One table successfully deleted
The following table is now deleted: "kindlefolder (db: capstonedb)"

AWS Glue > Crawlers

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

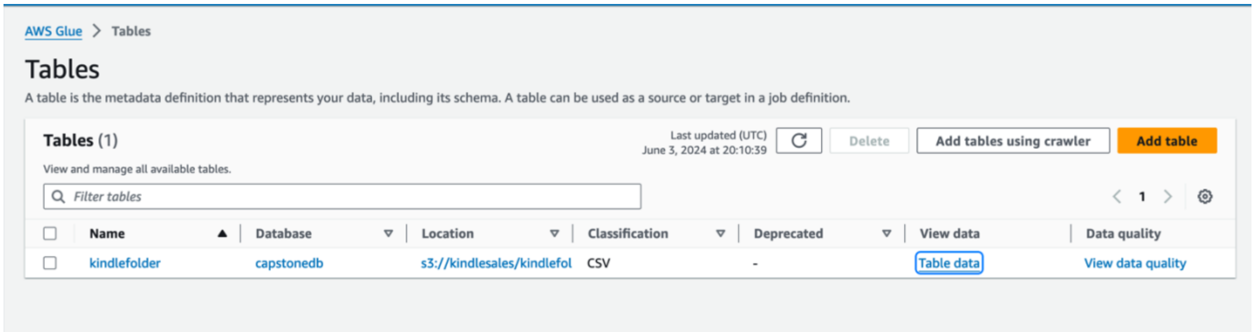
Crawlers (1) Info

View and manage all available crawlers.

Last updated (UTC)
June 3, 2024 at 20:06:53

□	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from...
□	capstonecrawl	✓ Ready		✓ Succeeded	June 3, 2024 at 20:0...	View log	1 created

The crawler is successfully created and successfully ran. Now we can see data in the database using AWS Athena. We can also see that a new table was created using the Crawler.



AWS Glue > Tables

Tables

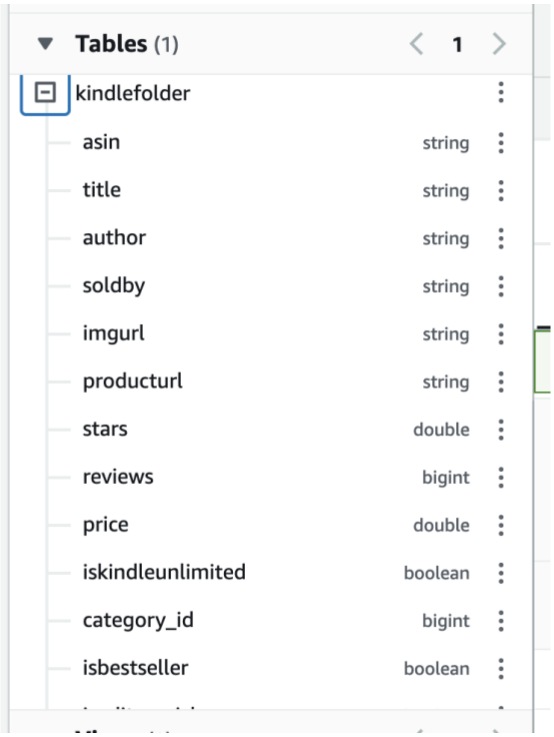
A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (1)

View and manage all available tables.

Last updated (UTC)
June 3, 2024 at 20:10:39

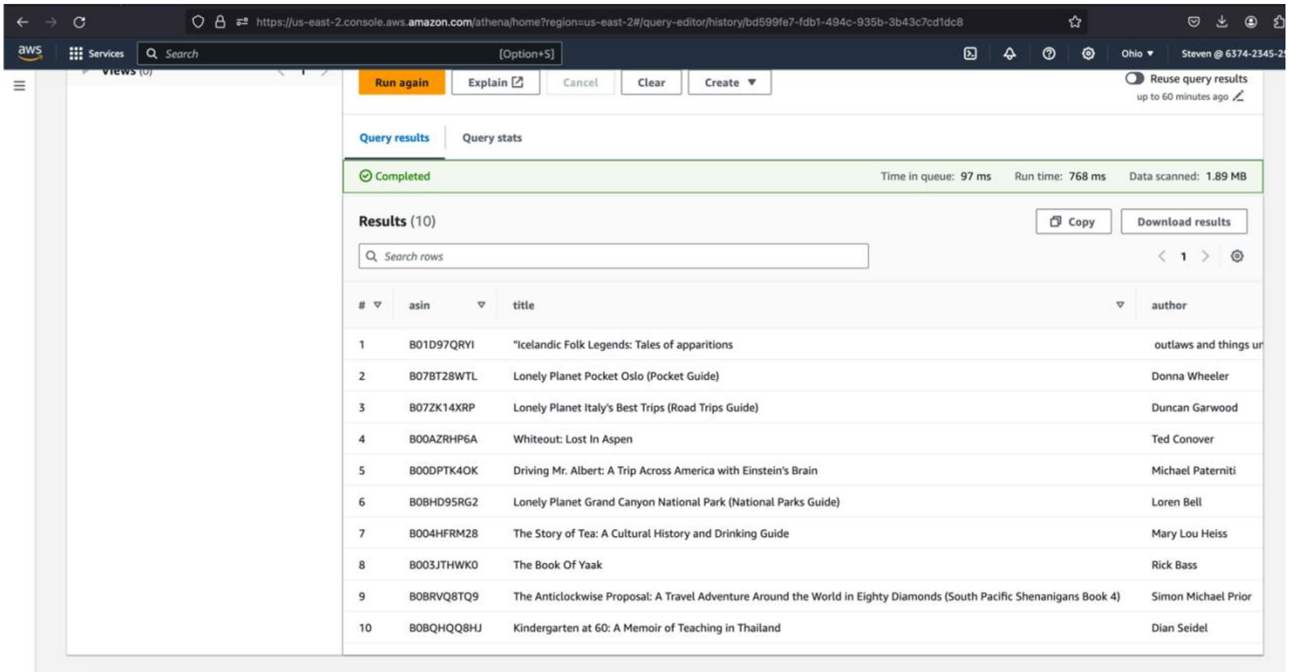
□	Name	Database	Location	Classification	Deprecated	View data	Data quality
□	kindlefolder	capstonedb	s3://kindlesales/kindlefol	CSV	-	Table data	View data quality



Tables (1)

kindlefolder

asin	string
title	string
author	string
soldby	string
imgurl	string
producturl	string
stars	double
reviews	bigint
price	double
iskindleunlimited	boolean
category_id	bigint
isbestseller	boolean



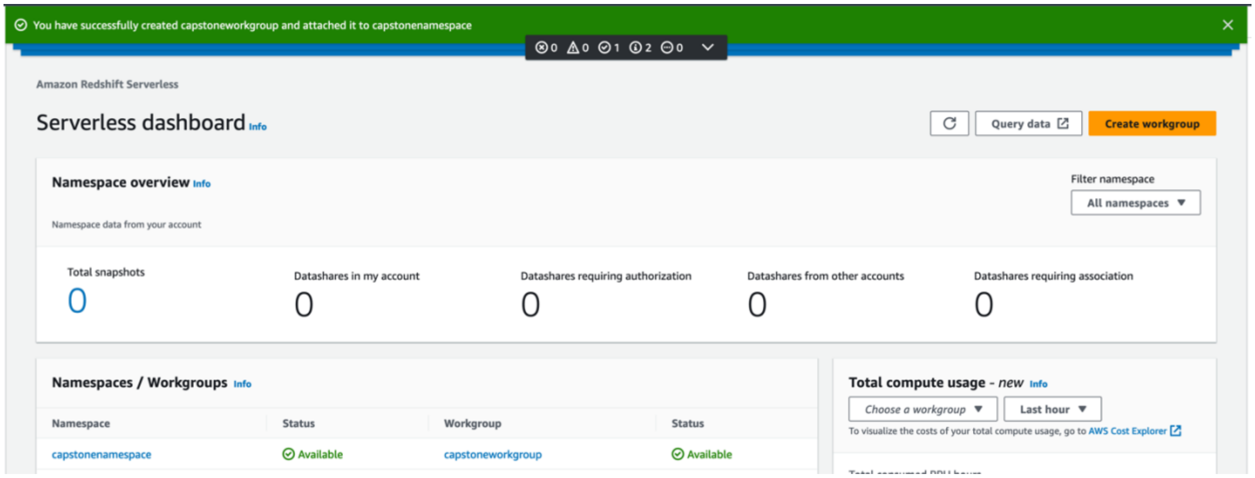
https://us-east-2.console.aws.amazon.com/athena/home?region=us-east-2#/query-editor/history/bd599fe7-fdb1-494c-935b-3b43c7cd1dc8

Completed Time in queue: 97 ms Run time: 768 ms Data scanned: 1.89 MB

Results (10)

#	asin	title	author
1	B01D97QRYI	"Icelandic Folk Legends: Tales of apparitions	outlaws and things ur
2	B07BT28WTL	Lonely Planet Pocket Oslo (Pocket Guide)	Donna Wheeler
3	B07ZK14XRP	Lonely Planet Italy's Best Trips (Road Trips Guide)	Duncan Garwood
4	B00AZRHP6A	Whiteout: Lost In Aspen	Ted Conover
5	B00DPTK4OK	Driving Mr. Albert: A Trip Across America with Einstein's Brain	Michael Paterniti
6	B0BHD95RG2	Lonely Planet Grand Canyon National Park (National Parks Guide)	Loren Bell
7	B004HFRM28	The Story of Tea: A Cultural History and Drinking Guide	Mary Lou Heiss
8	B003JTHWK0	The Book Of Yaak	Rick Bass
9	B0BRVQ8TQ9	The Anticlockwise Proposal: A Travel Adventure Around the World in Eighty Diamonds (South Pacific Shenanigans Book 4)	Simon Michael Prior
10	B0BQHQQ8HJ	Kindergarten at 60: A Memoir of Teaching in Thailand	Dian Seidel

Next is to transfer the data to the data warehouse using AWS Redshift. First a workspace and name group are created with Iam permissions given.



You have successfully created capstoneworkgroup and attached it to capstonenamespace

Serverless dashboard

Namespace overview

Namespace data from your account

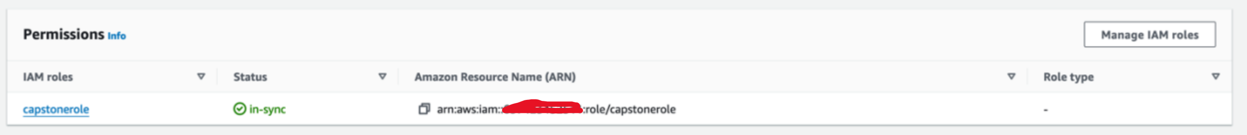
Total snapshots	Datashares in my account	Datashares requiring authorization	Datashares from other accounts	Datashares requiring association
0	0	0	0	0

Namespaces / Workgroups

Namespace	Status	Workgroup	Status
capstonenamespace	Available	capstoneworkgroup	Available

Total compute usage - *new*

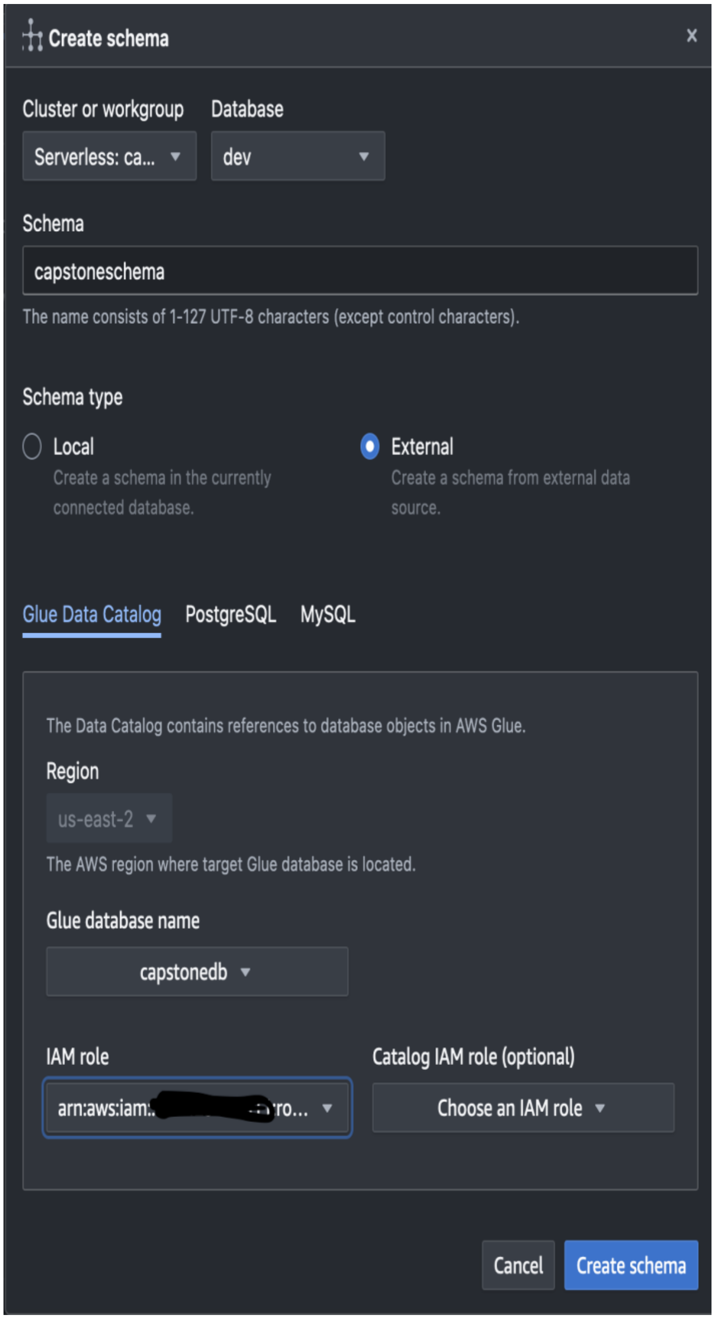
To visualize the costs of your total compute usage, go to AWS Cost Explorer



Permissions

IAM roles	Status	Amazon Resource Name (ARN)	Role type
capstonerole	in-sync	arn:aws:iam:[redacted]:role/capstonerole	-

Next is to create a schema and connect to and read from AWS Athena into the AWS Redshift data warehouse.



Cluster or workgroup: Serverless: ca...

Database: dev

Schema: capstoneschema

The name consists of 1-127 UTF-8 characters (except control characters).

Schema type

○ Local
Create a schema in the currently connected database.

● External
Create a schema from external data source.

The Data Catalog contains references to database objects in AWS Glue.

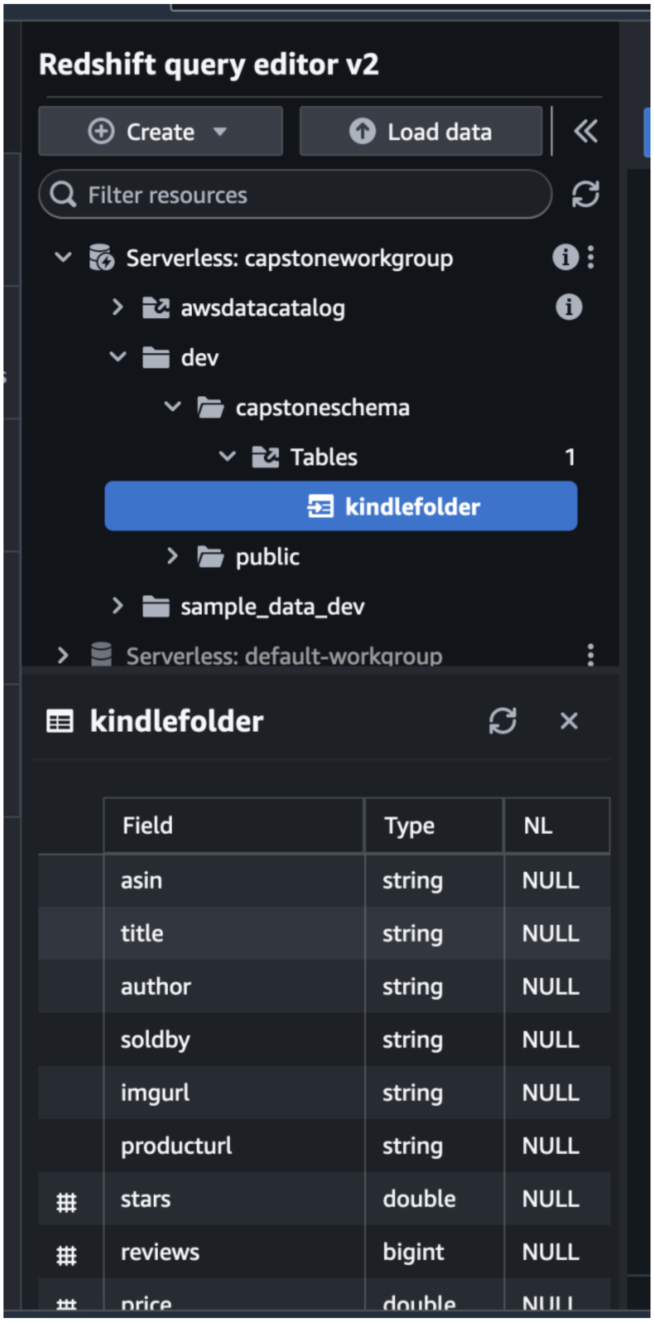
Region: us-east-2

The AWS region where target Glue database is located.

Glue database name: capstonedb

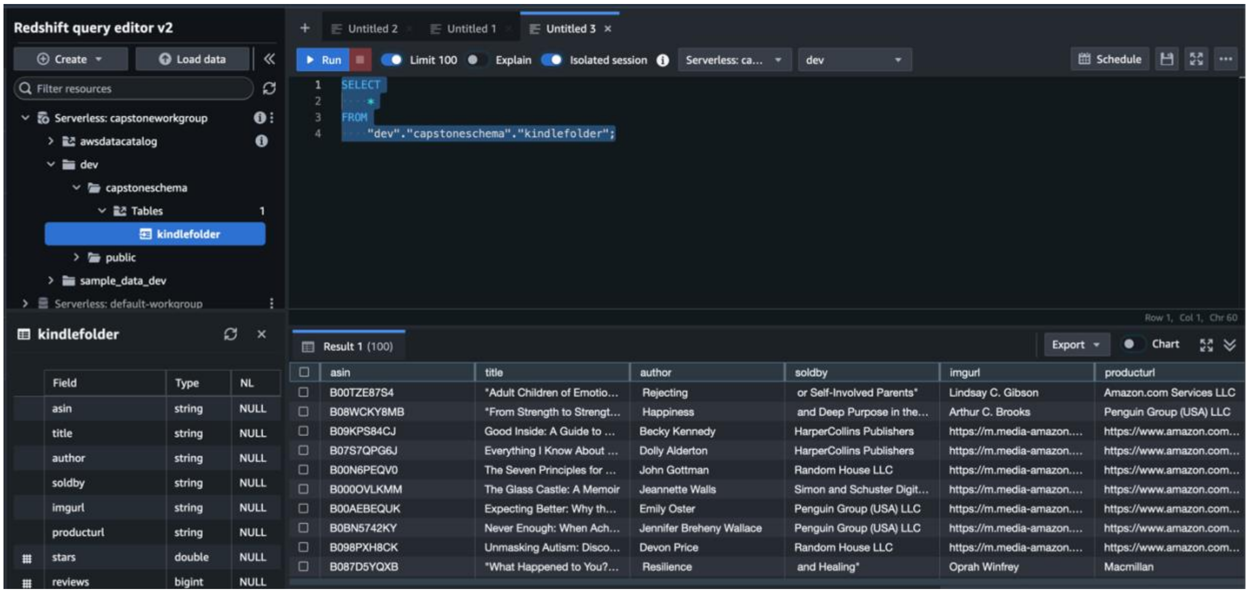
IAM role: arn:aws:iam:[redacted]ro...

Catalog IAM role (optional): Choose an IAM role



kindlefolder

Field	Type	NL
asin	string	NULL
title	string	NULL
author	string	NULL
soldby	string	NULL
imgurl	string	NULL
producturl	string	NULL
stars	double	NULL
reviews	bigint	NULL
price	double	NULL



```
SELECT
    *
FROM
    "dev"."capstoneschema"."kindlefolder";
```

kindlefolder

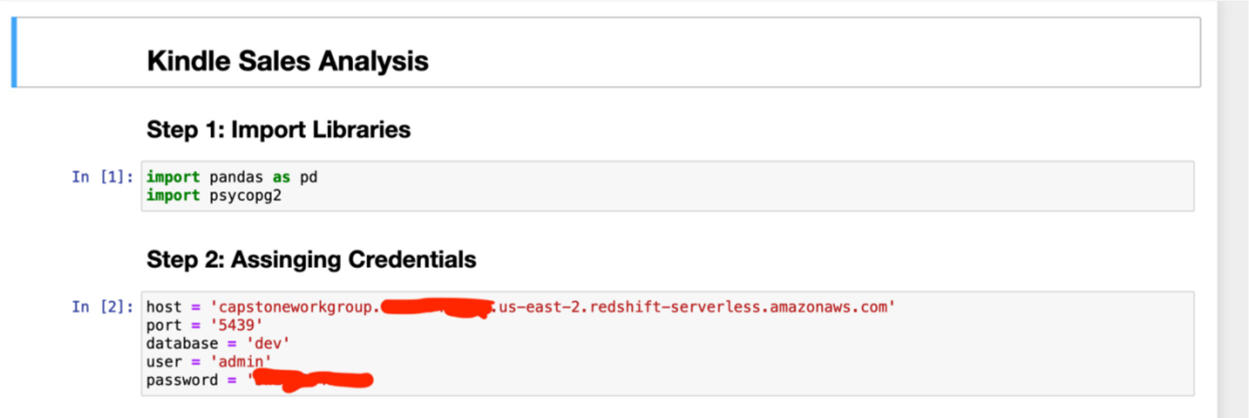
Field	Type	NL
asin	string	NULL
title	string	NULL
author	string	NULL
soldby	string	NULL
imgurl	string	NULL
producturl	string	NULL
stars	double	NULL
reviews	bigint	NULL

Result 1 (100)

asin	title	author	soldby	imgurl	producturl
B00TZE87S4	"Adult Children of Emotio...	Rejecting	or Self-Involved Parents"	Lindsay C. Gibson	Amazon.com Services LLC
B08WCKY8MB	"From Strength to Strengt...	Happiness	and Deep Purpose in the...	Arthur C. Brooks	Penguin Group (USA) LLC
B09KPS84CJ	Good Inside: A Guide to ...	Becky Kennedy	HarperCollins Publishers	https://m.media-amazon....	https://www.amazon.com...
B07S7QPG6J	Everything I Know About ...	Dolly Alderton	HarperCollins Publishers	https://m.media-amazon....	https://www.amazon.com...
B00N6PEQV0	The Seven Principles for ...	John Gottman	Random House LLC	https://m.media-amazon....	https://www.amazon.com...
B000OVLKMM	The Glass Castle: A Memoir	Jeannette Walls	Simon and Schuster Digit...	https://m.media-amazon....	https://www.amazon.com...
B00AEBEQUK	Expecting Better: Why th...	Emily Oster	Penguin Group (USA) LLC	https://m.media-amazon....	https://www.amazon.com...
B0BN5742KY	Never Enough: When Ach...	Jennifer Breheny Wallace	Penguin Group (USA) LLC	https://m.media-amazon....	https://www.amazon.com...
B098PXH8CK	Unmasking Autism: Disco...	Devon Price	Random House LLC	https://m.media-amazon....	https://www.amazon.com...
B087D5YQXB	"What Happened to You?...	Resilience	and Healing"	Oprah Winfrey	Macmillan

Part 3: Jupyter Notebook

Finally, we can connect to the Redshift data warehouse using Jupyter notebook.



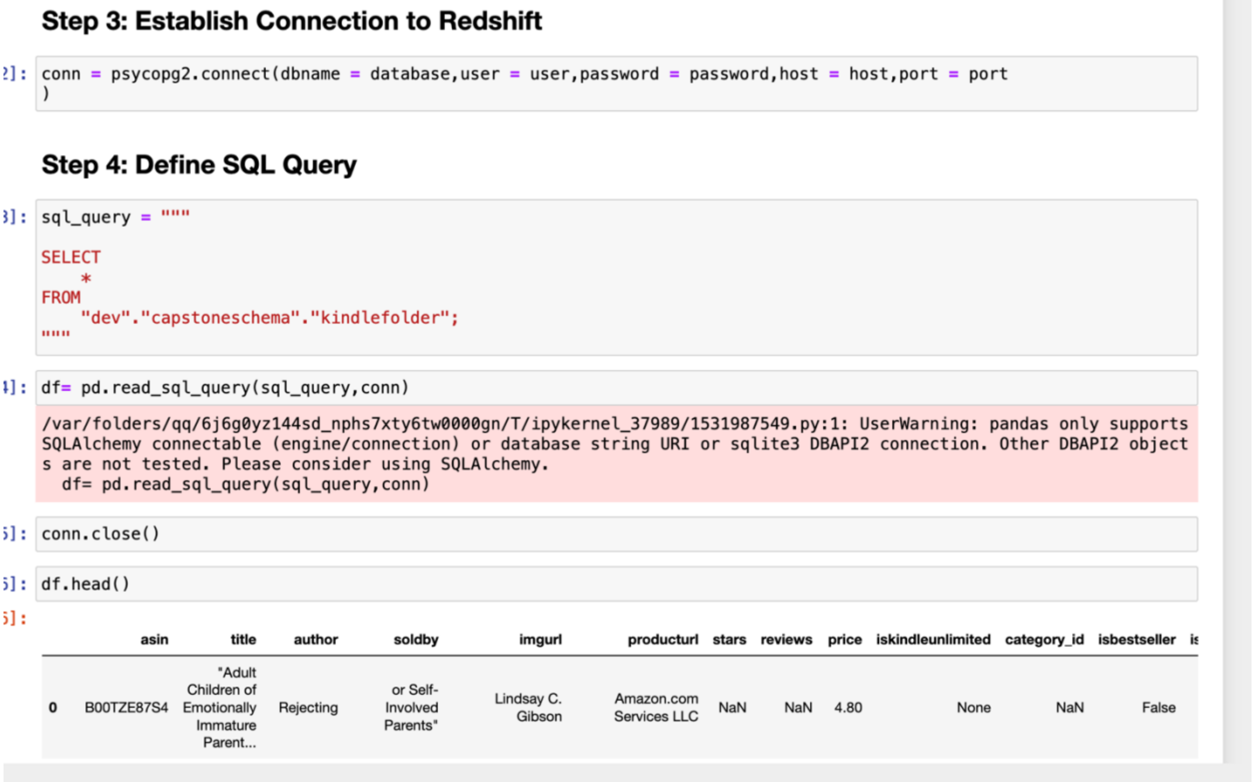
Kindle Sales Analysis

Step 1: Import Libraries

```
In [1]: import pandas as pd
        import psycopg2
```

Step 2: Assinging Credentials

```
In [2]: host = 'capstoneworkgroup.[redacted].us-east-2.redshift-serverless.amazonaws.com'
        port = '5439'
        database = 'dev'
        user = 'admin'
        password = '[redacted]
```



Step 3: Establish Connection to Redshift

```
2]: conn = psycopg2.connect(dbname = database,user = user,password = password,host = host,port = port
    )
```

Step 4: Define SQL Query

```
3]: sql_query = """

    SELECT
        *
    FROM
        "dev"."capstoneschema"."kindlefolder";
    """
```

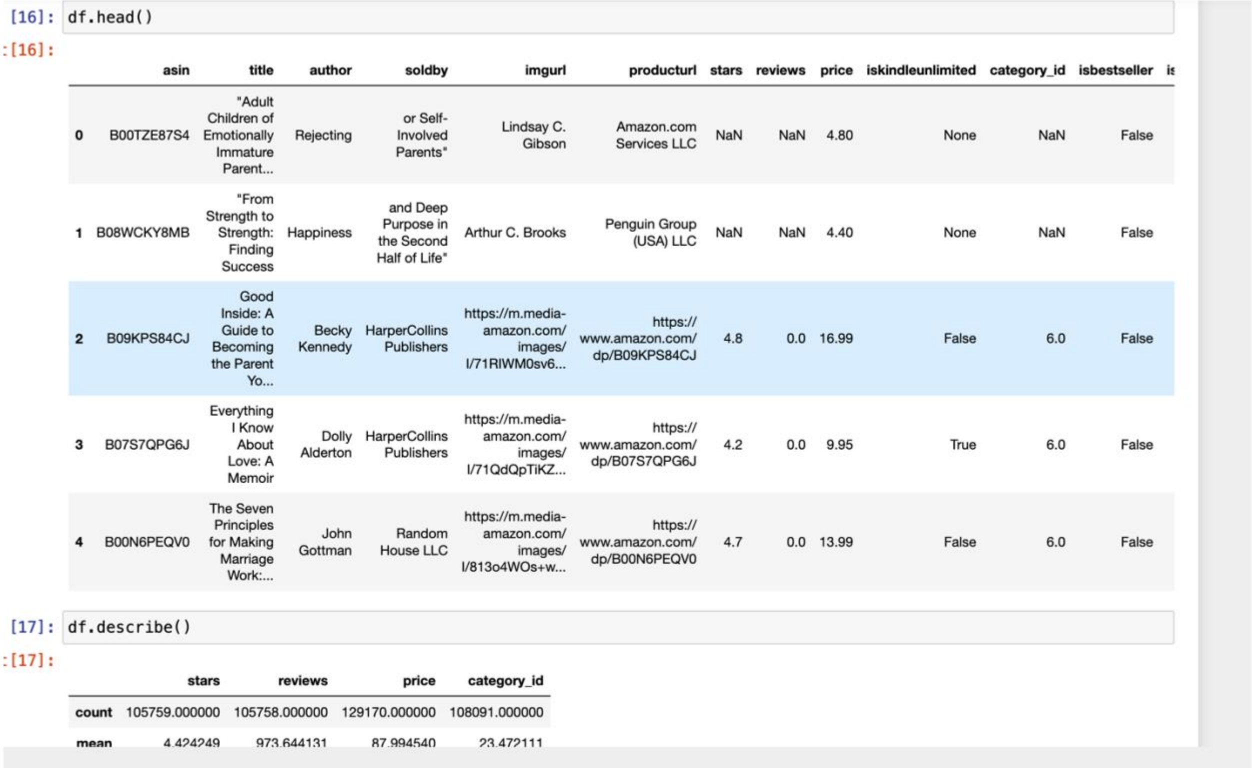
```
4]: df= pd.read_sql_query(sql_query,conn)

    /var/folders/qq/6j6g0yz144sd_nphs7xty6tw0000gn/T/ipykernel_37989/1531987549.py:1: UserWarning: pandas only supports
    SQLAlchemy connectable (engine/connection) or database string URI or sqlite3 DBAPI2 connection. Other DBAPI2 object
    s are not tested. Please consider using SQLAlchemy.
      df= pd.read_sql_query(sql_query,conn)
```

```
5]: conn.close()
```

```
5]: df.head()
```

5]:

	asin	title	author	soldby	imgurl	producturl	stars	reviews	price	iskindleunlimited	category_id	isbestseller	is
0	B00TZE87S4	"Adult Children of Emotionally Immature Parent...	Rejecting	or Self-Involved Parents"	Lindsay C. Gibson	Amazon.com Services LLC	NaN	NaN	4.80	None	NaN	False	



```
[16]: df.head()
```

:[16]:

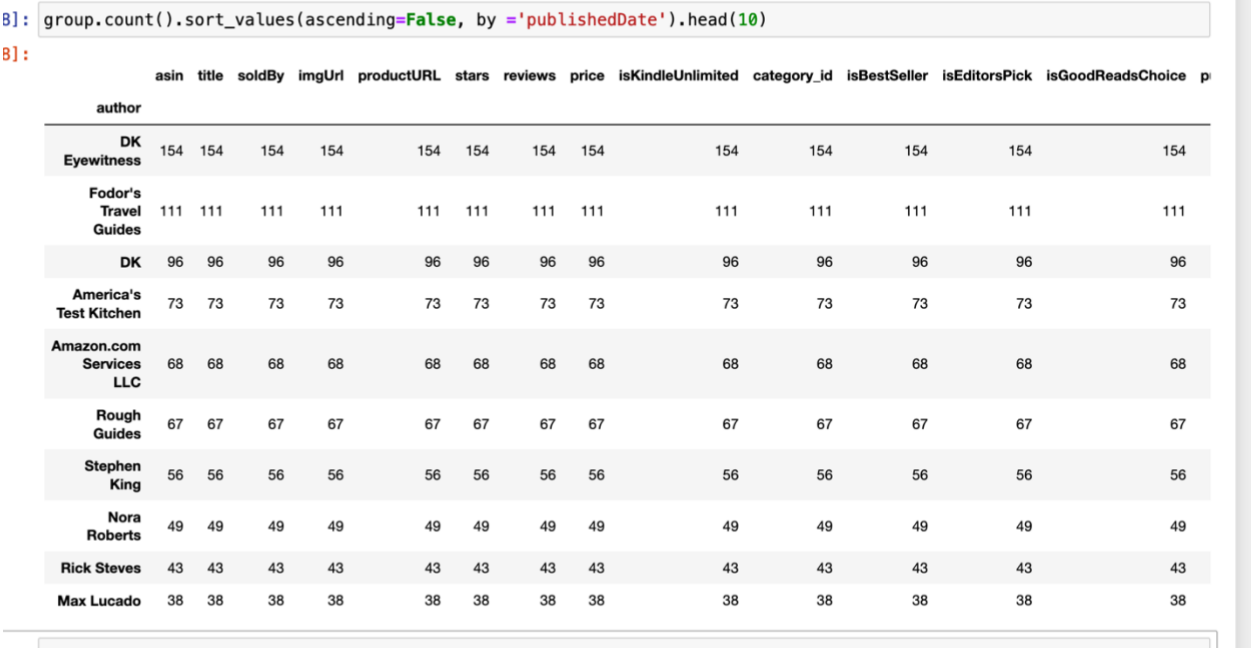
	asin	title	author	soldby	imgurl	producturl	stars	reviews	price	iskindleunlimited	category_id	isbestseller	is
0	B00TZE87S4	"Adult Children of Emotionally Immature Parent...	Rejecting	or Self-Involved Parents"	Lindsay C. Gibson	Amazon.com Services LLC	NaN	NaN	4.80	None	NaN	False	
1	B08WCKY8MB	"From Strength to Strength: Finding Success	Happiness	and Deep Purpose in the Second Half of Life"	Arthur C. Brooks	Penguin Group (USA) LLC	NaN	NaN	4.40	None	NaN	False	
2	B09KPS84CJ	Good Inside: A Guide to Becoming the Parent Yo...	Becky Kennedy	HarperCollins Publishers	https://m.media-amazon.com/images/I/71RIWM0sv6...	https://www.amazon.com/dp/B09KPS84CJ	4.8	0.0	16.99	False	6.0	False	
3	B07S7QPG6J	Everything I Know About Love: A Memoir	Dolly Alderton	HarperCollins Publishers	https://m.media-amazon.com/images/I/71QdQpTiKZ...	https://www.amazon.com/dp/B07S7QPG6J	4.2	0.0	9.95	True	6.0	False	
4	B00N6PEQV0	The Seven Principles for Making Marriage Work:...	John Gottman	Random House LLC	https://m.media-amazon.com/images/I/813o4WOs+w...	https://www.amazon.com/dp/B00N6PEQV0	4.7	0.0	13.99	False	6.0	False	

```
[17]: df.describe()
```

:[17]:

	stars	reviews	price	category_id
count	105759.000000	105758.000000	129170.000000	108091.000000
mean	4.424249	973.644131	87.994540	23.472111

Now we can begin analysis and answer the questions from the introduction. First, the 10 most published authors.

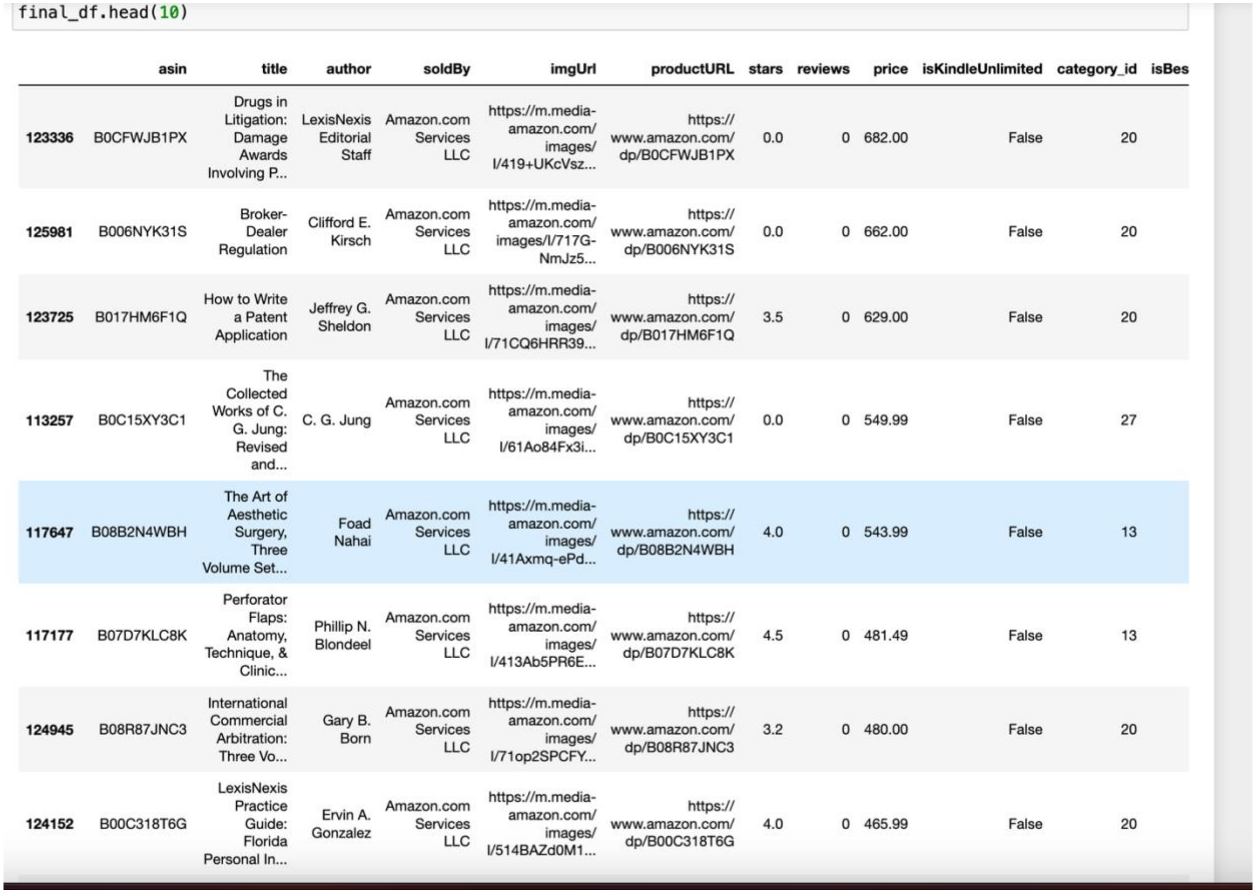


```
B]: group.count().sort_values(ascending=False, by ='publishedDate').head(10)
```

B]:

	asin	title	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	isEditorsPick	isGoodReadsChoice	p
author														
DK Eyewitness	154	154	154	154	154	154	154	154	154	154	154	154	154	
Fodor's Travel Guides	111	111	111	111	111	111	111	111	111	111	111	111	111	
DK	96	96	96	96	96	96	96	96	96	96	96	96	96	
America's Test Kitchen	73	73	73	73	73	73	73	73	73	73	73	73	73	
Amazon.com Services LLC	68	68	68	68	68	68	68	68	68	68	68	68	68	
Rough Guides	67	67	67	67	67	67	67	67	67	67	67	67	67	
Stephen King	56	56	56	56	56	56	56	56	56	56	56	56	56	
Nora Roberts	49	49	49	49	49	49	49	49	49	49	49	49	49	
Rick Steves	43	43	43	43	43	43	43	43	43	43	43	43	43	
Max Lucado	38	38	38	38	38	38	38	38	38	38	38	38	38	

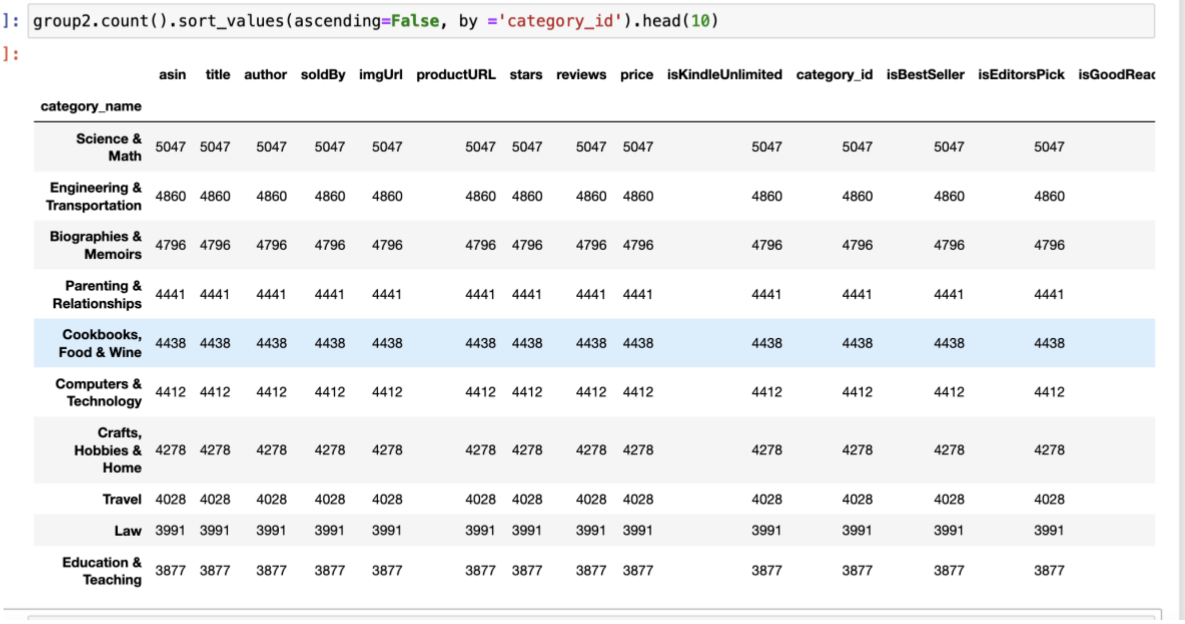
Next is the most expensive books.



```
final_df.head(10)
```

	asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBes
123336	B0CFWJB1PX	Drugs in Litigation: Damage Awards Involving P...	LexisNexis Editorial Staff	Amazon.com Services LLC	https://m.media-amazon.com/images/I/419+UKcVsz...	https://www.amazon.com/dp/B0CFWJB1PX	0.0	0	682.00	False	20	
125981	B006NYK31S	Broker-Dealer Regulation	Clifford E. Kirsch	Amazon.com Services LLC	https://m.media-amazon.com/images/I/717G-NmJz5...	https://www.amazon.com/dp/B006NYK31S	0.0	0	662.00	False	20	
123725	B017HM6F1Q	How to Write a Patent Application	Jeffrey G. Sheldon	Amazon.com Services LLC	https://m.media-amazon.com/images/I/71CQ6HRR39...	https://www.amazon.com/dp/B017HM6F1Q	3.5	0	629.00	False	20	
113257	B0C15XY3C1	The Collected Works of C. G. Jung: Revised and...	C. G. Jung	Amazon.com Services LLC	https://m.media-amazon.com/images/I/61Ao84Fx3i...	https://www.amazon.com/dp/B0C15XY3C1	0.0	0	549.99	False	27	
117647	B08B2N4WBH	The Art of Aesthetic Surgery, Three Volume Set...	Foad Nahai	Amazon.com Services LLC	https://m.media-amazon.com/images/I/41Axmq-ePd...	https://www.amazon.com/dp/B08B2N4WBH	4.0	0	543.99	False	13	
117177	B07D7KLC8K	Perforator Flaps: Anatomy, Technique, & Clinic...	Phillip N. Blondeel	Amazon.com Services LLC	https://m.media-amazon.com/images/I/413Ab5PR6E...	https://www.amazon.com/dp/B07D7KLC8K	4.5	0	481.49	False	13	
124945	B08R87JNC3	International Commercial Arbitration: Three Vo...	Gary B. Born	Amazon.com Services LLC	https://m.media-amazon.com/images/I/71op2SPCFY...	https://www.amazon.com/dp/B08R87JNC3	3.2	0	480.00	False	20	
124152	B00C318T6G	LexisNexis Practice Guide: Florida Personal In...	Ervin A. Gonzalez	Amazon.com Services LLC	https://m.media-amazon.com/images/I/514BAZd0M1...	https://www.amazon.com/dp/B00C318T6G	4.0	0	465.99	False	20	

Finally, the most popular genres/categories.



```
]: group2.count().sort_values(ascending=False, by ='category_id').head(10)
]:
```

category_name	asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	isEditorsPick	isGoodReac
Science & Math	5047	5047	5047	5047	5047	5047	5047	5047	5047	5047	5047	5047	5047	
Engineering & Transportation	4860	4860	4860	4860	4860	4860	4860	4860	4860	4860	4860	4860	4860	
Biographies & Memoirs	4796	4796	4796	4796	4796	4796	4796	4796	4796	4796	4796	4796	4796	
Parenting & Relationships	4441	4441	4441	4441	4441	4441	4441	4441	4441	4441	4441	4441	4441	
Cookbooks, Food & Wine	4438	4438	4438	4438	4438	4438	4438	4438	4438	4438	4438	4438	4438	
Computers & Technology	4412	4412	4412	4412	4412	4412	4412	4412	4412	4412	4412	4412	4412	
Crafts, Hobbies & Home	4278	4278	4278	4278	4278	4278	4278	4278	4278	4278	4278	4278	4278	
Travel	4028	4028	4028	4028	4028	4028	4028	4028	4028	4028	4028	4028	4028	
Law	3991	3991	3991	3991	3991	3991	3991	3991	3991	3991	3991	3991	3991	
Education & Teaching	3877	3877	3877	3877	3877	3877	3877	3877	3877	3877	3877	3877	3877	

References:

https://www.kaggle.com/datasets/asaniczka/amazon-kindle-books-dataset-2023-130k-books