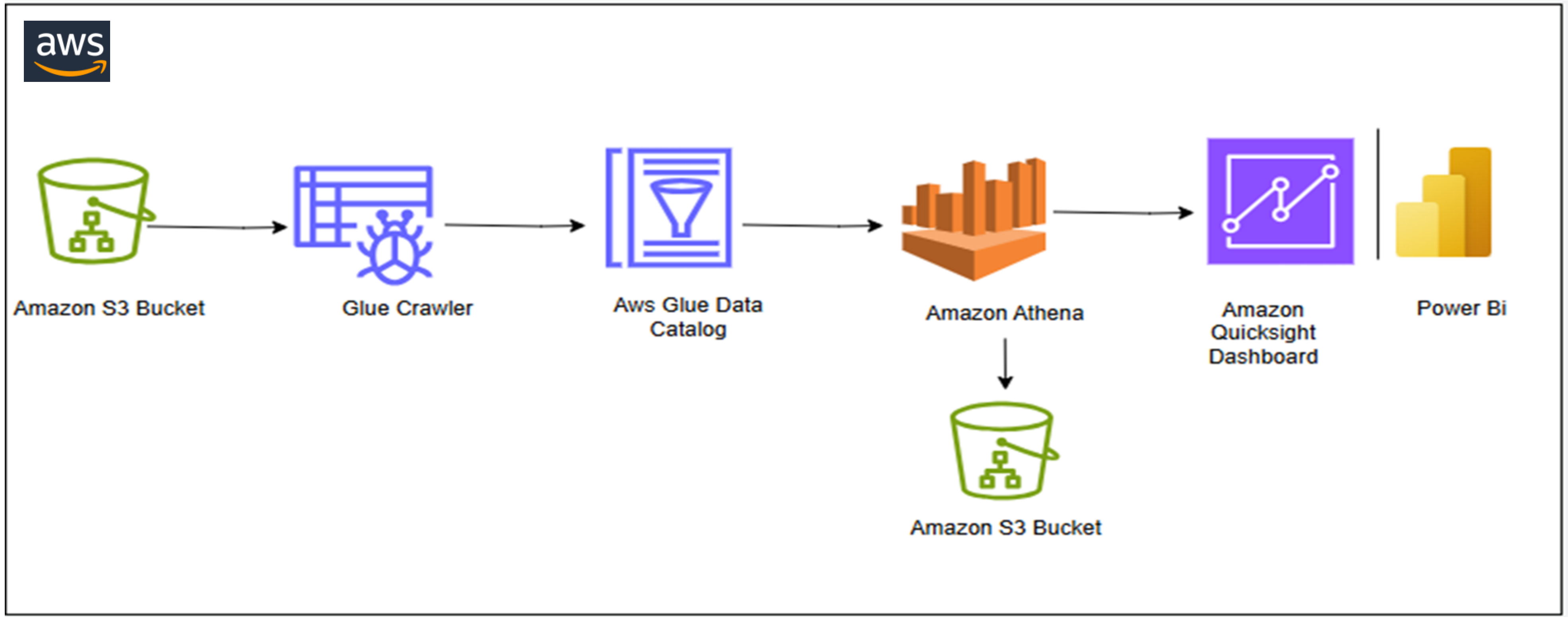
Case Study: Querying Data from S3 to Athena Using AWS Glue Crawler

Introduction

This case study demonstrates how to query data stored in Amazon S3 using Amazon Athena. We will use AWS Glue Crawler to automatically discover the schema of the data and make it available for querying in Athena. Additionally, we will configure Athena to store query results in an S3 bucket.



Prerequisites

Before querying data from Amazon S3 to Athena using AWS Glue Crawler, ensure you have the following ready for a smooth setup.

Active AWS Account with Permissions

You need an active AWS account with the necessary permissions to manage S3 buckets, run AWS Glue crawlers/jobs, and query data in Amazon Athena.

Amazon S3 Bucket with Data

An S3 bucket containing your data files (CSV, JSON, Parquet, ORC, Avro). For best Athena performance, consider partitioned Parquet or ORC data.

Chosen Access Method

You can use either the AWS Management Console or the AWS Command Line Interface (CLI). If using CLI, ensure it's installed and configured with your AWS credentials.

Basic Understanding of AWS Services

Familiarity with Amazon S3, AWS Glue, Amazon Athena, and basic SQL queries will help you follow this guide.

.

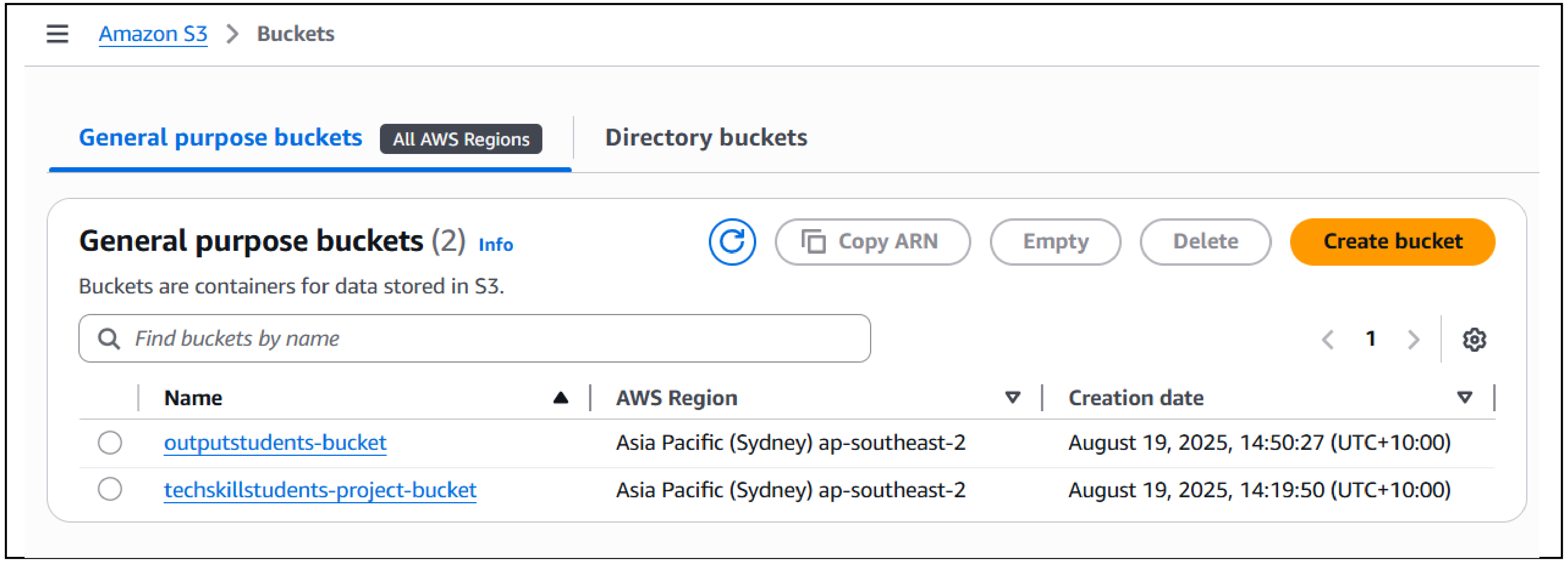
Step-by-Step Guide

Step 1: Set up an S3 Bucket

1. Open the Amazon S3 console.

2.Create a new S3 bucket or use an existing one.

3.Upload your data files to this S3 bucket.



General purpose buckets (2)

Buckets are containers for data stored in S3.

	Name ▲	AWS Region ▽	Creation date ▽
○	outputstudents-bucket	Asia Pacific (Sydney) ap-southeast-2	August 19, 2025, 14:50:27 (UTC+10:00)
○	techskillstudents-project-bucket	Asia Pacific (Sydney) ap-southeast-2	August 19, 2025, 14:19:50 (UTC+10:00)

The S3 bucket serves as the primary storage location for your data files that will be analyzed. Ensure your files are organized in a logical structure for easier crawling and querying.

Step 2: Create and Configure AWS Glue Crawler

1. Open the AWS Glue console.

2. Navigate to the 'Crawlers' section and click on 'Add crawler'.

3.Provide a name for your crawler and click 'Next'.

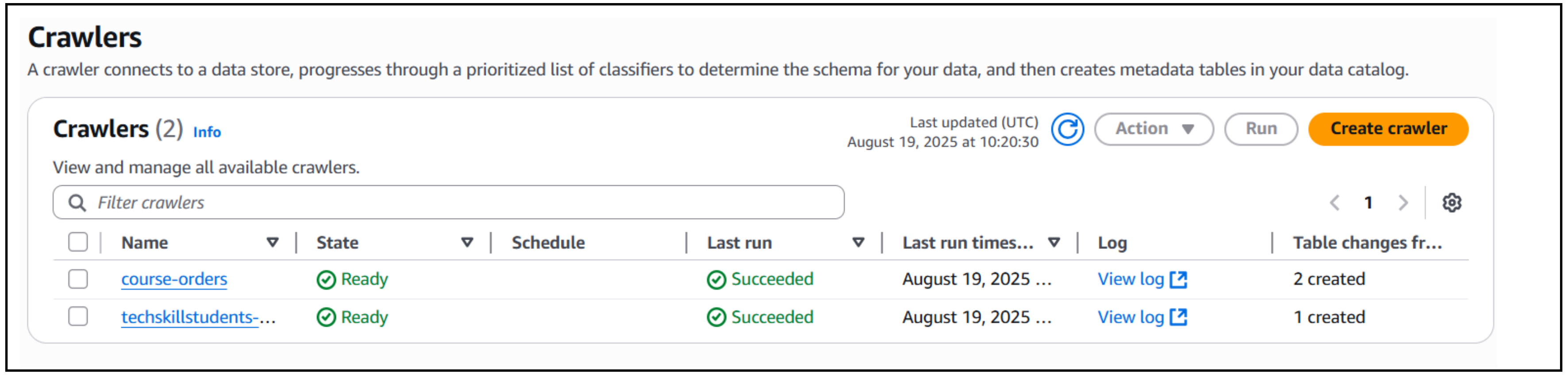
4.Define the data store by selecting 'S3' and specify the S3 bucket path where your data files are stored. Click 'Next'.

5.Choose or create an IAM role that has necessary permissions to access the S3 bucket and AWS Glue

6.Set the output database where the crawler results will be stored in the AWS Glue Data Catalog. If you don't have an existing database, create a new one.

7.Review the crawler configuration and click 'Finish'.

8.Start the crawler to analyze the data and populate the schema in the Data Catalog



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 (2) Info

Last updated (UTC)
August 19, 2025 at 10:20:30

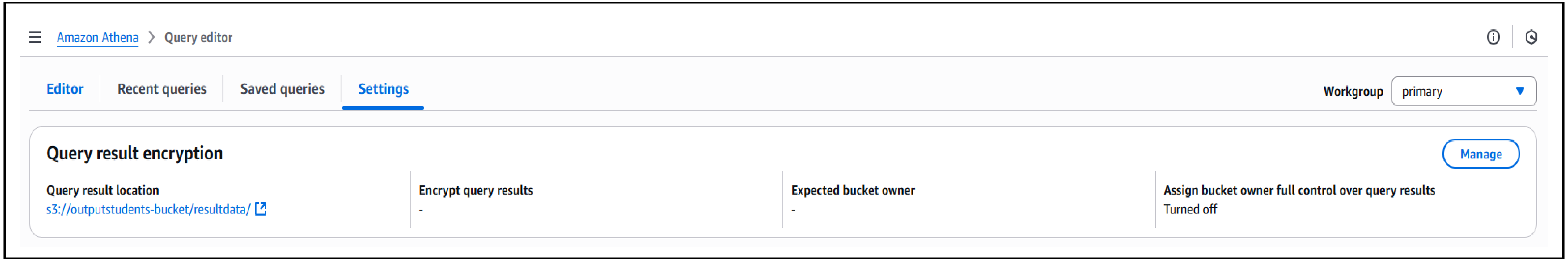
View and manage all available crawlers.

☐	Name	State	Schedule	Last run	Last run times...	Log	Table changes fr...
☐	course-orders	✓ Ready		✓ Succeeded	August 19, 2025 ...	View log	2 created
☐	techskillstudents-...	✓ Ready		✓ Succeeded	August 19, 2025 ...	View log	1 created

Step 3: Configure Query Result Location in Athena

1. Open the Amazon Athena console.
2. Go to 'Settings' and set the query result location to an S3 bucket where you want to store the query results.
3. Save the settings.

Configuring a query result location is essential as Athena stores all query results in S3. This allows you to maintain a history of query results and share them with others if needed.



Query result encryption

Query result location	Encrypt query results	Expected bucket owner	Assign bucket owner full control over query results
s3://outputstudents-bucket/resultdata/	-	-	Turned off

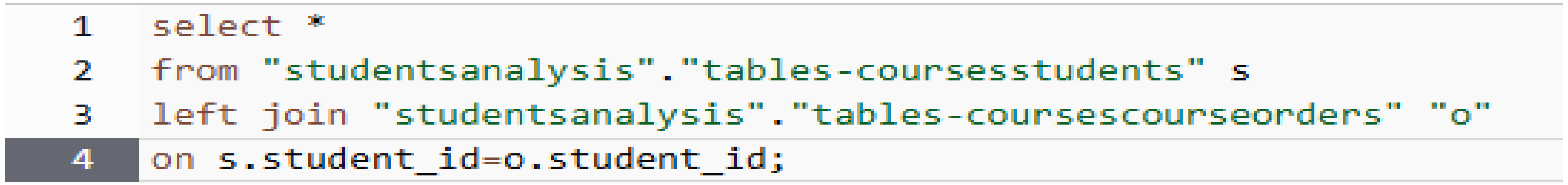
Step 4: Query Data Using Athena

1.In the Athena console, select the database created by the crawler from the Data Catalog.

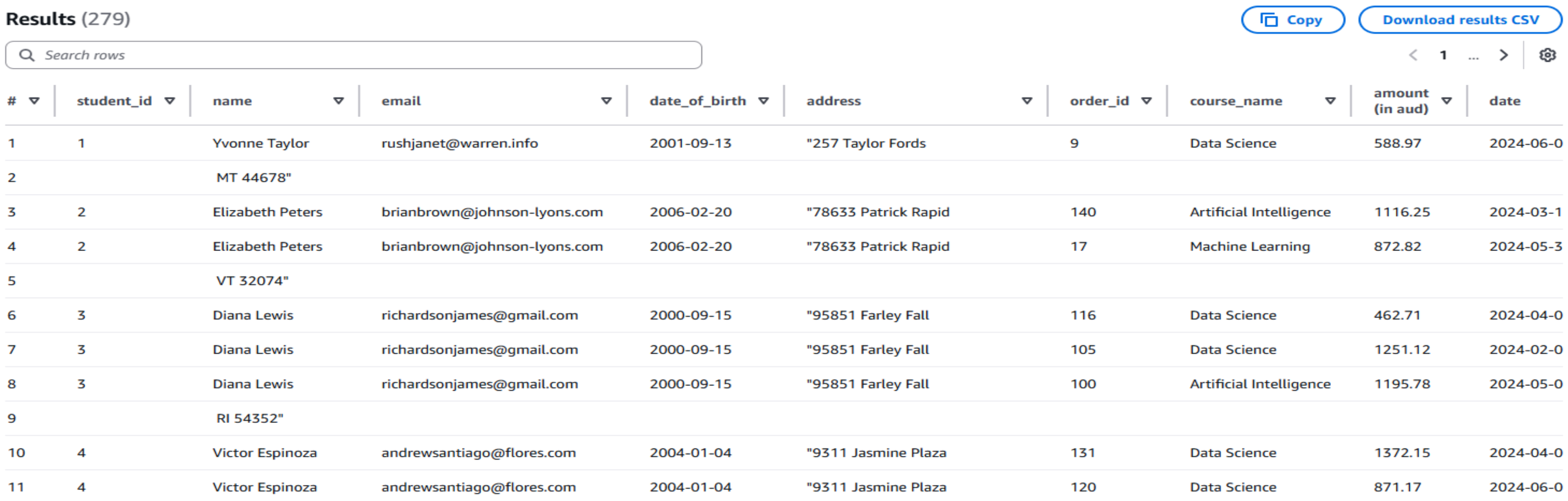
2.Write your SQL query to query the data stored in S3. For example:

3.Click on 'Run Query' to execute the SQL query.

4.View and analyze the query results in the Athena console or check the results stored in the specified S3 bucket.



```
select *
from "studentsanalysis"."tables-coursesstudents" s
left join "studentsanalysis"."tables-coursescourseorders" "o"
on s.student_id=o.student_id;
```



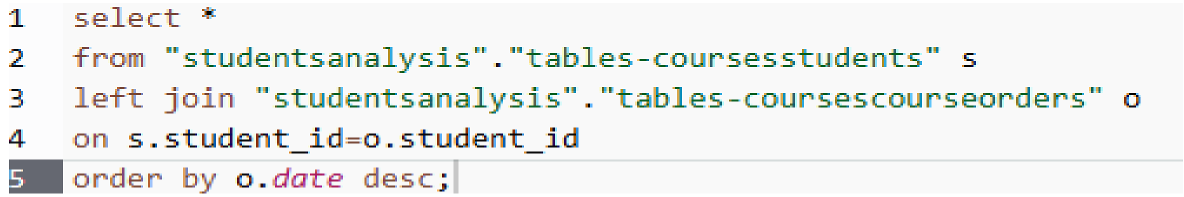
Results (279)

#	student_id	name	email	date_of_birth	address	order_id	course_name	amount (in aud)	date
1	1	Yvonne Taylor	rushjanet@warren.info	2001-09-13	"257 Taylor Fords	9	Data Science	588.97	2024-06-0
2		MT 44678"							
3	2	Elizabeth Peters	brianbrown@johnson-lyons.com	2006-02-20	"78633 Patrick Rapid	140	Artificial Intelligence	1116.25	2024-03-1
4	2	Elizabeth Peters	brianbrown@johnson-lyons.com	2006-02-20	"78633 Patrick Rapid	17	Machine Learning	872.82	2024-05-3
5		VT 32074"							
6	3	Diana Lewis	richardsonjames@gmail.com	2000-09-15	"95851 Farley Fall	116	Data Science	462.71	2024-04-0
7	3	Diana Lewis	richardsonjames@gmail.com	2000-09-15	"95851 Farley Fall	105	Data Science	1251.12	2024-02-0
8	3	Diana Lewis	richardsonjames@gmail.com	2000-09-15	"95851 Farley Fall	100	Artificial Intelligence	1195.78	2024-05-0
9		RI 54352"							
10	4	Victor Espinoza	andrewsantiago@flores.com	2004-01-04	"9311 Jasmine Plaza	131	Data Science	1372.15	2024-04-0
11	4	Victor Espinoza	andrewsantiago@flores.com	2004-01-04	"9311 Jasmine Plaza	120	Data Science	871.17	2024-06-0

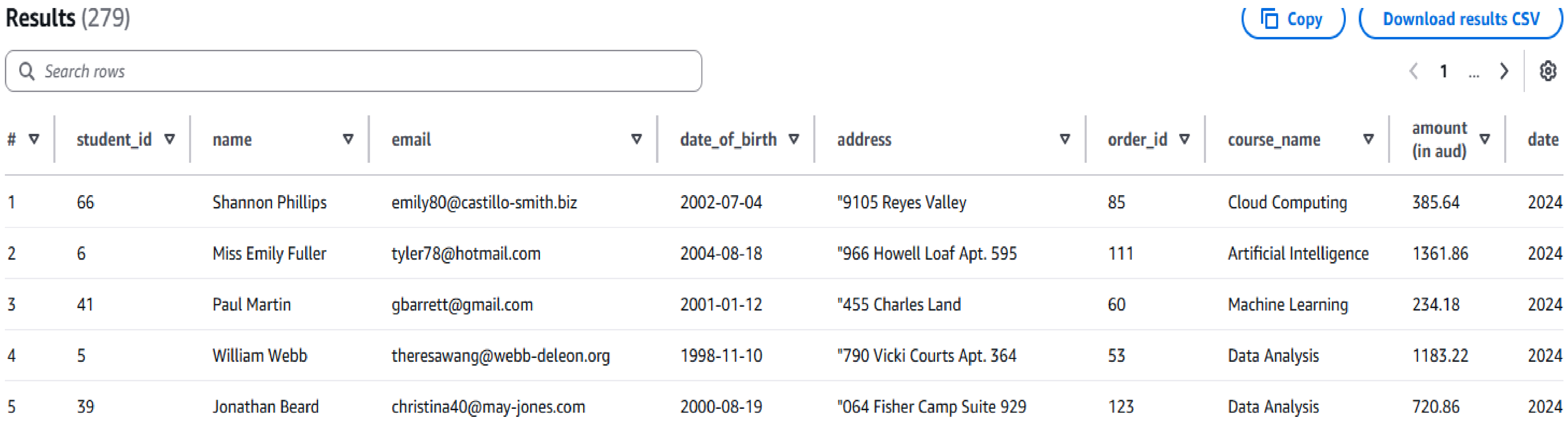
The results stored in the specified S3 bucket(Csv file):
409e5a71-3f91-4a22-9285-2fc821b20d50.xlsx

Queries:

1. Students with Their Orders



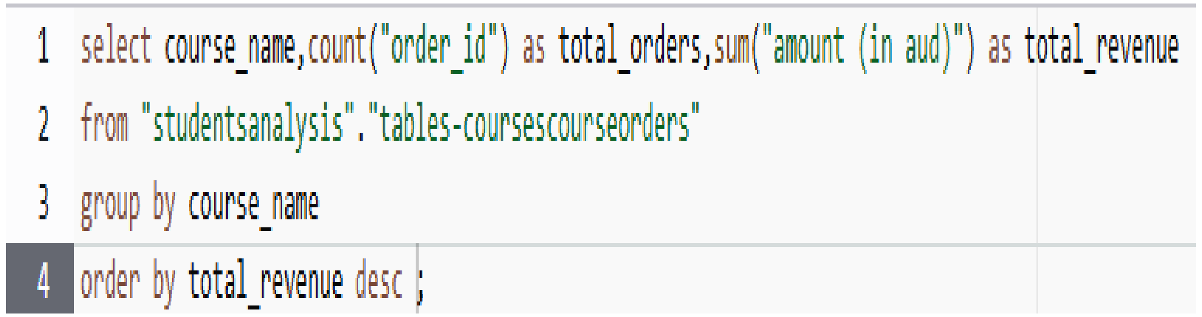
```
select *
from "studentsanalysis"."tables-coursesstudents" s
left join "studentsanalysis"."tables-coursescourseorders" o
on s.student_id=o.student_id
order by o.date desc;
```



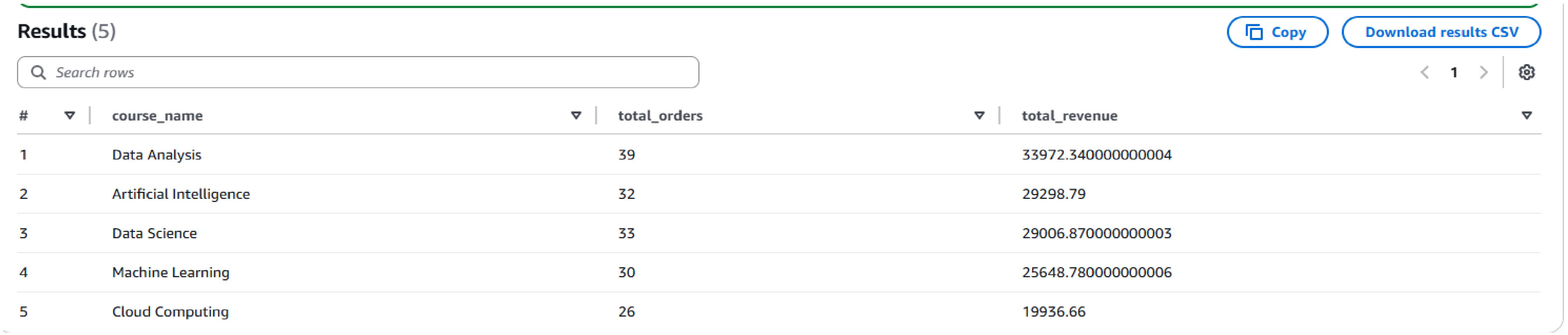
Results (279)

#	student_id	name	email	date_of_birth	address	order_id	course_name	amount (in aud)	date
1	66	Shannon Phillips	emily80@castillo-smith.biz	2002-07-04	"9105 Reyes Valley	85	Cloud Computing	385.64	2024
2	6	Miss Emily Fuller	tyler78@hotmail.com	2004-08-18	"966 Howell Loaf Apt. 595	111	Artificial Intelligence	1361.86	2024
3	41	Paul Martin	gbarrett@gmail.com	2001-01-12	"455 Charles Land	60	Machine Learning	234.18	2024
4	5	William Webb	theresawang@webb-deleon.org	1998-11-10	"790 Vicki Courts Apt. 364	53	Data Analysis	1183.22	2024
5	39	Jonathan Beard	christina40@may-jones.com	2000-08-19	"064 Fisher Camp Suite 929	123	Data Analysis	720.86	2024

2.Total Orders & Revenue by Course



```
select course_name,count("order_id") as total_orders,sum("amount (in aud)") as total_revenue
from "studentsanalysis"."tables-coursescourseorders"
group by course_name
order by total_revenue desc ;
```

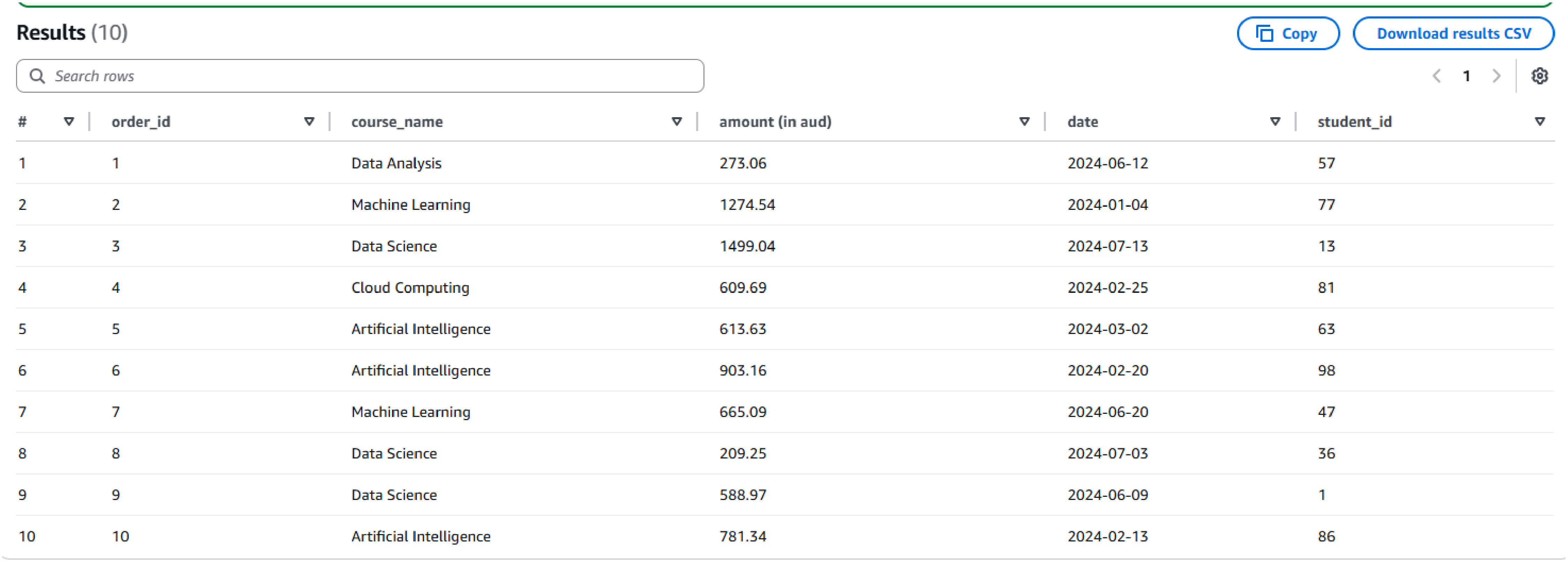


Results (5)

#	course_name	total_orders	total_revenue
1	Data Analysis	39	33972.340000000004
2	Artificial Intelligence	32	29298.79
3	Data Science	33	29006.870000000003
4	Machine Learning	30	25648.780000000006
5	Cloud Computing	26	19936.66

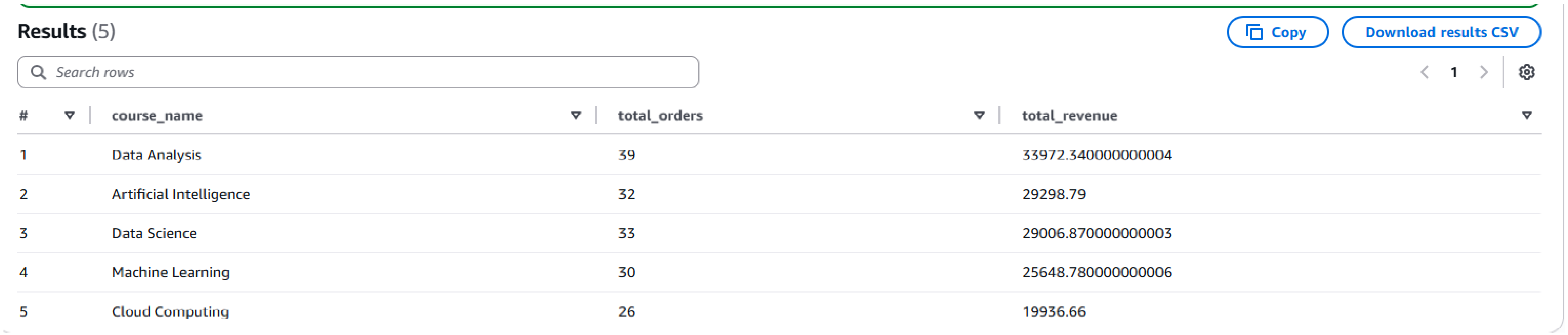
3. Top 10 Students by Orders

```
SELECT *
FROM "studentsanalysis"."tables-coursescourseorders"
LIMIT 10;
```



Results (10)

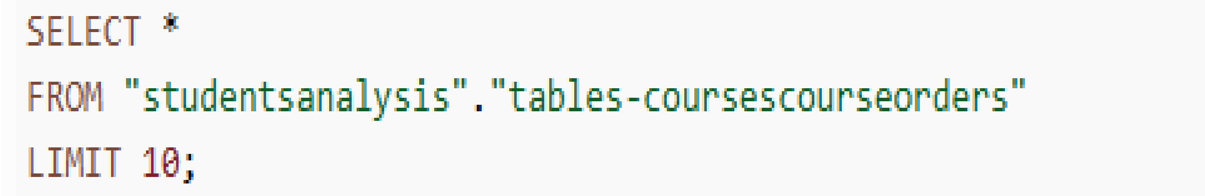
#	order_id	course_name	amount (in aud)	date	student_id
1	1	Data Analysis	273.06	2024-06-12	57
2	2	Machine Learning	1274.54	2024-01-04	77
3	3	Data Science	1499.04	2024-07-13	13
4	4	Cloud Computing	609.69	2024-02-25	81
5	5	Artificial Intelligence	613.63	2024-03-02	63
6	6	Artificial Intelligence	903.16	2024-02-20	98
7	7	Machine Learning	665.09	2024-06-20	47
8	8	Data Science	209.25	2024-07-03	36
9	9	Data Science	588.97	2024-06-09	1
10	10	Artificial Intelligence	781.34	2024-02-13	86



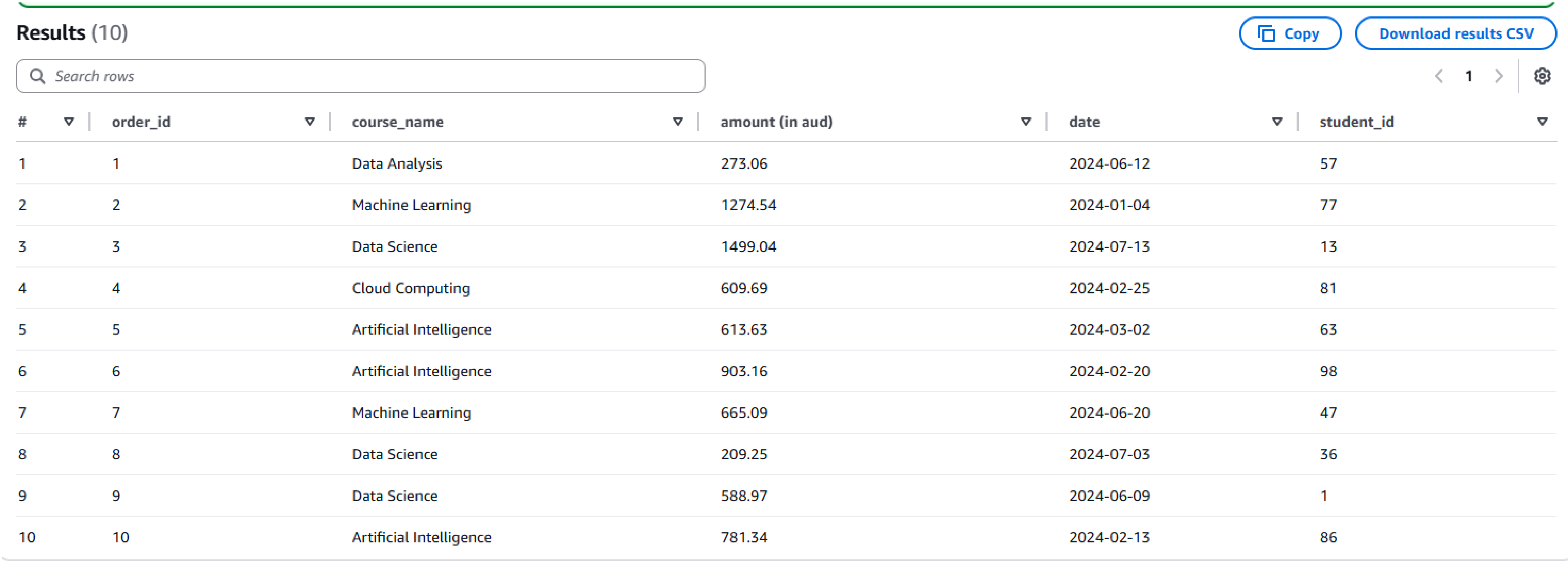
Results (5)

#	course_name	total_orders	total_revenue
1	Data Analysis	39	33972.340000000004
2	Artificial Intelligence	32	29298.79
3	Data Science	33	29006.870000000003
4	Machine Learning	30	25648.780000000006
5	Cloud Computing	26	19936.66

3. Top 10 Students by Orders



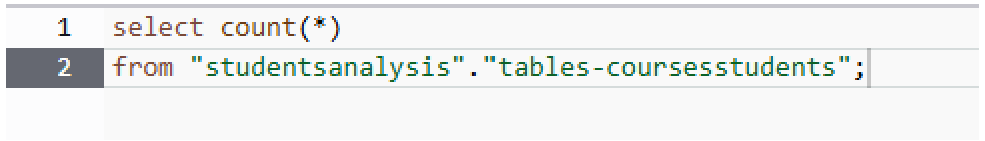
```
SELECT *
FROM "studentsanalysis"."tables-coursescourseorders"
LIMIT 10;
```



Results (10)

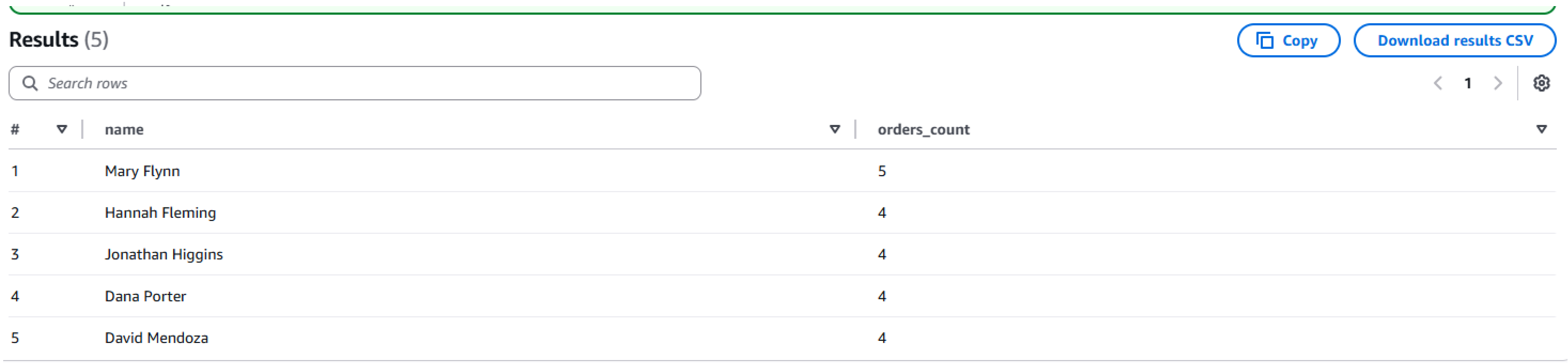
#	order_id	course_name	amount (in aud)	date	student_id
1	1	Data Analysis	273.06	2024-06-12	57
2	2	Machine Learning	1274.54	2024-01-04	77
3	3	Data Science	1499.04	2024-07-13	13
4	4	Cloud Computing	609.69	2024-02-25	81
5	5	Artificial Intelligence	613.63	2024-03-02	63
6	6	Artificial Intelligence	903.16	2024-02-20	98
7	7	Machine Learning	665.09	2024-06-20	47
8	8	Data Science	209.25	2024-07-03	36
9	9	Data Science	588.97	2024-06-09	1
10	10	Artificial Intelligence	781.34	2024-02-13	86

4. Number of Students



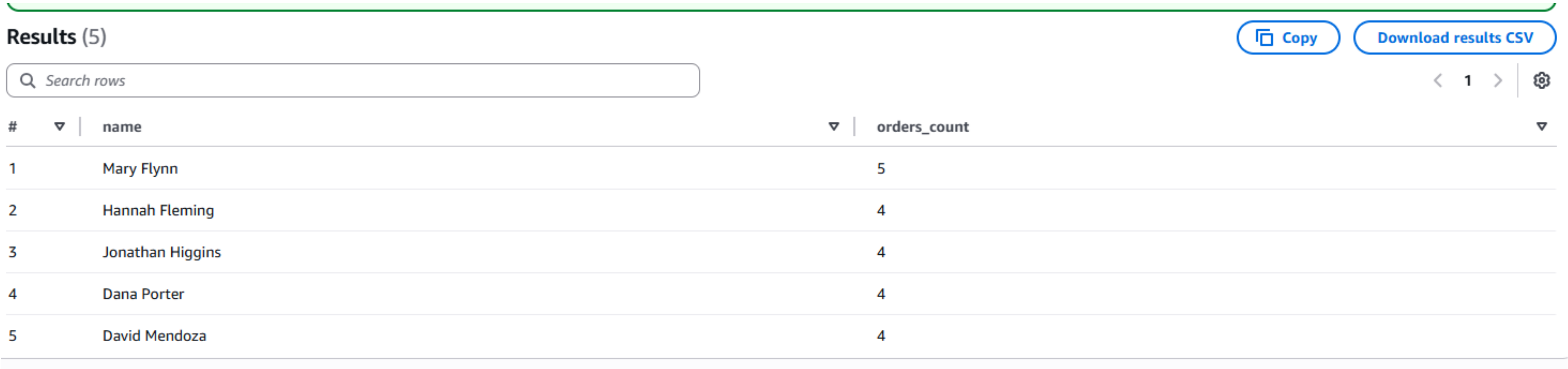
```
select count(*)
from "studentsanalysis"."tables-coursesstudents";
```

Results (1)



Results (5)

#	name	orders_count
1	Mary Flynn	5
2	Hannah Fleming	4
3	Jonathan Higgins	4
4	Dana Porter	4
5	David Mendoza	4



Results (5)

#	name	orders_count
1	Mary Flynn	5
2	Hannah Fleming	4
3	Jonathan Higgins	4
4	Dana Porter	4
5	David Mendoza	4

Conclusion

> By following these steps, you can efficiently query data stored in Amazon S3 using Amazon Athena, with the help of AWS Glue Crawler to automatically discover and catalog the schema. This setup allows for flexible and powerful data analysis without the need to set up and manage a traditional database infrastructure.

Store Data in S3

Maintain your data in cost-effective S3 storage

Crawl with AWS Glue

Automatically discover and catalog data schema

Query with Athena

Analyze data using standard SQL queries

Gain Insights

Make data-driven decisions without complex infrastructure

This serverless approach to data analysis provides a powerful yet simple way to extract insights from your data without the operational overhead of traditional database systems.