

Indication of how the exercise was carried out

1 - Preparation of the Environment

1.1 - Account and Group Creation.

- An AWS account and a specific user group was created to manage the resources needed for the pipeline.
- Users in the group were associated with all necessary administrative policies to facilitate access and management of the AWS services involved.

IAM > Gruppi di utenti

Gruppi di utenti (1) [Informazioni](#)

↻

Elimina

Crea gruppo

A user group is a collection of IAM users. Use groups to specify permissions for a collection of users.

Cerca

< 1 > ⚙

☐

Nome del gruppo ▲

☐

Utenti ▼

☐

Autorizzazioni ▼

☐

Ora creazione ▼

☐

[gruppo_sg](#)

1

Definito

2 giorni fa

utente_sg [Informazioni](#)

Elimina

Riepilogo

ARN
arn:aws:iam::575108947083:user/utente_sg

Accesso alla console
Disabilitato

Chiave di accesso 1
[Crea chiave di accesso](#)

Creato
October 14, 2024, 11:51 (UTC+02:00)

Ultimo accesso alla console
-

Autorizzazioni

Gruppi (1)

Tag

Credenziali di sicurezza

Last Accessed

Policy di autorizzazione (7)

↻

Rimuovi

Aggiungi autorizzazioni ▼

Le autorizzazioni sono definite da policy collegate all'utente direttamente o tramite gruppi.

Cerca

Filtra per Tipo
Tutti i tipi ▼

< 1 > ⚙

☐

Nome della policy [?](#) ▲

☐

Tipo ▼

☐

Collegato tramite [?](#)

☐

[AdministratorAccess](#)

Gestite da AWS - funzione lavorativa

[Gruppo \[gruppo_sg\]\(#\)](#)

☐

[AmazonRedshiftDataFullAccess](#)

Gestite da AWS

[Gruppo \[gruppo_sg\]\(#\)](#)

☐

[AmazonRedshiftFullAccess](#)

Gestite da AWS

[Gruppo \[gruppo_sg\]\(#\)](#)

☐

[AmazonRedshiftQueryEditor](#)

Gestite da AWS

[Gruppo \[gruppo_sg\]\(#\)](#)

☐

[AmazonS3FullAccess](#)

Gestite da AWS

Direttamente, Gruppo [gruppo_sg](#)

☐

[AWSGlueConsoleFullAccess](#)

Gestite da AWS

Direttamente, Gruppo [gruppo_sg](#)

☐

[AWSGlueServiceRole](#)

Gestite da AWS

[Gruppo \[gruppo_sg\]\(#\)](#)

1.2 - Role Configuration.

- Two specific roles were created: one for AWS Glue and one for Amazon Redshift, both with the necessary policies to access S3 and Redshift resources.

<input type="checkbox"/>	Glue-role	Servizio AWS: glue	3 ore fa
<input type="checkbox"/>	Redshift-role	Servizio AWS: redshift	-

Policy di autorizzazione (10) Informazioni			
Puoi collegare fino a 10 policy gestite.			
<div> <input type="text" value="Cerca"/> <div> Filtra per Tipo <div>Tutti i tipi</div> </div> </div> <div>< 1 > </div>			
<input type="checkbox"/>	Nome della policy ?	Tipo	Entità collegate
<input type="checkbox"/>	AdministratorAccess	Gestite da AWS - funzione lavorativa	4
<input type="checkbox"/>	AmazonDMSRedshiftS3Role	Gestite da AWS	3
<input type="checkbox"/>	AmazonGrafanaRedshiftAccess	Gestite da AWS	3
<input type="checkbox"/>	AmazonRedshiftAllCommandsFullAcc...	Gestite da AWS	2
<input type="checkbox"/>	AmazonRedshiftDataFullAccess	Gestite da AWS	4
<input type="checkbox"/>	AmazonRedshiftFullAccess	Gestite da AWS	4
<input type="checkbox"/>	AmazonRedshiftQueryEditorV2FullAc...	Gestite da AWS	3
<input type="checkbox"/>	AmazonS3FullAccess	Gestite da AWS	5
<input type="checkbox"/>	AWSGlueConsoleFullAccess	Gestite da AWS	5
<input type="checkbox"/>	my_personal_policy	Customer inline	0

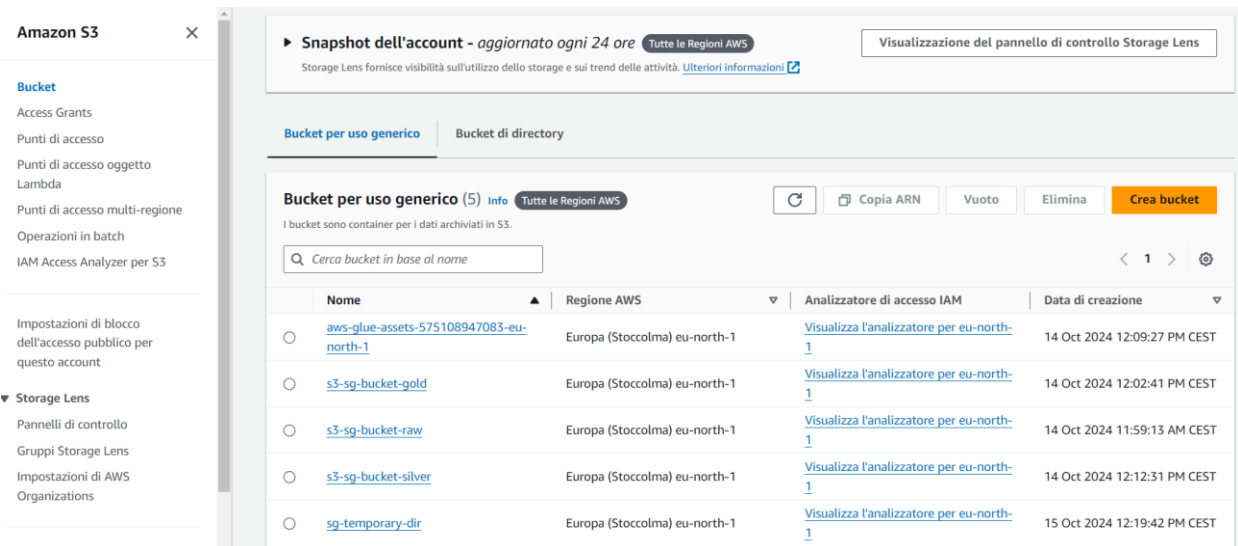
- A custom policy was added to ensure security and appropriate access to resources:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "VisualEditor0",
      "Effect": "Allow",
      "Action": [
        "redshift-serverless:*",
        "s3:*",
        "glue:*"
      ],
      "Resource": "*"
    }
  ]
}
```

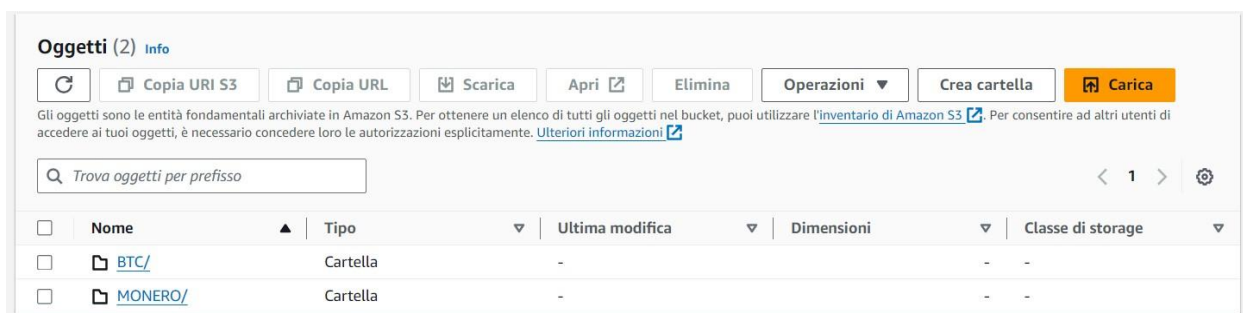
(Redundant but added for safety, as this is my first time interfacing with AWS).

1.3 - Creation of the S3 Buckets.

- Three buckets were created: s3-sg-bucket-raw for raw data, s3-sg-bucket-silver for clean and transformed data, and s3-sg-bucket-gold for finalized data.



- Within each bucket, 'subfolders' were created for BTC and MONERO to organize the data in a structured way.



- Necessary files were manually uploaded into the subfolders of s3-sg-bucket-raw.

2 – Pipeline Implementation with AWS Glue

2.1 - Cleanup and Transformation Scripts.

- Imported and configured the environment with Spark and Glue to process the data.
- Data were read from the S3 bucket in CSV format.
- Converted data to DataFrame to facilitate cleanup and transformation operations, handling missing values by averaging over the last 5 days.
- Saved the cleaned result in Parquet format to s3-sg-bucket-silver.

2.2 - Script T2: Calculating Moving Average and Join

- Clean data were read from the s3-sg-bucket-silver bucket.
- The 10-day moving average was calculated to reduce noise in the price data.
- Performed join with Google Trends data to create a unified dataset.
- Saved the final result in Parquet format to s3-sg-bucket-gold.

2.3 - Script L: Loading to Redshift

- Converted and mapped the data for inclusion in the Redshift database, ensuring proper typing.
- Uploaded the final dataset to Redshift Serverless to allow for future analysis.

- To make the upload to Redshift, a 'Connection' was created in Glue bringing back Redshift namespace and workgroup configuration data (see on next section).

AWS Glue > Connectors > my_redshift_connection

my_redshift_connection

Edit Delete Create job

Connection details Info

Connector type	JDBC	Connection URL	jdbc:redshift://sg-workgroup.575108947083.eu-north-1.redshift-serverless.amazonaws.com:5439/dev
Driver class name	-	Driver path	-
Username	sg-admin	Require SSL connection	true
Subnet	subnet-0d5cbcd1a0d78928b	Security groups	sg-03d48acb947288df5
Description	-	Created on	2024-10-15 16:26:30.735000
Last modified	2024-10-15 18:58:00.589000	Class name	-

3 - Configuring Amazon Redshift Serverless

3.1 - Creating the Namespace and Workgroup.

- Configured the namespace and workgroup of Redshift Serverless to manage database resources.

sg-namespace Info

Operazioni ▼ Esegui query sui dati

Informazioni generali

Spazio dei nomi	sg-namespace	Stato	Available	Nome dell'utente amministratore	sg-admin
ID dello spazio dei nomi	649e180d-3d5f-49b0-9f6d-a1bdf743c628	Data di creazione	October 14, 2024, 20:32 (UTC+02:00)	Nome database	dev
ARN dello spazio dei nomi	arn:aws:redshift-serverless:eu-north-1:575108947083:namespace/649e180d-3d5f-49b0-9f6d-a1bdf743c628	Archiviazione utilizzata	129MB	Numero totale di tabelle	2

Gruppo di lavoro Backup dei dati Database Sicurezza e crittografia Unità di condivisione dati Integrazioni Zero-ETL Policy delle risorse Tag

Nome del gruppo di lavoro

Configura le risorse di calcolo per il tuo gruppo di lavoro.

Operazioni ▼

Gruppo di lavoro	sg-workgroup	Stato	Available
------------------	--------------	-------	-----------

sg-workgroup

OperazioniEsegui query sui dati

Informazioni generali

Gruppo di lavoro
sg-workgroup

Spazio dei nomi
sg-namespace

ARN del gruppo di lavoro
arn:aws:redshift-serverless:eu-north-1:575108947083:workgroup/1eb77af0-1285-44be-bf7f-37d6c08e5fd4

Workgroup version
1.0.76645

Data di creazione
October 14, 2024, 20:32 (UTC+02:00)

Stato
Available

Configurazione
Produzione

Capacità di base
8 RPU

Nome di dominio personalizzato
-

Patch version
Patch 185

Endpoint
sg-workgroup.575108947083.eu-north-1.redshift-serverless.amazonaws.com:5439/dev

URL JDBC
jdbc:redshift://sg-workgroup.575108947083.eu-north-1.redshift-serverless.amazonaws.com:5439/dev

URL ODBC
Driver={Amazon Redshift (x64)}; Server=sg-workgroup.575108947083.eu-north-1.redshift-serverless.amazonaws.com; Database=dev

- Security rules have been set up to allow secure access to Redshift resources.

Accesso ai datiLimitiPerformanceTag

Rete e sicurezza

Virtual Private Cloud (VPC)
vpc-09b3b92befa2990db

ID endpoint VPC
vpce-04bd1a01a0178eb92

Gruppo di sicurezza VPC
sg-03d48acb947288df5

Sottorete
subnet-0d5cbcd1a0d78928b,
subnet-0cf2c0bc3a6ac35e,
subnet-03a6eef658d6b5da5,

Instradamento VPC avanzato
Disattivato

Accessibile pubblicamente
Permetti alle istanze e ai dispositivi esterni al VPC di connettersi al database attraverso l'endpoint del cluster.

Attivato

IP address type
IPv4

EC2 > Gruppi di sicurezza > sg-03d48acb947288df5 - default > Modifica le regole in entrata

Modifica le regole in entrata

Le regole in entrata controllano il traffico in entrata che può raggiungere l'istanza.

Regole in entrata

ID della regola del gruppo di sicurezza	Tipo	Protocollo	Intervallo porte	Origine	Descrizione - facoltativa
sgr-028cdd9cb99f0d934	Regola TCP personalizzata	TCP	5493	Person... 151.77.75.33/32	
sgr-0d82f1d0f257e9bfa	Tutto il traffico	Tutti	Tutti	Person... 0.0.0.0/0	
sgr-0db780ca45977ceab	Tutto il traffico	Tutti	Tutti	Person... sg-03d48acb947288df5	

Aggiungi regola

(I added more rules to make sure Glue's 'Connection' worked.)

3.2 - Table Creation

- The tables my_btc and my_monero were created with an optimized structure for analytical queries:

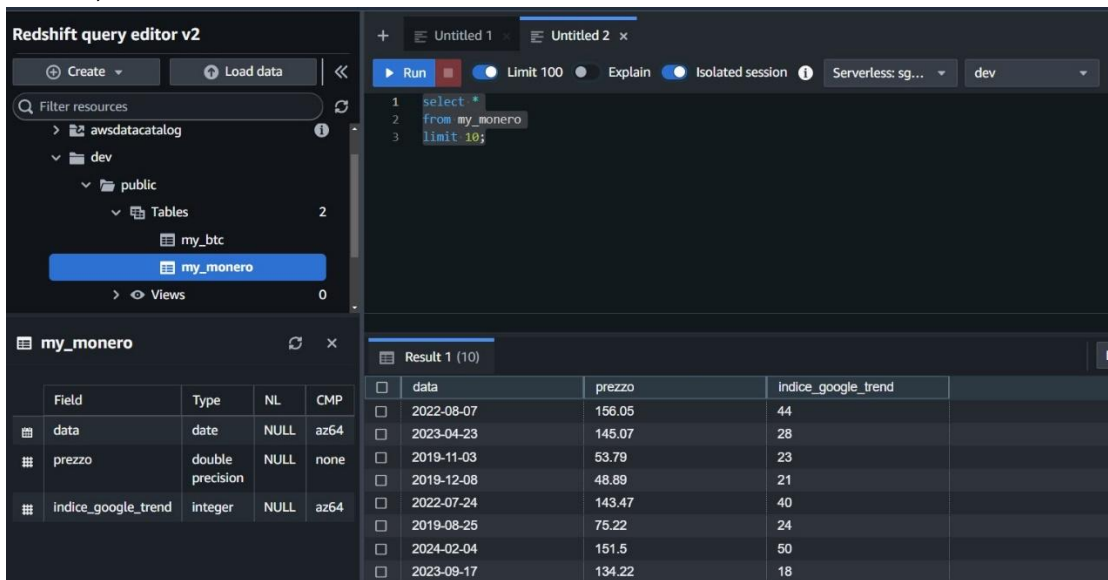
```
CREATE TABLE public.my_btc (  
  data date ENCODE az64,  
  prezzo double precision ENCODE raw,  
  indice_google_trend integer ENCODE az64  
) DISTSTYLE AUTO;
```

```
CREATE TABLE public.my_monero (
  data date ENCODE az64,
  prezzo double precision ENCODE raw,
  indice_google_trend integer ENCODE az64
) DISTSTYLE AUTO;
```

3.3 - Data Verification

- Basic verification queries were performed to ensure that the data were loaded correctly into the Redshift tables.

```
select *
from my_monero
limit 10;
```



Redshift query editor v2

Filter resources

awsdatacatalog

dev

public

Tables

my_btc

my_monero

Views

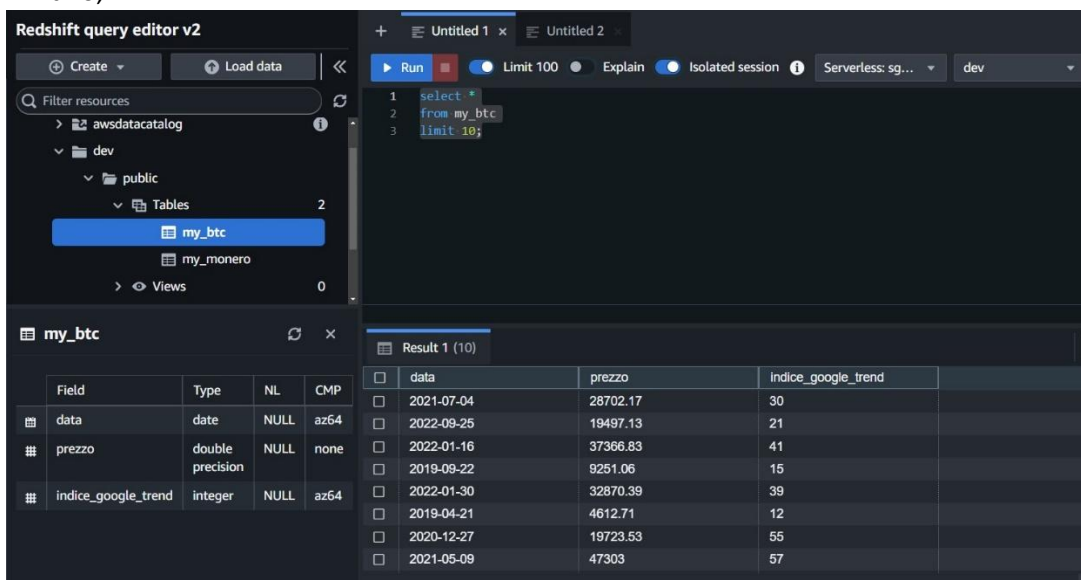
my_monero

Field	Type	NL	CMP
data	date	NULL	az64
prezzo	double precision	NULL	none
indice_google_trend	integer	NULL	az64

Result 1 (10)

data	prezzo	indice_google_trend
2022-08-07	156.05	44
2023-04-23	145.07	28
2019-11-03	53.79	23
2019-12-08	48.89	21
2022-07-24	143.47	40
2019-08-25	75.22	24
2024-02-04	151.5	50
2023-09-17	134.22	18

```
select *
from my_btc
limit 10;
```



Redshift query editor v2

Filter resources

awsdatacatalog

dev

public

Tables

my_btc

my_monero

Views

my_btc

Field	Type	NL	CMP
data	date	NULL	az64
prezzo	double precision	NULL	none
indice_google_trend	integer	NULL	az64

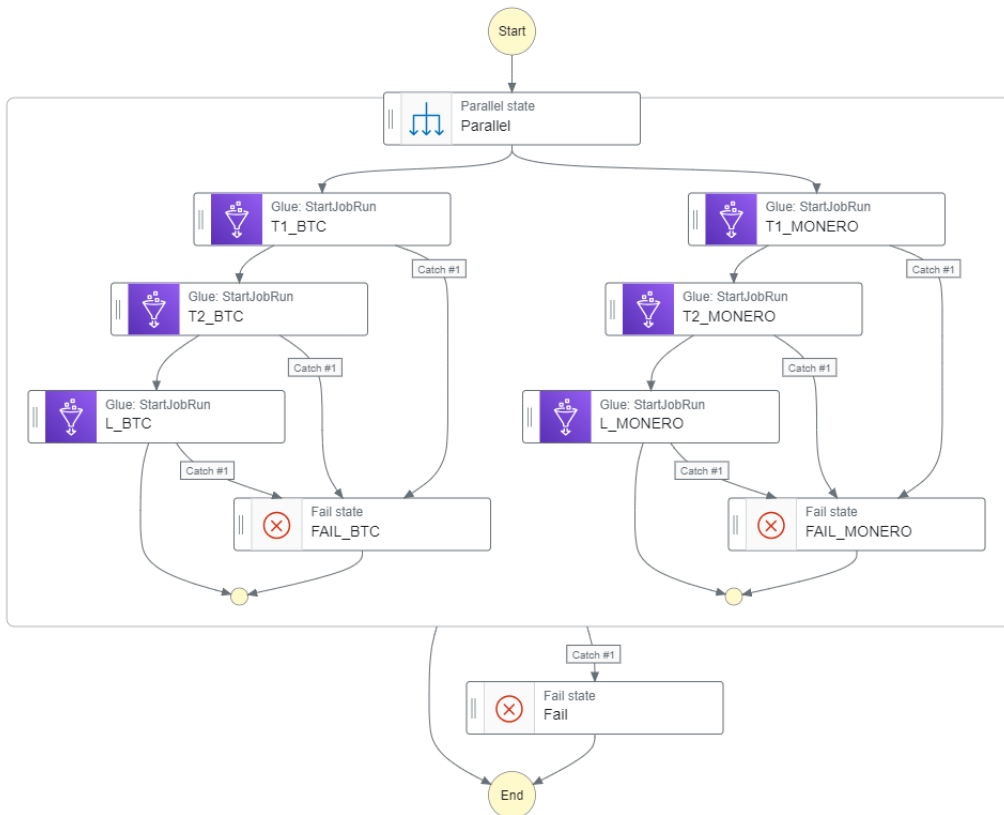
Result 1 (10)

data	prezzo	indice_google_trend
2021-07-04	28702.17	30
2022-09-25	19497.13	21
2022-01-16	37366.83	41
2019-09-22	9251.06	15
2022-01-30	32870.39	39
2019-04-21	4612.71	12
2020-12-27	19723.53	55
2021-05-09	47303	57

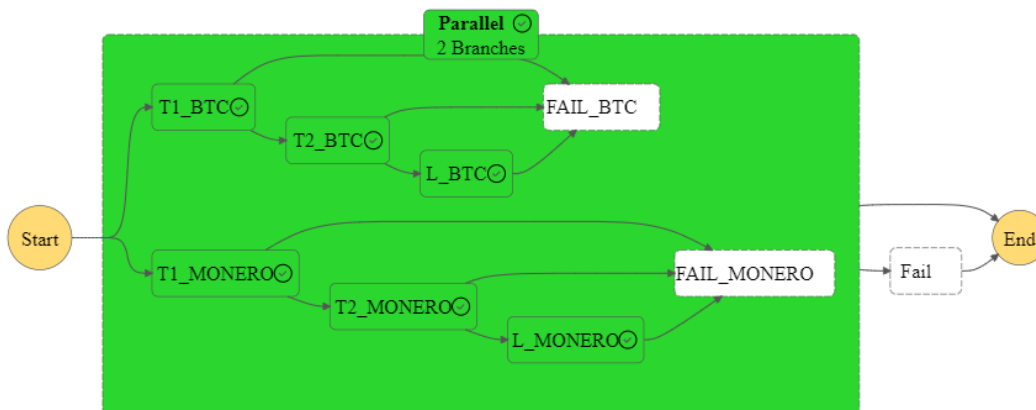
4 – Orchestration Step functions

4.1 – Workflow definition.

- A state machine was created using AWS Step Functions to orchestrate the sequential execution of AWS Glue scripts.
- The state definition included steps for the execution of cleanup, transformation and loading scripts.



- Execution has been initiated.



Vista tabella

	Nome	Tipo	Stato	Risorsa	Durata	Cronologia	Avviato dopo
<input checked="" type="radio"/>	Parallel	Parallel	✓ Riuscito	-	00:00:01.032	<div><div></div></div>	00:00:00.037
<input type="radio"/>	#0	ParallelBra...	✓ Riuscito	-	00:00:01.032	<div><div></div></div>	00:00:00.037
<input type="radio"/>	T1_BT	Task	✓ Riuscito	Blue job	00:00:00.375	<div><div></div></div>	00:00:00.037
<input type="radio"/>	T2_BT	Task	✓ Riuscito	Blue job	00:00:00.295	<div><div></div></div>	00:00:00.412
<input type="radio"/>	L_BTC	Task	✓ Riuscito	Blue job	00:00:00.362	<div><div></div></div>	00:00:00.707
<input type="radio"/>	#1	ParallelBra...	✓ Riuscito	-	00:00:00.962	<div><div></div></div>	00:00:00.037
<input type="radio"/>	T1_M	Task	✓ Riuscito	Blue job	00:00:00.375	<div><div></div></div>	00:00:00.037
<input type="radio"/>	T2_M	Task	✓ Riuscito	Blue job	00:00:00.295	<div><div></div></div>	00:00:00.412
<input type="radio"/>	L_MO	Task	✓ Riuscito	Blue job	00:00:00.292	<div><div></div></div>	00:00:00.707

5 - Viewing with Amazon QuickSight

5.1 - Account Configuration

- An Amazon QuickSight instance has been created and configured for visual data analysis.
- A connection to the Redshift database has been established to allow access to the transformed data.

5.2 - Creating Graphs.

- Charts have been created to explore the potential of QuickSight that can help interpret results and make data-driven decisions.

