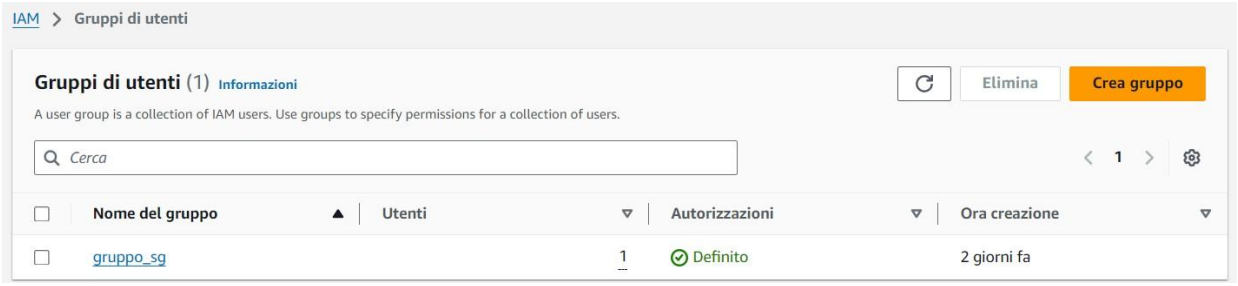
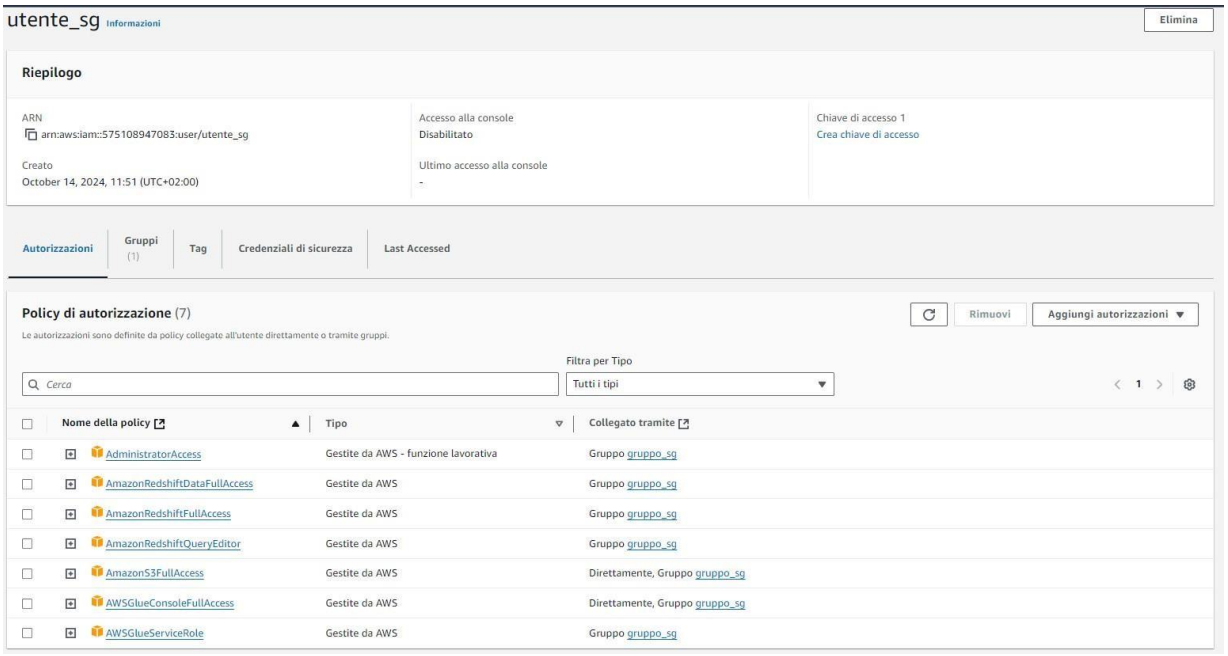
Indicazione su come è stata svolta l'esercitazione

1 - Preparazione dell'Ambiente

1.1 - Creazione dell'Account e Gruppo

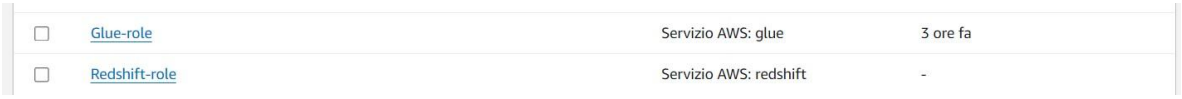
- È stato creato un account AWS e un gruppo di utenti specifico per gestire le risorse necessarie alla pipeline.
- Gli utenti del gruppo sono stati associati a tutte le policy amministrative necessarie per facilitare l'accesso e la gestione dei servizi AWS coinvolti.



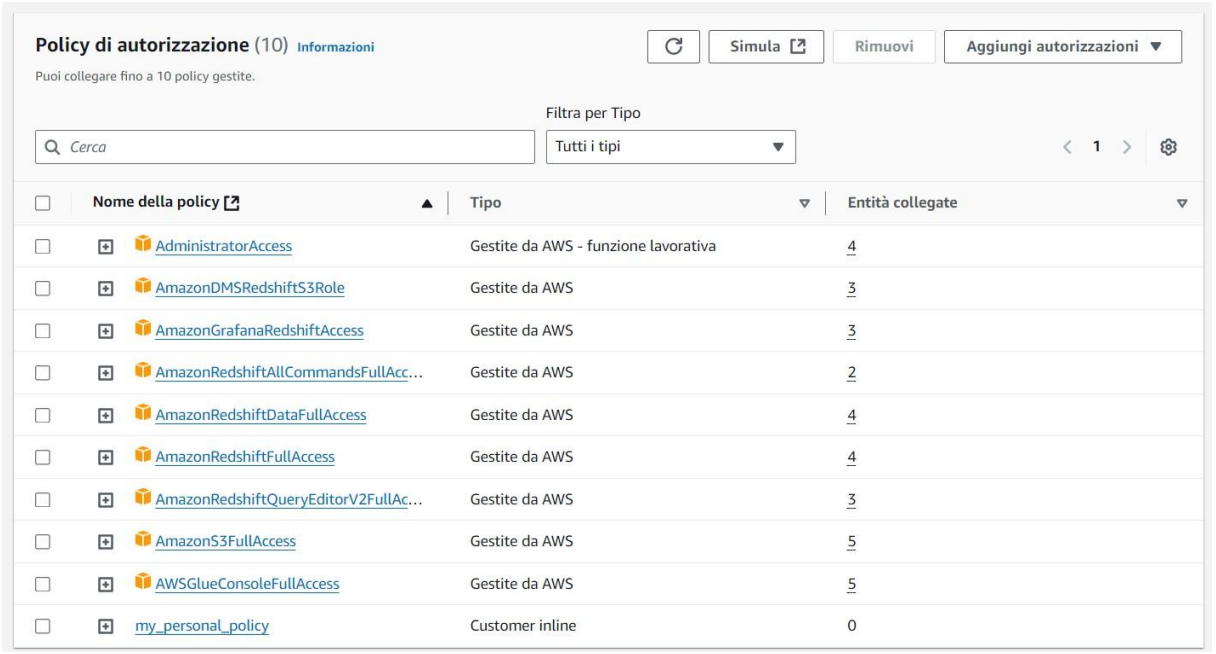


1.2 - Configurazione dei Ruoli

- Sono stati creati due ruoli specifici: uno per AWS Glue e uno per Amazon Redshift, entrambi con le policy necessarie per accedere alle risorse S3 e Redshift.



- È stata aggiunta una policy personalizzata per garantire la sicurezza e l'accesso appropriato alle risorse:



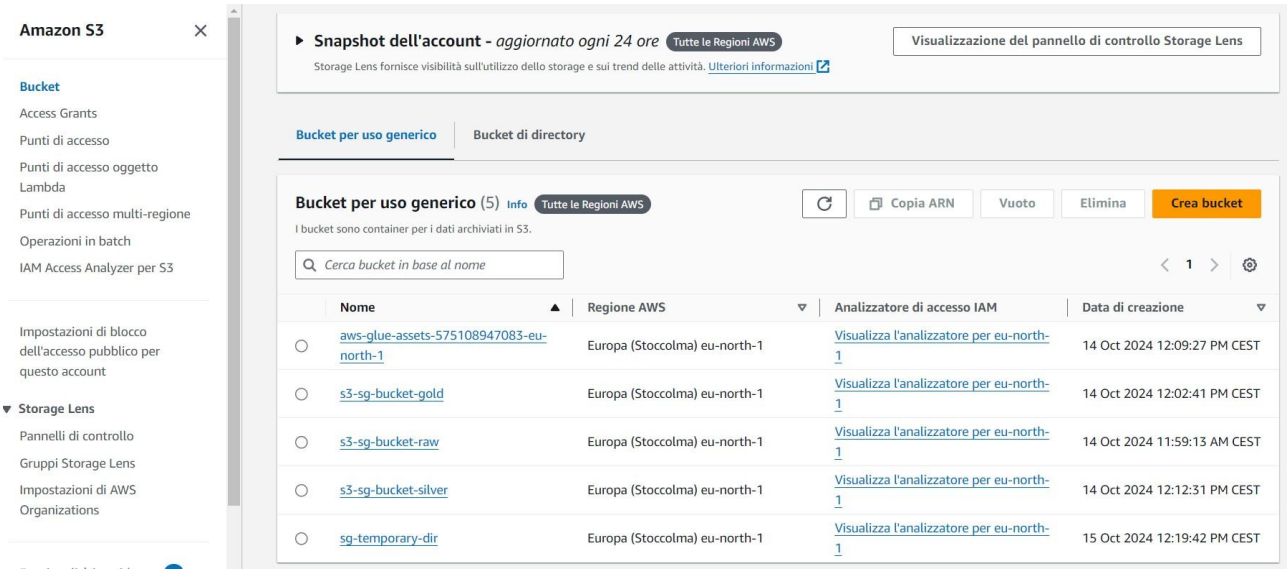
{

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

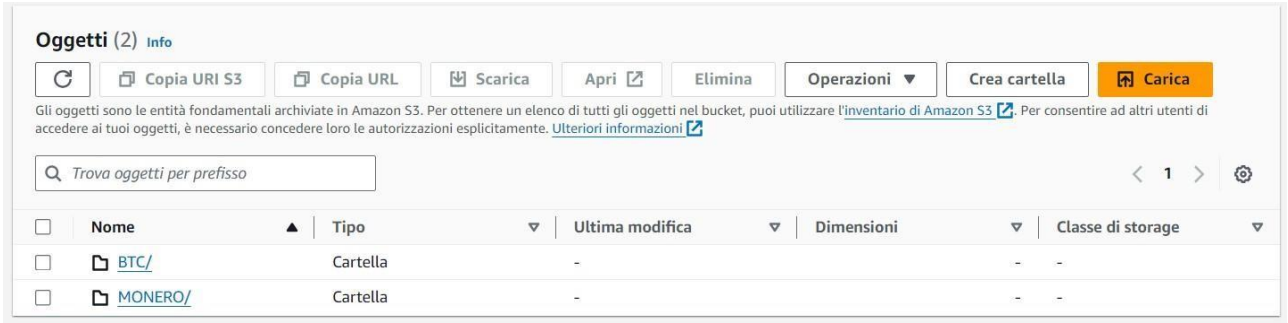
(Ridondante ma aggiunta per sicurezza, essendo la prima volta che mi interfaccio con AWS).

1.3 - Creazione dei Bucket S3

- Sono stati creati tre bucket: s3-sg-bucket-raw per i dati grezzi, s3-sg-bucket-silver per i dati puliti e trasformati, e s3-sg-bucket-gold per i dati finalizzati.



- All'interno di ogni bucket, sono state create 'sottocartelle' per BTC e MONERO per organizzare i dati in modo strutturato.



- I file necessari sono stati caricati manualmente nelle sottocartelle di s3-sg-bucket-raw.

2 - Implementazione della Pipeline con AWS Glue

2.1 - Script di Pulizia e Trasformazione

- Importato e configurato l'ambiente con Spark e Glue per elaborare i dati.
- I dati sono stati letti dal bucket S3 in formato CSV.
- Convertiti i dati in DataFrame per facilitare le operazioni di pulizia e trasformazione, gestendo i valori mancanti tramite media degli ultimi 5 giorni.
- Salvato il risultato pulito in formato Parquet su s3-sg-bucket-silver.

2.2 - Script T2: Calcolo Media Mobile e Join

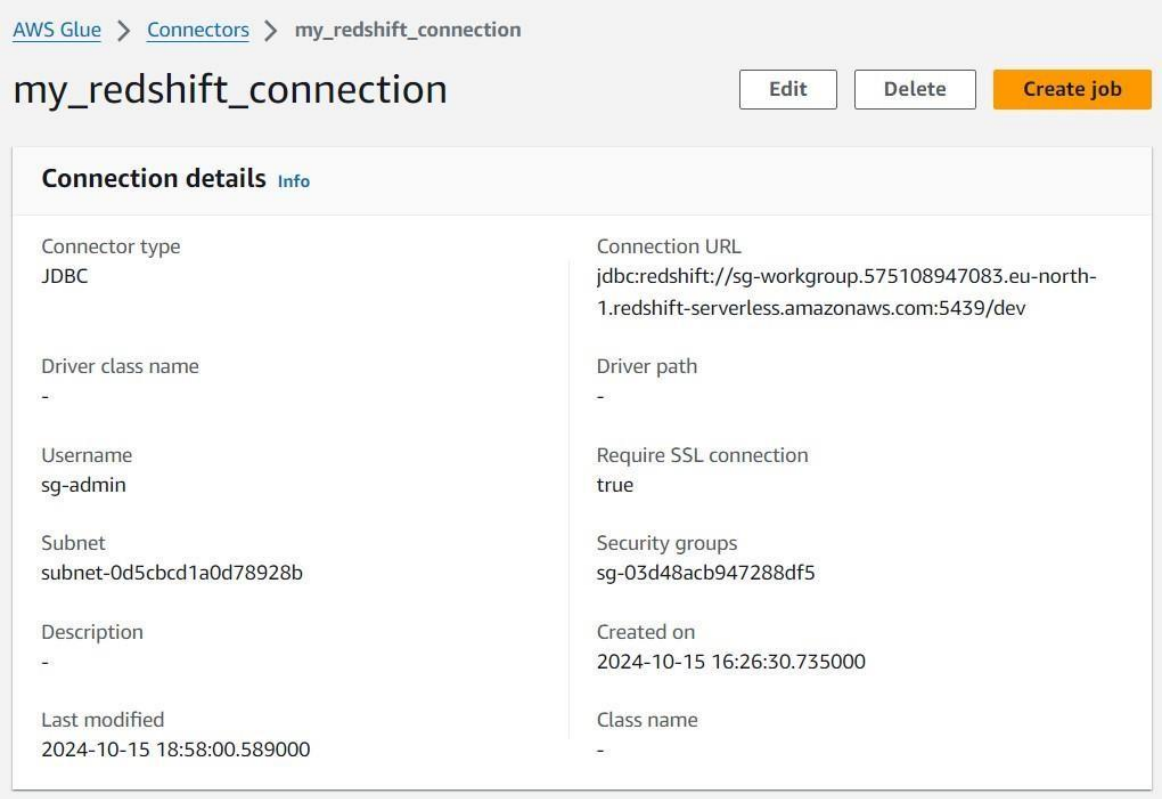
- I dati puliti sono stati letti dal bucket s3-sg-bucket-silver. ○ È stata calcolata la media mobile a 10 giorni per ridurre il rumore nei dati di prezzo.

- Eseguito il join con i dati di Google Trends per creare un dataset unificato. ○ Salvato il risultato finale in formato Parquet su s3-sg-bucket-gold.

2.3 - Script L: Caricamento su Redshift

○ I dati sono stati convertiti e mappati per l'inserimento nel database Redshift, assicurando la corretta tipizzazione.

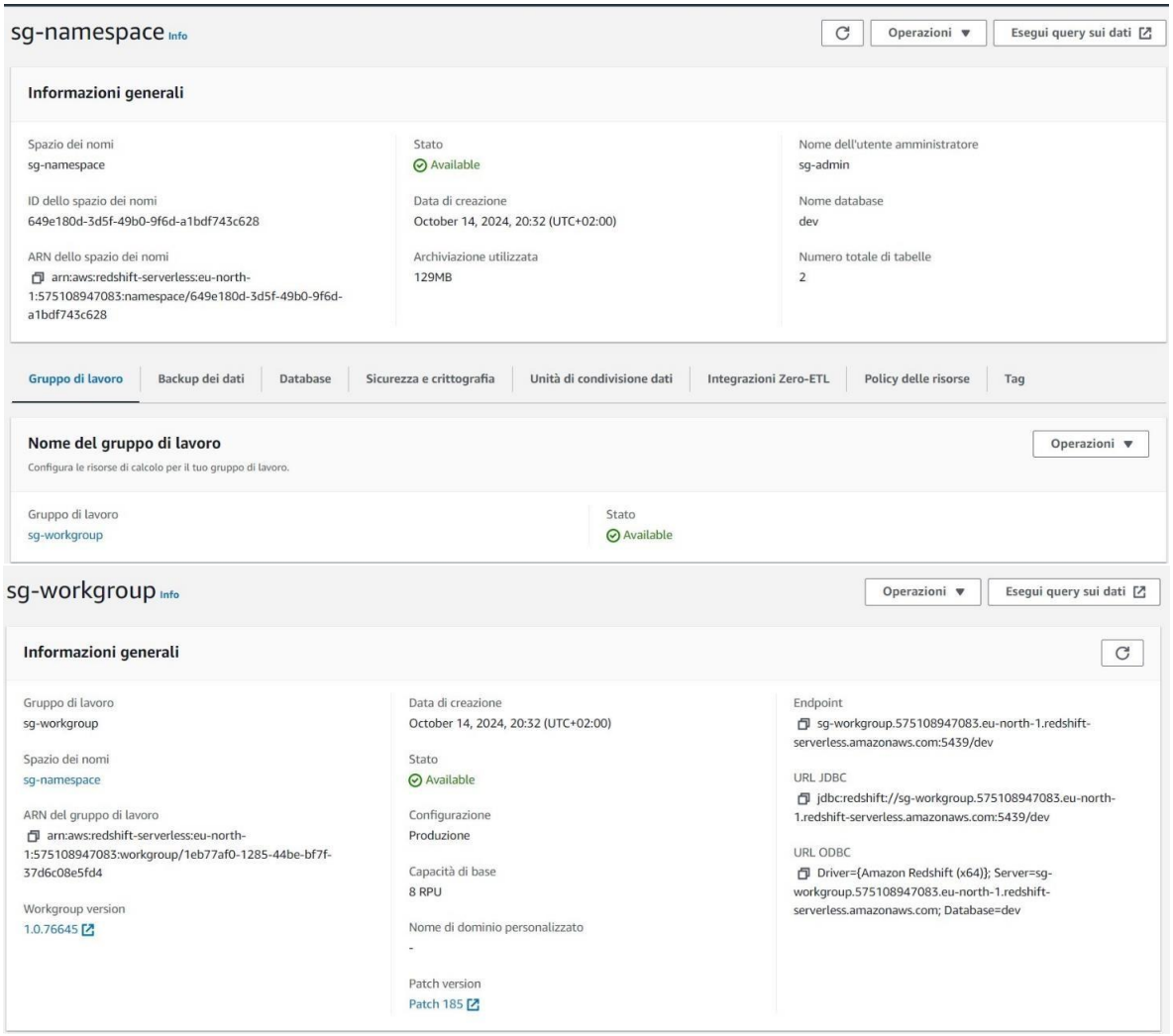
- Caricato il dataset finale su Redshift Serverless per consentire analisi future.
- Per effettuare il caricamento su Redshift è stata creata una 'Connection' in Glue riportando i dati della configurazione del namespace e del workgroup di Redshift (vedi dopo).



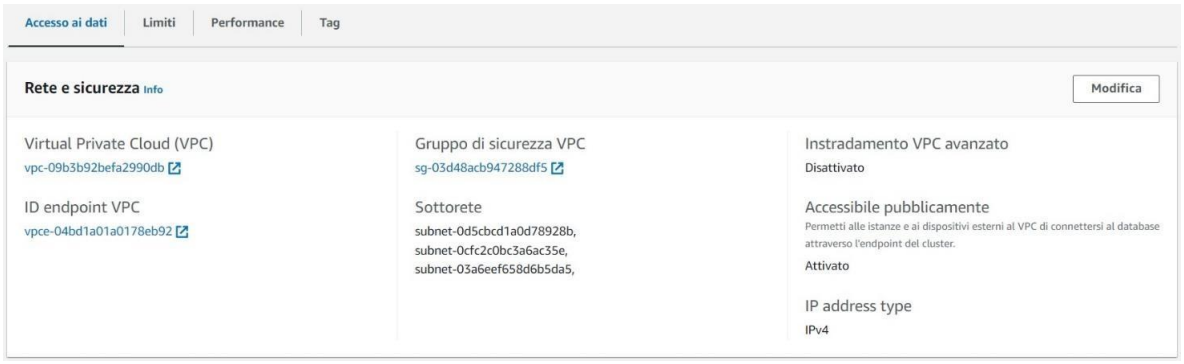
3 - Configurazione di Amazon Redshift Serverless

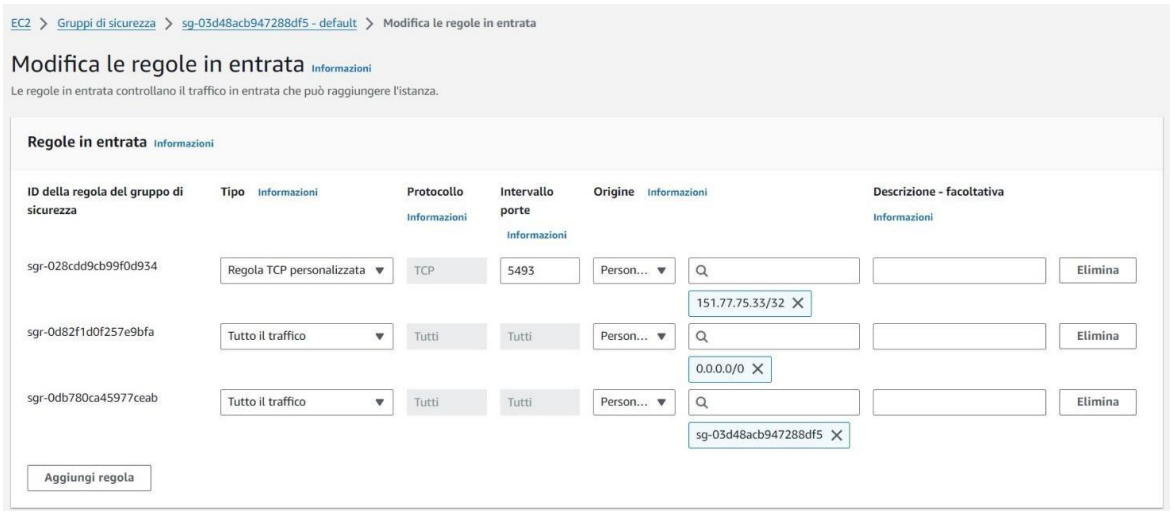
3.1 - Creazione del Namespace e Workgroup

○ Configurato il namespace e il workgroup di Redshift Serverless per gestire le risorse del database.



- Sono state impostate le regole di sicurezza per permettere l'accesso sicuro alle risorse Redshift.





(Ho aggiunto altre regole per essere sicuro che la 'Connection' di Glue funzionasse.)

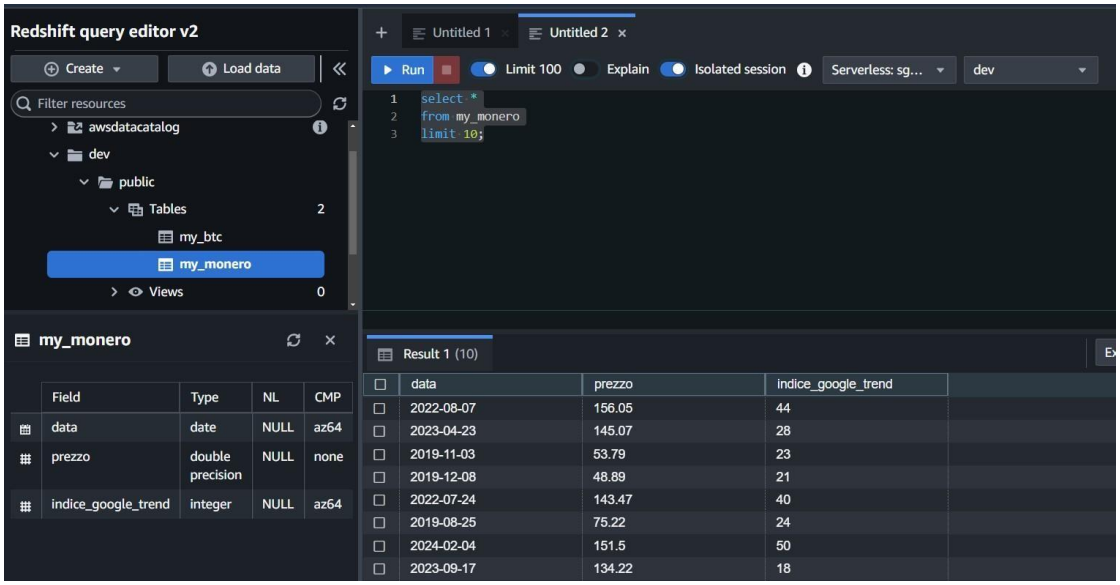
3.2 - Creazione delle Tabelle

- Sono state create le tabelle my_btc e my_monero con una struttura ottimizzata per le query analitiche:

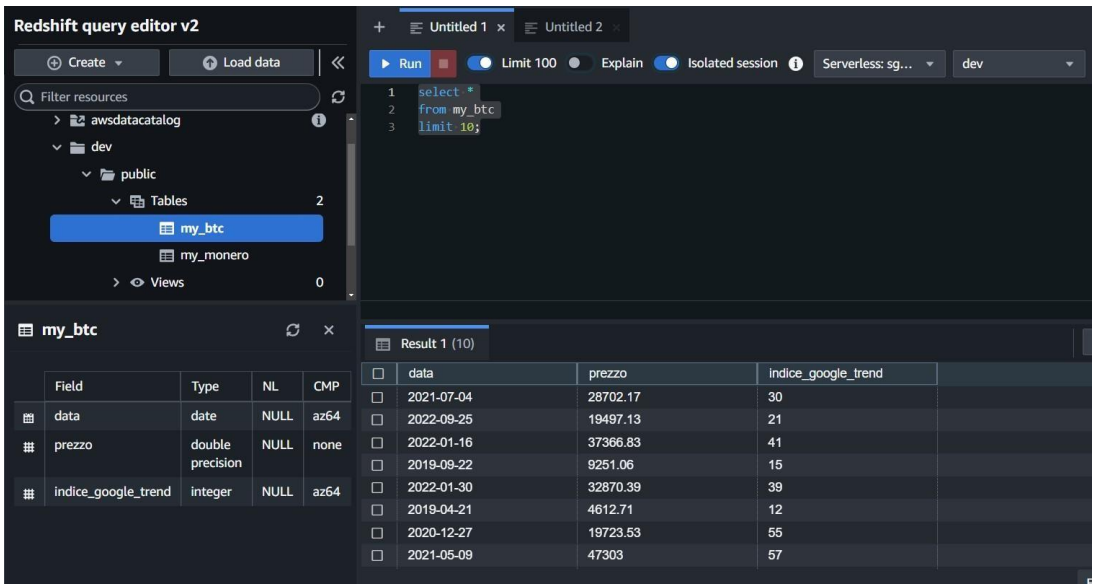
```
CREATE TABLE public.my_btc (
   data date ENCODE az64,
   prezzo double precision ENCODE raw,
   indice_google_trend integer ENCODE az64
) DISTSTYLE AUTO;

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

3.3 - Verifica dei Dati ○ Sono state eseguite query di verifica basiche per assicurarsi che i dati siano stati caricati correttamente nelle tabelle di Redshift.

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



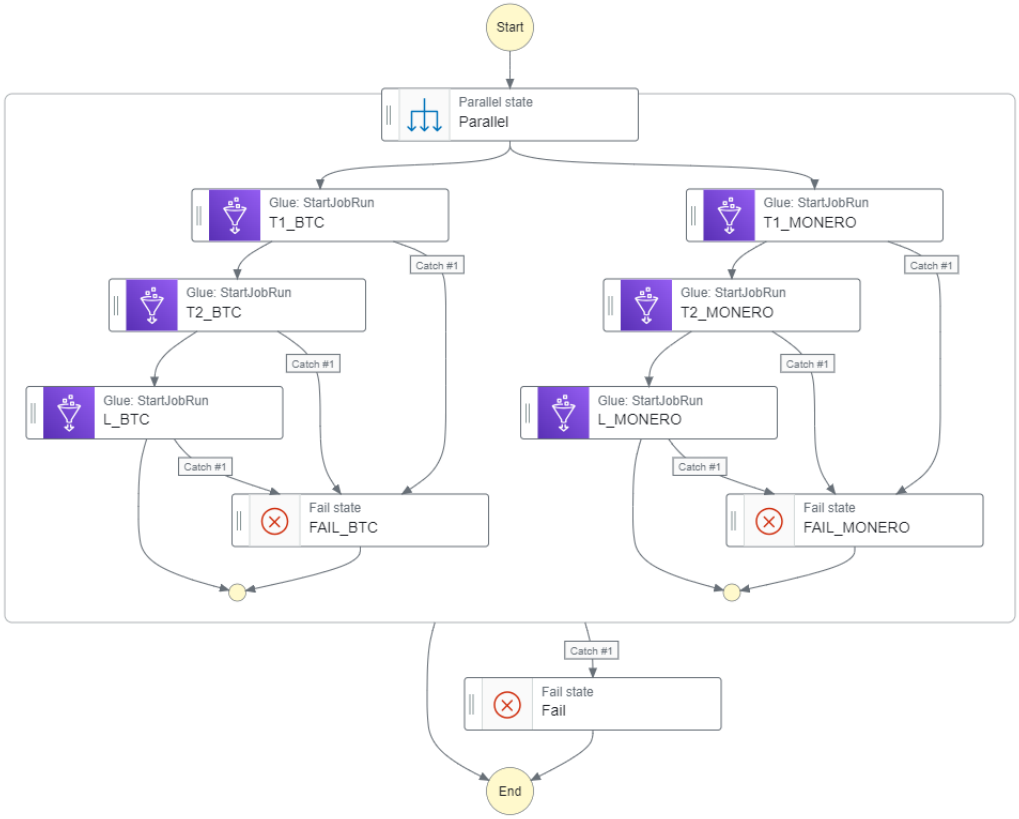
select *
from my_btc
limit 10;



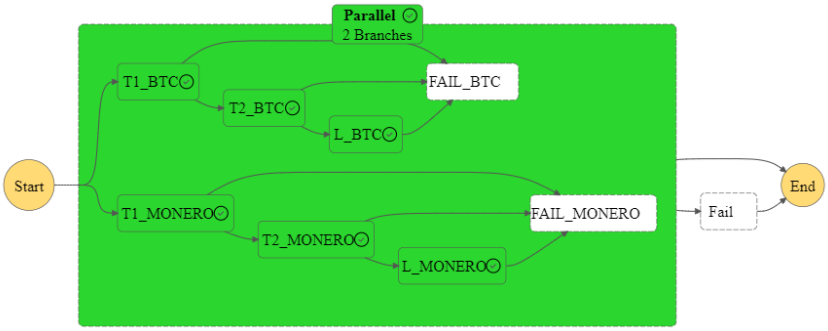
4 – Orchestrazione Step functions

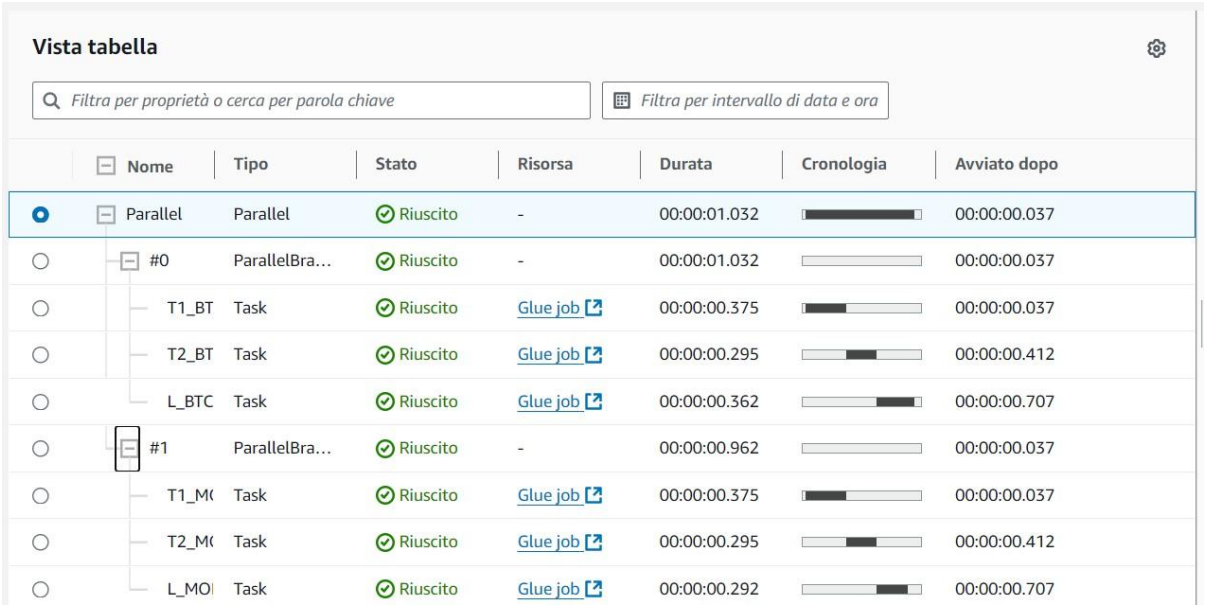
4.1 – Definizione del flusso di lavoro

- È stata creata una macchina a stati utilizzando AWS Step Functions per orchestrare l'esecuzione sequenziale degli script di AWS Glue.
- La definizione dello stato includeva passi per l'esecuzione degli script di pulizia, trasformazione e caricamento.



- E’ stata avviata l’esecuzione.





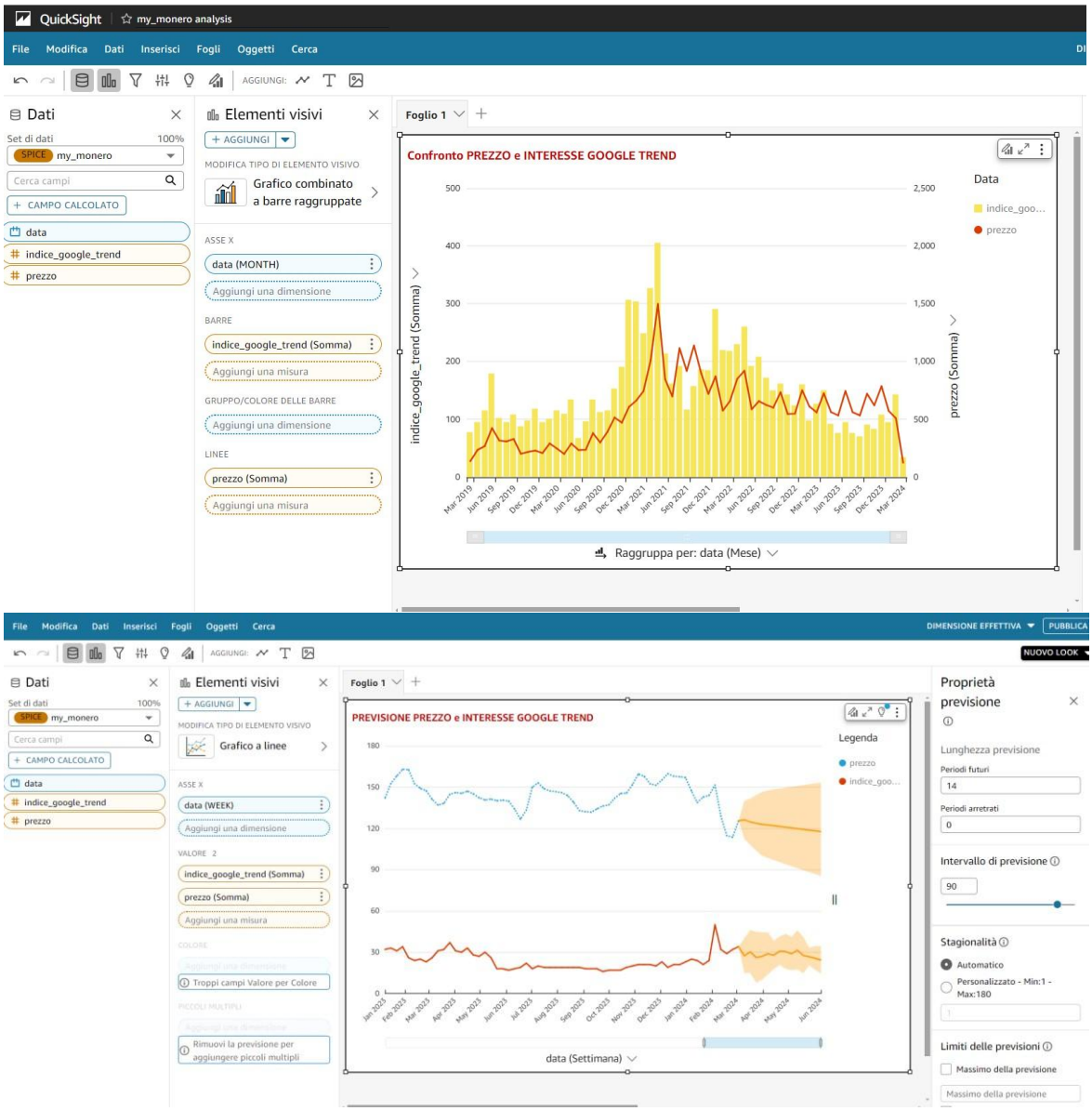
5 - Visualizzazione con Amazon QuickSight

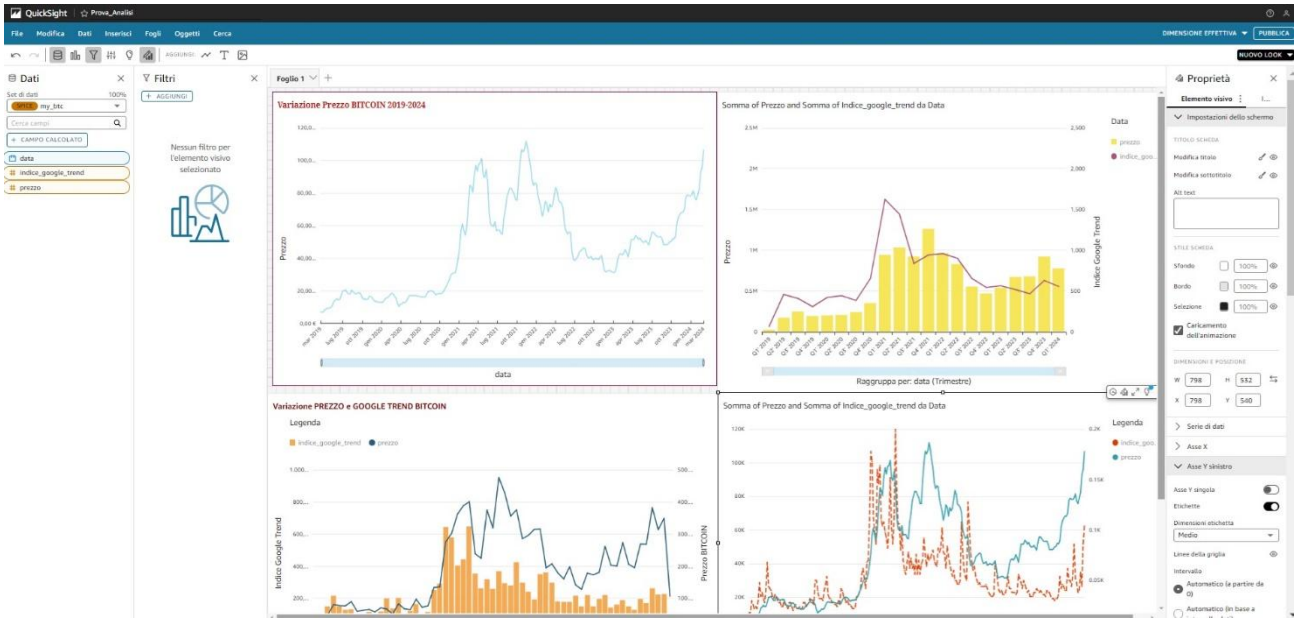
5.1 - Configurazione dell'Account

- È stata creata e configurata un'istanza di Amazon QuickSight per l'analisi visiva dei dati.
- È stata stabilita la connessione al database Redshift per consentire l'accesso ai dati trasformati.

5.2 - Creazione di Dashboard

- Sono stati creati grafici per esplorare le potenzialità di QuickSight che possono aiutare a interpretare i risultati e a prendere decisioni basate sui dati.




QuickSight
Prova_Analisi
File Modifica Dati Inserisci Fogli Oggetti Cerca
PUBBLICA
NUOVO LOOK
Dati
Filtri
AGGIUNGI
Nessun filtro per l'elemento visivo selezionato
Foglio 1
Variazione Prezzo BITCOIN 2019-2024
Prezzo
data
Somma of Prezzo and Somma of Indice_google_trend da Data
Data
Indice Google Trend
Raggruppa per: data (Trimestre)
Variazione PREZZO e GOOGLE TREND BITCOIN
Legenda
indice_google_trend
prezzo
Indice Google Trend
Prezzo BITCOIN
Somma of Prezzo and Somma of Indice_google_trend da Data
Legenda
Proprietà
Elemento visivo
Impostazioni dello schermo
Serie di dati
Asse X
Asse Y sinistro