

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.

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 >

<input type="checkbox"/>	Nome del gruppo	Utenti	Autorizzazioni	Ora creazione
<input type="checkbox"/>	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)

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

Cerca

Filtra per Tipo
Tutti i tipi

< 1 >

<input type="checkbox"/>	Nome della policy	Tipo	Collegato tramite
<input type="checkbox"/>	AdministratorAccess	Gestite da AWS - funzione lavorativa	Gruppo gruppo_sg
<input type="checkbox"/>	AmazonRedshiftDataFullAccess	Gestite da AWS	Gruppo gruppo_sg
<input type="checkbox"/>	AmazonRedshiftFullAccess	Gestite da AWS	Gruppo gruppo_sg
<input type="checkbox"/>	AmazonRedshiftQueryEditor	Gestite da AWS	Gruppo gruppo_sg
<input type="checkbox"/>	AmazonS3FullAccess	Gestite da AWS	Direttamente, Gruppo gruppo_sg
<input type="checkbox"/>	AWSGlueConsoleFullAccess	Gestite da AWS	Direttamente, Gruppo gruppo_sg
<input type="checkbox"/>	AWSGlueServiceRole	Gestite da AWS	Gruppo gruppo_sg

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.

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

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

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	<u>4</u>
<input type="checkbox"/>	AmazonDMSRedshiftS3Role	Gestite da AWS	<u>3</u>
<input type="checkbox"/>	AmazonGrafanaRedshiftAccess	Gestite da AWS	<u>3</u>
<input type="checkbox"/>	AmazonRedshiftAllCommandsFullAcc...	Gestite da AWS	<u>2</u>
<input type="checkbox"/>	AmazonRedshiftDataFullAccess	Gestite da AWS	<u>4</u>
<input type="checkbox"/>	AmazonRedshiftFullAccess	Gestite da AWS	<u>4</u>
<input type="checkbox"/>	AmazonRedshiftQueryEditorV2FullAc...	Gestite da AWS	<u>3</u>
<input type="checkbox"/>	AmazonS3FullAccess	Gestite da AWS	<u>5</u>
<input type="checkbox"/>	AWSGlueConsoleFullAccess	Gestite da AWS	<u>5</u>
<input type="checkbox"/>	my_personal_policy	Customer inline	0

```
{
  "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.

Amazon S3

Bucket

Access Grants

Punti di accesso

Punti di accesso oggetto

Lambda

Punti di accesso multi-regione

Operazioni in batch

IAM Access Analyzer per S3

Impostazioni di blocco dell'accesso pubblico per questo account

Storage Lens

Pannelli di controllo

Gruppi Storage Lens

Impostazioni di AWS Organizations

Snapshot dell'account - aggiornato ogni 24 ore

Tutte le Regioni AWS

Visualizzazione del pannello di controllo Storage Lens

Storage Lens fornisce visibilità sull'utilizzo dello storage e sui trend delle attività. [Ulteriori informazioni](#)

Bucket per uso generico

Bucket di directory

Bucket per uso generico (5)

Info

Tutte le Regioni AWS

↻

Copia ARN

Vuoto

Elimina

Crea bucket

I bucket sono container per i dati archiviati in S3.

Q Cerca bucket in base al nome

< 1 > ⚙

	Nome	Regione AWS	Analizzatore di accesso IAM	Data di creazione
<input type="radio"/>	aws-glue-assets-575108947083-eu-north-1	Europa (Stoccolma) eu-north-1	Visualizza l'analizzatore per eu-north-1	14 Oct 2024 12:09:27 PM CEST
<input type="radio"/>	s3-sg-bucket-gold	Europa (Stoccolma) eu-north-1	Visualizza l'analizzatore per eu-north-1	14 Oct 2024 12:02:41 PM CEST
<input type="radio"/>	s3-sg-bucket-raw	Europa (Stoccolma) eu-north-1	Visualizza l'analizzatore per eu-north-1	14 Oct 2024 11:59:13 AM CEST
<input type="radio"/>	s3-sg-bucket-silver	Europa (Stoccolma) eu-north-1	Visualizza l'analizzatore per eu-north-1	14 Oct 2024 12:12:31 PM CEST
<input type="radio"/>	sg-temporary-dir	Europa (Stoccolma) eu-north-1	Visualizza l'analizzatore per eu-north-1	15 Oct 2024 12:19:42 PM CEST

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

Oggetti (2) Info

↻

Copia URI S3

Copia URL

Scarica

Apri

Elimina

Operazioni ▼

Crea cartella

Carica

Gli oggetti sono le entità fondamentali archiviate in Amazon S3. Per ottenere un elenco di tutti gli oggetti nel bucket, puoi utilizzare [l'inventario di Amazon S3](#). Per consentire ad altri utenti di accedere ai tuoi oggetti, è necessario concedere loro le autorizzazioni esplicitamente. [Ulteriori informazioni](#)

Q Trova oggetti per prefisso

< 1 > ⚙

<input type="checkbox"/>	Nome	Tipo	Ultima modifica	Dimensioni	Classe di storage
<input type="checkbox"/>	BTC/	Cartella	-	-	-
<input type="checkbox"/>	MONERO/	Cartella	-	-	-

- 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).

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 - 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.

sg-namespace

Info

Operazioni

Esegui query sui dati

Informazioni generali

Spazio dei nomi

sg-namespace

ID dello spazio dei nomi

649e180d-3d5f-49b0-9f6d-a1bdf743c628

ARN dello spazio dei nomi

arn:aws:redshift-serverless:eu-north-1:575108947083:namespace/649e180d-3d5f-49b0-9f6d-a1bdf743c628

Stato

Available

Data di creazione

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

Archiviazione utilizzata

129MB

Nome dell'utente amministratore

sg-admin

Nome database

dev

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

Operazioni

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

Gruppo di lavoro

sg-workgroup

Stato

Available

sg-workgroup

Info

Operazioni

Esegui 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

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

Accesso ai dati | Limiti | Performance | Tag

Rete e sicurezza [info](#) Modifica

Virtual Private Cloud (VPC)

[vpc-09b3b92bfa2990db](#)

ID endpoint VPC

[vpce-04bd1a01a0178eb92](#)

Gruppo di sicurezza VPC

[sg-03d48acb947288df5](#)

Sottorete

subnet-0d5cbcd1a0d78928b,
subnet-0cfc2c0bc3a6ac35e,
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 [Informazioni](#)

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

Regole in entrata [Informazioni](#)

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

Aggiungi regola

(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;
```

Redshift query editor v2

Create Load data

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_monero
limit 10;
```

Redshift query editor v2

Create Load data

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

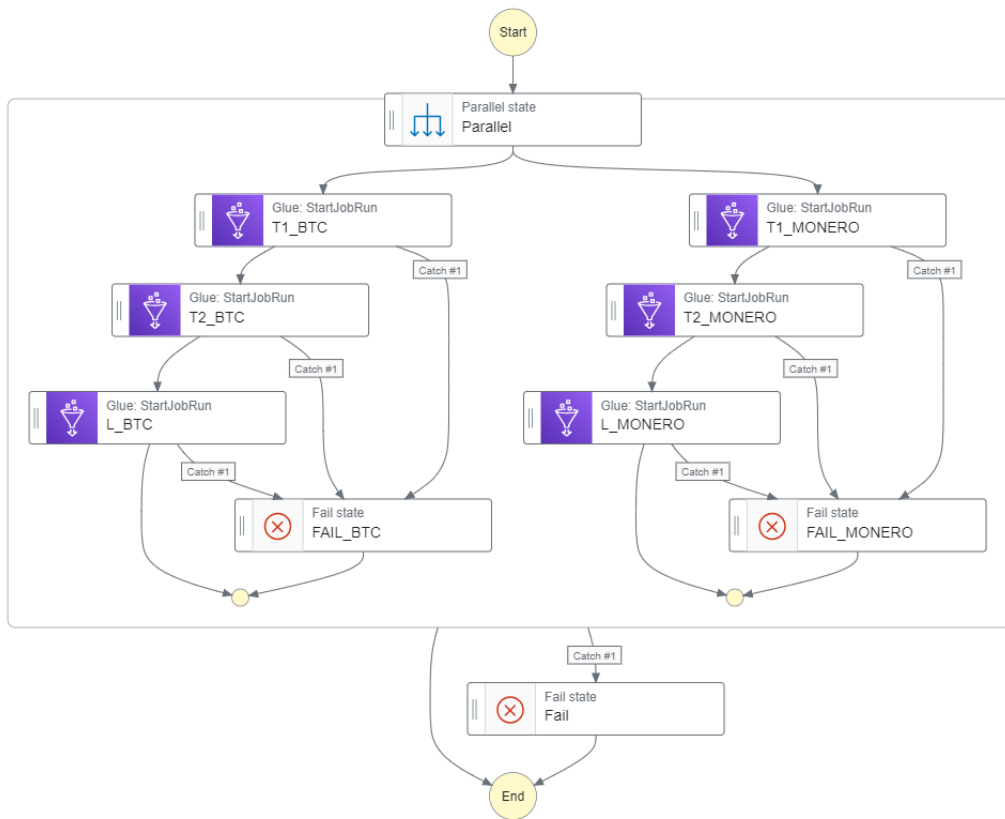
Result 1 (10)

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

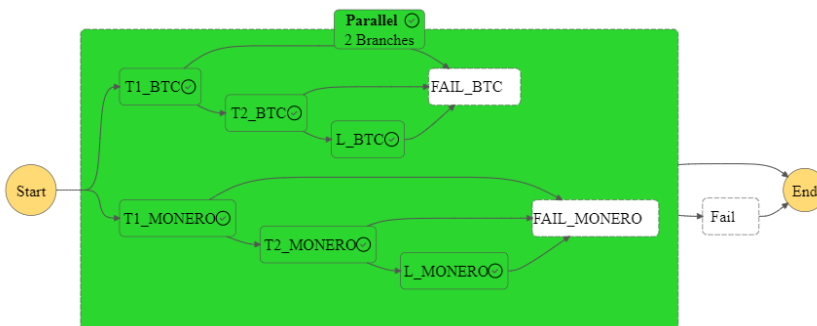
4 – Orchestrazione Step functions

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.



Vista tabella

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

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.

