Una pipeline E2E AWS per analizzare BTC e XMR

Il progetto ha l'obiettivo di sviluppare una pipeline E2E per la trasformazione di dati grezzi in tabelle ottimizzate su Amazon Redshift, focalizzandosi su due criptovalute: Monero (XMR) e Bitcoin (BTC). Questi dati, suddivisi in quattro file relativi ai prezzi e ai trend di Google, devono essere caricati dagli studenti su Amazon S3 per avviare la pipeline. Il processo culmina con la creazione di due tabelle Redshift, pronte per analisi avanzate e visualizzazioni.

Fasi della Pipeline

1. Caricamento su S3: Gli studenti devono inizialmente caricare i quattro file grezzi nel bucket "raw" su Amazon S3.
2. Estrazione e Pulizia: I dati vengono estratti, puliti e preparati per l'analisi.
3. Caricamento su Redshift: I dati puliti sono caricati in tabelle su Amazon Redshift, ottimizzate per l'analisi.
4. Visualizzazione (Opzionale): L'utilizzo di Amazon QuickSight per la creazione di dashboard e visualizzazioni è facoltativo ma consigliato per esplorare i dati trasformati.

I files
Come detto abbiamo una coppia di files per ogni crypto valuta. Le due cryptovalute e le relative pipeline possono essere considerate indipendenti, e possono dunque procedere in parallelo generando una tabella finale ciascuno. Se lo studente, lo desidera, può anche, a valle, joinare le due tabelle per ulteriori analisi, ma questo non è parte di questo progetto.
Prendendo Bitcoin come esempio abbiamo la seguente coppia di files.

File di prezzo (BTC/EUR).
Questo file CSV, ha diverse colonne ma quelle interessanti, per questo progetto, sono le prime due:

- ‘Date’, una stringa di formato (esempio) “03/12/2024”
- ‘Price’ un numero (appunto il prezzo), esempio: 145.4

Occorre notare che alcuni valori del prezzo sono mancanti (hanno valore -1), quindi lo studente deve cercare di eseguire alcune azioni su di essi, come, per esempio:

- Eliminare totalmente la riga

- Fillare il valore con il precedente
- Fillare il valore con la media/mediana dei 5 precedenti

Ovviamente, le azioni di cui sopra, vanno eseguite durante la pipeline.

File di Google trend.
Questo file CSV, ha due sole colonne:

- Settimana: una colonna che raffigura l'indice temporale per il valore interesse.
- Interesse bitcoin: un valore intero, che va fra 0 e 100, dove 100 significa massimo interesse, in accordo alle ricerche Google degli utenti di tutto il mondo.

Il file di Google trend, ha una granularità minore di quello di prezzo, il quale è giornaliero. In fase di Join, dunque alcuni valori del prezzo saranno persi. Si consiglia quindi di smussare il prezzo stesso mediante una media mobile.

Struttura della pipeline

Il nostro progetto si articola in due pipeline distinte, dedicate rispettivamente a Bitcoin e Monero. Queste pipeline sono progettate per essere eseguite in parallelo, massimizzando l'efficienza del processo.

Preparazione dei Dati

I dati di partenza sono situati su Amazon S3, all'interno di un bucket denominato "raw". Il primo step di ciascuna pipeline consiste nella pulizia dei file relativi ai Google trend e ai prezzi delle due criptovalute. Una volta puliti, i dati sono salvati in formato Parquet su un nuovo bucket S3, identificato come "argento", preparandoli per l'analisi successiva.

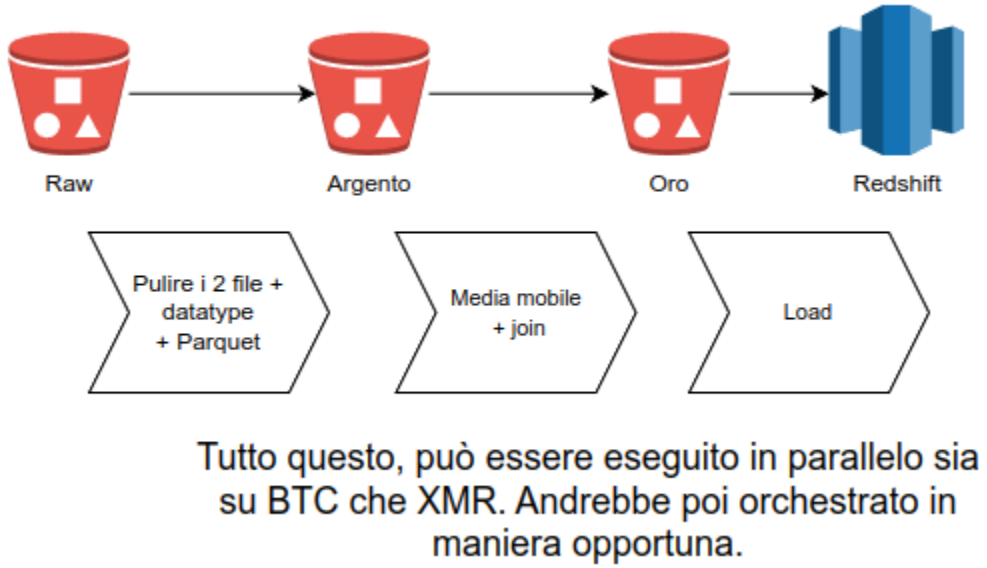
Analisi e Unificazione dei Dati

La seconda fase del processo prevede la lettura degli artefatti puliti per calcolare la media mobile a 10 giorni dei prezzi, con lo scopo di minimizzare il rumore. Successivamente, avviene il join tra i dataset di prezzo e Google trend, risultando in un unico file strutturato come segue: data, prezzo, indice_google_trend.

Caricamento su Redshift e Visualizzazione

Infine, un'ulteriore pipeline trasferisce il file unificato su Amazon Redshift, rendendo i dati pronti per l'analisi approfondita. Come passaggio opzionale, gli utenti possono sfruttare Amazon QuickSight per visualizzare e interpretare i risultati ottenuti, facilitando l'esplorazione dei dati e la condivisione degli insight.

Qui una raffigurazione visuale della pipeline:



Per quanto riguarda le trasformazioni dei dati nelle pipeline di questo progetto, lo studente ha la possibilità di scegliere tra due servizi AWS: **GLUE ETL e EMR**. Ogni servizio offre specifici vantaggi e si adatta a diverse esigenze di elaborazione dei dati.

Una volta selezionato il servizio per le trasformazioni, è essenziale prestare attenzione all'orchestrazione del processo. **AWS Step Functions** si rivela uno strumento cruciale in questa fase, consentendo di coordinare le attività e garantire che le due pipeline possano eseguire in parallelo senza intoppi. Utilizzando Step Functions, lo studente può definire facilmente il flusso di lavoro e monitorare lo stato di esecuzione delle pipeline, assicurandosi che ogni passo sia completato come previsto.