Guida operativa e note metodologiche per la replicazione del progetto

Dati storici

- Elon musk tweets

  Dati reperibili al link: https://data.world/barbaramaseda/elon-musk-tweets . Questo sito raccoglie tutti i tweets di Elon Musk raccolti in tempo reale con le applet IFTTT. Una volta scaricato il dataset è risultato necessario pulirlo e filtrare specifiche parole chiave nel testo per mantenere solamente i dati di nostro interesse. Lo script che svolge questo compito è nominato "PutMongoElonTweets.ipynb" (cartella "script"). Successivamente i dati sono stati caricati sul database MongoDB.

- Dati storici dogecoin

  Per ottenere le serie storiche di dogecoin abbiamo utilizzato le API di Coingeko (documentazione: https://www.coingecko.com/en/api). Tramite invocazione get https è stato creato un dataset in locale contenente i dati nel periodo 1/9/2020 - 1/6/2021. Anche in questo caso i dati sono stati puliti e gestiti prima di venire caricati su MongoDB. Lo script responsabile di questo processo si chiama "CoinGekoHistorical". Il dataset è stato inoltre arricchito con una funzione di calcolo per ottenere la variazione percentuale di prezzo giornaliera (non fornita direttamente dalle API)

L'integrazione dei due dataset è stata svolta con un join sull'attributo data, in modo da avere per ogni giorno il prezzo di doge e i relativi tweets.

Dati veloci

Tutte le varie componenti della nostra architettura sono state implementate in locale data la natura simulativa del progetto. Con semplici modifiche comunque sarebbe possibile distribuire su macchine diverse le varie componenti.
Nello specifico, le componenti installate sono state:

- Apache Zookeeper
- Apache Kafka (porta 9092)
- Apache NiFi (porta 9090 per non andare in conflitto con Zookeeper)
- MongoDB

Di seguito la guida operativa per l'avvio della pipeline di estrazione, storage e integrazione dei dati:

1. Avvio di Zookeeper, Kafka, MongoDB e NiFi.
2. Creazione delle collezioni "Coingecko", "Twitter" e "CoinTweetData" su MongoDB.
3. Caricamento del template "Estrazione dati (Template NiFi).xml" (presente nella cartella "Script" nella cartella condivisa) su NiFi.
4. Avvio di tutti i nodi nel workflow NiFi.
5. Scheduling dello script "Aggregator.ipynb" in modo che venga eseguito una volta al giorno per l'integrazione dei dati raw. (Lo script integra tutti i dati del giorno

precedente al momento del suo avvio, per eseguire dei test è possibile modificare la variabile “backDays”, per esempio impostandola a zero è possibile integrare i dati del giorno corrente). Lo scheduling può essere eseguito in diversi modi, ad esempio attraverso il semplice Task Scheduler di Windows.

6. Tutti i dati prodotti dal processo descritto sopra saranno visionabili nelle collezioni create al punto 2 su MongoDB.