Hard Disk Failure Data Processing

Corso di Sistemi ed Architetture per Big Data

Luca Falasca **0334722**

Matteo Conti **0323728**

Progetto 1 - Batch processing

2024

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Indice

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Introduzione

Introduzione-Obbiettivi

Il progetto verte sull'analisi di un dataset contenente dati riguardanti il monitoraggio di dischi rigidi installati all'interno di un cluster di server gestito da un cloud provider, in particolare si vuole:

- Realizzare una pipeline di elaborazione dati
- Eseguire le query richieste dalla specifica
- Visualizzare i risultati
- Analizzare le performance ottenute con i formati dati CSV e Parquet

Introduzione-Dataset

Il dataset fornito è una versione ridotta di quello presentato nel Grand Challenge della conferenza ACM DEBS 2024. Delle numerose colonne presenti nel dataset, ne verranno selezionate solamente cinque, in particolare:

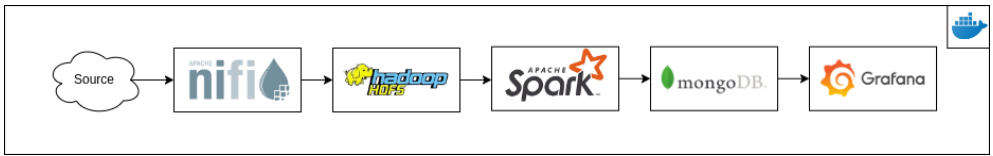
- *date*: data della misurazione nel formato 'YYYY-MM-DD'
- *serial_number*: identificativo del disco rigido
- *model*: modello del disco rigido
- *s9_power_on_hours*: tempo di accensione del disco rigido in ore
- *vault_id*: identificativo del gruppo di storage server a cui il disco rigido appartiene
- *failure*: flag che indica se il disco rigido ha subito una failure o meno

Pipeline

Pipeline

La pipeline di elaborazione dati è stata containerizzata e deployata tramite docker compose ed è composta da cinque componenti:

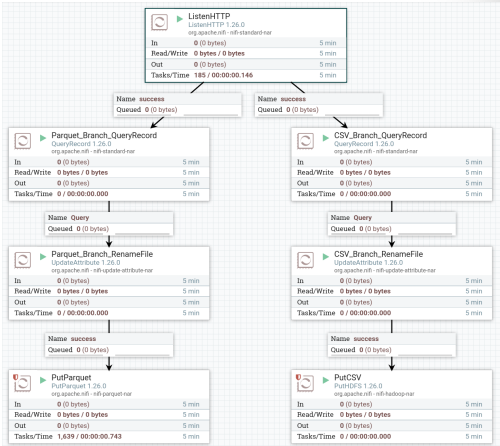
- Data ingestion
- Data storage
- Data processing
- Analytical data storage
- Data visualization



Pipeline - Data ingestion

Per la data ingestion è stato utilizzato il framework Apache NiFi, il quale si occupa di:

- Ricevere tramite HTTP il file CSV contenente i dati
- Filtrare i dati
- Memorizzare i dati su HDFS in formato Parquet e CSV



Conclusione

Clean layout and two-column text

This is a text in first column.

$$E = mc^2$$

$$1 + 2 + \cdots + k = \frac{k \cdot (k+1)}{2}.$$

- First item
- Second item

This text will be in the second column and on a second tought this is a nice looking layout in some cases.

1. First
2. Second

Sample frame title

In this slide, some important text will be **highlighted** because it's important. Please, don't abuse it.

Remark
Sample text

Important theorem
Sample text in alert box

Examples 1
Sample text in green box. The title of the block is "Examples".

Preliminary Empirical Study

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2
  - text subitem

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2
  - text subitem
- Text visible on slides 3

Sample frame title

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1
- Text visible on slide 2
  - text subitem

- Text visible on slide 4

Hard Disk Failure Data Processing

Corso di Sistemi ed Architetture per Big Data

Luca Falasca **0334722**

Matteo Conti **0323728**

Progetto 1 - Batch processing

2024

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA