Analisi di dati di monitoraggio con Apache Spark

2023-2024

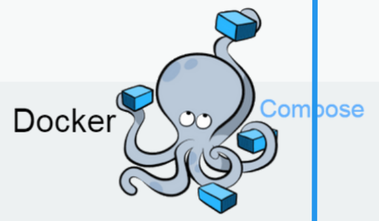
Marina Calcaura, Valerio Brauzi

Dataset di riferimento

date	serial_number	model	failure	vault_id	s9 power on hours	...
2023-04-01T00:00:00.000000	8HK2SSMH	HGST HUH721212ALN604	0	1113	38445.0	...
2023-04-01T00:00:00.000000	10B0A01UF97G	TOSHIBA MG07ACA14TA	0	1067	27425.0	...
2023-04-01T00:00:00.000000	5080A117F97G	TOSHIBA MG07ACA14TA	0	1095	18029.0	...
2023-04-01T00:00:00.000000	ZL2NG0QS	ST16000NM001G	0	2010	6059.0	...
...	...	...	...	...	...	...

Dati forniti da Backblaze

Architettura



Data-acquisition & Data-ingestion

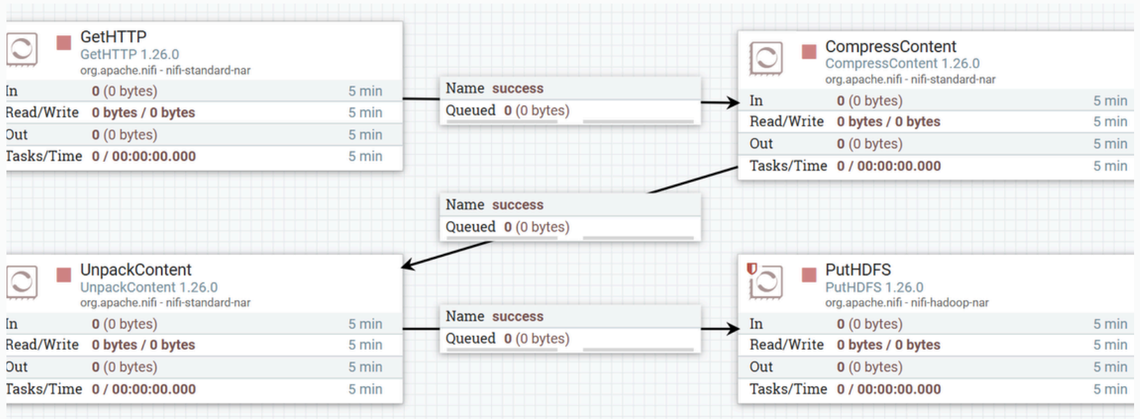


figura1: template per il Download e la Conversione dei dati

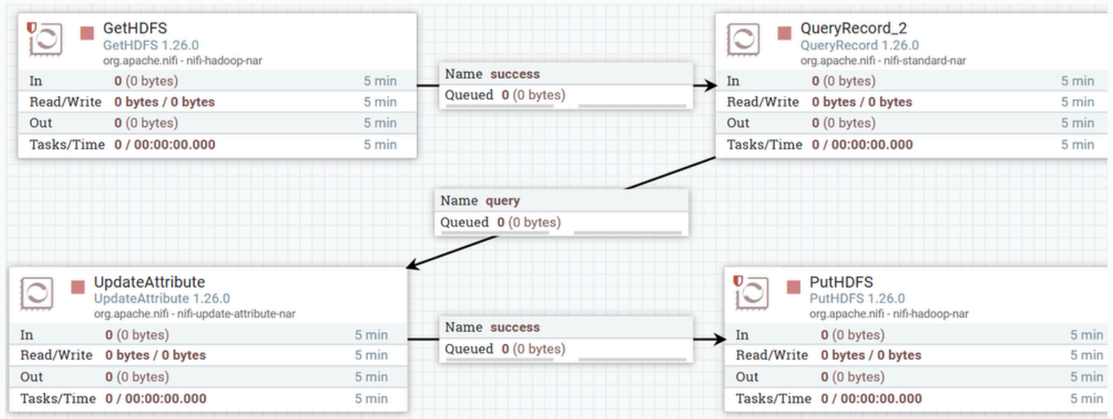
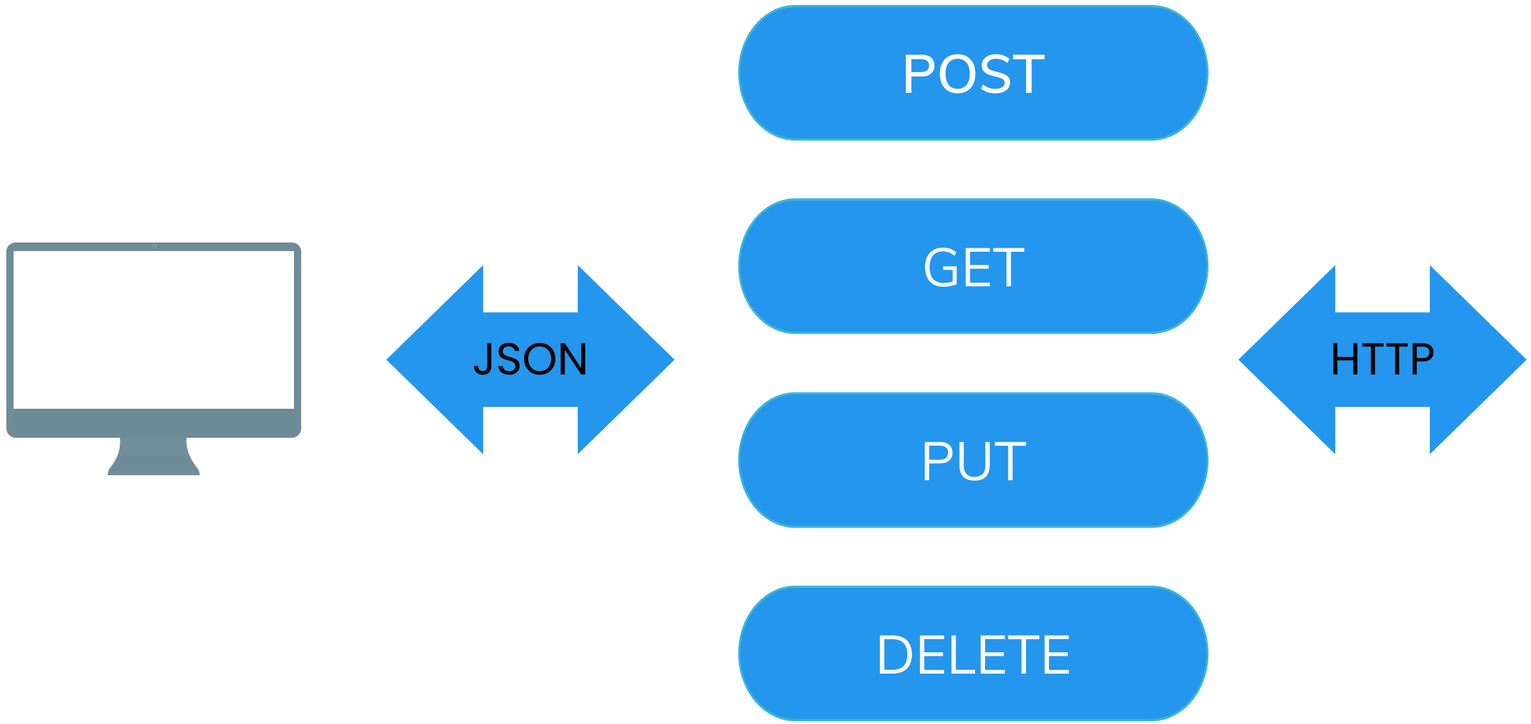


figura2: template per la Pulizia e il Filtraggio dei dati

REST API per l’automatizzazione del flusso



Query1

read()

creazione dataframe iniziale.

groupBy()

calcola il numero totale di fallimenti per ogni giorno e per ogni vault_id

filter()

filtra i vault che hanno subito esattamente 4, 3 e 2 fallimenti

orderBy()

show()

attuazione query

Query2.1

read()

creazione dataframe iniziale.

groupBy().agg(sum())

calcola fallimenti totali per ogni modello

orderBy()

crea classifica in base ai fallimenti totali

show()

attuazione query

Query2.2

filter()

dal dataframe iniziale filtra solo le entry con failure = 1

groupBy().agg (count())

raggruppa per 'vault_id' e calcola il numero totale di fallimenti per ogni vault

groupBy().agg (collect_set())

raggruppa per 'vault_id', raccogliendo i modelli distinti di hard disk soggetti ad almeno un fallimento

join()

unisce i due dataframe precedenti

orderBy()

ordina il dataframe precedente per creare una classifica

show()

attuazione query

Query3

read()

creazione dataframe iniziale.

filter().select().distinct()

ottenere solo i serial_number che hanno failure uguale a 1

groupBy()

trova la data massima per ogni "model"

join()

unisce i risultati del dataframe precedente con il dataframe originale

leftJoin()

unisce il dataframe precedente con quello generato dalla filter

calculateStatistics()

calcola le statistiche sui “failed hard disk”

calculateStatistics()

calcola le statistiche sui “not failed hard disk”

union()

unisce i dataframe delle statistiche

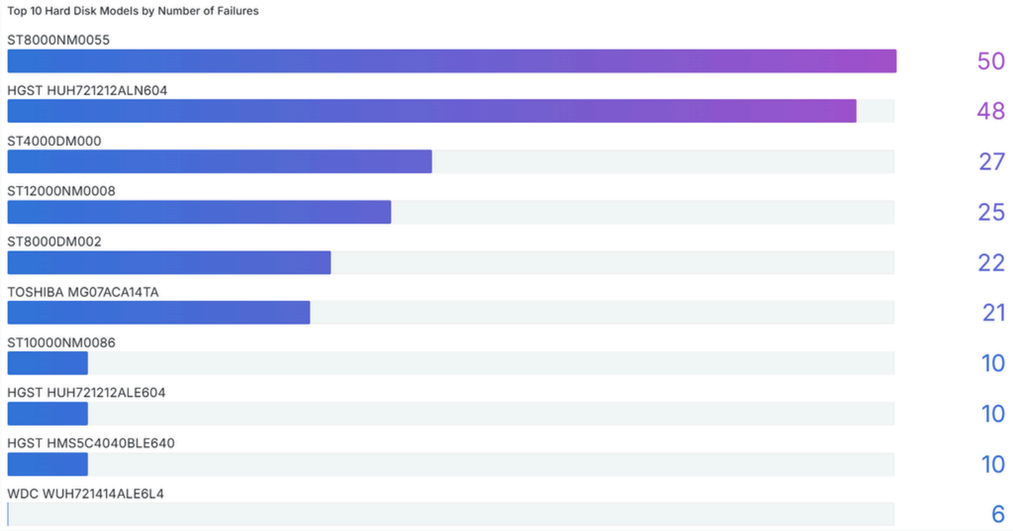
show()

attuazione query

Data-visualization

Query2 - Top 10 Vaults by Number of Failures

vaultId	totalFailures ↓	uniqueModels
1113	15	HGST HUH721212ALN604
1120	10	HGST HUH721212ALN604
1093	9	ST10000NM0086
1053	8	ST8000NM0055, TOSHIBA MQ01ABF050M
1118	8	HGST HUH721212ALN604
1032	7	ST8000DM002
1090	7	ST8000NM0055, WDC WD5000LPVX, TOSHIBA MQ01ABF050
1124	6	HGST HUH721212ALN604, HGST HUH721212ALE604
1066	6	TOSHIBA MG07ACA14TA
1055	6	ST8000NM0055



Distribution of Operating Hours for Hard Disks with and without Failures

# failure	min	25th_percentile	50th_percentile	75th_percentile	max	count
1	522	27781	38702	51965	71608	265
0	0	15119	22650	42066	87702	242644

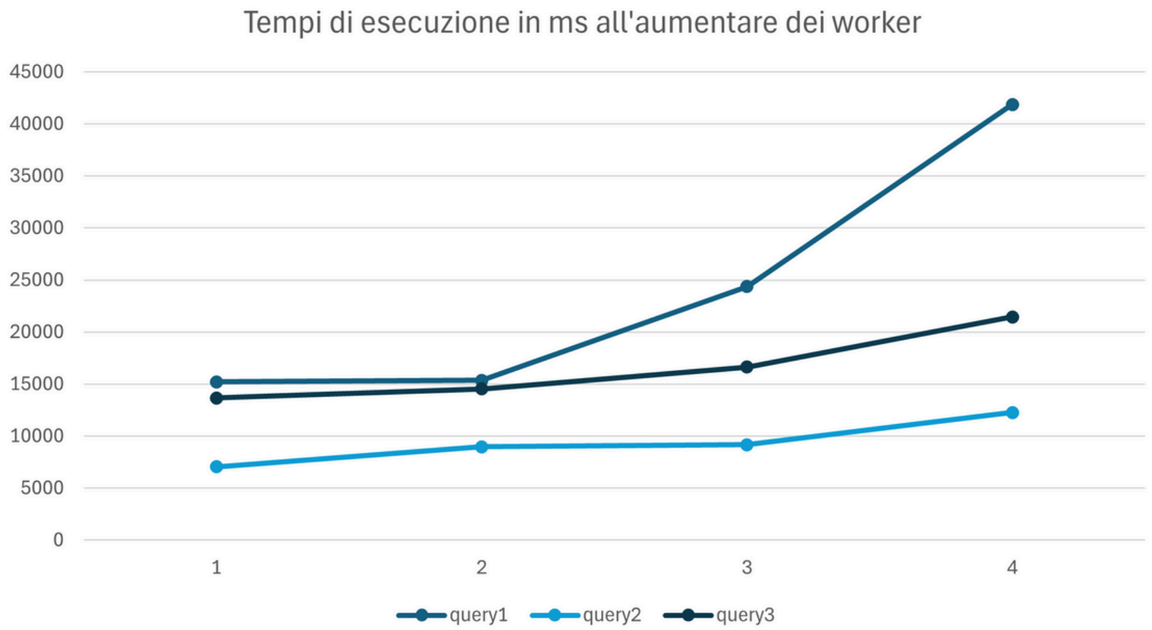
Valutazione delle Performance

Specifiche:

- 5 esecuzioni di ogni query
- configurazione con 1,2,3,4 worker

Caratteristiche macchina:

- {Processore}: 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz, 2803 Mhz, 4 core, 8 processori logici
- {Memoria fisica installata (RAM)}: 16,0 GB
- {SO}: Microsoft Windows 11 Pro



Grazie per l’attenzione!