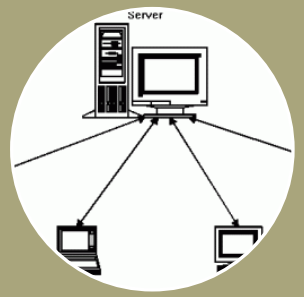
SABD PROGETTO 2

Dissan Uddin Ahmed

TOPICS

Obiettivo



Architettura

Queries

OBIETTIVO

Lo scopo del progetto `e usare il framework di data stream processing Apache Flink per rispondere a due query su dati di telemetria di circa 200k hard disk nei data center gestiti da Backblaze. Essendo i dati già disponibili è necessario simulare la produzione di 23 giorni di monitoraggio.

DATASET

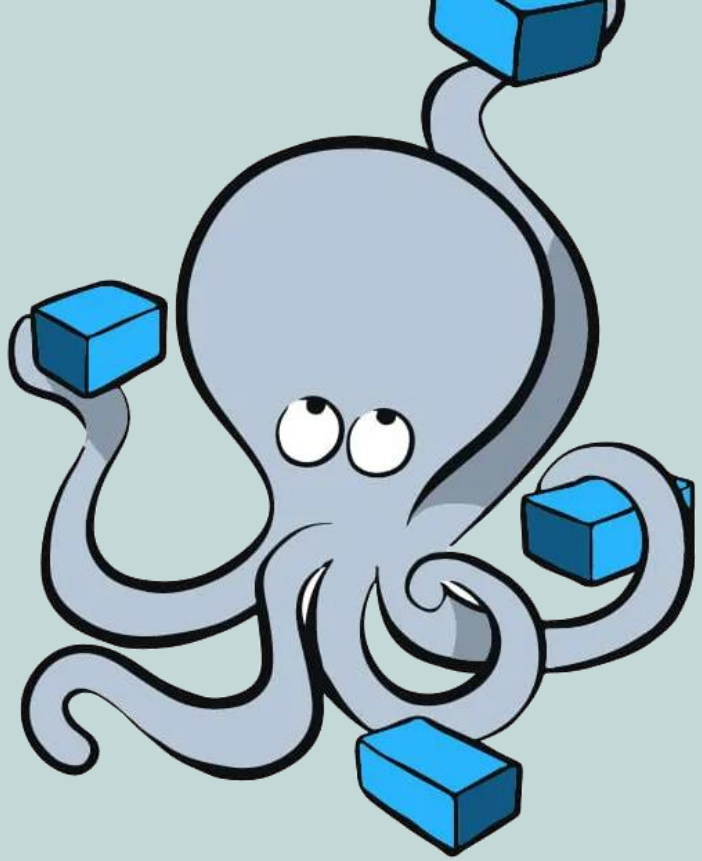
Dataset di blackbaze che contiene dati riguardanti 23 giorni di monitoraggio.

Le query verranno eseguite su questo sottoinsieme di campi.

Nome	Tipo
Date	Format: yyyy-mm-ddTHH:MM:ss.SSSSSS
Serial_number	String
Model	String
Failure	Boolean
Vault_id	Int64
S194_temperature_Celsius	Int64

DEPLOYMENT

Docker compose

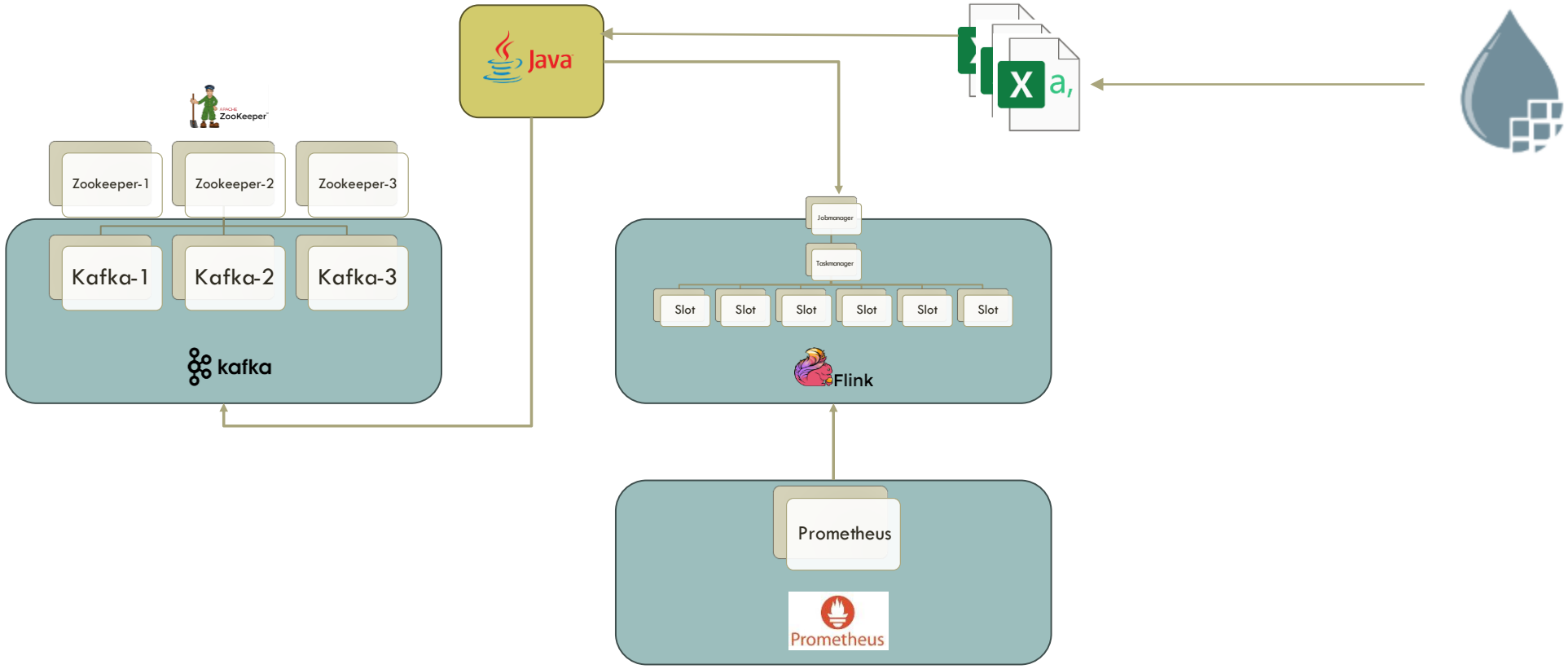


docker Compose

ARCHITETTURA

Supported by docker-engine

ARCHITETTURA

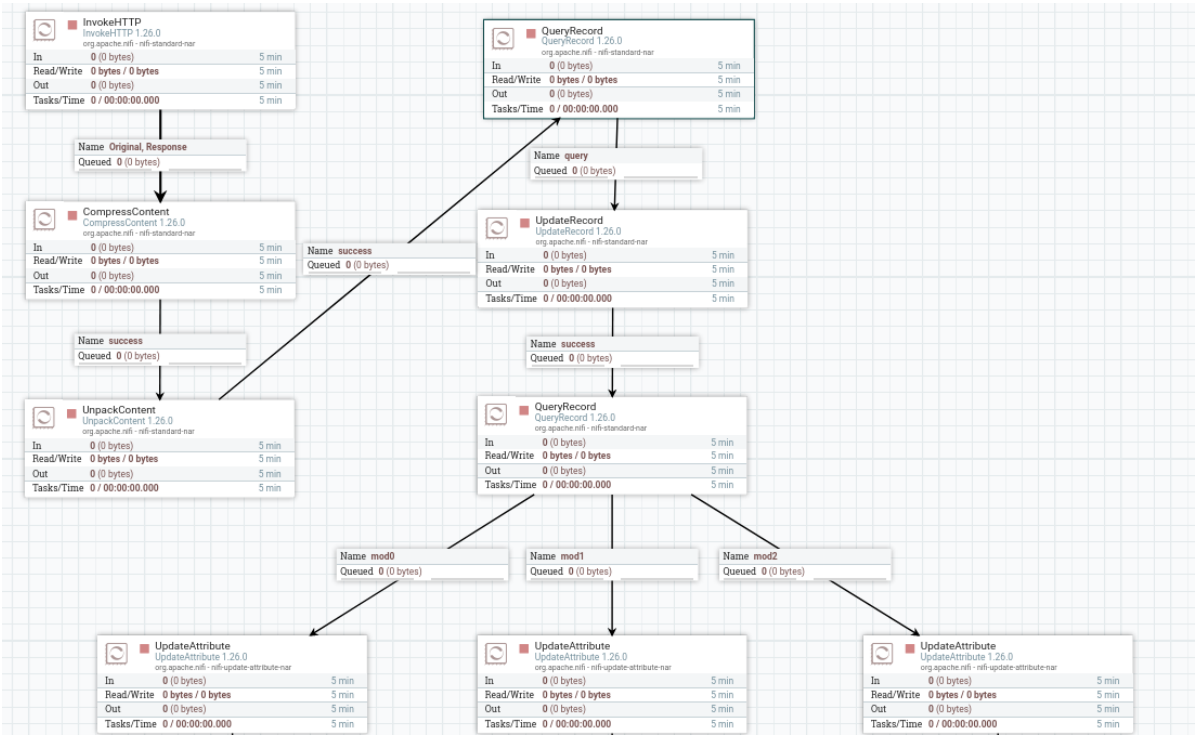


INGESTION

Nifi + Kafka

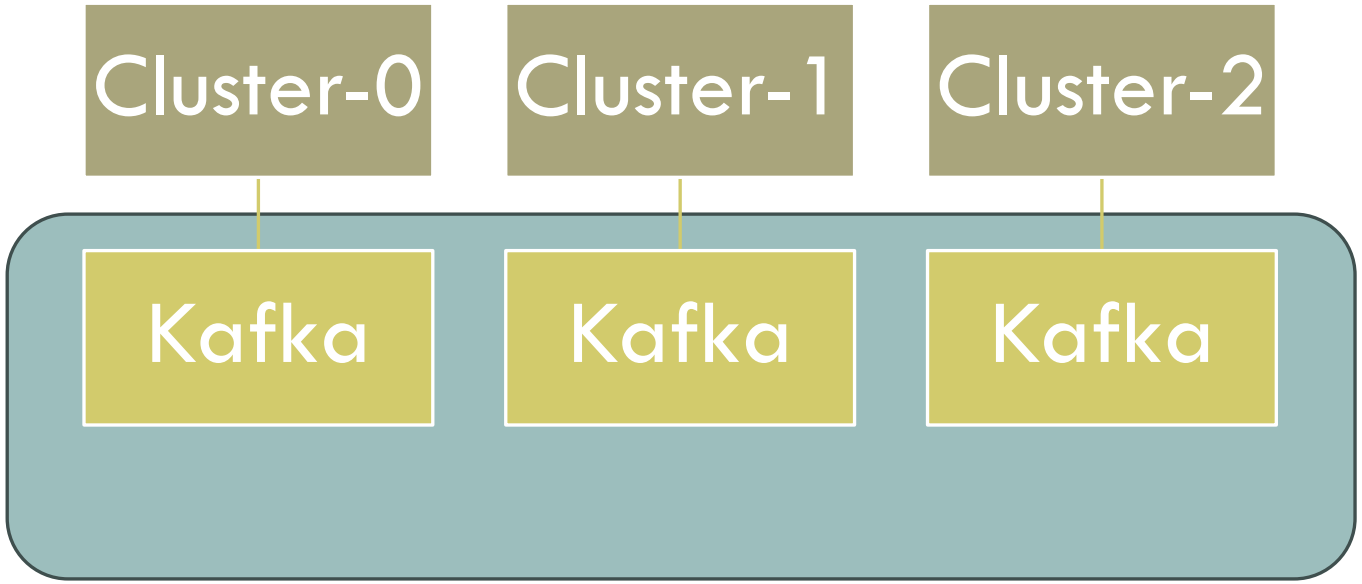
NIFI

Sono stati notati alcuni campi duplicati nell'header quindi sono stati modificati ed è stato cambiato il formato del date in timpestamp. È Servito per dividere il dataset in 3 parti così da simulare i dati provenienti da 3 cluster differenti



PRODUCER KAFKA

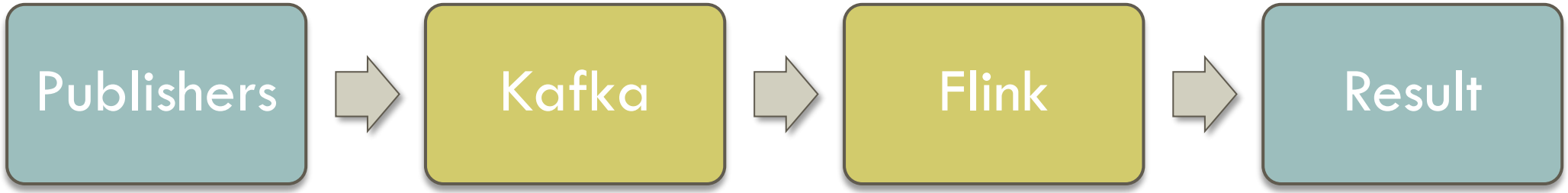
È un applicazione java che si occupa di leggere il proprio file di prendere le tuple creare un gruppo di 500 tuple e mandarlo a kafka facendo una sleep di 240 ms tra uno scambio e un altro. Infine vengono mandate delle tuple fasulle per far partire la finestra di 23 giorni.



QUERY

Query 1 & Query 2

PIPELINE

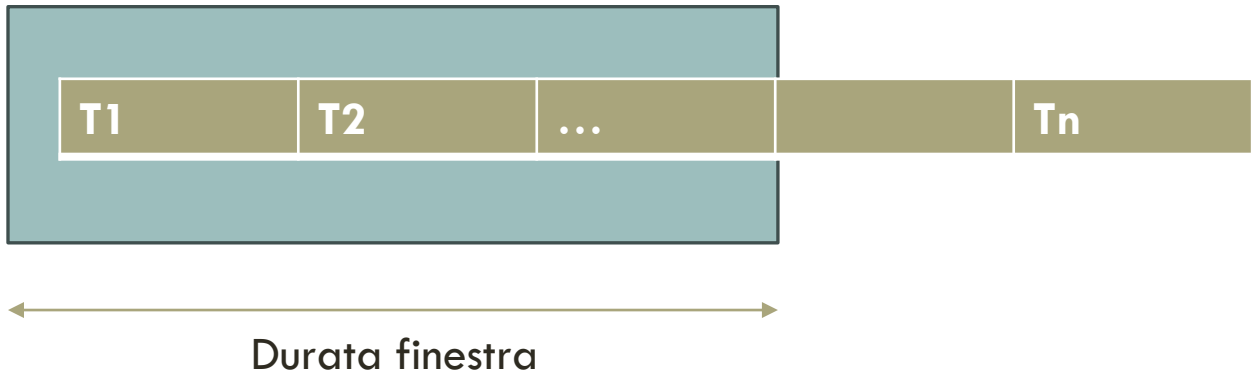


WATERMARK

Strategia usata per calcolare le finestre tumbling di durata EnvenTime 1, 3, 23 giorni,

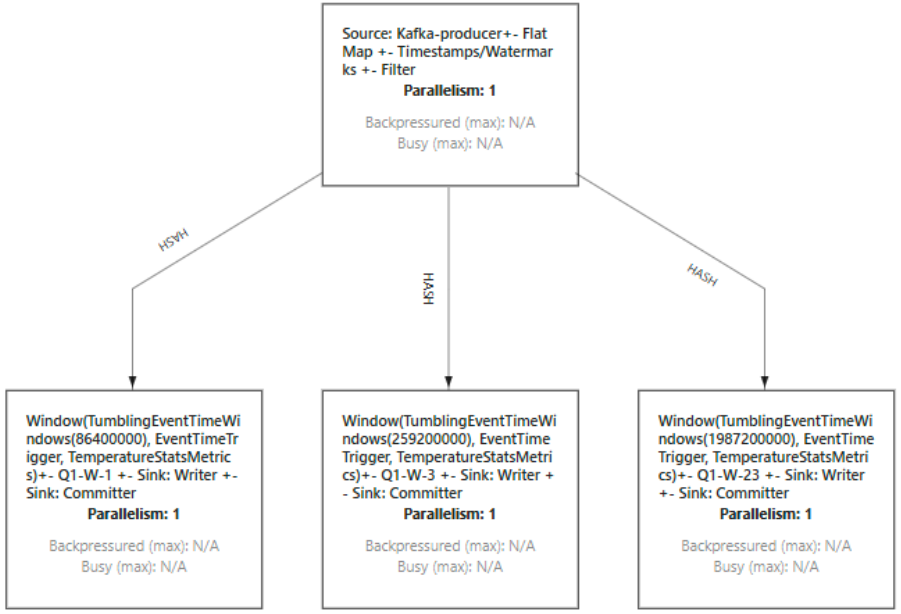
forBoundedOutOfOrderness di giorni 1

In Quanto avendo 3 cluster è possibile che qualcuno mandi un timestamp t1 > t2 e quindi che la finestra parta la scartando alcune tuple del giorno precedente arrivi



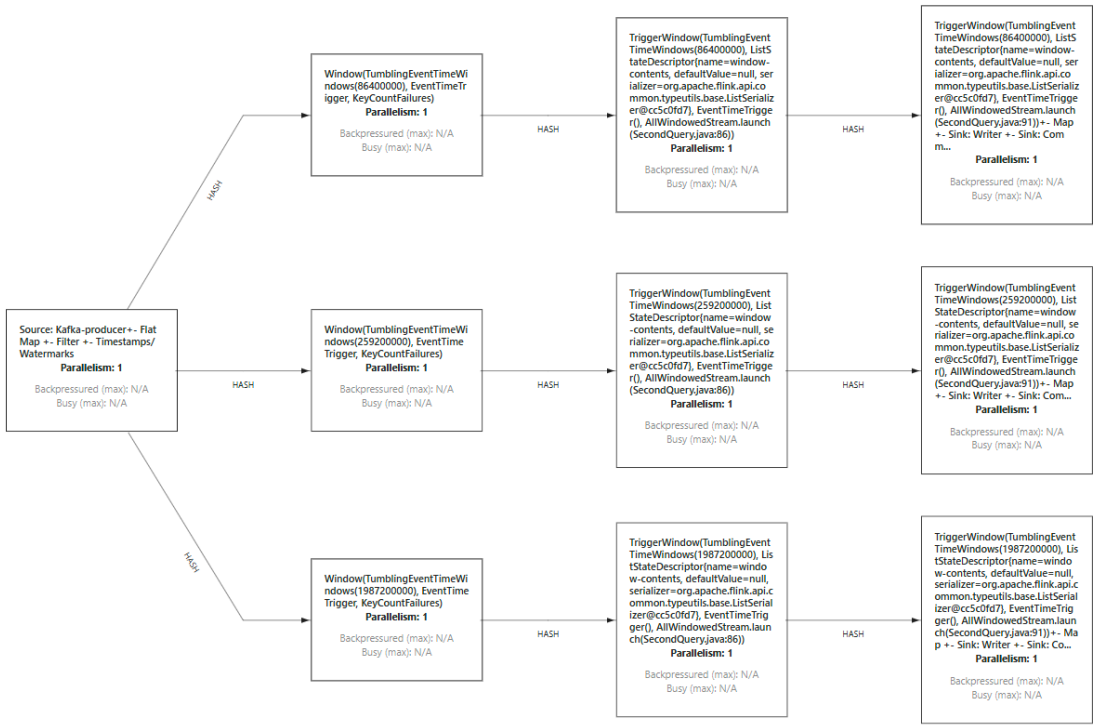
QUERY 1

Per i vault (campo vault temperature id) con identificativo compreso tra 1000 e 1020, calcolare il numero di eventi, il valor medio e la deviazione standard della temperatura misurata sui suoi hard disk (campo s194 celsius). Si faccia attenzione alla possibile presenza di eventi che non hanno assegnato un valore per il campo relativo alla temperatura. Selezione delle Colonne Rilevanti: Sono state selezionate solo le colonne necessarie per l'analisi: data, vault_id e failure



QUERY 2

Calcolare la classifica aggiornata in tempo reale dei 10 vault che registrano il piu alto numero di fallimenti nella stessa giornata. Per ogni vault, riportare il numero di fallimenti ed il modello e numero seriale degli hard disk guasti



RISULTATI

Latenza e Throughput

QUERY 1 & 2

Latenza

Throughput

QUERY 1

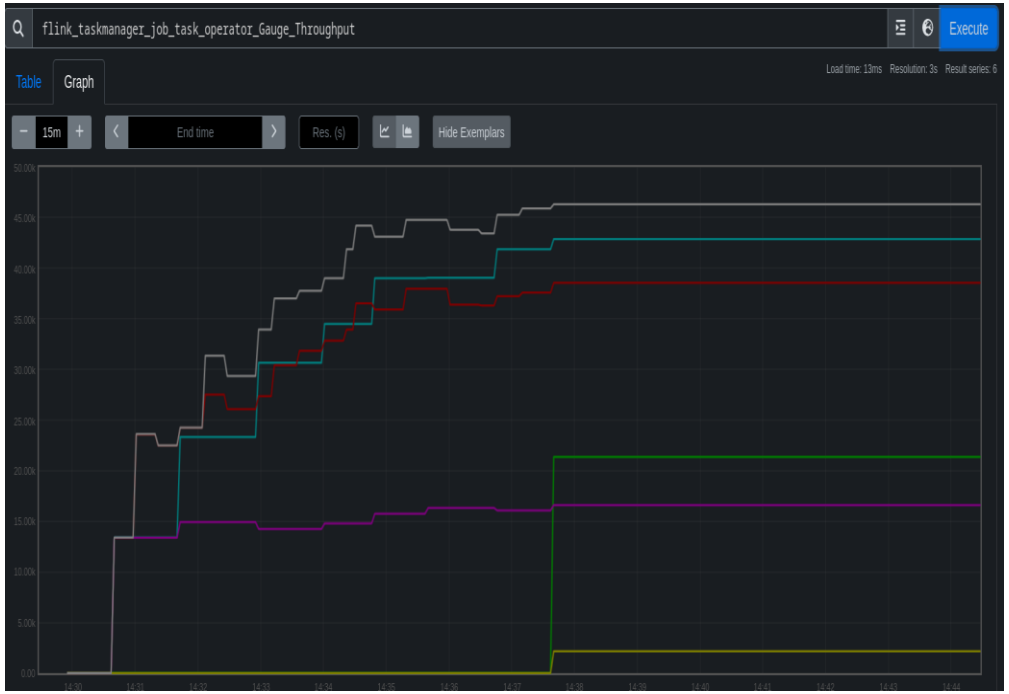
Latenza

Throughput

QUERY 2

Latenza

Throughput



GRAZIE PER L'ATTENZIONE

A cura di Dissan Uddin Ahmed