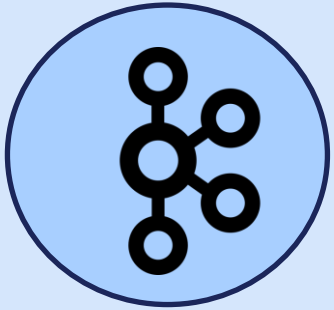

SABD-Project 2
Pieraldo Santurro
0353517

Introduzione

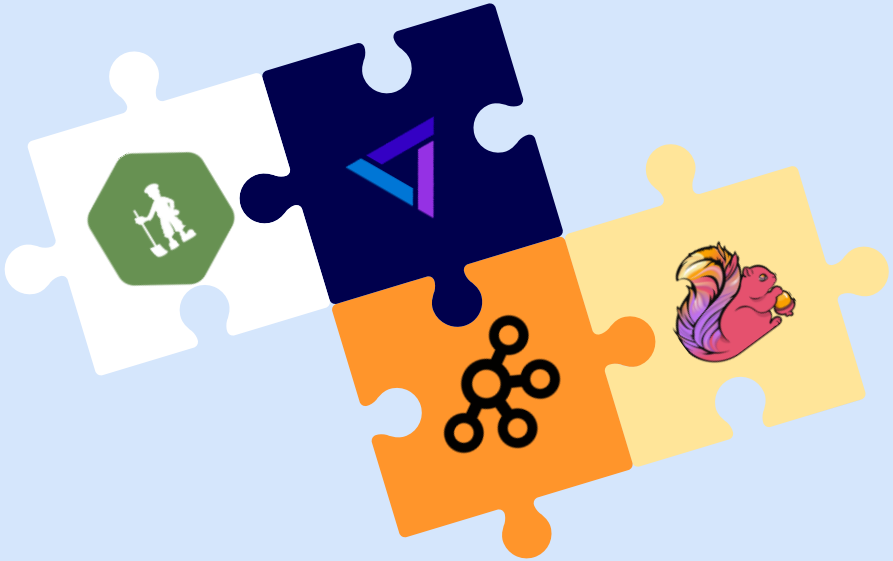
Analisi di dati forniti da Blackblaze relativi a fallimenti di Hard Disk



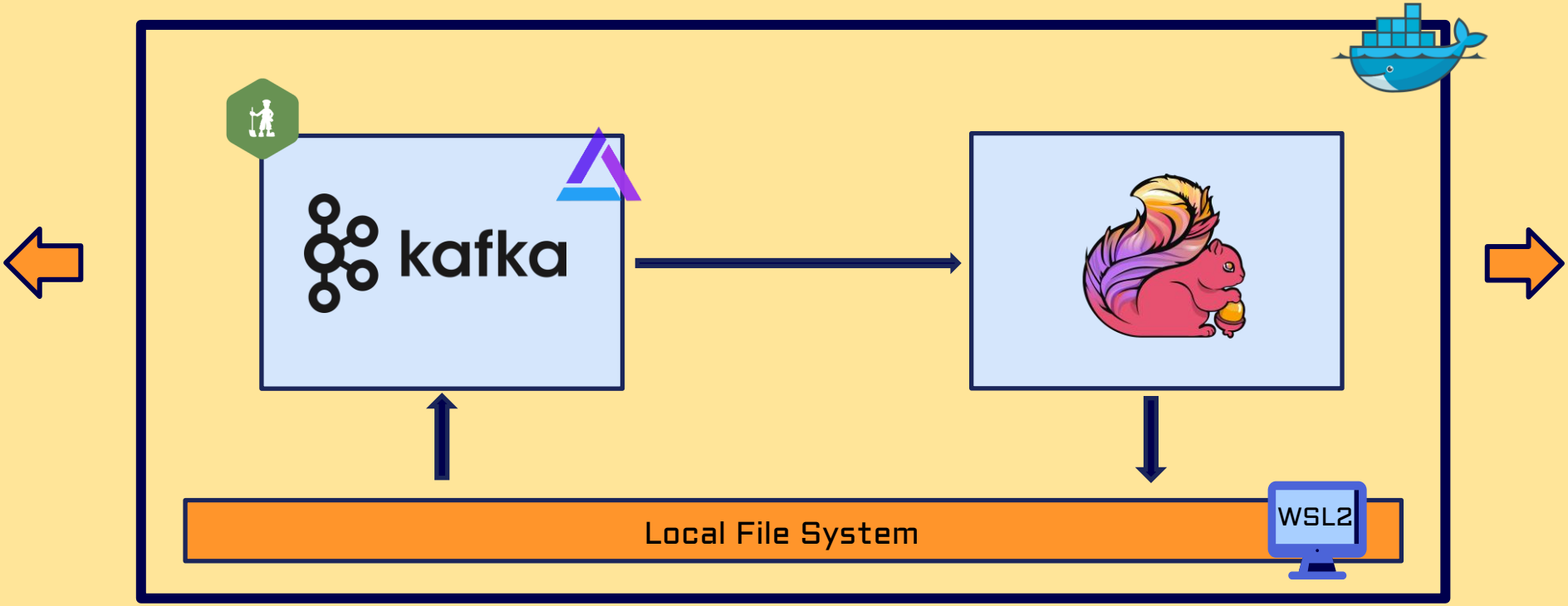
Caricamento del dataset filtrato su topic Kafka

Esecuzione delle Query su Flink e salvataggio sul Local File System

Architettura

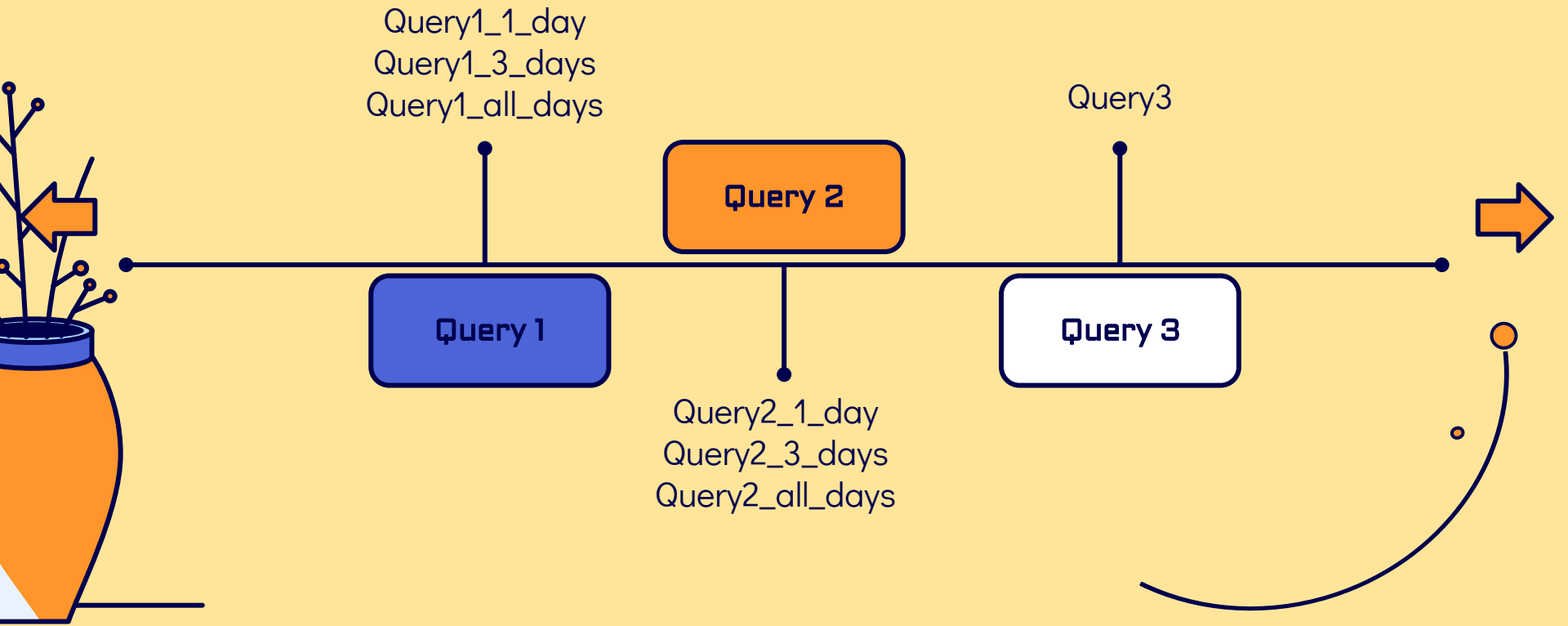


Architettura



Queries

Queries


Query1_1_day
Query1_3_days
Query1_all_days
Query 2
Query3
Query 1
Query 3
Query2_1_day
Query2_3_days
Query2_all_days

Queries

Tumbling Windows

5 secondi

Tempo simulato

23 giorni ⟶ 33120 minuti

Tempo simulazione

18 minuti

Fattore di Accelerazione

$$\frac{Tempo\ simulato\ totale}{Tempo\ reale\ totale} = 1840$$

Un giorno

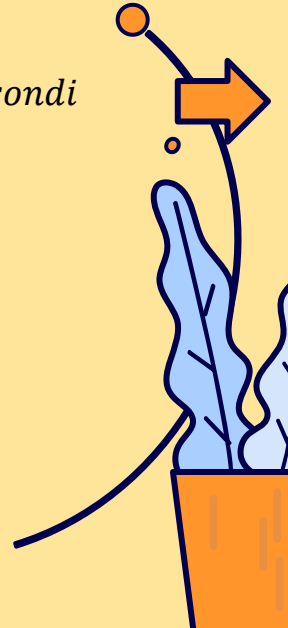
$$\frac{Minuti\ in\ un\ giorno = 1440}{1840} = 49.96\ secondi$$

Un'ora

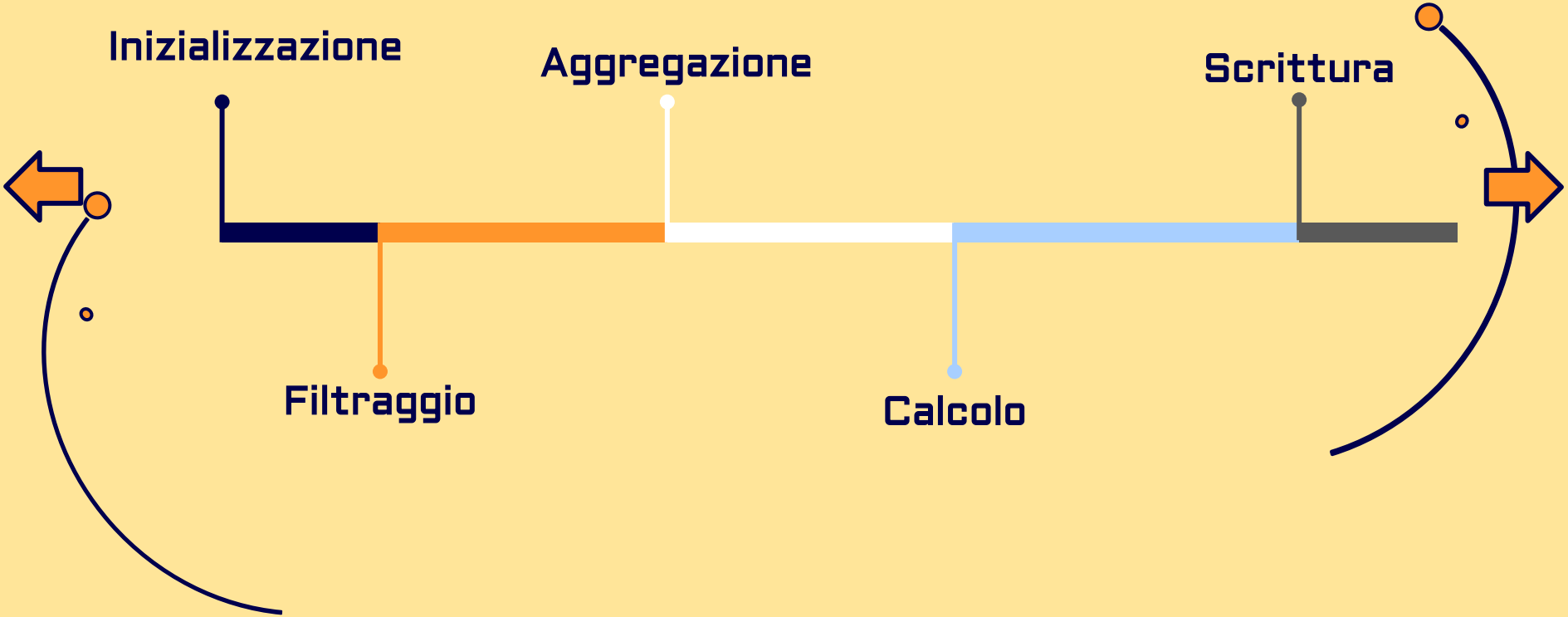
$$\frac{Minuti\ in\ un'ora = 60}{1840} = 1.96\ secondi$$

Mezz'ora

$$\frac{30}{1840} = 0.98\ secondi$$



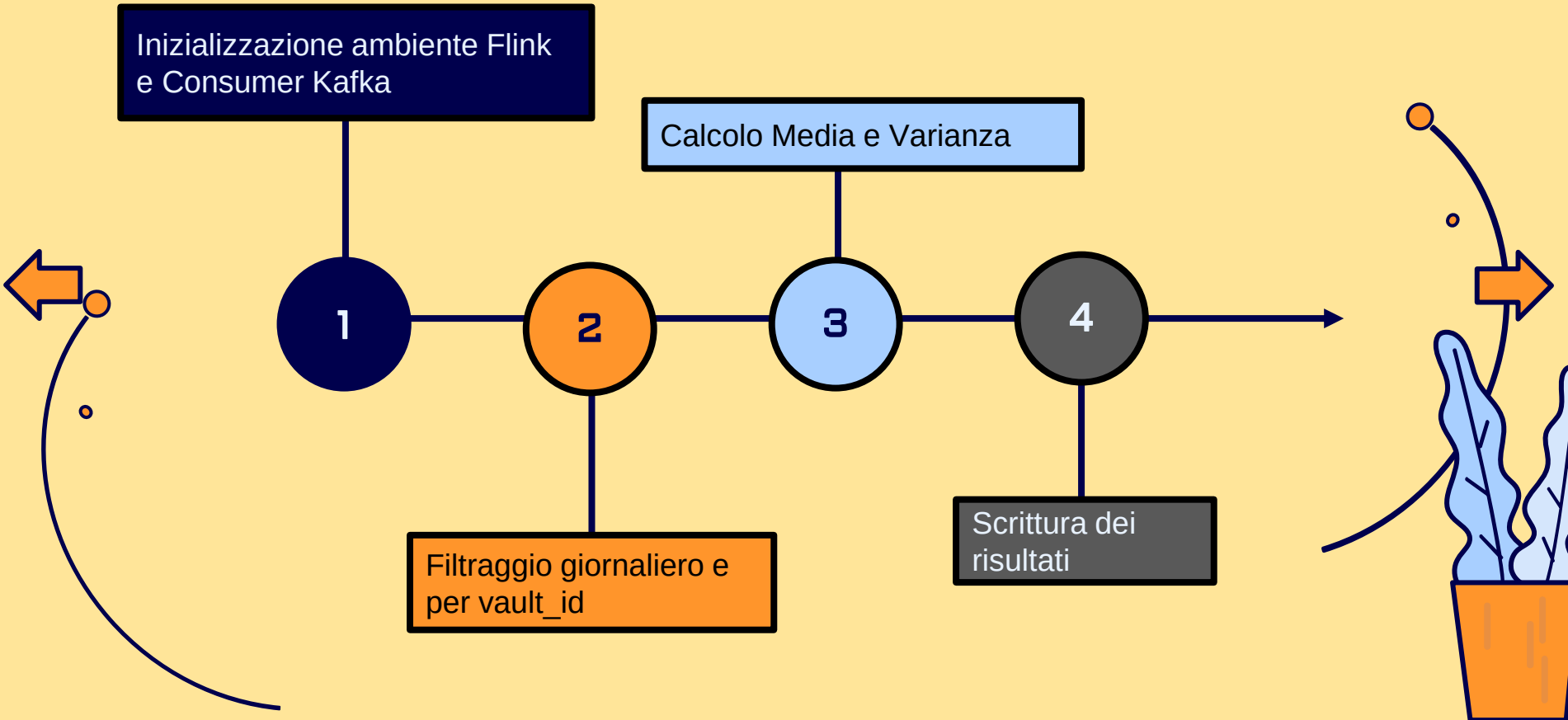
Queries


Inizializzazione
Aggregazione
Scrittura
Filtraggio
Calcolo

Query 1

Per i vault (campo vault 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 temperature celsius). Si faccia attenzione alla possibile presenza di eventi che non hanno assegnato un valore per il campo relativo alla temperatura.

Query 1


Inizializzazione ambiente Flink e Consumer Kafka
Calcolo Media e Varianza
1
2
3
4
Filtraggio giornaliero e per vault_id
Scrittura dei risultati

Query 1

Media

$$\text{mean} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Deviazione Standard

$$\text{stDev} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - mean)^2}$$

Query 1

window,vault_id,count,mean_s194,stddev_s194,latency,throughput
1719670425000 - 1719670430000,1000,317,22.179810725552052,2.3217194138677457,5001,1296.8
1719670425000 - 1719670430000,1001,316,18.29113924050633,1.8411739426086813,5001,1296.8
1719670425000 - 1719670430000,1002,332,22.572289156626507,2.439178868993142,5001,1296.8
1719670425000 - 1719670430000,1003,305,18.236065573770492,2.2272553709674514,5001,1296.8
1719670425000 - 1719670430000,1004,308,19.7012987012987,2.8137645673808005,5001,1296.8
1719670425000 - 1719670430000,1005,307,21.257328990228014,2.7327611943207963,5001,1296.8
1719670425000 - 1719670430000,1006,309,22.57281553398058,3.8447255723250375,5001,1296.8
1719670425000 - 1719670430000,1007,321,25.822429906542055,4.738283962103057,5001,1296.8
1719670425000 - 1719670430000,1008,317,29.533123028391167,4.122303040808419,5001,1296.8
1719670425000 - 1719670430000,1009,297,27.64983164983165,3.340186133436527,5001,1296.8
1719670425000 - 1719670430000,1010,317,22.082018927444796,2.738824443236922,5001,1296.8
1719670425000 - 1719670430000,1011,283,26.374558303886925,6.562635647539818,5001,1296.8
1719670425000 - 1719670430000,1012,315,30.16190476190476,7.532099110261543,5001,1296.8
1719670425000 - 1719670430000,1013,289,26.737024221453286,4.089827806981952,5001,1296.8
1719670425000 - 1719670430000,1014,318,25.965408805031448,5.14830969905723,5001,1296.8
1719670425000 - 1719670430000,1015,294,26.156462585034014,4.064447951021061,5001,1296.8
1719670425000 - 1719670430000,1016,301,26.627906976744185,4.774073663045416,5001,1296.8
1719670425000 - 1719670430000,1017,304,28.14144736842105,5.4231829855351865,5001,1296.8
1719670425000 - 1719670430000,1018,310,27.10322580645161,3.887405957453338,5001,1296.8
1719670425000 - 1719670430000,1019,306,27.76143790849673,5.016829509219544,5001,1296.8
1719670425000 - 1719670430000,1020,318,26.569182389937108,5.609958011541817,5001,1296.8
1719670430000 - 1719670435000,1000,49,22.244897959183675,2.5114481735371688,5000,189.2
1719670430000 - 1719670435000,1001,41,18.024390243902438,1.4564876709105967,5000,189.2
1719670430000 - 1719670435000,1002,40,22.775,2.6219029348928995,5000,189.2
1719670430000 - 1719670435000,1003,47,18.93617021276596,2.823942131580776,5000,189.2

Query 1

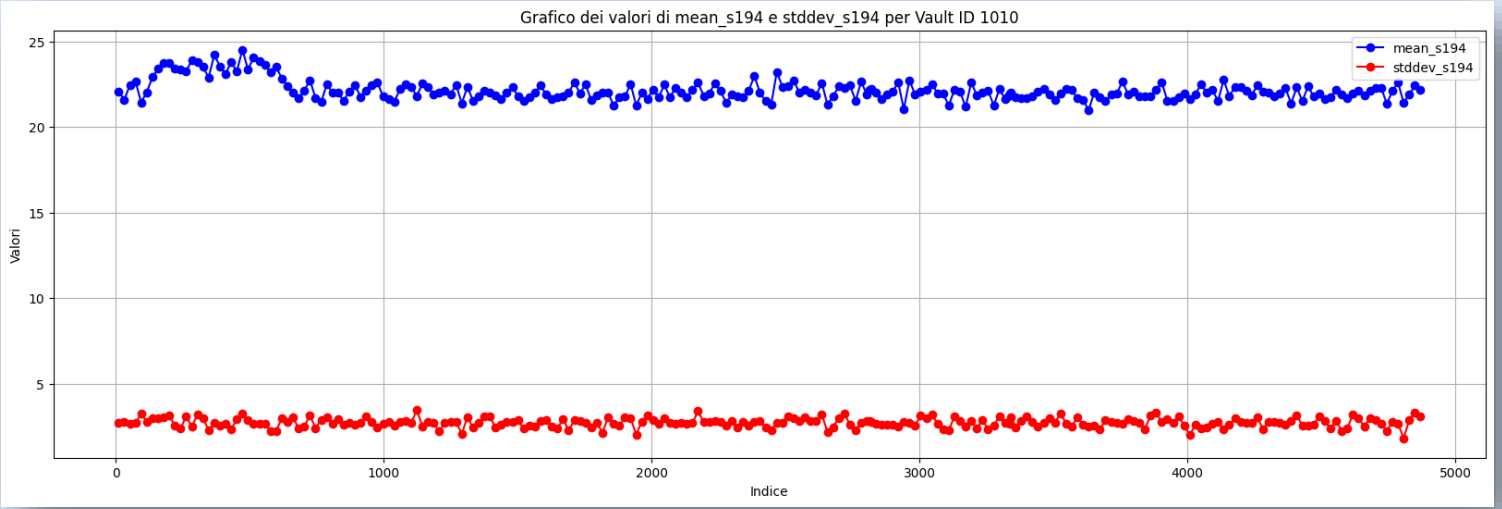
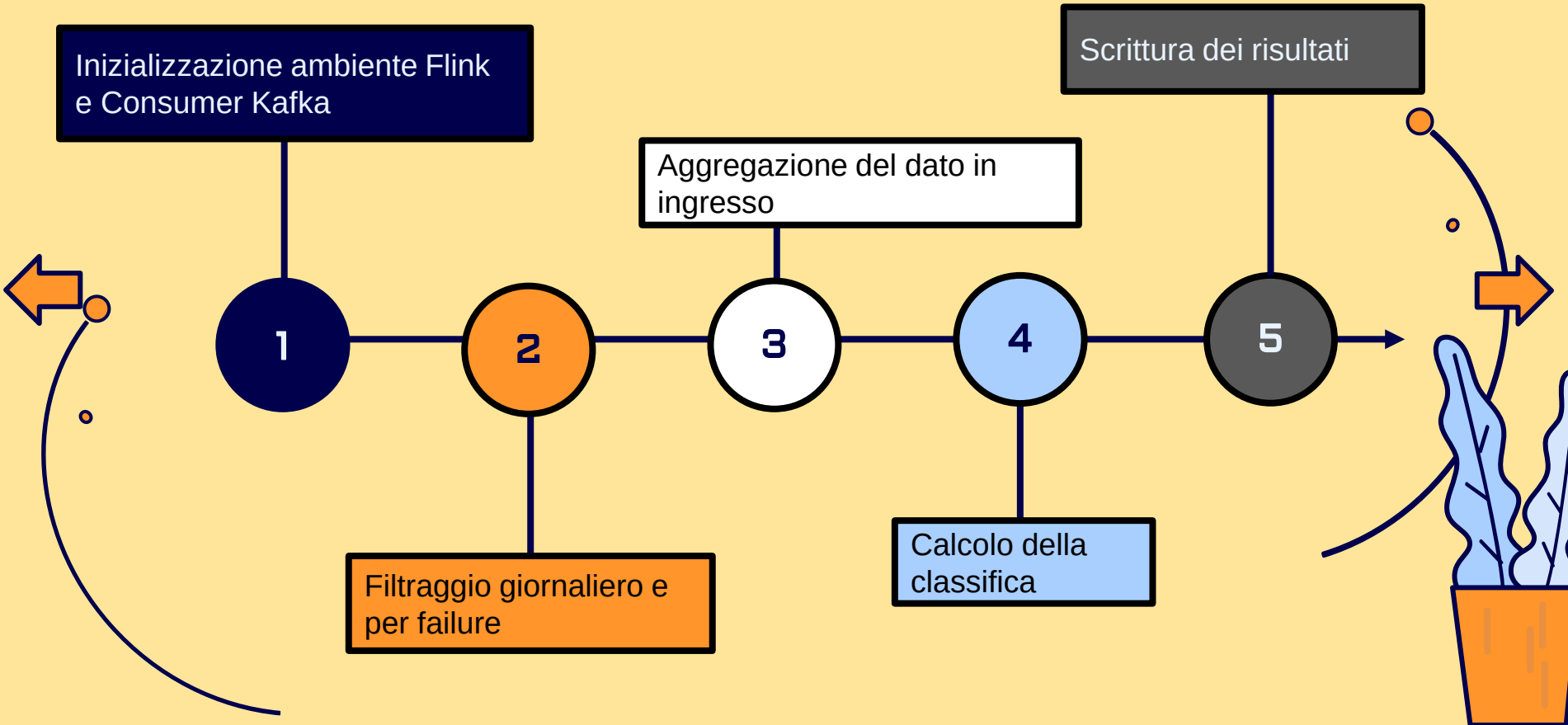


Grafico dei valori di mean_s194 e stddev_s194 per Vault ID 1010
mean_s194
stddev_s194
Valori
Indice
25
20
15
10
5
0
1000
2000
3000
4000
5000

Query 2

Calcolare la classifica aggiornata in tempo reale dei 10 vault che registrano il più 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.

Query 2


Inizializzazione ambiente Flink e Consumer Kafka
Aggregazione del dato in ingresso
Scrittura dei risultati
1
2
3
4
5
Filtraggio giornaliero e per failure
Calcolo della classifica

Query 2

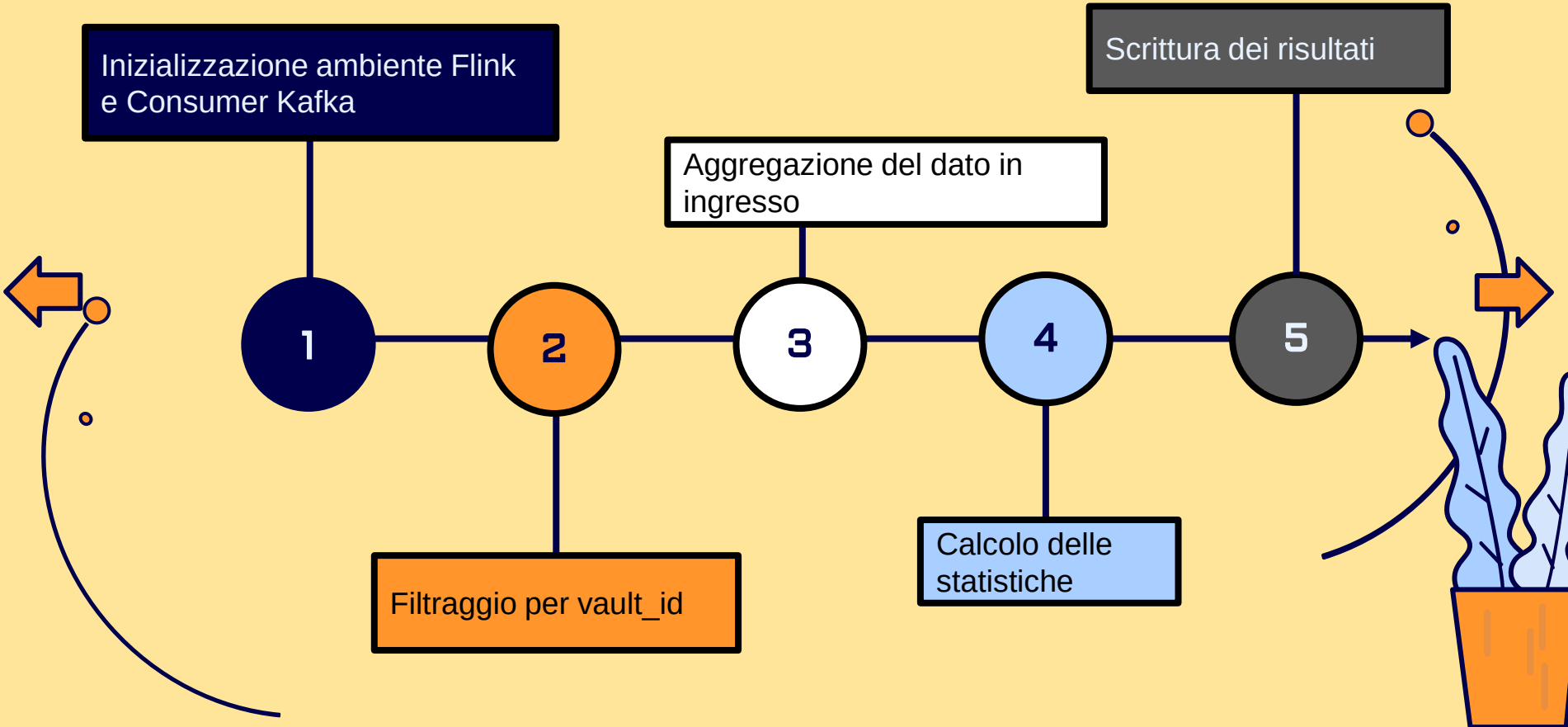
```
window,latency,throughput,vault_id1,failures1,model_serial_list1,...,vault_id10,failures10,model_serial_list10
1719670430000,5001,0.8,1047,1,[ST4000DM000, Z304NT6A],1145,1,[HGST HUH721212ALE604, 8HGXNP2H],1099,1,[ST12000NM001G,
ZTN0AGRW],1055,1,[ST8000NM0055, ZA1814BV], , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , ,
1719670440000,5000,0.2,1106,1,[TOSHIBA MG07ACA14TA, 70K0A08ZF97G],1047,1,[ST4000DM000, Z304NT6A],1145,1,[HGST HUH721212ALE604,
8HGXNP2H],1099,1,[ST12000NM001G, ZTN0AGRW],1055,1,[ST8000NM0055, ZA1814BV], , , , , , , , , , , , , , , , , , , , , , , ,
1719670445000,5000,0.2,1106,1,[TOSHIBA MG07ACA14TA, 70K0A08ZF97G],1047,1,[ST4000DM000, Z304NT6A],1145,1,[HGST HUH721212ALE604,
8HGXNP2H],1099,1,[ST12000NM001G, ZTN0AGRW],1118,1,[HGST HUH721212ALN604, AAG68EWH],1055,1,[ST8000NM0055,
ZA1814BV], , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , ,
```

```
1719671590000,5000,0.2,1093,3,[ST10000NM0086, ZA21HD3Y], [ST10000NM0086, ZA21E1Z8], [ST10000NM0086, ZA20ZBAR],1032,3,[ST8000DM002,
ZA136191], [ST8000DM002, ZA12QCLA], [ST8000DM002, ZA12YMZ6],1041,2,[ST8000NM0055, ZA16V744], [ST8000NM0055, ZA171VP2],1010,2,[ST4000DM000,
Z304KA0X], [ST4000DM000, Z304JW59],1044,2,[ST8000NM0055, ZA17HPSZ], [ST8000NM0055, ZA18BLRJ],1014,2,[ST4000DM000, Z304LAP7], [ST4000DM000,
Z304KCMR],1152,1,[WDC WUH721414ALE6L4, QGK8T2WT],1120,1,[HGST HUH721212ALN604, AAG63XJH],1089,1,[TOSHIBA MG07ACA14TA,
61A0A0MUF97G],1034,1,[ST8000DM002, ZA11R9DT]
1719671715754,0,0.0,1093,3,[ST10000NM0086, ZA21HD3Y], [ST10000NM0086, ZA21E1Z8], [ST10000NM0086, ZA20ZBAR],1032,3,[ST8000DM002, ZA136191],
[ST8000DM002, ZA12QCLA], [ST8000DM002, ZA12YMZ6],1041,2,[ST8000NM0055, ZA16V744], [ST8000NM0055, ZA171VP2],1010,2,[ST4000DM000, Z304KA0X],
[ST4000DM000, Z304JW59],1044,2,[ST8000NM0055, ZA17HPSZ], [ST8000NM0055, ZA18BLRJ],1014,2,[ST4000DM000, Z304LAP7], [ST4000DM000,
Z304KCMR],1152,1,[WDC WUH721414ALE6L4, QGK8T2WT],1120,1,[HGST HUH721212ALN604, AAG63XJH],1089,1,[TOSHIBA MG07ACA14TA,
61A0A0MUF97G],1034,1,[ST8000DM002, ZA11R9DT]
```

Query 3

Calcolare il minimo, 25-esimo, 50-esimo, 75-esimo percentile e massimo delle ore di funzionamento (campo s9 power on hours) degli hark disk per i vault con identificativo tra 1090 (compreso) e 1120 (compreso). Si presti attenzione, il campo s9 power on hours riporta un valore cumulativo, pertanto le statistiche richieste dalla query devono far riferimento all'ultimo valore utile di rilevazione per ogni specifico hard disk (si consideri l'uso del campo serial number). I percentili devono essere calcolati in tempo reale, senza ordinare tutti i valori e possibilmente senza accumularli; si utilizzi pertanto un algoritmo approssimato che consente di calcolare i percentili riducendo la quantita di memoria occupata al prezzo di una minore accuratezza, e.g., [4, 6, 7, 3, 5].

Query 3


Inizializzazione ambiente Flink e Consumer Kafka
Scrittura dei risultati
Aggregazione del dato in ingresso
1
2
3
4
5
Filtraggio per vault_id
Calcolo delle statistiche

Query 3

window,vault_id,min,percentile25,percentile50,percentile75,max,count,latency,throughput
1719670446000,1090,333.000000,49421.953125,49447.000000,49455.140625,74326.000000,367,2000,5954.000000
1719670446000,1091,314.000000,49142.000000,49146.000000,49150.700000,61178.000000,369,2000,5954.000000
1719670446000,1092,502.000000,48858.160000,48864.000000,48868.160000,60002.000000,398,2000,5954.000000
1719670446000,1093,215.000000,48494.870000,48500.347222,48505.560000,48514.000000,409,2000,5954.000000
1719670446000,1094,8207.000000,16598.810000,17781.458333,21518.110000,44624.000000,395,2000,5954.000000
1719670446000,1095,3876.000000,16507.880000,18136.666667,22355.800000,45442.000000,395,2000,5954.000000
1719670446000,1096,8064.000000,47051.680000,47056.720000,47063.520000,69680.000000,382,2000,5954.000000
1719670446000,1097,78.000000,18312.156250,21198.200000,23043.093750,77804.000000,358,2000,5954.000000
1719670446000,1098,4021.000000,17695.391667,20631.680000,23011.250000,47505.000000,373,2000,5954.000000
1719670446000,1099,725.000000,15037.000000,17981.480000,22721.312500,69102.000000,364,2000,5954.000000
1719670446000,1100,5187.000000,18716.250000,20498.500000,23663.875000,72014.000000,352,2000,5954.000000
1719670446000,1101,356.000000,15621.980000,18201.500000,22277.100000,45754.000000,376,2000,5954.000000
1719670446000,1102,2773.000000,17953.340000,21760.472222,23585.660000,71085.000000,404,2000,5954.000000
1719670446000,1103,1283.000000,14362.600000,18042.388889,21092.080000,45518.000000,390,2000,5954.000000
1719670446000,1104,111.000000,14788.560000,17490.600000,21584.360000,45042.000000,382,2000,5954.000000
1719670446000,1105,3061.000000,17395.100000,19511.430556,22287.960000,43243.000000,399,2000,5954.000000
1719670446000,1106,6560.000000,16489.533333,18357.060000,21020.960000,76511.000000,372,2000,5954.000000
1719670446000,1107,558.000000,16268.640000,17680.416667,22266.060000,42896.000000,401,2000,5954.000000
1719670446000,1108,696.000000,16796.800000,19218.666667,22888.360000,41573.000000,384,2000,5954.000000
1719670446000,1109,95.000000,16910.180000,19382.333333,21713.500000,84375.000000,392,2000,5954.000000
1719670446000,1110,267.000000,17223.970000,18659.569444,22581.590000,69759.000000,403,2000,5954.000000
1719670446000,1111,570.000000,39380.820000,39398.750000,39406.540000,39417.000000,388,2000,5954.000000
1719670446000,1112,142.000000,17253.410000,18663.319444,22620.090000,39065.000000,403,2000,5954.000000
1719670446000,1113,80.000000,38440.000000,38448.680000,38460.208333,42718.000000,371,2000,5954.000000
1719670446000,1114,1660.000000,17534.125000,18721.460000,21263.281250,81516.000000,362,2000,5954.000000
1719670446000,1115,602.000000,14966.670000,17036.160000,21351.760000,38060.000000,379,2000,5954.000000

Query 3

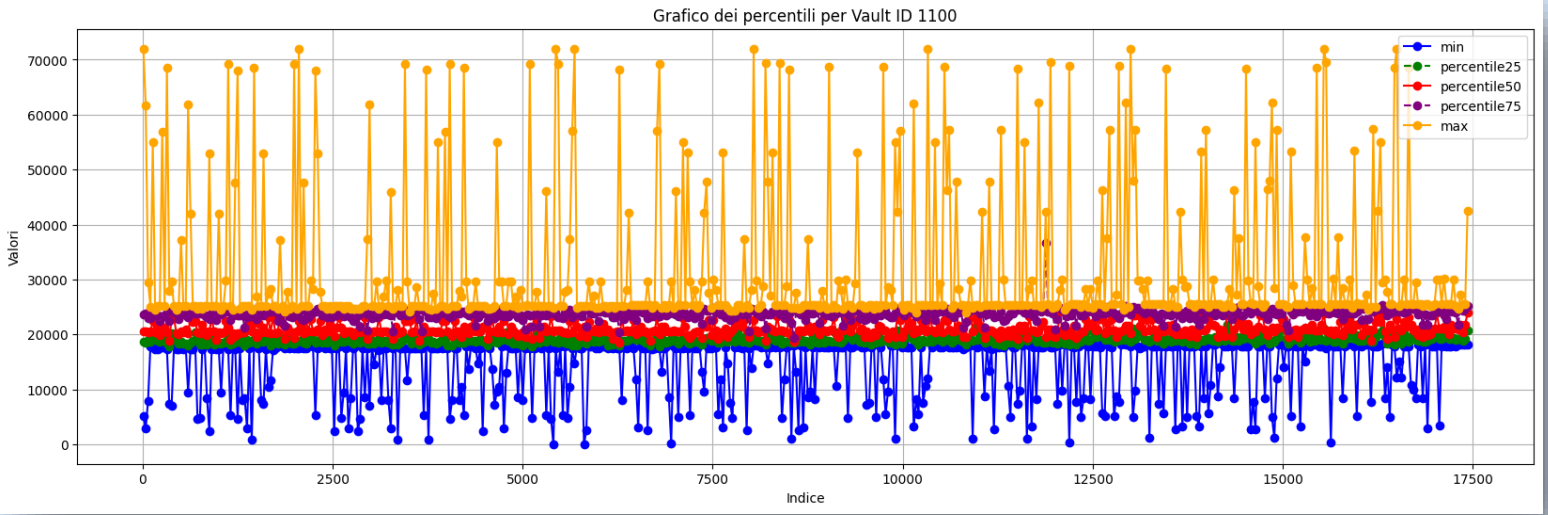


Grafico dei percentili per Vault ID 1100
min
percentile25
percentile50
percentile75
max
Valori
Indice
0
10000
20000
30000
40000
50000
60000
70000
0
2500
5000
7500
10000
12500
15000
17500

Latency & Throughput

- **Latency**: *tempo che intercorre tra quando i dati sono disponibili e quando vengono processati*

- **Throughput**: *quantità di dati processati nella finestra temporale*

Query 1

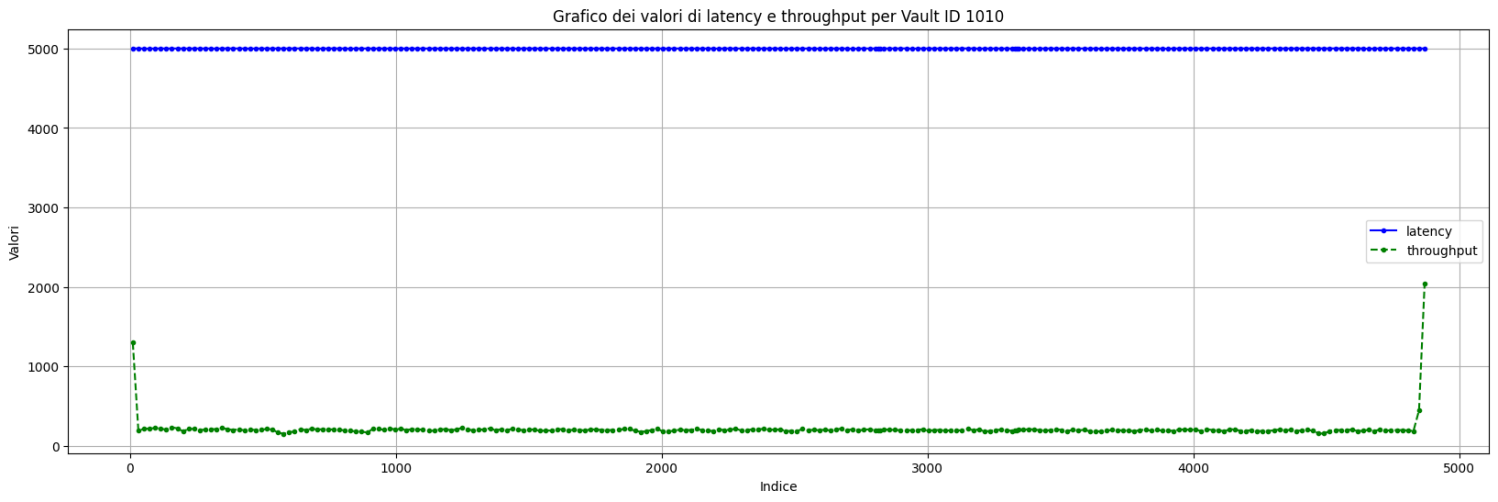


Grafico dei valori di latency e throughput per Vault ID 1010
Valori
Indice
latency
throughput
0
1000
2000
3000
4000
5000

Query 2

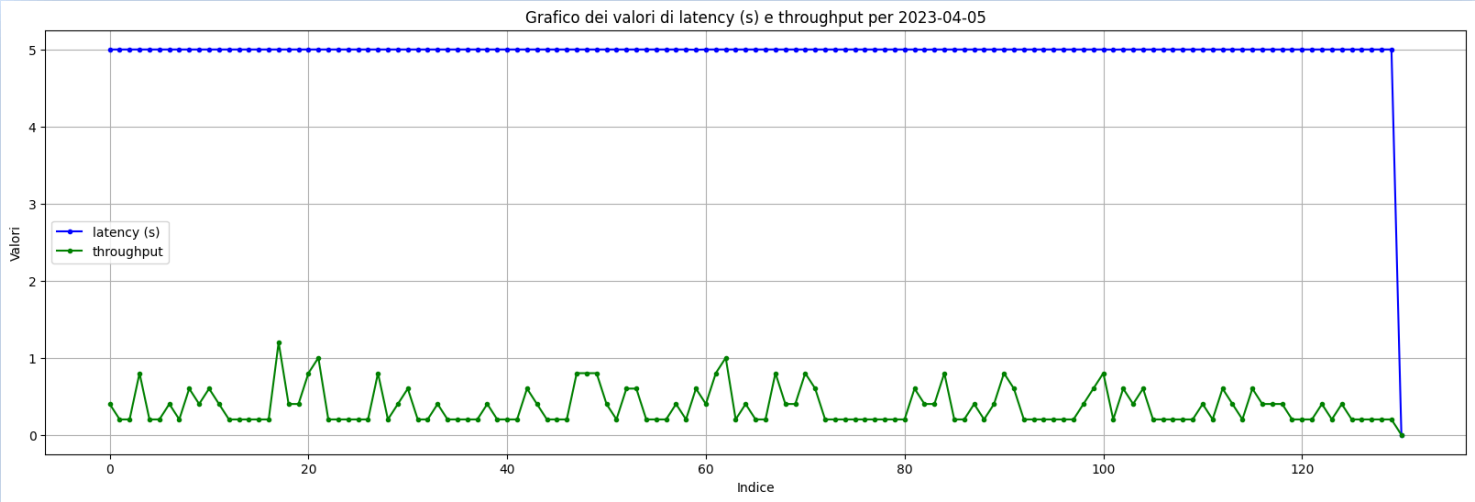


Grafico dei valori di latency (s) e throughput per 2023-04-05
latency (s)
throughput
Valori
Indice
0
1
2
3
4
5
0
20
40
60
80
100
120

Query 3

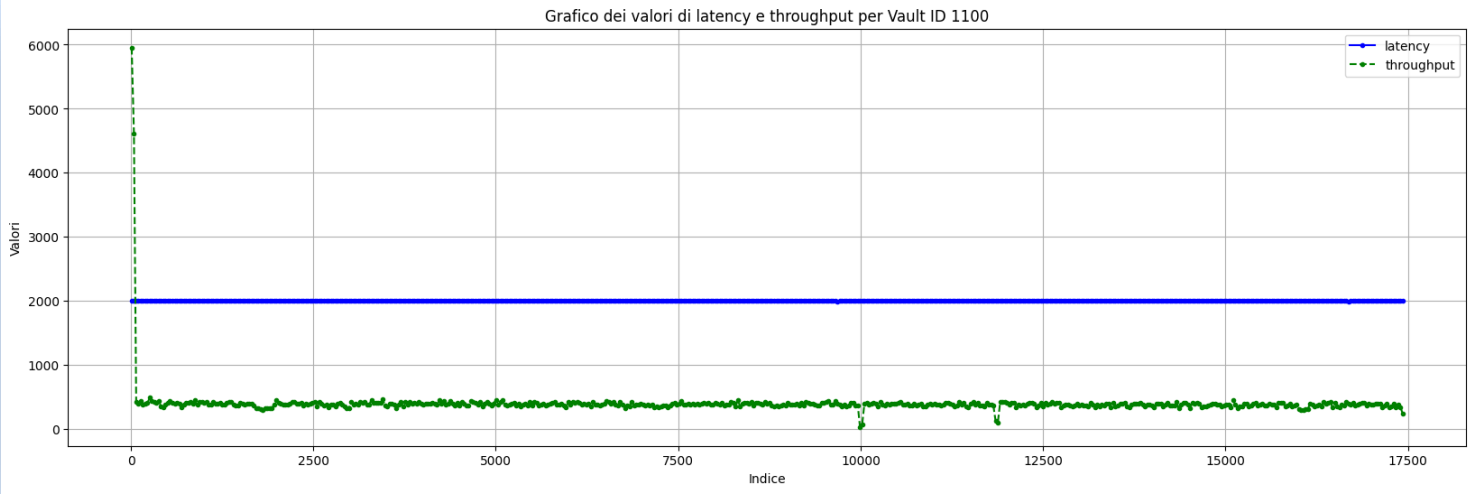
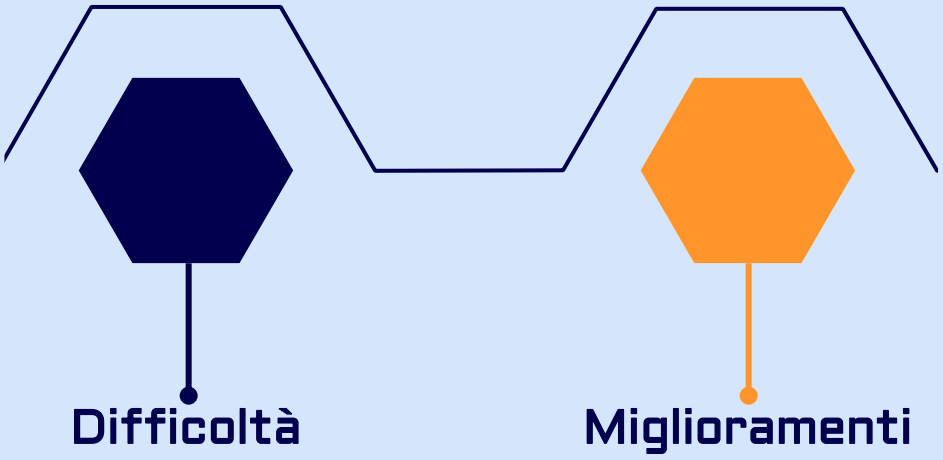


Grafico dei valori di latency e throughput per Vault ID 1100
latency
throughput
Valori
Indice
6000
5000
4000
3000
2000
1000
0
0
2500
5000
7500
10000
12500
15000
17500

Conclusioni



https://github.com/Pi3raldoSanturro/PyFlink

Grazie per l'attenzione!

https://github.com/Pi3raldoSanturro/SABD_Project2