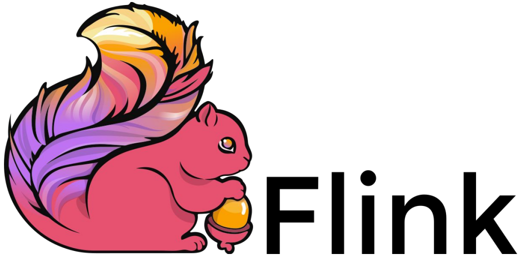
Progetto 2

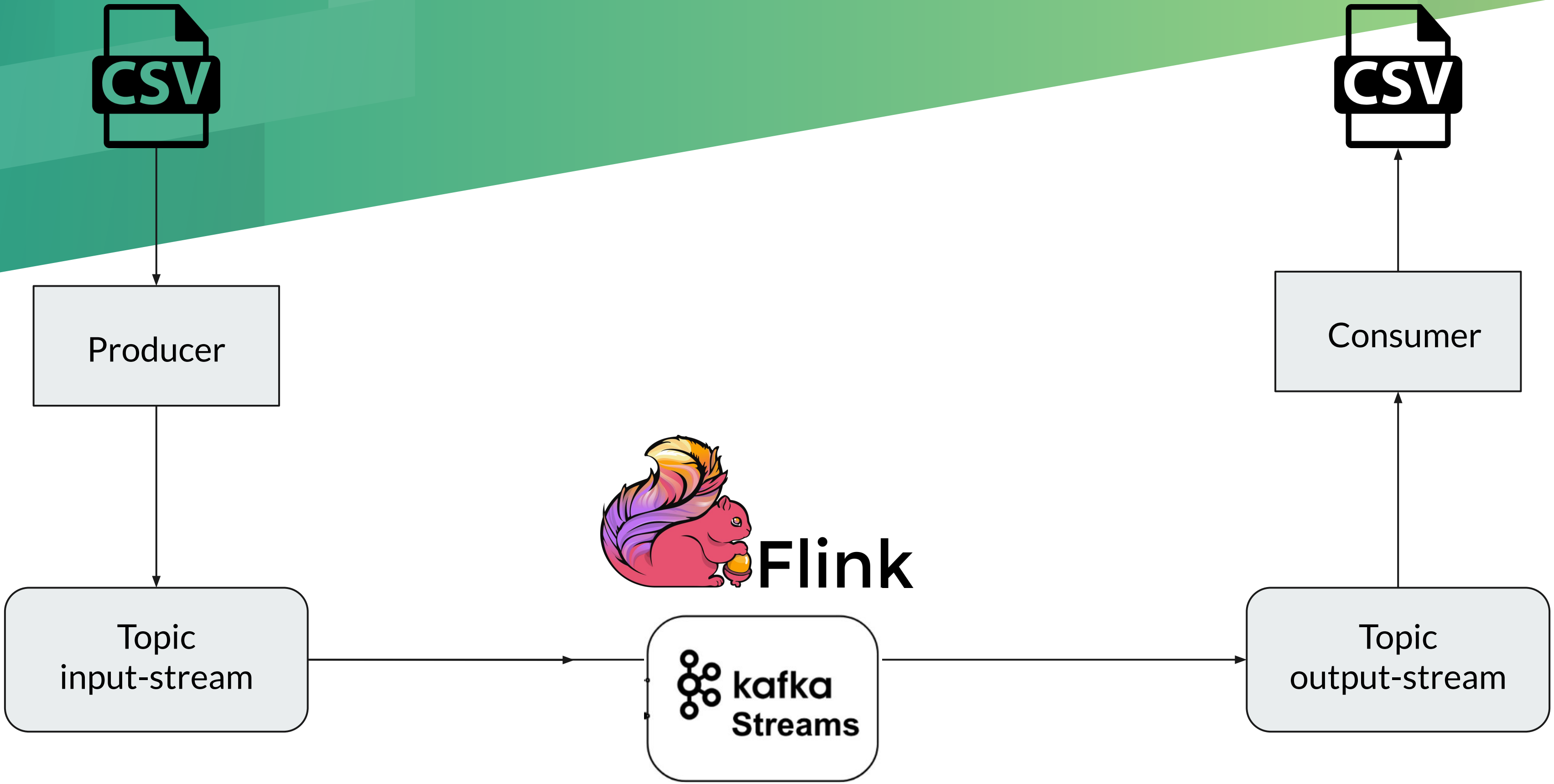
Damiano Nardi

Processamento delle query



kafka
Streams

Architettura



CSV

Producer

Topic
input-stream

Flink

kafka
Streams

Topic
output-stream

Consumer

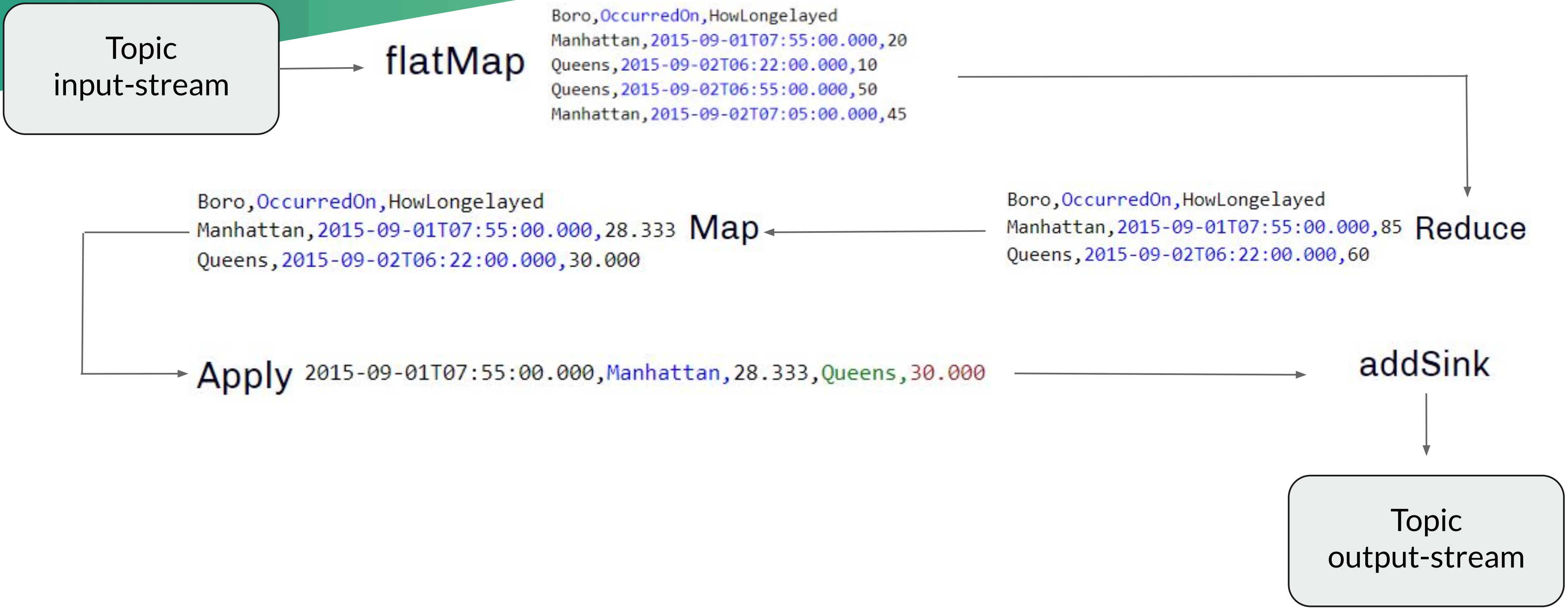
CSV

Query

Query1

1. Stream da Kafka topic
2. flatMap
3. put EventTime & TumblingEventTimeWindows(1,7,30)
4. Reduce
5. Map
6. TumblingEventTimeWindows(1,7,30)
7. Apply
8. addSink

Query1



```
Boro,OccurredOn,HowLongelayed
Manhattan,2015-09-01T07:55:00.000,20
Queens,2015-09-02T06:22:00.000,10
Queens,2015-09-02T06:55:00.000,50
Manhattan,2015-09-02T07:05:00.000,45
```

```
Boro,OccurredOn,HowLongelayed
Manhattan,2015-09-01T07:55:00.000,85
Queens,2015-09-02T06:22:00.000,60
```

```
Boro,OccurredOn,HowLongelayed
Manhattan,2015-09-01T07:55:00.000,28.333
Queens,2015-09-02T06:22:00.000,30.000
```

```
2015-09-01T07:55:00.000,Manhattan,28.333,Queens,30.000
```

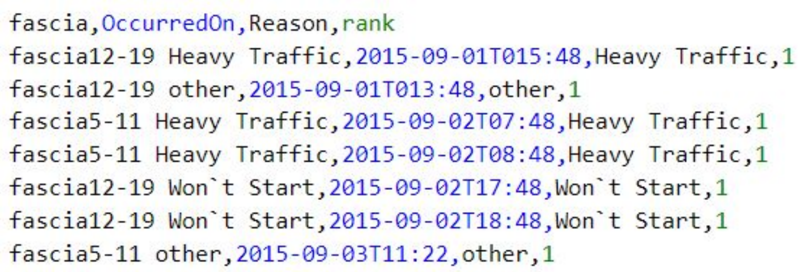
Query2

1. Stream da Kafka topic
2. flatMap
3. EventTime e TumblingEventTimeWindows(1,7)
4. Reduce
5. Map
6. TumblingEventTimeWindows(1,7)
7. Reduce
8. Map
9. TumblingEventTimeWindows(1,7).
10. Reduce & Map
11. addSink

Query2

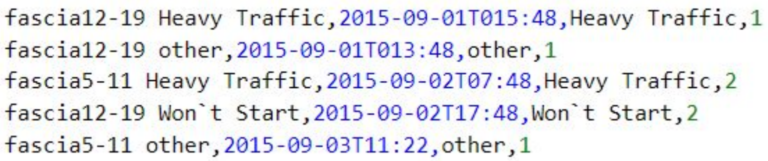
Topic input-stream

flatMap



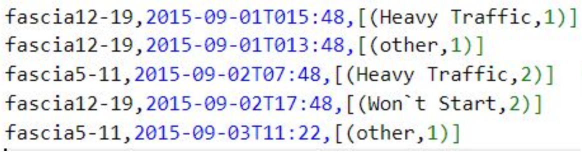
```
fascia,OccurredOn,Reason,rank
fascia12-19 Heavy Traffic,2015-09-01T015:48,Heavy Traffic,1
fascia12-19 other,2015-09-01T013:48,other,1
fascia5-11 Heavy Traffic,2015-09-02T07:48,Heavy Traffic,1
fascia5-11 Heavy Traffic,2015-09-02T08:48,Heavy Traffic,1
fascia12-19 Won`t Start,2015-09-02T17:48,Won`t Start,1
fascia12-19 Won`t Start,2015-09-02T18:48,Won`t Start,1
fascia5-11 other,2015-09-03T11:22,other,1
```

Reduce



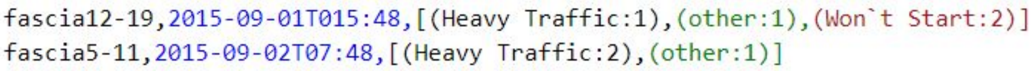
```
fascia12-19 Heavy Traffic,2015-09-01T015:48,Heavy Traffic,1
fascia12-19 other,2015-09-01T013:48,other,1
fascia5-11 Heavy Traffic,2015-09-02T07:48,Heavy Traffic,2
fascia12-19 Won`t Start,2015-09-02T17:48,Won`t Start,2
fascia5-11 other,2015-09-03T11:22,other,1
```

Map



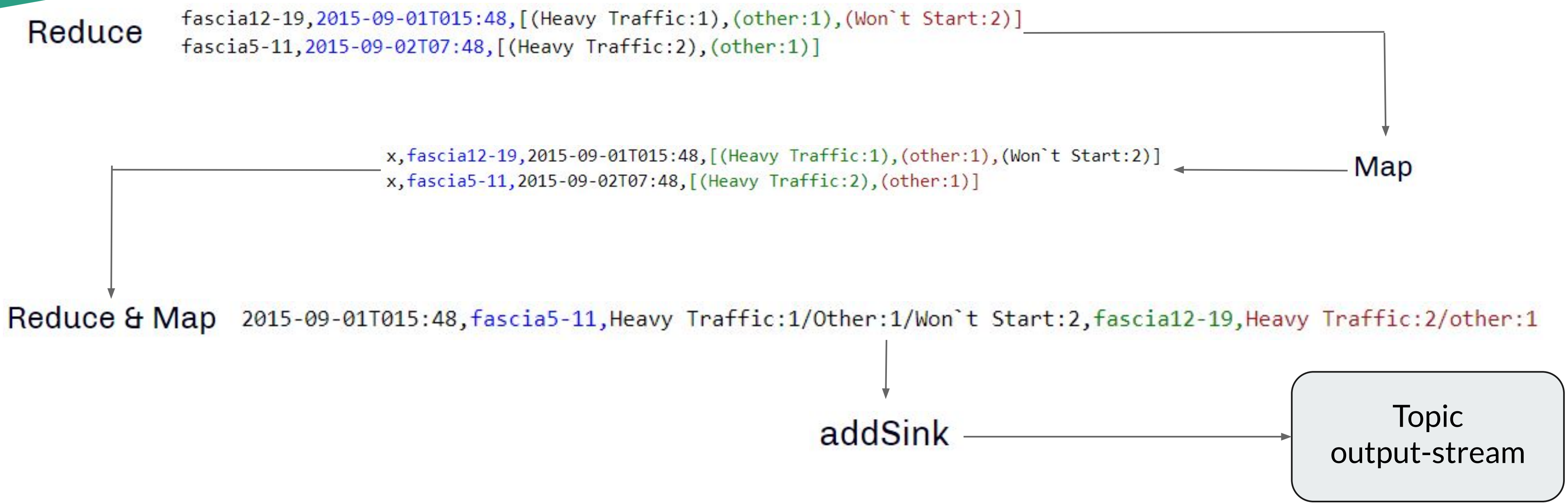
```
fascia12-19,2015-09-01T015:48,[(Heavy Traffic,1)]
fascia12-19,2015-09-01T013:48,[(other,1)]
fascia5-11,2015-09-02T07:48,[(Heavy Traffic,2)]
fascia12-19,2015-09-02T17:48,[(Won`t Start,2)]
fascia5-11,2015-09-03T11:22,[(other,1)]
```

Reduce



```
fascia12-19,2015-09-01T015:48,[(Heavy Traffic:1),(other:1),(Won`t Start:2)]
fascia5-11,2015-09-02T07:48,[(Heavy Traffic:2),(other:1)]
```

Query2



Reduce

```
fascia12-19,2015-09-01T015:48,[(Heavy Traffic:1),(other:1),(Won`t Start:2)]
fascia5-11,2015-09-02T07:48,[(Heavy Traffic:2),(other:1)]
```

Map

```
x,fascia12-19,2015-09-01T015:48,[(Heavy Traffic:1),(other:1),(Won`t Start:2)]
x,fascia5-11,2015-09-02T07:48,[(Heavy Traffic:2),(other:1)]
```

Reduce & Map

```
2015-09-01T015:48,fascia5-11,Heavy Traffic:1/Other:1/Won`t Start:2,fascia12-19,Heavy Traffic:2/other:1
```

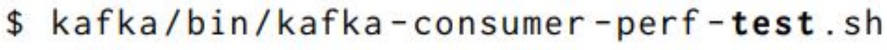
addSink

Topic
output-stream

Tempi

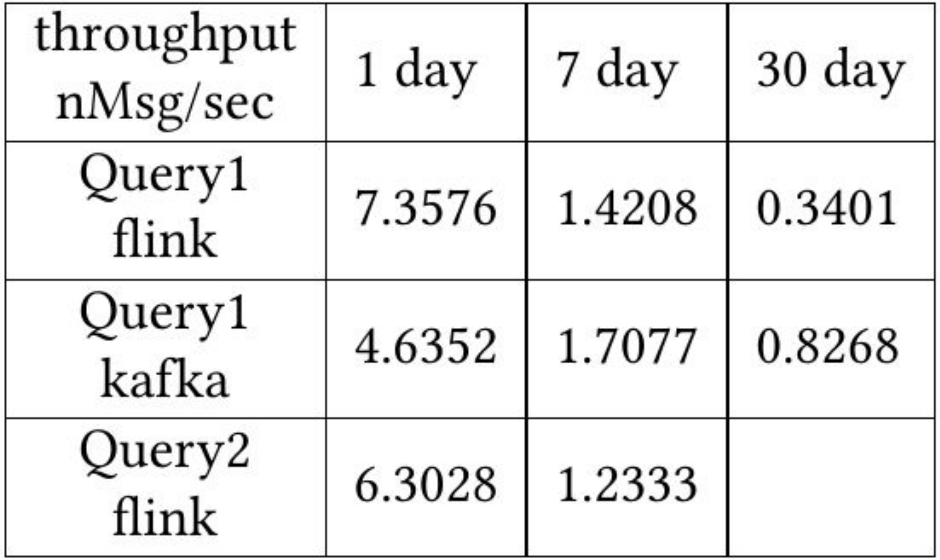
Throughput

Per misurare il throughput ho utilizzato questo comando di kafka:



```
$ kafka/bin/kafka-consumer-perf-test.sh
```

throughput del producer: 419.0764 nMsg/sec



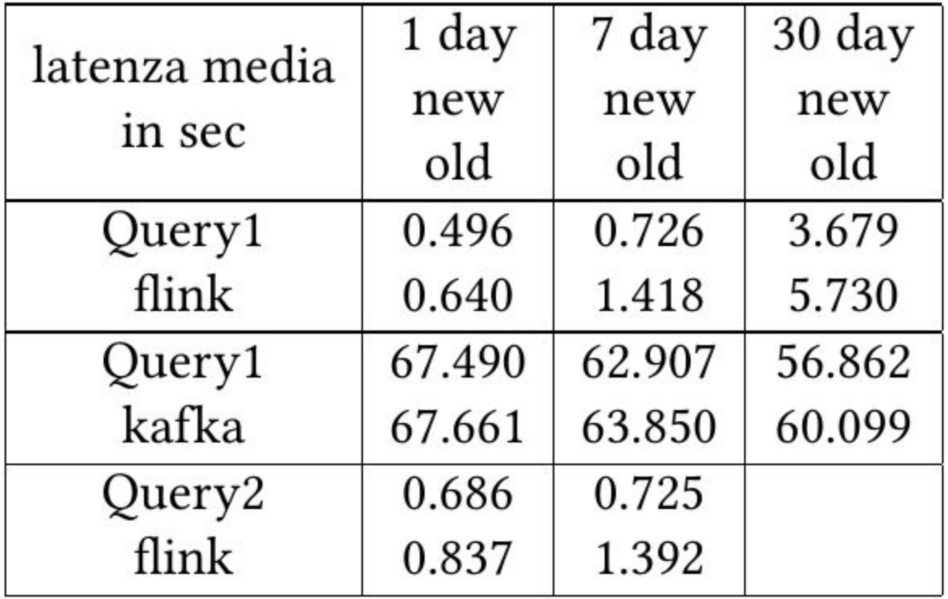
throughput nMsg/sec	1 day	7 day	30 day
Query1 flink	7.3576	1.4208	0.3401
Query1 kafka	4.6352	1.7077	0.8268
Query2 flink	6.3028	1.2333	

Latenza

Quando viene fatta una reduce andiamo ad unificare più righe del dataset che sono state inserite nella topica a tempi di processamento diversi

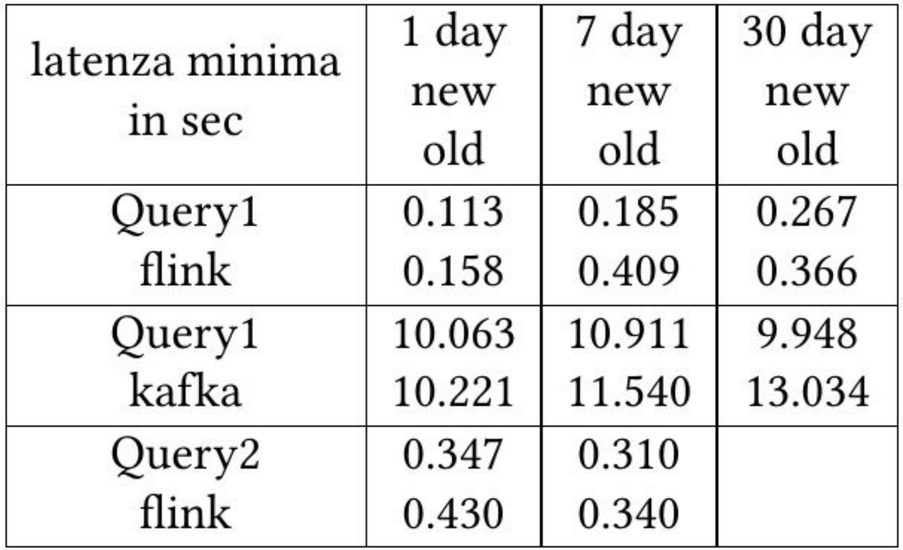
2 tipi di latenza:

1. Latenza “nuova”
2. Latenza “vecchia”

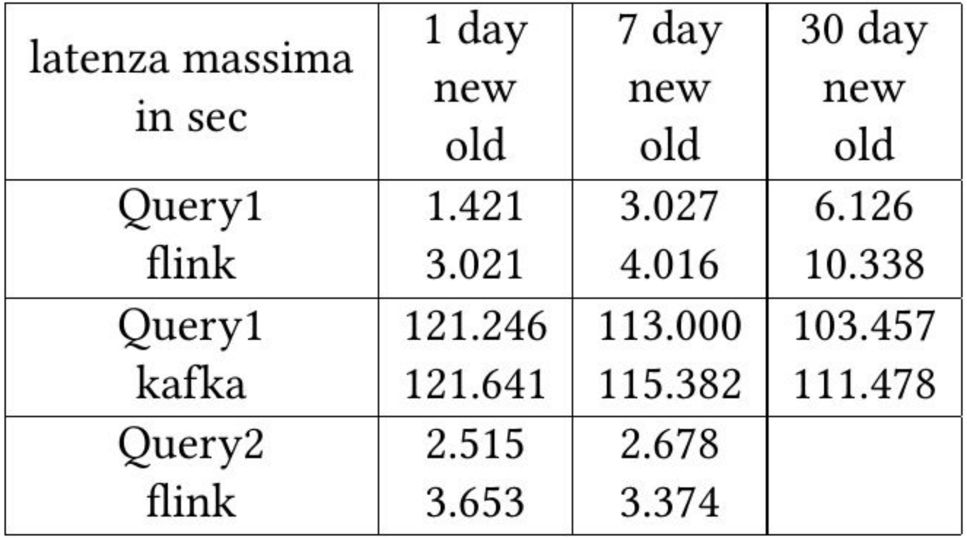


latenza media in sec	1 day new old	7 day new old	30 day new old
Query1 flink	0.496 0.640	0.726 1.418	3.679 5.730
Query1 kafka	67.490 67.661	62.907 63.850	56.862 60.099
Query2 flink	0.686 0.837	0.725 1.392	

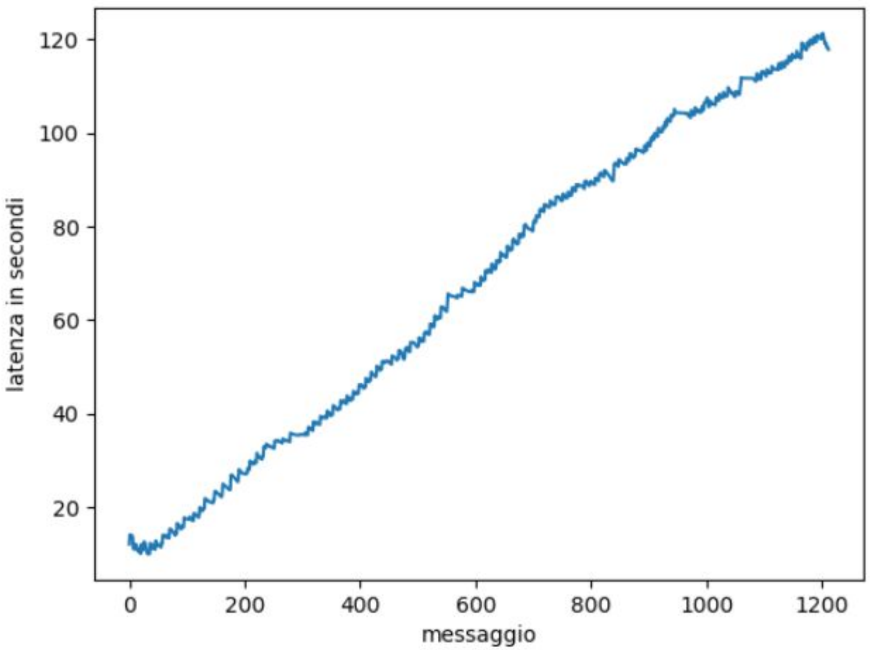
Latenza minima & massima



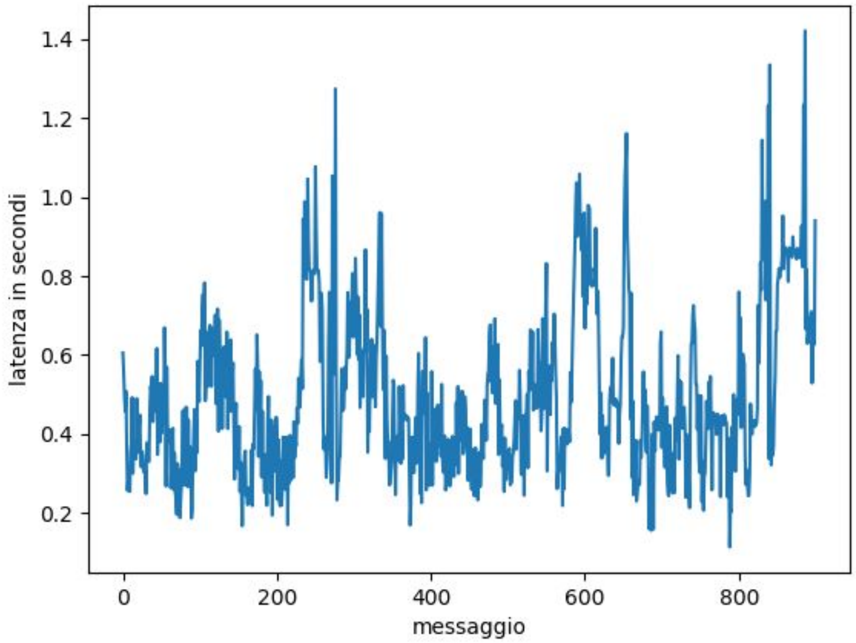
latenza minima in sec	1 day new old	7 day new old	30 day new old
Query1 flink	0.113 0.158	0.185 0.409	0.267 0.366
Query1 kafka	10.063 10.221	10.911 11.540	9.948 13.034
Query2 flink	0.347 0.430	0.310 0.340	



latenza massima in sec	1 day new old	7 day new old	30 day new old
Query1 flink	1.421 3.021	3.027 4.016	6.126 10.338
Query1 kafka	121.246 121.641	113.000 115.382	103.457 111.478
Query2 flink	2.515 3.653	2.678 3.374	

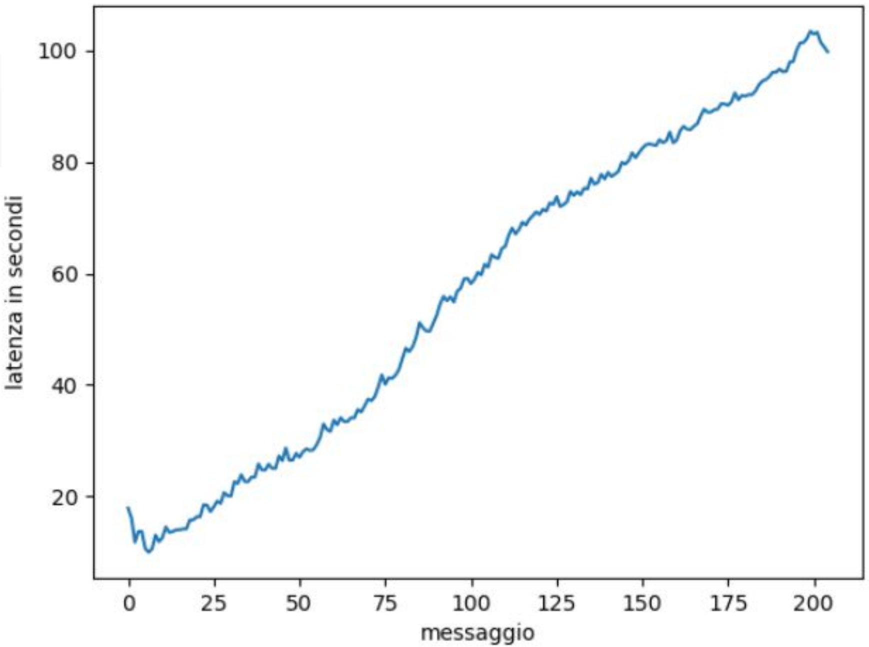


new-latency-1Day-Kafka-query1.csv

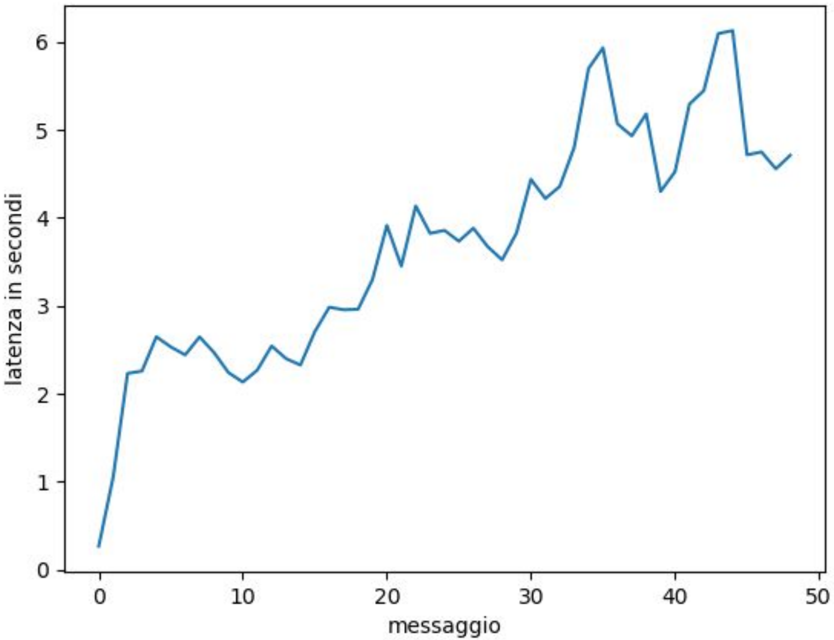


new-latency-1Day-Flink-Query1.csv

new-latency-30Day-Kafka-query1.csv



new-latency-30Day-Flink-query1.csv



Grazie dell’attenzione