FLIGHT DATA ANALYSIS

SECONDO PROGETTO BIG DATA

Alessandro Wood & Luca Gregori

Contesto dei Dati

Dati relativi ai voli domestici negli USA dal 2009 al 2019.

Raccolti dal Dipartimento dei Trasporti americano.

Un Dataset per ogni anno.

TECNOLOGIE UTILIZZATE

Spark, Spark Streaming, Spark MLlib

Kafka

HDFS Hadoop

Cassandra

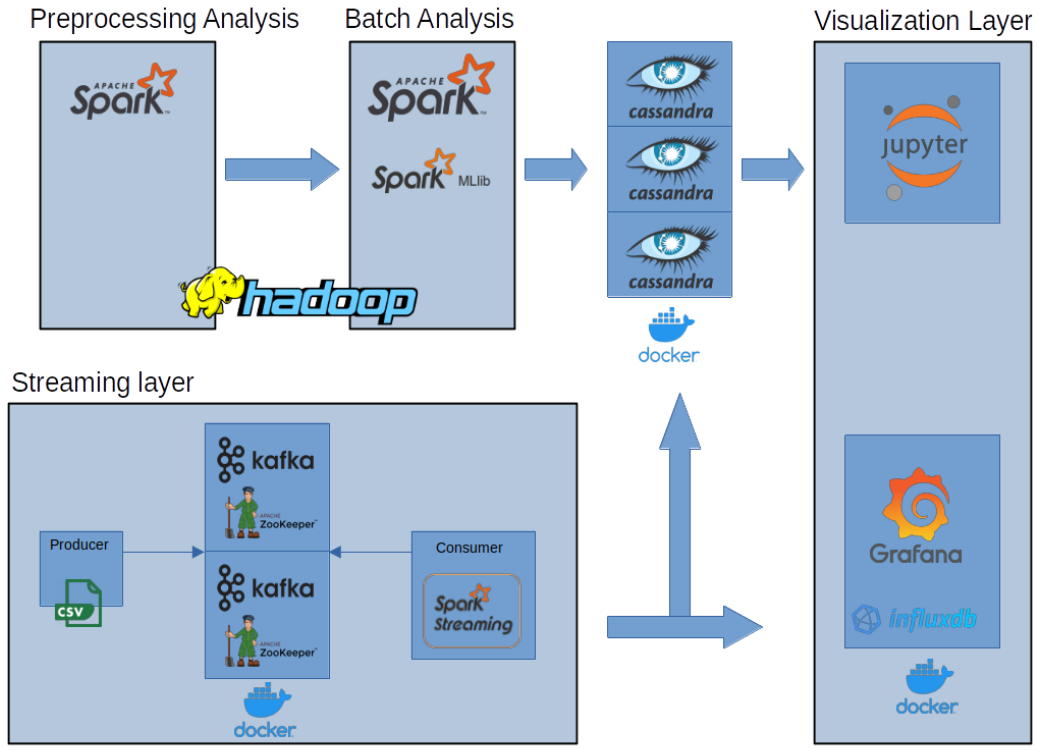
Docker

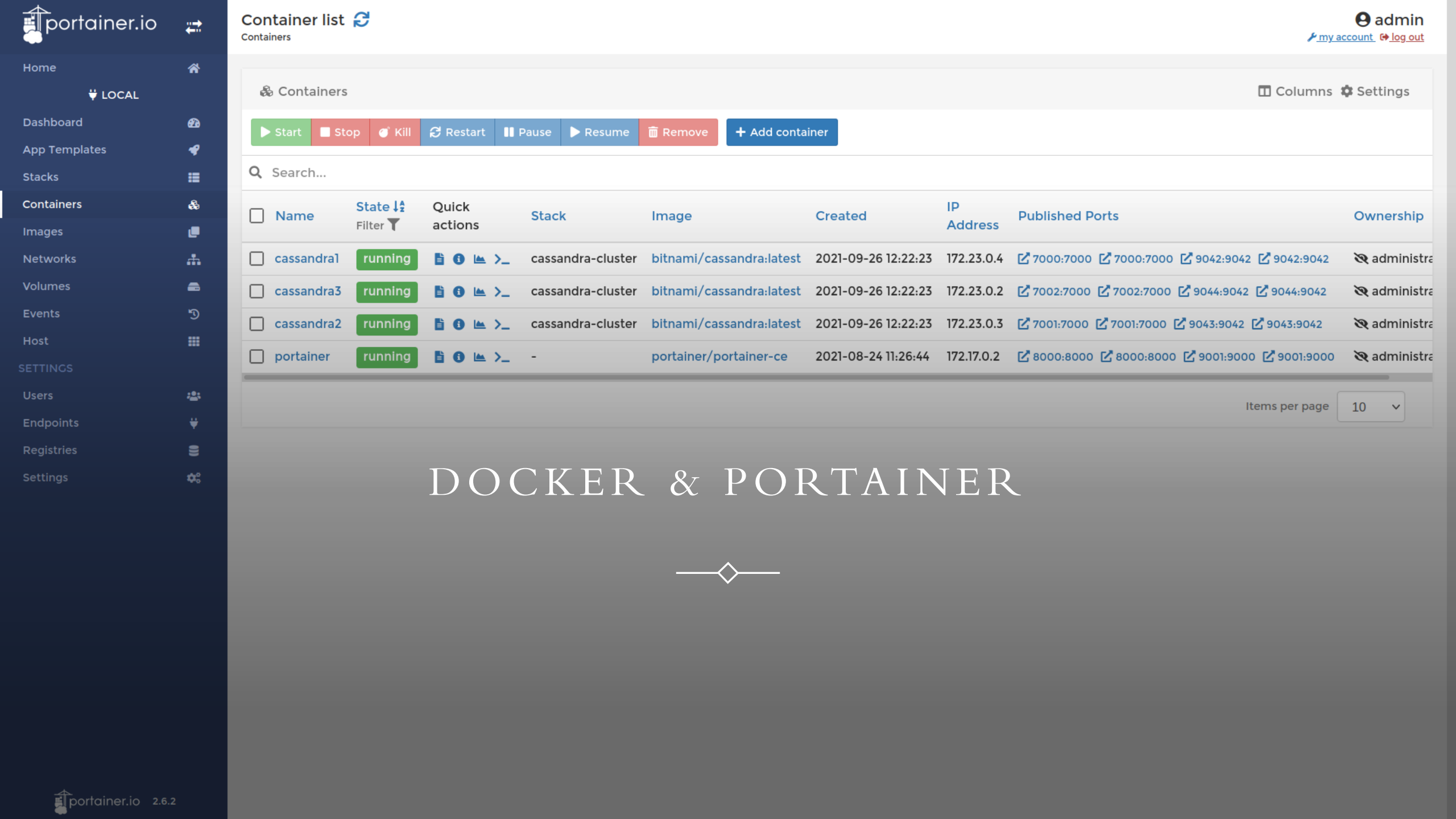
Portainer

Grafana

InfluxDB

ARCHITETTURA LAMBDA





DOCKER & PORTAINER

Preprocessing & Batch Layer

Pulizia dei Dati e Merge dei Dataset.

Batch Analysis : analisi effettuate con 4 diverse granularità temporali.

Salvataggio dei report su Cassandra.

Batch Layer: Spark MLlib

Preprocessamento dei dati tramite pipeline: applicazione di label encoder, vettorizzazione e standardizzazione.

Addestramento di un modello per predire voli in ritardo/cancellati.

STREAMING LAYER

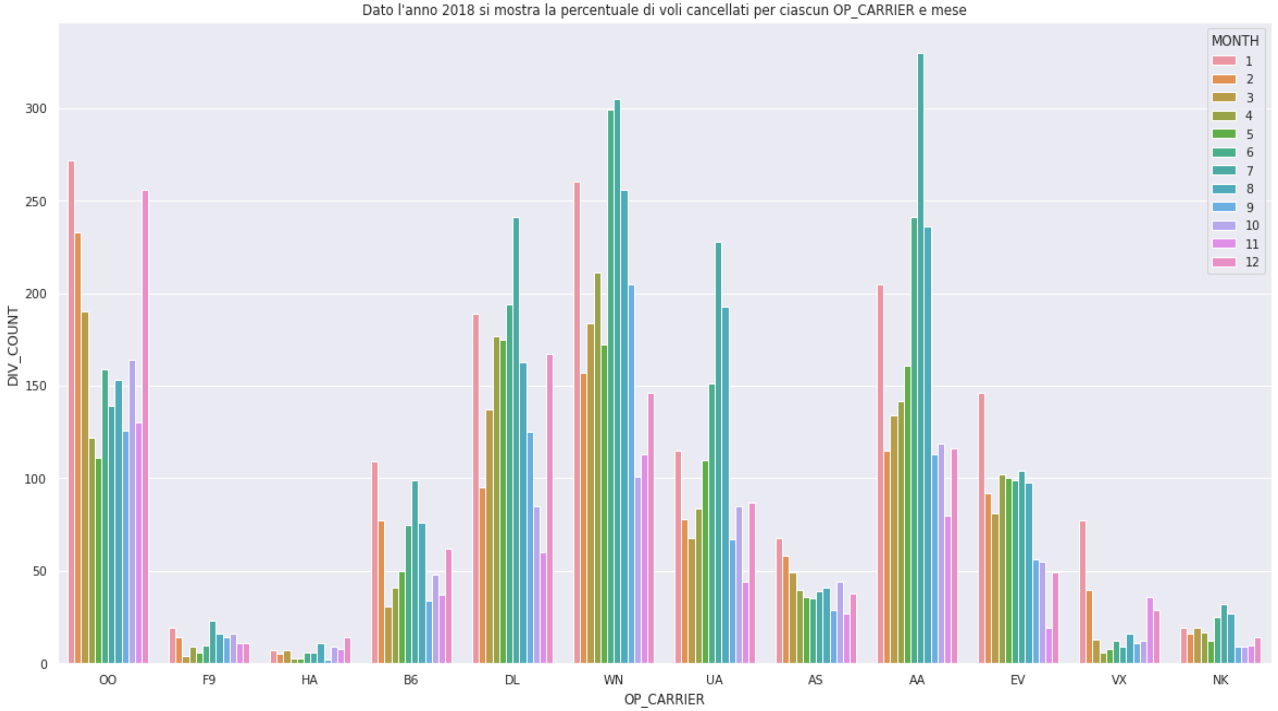
Kafka come modulo di data ingestion, simulando un feed di live-data.

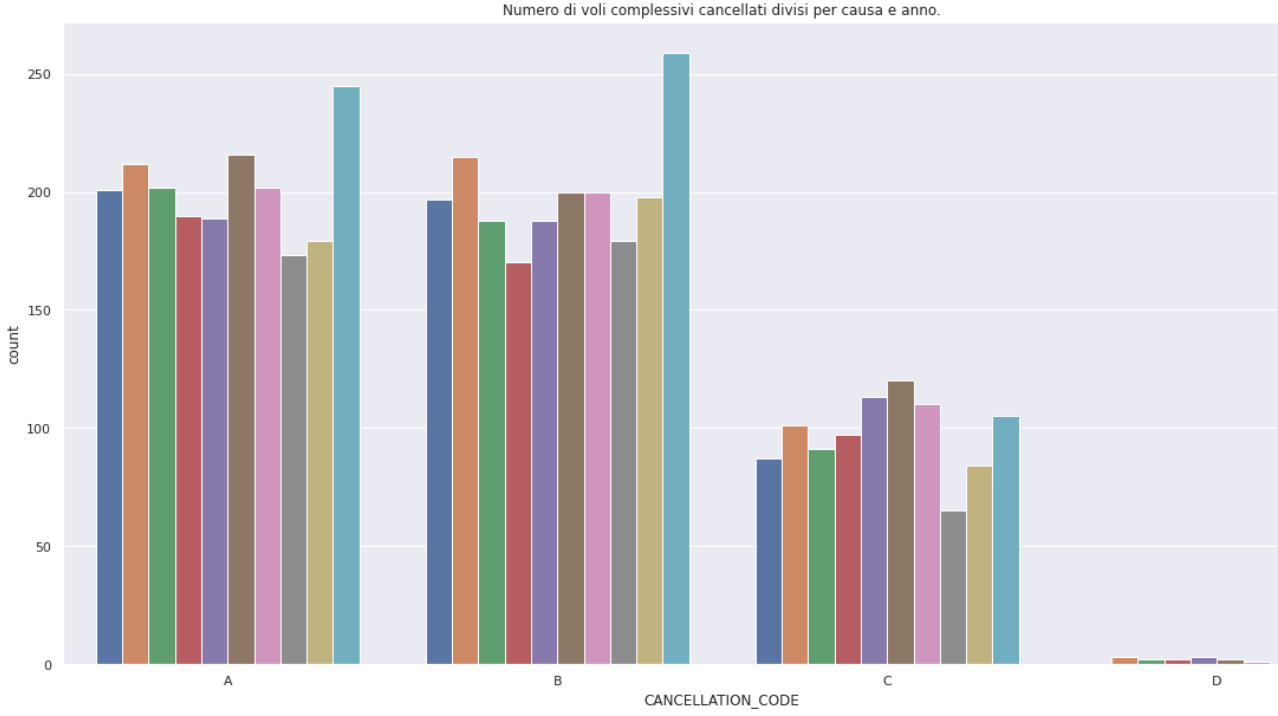
Delay Report salvati su Cassandra.

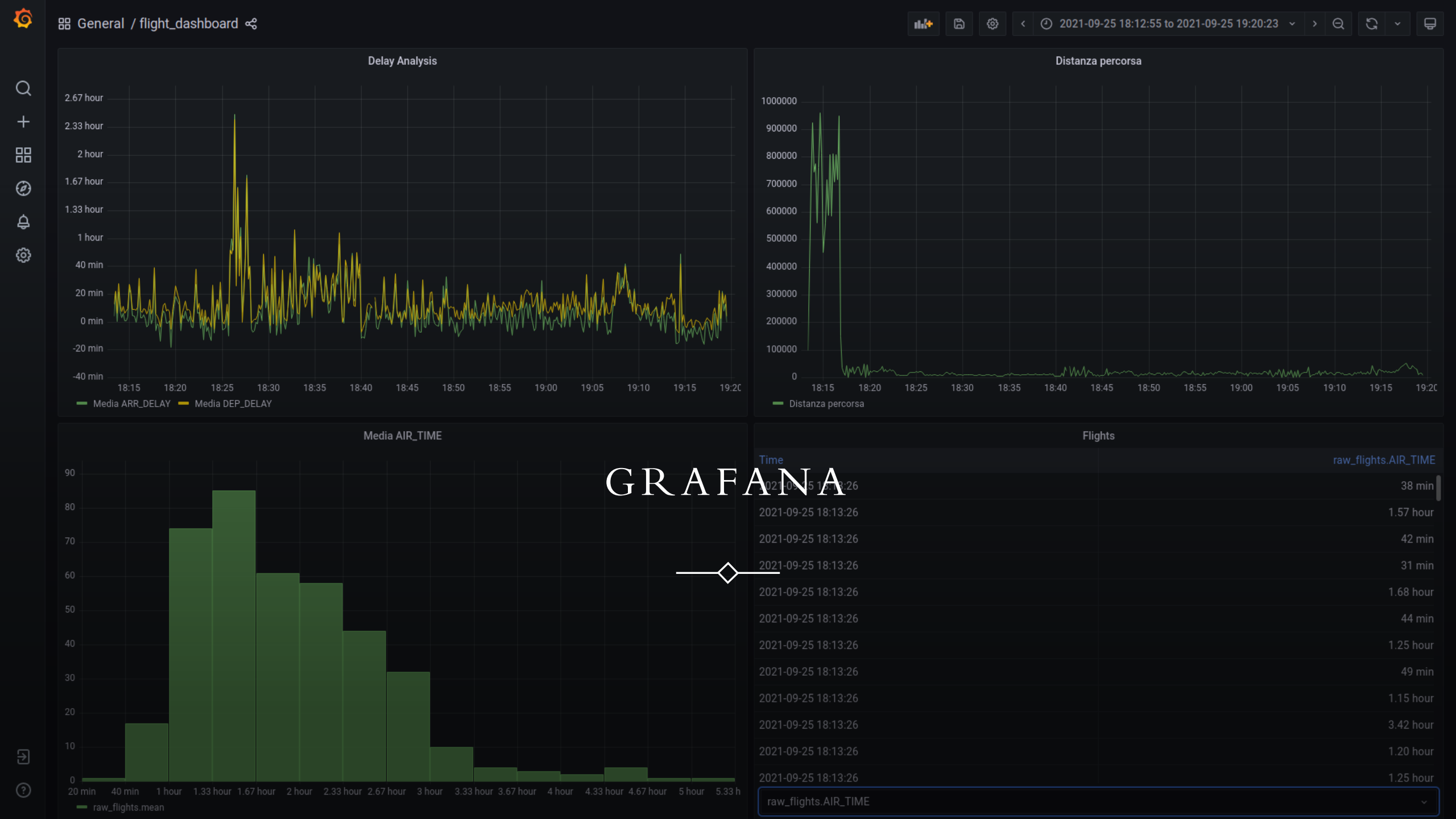
Preprocessing dei Dati Grezzi salvati sottoforma di Time-Series su InfluxDB per successiva visualizzazione.

VISUALIZATION LAYER

JUPYTER FOR BATCH ANALYSIS VISUALIZATION






General / flight_dashboard
2021-09-25 18:12:55 to 2021-09-25 19:20:23
Delay Analysis
2.67 hour
2.33 hour
2 hour
1.67 hour
1.33 hour
1 hour
40 min
20 min
0 min
-20 min
-40 min
18:15 18:20 18:25 18:30 18:35 18:40 18:45 18:50 18:55 19:00 19:05 19:10 19:15
Media ARR_DELAY
Media DEP_DELAY
Distanza percorsa
1000000
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
18:15 18:20 18:25 18:30 18:35 18:40 18:45 18:50 18:55 19:00 19:05 19:10 19:15
Distanza percorsa
Media AIR_TIME
90
80
70
60
50
40
30
20
10
0
20 min 40 min 1 hour 1.33 hour 1.67 hour 2 hour 2.33 hour 2.67 hour 3 hour 3.33 hour 3.67 hour 4 hour 4.33 hour 4.67 hour 5 hour
raw_flights.mean
GRAFANA
Flights
Time
raw_flights.AIR_TIME
2021-09-25 18:13:26 38 min
2021-09-25 18:13:26 1.57 hour
2021-09-25 18:13:26 42 min
2021-09-25 18:13:26 31 min
2021-09-25 18:13:26 1.68 hour
2021-09-25 18:13:26 44 min
2021-09-25 18:13:26 1.25 hour
2021-09-25 18:13:26 49 min
2021-09-25 18:13:26 1.15 hour
2021-09-25 18:13:26 3.42 hour
2021-09-25 18:13:26 1.20 hour
2021-09-25 18:13:26 1.25 hour
raw_flights.AIR_TIME

GRAZIE PER L'ATTENZIONE!