



SISTEMI E ARCHITETTURE PER BIG DATA

PROGETTO 2

ANALISI DEI RITARDI E DEI
GUASTI DEL TRASPORTO
SCOLASTICO DI NYC

Marco Balletti, Francesco Marino

SOMMARIO



01

Architettura

Kafka, Consumers,
Producer, Client, Kafka
Streams, Flink

02

Queries

Risoluzione delle query
con Flink e Kafka Streams

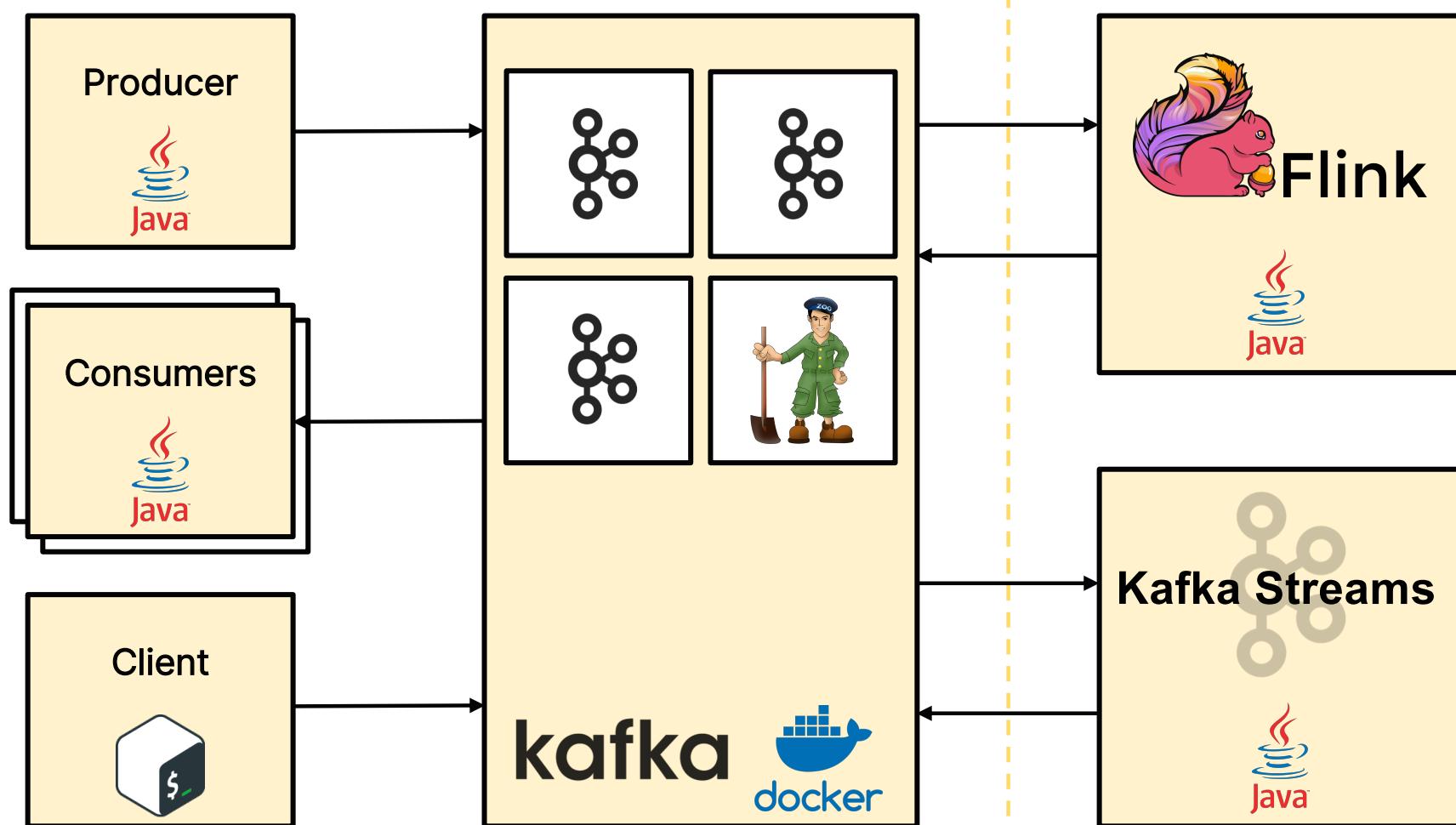
03

Benchmark

Confronto tra Flink e
Kafka Streams

ARCHITETTURA

01



INGESTIONE E SALVATAGGIO

D.S.P.

CONTAINER

BROKER

effeerre/kafka

Funzione:
Kafka Cluster

Modalità di esecuzione:
Tre container

ZOOKEEPER

zookeeper

Funzione:
Coordinazione del cluster
Kafka

Modalità di esecuzione:
Singolo container

PRODUCER

Sorgente real-time

Funzione:

Pubblicazione delle tuple
in maniera ritardata per
simulare una sorgente
real-time

Modalità di esecuzione:
Singola entità

CONSUMER

Salvataggio output

Funzione:

Salvataggio dei risultati in
formato CSV o stampa
delle tuple a schermo

Modalità di esecuzione:
Molteplici thread

02

ESECUZIONE DELLE QUERY

PARSING DEI RITARDI

Traduzione di stringhe indicanti intervalli temporali in linguaggio naturale.

Libreria locale costruita ad-hoc per lo scopo:

- Maggiore **leggerezza** del programma
- **Performance**
- **Accuratezza**
- Tolleranza agli **errori di battitura**

PARSING DEI RITARDI

1) Ore



$([0-9]^+)(h.^*)$

2) Minuti



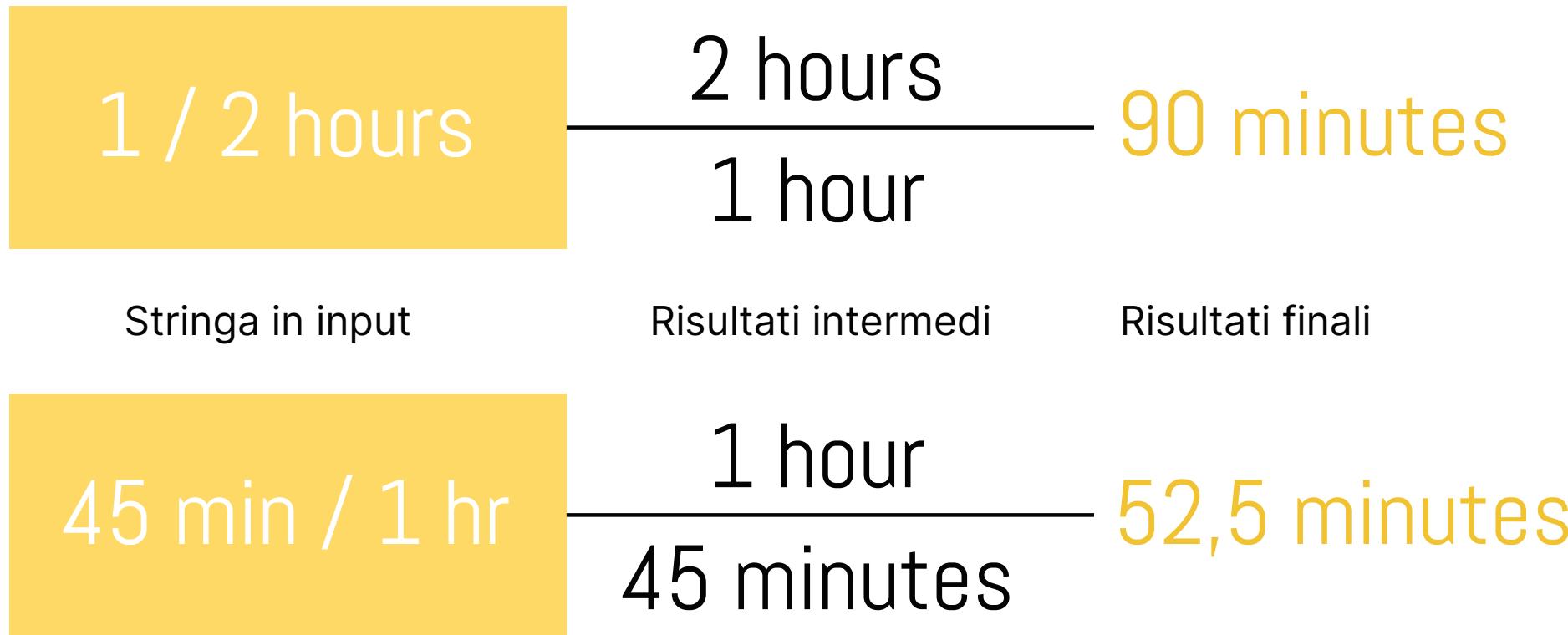
$([0-9]^+)(m.^*)$

3) Valori isolati



$([0-9]^+\$)$

PARSING DEI RITARDI



PARSING DEI RITARDI

11

30min0

30 minutes

20

20 minutes

235d

10-12

10-dic

TEMPO

Flink

Assegnamento **event time** e gestione dei **watermark** sulla base di un campo interno alla tupla.

Kafka Streams

Assegnamento **event time** affidato al producer sulla base di un campo interno alla tupla.

FINESTRE TEMPORALI CUSTOM

Flink

Finestre temporali **mensili** personalizzate:

- **Tumbling window**
- Size dinamica
- **Assigner** personalizzato:
 - Tupla → Event time → Mese di riferimento → Window
 - Corretto funzionamento a prescindere dal numero di giorni dello specifico mese

FINESTRE TEMPORALI CUSTOM

Kafka Streams

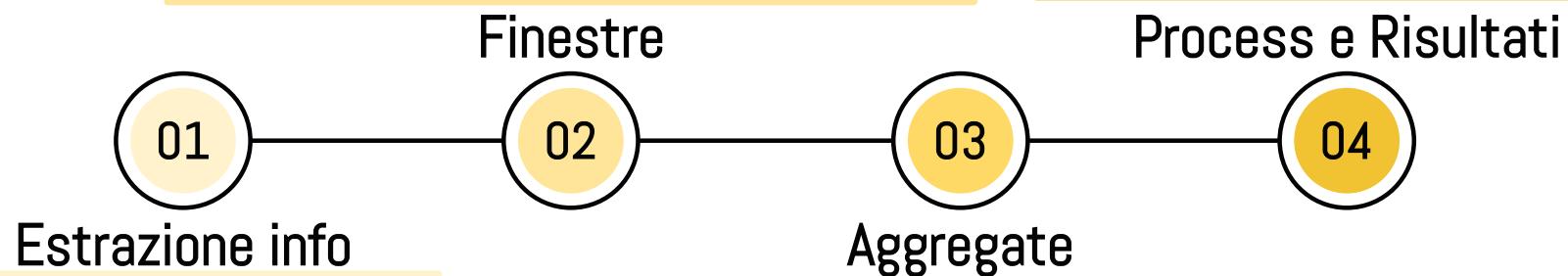
Finestre temporali personalizzate:

- Allineate alla mezzanotte di una time-zone specificata
- **Giornaliere**
- **Settimanali**
 - Allineate al lunedì (ore 00:00.000)
 - con termine di domenica (ore 23:59.999)
- **Mensili**
 - Allineate al primo giorno del mese (ore 00:00.000)
 - con termine ultimo giorno del mese (ore 23:59.999)

QUERY 1 - FLINK

- Giornaliere → Tumbling window di 1 giorno con offset pari a 4 ore
- Settimanali → Tumbling window di 7 giorni con offset pari a 4 ore
- Mensili → Custom

Reperimento corretta data di inizio finestra, conversione risultato in stringa e pubblicazione sul Kafka topic



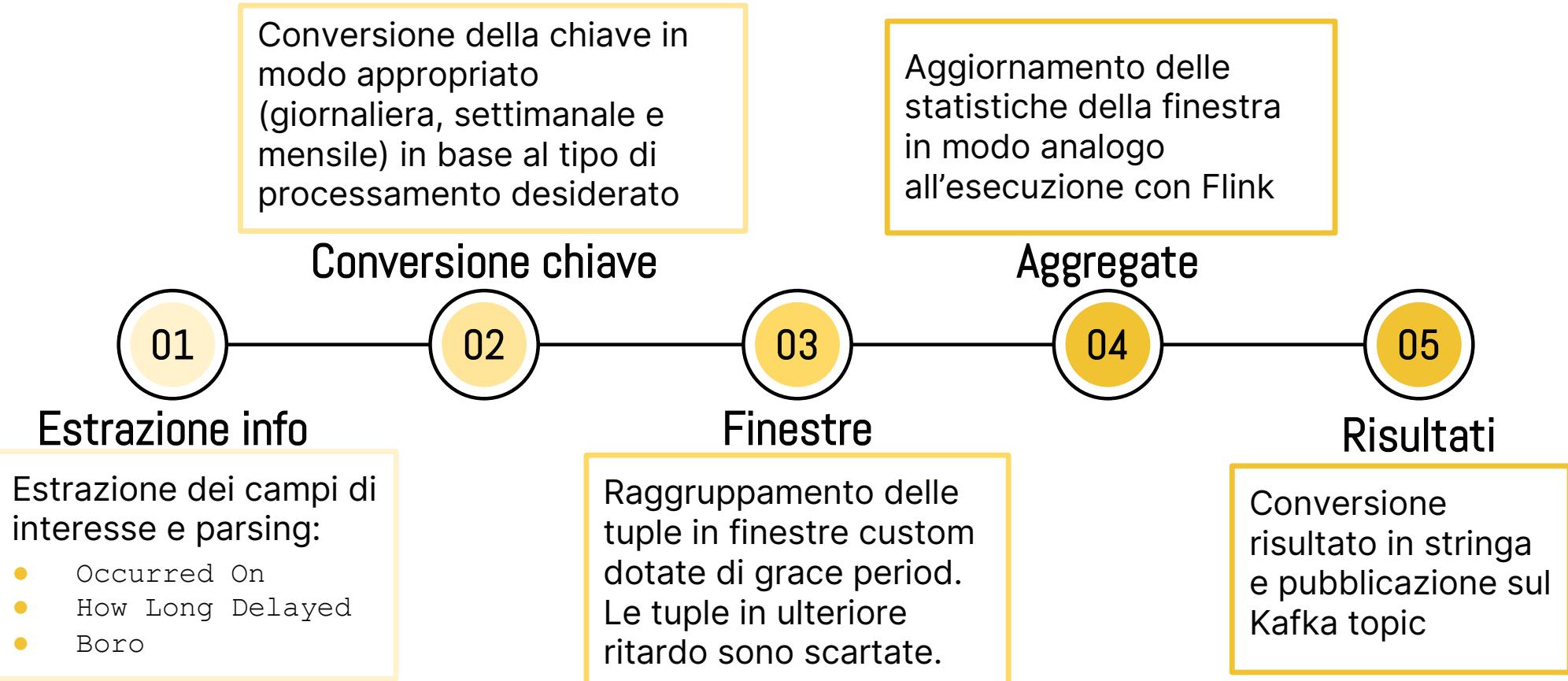
Estrazione dei campi di interesse e parsing:

- Occurred On
- How Long Delayed
- Boro

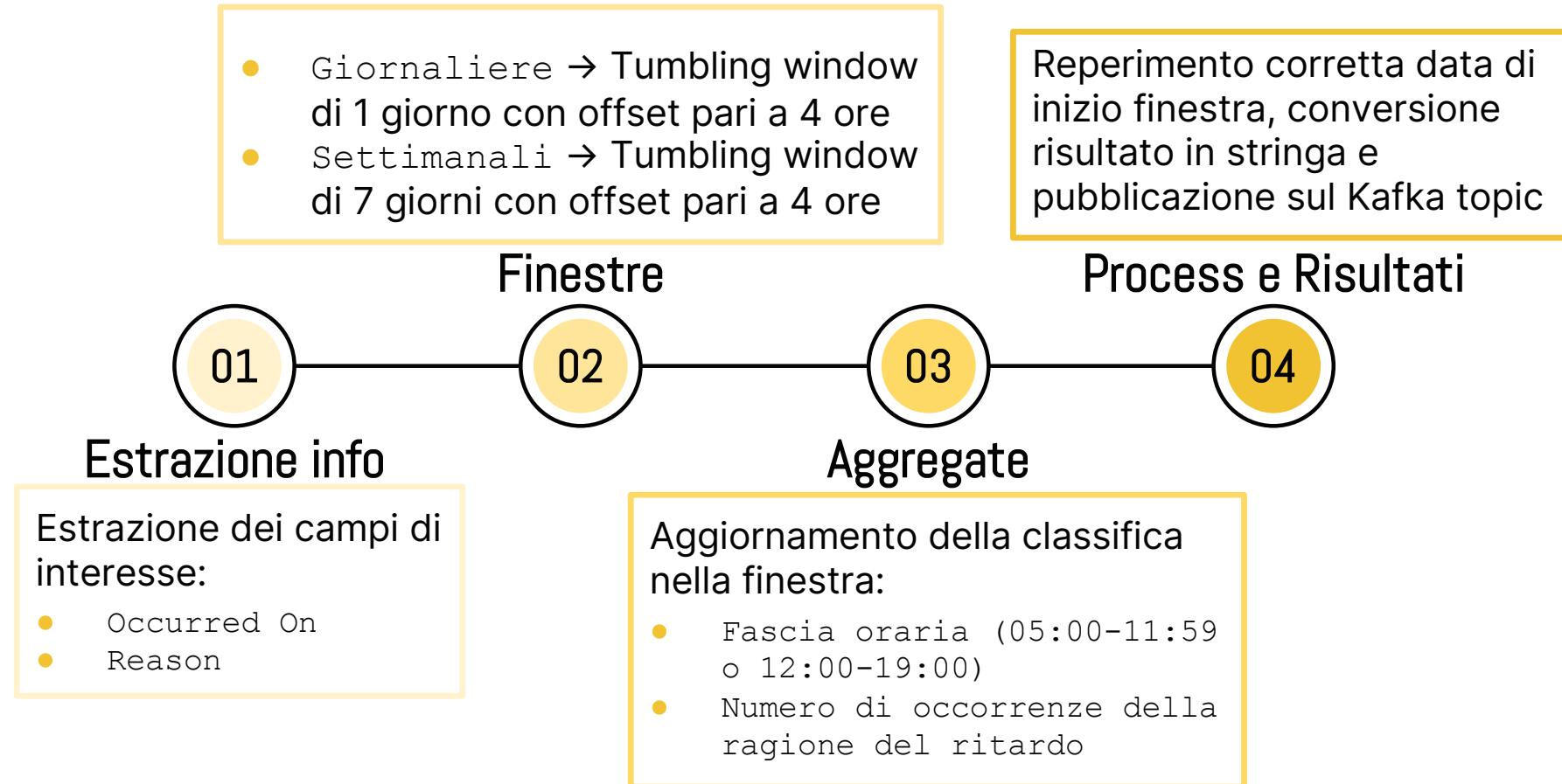
Aggiornamento delle statistiche della finestra:

- Quartiere/contea
- Tempo ritardo totale
- Totale disservizi

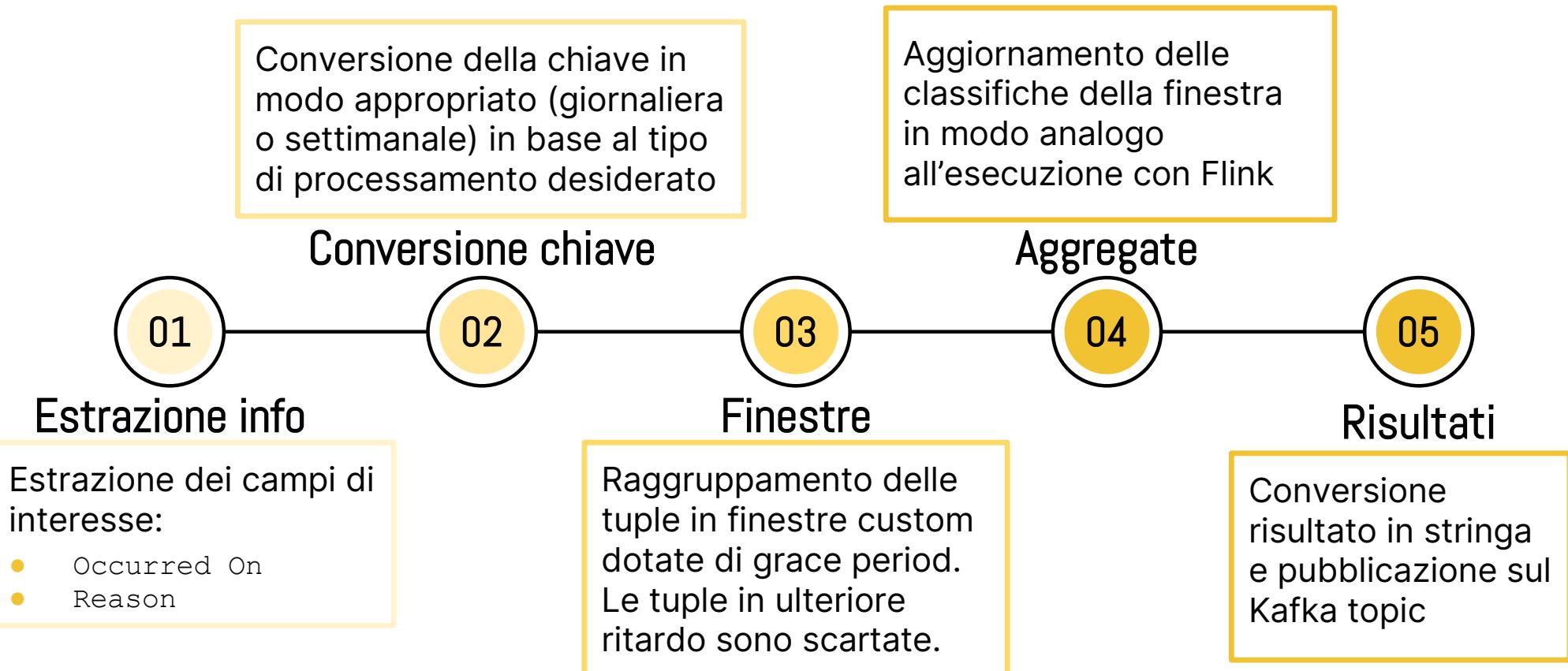
QUERY 1 – KAFKA STREAMS



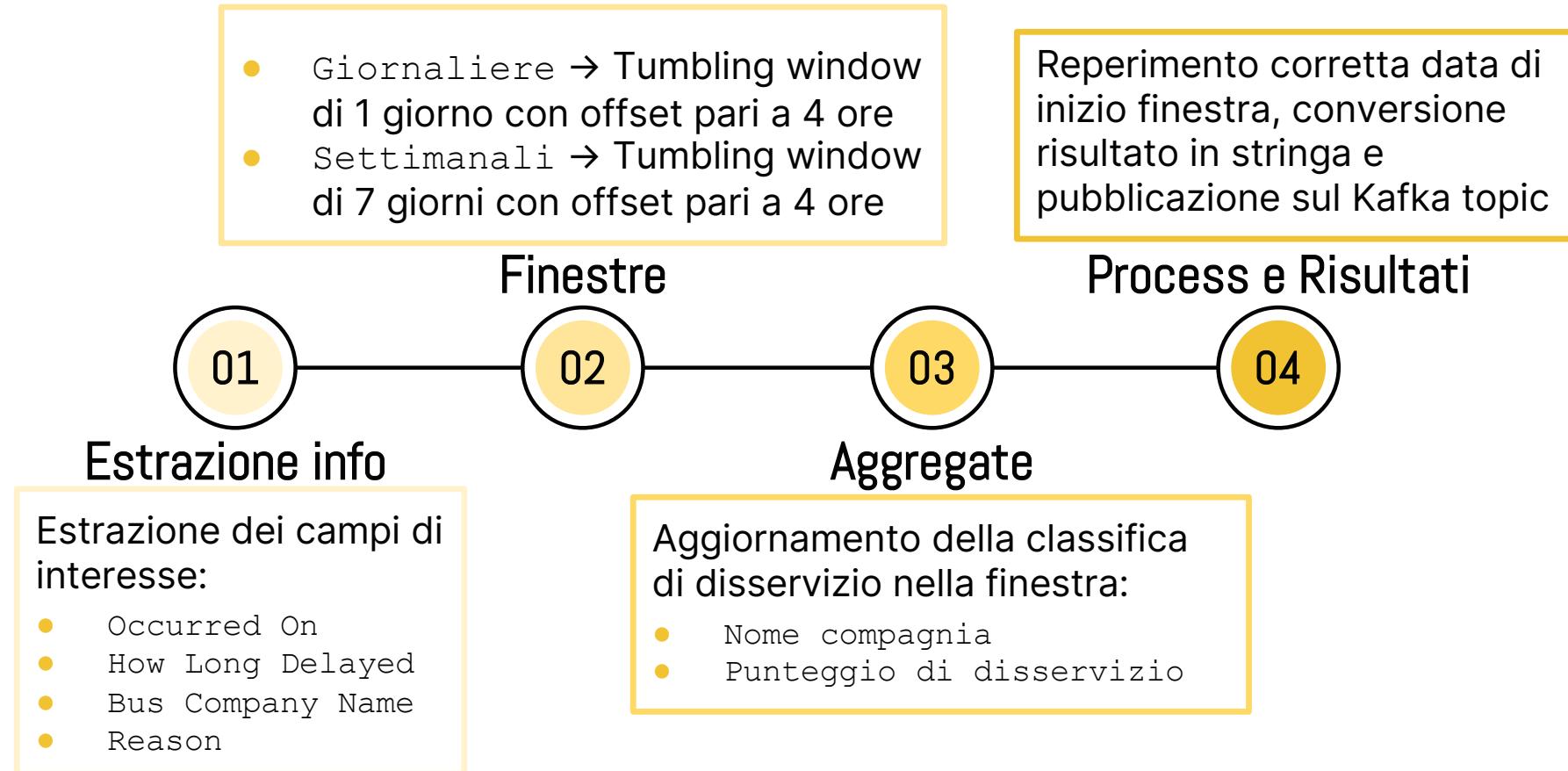
QUERY 2 - FLINK



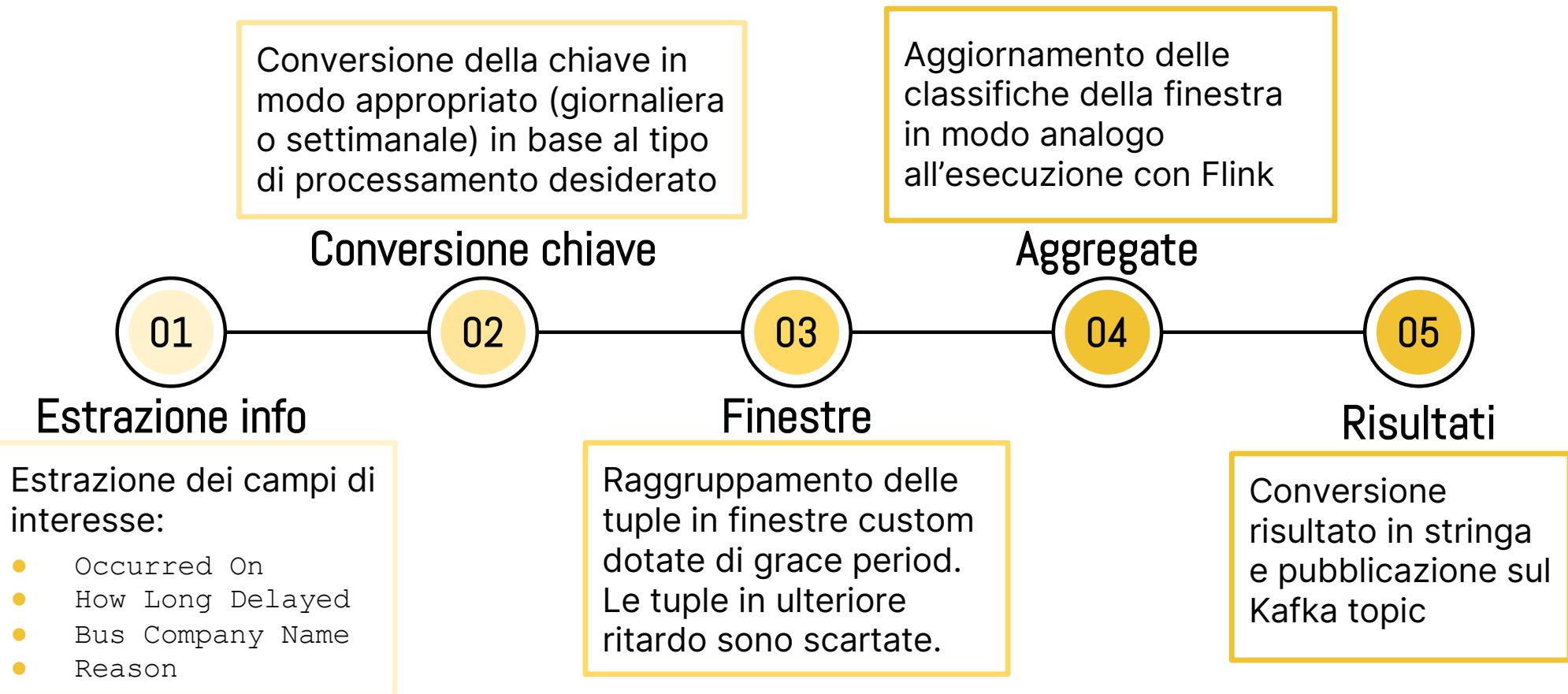
QUERY 2 - KAFKA STREAMS



QUERY 3 - FLINK



QUERY 3 - KAFKA STREAMS



BENCHMARK

03

STRUMENTI DI VALUTAZIONE DI PERFORMANCE



Auto-Monitoring

Meccanismi di valutazione di throughput e latenza offerti dai framework

Pro:

Meccanismo automatizzato

Contro:

Potrebbe non fornire metriche precise (Kafka Streams)



Sperimentale

Utilizzo di una struttura statica con metodi `synchronized` per l'aggiornamento di contatori di stato

Pro:

Flessibilità nella definizione delle metriche

Contro:

Integrazione (Sink alternativi o in funzioni di map)

FLINK	THROUGHPUT FLINK UI (tuple/sec)	THROUGHPUT SPERIMENTALE (tuple/sec)	LATENZA SPERIMENTALE (sec/tupla)
QUERY 1 DAILY	0,338	0,285	3,508
QUERY 1 WEEKLY	0,071	0,062	16,05
QUERY 1 MONTHLY	0,016	0,016	60,8
QUERY 2 DAILY	0,344	0,303	3,35
QUERY 2 WEEKLY	0,071	0,075	15,7
QUERY 3 DAILY	0,338	0,283	3,575
QUERY 3 WEEKLY	0,071	0,064	15,83

KAFKA STREAMS

THROUGHPUT SPERIMENTALE
(tuple/sec)

LATENZA SPERIMENTALE
(sec/tupla)

QUERY 1 DAILY

0,344

3,04

QUERY 1 WEEKLY

0,064

15,8

QUERY 1 MONTHLY

0,019

52,05

QUERY 2 DAILY

0,332

3,01

QUERY 2 WEEKLY

0,076

15,2

QUERY 3 DAILY

0,321

3,12

QUERY 3 WEEKLY

0,066

15,23

CONCLUSIONI

FLINK

Prestazioni leggermente inferiori rispetto a Kafka Streams

Garanzie su ordinamento delle tuple e processamento

Maggiore accuratezza

KAFKA STREAMS

Throughput e latenze migliori

Assenza di garanzie di ordinamento

Minore accuratezza

GRAZIE PER L'ATTENZIONE!

Referenze:

1. http://www.ce.uniroma2.it/courses/sabd1920/projects/prj2_dataset.zip
2. <https://ci.apache.org/projects/flink/flink-docs-stable/dev/stream/operators/windows.html>
3. [Template Powerpoint](#)

Marco Balletti
Francesco Marino

