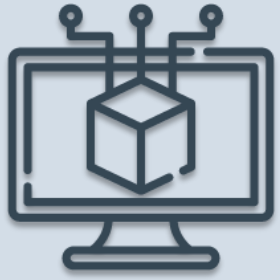
Sistemi e architetture per big data

Progetto 2 - Analisi dei dati provenienti da dispositivi Automatic Identification System (AIS)

Giuseppe Lasco

Marco Marcucci

Sommario



Architettura

- Producer
- Kafka
- Flink
- Consumer

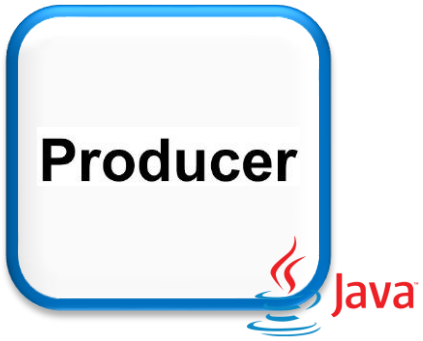
Query

- Query 1
- Query 2
- Query 3

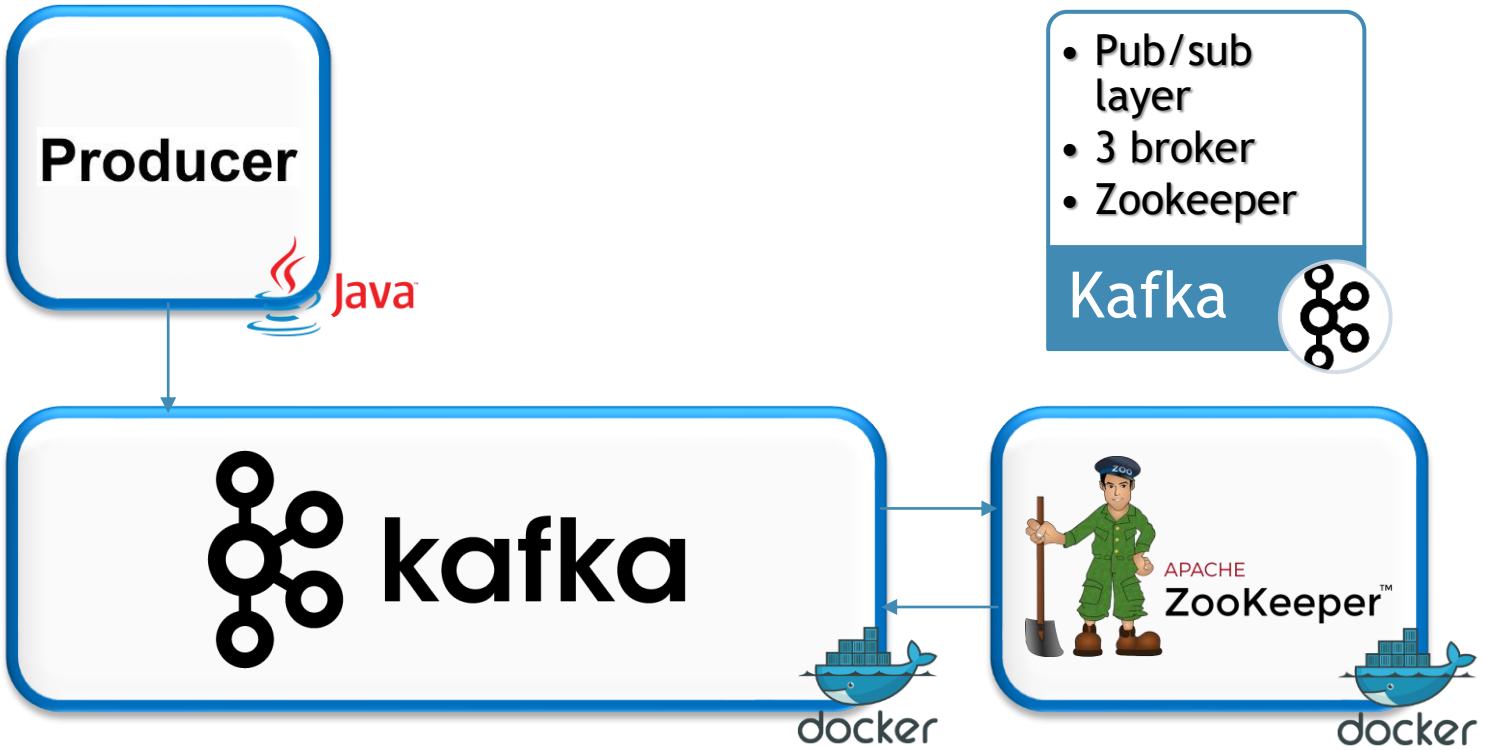
Benchmark

- Latenze
- Throughput

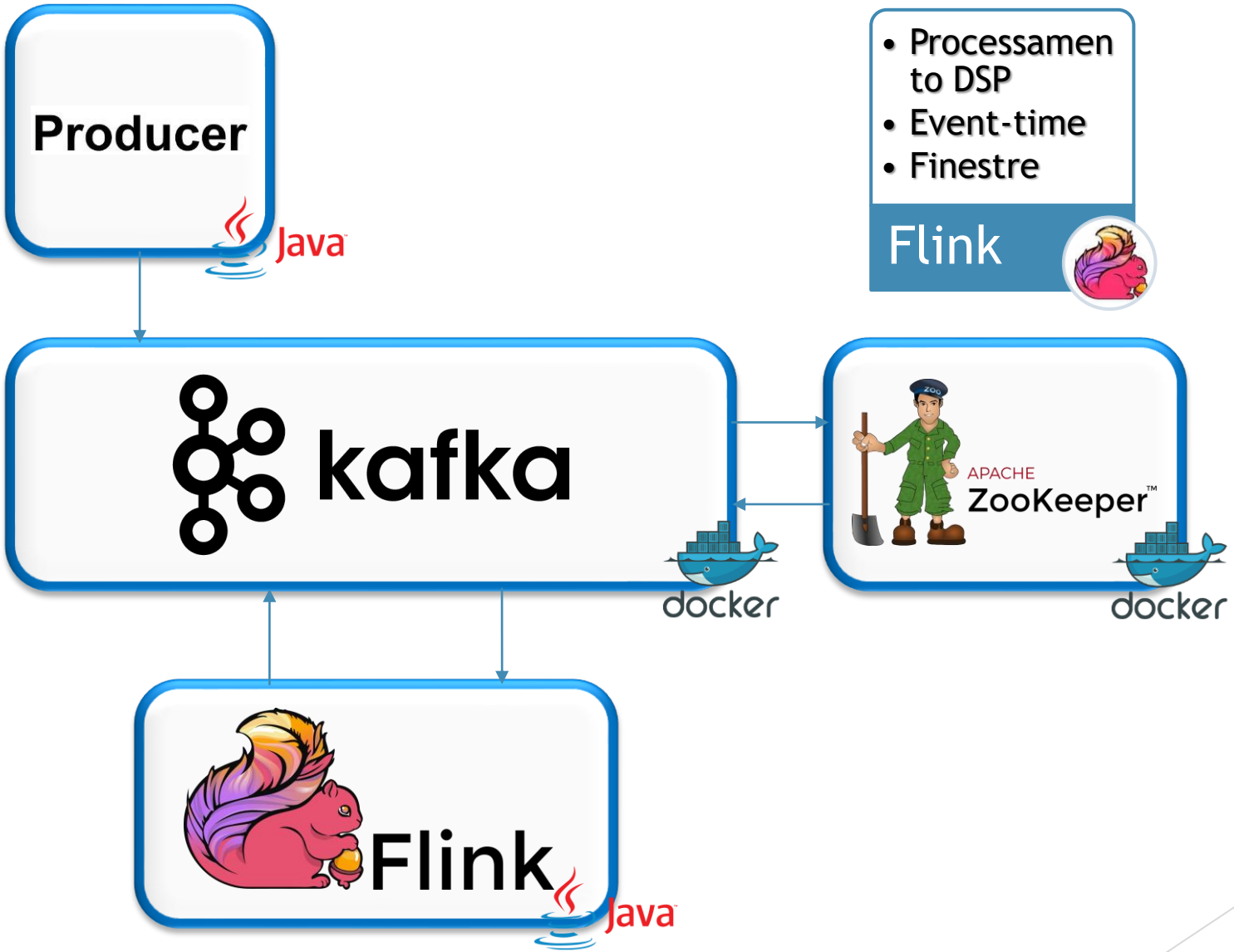
Architettura



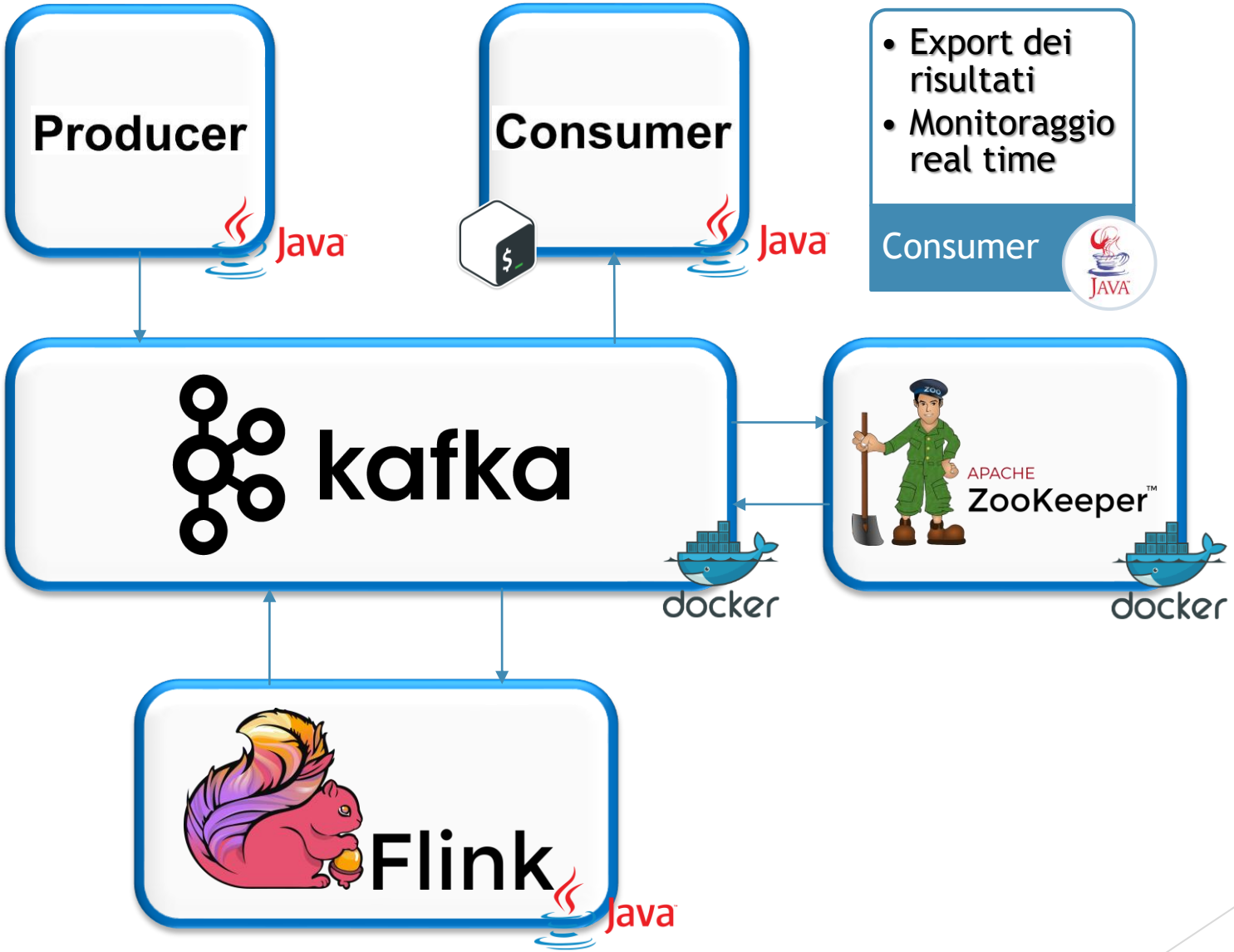
Architettura



Architettura

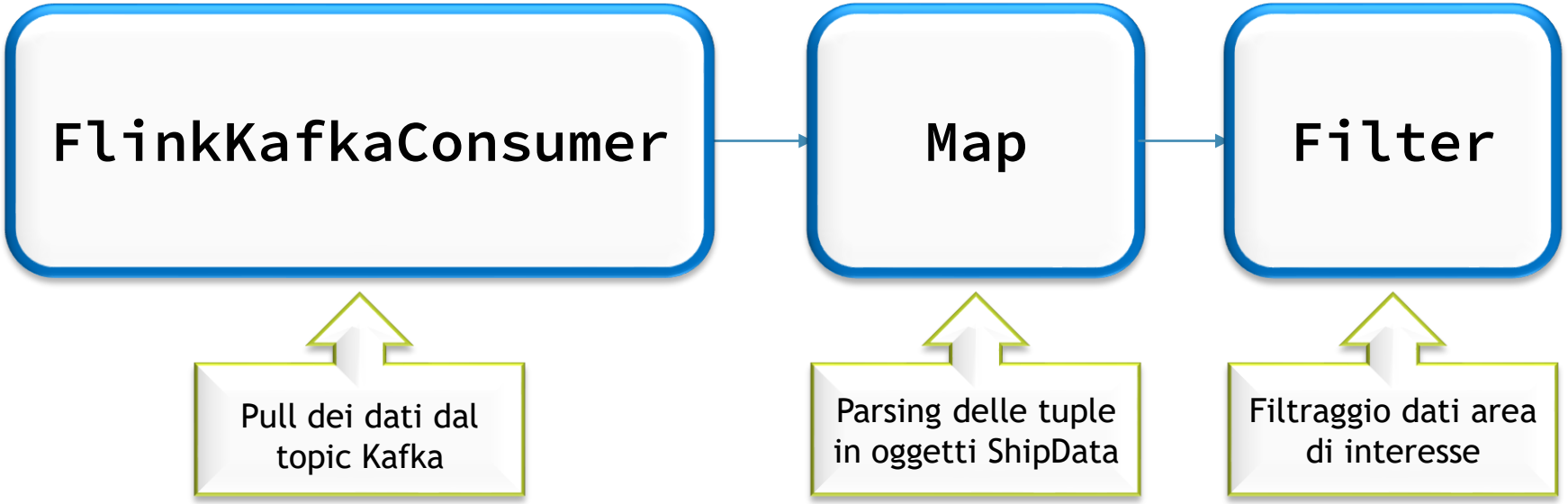


Architettura



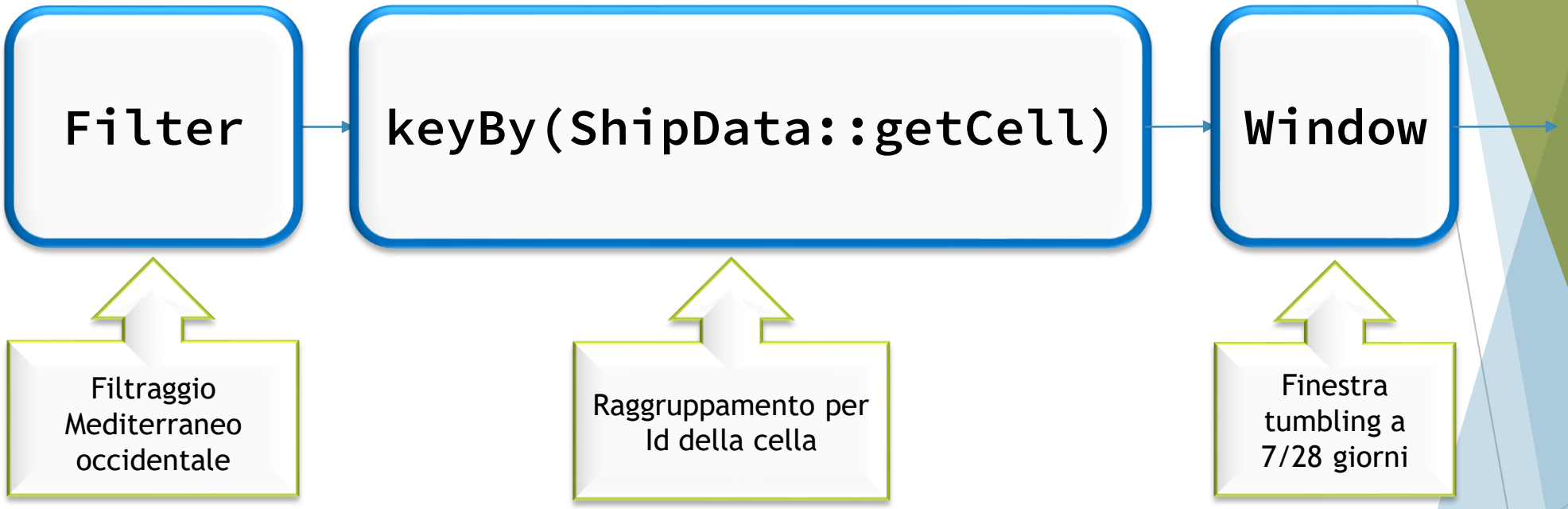
Query

Preprocessing



Query

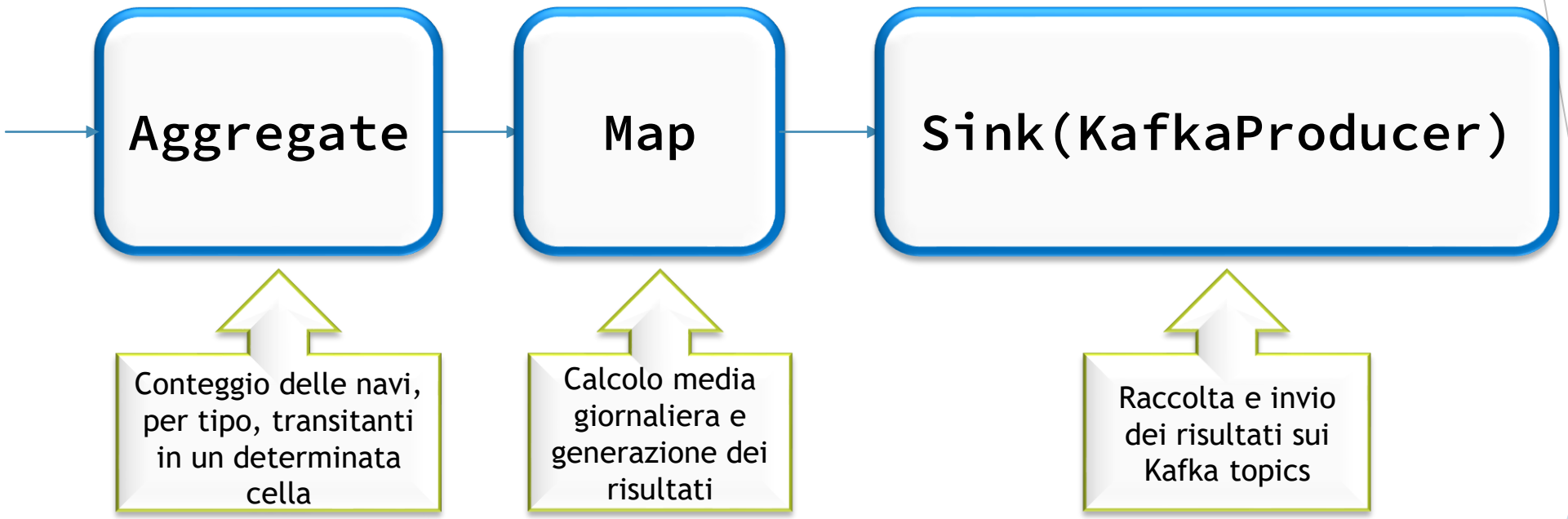
Query 1



Obiettivo: Calcolare per ogni cella del Mar Mediterraneo Occidentale, il numero medio giornaliero di navi militari (SHIPTYPE = 35), navi per trasporto passeggeri (SHIPTYPE = 60-69), navi cargo (SHIPTYPE= 70-79) e others (tutte le navi che non hanno uno SHIPTYPE che rientri nei casi precedenti) negli ultimi 7 giorni (di event time) e 1 mese (di event time).

Query

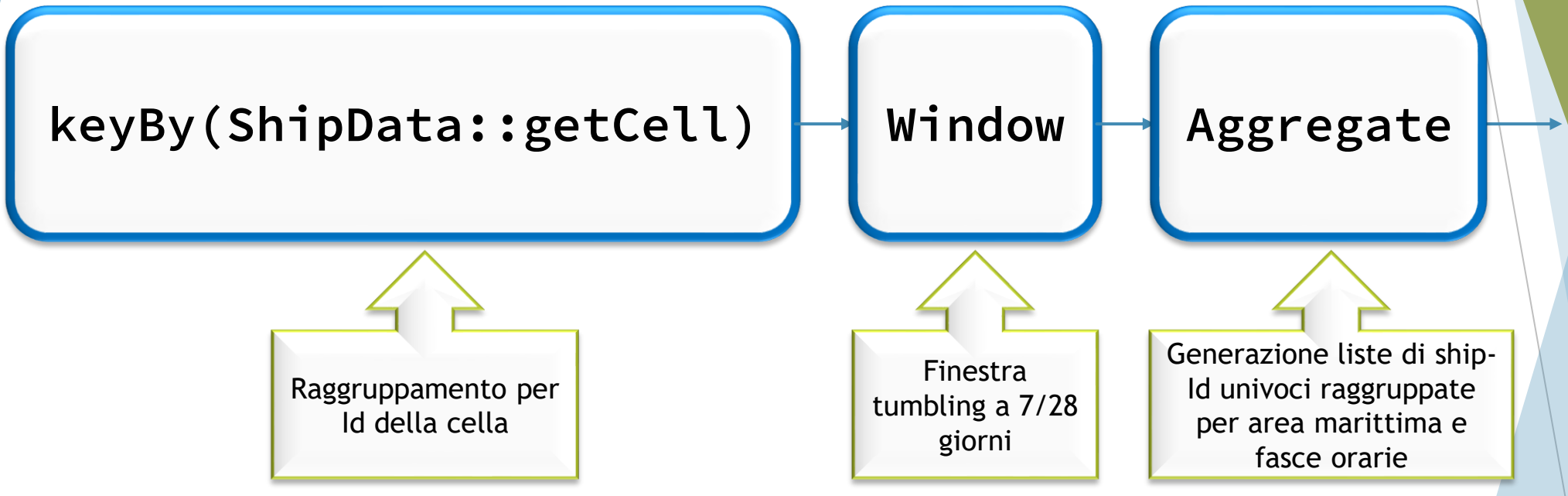
Query 1



Obiettivo: Calcolare per ogni cella del Mar Mediterraneo Occidentale, il numero medio giornaliero di navi militari (SHIPTYPE = 35), navi per trasporto passeggeri (SHIPTYPE = 60-69), navi cargo (SHIPTYPE= 70-79) e others (tutte le navi che non hanno uno SHIPTYPE che rientri nei casi precedenti) negli ultimi 7 giorni (di event time) e 1 mese (di event time).

Query

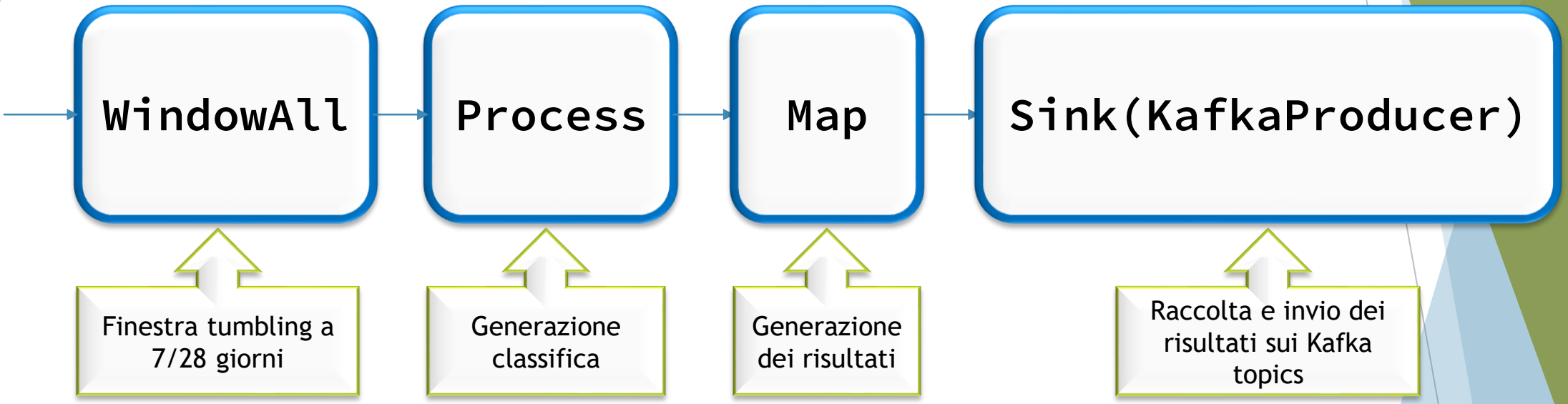
Query 2



Obiettivo: Per il Mar Mediterraneo Occidentale ed Orientale fornire la classifica delle tre celle più frequentate nelle due fasce orarie di servizio 00:00-11:59 e 12:00-23:59, in una settimana e in un mese.

Query

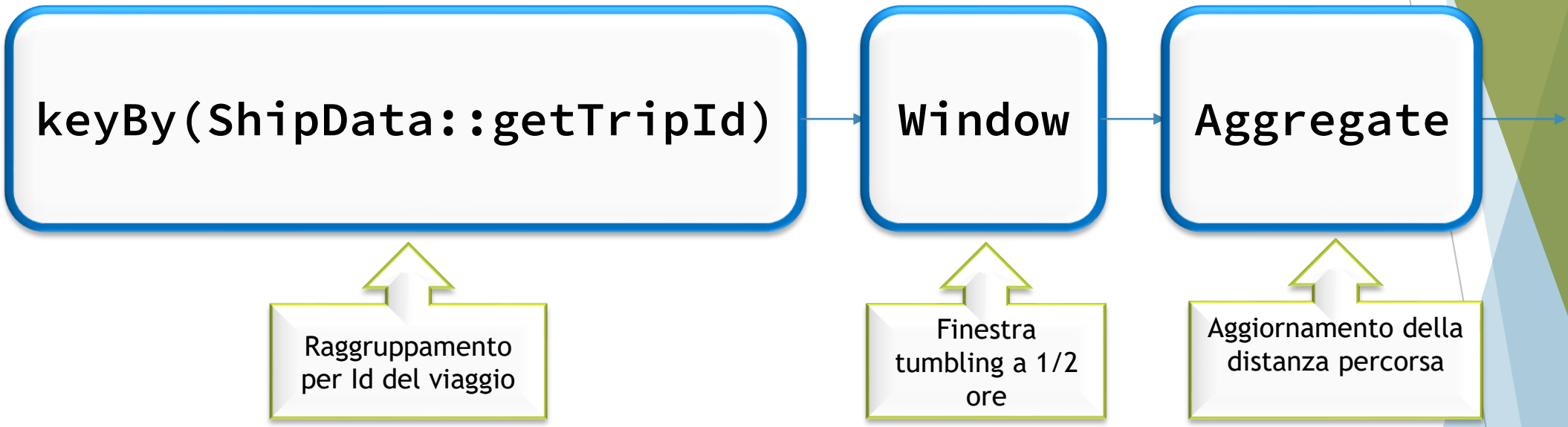
Query 2



Obiettivo: Per il Mar Mediterraneo Occidentale ed Orientale fornire la classifica delle tre celle più frequentate nelle due fasce orarie di servizio 00:00-11:59 e 12:00-23:59, in una settimana e in un mese.

Query

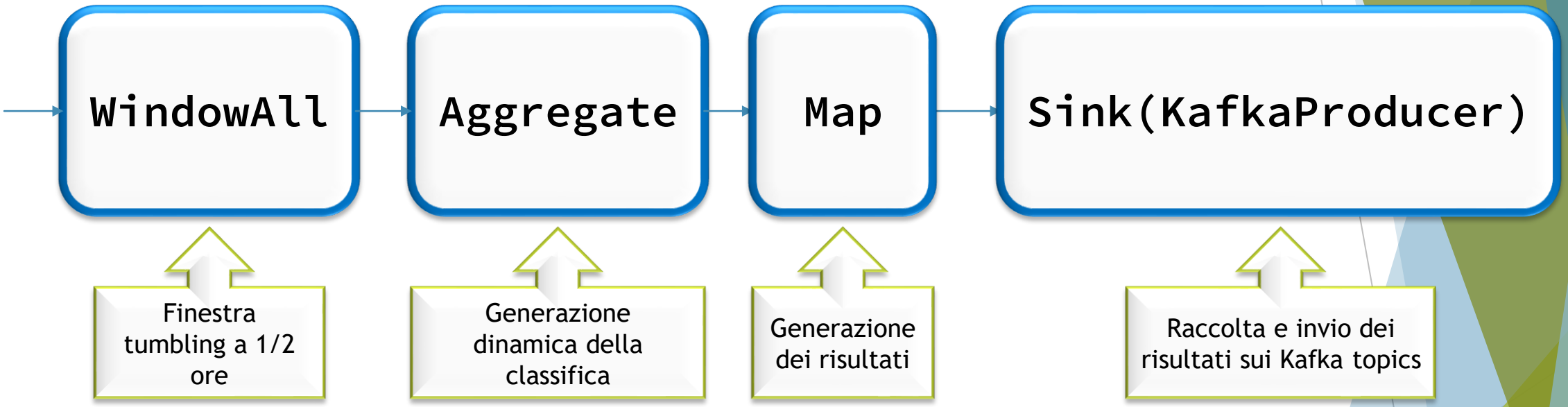
Query 3



Obiettivo: Fornire la classifica in tempo reale dei 5 viaggi che hanno il punteggio di percorrenza più alto, in un'ora e due ore. Il punteggio di percorrenza viene calcolato come pari alla distanza percorsa fino a quel momento del viaggio.

Query

Query 3



Obiettivo: Fornire la classifica in tempo reale dei 5 viaggi che hanno il punteggio di percorrenza più alto, in un'ora e due ore. Il punteggio di percorrenza viene calcolato come pari alla distanza percorsa fino a quel momento del viaggio.

Benchmark

Prestazioni: Throughput e Latenze

Query	Throughput (tuple/sec)	Latenza (sec/tupla)
Query1 Weekly	41.598	0.024
Query1 Monthly	29.987	0.033
Query2 Weekly	2.0	0.501
Query2 Monthly	0.438	2.281
Query3 One Hour	206.567	0.005
Query3 Two Hour	104.713	0.009

* Le metriche si riferiscono alla durata complessiva del replay di 5 secondi

Specifiche HW: L'esecuzione del progetto e la valutazione delle prestazioni sono state eseguite su Linux Mint 20.1 Cinnammon, CPU AMD Ryzen 5 3600, 6 core, 12 thread e 16 GB di RAM, con archiviazione su SSD.

Grazie per l’attenzione