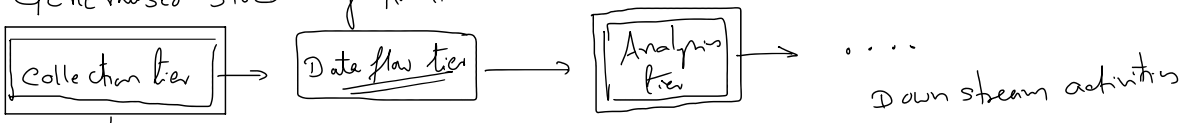
Agenda

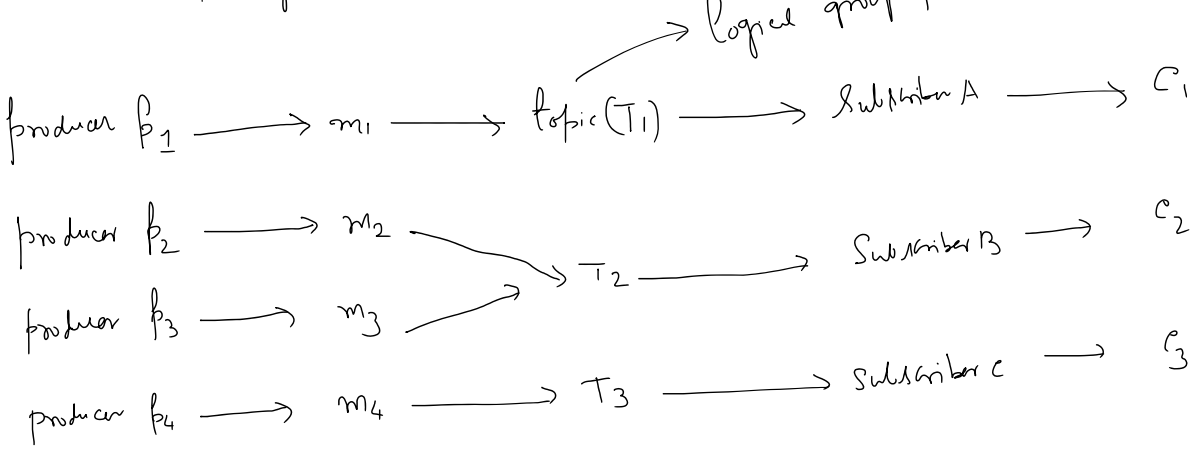
1. Data flow layer

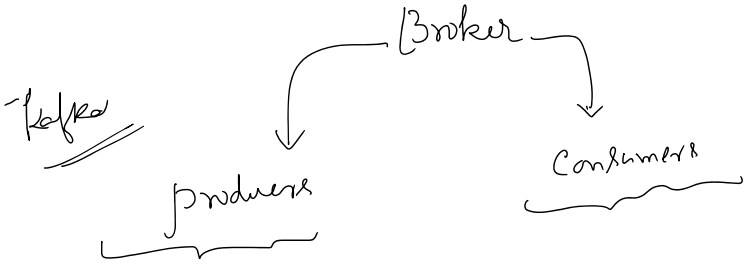
Generalized streaming Architecture



Common-interaction patterns

- client-server (Request-Response pattern) : [when client must have an immediate response
- pub-subscriber (producer-consumer pattern)
  - complexity in terms of variety of Data Sources / formats





Data flow tier

message Delivery Semantics

→ Analytics

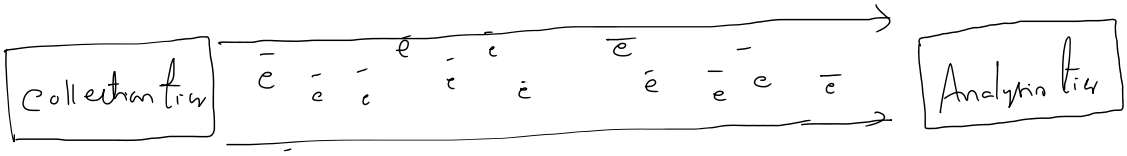
To decouple collection tier and Analysis tier

Considered as message queue

The collection tier produces events at much faster pace compared to the analysis tier

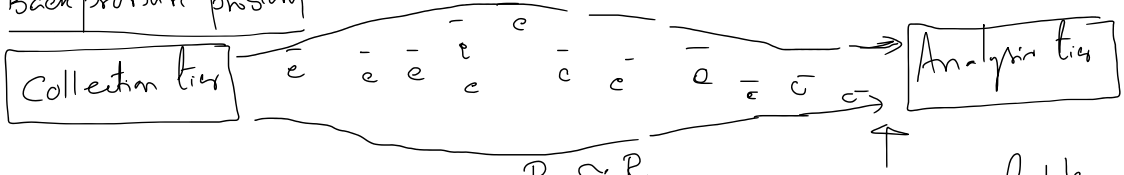
Rate of ingestion (R_i)

Rate of processing (R_p)



Normal Data flow $R_i \leq\simeq R_p$

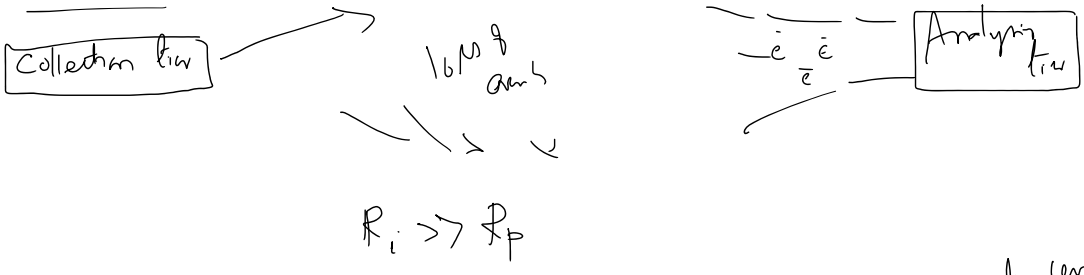
Back pressure problem



$R_i \gtrsim R_p$

Choosing problem

Data loss

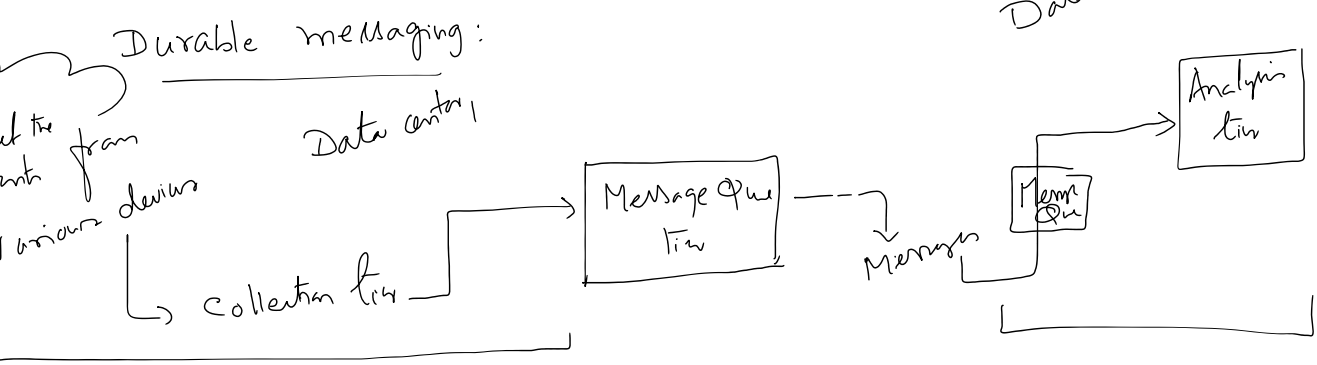


$R_i >> R_p$

Durable messaging:

Data center 1

Collect the events from various devices



Durable messages provide a degree of fault tolerance allow offline consumption of events

Usecase

What if the events are lost?

Architecture Design perspective

(1) Impact on Business if the communication between collection tier and analysis tier is interrupted significantly

(2) How many days (Tolerance factor – message loss) of Data the business can tolerate

(3) Need for storage of historical data

Credit card txns data

Criteria	Inference
(i) Business Impact	
(ii) Tolerance for event loss	
(iii) Need for storing Historical Data	

Semantics :–

Number of partitions Needed = $\max\left\{\frac{T_t}{T_p}, \frac{T_t}{T_c}\right\}$

T_t: Throughput of the system

T_p: Max. throughput of producer writing events onto single partition

T_c: Max Throughput of a Consumer reading events from a single partition

System Throughput = 6GB/min = 6×10^3 MB/min = $6 \times 10^3 / (60\ sec)$
= 100 MB/sec

T_p = 5MB/sec

$\frac{T_t}{T_p} = \frac{100\ MB/sec}{5\ MB/sec} = 20$

Estimated no. of producers

$20 = \max\{20, \frac{T_t}{T_c}\}$ ← Estimated number of consumers

$\frac{T_t}{T_c} \leq 20$

$\frac{T_t}{20} \leq t_c$

$t_c \geq \frac{T_t}{20} \Rightarrow$ 5MB/sec

Ex²

R_i = 24 MB/hr R_p = 2 MB/min

$R_i = 24\ MB / 60\ min$ R_p = 2 MB/min

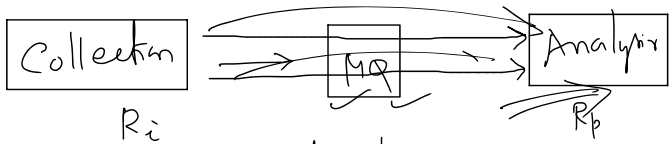
= 0.4 MB/min

$\frac{R_i}{R_p} = \frac{0.4}{2} = 0.8 \Rightarrow \frac{R_i}{R_p} = \frac{8}{10} = \frac{4}{5}$

$$R_i = \frac{4}{5} R_p$$

Avg msg size = 64kB

$R_i \leq R_p$



Exactly once, atleast once using auxiliary message queue

$|R_i - R_p| = \frac{1}{5} \times 2\ MB/min$

1 min 0.4 MB

$\frac{400}{25}$ kB ≃ 6kB

1 min $\quad \frac{0.4\ MB}{64\ kB} = \frac{400\ kB}{64} \simeq 6\ kB$

1 min $\longrightarrow$ 6 kB data

1 hr : 360 kB

(ii)

$R_i = 120\ MB/min$

$R_p = 24\ MB/hr$

$24/60\ MB/min$

$= \frac{4}{10} = 0.4\ MB/min$

$\frac{R_i}{R_p} = \frac{120\ MB/min}{0.4\ MB/min} = \frac{1200}{4} = 300$

$R_i = 300\ R_p$

$R_i >> R_p$

$n_i = \frac{120}{64\ kB}/min = \frac{30}{16}(10^3) \simeq 2000$ ✓

$n_p = \frac{24\ MB}{64\ kB}/60\ Min = \frac{4 \times 10^3}{64 \times 10} = \frac{400}{64} = 6.25$

R_p 7 mg/min $\Rightarrow$ $7 \times 64\ kB/min$

$448\ kB/min$

$R_i = 2000$

$2000 - 7 = (1093) \times 64\ kB = 1.1 \times 64\ MB/min$

Auxiliary storage $\simeq 70\ MB/min$

Refresh hrly: $70 \times 60 = 4200\ MB$