Stream processing and analytics

Quiz 1
- ~ 6th Contact Session (7th Contact Session)
- ~ 30min { No. of questions = 10 ; each 0.5 mark }
- Window { Couple of days including weekend }

Quiz 2
- ~ 13th Contact Session
- 30min { No. of questions = 20 ; each $\frac{1}{4} = 0.25$ marks }
- Window (3 days { Fri, Mon })

Content: Apache Kafka, Apache Spark open access documentation
⊕
Slides

Assignment 1 (Experiential learning)
↳ Explore Streaming platforms
↳ mapped to individual groups

Gr1 → Samza
Gr1 → flink
Gr2 → Kevinio
ksqlDB
flume

Assignment II
↳ Databricks Spark
↳ ML module → Spark MLlib
Python

10 M
~~20 M~~

Quiz I	Quiz II	A_1	A_2	Total
5	5	15	10	30

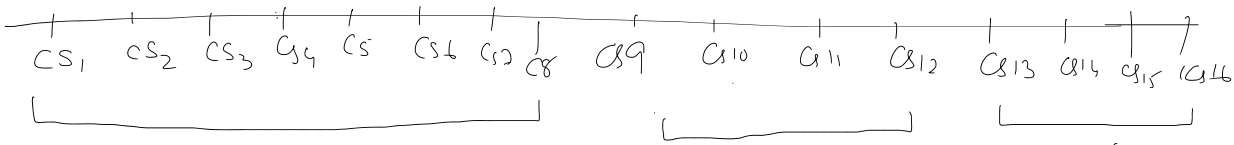
⇒ group allocation (as per your choice) 2-3 weeks time : groups finalization ~ 5th Contact Session

30 min/day {Excluding assignment queries}

↳ BF / Coffee time

Data Engineering / Data Architect

CS: 16



- Understanding reqs
- Proposing Data Arch.
- Evaluation of Data Arch.
- Aspects — Non-functional Requirements
- Kafka Arch.

Kafka graph aspects ↳ streaming algorithms and time/space complexity issues

PySpark

~~DStream~~ obsolete deprecated

↳ Spark Structured Stream

↳ Architecture and execution of jobs in Spark

Stream processing:

↳ Computations on potentially endless and constantly evolving source of data

Tried

Computation model

Create

Retrieve

Functions

Data representation → Data types

↳ String

↳ float

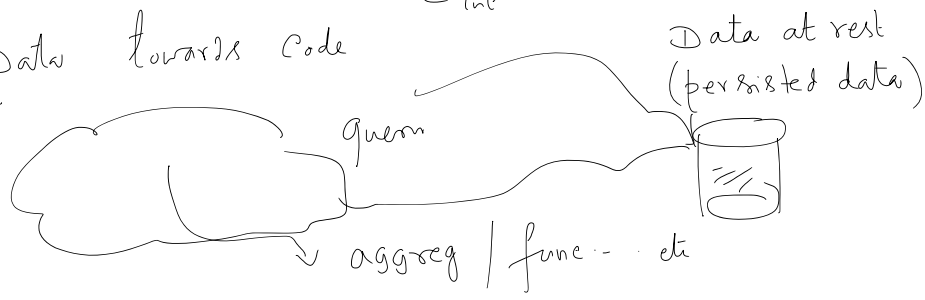
↳ int

Create
Retrieve
Update
Delete

changing original contents/Data

Data types
- String
- float
- int

Moving Data towards Code



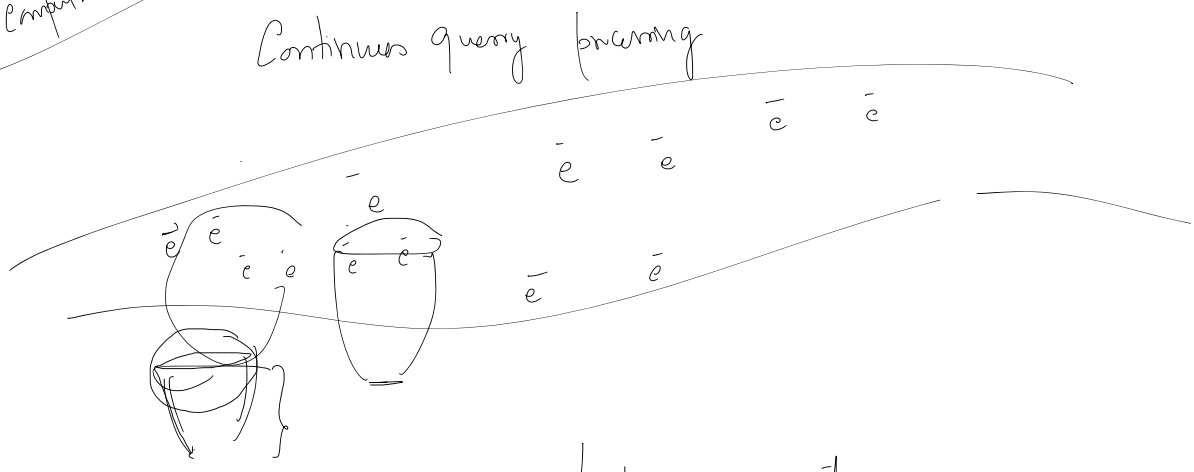

```
def computeSum(list):
    sum = 0
    for element in list:
        sum += element
```

Data fetched either from file/Data store

relies on sum

streaming as computational model

Continuous query processing



Code is executed on fast moving Data

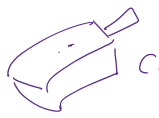
Data stream:

It comprises of endless stream of immutable data (events)

Size of event ~ kB (< 1MB)

Event is an immutable fact regarding something that occurred within the System Under consideration

Data records in the context of data streaming are called events



card swipe at pos

Card ending in xx 4593
Spent 20,000 at Starbucks
time stamp

⇒ Un-mutable { Cannot modify content of the event }

- Transformations
- Aggregations
- filtering

} Operations on events

Examples of Stream processing

① Appln logs : timestamp : <context> | status

{ 2025:07:27 :18:00:05:10 : exception | failed opening | warning | Long transaction | error

② Web Analytics :

③ Real time pricing of a stock

Data Types

- Structured → Representation of data has both Schema ; type information of each col. (Sql DB)
- Semi structured → Some level of Schema structure but Type information missing ; → XML, JSON ; HTML
- Un structured Data
  - hierarchical in nature
  - neither Schema compliance Nor Type semantics
  - text, video, audio → facebook post

Size of the data

Size of the data

1 Byte = 8 bits

kB $\simeq$ 1024 Bytes $\simeq 10^3$ Bytes

kB MB GB TB -

1024 kB 1024 MB 1024 GB -

$\simeq 10^3$ kB 10^3 MB 10^3 GB

1 GB $\simeq 10^3$ kB

1 TB $\simeq 10^3$ GB $= 10^3 (10^3$ MB$) = \underline{10^6}$ MB