## Lecture 1

- If single computer is enough dont use DS.
- Motives for DS:

① Parallelism

② Fault tolerance : If one fails, another is backup

③ Physical reasons : Systems inherently are physically distributed

④ Security : require isolation

- Unexpected failure patterns
- Partial failures
- Performance

} Basic challenges

- Infrastructure

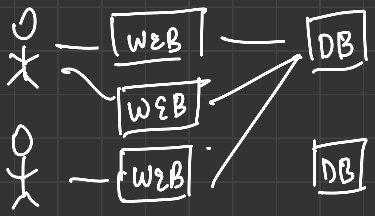
- Storage
- Communication
- Computation

} Abstractions → Implementations:

- RPC, threads - way of structuring concurr. ops
- concurrency control

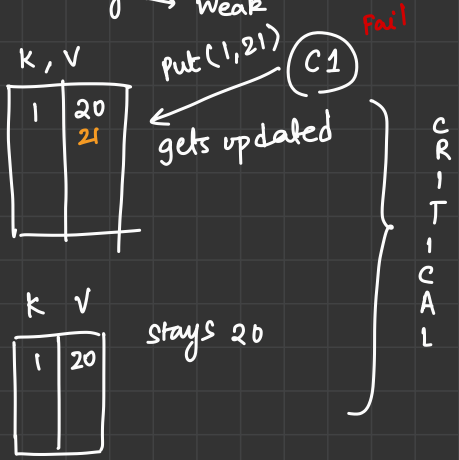
- Performance

- High level goal : scalable speed up  2x machines = 2x throughput



- Fault Tolerance ] Big tools
  - Availability | clever ways of avoiding to write non-volatile storage
  - Recoverability | management of replication

- Consistency → Strong : guarantee of most recent write is expensive
  → Weak



- Replication for fault tolerance should have independent failure probabilities.

↓

might make communiation expensive

map (k,v)

split v into words
for each word w
emit (w, 1)

reduce (k,v)

emit ( len(v))
↑
array

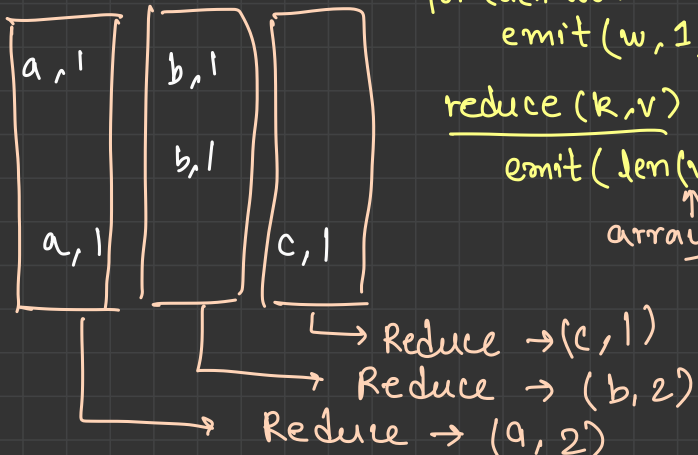
Jan 22, 2023

→ map Reduce

Input File 1 → map

Input File 2 → map

Input File 3 → map


a , 1
b , 1
b , 1
a , 1
c , 1
Reduce → (c, 1)
Reduce → (b, 2)
Reduce → (a, 2)