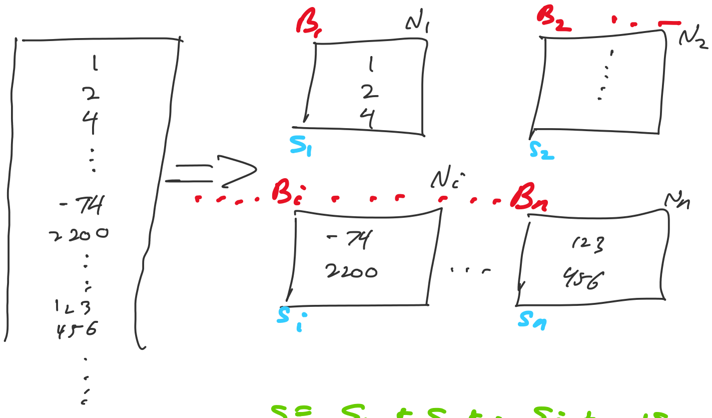
Hadoop Ecosystem

Thursday, January 26, 2023 3:22 PM

- Data Processing
  - Data Parallelism
    - Multiple Nodes / Machines (Cluster): Distributed Storage
Distributed Compute

* Storage
- File : Sequence of Bytes (Text) Code & Data
  Typical Filesystem ⇒ Inodes

  DFS ⇒ Blocks



$S = S_1 + S_2 + \dots S_i + \dots + S_n$ Linear

* Fault-Tolerance !
↳ K-safety

Safety

* Performance
↳ Storage : Redundancy / IOPS
Compute : Locality

Efficiency

* Compute
- Parallelism & Concurrency (HPC ⇒ Schedulers / Task Management)
- Task / Job : Process ⇒ scheduling
  - Stream of Bytes : Code & Data (Binary)
- Psuedo Code → Adding a list of numbers

```
m() {                         ← input blocks
    partial_total = 0
    for (each block) {
        read_block()
        partial_total += record
    }
    return partial_total
}
```

```
r() {                         → output blocks
    final_total
    for (each block) {
        read_block()
        final_total += record
    }
    return final_total
```

intermediary
blocks

↑
call in parallel
across each node N_i !

parallel

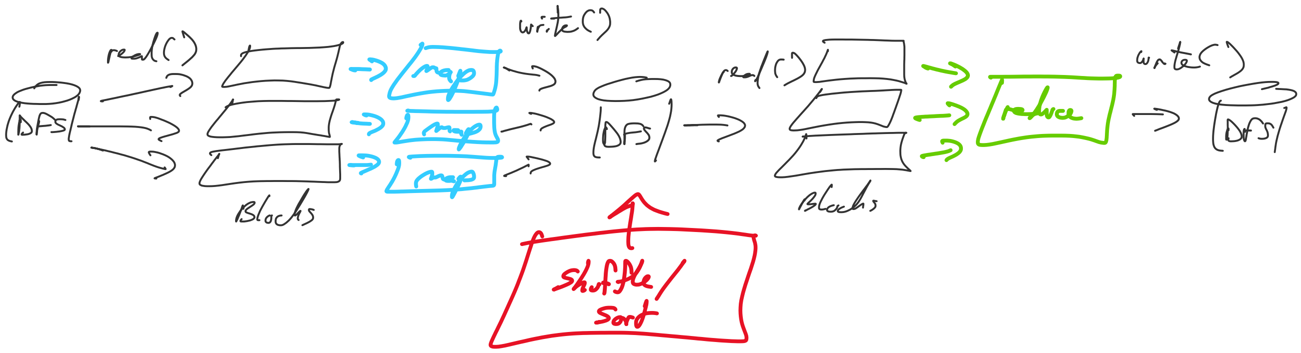
map

↑
single call for
across all partitioned blocks!

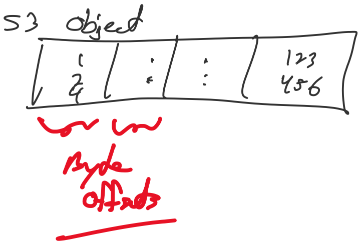
serial

reduce

* Hadoop cluster



* cloud: AWS
- Hadoop (EMR): HDFS (dynamic) / YARN (single tenant)
- S3: Object Storage


S3 object
1
2
3
123
456
Byte
offsets