Introduction to Bigdata

Distributed Systems

Hadoop and its role in Bigdata

Spark Introduction.

Benefits of using PySpark

What is Bigdata ? ⟶ data with huge size.

201201hourly.txt ⟶ 507 MB

Notepad ⟶ Failed ⟶ Bigdata

Wordpad ⟶ Had some latency but ⟶ NORMAL DATA
loading was successful

Distributed Architectures.

Hadoop (Its a distributed processing engine with a distributed storage facility)

- Storage (HDFS)
- Processing MR
- Resource management YARN.

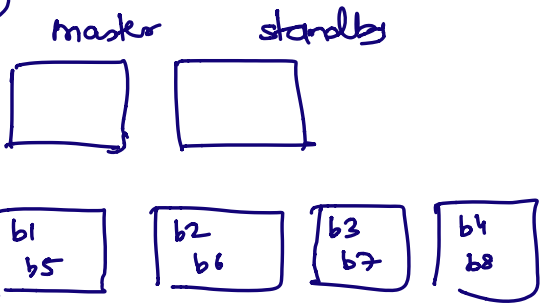
- Benefit. (Parallel Processing)

file1.txt → 512 MB (8 blocks)

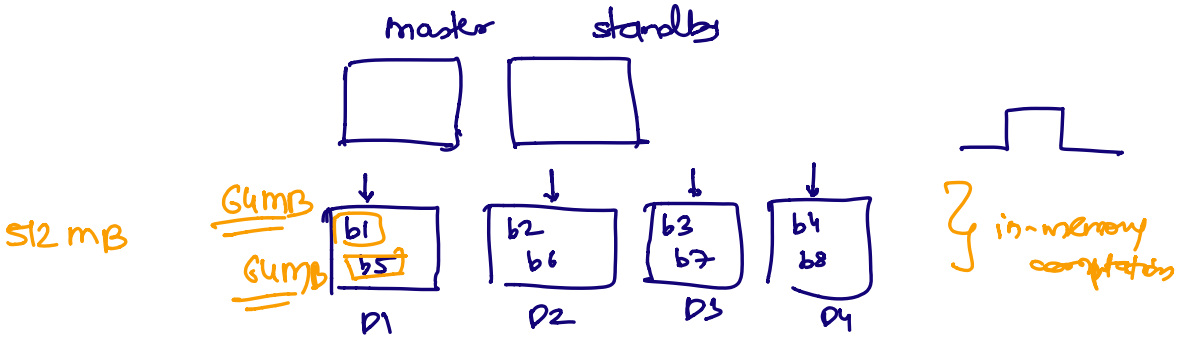
Hadoop splits your data in the form of blocks

B.S = 64 MB

maintain data and perform processing



When it comes to Processing, all systems will participate



This I advise parallel processing

Apache Spark. (distributed Processing Engine)