

# BASIC CONCEPTS

prepare to use tools

## DISTRIBUTED FILE SYSTEM

file system: how OS manages files

store info long-term (in hard disks)

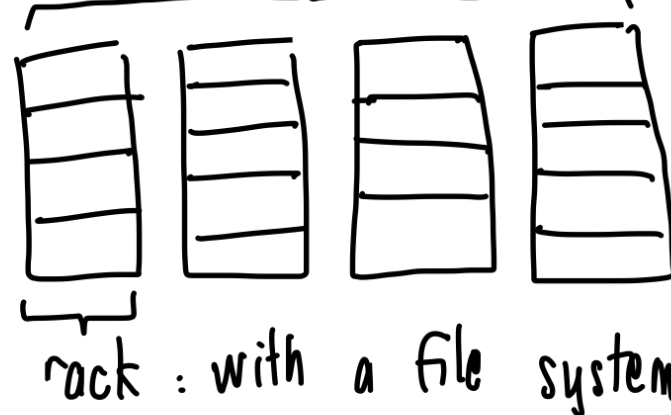
device capacity

with more data? big volumes variety

replicate data among racks

DISTRIBUTED FILE SYSTEM: many storage computers are connected through network

**fault tolerance:** data can be found somewhere else in case of error multiple users/readers + increase system performance x (hard to maintain changes)



rack: with a file system

provide high concurrency data scalability

data partitioning through data replication

- contents:
- numeric
  - alphabetic
  - alphanumeric
  - binary
- name extension

how? efficiency impact flat database

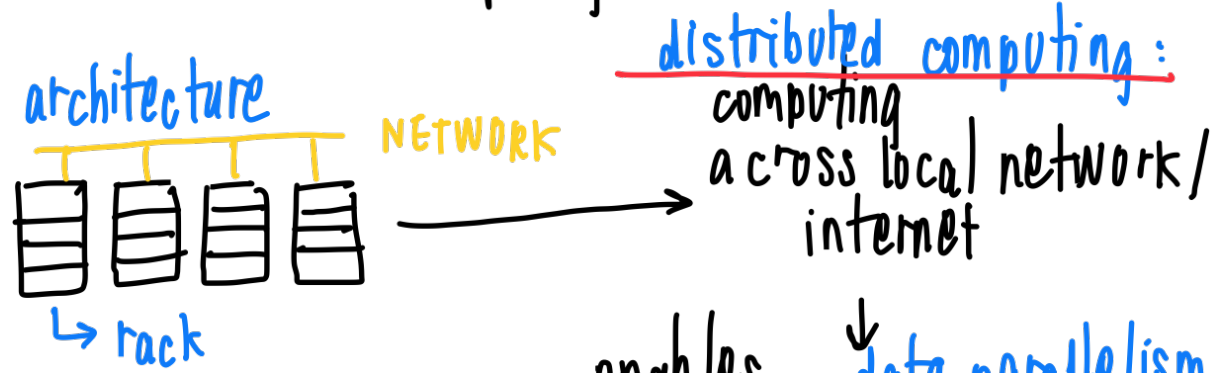
## SCALABLE COMPUTING OVER THE INTERNET

computing nodes one (1): most multiple (1+): parallel computer

large number of single computers with specialized capabilities connected via the network

cousin: commodity cluster: affordable parallel computers less specialized (average number of nodes) pushed for this reduce cost of computing more generic

common failures: node or rack fails connectivity of rack to network stops connection between individual nodes breaks X restart: ability to recover from failures solutions: redundant data storage data-parallel job restart



distributed computing: computing across local network/internet

enables data parallelism: jobs that share nothing can work on different data sets parts of a data set

## big data analysis

large volumes of data can be analyzed using parallelism achieves scalability performance cost reduction