BASIC CONCEPTS → prepare to use tools

DISTRIBUTED FILE SYSTEM → device capacity → with more data? → big volumes / variety

file system: how O's manages files → how? → name. extension → contents: - numeric - alphabetic - alphanumeric - binary

efficiency → flat / impact → database

store info long-term → (in hard disks)

DISTRIBUTED FILE SYSTEM: many storage computers are connected throug network

rack: with a file system

replicate data among racks

fault tolerance: data can be found somewhere else in case of error; multiple users/ readers

+ increase system performance

x(hard to mantain changes)

provide → high concurrency

provide → data scalability

data partitioning

through

data replication

SCALABLE COMPUTING OVER THE INTERNET

mputing nodes → one (1): most

mputing nodes → multiple (1+): parallel computer → large number of
- single computers with
- specialized capabilities
- connected via the network

cousin: commodity cluster: parallel computers
- affordable
- less specialized
- (average number of nodes)

commodity cluster → more generic

pushed for this → reduce cost of computing

common failures ← node or rack fails; connectivity of rack to network stops; connection between individual nodes breaks

common failures → fault tolerance: ability to recover from failures → x restart

solutions → redundant data storage; data-parallel job restart

architecture: NETWORK → rack

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