Definition and Characteristics of Big Data

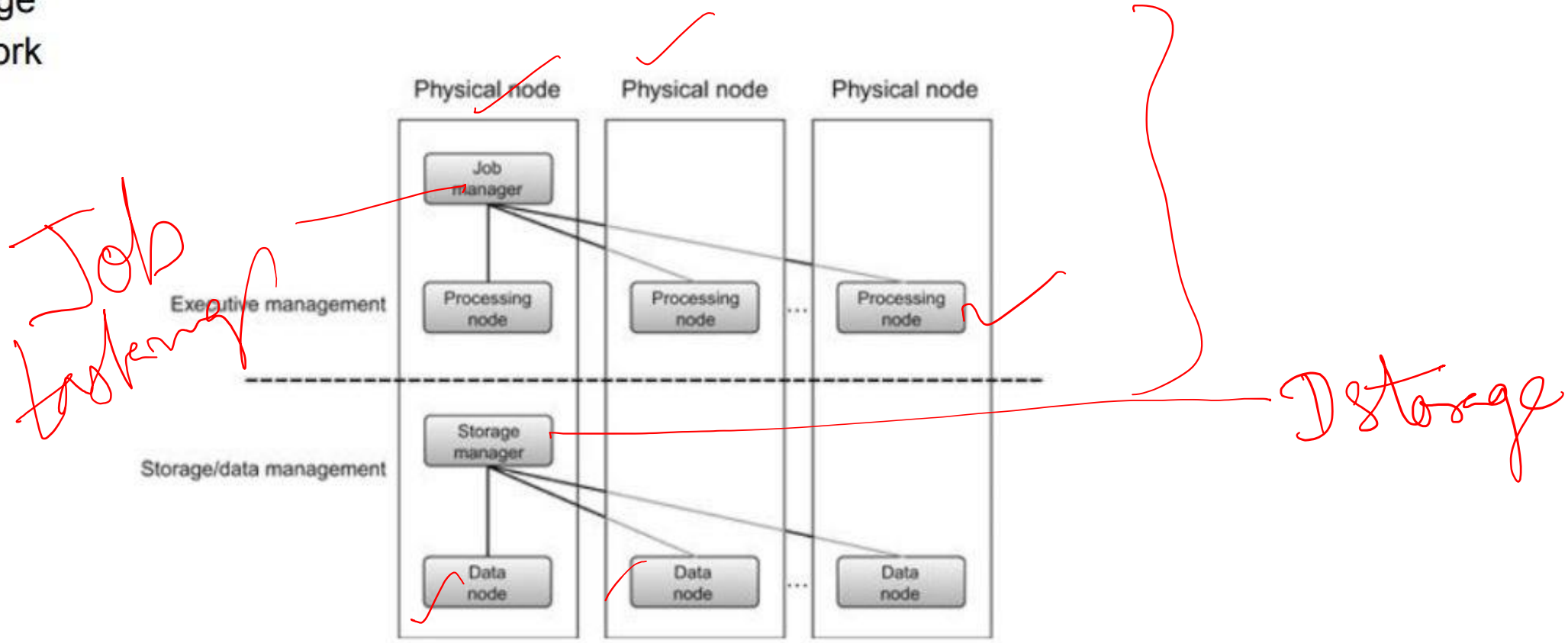
"*Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**.*" -- Gartner

which was derived from:

"*While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: **volumes, velocity and variety**. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each.*" – Doug Laney

Key Computing Resources for Big Data

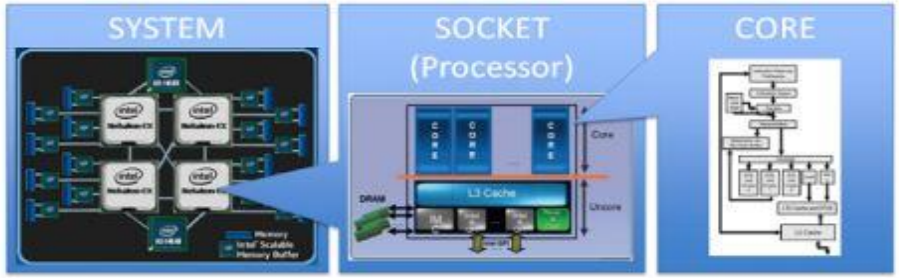
- Processing capability: CPU, processor, or node.
- Memory
- Storage
- Network

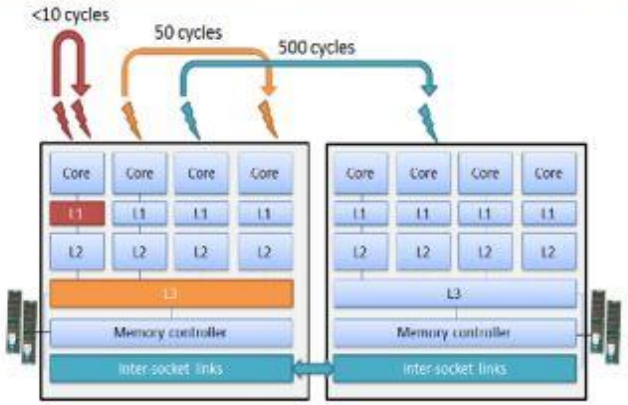


"Big Data Analytics", David Loshin

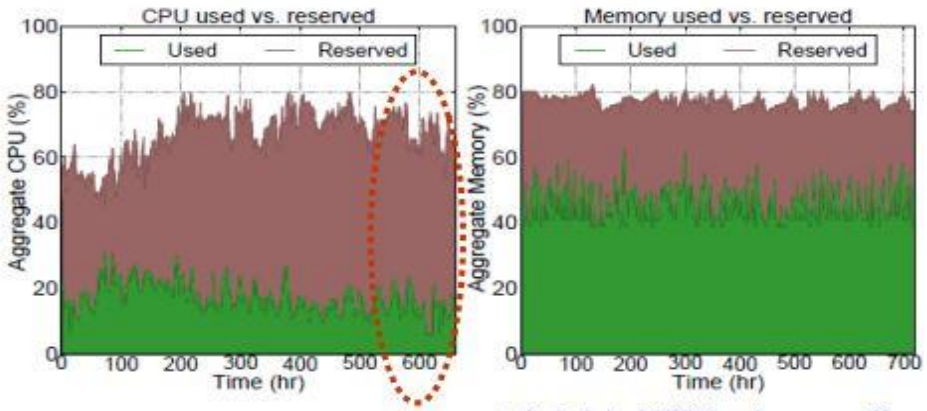
Scalability — Scale Up & Scale Out

- Scale out
  - Use more resources to distribute workload in parallel
  - Higher data access latency is typically incurred
- Scale up
  - Efficiently use the resources
  - Architecture-aware algorithm design





Example: Resource utilization for a large production cluster at Twitter data center



www.stanford.edu/~cdel/2014.asplos.quasar.pdf

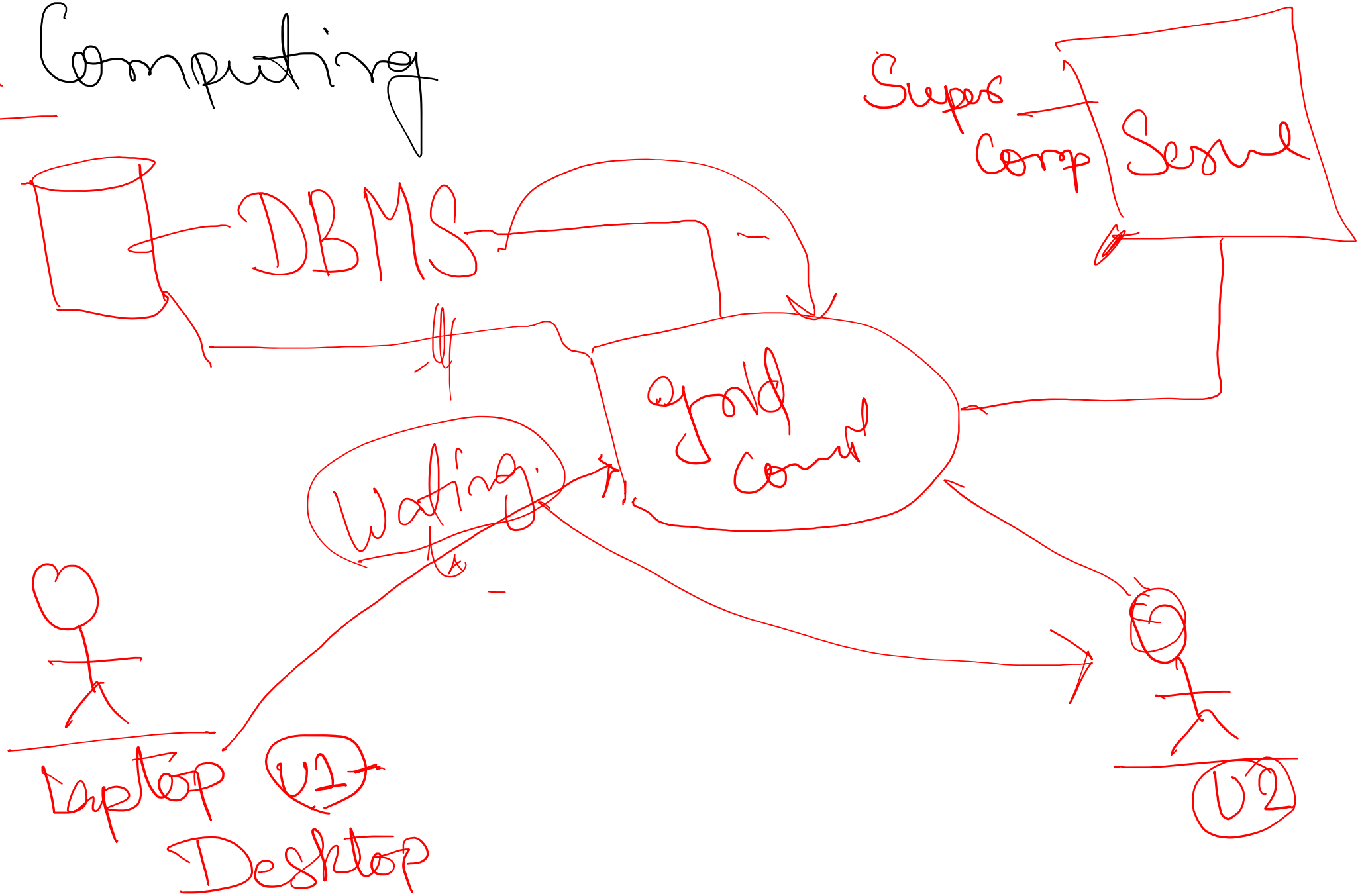
- For independent data ==> scale up may not have obvious advantage than scale out
- For linked data ==> utilizing scale up as much as possible before scale out

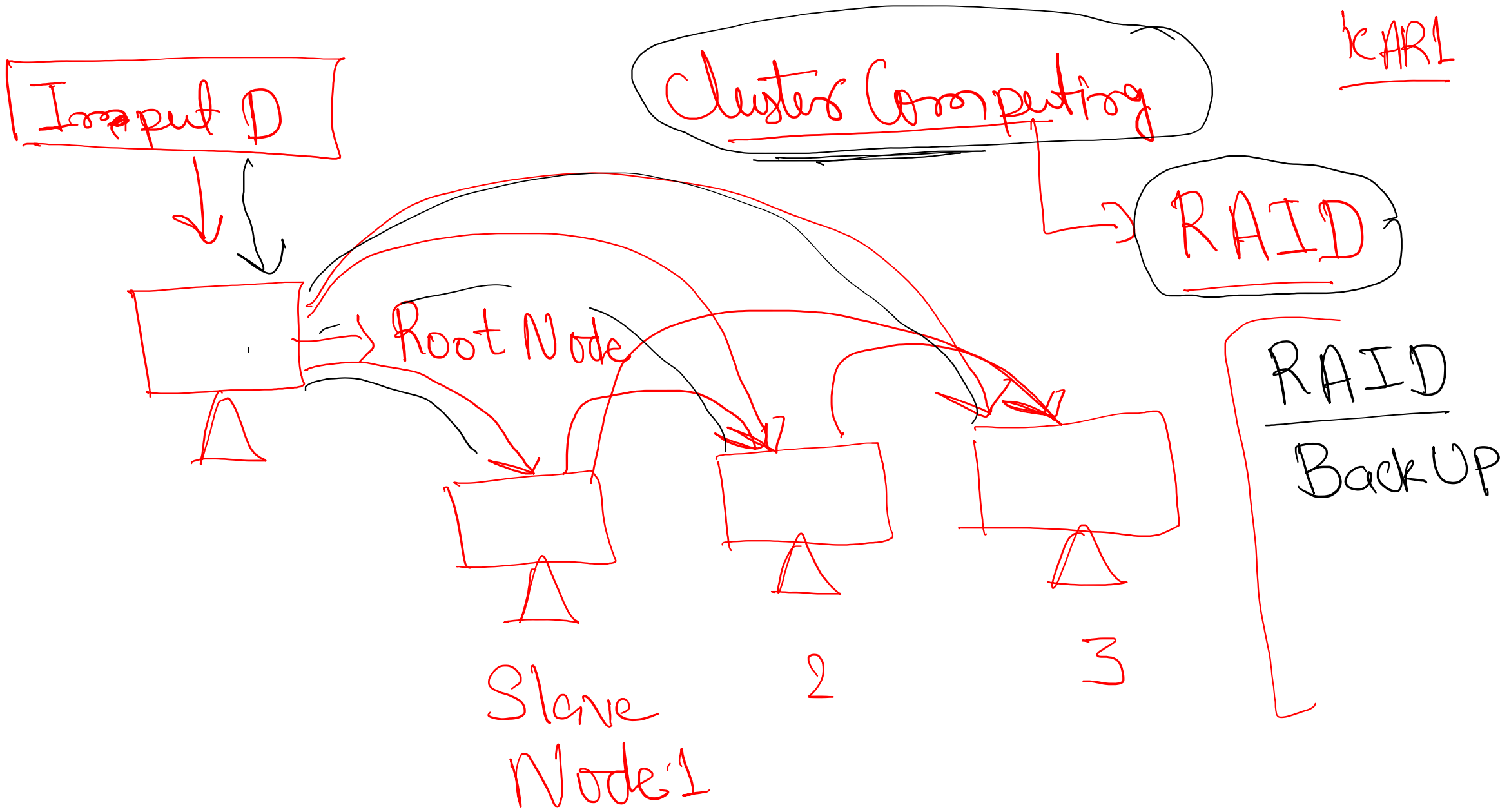
Techniques towards Big Data

- Massive Parallelism (MPP)
- Huge Data Volumes Storage
- Data Distribution
- High-Speed Networks
- High-Performance Computing
- Task and Thread Management
- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

→ Techniques exist for years to decades. Why is Big Data hot now?

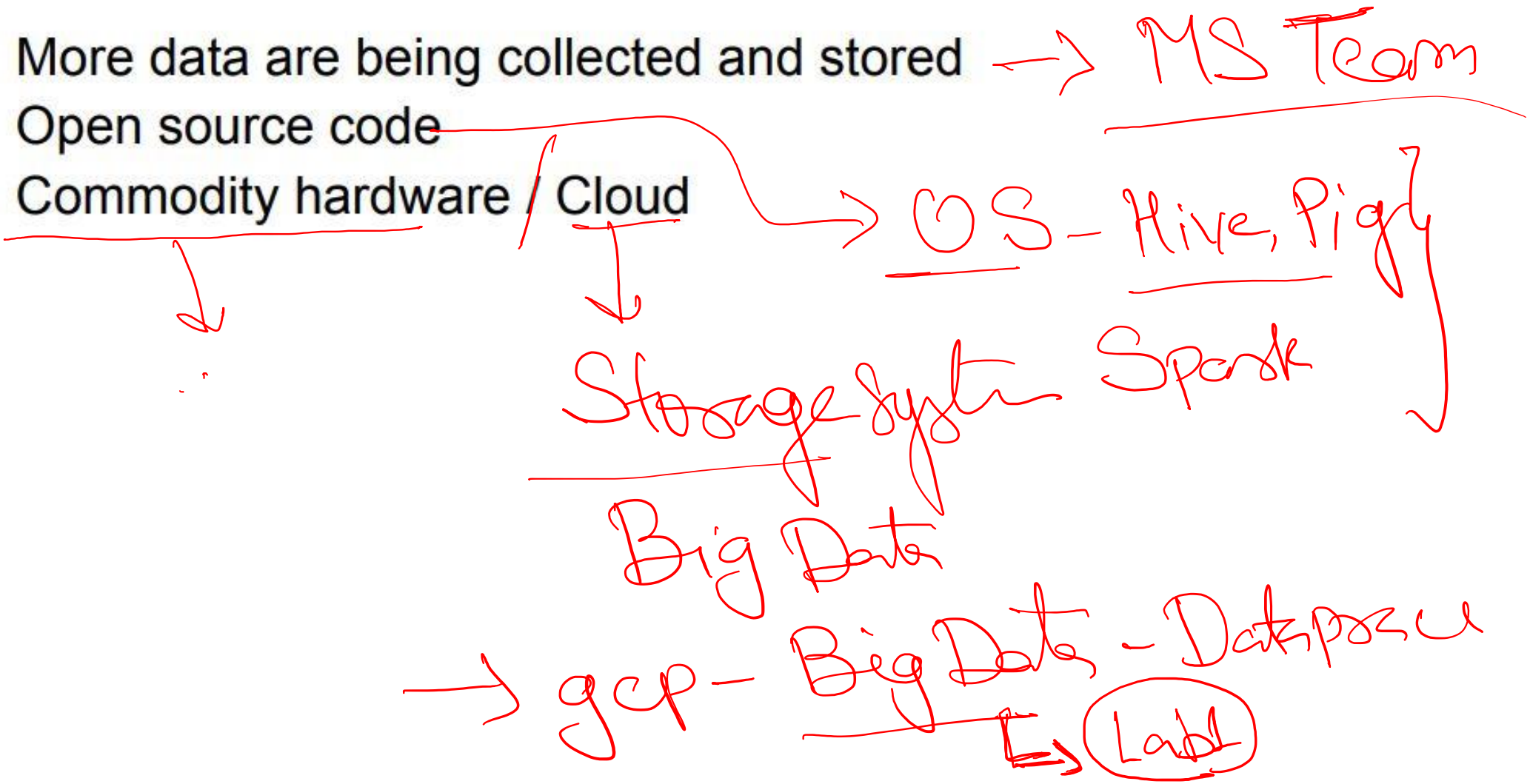
Grid Computing




Input D
Cluster Computing
KARL
RAID
Root Node
RAID
BackUP
Slave
Node:1
2
3

Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

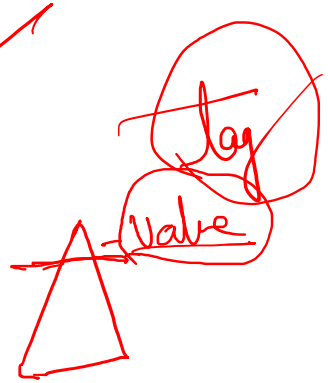


Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

→
- High-Volume
- High-Velocity
- High-Variety

5 Veracity Value



→ Artificial Intelligence