Big Data Analytics
8CS421

GitHub : https://github.com/technoindianjr/Big-Data-Analytics-Lab_8cs4-21

With the advance of IT storage, processing, computation, and sensing technologies, Big Data has become a novel norm of life. Only until recently, computers are able to capture and analysis all sorts of large-scale data from all kinds of fields -- people, behaviour, information, devices, sensors, biological signals, finance, vehicles, astrology, neurology, etc. Almost all industries are bracing into the challenge of Big Data and want to dig out valuable information to get insight to solve their challenges.

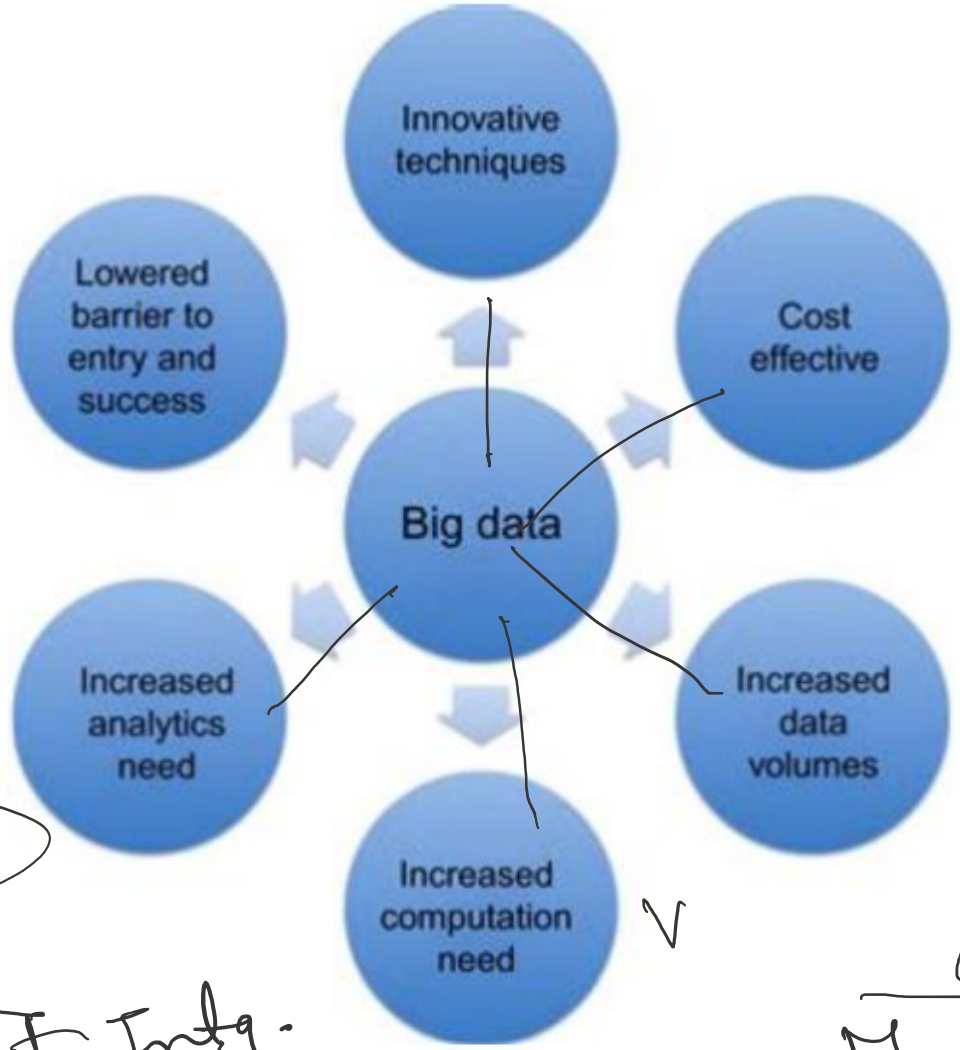
Definition and Characteristics of Big Data

"*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

What made Big Data needed?



SPARK

AWS EMR

MongoDB

Hadoop

V3

PA

SA

DA

NoSQL

V

V

Data Intg. Software

Data Visual.

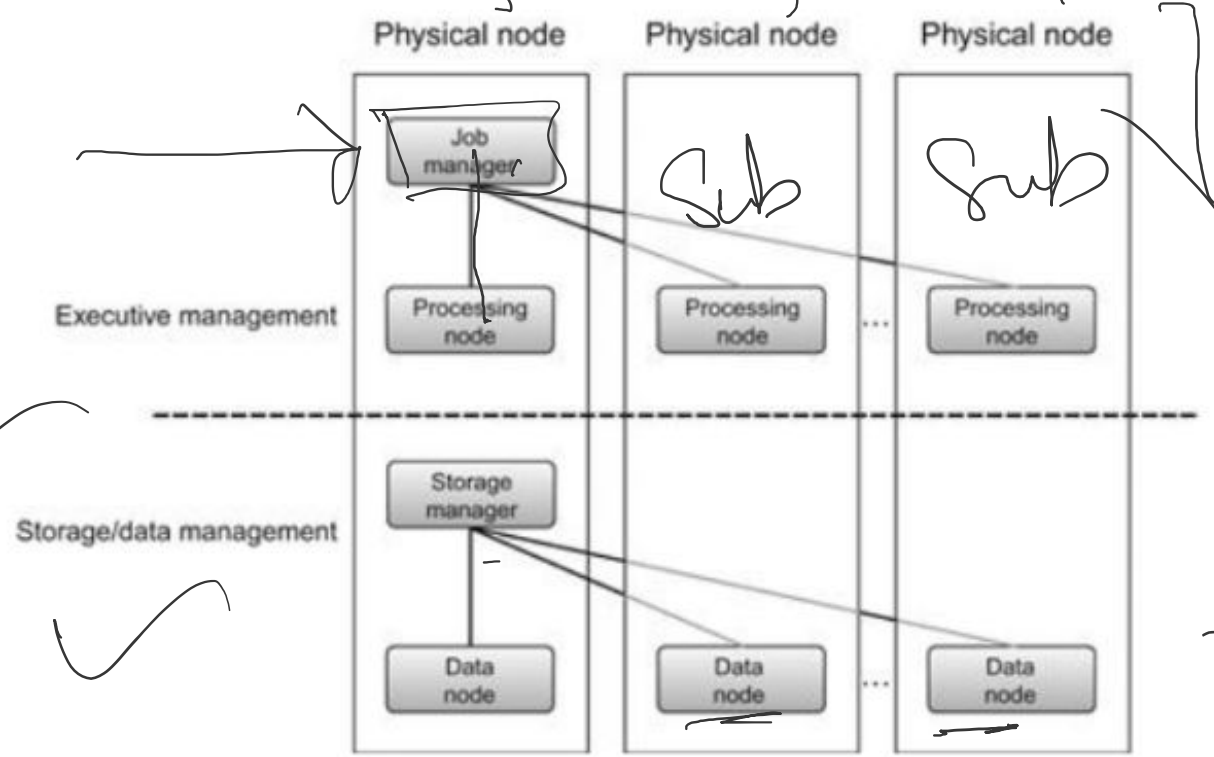
BI

Mining tools

Data Lake

Key Computing Resources for Big Data

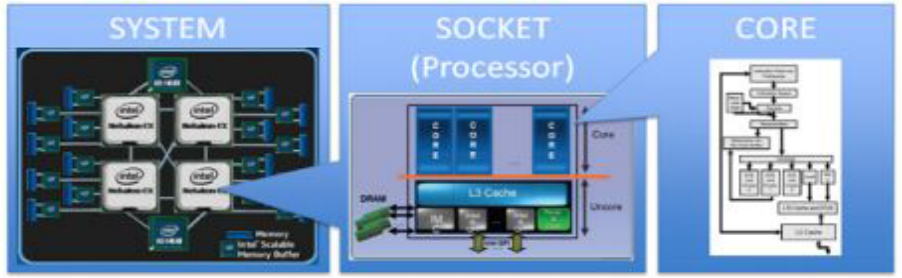
- Processing capability: CPU, processor, or node.
- Memory
- Storage
- Network

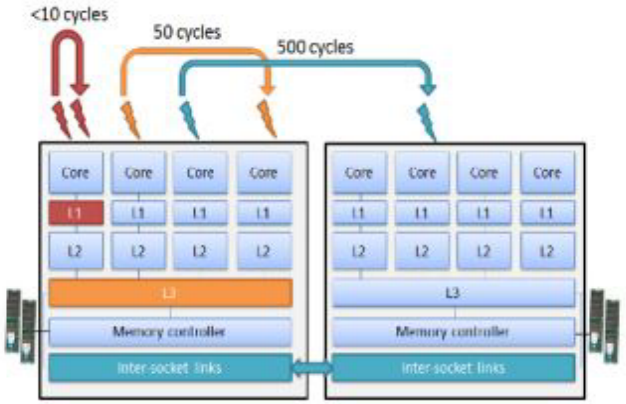


"Big Data Analytics", David Loshin

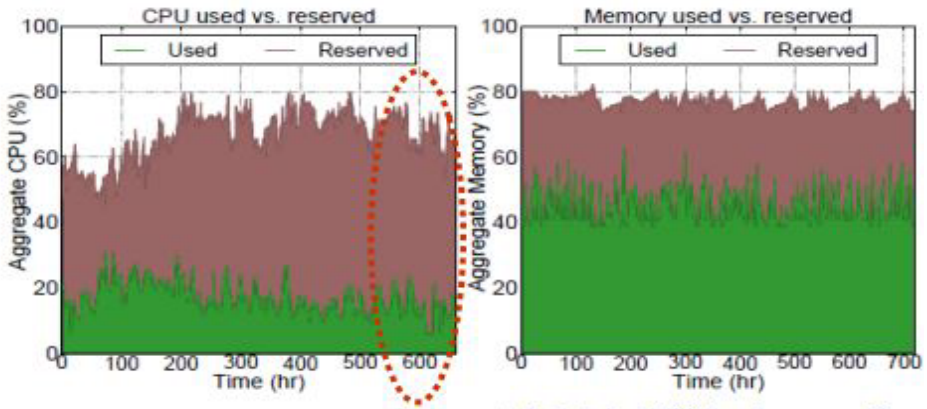
Scalability — Scale Up & Scale Out

- 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