# Why Parallelism: Comprehensive Review

DSC 208R - Data Management for Analytics

June 2025

## Why Bother with Processing Data on Clusters?

The fundamental question addressed is why it is necessary to process data on clusters rather than relying solely on single-node processing or sampling.

### Why Sampling Does Not Suffice?

- Sampling involves taking a small, representative subset of a large dataset for analysis.
- While sampling can be useful for quick insights, it often does not suffice for detailed analysis, especially when:
  - The insights are found in the "long tail" of data distributions, meaning rare events or minority groups that are unlikely to be captured by a sample.
  - Granular phenomena, such as individual user behaviors or specific scientific observations, need to be studied.
  - The analysis requires high precision or accuracy across the entire dataset.

## The Rise of Large-Scale Data

Large-scale data, often referred to as "Big Data," has become a transformative force across various domains.

### Illustrative Examples of Large-Scale Data

- **Astronomy (SDSS)**: The Sloan Digital Sky Survey has generated over 1PB+ of high-resolution images since 2000, at a rate of approximately 200 GB per day, enabling astronomers to study complex galactic evolution behaviors.
- **Genomics (Precision Medicine)**: Analyzing human genomes across cohorts for precision medicine involves data on the order of 3GB per human genome, leading to exabytes (EB) of data for a country like the USA.
- **E-commerce (Recommendations)**: Platforms like Netflix log extensive user behavior (views, clicks, pauses, searches) and use recommender systems that combine terabytes (TBs) of data from all users and content to provide tailored experiences. Over 80% of what people watch on Netflix comes from recommendations.
- **Computer Vision (ImageNet)**: ImageNet, a dataset with over 10 million labeled images across 20,000 classes, exceeds 500GB uncompressed as tensors and was a harbinger of the deep learning revolution.

### Why Large-Scale Data is a Game Changer

Large-scale data enables significant advancements in data science by:

- Allowing the study of granular phenomena in sciences and businesses previously impossible.
- Facilitating new applications and enabling personalization/customization.

- Enabling more complex Machine Learning (ML) prediction targets and mitigating variance to offer high accuracy.

This shift is supported by concurrent advancements in hardware:

- Storage capacity has exploded, with petabyte (PB) clusters becoming common.
- Compute capacity has grown with multi-core processors, GPUs, and other accelerators.
- DRAM capacity has expanded from gigabytes (GBs) to terabytes (TBs).
- Cloud computing has "democratized" access to powerful hardware through Software as a Service (SaaS) models.

# The "Big Data" Concept

"Big Data" is a marketing term that became popular in the late 2000s to early 2010s. It is typically characterized by the "3 Vs":

- **Volume**: Data that is larger than what can be processed or stored on a single node's DRAM.
- **Variety**: Data that comes in diverse forms, including relational data, documents, tweets, multimedia, etc..
- **Velocity**: Data generated at a high rate, such as from sensors or surveillance systems.

Wikipedia defines "Big Data" as "data that is so large and complex that existing toolkits [e.g., RDBMSs] are not adequate".

## Why "Big Data" Now?

The prevalence of "Big Data" is driven by several factors:

1. **Applications**: A new "data-driven mentality" permeates almost all human endeavors, including web services (search, e-commerce, social media), science (satellite imagery, CERN), medicine (pharmacogenomics), logistics (IoT), finance (high-throughput trading), humanities (digitized literature), and governance.
2. **Storage**: The digitization of the world has led to an exponential increase in worldwide byte shipments across various storage media types like HDD, SSD, and NVM-NAND.

Ultimately, to analyze this large-scale data effectively, parallel and scalable data systems are indispensable.