1.1 — Introduction & Motivation

Data Processing at Scale

2026-03-27

Data Processing at Scale — Day 1

Course website

https://vbergeron.github.io/data-processing-at-scale/

This presentation

https://vbergeron.github.io/data-processing-at-scale/1.1-introduction.pdf

Who am I?

- **Valentin Bergeron** — Engineering Manager & Tech Lead @ Ledger
- Worked in AdTech / CRM with high volume
- What I care about
  - Building data intensive applications
  - Functional programing and craft
  - Proving things about programs

What this course teaches

You have access to AI

This lecture focus on what it can’t do.

What this course teaches

- **Vocabulary** — The precise technical terms that let you reason, communicate, and prompt.
- **Judgment** — When to use what, and why.
- **Intuition** — How to know when something is wrong.

How this course was made

- Slides written in **Typst** with the **Touying** presentation framework
- Content structured by an **LLM**, *driven*, *curated* and *edited* by a human
- Source on **GitHub**

Course overview

- **Day 1** — Foundations: distributed systems & scala
- **Day 2** — Storage formats, batch processing (Spark & SparkSQL)
- **Day 3** — Stream processing (Spark & Flink), OLAP engines (Clickhouse)
- **Day 4** — Advanced topics & project briefing

Evaluation

- 100% project presentation (Demo Day 24 / 04)
- Pick a dataset, build a system, defend your architectural choices
- No exam — your understanding shows in what you build and what you can explain

Day 1 — Foundations

1. Introduction & Motivation
2. Distributed Programming with Scala
3. Distributed Systems Fundamentals

Data Processing at Scale

What big data made possible

- **Web search** — indexing and ranking the entire internet in milliseconds
- **Recommendations** — Netflix, Spotify, ByteDance's TikTok: personalized feeds from billions of items
- **Live traffic routing** — GPS recalculating paths from millions of drivers in real time
- **Fraud detection** — scoring millions of card transactions per second
- **Genomics** — sequencing cost from $100M to $100, powered by parallel processing

The data explosion

Year	Global data	Largest HDD	Drives needed
2010	2 ZB	2 TB	1 billion
2015	15 ZB	10 TB	1.5 billion
2020	59 ZB	20 TB	3 billion
2025	175 ZB	36 TB	4.9 billion

Sources: IDC Global DataSphere (Statista); Wikipedia *History of hard disk drives*

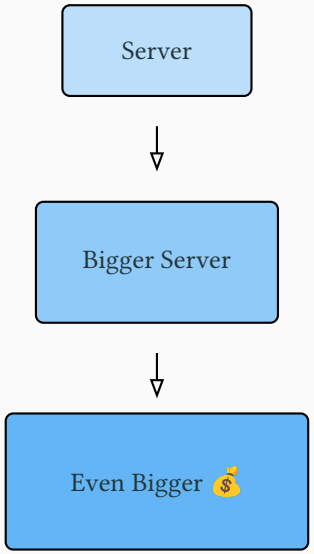
Throughput — the unifying measure

at 1 GB/s sustained throughput

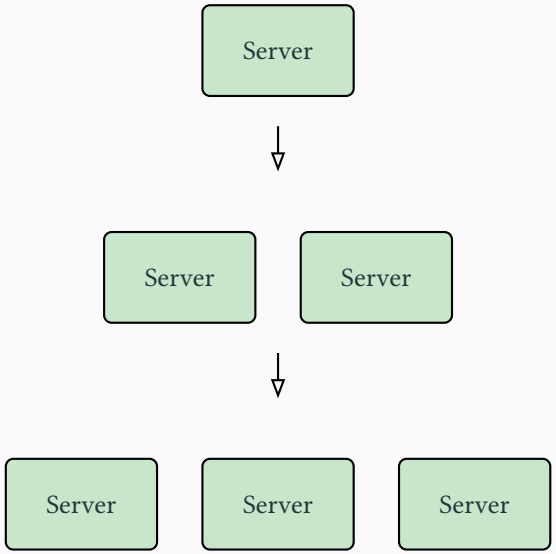
- Throughput is how you reason about data regardless of scale
- Even at excellent throughput, growing data makes wall-clock time explode

Why not just get a bigger machine?

Vertical scaling

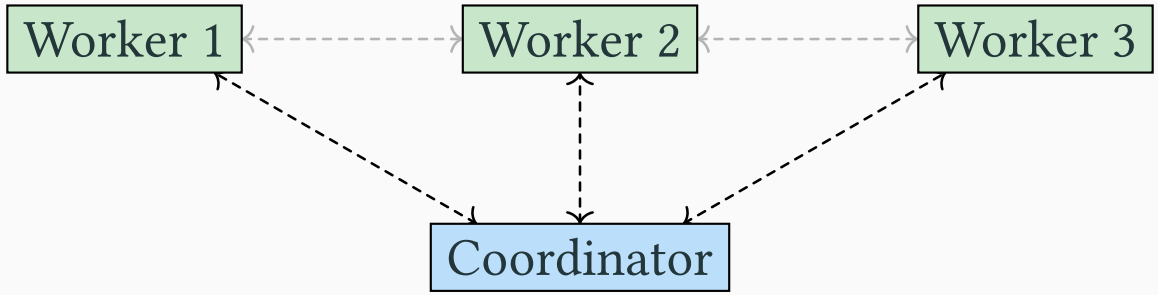


Horizontal scaling



- Moore's Law is flattening — diminishing returns, single point of failure
- **But**: below ~10 TB, a single beefy server is probably the right call
- Most companies don't actually have big data

So we distribute



- Parallelism gives throughput, but coordination has a price
- A network round-trip costs **millions** of CPU cycles
- This tension is the subject of this course

Can software ease the pains of hardware?

Lab

https://vbergeron.github.io/data-processing-at-scale/lab-1.2-single-node-benchmark.pdf

Next

https://vbergeron.github.io/data-processing-at-scale/1.2-scala-fp.pdf