Data-Intensive Distributed Computing

CS 451/651/431/631 (Fall 2021)

Part 1: Introduction to Big Data

Ali Abedi

These slides are available at https://www.student.cs.uwaterloo.ca/~cs451/

Agenda for Today

Who am I?
What is big data?
Why big data?
Course structure?

2

Who am I?

PhD from Waterloo (2017)
Systems and Networking Research Group
5'th time teaching this course

Big Data

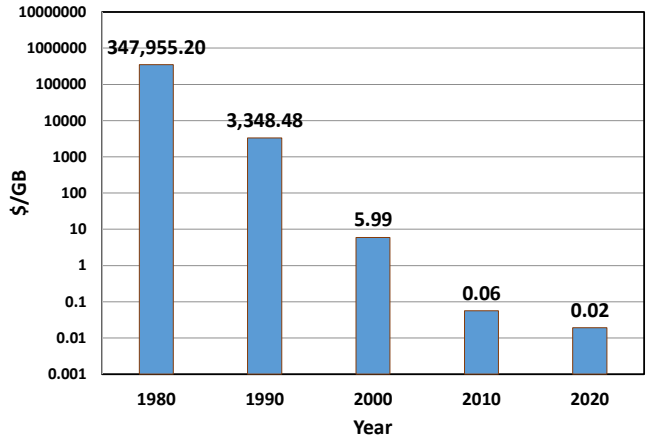
Let’s see what big data is and where it came from.

Two Questions:
1- How much does data storage cost?
2- How much data do we generate?

5

The storage cost has decreased dramatically over the years.
The first hard disk drive, like so many innovations in computing, came from IBM. It was called the IBM Model 350 Disk File and was a huge device. It had 50 24-inch disks contained inside a cabinet that was as large as a cupboard and anything but lightweight. This hulk of a storage unit could store a whopping 3.75 MB of data.

How much does data storage cost?



Two Questions:
1- How much does data storage cost?
2- How much data do we generate?

8

How much data do we generate?

- 4 PB is generated on Facebook everyday
- 500 M tweets on Twitter everyday
- 720,000 h video uploaded to YouTube everyday
- 75 billion IoT devices by 2025

This is tiny sample of the data generated everyday.
Every day over 80 years of video is uploaded to YouTube!
Soon billions of Internet of Things (IoT) devices will generate a lot of data even if each only generates 10s of bytes each day.

How much data do we generate?

- 2.5 exabytes (2,500,000 TB) of data is generated each day

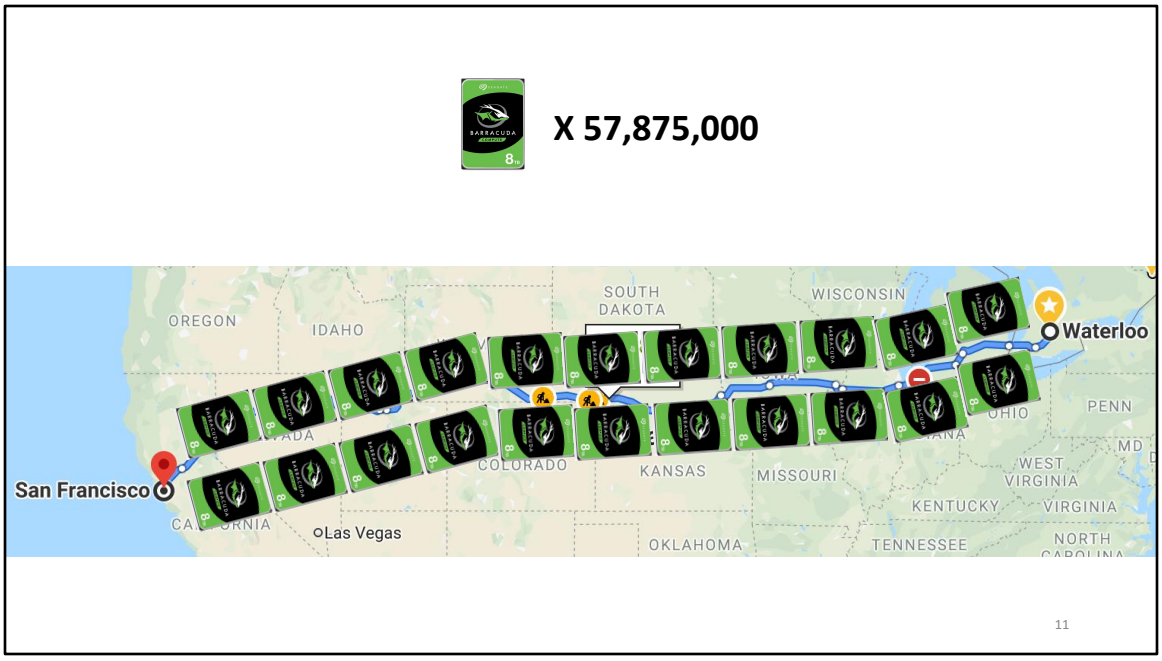
X 312,500

- 90% of all data has been created in the last two years
- 463 exabytes of data will be generated each day in 2025

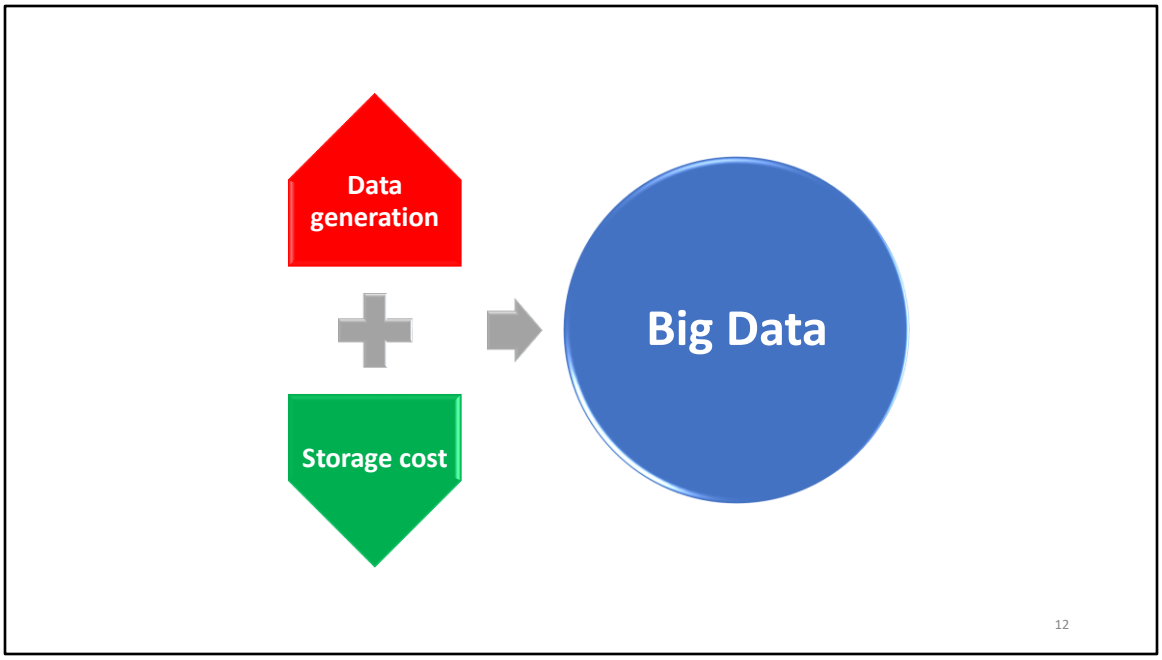
X 57,875,000

10

Every person generates 1.7 megabytes in just a second.
Although we generate a lot of data today, it is nothing compared to what we will generate in the near future!



This is how 50M HDDs look like ☺



The combination of low storage cost and high rate of data generation has created big data.

We now talk about why big data is important.

Big data has significant impacts on business, science, and society.

Business

- Data-driven decisions
- Data-driven products

15

Business Intelligence

An organization should retain data that result from carrying out its mission and exploit those data to generate insights that benefit the organization, for example, market analysis, strategic planning, decision making, etc.

Duh!?

16

This is not a new idea!

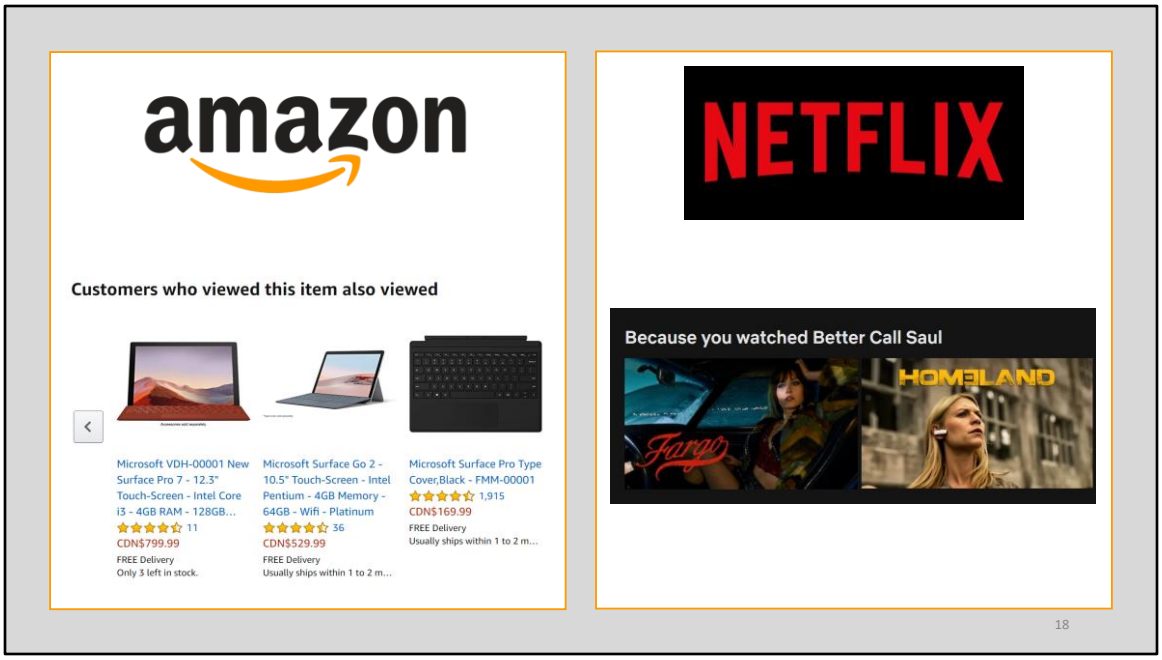
In the 1990s, Wal-Mart found that customers tended to buy diapers and beer together. So they put them next to each other and increased sales of both.*

So what's changed?

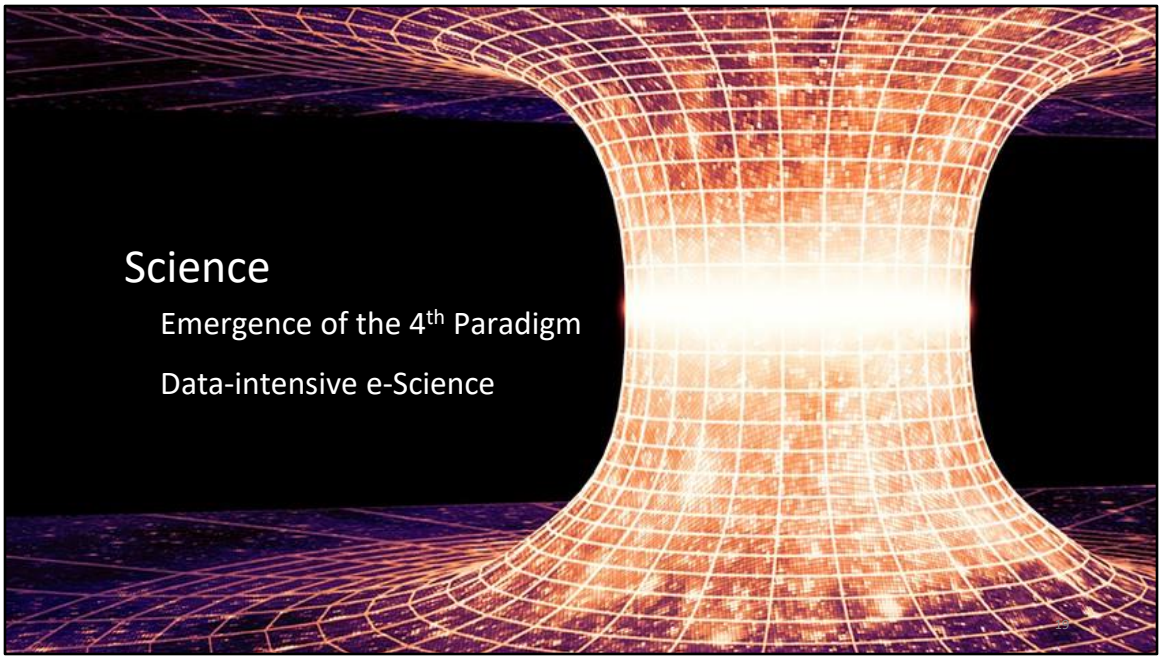
More compute and storage

Ability to gather behavioral data

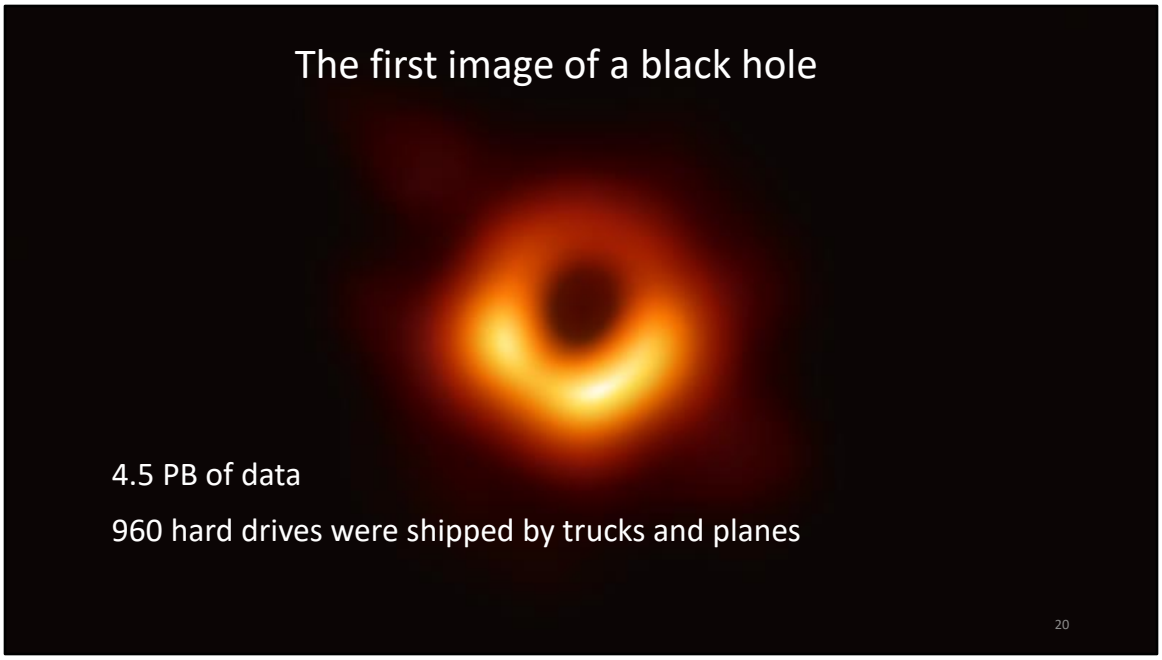
* BTW, this is completely apocryphal. (But it makes a nice story.)



For example, amazon has an online shopping service. By analysing user behavior (data science), it finds out that customers tent to buy certain items together. By gaining this insight, it develops a product recommendation system (data product). This cycle continues and hopefully the company makes money.

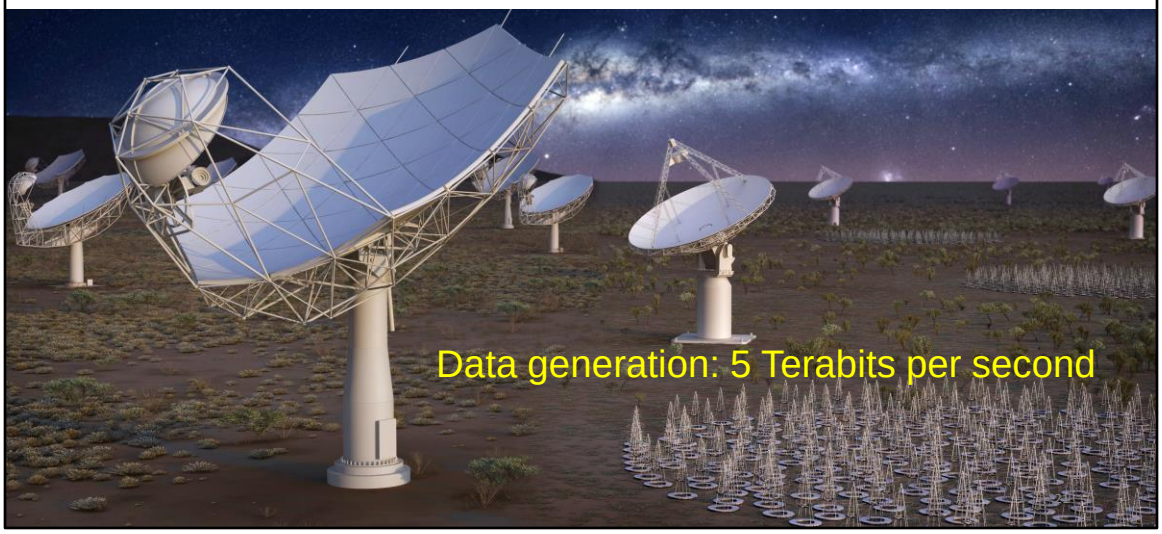


New experimentation tools generate a lot of data which makes data processing very challenging. Next, we see a few examples.



Used 8 telescopes over several days to collect the data. The volume of data was so much that it would take around 25 years to transfer it over the Internet. So they used trucks and planes to move the data. Apparently they didn't take the big data course ☺

Square Kilometre Array (SKA) telescope
a huge big data challenge



Expected to be operational in 2027. This gigantic telescope will generate so much data that is impossible to process today.
The big data challenge is one of the main outstanding problems of this project.

Society

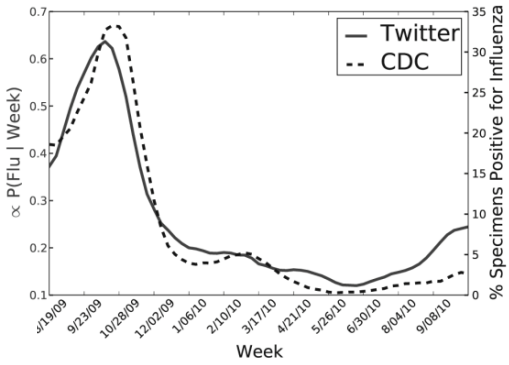
Humans as social sensors

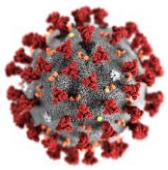
Computational social science

22

Let's now review the impact of big data on society. Thanks to social networks people create a lot of content on the Internet. They are like sensors that report their observations and thought on social media platforms. How can we process this data?

Predicting *X* with Twitter





Fall 2020, CS451 project: Use data sources such as Twitter to predict the spread of COVID-19

23

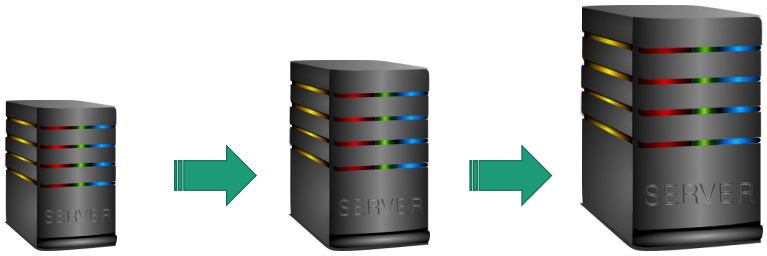
There are many studies that try to predict something (X) from Twitter data. For example, there are studies on estimating the spread of a disease only using Twitter. The graph shows a good match between the estimated data and ground truth (CDC data).

And that’s how big data became the new hot topic!


Tackling Big Data!
25

Vertical scaling (scaling up)

Super expensive!



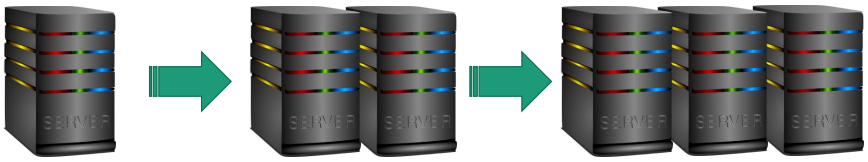
But this is expensive and limited!

26

To deal with big data we need more and more processing power.
One way to achieve this is to upgrade our server for example by putting more RAM modules in it. Or replace it with a more powerful server.
This approach is very expensive and does not scale well because there is a limit on how powerful a server can be today.

Horizontal scaling (scaling out)

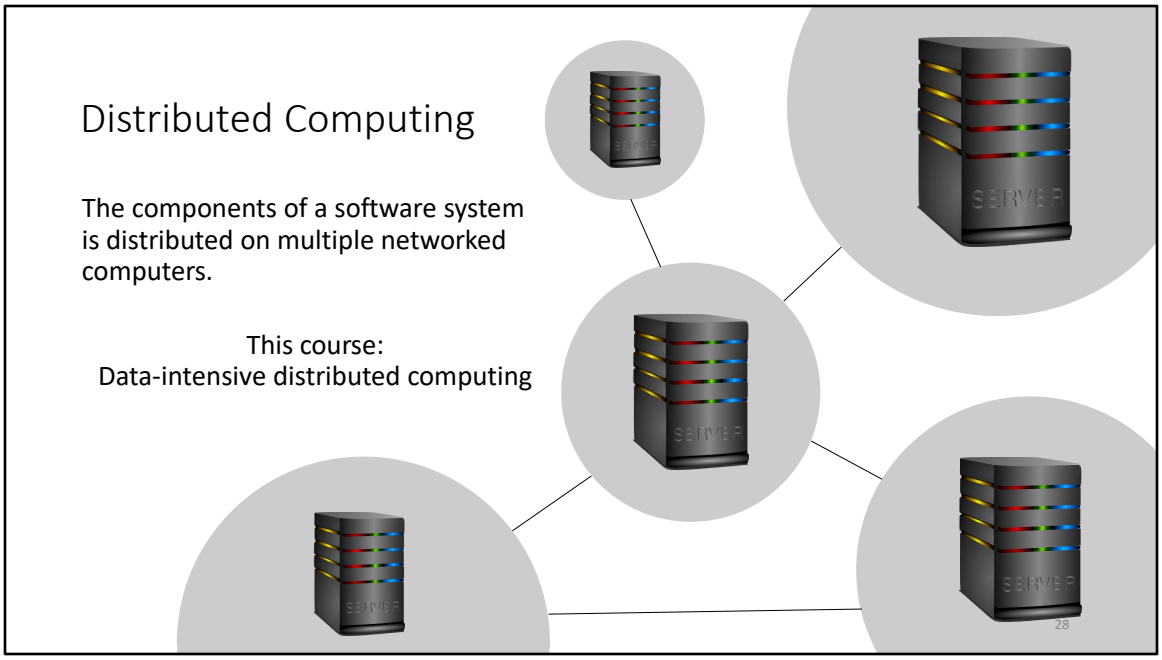
Inexpensive computers



Nice! Enters distributed computing …

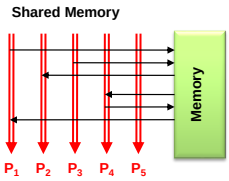
27

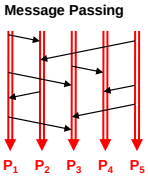
On the other hand, instead of making the server more powerful we can buy more cheap servers! This is really cool but it brings it own challenges (hence this course).



In this course, we study how we can process big data files on many cheap commodity servers.

Parallelization on even a single machine is challenging





⚠ CS350 reminder! ⚠

Basic primitives: Locks, condition variables, semaphores

Problems: Deadlock, livelock, race condition

29

But running a program needs parallelizing processing over multiple servers.
We know that parallelization is so challenging even on a single machine!

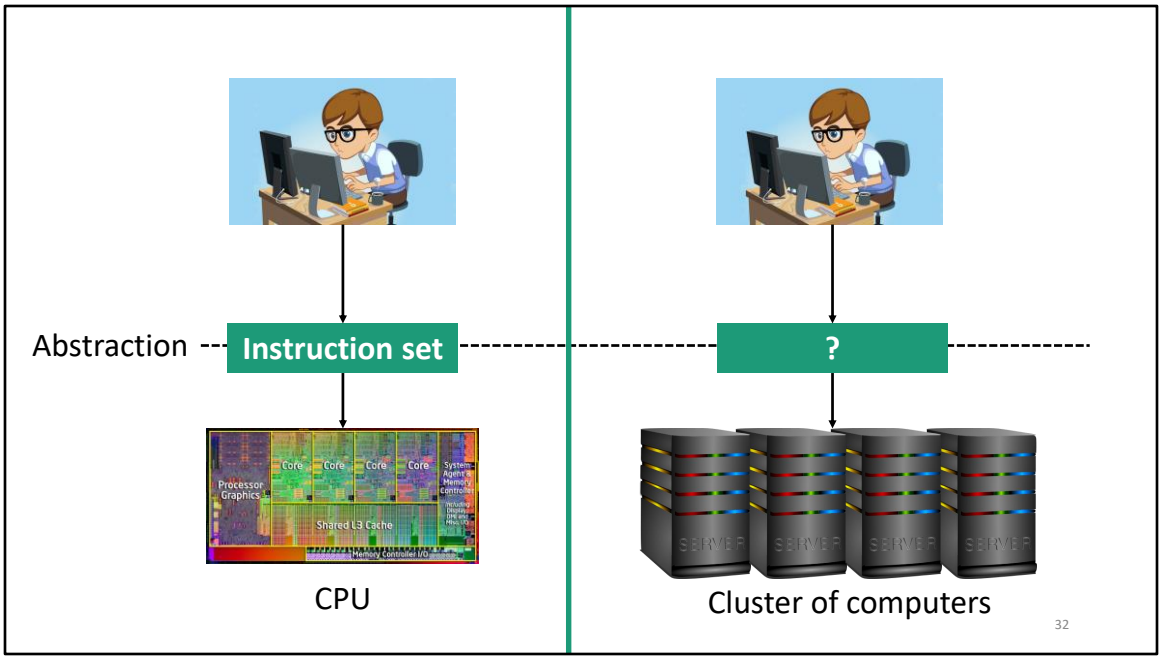
Parallelization on multiple computers

The scale of clusters and (multiple) datacenters
The presence of hardware failures and software bugs
The presence of multiple interacting services

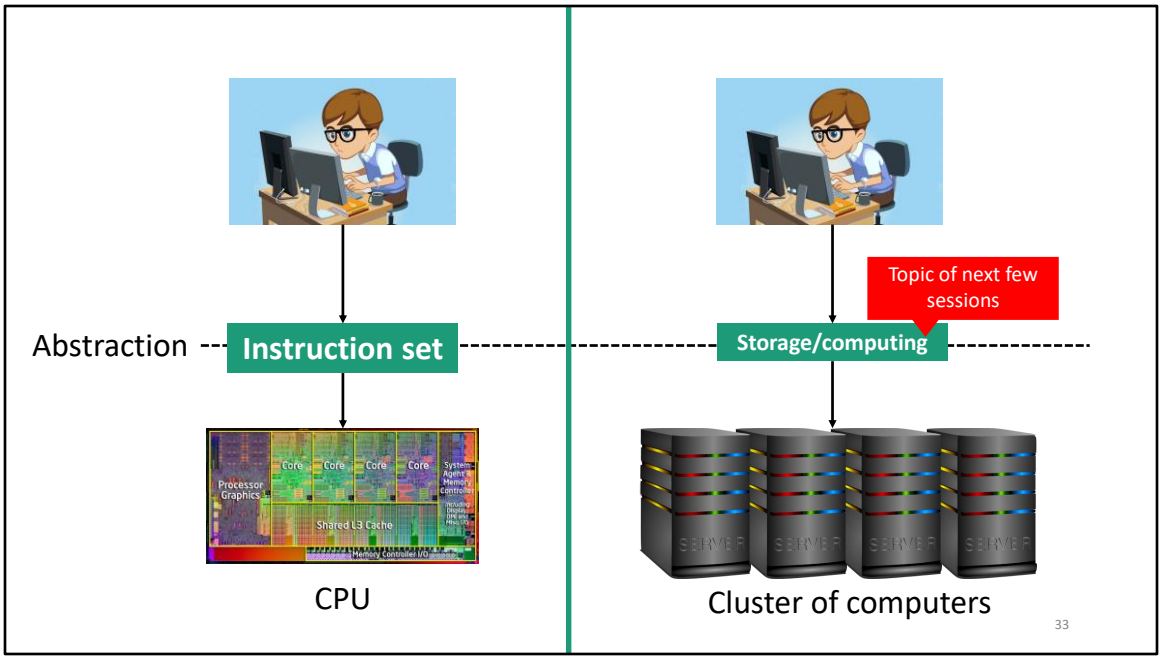
It is difficult!

30

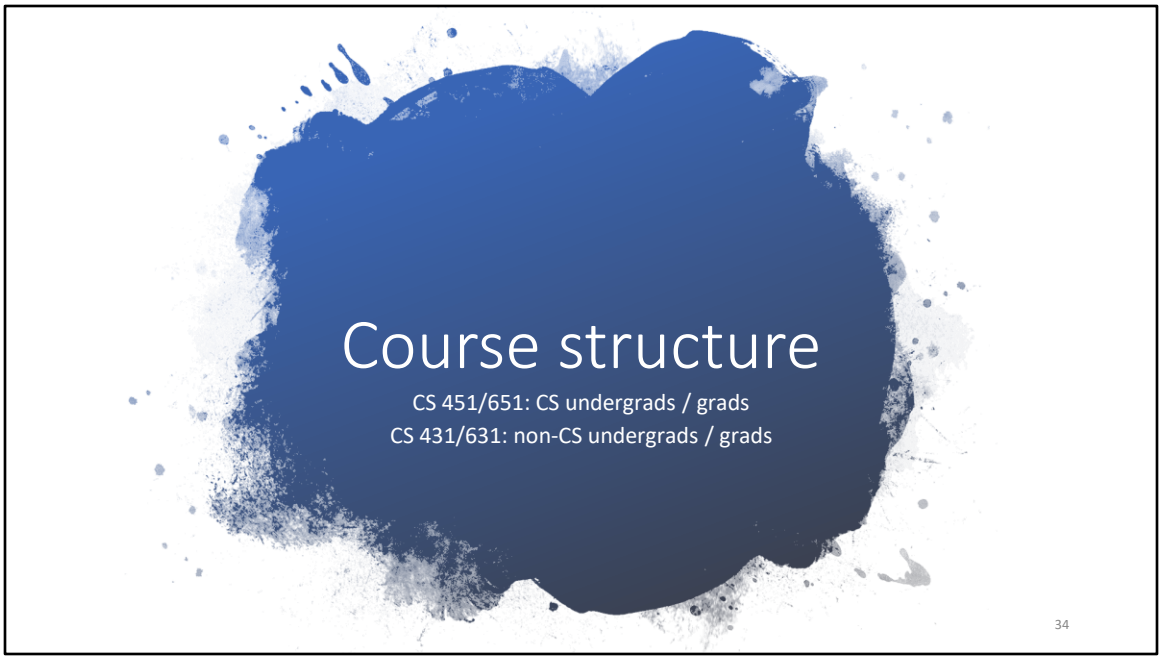
Now add the complexities of a cluster of servers!
Bottom line: it is very difficult to do.



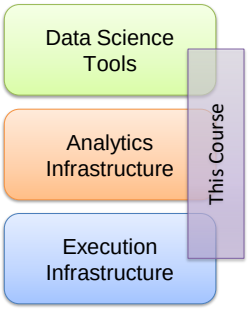
Abstraction comes to rescue.
The instruction set of a CPU provides an abstraction layer that hides away the complexity of the architecture of a CPU.
When we add 2 variables in our problem we often have no idea how it's actually done in the CPU.
Similarly, we need an abstraction layer to hide away the complexities of a cluster of computers (or even a datacenter)



We need a solution for both storage and computing.


Course structure
CS 451/651: CS undergrads / grads
CS 431/631: non-CS undergrads / grads
34

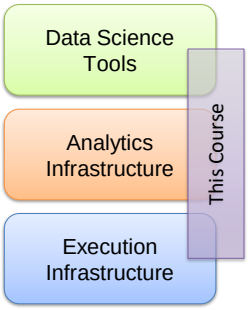
What is this course about?



"big data stack"

Buzzwords

data science, data analytics, business intelligence, data warehouses and data lakes



MapReduce, Spark, Pig, Hive, noSQL, Pregel, Giraph, Storm/Heron

"big data stack"

Text: frequency estimation, language models, inverted indexes

Graphs: graph traversals, random walks (PageRank)

Relational data: SQL, joins, column stores

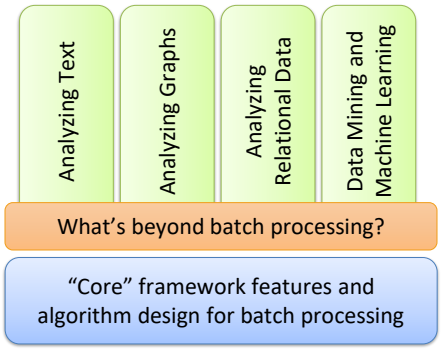
Data mining: hashing, clustering (*k*-means), classification, recommendations

Streams: probabilistic data structures (Bloom filters, CMS, HLL counters)

This course focuses on algorithm design and "thinking at scale"

36

Structure of the Course



37

Time spent on the course

W21 course evaluation

CS451/651: 9.5 hours/week
CS431/631: 7.5 hours/week

Important Coordinates

Course website:
https://www.student.cs.uwaterloo.ca/~cs451/

Lots of info there, read it!
("I didn't see it" will not be accepted as an excuse)

Communicating with us:
Piazza for general/private questions (link on course homepage)

Lectures and Office hours

39

Course Design

This course focuses on algorithm design and "thinking at scale"

Not the "mechanics" (API, command-line invocations, et.)
You're expected to pick up MapReduce/Spark with minimal help

Components of the final grade:

CS451/651: 8 *individual* assignments
CS431/631: 7 *individual* assignments
group final project (up to 3 students)
Weekly quizzes

Please set an alarm for weekly quizzes!

40

Expectations

Your background:

CS451/651 : Comfortable in Java and Scala (or be ready to pick it up quickly)
CS431/631 : Comfortable in Python (or be ready to pick it up quickly)

You are:

Genuinely interested in the topic
Be prepared to put in the time
Comfortable with rapidly-evolving software

41

Academic Integrity

All assignments will be checked for cheating!
0 on assignment + penalty on final grade

Assignment Mechanics (CS451/651)

Java

Scala 

We'll be using <u>private</u> Git repos for assignments

Complete your assignments, push to GitLab
We'll pull your repos at the deadline and grade

Late assignments will get 0

Assignment Mechanics (CS431/631)

Assignments will use Python and Jupyter (Google Colab)
Everything you need to know is in the assignment itself

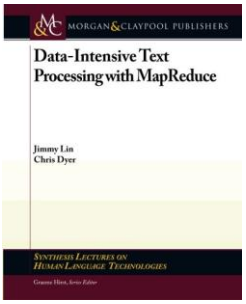
Assignments will generally be submitted using Git
Details are on the course website for the appropriate assignment

Python

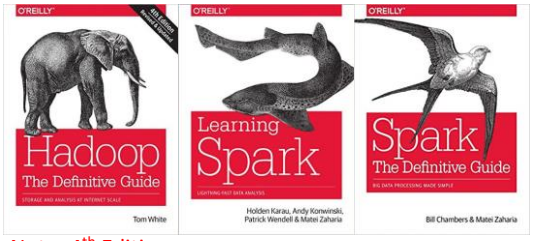
Late assignments will get 0

Course Materials

One (required) textbook +
Three (optional but recommended) books +
Additional readings from other sources as appropriate



(optional but recommended)



Note: 4th Edition