BIG DATA COMPUTING

ID's last digit: 0 – 4

Andrea Pietracaprina

Department of Information Engineering

University of Padova

andrea.pietracaprina@unipd.it

OUTLINE

1. Big Data Phenomenon
2. Computational Challenges
3. Organization of the Course
4. Administrative Issues

Big Data Phenomenon

DATA

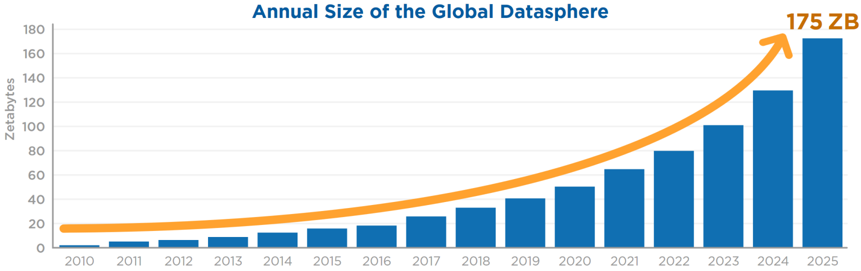
“~~Space~~ is big. Really big”

(*Douglas Adams, The Hitchhiker’s Guide to the Galaxy*)

Why is DATA growing so much?

- Technological progress:
  - Growth of storage capacity
  - Growth of comunication bandwith
  - Growth of computing capacity
- Reduction of ICT costs
- Pervasiveness of digital technologies: scientific research, health, business, politics, social interactions, ...

Big Data Phenomenon



From: *The Digitization of the World (IDC, 2018)*

How big is 175ZB?:

- 1 ZettaByte (ZB) $=$ 1 trillion GB $= 10^{12}$ GB;
- 175 ZB $\equiv$ 23 parallel stacks of DVD from Earth to Moon;
- Downloading 175 ZB at 1Gb/s takes $>$ 43 million years

Big Data Phenomenon

The world continuously collects huge amounts of:

SCIENCE

- Physical data: from sensors, telescopes, particle physics experiments.
- Biological/medical data: from genetic studies, patient monitoring, epidemic evolution analyses.

SOCIAL INTERACTIONS

- Human activity data: from social networks, mobile devices, internet/web traffic, IoT systems.

BUSINESS

- Business data: from online stores, customer profiling, bank/credit-card/financial services, quality-of-service monitoring.

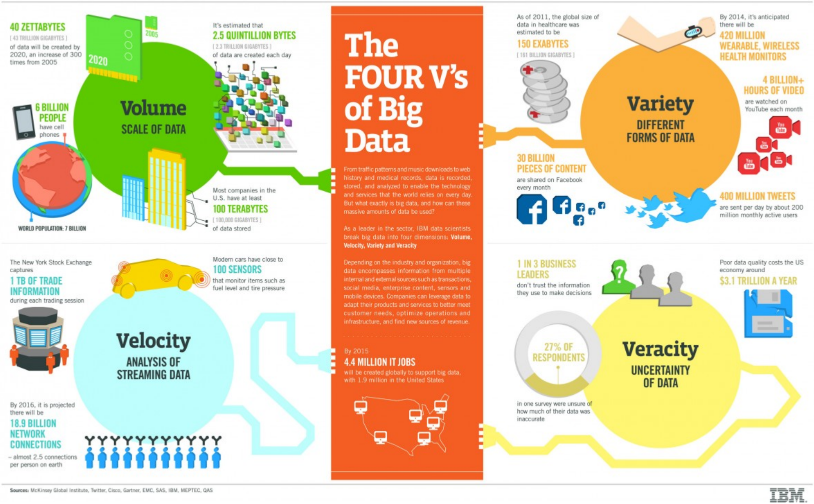
Big Data Phenomenon

The term Big Data relates to two distinct issues:

- ISSUE 1:
  - Data produced everywhere;
  - Need for automated analytics (vs human inspection);
  - Challenges: identification of suitable analysis tools, data selection/preparation.
- ISSUE 2: Focus: feasibility and efficiency of the computation
  - Massive datasets need to be processed;
  - Traditional (algorithmic) approaches are unsuited;
  - Challenges: development of novel computing frameworks, novel solutions

This course focuses on ISSUE 2!

Computing Challenges



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Source: IBM Big Data & Analytics Hub

Computational Challenges

- **Volume:** processing huge datasets poses several challenges and requires a data-centric perspective.
- **Veracity:** large datasets coming from real-world applications are likely to contain *noisy, uncertain data*, hence accuracy of solutions must be reconsidered.
- **Velocity:** sometimes, the data arrive at such a high rate that they cannot be stored and processed offline. Hence stream processing is needed.
- **Variety:** large datasets arise in *very different scenarios*. More effective processing is achieved by adapting to the actual characteristics of data.

The above issues require a

paradigm shift w.r.t. traditional computing.

Computational Challenges

To tackle the above challenges effectively, one needs:

- Platforms with:
  - High storage capacity and computing power
    $\Rightarrow$ parallel/distributed architectures
  - Moderate costs
  - Ease of programming and management
- Focus on accuracy-resource tradeoffs, to cope with size, noise, and uncertainty of data
- Data-centric view
- Data stream processing (sometimes)

Big Data Computing Course

What will we learn?

1. Novel computing/programming frameworks for big data processing: theory and practice
2. Key techniques to process large-scale data
   - Rigorous setting (provable guarantees)
   - Application to fundamental data analysis primitives

Specific topics

1. Frameworks: Distributed (MapReduce, Apache Spark) and Streaming
2. Techniques:
   - Coresets (application to clustering);
   - Sampling (application to frequent itemsets);
   - Locality sensitive hashing (application to similarity search);
   - Sketches (application to estimating of moments)

Organization of the Course

Subdivision into classes

The students are subdivided into two parallel classes based on their ID's last digit (*same syllabus, homeworks, and exams*)

- Class A (prof. Pietracaprina): last digit 0-4
- Class B (prof. Silvestri): last digit 5-9

Lectures

Lectures will be in presence and online via Zoom. For Class A, the Zoom link is

https://unipd.zoom.us/j/84232837748

- For each topic, partial slide sets are made available in advance. Final versions (together with solutions to exercises) are uploaded after the topic is fully covered.
- *Attendance and active participation are strongly encouraged*.

Organization of the Course

Exam

- Final written exam (26 points)
- Homeworks: programming assignments (6+1 points)
  - Groups of 2-3 students (even from different classes)
  - 3 homeworks, approx. one every 3 weeks.
  - Use of Apache Spark on individual PCs (Homeworks 1-2) and on CloudVeneto (Homework 3)
  - All group members receive the same grade. The extra point is given if all homeworks submitted by the respective deadlines.
- Oral exam: at teacher's discretion (see rules in Moodle)

Organization of the Course

Required background

- Java (preferred) or Python programming
- Basic algorithmics: asymptotic, worst-case analysis; fundamental algorithms and data structures; (e.g., lists, queues, stacks, hash tables, maps/dictionaries)
- Basic math tools, combinatorics, and probability.

Administrative Issues

Online tools

- **Course Moodle:**
  - Announcements and student forum.
  - Infos: Zoom, contacts, textbooks, exam rules and sessions.
  - Lectures diary.
  - Material: slides, videos, exercises, articles.
  - Preliminary exams grades.
- **Uniweb:** Official exam lists and final grades.
- **Exam Moodle (only one for the two classes):**

  https://esami.elearning.unipd.it/course/view.php?id=3801

  - Formation of groups for Homeworks
  - Submission of Homeworks.
  - Written tests (if online).

Administrative Issues

Contacts and office hours

- Teacher (prof. Andrea Pietracaprina): andrea.pietracaprina@unipd.it
- TAs (dott. Paolo Pellizzoni and dott. Nicolò Penzo): bdc-course@dei.unipd.it

Office hours are by appointment (via Email). Teaching assistants should be contacted only for questions related to homeworks.

TODO: As soon as possible

- Register in the Course Moodle (no password)
- Register in the Exam Moodle (password: bigdatalab)
- Form groups of size at most 3 for the homeworks **by March 15**. Once a group is formed, it must be registered in the Exam Moodle using the Group registration link.