DMIF, University of Udine

Data Management for Big Data

Course Introduction

Andrea Brunello
andrea.brunello@uniud.it

April 2022

Introduction

Andrea Brunello, Postdoctoral Researcher at the University of Udine, and member of the Data Science and Automatic Verification Laboratory `https://datasciencelab.dimi.uniud.it/`

My research interests are in the fields of Explainable AI, Virtual Sensing, Machine Learning for Healthcare and, broadly speaking, Data Science. I have also been working on Business Analytics, Data Warehousing and Data Integration projects.

Our laboratory works with local, national, and international partners on Data Science and Automatic Verification projects, with a particular emphasis on the cross-contamination and integration between the two areas.

Content Outline

We are going to spend 12 hours together, covering the topics:

- Relational Databases and SQL
- NoSQL Databases
- Data Warehousing
- Big Data and Business Analytics

A Brief History of Information Systems

The Importance of Data

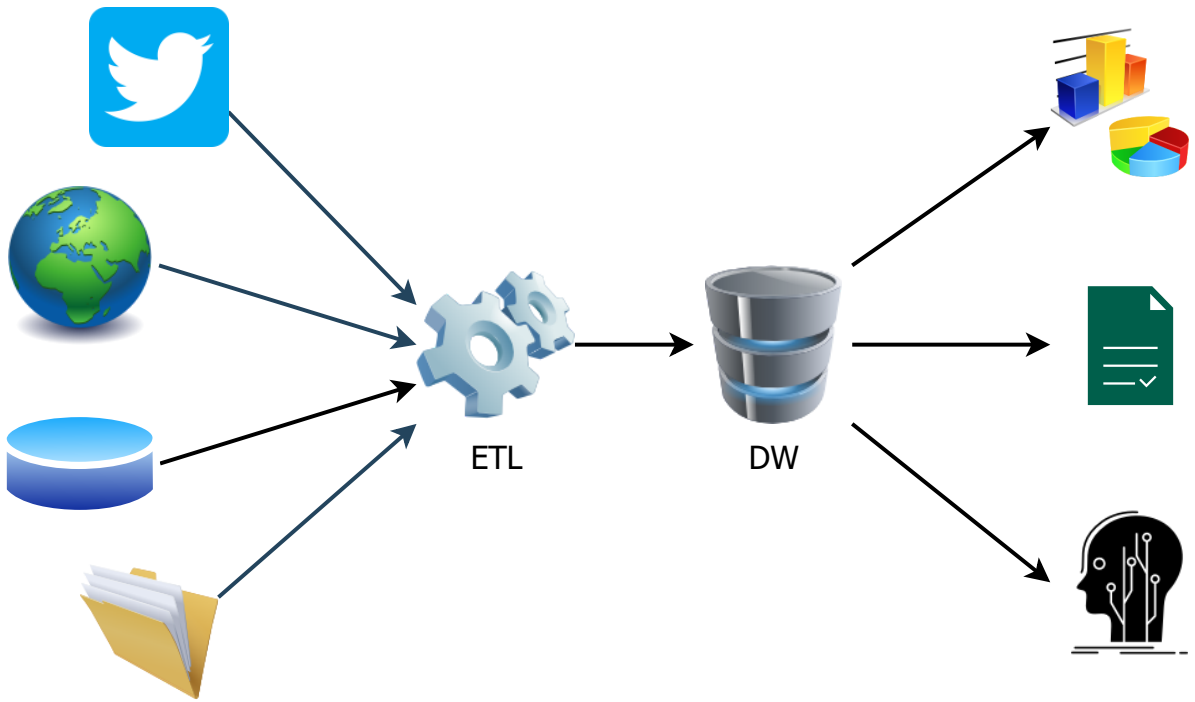
Regardless of the domain, data is driving the future of IT systems and a massive number of technologies across multiple industries heavily depend on it to thrive.

Data can be defined as a collection of raw, unanalyzed, unorganized material.

Information is data that has been processed, aggregated and organized into a human-friendly format that provides more context (data visualizations, reports, dashboards, . . .).

Knowledge derives from a combination of information, experience and intuition. It allows one to draw inferences and develop insights and thus it can assist in decision making.

The Data Analytics Workflow

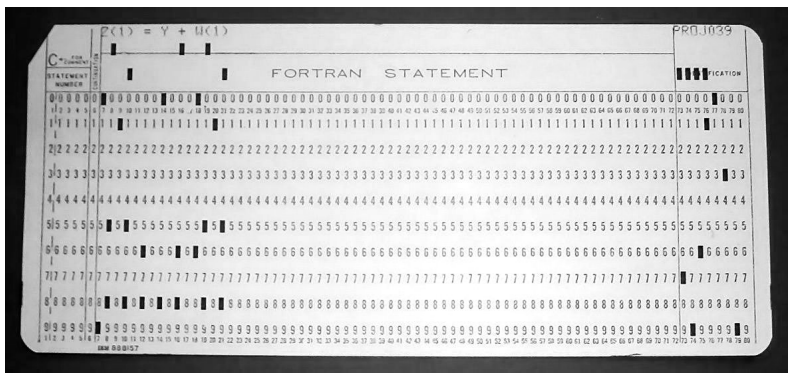


From Bookkeeping to Magnetic Tapes

Before electronic data processing, companies used to manage their customers, purchases and inventory using traditional bookkeeping methods.

Early electronic data processing came about in the 1950s; initial systems were based on punch cards, naturally exposed to damage and loss of data.

From 1960s magnetic tapes provided better data storage, but sequential access required full tape scan even for 5% of data. Many dedicated hardware as many formats available.



From batch to interactive processing

In the decades 50s-60s we have also the switch from batch to interactive processing.

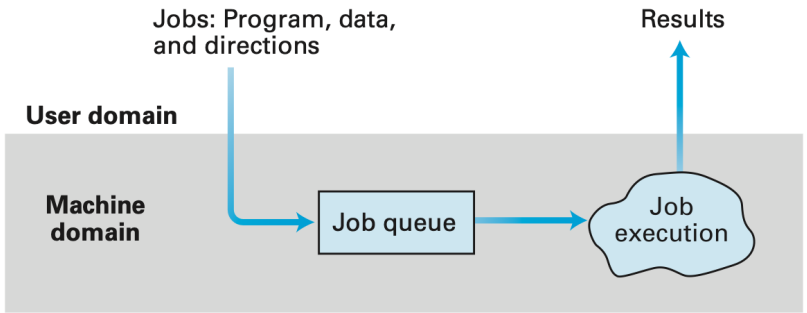


Figure: Batch processing

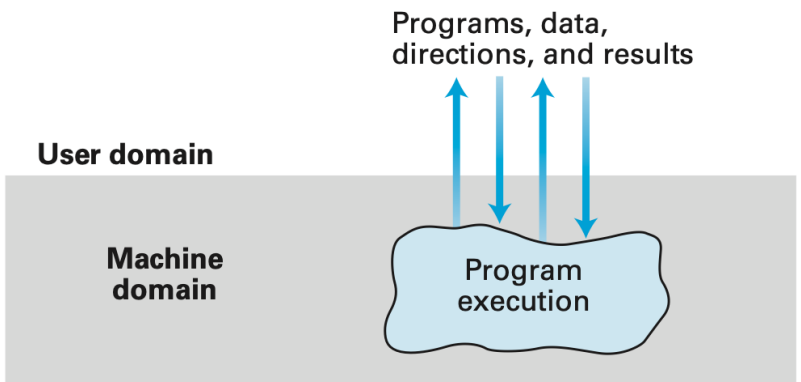


Figure: Interactive processing

Direct Access Storage Device

The 1970s saw the advent of disk storage, also referred to as direct access storage device (DASD).

No need to go through records $1, 2, 3, ...n$ to get to record $n + 1$ once the location address of $n + 1$ is known.

The time to locate a record is measured in milliseconds.

New, more complex data structures were developed, such as lists and trees to be stored on disk.

Along with the DASD, it came a new type of system software known as a database management system (DBMS).

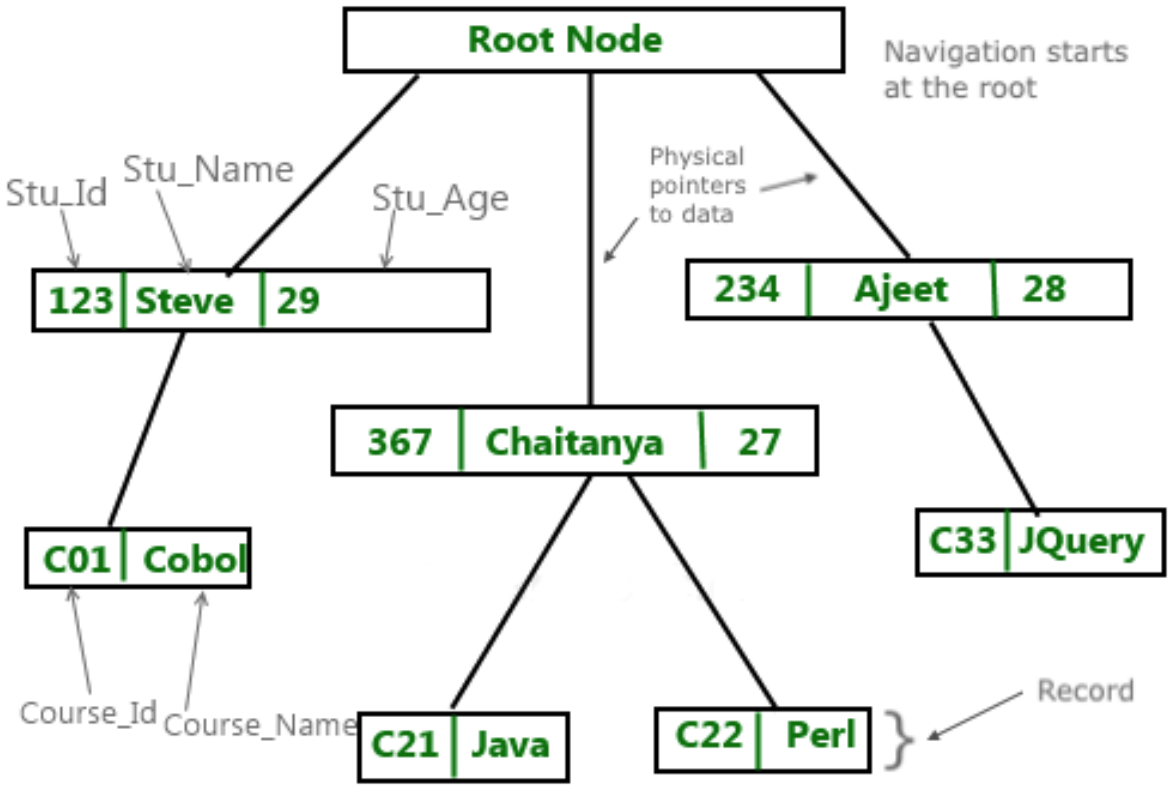
Online transaction processing

With DBMS it was easier to store and access data on a DASD; moreover, the DBMS took care of tasks such as storing data on a DASD, indexing data, managing access rights, and so forth.

By the mid-1970s, online transaction processing (OLTP) made even faster access to data possible. Applications like bank teller systems and manufacturing control systems became possible.

At this point, network and hierarchical data models became of widespread use: they rely on graphs and trees as structures to store data.

Hierarchical data model



Major drawback: there can be only one-to-many relationships between nodes

The advent of the relational model

Meanwhile, Ted Codd defines the relational data model:

- Would win the ACM Turing Award for this work
- Points of strength: simplicty and possibility of hiding physical details
- IBM Research begins work on System R prototype
- UC Berkeley begins work on Ingres prototype
- They both still exhibit computationally worse performance with respect to network and hierarchical data models
- By the 80s prototypes evolve into commercial systems
- SQL becomes industrial standard

Wal-Mart and the Birth of the DW

Around 1990 Wal-Mart retail corporation began to achieve wide acclaim for its mastery of supply chain management.

Behind this success was Wal-Mart's data warehouse, and a new way of interacting with data, called OLAP.

Data is collected by its point-of-sales systems to achieve unprecedented insight into the purchasing habits of its 100 million customers and the logistics guiding its 25,000 suppliers.

Wal-Mart's data warehouse was the first commercial Enterprise Data Warehouse to reach 1 terabyte of data in 1992.

Large, Rapid and Heterogeneous Data

Around 2000s, the types of data stored in database systems evolved rapidly, pushed by an ever increasing usage of the Internet and multimedia.

The variety of new data-intensive applications led to NoSQL systems, which gave programmers greater flexibility to work with new types of data, but lacked a high level query language.

Distributed storage and computing frameworks were developed, such as Hadoop.

To allow for the interchange of information between systems, formats such as XML and JSON became of widespread usage.

Data Mining applications started to emerge.

Nowadays, most of large and medium sized organizations rely on decision support systems.

From DSS to Decision Management Systems

- A decision is the selection of a course of action from a set of alternatives
- A DSS recommends such an action by offering managers information upon which to build ideas so to come up with the final choice
- A DMS is an "action-oriented" evolution of a DSS
- It makes one step more and takes decisions without human intervention based on known information and a set of coded business rules
- Of course, not all judgments may be automated (strategic vs operational decisions)

References

A. Silberschatz, H.F. Korth, S. Sudarshan *Database system concepts*, 7th Edition, 2020.

W.H. Inmon, *Building the Data Warehouse*, 4th Edition, 2005.