Data Stewards and Data Management Principles

(LTAT.02.014)

Priit Adler
Research Fellow, University of Tartu
Elixir EE, BIIT

2021.05.11
Recap
Data Literacy

the ability to read, understand, create, and communicate data as information.
Data Practices Courseware

1.5 Analyze and Report - [https://datapRACTICES.org/courseware/1_5.html](https://datapRACTICES.org/courseware/1_5.html)
Big data

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.
Every minute of the day:

- Pinterest users pin 3,472 images.
- Vine users share 8,333 videos.
- Skype users connect for 23,300 hours.
- Yelp users post 26,380 reviews.
- Apple users download 48,000 apps.
- Pandora users listen to 61,141 hours of music.
- YouTube users upload 72 hours of new video.
- Email users send 204,000,000 messages.
- Facebook users share 2,460,000 pieces of content.
- Tinder users swipe 416,667 times.
- WhatsApp users share 3,472,222 photos.
- Amazon users make $83,000 in online sales.
- Instagram users post 216,000 new photos.
- Twitter users tweet 277,000 times.
Transactions + Interactions + Observations = BIG DATA

At the same time ...

https://www.slideshare.net/welkaim/introduction-to-big-data-65870623
The Vs of Big Data

- **Volume**
  - 40 Zettabytes (43 Trillion Gigabytes) of data will be created by 2020, an increase of 300 times from 2005.
  - 2.5 Quintillion Bytes 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 of data stored.

https://www.slideshare.net/welkaim/introduction-to-big-data-65870623
The Vs of Big Data

- **Velocity**: Analysis of streaming data.
- **Volume**: 1 TB of trade information during each trading session.
- **Variety**: Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.
- **Veracity**: By 2016, it is projected there will be 18.9 billion network connections – almost 2.5 connections per person on earth.

https://www.slideshare.net/welkaim/introduction-to-big-data-65870623
The Vs of Big 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

DIFERENT FORMS OF DATA

30 BILLION PIECES OF CONTENT are shared on Facebook every month

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

400 MILLION TWEETS are sent per day by about 200 million monthly active users
The Vs of Big Data

1 in 3 business leaders don't trust the information they use to make decisions.

27% of respondents in one survey were unsure of how much of their data was inaccurate.

Poor data quality costs the US economy around $3.1 trillion a year.

https://www.slideshare.net/welkaim/introduction-to-big-data-65870623
The Vs of Big Data

The ‘Datafication’ of our World:
- Activities
- Conversations
- Words
- Voice
- Social Media
- Browser logs
- Photos
- Videos
- Sensors
- Etc.

Analysis

Analysing Big Data:
- Text analytics
- Sentiment analysis
- Face recognition
- Voice analytics
- Movement analytics
- Etc.

Value
Big Data Management
(6 ECTS) LTAT.02.003

The objective of this course is to introduce students to the principles and methods of advanced data management and processing. The course will cover the techniques of storing and processing different types of data (structured, semi-structured and unstructured). It will cover the state-of-the-art in different types of big data processing systems (e.g., stream processing, graph data processing, scalable machine and deep learning systems).
Distributed Systems

(6 ECTS) LTAT.06.007

The student will get the main idea of construction and basic principles of work of distributed systems. After passing the course the student is able to create the distributed systems, analyse and improve the existing systems.
Homework 8

Compile detailed DMP

• OP1: Continue with the project from HW7
  Improve the plan to reduce bad practices. Write the plan, that would have saved time and effort for the researchers.

• OP2: Write one for one of your own projects
  Ideally for running project and something you can use from now on

• Deadline 31st of May