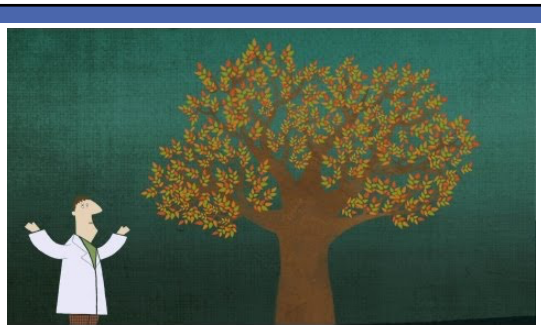
Curso de extensão em Data Science

GERÊNCIA DE INFRAESTRUTURA PARA BIG DATA

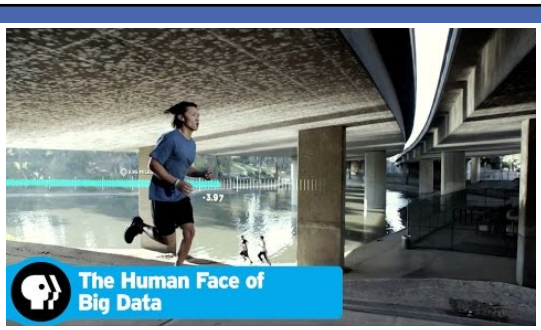
Prof. Tiago Ferreto – tiago.ferreto@pucrs.br

ESCOLA POLITÉCNICA

BIG DATA



Big Data - Tim Smith - https://www.youtube.com/watch?v=j-0cUmUyb-Y



THE HUMAN FACE OF BIG DATA | Trailer | PBS - https://www.youtube.com/watch?v=kAZ8IK224Kw

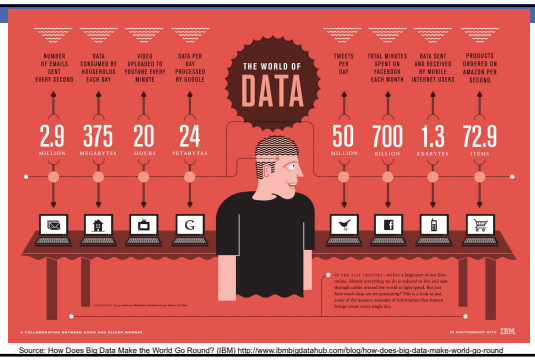
Data age!

- IDC estimation of the "digital universe" size
  - 4.4 zettabytes in 2013 → 44 zettabytes in 2020
- Flood of data
  - New York Stock Exchange generates about 4−5 terabytes of data per day
  - Facebook hosts more than 240 billion photos, growing at 7 petabytes per month
  - Ancestry.com, the genealogy site, stores around 10 petabytes of data
  - The Internet Archive stores around 18.5 petabytes of data
  - The Large Hadron Collider near Geneva, Switzerland, produces about 30 petabytes of data per year
- Machines (IoT) will generate more data than people
  - Machine logs, RFID readers, sensor networks, vehicle GPS traces, retail transactions
- Goal: More important than storing data is how to extract value from it!

Big Data

"Data that is massive in volume, with respect to the processing system, with a variety of structured and unstructured data containing different data patterns to be analyzed."

- Traditional approaches for storing and processing doesn't work due to the huge volume of data
- Data may contain valuable information → needs to be processed in a short span of time
- This valuable information can be used to make predictive analyses, as well as for marketing and many other purposes
- Using the traditional approach, analysis won't complete within the given time frame.



Source: How Does Big Data Make the World Go Round? (IBM) http://www.ibmbigdatahub.com/blog/how-does-big-data-make-world-go-round

What is big?

- 0.43 x 10 18 seconds: The Age of the Universe (13.77 billion years)
- 5 Exabytes: All words ever spoken by human beings (in text) Roy Williams (Caltech, 1993)
- 21 Exabytes/month: Global Internet traffic in 2007 Padmasree Warrior (CISCO, March 2010)
- 160 Exabytes: Digital information created, captures, and replicated world wide in 2007 (International Data Corporation, 2007)
- 42 Zettabytes: All words ever spoken by human beings (if digitized in 6kHz 16 bit audio) Mark Lieberman (U. Penn, 2003)

1 Bit	Binary digit
8 Bits	1 byte
1,024 Bytes	1 KB (kilobyte)
1,024 KB	1 MB (megabyte)
1,024 MB	1 GB (gigabyte)
1,024 GB	1 TB (terabyte)
1,024 TB	1 PB (petabyte)
1,024 PB	1 EB (exabyte)
1,024 EB	1 ZB (zettabyte)
1,024 ZB	1 YB (yottabyte)
1,024 YB	1 brontobyte
1,024 brontobyte	1 geopbyte

Big Data Dimensions

- 3Vs (Gartner interpretation – published in a white paper by Douglas Laney in 2001)
  - Volume: scale of data (incoming and cumulative)
  - Velocity: analysis of streaming data
  - Variety: different forms of data
- 4Vs (IBM definition)
  - Volume, Velocity, Variety
  - + **Veracity: uncertainty of data**
- 6Vs (Microsoft)
  - Volume, velocity, variety, veracity
  - + **Variability: complexity of data set (lack of fixed patterns)**
  - + **Visibility: requires full picture of data to make informative decision**
- 5Vs (Yuri Demchenko, 2014)
  - Volume, velocity, variety, veracity
  - + **Value: importance of the results obtained through storing, processing and analyzing**

Volume

- Past: companies used only data created by their employees
- Now: data generated by devices and customers

- Social media enhances the amount of data generated (videos, photos, tweets, etc)
  - 6 billion (from 7 billion – world population) have cell phones
  - Cell phones have sensors that generate data for each event, which is stored and analyzed (example, health reports)
- The volume in big data is of an amount that cannot be gathered, stored, and processed using traditional approaches

- Examples
  - Facebook
    - 2 billion active users
    - 600 TB of data is ingested into Facebook's database
  - Jet airplane
    - 10TB of data are generated for every hour of flight time (scale of petabytes generated per day for all flights)

Velocity

- Rate at which data is being generated – how fast the data is coming in

- Examples
  - New York stock exchange captures 1TB of data during each trading session
  - 120 hours of video are being upload to YouTube every minute
  - Modern cars have almost 100 sensors to monitor each item from fuel and tire pressure to surrounding obstacles
  - 200 million emails are sent every minute

- Another dimension of velocity is the period of time during which data will make sense and be valuable
  - Will it age and lose value over time, or will it be permanently valuable?

Variety

- Classification of data: **structured** or **unstructured**
- **Structured data**
  - Information with predefined schema or data model (predefined columns, data types, etc)
  - Examples: relational databases
- **Unstructured data**
  - No predefined schema or data model
  - Examples: documents, emails, social media text messages, videos, still images, audio, graphs, output of sensors, devices, RFID tags, machine logs, cell phone GPS signals, etc
- Variety of data – examples of unstructured data generated
  - 400 million Tweets are sent per day
  - 4 billion hours of videos are watched on YouTube every month
- Independently of its "structureness" – data must be processed in order to generate a better user experience or revenue for the companies itself

Veracity

- Uncertainty of data
- May be due to poor data quality or noise in data
- Veracity is important when **analyzing and making decisions**
- In certain situations, there is no time to clean streaming data or high velocity data to eliminate uncertainty
  - Data may lose value if it takes too long to be analyzed (for example, data generated by sensors)
  - Uncertainty must be taken into account

Variability

- **Lack of consistency or fixed patterns in data**
- Different from variety
- If the meaning and understanding of data keeps on changing, it will have a huge impact on your analysis and attempts to identify patterns

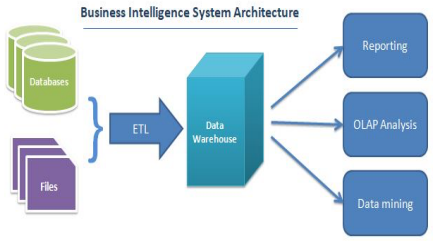
Value

- Most important characteristic!
- Also applies in small data
- It evaluates if it's worth storing the data and investing in infrastructure to do so
- One aspect of value is that it is necessary to store a huge amount of data before it can be used to give valuable information
- There is no gain if the stored data does not return in value to the organization

Traditional approaches to data storage

- **Batch jobs** are scheduled to migrate data into **data warehouses** in a day, week or month period
- Data has a **schema** and is categorized as **structured data**
- It goes through a **processing analysis cycle** to **create datasets** and **extract meaningful information**
- Ingestion in the data warehouse through **ETL (Extract, Transform, and Load)** operations – take raw data and process it for analytical and reporting purposes
- Data warehouse is **optimized for reporting and analytics purposes** only
  - Core of Business Intelligence (BI) systems

Architecture of a traditional BI system



Problems using the traditional approach

- **Latency rate** – reports are not in real time (take days or weeks to be generated)
  - Organizations need to respond quickly
- **Limited sources** – relies on structured data – nowadays most information is not structured (blog posts, web reviews, social media sources, posts, tweets, photos, videos, audio, sensor logs, etc)
  - Currently fewer than 20% of data has a definite schema – 80% is raw data without any specific pattern that can be transformed to be placed in a traditional relational database system
- **Limited scale** – cannot handle an increasing amount of data that is generated fast
  - For example, autonomous cars generate and consume 4TB of data each day for an hour of driving

Cluster Computing

- Set of computers (nodes) connected to each other in such a way that they act as a single server to the end user
- Characteristics
  - **High availability** – data must be available at all times even when hardware or software failures occur
  - **Resource pooling** – multiple computers are connected to each other to act as a single computer. Besides data storage, CPU and memory are also shared and can be utilized in individual computers to process different tasks independently and then merge outputs to produce a result.
  - **Easy scalability** – additional storage capacity or computational power can be easily added with new machines to the group (**scale out** instead of **scale up**)
  - **Parallel processing** – each node executes a task in parallel
  - **Commodity hardware** – its uses COTS hardware with reasonable storage and computation power – much less expensive compared to dedicated processing server with powerful hardware

Cloud versus on-premises infrastructure

- There are several public cloud solutions to implement Big Data – Microsoft, Google and Amazon

	On-premises	Cloud
Cost	Big start up cost. Needs to set up high-end servers, networking, storage.	Avoids the start up costs. Increase on demand and pay per use.
Security	Sense of increased security. Control over who is accessing their data, when it is uses, and for what purpose.	There are no assurances about where is the data, how is it being managed, or who can access it. Providers aim at making sure that every bit of information put in the cloud is safe and secure (e.g. encryption mechanisms, redundancy and geo replication).
Capabilities	Require local personnel to manage the implementation on site – for setup, support.	Also requires staff, but with a focus on the problem.
Scalability	Requires evaluating further the required capacity in order to acquire hardware. Depending on the capacity demand analysis, there can be a high amount of idle resources.	Resources can be added on demand

Public Data Sets

- AWS Public Dataset Program - https://aws.amazon.com/opendata/public-datasets/
- Google Cloud Pubilc Datasets - https://cloud.google.com/public-datasets/
- Registry of Research Data Repositories - https://www.re3data.org/
- Wikipedia downloadable dataset – http://en.wikipedia.org/wiki/Wikipedia:Database_download
- Million Song Dataset - https://labrosa.ee.columbia.edu/millionsong/
- U.S. Government's open data - https://www.data.gov/
- Open Data Brazilian Portal - http://dados.gov.br/dataset
- https://github.com/awesomedata/awesome-public-datasets
- Kaggle Datasets - https://www.kaggle.com/datasets
- DataHub - https://datahub.io/
- NYC Taxi & Limousine Commission Trip Record Data - http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
- YFCC100M - https://multimediacommons.wordpress.com/yfcc100m-core-dataset/

Examples of Applications using Big Data datasets

- Astrometry.net (http://astrometry.net/)
  - Watches the Astrometry group on Flickr for new photos of the night sky
  - Analyzes each image and identifies which part of the sky it is from, as well as any interesting celestial bodies, such as stars or galaxies
- Google Books Ngram Viewer (https://books.google.com/ngrams/)
  - Uses Google Books dataset to displays a graph showing how phrases have occurred in a corpus of books (e.g., "British English", "English Fiction", "French") over selected years.
- OpenFlights.org (https://openflights.org/)
  - Map flights around the world, search and filter in all sorts of ways, calculate statistics automatically, and share flights and trips.

BIG DATA – HANDS-ON