Large Scale Data Management

Ricardo Vilaça

rmvilaca@di.uminho.pt

https://rmpvilaca.github.io/

Goals

- Provide hands-on experience on modern large-scale data analysis systems and database systems
- The project involves defining a problem, understanding the storage and processing needs (short term and long term) of a big data solution and select an appropriate technical approach
- You are expected to define the problem, find the appropriate data, and create an initial proof-of-concept implementation of the solution
- This will equip you with necessary skills and knowledge for applying a state-of the-art data science approach in real-world problems, as they arise in industry or research settings

Guidelines

- This is a team project, where each team should have 2 or 3 members
- You need to identify a problem based on an existing or new data set. There are no constraints on the data as long as all privacy or confidentiality constraints are met
- You need to implement a process that computes the result in a repeatable fashion
- It should be sufficiently easy to kick-off a new end to end execution of the process

Guidelines

- It should also be easy to analyze larger amounts of the data as they become available over time
- The result of the processing should be accessible for review through some kind of serving layer and presented in a form that would make sense in an intended real world scenario
- You can pick any of the technical solutions discussed in the course as long as you can justify why you picked that solution

Definition of the problem

- You should formulate the specific problem and use case for the system/application
- Needs to be a big data problem, involve complexity along some dimensions such as the size of the data (Volume), the quality and variety of the data (Variety), the speed at which data arrives and need to be analyzed (Velocity)
- It is permitted, but not required, to select a problem that requires advanced processing such as data mining or machine learning algorithms

Definition of the problem

- You can make assumptions about the data, number of processes, users, etc. One such assumption could be, for example, that the data is cleaned to a certain degree
- All such assumptions must be explicitly defined, and when appropriate reflected in the solution and the report
- In other words, the focus of the problem is not on accuracy or importance of the findings, but rather on the big data solution provided

Project Deliverables and Due Dates

- Project proposal (~3 pages) 31 March 2019
- Project report (~10 pages) and source code 19 May 2019
- Project In-class presentation (~10 minutes) 24 May 2019

Project Proposal

- The project proposal should build on one or more large-scale open dataset(s) that are good candidates for analysis
- The idea is to survey the data sets freely available and identify what are their strengths and weaknesses for the needs of the analysis
- The proposal should then focus on what are some promising intuitions or questions that can be answered through rigorous analysis of the data
- You should try to provide a concrete proposal for a big data solution that can potentially be used for analysis of the same dataset over time
- Emphasis should be given on how the solution scales

Project Proposal Contents

- Domain Description and Motivation: What is the data domain? What is the goal of your project?
- What is the motivation for the proposal? What are the questions you want to answer? Why the analysis is important? What are a few potential applications?
- Architecture of Proposed Solution: Provide a draft of your overall data analytics architecture.
- Describe the data collection/ingestion process, data storage, data processing, data serving, and (optionally) data visualization.
- Provide a draft figure that depicts the overall architecture and data flow in your system.

Project Proposal Contents

- Describe anticipated limitations or difficulties with your approach.
- System Evaluation and Data Analysis: How will you evaluate your system and architecture?
- What results you plan to obtain? What type of data analysis you will perform? How this type of analysis is adequate for the data, problem and questions posed? What are the steps you need to take to scale your solution?

Project Report and Source Code

- The project report should represent all the completed work. The expectation is that most of the work has been completed and any major results are available
- You should be able to provide a complete description of the project

Project Report Structure

- Abstract: The abstract (limited to 150-200 words) should be a comprehensive but concise description of your project that aims to attract potential readers. It should briefly discuss the motivation, problem of interest, technical approach to solve it and main results of your work.
- Introduction/Motivation: What is the project about? What is the data domain? What is the goal of your project? What is the motivation for rigorous data analysis in this domain? What are the questions you want to answer? Why the analysis is important? What are potential applications?
- Data and Data Analysis: What are the data dimensions and processing dimensions of your solution? What are the pre-processing steps you need to perform? Are there any data cleaning, data transformation steps that are required? What are the data analytics methods you have employed? What type of data analysis you have performed? How this type of analysis is adequate for the data, problem and questions posed?

Project Report Structure

- Architecture of Proposed Solution: Describe your overall data analytics architecture. Describe the data collection/ingestion process, data storage process, data processing, data serving, and (optionally) data visualization. Provide a figure that clearly depicts the overall architecture and data flow in your system. Describe limitations or difficulties with your approach.
- Evaluation/Results: How did you evaluate/test your work? What analysis did you perform? What datasets have been used? Provide summary statistics of your dataset. How your evaluation provides support (or not) of your solution. Show results of your analysis, discuss important findings, discuss implications of your analysis to applications.

Project Report Structure

- Conclusions: What are the conclusions of your work? Is your solution adequate for the problem? Are there any highlights of the analysis? What are some ideas for future work?
- References: The final report should include the full reference of the libraries, papers, code, tutorials that you have based your project on, your approach to solve the problems, and the tools and datasets that you have employed. Full citation is required. References should be specific and found inside the text, as appropriate.

Presentation

- The in-class presentation should be seen as your opportunity to present your hard work in class, your ideas, your approach and solution, your results, and discuss further implications for future work.

Available Big Data Sets

- Generic repositories
  - AWS Public Datasets, https://registry.opendata.aws/
  - Comprehensive Knowledge Archive Network, https://ckan.org/
  - Stanford Large Network Dataset Collection, http://snap.stanford.edu/data/index.html
  - Open Flights, https://openflights.org/
  - ASA Flight data, http://stat-computing.org/dataexpo/2009/the-data.html

Available Big Data Sets

- Geo Data
  - OpenStreetMap, https://planet.openstreetmap.org/
  - Natural Earth Data, http://www.naturalearthdata.com/downloads/
  - GeoNames, http://www.geonames.org/
  - Libre Map Project, http://libremap.org/
- Web Data
  - Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Database_download
  - Public Terabyte Dataset Project, http://www.scaleunlimited.com/datasets/public-terabyte-dataset-project/
  - StackOverflow, https://stackoverflow.blog/tags/cc-wiki-dump/

Available Big Data Sets

- Government Data
  - European Parliament Proceedings, http://www.statmt.org/europarl/
  - Public health data sets, https://phpartners.org/health_stats.html
  - UN Data, http://data.un.org/Explorer.aspx
  - UK Government data, https://data.gov.uk/
  - US Patent and Trademark Office, https://www.google.com/googlebooks/uspto.html

Project Ideas

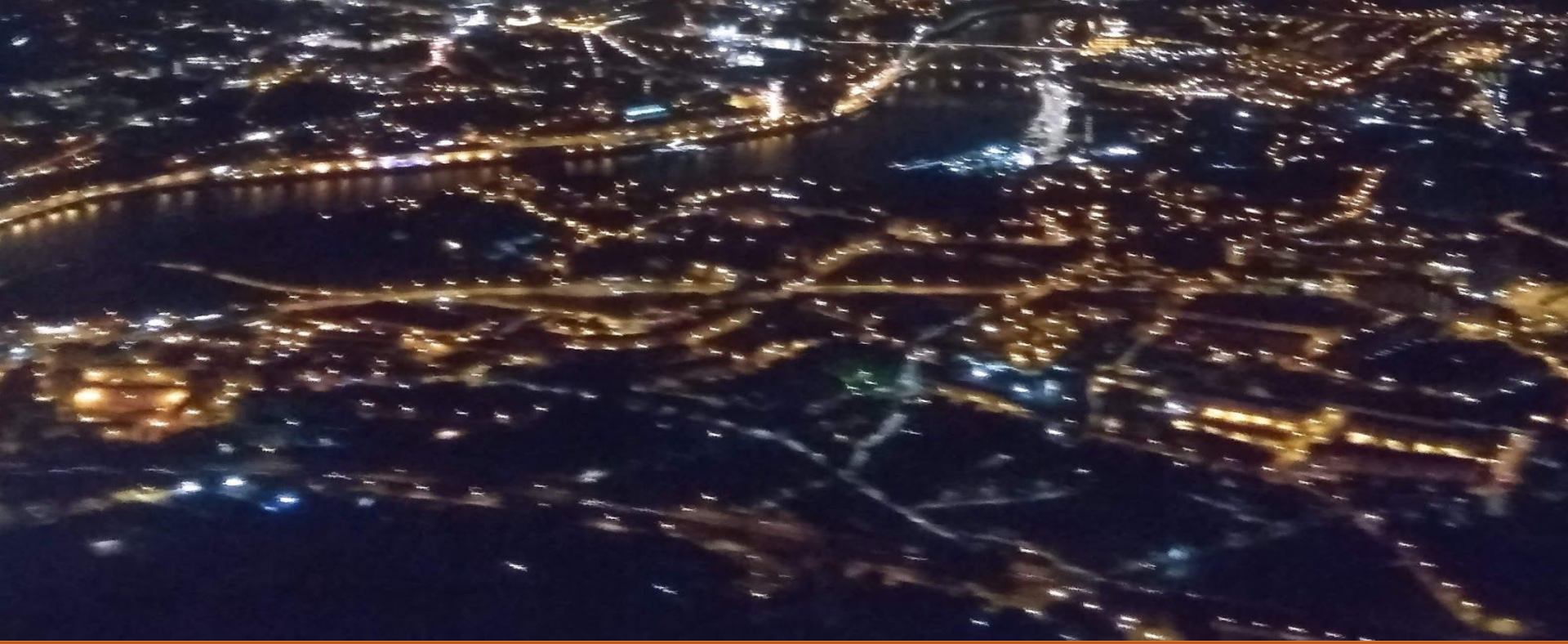
- Build a web interface that allows users to search in big data (e.g., Health records, census data, ... etc.)
- Collect tweets and use them to run some correlation analysis or sentiment analysis, e.g., how do people in different states perceive brands (car brands, food brands, ... etc.)
- Download satellite data and find the correlation between temperature, vegetation, precipitation, and fires. For example, you can compute the average per day/week/month/season/year and show how the averages change over time

Project Ideas

- Collect census data, POI data, lakes, parks, ... etc. and try to rank cities by their quality of life.
- Trend prediction in fashion
- Correlating price/volume of low volume stocks with social media
  - search information related to future price and volume movements
  - find indicators to predict abnormal price or volume changes
- Analysis on the cancer genome (big) data
  - TCGA data set
  - patient-based treatment recommendation

Project Ideas

- Music recommendation system with geospatial information
  - MMTD - Million Musical Tweets Dataset
- How to name your new-born baby?
  - prediction of trends in baby names around the world
- Movie exploration/recommendation system
- Best transport choice
- Fake reviews detection



Large Scale Data Management

rmvilaca@di.uminho.pt