Large Scale Data Management

Ricardo Vilaça

rmvilaca@di.uminho.pt

https://rmpvilaca.github.io/

Presentation

- HASLab Researcher @ UM & INESC TEC
  - https://dbr-haslab.github.io/
- Research Interests:
  - Secure Databases
  - Scalable Transactions
  - Dependable Distributed Systems
  - Large Scale Data Stores
  - Serverless Computing
  - Hands-Free Query Optimizer

Course Information

Course Context

- Data science is intended to extract value in the form of knowledge through the analysis of large datasets
  - Sets which are usually heterogeneous, complex, contain noise and, not infrequently, still incomplete
- Data management has steadily moved away from relational monolithic systems
  - Hybrid solutions are increasingly being adopted, combining different storage components, including unstructured files, documents and graphs, and different interrogation and processing systems
  - Graphical representations for summarising the most important characteristics of vast amounts of data are increasingly desirable
  - Data security and privacy guarantees are increasingly important from the point of view of ethical manipulation, integrity and legal compliance

The purpose of this lecture is to provide the fundamental concepts and computational paradigms of large-scale data management and Big Data, including methods for storing, updating, querying, and analyzing large datasets including user interfaces with interactive results visualization, with data security and privacy guarantees.

Course Organization

- Lecture
  - Fundamental concepts and computational paradigms of large-scale data management and Big Data
  - Methods for storing, updating, querying, and analyzing large datasets
  - Foundations to build distributed systems using data centric programming and large-scale data processing
- Practical assignment
  - Provide hands-on experience on modern large-scale data analysis systems and database systems
- Final exam
- Evaluation
  - 60 % Final Exam + 40% Practical assignment
  - Minimum of 8 at each

Course Prerequisites

- Java Programming
- Python Programming
- Databases
- Data Structures and Algorithm
- Distributed Systems

Tentative Schedule

Date	Theory	Practice	Lecturer
8/2/2019	Introduction	X	RMV
15/2/2019	Cloud Computing	Docker/Postgres	RMV
22/2/2019	Challenges and Foundations	GCP	RMV
1/3/2019	Scalable Storage	Hadoop/HDFS	JTP
8/3/2019	Traditional DBs	Practical Assignment	RMV
15/3/2019	NoSQL	HBase	RMV
22/3/2019	A new life to SQL	BigQuery	RMV
29/3/2019	Cloud Data Management Systems	Cloud Firestore	RMV
5/4/2019	Distributed Computation	Cloud Dataproc/Spark	RMV
12/4/2019	Streaming	Flink	RMV
26/4/2019	Big data machine learning	TensorFlow	RMV
3/5/2019	Big Data visualization	DataLab and Colaboratory	RMV
10/5/2019	Security and Privacy	Safe HBase	RMV
24/5/2019	Practical Assignment Presentation		RMV

Course Contents

- Introduction
  - Big Data, Data Economy
  - Supercomputing vs Cluster Computing vs Cloud Computing
  - Containers/Docker
- Cloud Computing
  - Deployment environments: data centres, internet-wide systems
  - Challenges: scalability, high availability, performance (throughput, tail latency), consistency
  - The OS for the Data-centre and Cloud Computing (IaaS, PaaS, SaaS, FaaS)

Course Contents

- Scalable Services and Programming methods
  - Design Reliable, Scalable Services and Applications
  - Distributed systems architectures
  - Programming methods
- Distributed storage
  - From local to distributed to programmable software-defined storage
  - Storage optimisations for handling large scale data
  - Hadoop HDFS case study

Course Contents

- Traditional DBs
  - Storage Data Structures
  - Query optimization
  - Introduction to transaction processing: purpose, anomalies, serializability, snapshot isolation, concurrency
  - Commits and consensus
  - CAP Theorem/Difficulty of scaling while maintaining ACID properties
  - One size does not fit all
    - Tradeoffs
- A new life to SQL
  - NewSQL Databases
  - SQL on Big Data
  - SQL-in-the-cloud systems

Course Contents

- NoSQL
  - Eventual consistency
  - Schema free
  - Graph Databases, Key-value Stores, Document Databases, Column Stores and TimeSeries
  - Scalable Transactions
- Cloud Data Management systems
  - NoSQL
  - SQL
  - Serverless

Course Contents

- Distributed Computation
  - MapReduce
  - Spark
  - Dataflow processing models
  - Graph processing models
- Online / Streaming / Real-time analytics
  - Data stream processing
  - Incremental and online query processing
  - Lambda Architecture
  - Google DataFlow

Course Contents

- Big data machine learning systems
  - TensorFlow/Keras/PyTorch
  - Hardware/scalability challenges
  - Optimizations to reduce computational needs
- Big Data Visualization
  - Jupyter Notebooks
  - Google DataLab and Colaboratory

Course Contents

- Security and privacy
  - Ethical and Legal issues
  - Data protection concepts: access control, encryption, compartmentalization
  - Data anonymization and de-anonymization techniques
  - Differential privacy
  - Cryptographic tools for data security and privacy
  - Secure Processing databases
  - Privacy preserving deep learning

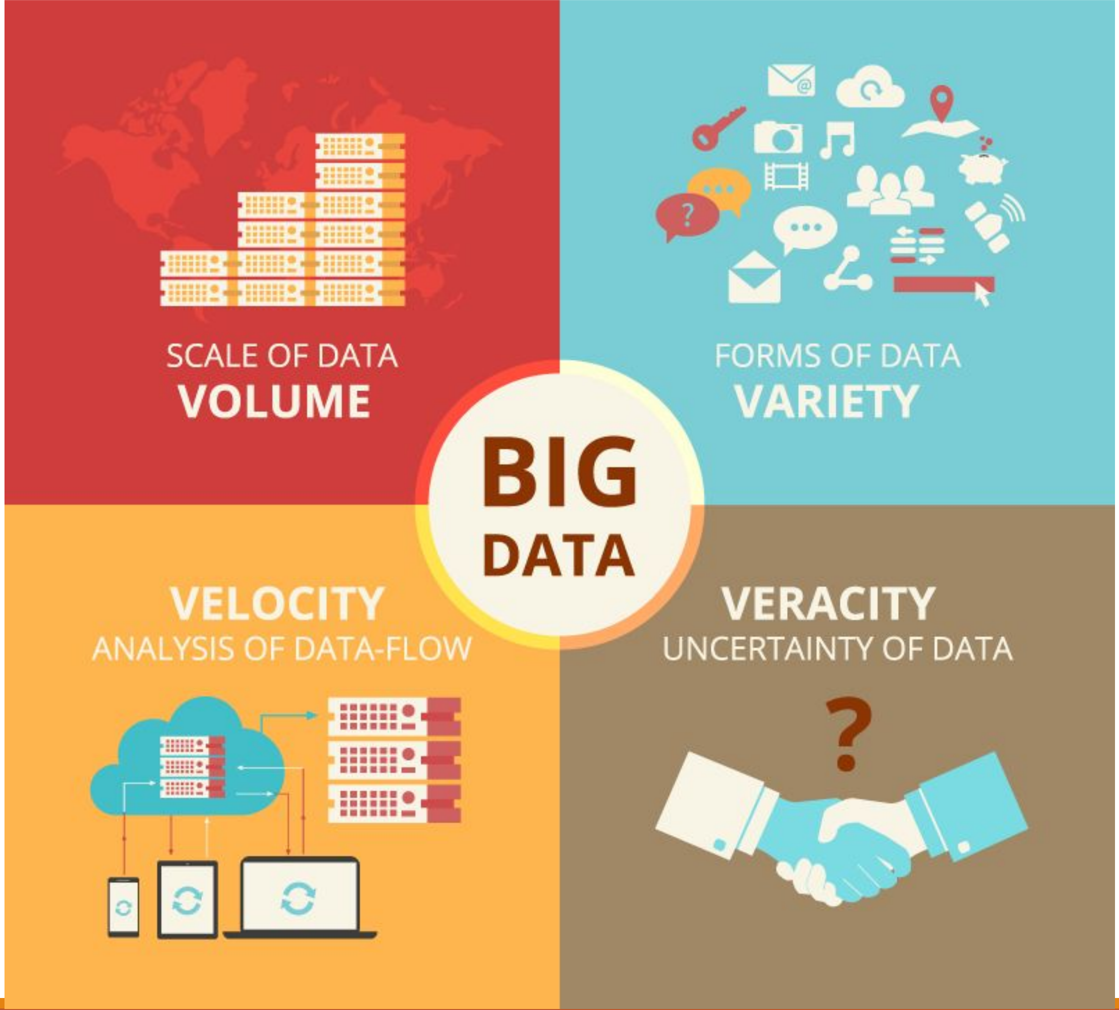
Materials and sources

- **M. Kleppmann, Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, OReilly, 1st Edition, 2017**
- L. Wiese, Advanced Data Management: For SQL, NoSQL, Cloud and Distributed Databases, De Gruyter, 2015
- Joseph M. Hellerstein and Michael Stonebraker. Readings in Database Systems: Fourth Edition. The MIT Press, 2005
- T. Öszu, P. Valduriez, Principles of Distributed Database Systems, Springer, 3rd ed., 2011
- T. White, Hadoop – The Definitive Guide, O'Reilly, 4th ed., 2015
- L. Barroso, J. Clidaras, U. Hölzle, The Datacenter as a Computer - An Introduction to the Design of Warehouse-Scale Machines, 2013

Big Data

Big Data

- Big Data is a relative term
  - If things are breaking, you have Big Data
  - Big Data is not always Petabytes in size
- Big Data is often hard to understand
  - A model explaining it might be as complicated as the data itself
- The game may be the same, but the rules are completely different
  - What used to work needs to be reinvented in a different context


SCALE OF DATA
VOLUME
FORMS OF DATA
VARIETY
BIG
DATA
VELOCITY
ANALYSIS OF DATA-FLOW
VERACITY
UNCERTAINTY OF DATA
?

Big Data 4Vs

- Volume
  - Data larger than a single machine (CPU,RAM,disk)
  - Infrastructures and techniques that scale by using more machines
- Velocity
  - Endless stream of new events
  - No time for heavy indexing (new data keeps arriving always)
  - Led to development of data stream technologies
- Variety
  - Different data formats, data semantics and data structures types
- Veracity
  - Uncertainty of data
  - Untrusted
  - Uncleansed

Scalability

- Popular solution for massive data processing
  - scale and build distribution, combine theoretically unlimited number of machines in single distributed storage
- Scale-up: add resources to single node (many cores) in system (e.g. HPC)
- Scale-out: add more nodes to system (e.g. Amazon EC2)

Supercomputing

- Focus on performance (biggest, fastest). At any cost!
- Programming effort seems less relevant
- Fortran + MPI: months do develop and debug programs
- GPU, i.e. computing with graphics cards
- FPGA, i.e. casting computation in hardware circuits
- Assumes high-quality stable hardware

Cluster Computing

- Use a network of many computers to create a 'supercomputer'
- Oriented towards business applications
- Use cheap servers (or even desktops), unreliable hardware
  - software must make the unreliable parts reliable
- Solving large tasks with more than one machine
  - Parallel database systems (e.g. Teradata, Vertica)
  - NoSQL systems
  - Hadoop / MapReduce / Spark

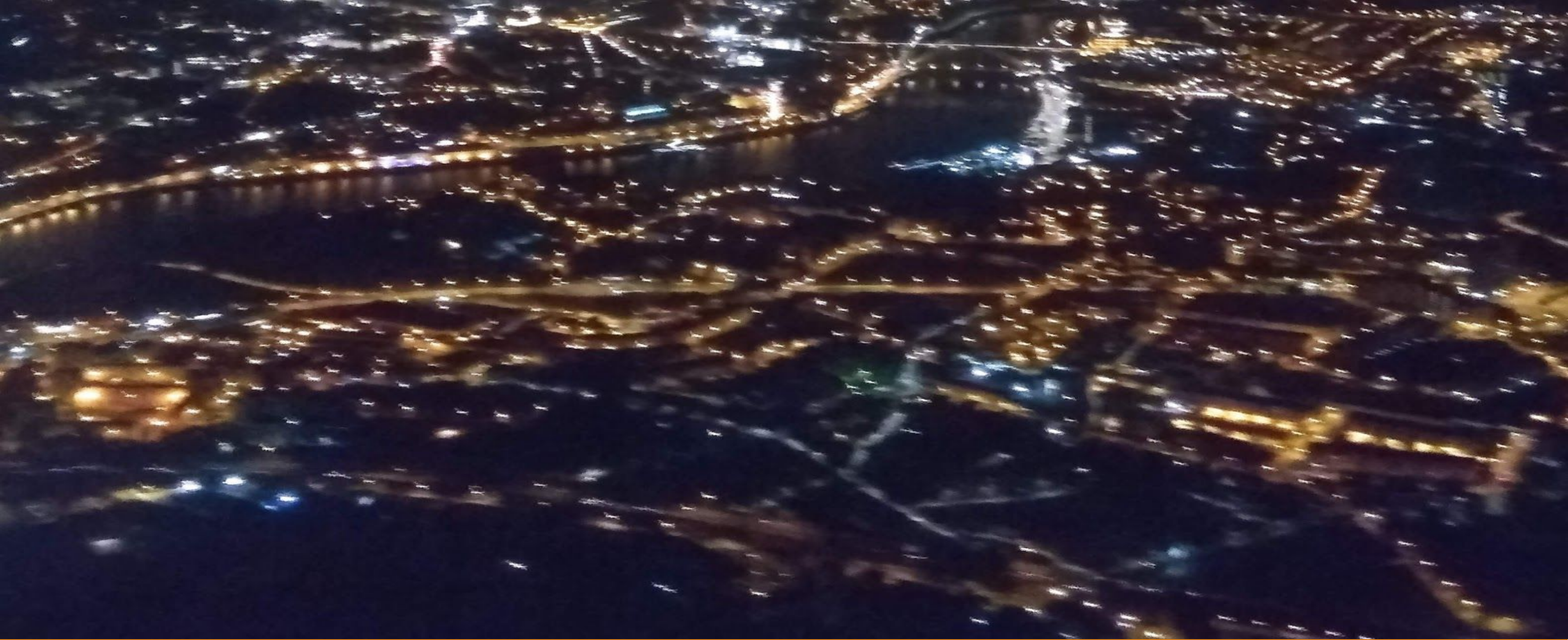
Cloud Computing

- Machines operated by a third party in large data centers
  - sysadmin, electricity, backup, maintenance externalized
- Rent access by the hour
  - Renting machines (Linux boxes): Infrastructure as a Service
  - Renting systems (Redshift SQL): Platform-as-a-service
  - Renting an software solution (Salesforce): Software-as-a-service
- {Cloud,Cluster} are independent concepts, but they are often combined!
  - Hadoop on Google Cloud Platform

GCP

Google Cloud Platform

- https://cloud.google.com
  - Starting next week
  - Each account has 50$
  - Please send me email to rmvilaca@di.uminho.pt with:
    - First name
    - Last Name
    - School Email address from the @alunos.uminho.pt domain



Large Scale Data Management

rmvilaca@di.uminho.pt