CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Big Data Processing

L01: Module Introduction

Dr. Ignacio Castineiras
Department of Computer Science

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Presentation

- Ignacio Castiñeiras.
    - Lecturer at the Department of Computer Science.
    - Email: Ignacio.Castineiras@cit.ie
    - Office Room: C131
    - Telephone: +353 21 433 5857
- Qualification:
    - PhD. in Computer Science: 2014.
    - MEd. in Computer Science: 2011.
    - MSc. in Computer Science: 2009.
    - BSc. in Computer Science: 2007.

Presentation

[2018 -] Cork Institute of Technology
Lecturer at Dept. Computer Science

- Research Group Ríomh

[2015 - 2018] Cork Institute of Technology
Assistant Lecturer at Dept. Computer Science

- Research Group Ríomh

[2014 - 2015] University College Cork
Postdoc at Insight Centre for Data Analytics

- EU FP7 Project GENiC

[2007 - 2014] Complutense University of Madrid
PhD. & MSc. at Declarative Programming Group

- Spanish National Projects FAST & MERIT

Background: Optimisation and decision analytics.
Application of **Constraint Programming** to real-life Constraint Satisfaction and Optimisation Problems.

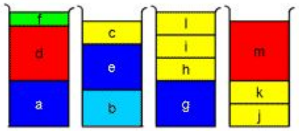
Presentation

Research Background

- Constraint satisfaction and optimisation problems:
  - Examples: Manufacturing & service industries:
  - Feasible/optimal allocation/scheduling of company resources.
  - Challenge: Combinatorial nature.
- Constraint Programming:
  - Subfield of Artificial Intelligence.
  - High-level declarative problem formulation.
  - Problem solving: Inference process + search on top of it.

Presentation

PhD Research Experience



- Tackle real-life problems with Constraint Programming
  - Employee Timetabling Problem.
  - Bin Packing Problem.
- Comparison among multiple paradigms and solvers.
  - Algebraic - Object Oriented - (Functional) Logic Programming
  - C++, Python, SICStus Prolog, Haskell, TOY, etc.
- Implementation of constraint solvers:
  - Adapt object-oriented solver library to a logic programming environment.
  - Extend solver with high-level user defined search strategy specification.

Presentation

Postdoc Research Experience

- GENiC: Globally Optimised Energy Efficient Data Centres.
  - European Union FP7 Programme: http://projectgenic.eu/
  - Green computing.
  - Sustainable DCs.
  - Renewable energy sources.

CIT CORK INSTITUTE OF TECHNOLOGY INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh

TU/e Technische Universiteit Eindhoven University of Technology

IBM

United Technologies Research Center

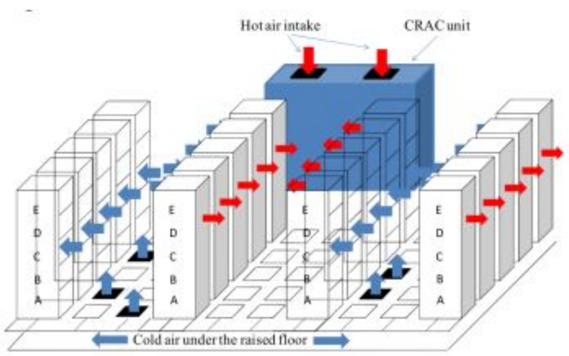
AtoS

acciona Infraestructuras

Presentation

Postdoc Research Experience

- Develop scalable decision support tools:
  Optimise workload allocation of single and distributed DCs
  - Single DC: Reduce power consumption.
  - Geographically Dcs: Reduce overall energy consumption.

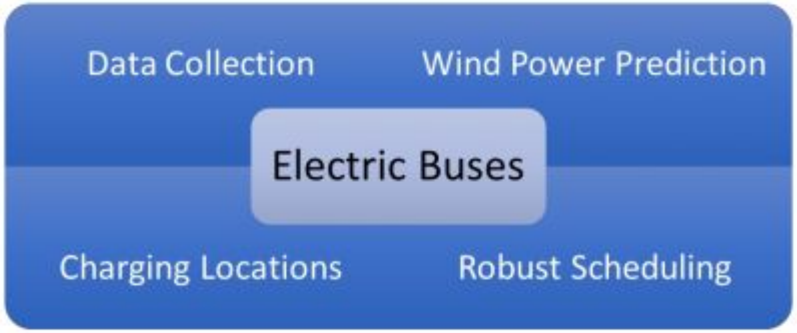


Greenland
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean

Presentation

Lecturer Research Experience

- SMART Electric Buses.
  - Ireland SEAI Programme: https://smartebuses.github.io/web/index.html
  - Green computing.
  - Sustainable Transportation.
  - Renewable energy sources.



UCC
University College Cork, Irelan
Coláiste na hOllscoile Corcaig

CIT CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Presentation

Ríomh: Intelligent Secure Systems Group.

Research Areas:

- Future Networks & Internet of Things
- Virtualisation Technologies
  - Cloud Computing
  - Network and Information Security
- Data Analytics
  - Machine Learning
  - Optimisation Techniques

Contact us: Donna.Oshea@cit.ie (Head)

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Learning Outcomes

Module Descriptor:
https://courses.cit.ie/index.cfm/page/module/moduleId/13442

- LO1: Appraise how the velocity, volume and variety of data will impact how data is stored, managed and analysed.
- LO2: Survey the different tools that constitute a big data framework.
- LO3: Process large-scale temporal, geospatial, text and graph datasets using descriptive and analytical tools.
- LO4: Design and develop a real-time streaming algorithm for performing large scale distributed computation.

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Syllabus Week Plan

Week 1: September 21st – September 27th.

Lectures

- L01. Module Introduction.
- L02. Big Data Motivation.
- L03-04. Distributed Programming.

Lab

- Lab01: Sequential-solving Programming Exercise: Minesweeper.

Big Data Mindset.

- *Introductory example of a Big-Data driven society.*

Syllabus Week Plan

Week 2: September 28th – October 4th.

Lectures

- L05. Distributed Programming.
- L06. Spark Core Model of Parallel Computing: RDDs.

Lab

- Lab02. Distributed Programming (Lab Demonstration).

Big Data Mindset.

- *The thinking or mental shift big data requires: Sampling => All data.*

Syllabus Week Plan

Week 3: October 5th – October 11th.

Lectures

- L07-08. Spark Core Model of Parallel Computing: RDDs.

Lab

- Lab03. Databricks: A Tutorial.

Big Data Mindset.

- *The thinking or mental shift big data requires: Causation (Why?) => Correlations (What?)*

Syllabus Week Plan

Week 4: October 12th – October 18th.

Lectures

- L09-10. Spark Core Model of Parallel Computing: RDDs.

Lab

- Lab04. Spark Core - Introductory Exercises.

Big Data Mindset.

- *Datification (or the art of extracting data from the most surprising places).*

Syllabus Week Plan

Week 5: October 19th – October 25th.

Lectures

- L11. Spark Core Model of Parallel Computing: RDDs.
- L12. Spark SQL.

Lab

- Lab05. Spark Core - Advanced Exercises.

Big Data Mindset.

- *Data Reuse: Data's multiple lives.*

Syllabus Week Plan

Week 6: November 2nd – November 8th.

Lectures

- L13-14. Spark SQL.

Lab

- Lab06. Spark SQL - Introductory Exercises.

Big Data Mindset.

- *Data regulations: Data ownership and its accountability.*

Syllabus Week Plan

Week 7: November 9th – November 15th.

Lectures

- L15-16. Spark SQL.

Lab

- Lab07. Spark SQL - Advanced Exercises.

Big Data Mindset.

- *The dark side of big data: I know who you are. I guess what would you do.*

Syllabus Week Plan

Week 8: November 16th – November 22nd.

Lectures

- L17-18. Spark Streaming.

Lab

- Lab08. Spark Streaming - Introductory Exercises.

Big Data Mindset.

- *Big data industry revolution: Education as a use-case: Get to know students better.*

Syllabus Week Plan

Week 9: November 23rd – November 29th.

Lectures

- L19. Spark Streaming.
- L20. Spark Structured Streaming.

Lab

- Lab09. Spark Streaming - Advanced Exercises.

Big Data Mindset.

- *Big data industry revolution: Education as a use-case: Adaptative learning.*

Syllabus Week Plan

Week 10: November 30th – December 6th.

Lectures

- L21. Spark Structured Streaming.
- L22. Anatomy of the Execution of a Spark Core Program.

Lab

- Lab10. Spark Structured Streaming - Advanced Exercises.

Big Data Mindset.

- *Big data industry revolution: Education as a use-case: The dark side again.*

Syllabus Week Plan

Week 11: **December 7th – December 13th.**

Lectures

- L23. Anatomy of the Execution of a Spark Core Program.
- L24. Big Data Storage.

Lab

- Lab11. Distributed Solving of Minesweeper.

Big Data Mindset.

- *Big data: What do you think?*

Syllabus Week Plan

Week 12: December 14th – December 20th.

Lectures

- L25. Big Data Storage.
- L26. Module Wrap-Up: 24 Ideas for 24 Lectures.

Lab

- Lab12. Big Data Storage (Lab Demonstration).

Big Data Mindset.

- *Big data: What do you think?*

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Methodology

- 2h lecture (once per week):
  - Concepts explanation.
  - Application via code examples.
  - Put together to extract conclusions.

- 2h lab session (once per week):
  - Reinforce the concepts seen in the lectures.
  - Weekly exercises to practice: Attempt a bunch of exercises for which the solution is provided.

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Evaluation

Module Descriptor:
https://courses.cit.ie/index.cfm/page/module/moduleId/13442

Assignment 1:

- Use Spark Core and Spark API to perform descriptive analytics of an real-world open source dataset.
- Compare and contrast the efficiency and expressiveness of both approaches.
- Write a report of up to 1,000 words with a novel data analysis exercise proposed by yourself.

Marks: 50
Deadline: Week 8, Sunday 22nd of November

Evaluation

Module Descriptor:
https://courses.cit.ie/index.cfm/page/module/moduleId/13442

Assignment 2:

- Use Spark Streaming and Spark Structured Streaming to perform offline and online analytics of an real-world open source dataset.
- Compare and contrast the efficiency and expressiveness of both approaches.
- Write a report of up to 1,000 words with a novel dataset available in the internet and compare it to our existing one.
- Write a report of up to 1,000 words with a use-case for the Spark libraries on Graphs or on Machine Learning.

Marks: 50
Deadline: Week 10, Sunday 6th of December

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Motivation

Why is important to study this module?

For any module I teach I usually take some minutes during lecture 1 to justify/motivate why to study the module.

In this case, the motivation has grown so much that it has become part of the indicative content: Big Data Mindset.

But, in a single point: Why to study big data?
Because it is transforming our society.

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Thank you for your attention!