Summary

Syllabus

Lecture Topics (Science)	Lab Topics (Technology)	Remarks
Course Outline; Intro. to Big Data	R Tutorial	
Data Computing Cycles; R Examples	Lab1 - R Lab Submission	
Data Computing - Text Data	Python Tutorial	
Data Computing - Time Series Data	Lab2 - Python Lab Submission	Project Grouping Deadline
Midterm	Hadoop Tutorial	
Parallel Computing - Theory	Lab3 - Hadoop Lab Submission	
Parallel Computing - Hadoop	Pig Latin Tutorial	
Parallel Computing - Spark	Lab4 - Pig Latin Lab Submission	
Parallel Computing - Others	Spark Tutorial	
Summary	Lab5 - Spark Lab Submission	
(Project Time)	(Project Time)	
(Project Time)	(Project Time)	
(Project Time)	(Project Time)	Project Presentation File Deadline Project Report Deadline

(Please click "Files" on the left menu for further details.)

	Percentage
Project	40%
In-class	5%
Midterm	**15%**
Exam	40%

Real Job Ads

- From JobsDB (HK),

- Bachelor in Computer Science, Engineering or Big Data or relevant disciplines
- At least 2-year experience with Data Engineering or Big Data Technologies, or Data Transformation, and modeling
- Experience in architecting and building scalable data platforms
- Experience with Cloud Technologies (Data Lake, Azure, Google, AWS etc.) or experience with open source technologies (Spark, Kafka, Presto, Hive, Cassandra etc.)
- Experience with SQL and/or NOSQL databases
- Experience with 2 of 3 - Java, Scala, and Python programming languages
- Machine learning experience with Spark or similar
- Self-motivated and team player, able to work under dynamic work environment and flexible to changes

- Degree in Computer Science, Information System, Actuarial or related discipline.
- At least 5 years of working experience in data warehouse, ETL, BI, Big data areas and broad exposure to all its sub-disciplines.

- Strong in SQL, Python/Scala.
- Experience working with large, complex and multiple data sets from various sources.
- Expert in data architecture, data modelling and design, data pipeline and data integration.
- Hands-on experience in using big data components such as Hive, Spark, Presto, Python and Airflow.
- Experience in AWS (EC2, S3, Kinesis/Kafka, Athena, Redshift), Google Cloud or other Public cloud environments is a must.
- Experience leading a small team of data engineers is a plus.
- Working experience in the digital, e-commerce industry, mobile app and web environment is highly desirable.
- Candidates with less experience will be considered as Data Engineer.

Real Job Ads

- From Linkedin (HK),

What You'll Need To Succeed

- Important! Experience in dynamic pricing optimiza
  and others)
- 4+ years of experience in Data Science
- Coding proficiency using Python and/or Scala
- Knowledge of applied statistics and optimization t
- SQL working proficiency
- Good communication and interpersonal skills that
  multi-cultural environment

It's Great if You Have

- Working proficiency with Big Data stack: Spark, Hadoop

We Are Looking For Someone With

- Rich experience on dashboard development and data modelling using Microsoft PowerBI
- Understanding of relational and warehousing database technology working with at least one of the major databases platforms (e.g., Oracle, SQLServer, or Postgres)
- Knowledge of big data processing frameworks and techniques such as HDFS, MapReduce, Stream processing, etc. will be an advantage
- Practical working knowledge of data processing tools using SQL, Spark, Nifi, etc.
- Experience with integrating to back-end/legacy environments
- Knowledge of Oracle Business Intelligence (OBIEE), Hyperion Interactive Reporting (BRIO) and Oracle Data Integrator (ODI) is desired.
- Collaborative attitude, willingness to work with team members; able to coach, participate in code reviews, share skills and methods
- Good verbal and written communication; effectively articulates technical vision, possibilities, and outcomes
- Provides insight on business problems through advanced data analysis and visualization

Textbook Titles

- Data Science and Big Data Analytics
- Hadoop: The Definitive Guide
- Learning Spark Lightning-Fast Big Data Analysis
- MapReduce: a flexible data processing tool
- The Hadoop Distributed File System
- Beginning Apache Pig: Big Data Processing Made Easy

Course Materials

- All the course related content, communication, and grading have been posted on CANVAS
- https://canvas.cityu.edu.hk

To be Successful at Data Science, Think Batman, Not Superman

Apr 23, 2018 | 9000 Views

Kimberly Cook

...

FULL BIO

analytics

Data science

IBM

I recently made a Batman analogy when discussing the topic of data science with some colleagues. In this post, I will explore this analogy further.

- http://houseofbots.com/news-detail/2775-4-to-be-successful-at-data-science-think-batman-not-superman

Final Exam (40%)

- 30% of the final exam mark must be obtained to pass the course. (i.e. 30/100)
- Based on the lecture notes and tutorial / lab materials.
- Announced by the university administration.

- Objectives:
  - To assess the capability of students to
    - Identify data computing problems
    - Review the existing concepts in data computing
    - Review the existing technology in data computing
    - Develop data computing solutions
    - Accelerate data computing solutions by parallel computing
    - Apply data computing solutions with specific case studies

Open Project (40%)

- To be consistent with the CityU discovery-enriched curriculum, each group has to identify an interesting problem and propose a data computing solution to solve the problem with parallel computing elements.
- A project cover sheet template and project report template have been provided for you on CANVAS.
- Deliverables:
  - Project Cover Sheet
  - Project Report
  - Supporting Materials
- Please submit your project deliverables on CANVAS https://canvas.cityu.edu.hk
- Late submissions are not graded and will be given 0 mark.

Open Project (40%)

Report	
Real World Impact / Creativity	/ 5
Solid Works and Output Amount	/ 20
Technical Depth and Correctness	/ 20
Parallel Computing Elements	/ 20
Use of Written English	/ 5
Presentation	
Technical Presentation Amount	/ 20
Technical Presentation Skills	/ 5
Question and Answer (Q&A)	/ 5
	/ 100

Open Project (40%)

Project Example: (More past projects in CANVAS)

"Big Data Computing Solutions to Hong Kong Real Estate Data"

1. Collect the Hong Kong real estate data from several sources.
- Document the source of the data clearly in the report (e.g. https://data.gov.hk/en/).

2. Preprocess and Visualize the data with histograms, scatterplots, and other diagrams you have learned;
- Preprocess the data so that you can visualize it.
- Implement data visualizations so that we know better about the data.

3. Analyze the data and discuss your own findings
- Perform advanced analysis on the data (e.g. data clustering and association rule mining)
- Explain the findings, and try to make conjectures about the findings you obtained.

4. Discuss how parallel computing is applied to accelerate the data computing process
- Describe what kind of parallel computing strategy you have implemented (e.g. parallel for loop)
- Explain why such a parallel computing strategy has been adopted (e.g. memory hierarchy)

5. Conclusion and Future work.
- State your conclusions and the related pros / cons.
- If you have enough time, what you can do? What problems are there to be investigated further?

Open Project (40%)

- Possible Data Sources: (but not limited to)
- (You are encouraged to find your own datasets you are interested in; below are just examples that you can choose.)

- Hong Kong Government Data: https://data.gov.hk/en/
- US Government Data: https://www.data.gov/
- Singapore Government Data: https://data.gov.sg/
- UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/
- Panama Papers Graph Data (i.e. Network): https://github.com/amaboura/panama-papers-dataset-2016
- Stanford Large Network Dataset Collection: https://snap.stanford.edu/data/
- Offshore Leaks Database (i.e. Text Data): https://offshoreleaks.icij.org/
-
- Miscellaneous:
- http://www.kdnuggets.com/2011/02/free-public-datasets.html
- https://r-dir.com/reference/datasets.html
- https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- http://www.datasciencecentral.com/page/search?q=data+sets

Open Project (40%)

- Possible Project Ideas: (but not limited to)
  - Analyze factors relating the gaming performance in League of Legends
  - Exploration of Factors Relating to Movie Box Office Performance
  - Historical Buildings in Hong Kong
  - FIFA players' statistics and Professional Football Clubs' Seasonal Performance
  - A visual exploration of aircraft crashes since 1908
  - NBA in Data: An analytical report on Los Angeles Lakers
  - Hong Kong Housing Trend
  - Gastronomy and Ingredients Matching Across the World
  - Exploring of factors relating to League of Legend world championship performance
  - The frequency of earthquakes
  - Homeless, Hong Kong
  - The Relationship among Gender, Education and Employment in Hong Kong
  - Renewable energy in the European Union
  - Flight Networking and On-time Performance Analysis
  - Analysis of Factors Affecting Global Temperature Rise

Open Project (40%)

- Possible Project Ideas: (but not limited to)
  - Secondary School in Hong Kong
  - World University Rankings and Statistics
  - Exploring currency exchange rate
  - Mass Shooting in America
  - An evaluation of workplace environment in Hong Kong
  - Shootings in NBA
  - Exploration of typhoon in Hong Kong in 21st century
  - IMDB Movie Analysis
  - Data mining in conditions and predictions of G20 countries by continent
  - The Analysis of Mandatory Provident Fund (MPF) Schemes
  - Understanding people's reactions to new movies cia Twitter and film review websites
  - Mobile Application (ios and android system) Ranking and the relevant factors on America market
  - Unemployment rate and major indices of US, Germany and Japan
  - Analysis on the 2016 Legislative Council Election

Q&A

Any question ?