ST446 Distributed Computing for Big Data

Week 1, LT 2018

Instructor: Milan Vojnovic

Teaching assistant: Christine Yuen

Your team

- Instructor: Milan Vojnovic (mailto:m.vojnovic@lse.ac.uk)
- Teaching instructor: Christine Yuen (mailto:l.t.yuen@lse.ac.uk)

Plan today

- Administration and logistics (this slide deck)
- Introduction to big data analytics
- Lab preview

Why take this course?

- It will provide you with knowledge about principles and how to use modern computing systems for processing big data
- You will learn
  - basic principles, system architectures, and APIs for storing and processing big data
  - how to conduct batch data processing at scale
  - how to query data at scale using structured query languages
  - how to conduct graph data analytics
  - how to conduct stream data processing
  - how to solve large-scale machine learning tasks

Course outline

Week	Topic	Week	Topic
1	Introduction	7	Stream data processing
2	Distributed file systems and key value stores	8	Scalable machine learning I
3	Computation models	9	Scalable machine learning II
4	Structured data management	10	AI and deep learning tasks
5	Graph data processing	11	Numerical computations using data flows
6	*Reading week*		

Guest lecturers:

- Week 7: Eno Thereska, *Principal Scientist*, Amazon
- Week 9: Ryota Tomioka, *Researcher*, Microsoft Research
- Week 10: Marc Cohen, *Software Engineer*, Google

Prerequisites, software and services

- Basic knowledge of Python
    - Mirrors similar tool usage and learning in ST445
- Software and services:
    - Python and jupyter notebooks
    - Hadoop
    - Bigtable
    - Hive
    - SQL like APIs (e.g. BigQuery)
    - Spark (PySpark interface)
    - Google Cloud Platform (GCP) (**ACK: our sponsor**)
    - Github to share course documents and assignments

Readings

- Mixed set of readings, specific to each week
  - Sources include books, white papers and research papers that describe the design of various systems for big data analytics
  - Available from LSE library or purchase from Amazon
- Often linked to Internet sources
  - References to research and white papers are provided

Course meetings

- Ten two-hour lectures: Monday 10:00–12:00 in TW2.2.04
- Ten 1.5-hour classes (computer labs): Thursdays 12:30–14:00 in TW2.4.01
- No lecture/class in Week 6
- Office hours
  - Milan: Tuesdays 14:00-15:00
  - Christine: Mondays 13:00-14:00 (from week 2, computer lab related questions only)

Assessment

- Weekly assignments (20%)
  - 2 problem sets will be accessed (each 10%)
  - Other problem sets will be assigned but not assessed
- Project (80%)
  - Work on conducting a big data analytics task
  - We encourage you to *work on your projects throughout the course*

Project topics will be provided

Examples:

- Graph queries on Yago knowledge base / MusicBrainz
- Queries on Yelp data reviews
- Movie recommendations using Netflix / MovieLens datasets
- Using Spark on Azure
- ML using Microsoft Cognitive Toolkit
- Querying using Microsoft Cosmos DB
- Amazon graph database Neptun
- Querying graphs using Neo4J database
- Optimizing Hive queries
- Querying key-value pairs using Apache Cassandra
- Stream processing using Amazon Kinesis
- Approximate query answering in Spark SQL
- Working with Apache Kafka
- Deep neural network training using Tensorflow
- Create activity charts for big data technologies (e.g. Spark, Hive, Cassandra) over time, using Stack Overflow / github archive data on BigQuery
- Financial times dataset basic text analysis using Spark / stock price prediction

Collaborations

- All assignments are individual unless we instruct you otherwise
- For individual assignments:
  - You can discuss solutions with peers
  - You are not allowed to copy-and-paste the code
  - You need to write the code yourself
- You can use online resources but always give credit in comments if you borrow code or solution

Remarks on software tools

- We advise you that you use your own laptops (and GCP when needed)
- We encourage you to install various software locally
  - This will allow you to gain hands-on experience and learn about various system components and configuration options
- You may use software installed locally in your system or run it within a docker container or on GCP
  - We will provide you with instructions how to do this
- We will use GCP in the course
  - You need to send us your gmail account, so we can add you as a user in our GCP account
- We will use GitHub Classroom for the course
  - You need to send us your GitHub username, so we can sign you up for the lse-st446 organization