REVISION LECTURE

BIG DATA PROCESSING

Félix Cuadrado

felix.cuadrado@qmul.ac.uk

Queen Mary University of London

School of Electronic Engineering and Computer Science

Exam format

- 4 questions (25 marks each)
  - Multiple subquestions
- Practical assignments
  - Write pseudocode (MapReduce)
  - Interpret code (MapReduce/Spark)
- Short essay questions
  - Parallel Computing Performance / Reliability
  - MapReduce/YARN/HDFS
  - Spark / Stream Processing / Graph processing

Practical Questions (MapReduce)

- Write pseudocode to solve problem on given dataset
    - Numerical summarization, top k, filters
    - Joins
    - Iterative jobs
    - Combiners
- Interpret pseudocode
- Analyse/discusss behavior/performance

Practical Questions (Spark)

- Interpret pseudocode
  - Documentation on the semantics of transformations/actions will be provided as part of the question
- Write pseudocode to modify existing code
  - Based on given documentation
- Compare with equivalent MapReduce code
  - Performance, e.g. number of jobs, applicability of unique Spark features

Hadoop Map/Reduce

- Map/Reduce programming model
- Apache Hadoop architecture
  - HDFS, YARN, Anatomy of a job
- Map/Reduce performance – optimizations
- Performance
  - Speedup, Amdahl's Law
- Reliability, fault tolerance

Big Data Landscape

- In-memory data processing
  - Spark, RDDs, Transformations, actions
- Stream processing
  - windowing
- Big Graphs
  - Graph Management
  - Graph Processing