COMP9313: Big Data Management

Revision

Final Exam

- 19th Aug, 2020
  - Start from 11:59:59am (noon)
  - Last for 24 hours
- Questions will be published on Webcms3
  - https://webcms3.cse.unsw.edu.au/COMP9313/20T2/
  - Notices will be published on Webcms3, Piazza and the course's website
  - Submit your answer sheet using give
- If you need a special consideration, apply it now!
  - https://student.unsw.edu.au/special-consideration

Final Exam

- If you have clarification questions, please post in course WebCMS forum.
  - NO course staff available between 11pm and 8am.
- Open book exam
  - All of the work you submit for this exam should be completed by yourself without assistance from anyone else.
- Start the exam early and do not submit the answer in last minute
- Question will be formed like those in Assignment 1 and Sample Exam

Final Exam

- Proj2 and ass1 will be marked ASAP; we aim at delivering the result before the exam
- Pre-exam consultations:
  - 17 Aug (Mon): 1400-1600
- Pre-exam rehearsal

- Tips: Write down intermediate steps.

- Disclaimer:
  - We will go through the main contents of each lecture. However, note that it is by no means exhaustive.

Introduction to Big Data Management

- 3V's -> 7V's
  - What are they and what's the challenge?
- Applications of big data
- Big data management
  - Data Acquisition
  - Data Storage
  - Data Processing
    - Preparation, exploration, cleansing, integration, …

- Not needed:
  - Data curation

Hadoop and HDFS

- Master-Slave architecture
- Architecture of HDFS
  - Namenode, secondary namenode, datanode
- Fault-tolerance
  - Replication management
  - Erasure Coding

Spark and RDD

- Spark
  - Features
    - E.g., Lazy evaluation
  - Architecture

- RDD
  - Transformations and actions
  - Lineage
  - DAG

MapReduce

- Idea of Map and Reduce
- Shuffle

- Write MapReduce in Spark
  - Operations
    - E.g., combineByKey, …
  - Correctness
  - Efficiency

- Not needed:
  - TF-IDF

High Dimensional Similarity Search

- LSH
  - LSH functions for different similarity/distance functions
  - AND-OR composition
  - FP and FN
- C2LSH
- PQ
  - Framework
  - Distance Estimation

- Not needed:
  - Similarity search in low dimensional spaces
  - Multi-probe LSH
  - Detail about K-means

Spark SQL

- Dataframe
  - Features
  - Relationship to Spark RDD
- Operations
  - E.g., join, groupBy, …
- Columnar Storage

- Not needed:
  - Plan Optimization & Execution

PySpark MLlib

- Classification/machine learning
  - Definition, training, test, evaluation, …
  - Two-Step process
  - Cross-validation
- MLlib
  - Pipeline
- Ensemble learning
  - Stacking

- Not needed:
  - Naïve bayes Classification

Mining Data Streams

- The stream model/DSMS
- Sampling
- Sliding window query
- Filtering

- Not needed:
  - Sketches

Recommender System

- RS Model
- Key problems
- Content based Recommendation
- Collaborative Filtering
  - User-user
  - Item-item
- Factorization
- Modelling Local & Global Effects

- Not needed:
  - Knowledge-based RS

Thank You and Good Luck!