CS4225/CS5425 Big Data Systems for Data Science

Mid-Term Revision

Bingsheng He
School of Computing
National University of Singapore
hebs@comp.nus.edu.sg

School of Computing

Mid-term Test Instructions

- Time: 14-15:30pm March 18, Saturday.
  - Actual paper time will be around 1 hour.
- For both Grp L1 & L2.
- Venue: UTOWN AUDITORIUM 1/2.
- Held in person; open book + notes, but no electronics usage
- Seating plan will be given later.

Test

- Focus is on understanding and application, not facts / memorization
- Example questions
  - Integrative: Require you to combine knowledge from different chapters of the textbook
  - "Why not": Example, Tommy proposed a solution A to solve problem B in the lecture. Tell me what is the problem with solution A and how to overcome this problem
  - "From the book": Answerable as long as you attend the lecture and/or read the slide

Scope of Test

- **Scope**: the content in the lecture
- **Out of scope**:
  - The lab specific content
  - Your project
  - Additional information/note in the comment box
  - The content marked as “optional”
- In the following, I will
  - Have a revision on the **key points** that we learnt in this semester.
  - Go through several example questions.

Introduction to Data Science

- What is (big) data science?
  - 4V big data challenges
- Cloud computing and (big) data science?
  - Why cloud computing leads to big data
  - Why cloud computing is also a solution
- *Infrastructure* for big data
  - Data center architecture
  - "Big Ideas": past, present and future
  - **Given a particular algorithm/system, can you analyze it according to the "big ideas"?**

MapReduce

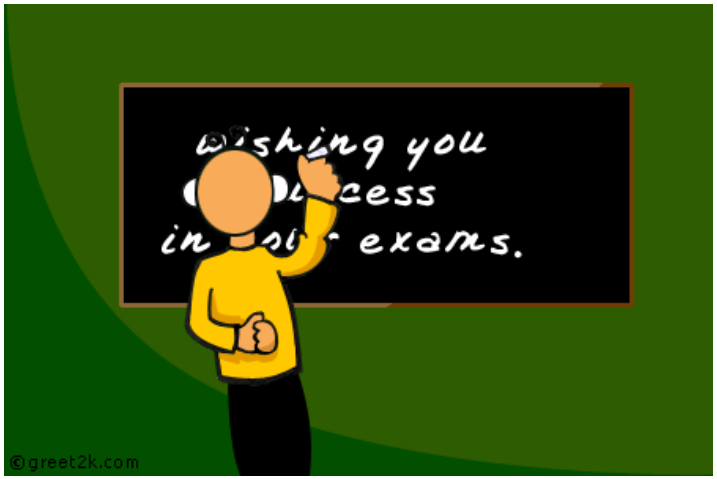
- System design principles
  - Why MapReduce?
  - System internal of MapReduce
  - Why or why not: e.g., why HDFS chooses three replicas?
- Basic algorithmic design
  - Performance analysis: parallelism, network and disk I/O
  - Given an algorithm, you need to conduct performance analysis and identify the performance issues for further improvement.
  - Given a problem, you need to design the solution in MapReduce and conduct performance analysis.

MapReduce

- Relational databases
  - ~~Value-to-Key Conversion~~
  - Joins
  - Algorithm designs and performance analysis with similar data operators.
- Large-scale machine learning
  - Similarity and clustering
    - How to calculate?
  - Their implementations on MapReduce
  - Algorithm designs and performance analysis with their implementations
  - Single job vs. multiple jobs

Test Paper Format

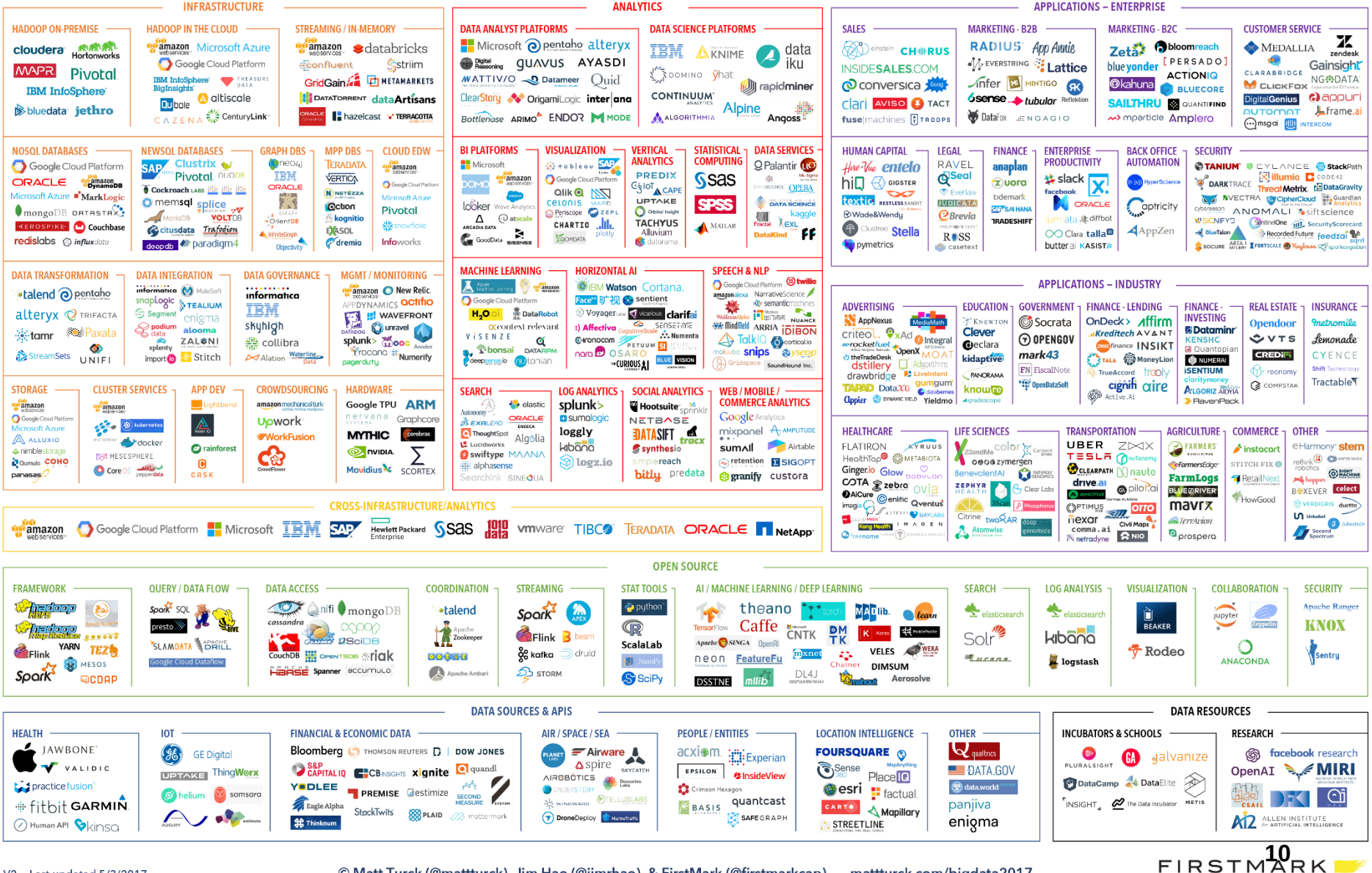
- Question structures (total 25 marks)
  - MCQ
  - Like what we have tried in the class quiz (we have tried more relevant question in the past year; previous years have different formats).
- MCQ:
  - Shade your answers on the OCR Answer sheet using a 2B pencil. You need to hand in both the OCR sheet AND this paper at the end of the test.
  - When multiple options can be the answer, choose the most appropriate combination from the available options.


©greet2k.com

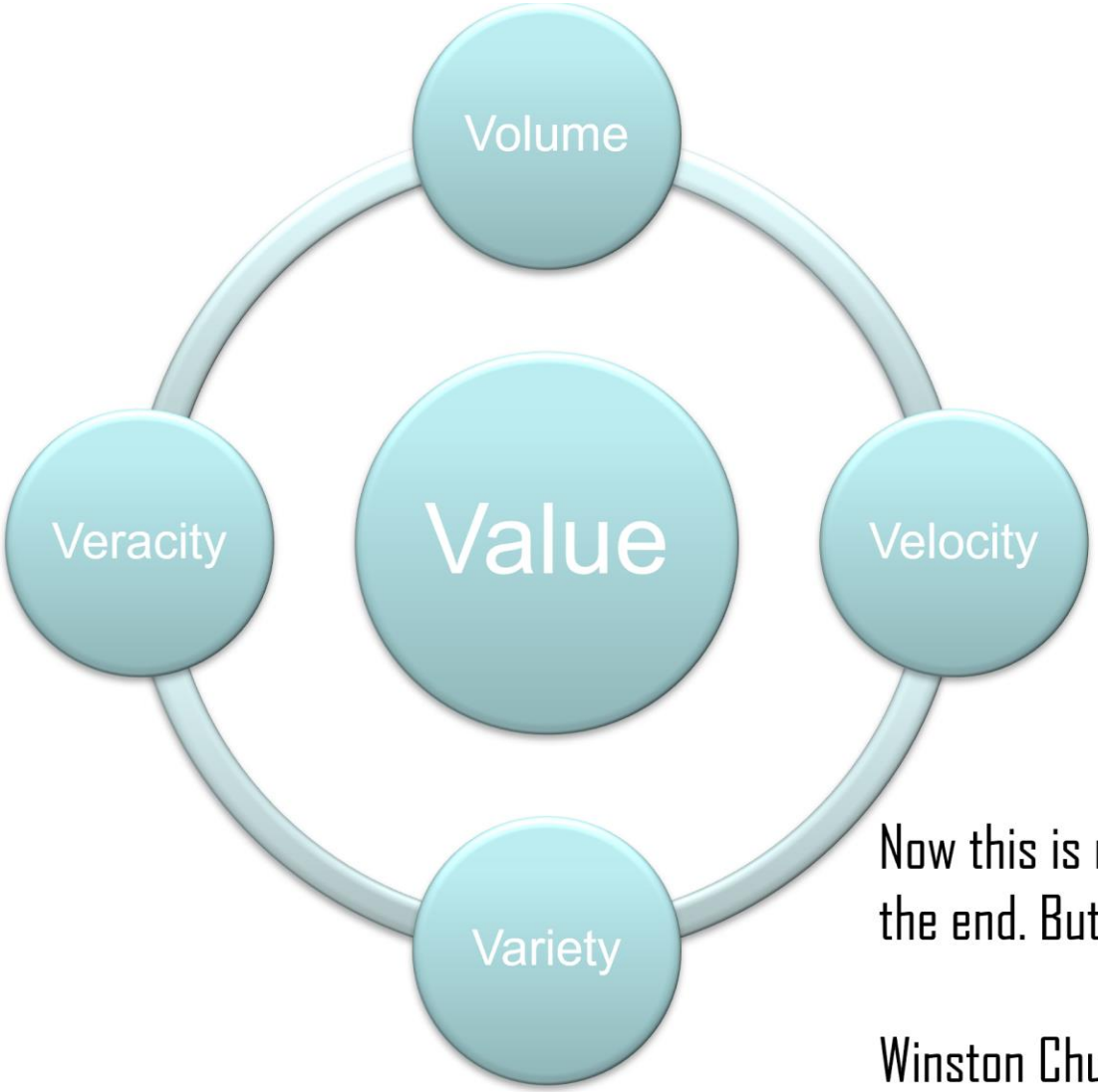

study bunnies
chibird

“One size does not fit all”

BIG DATA LANDSCAPE 2017



Future Big Data Systems



Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.

Winston Churchill