**City University of Hong Kong**
**Course Syllabus**

**offered by School of Data Science**
**with effect from Semester A 2021/22**

---

## Part I Course Overview

| | |
|---|---|
| **Course Title:** | **Big Data: The Arts and Science of Scaling** |
| **Course Code**: | **SDSC3001** |
| **Course Duration:** | **One Semester** |
| **Credit Units**: | **3** |
| **Level**: | **B3** |
| **Proposed Area:** *(for GE courses only)* | ☐ Arts and Humanities<br>☐ Study of Societies, Social and Business Organisations<br>☐ Science and Technology |
| **Medium of Instruction:** | **English** |
| **Medium of Assessment:** | **English** |
| **Prerequisites**: *(Course Code and Title)* | **CS3402 Database system** |
| **Precursors**: *(Course Code and Title)* | **Nil** |
| **Equivalent Courses**: *(Course Code and Title)* | **Nil** |
| **Exclusive Courses**: *(Course Code and Title)* | **Nil** |

## Part II Course Details

### 1. Abstract

*(A 150-word description about the course)*

This course aims at teaching students how to tame massive data which are intensively used in high-impact industrial applications. Students will learn two mainstream categories of technical solutions for big data, namely algorithmic approaches and systems approaches. For algorithm approaches, some popular stream algorithms such as heavy hitters and sketching algorithms used when we have a limited memory will be introduced. To deal with huge amount of data, the instructor will also teach sampling-based algorithms, such as approximate counting, that tame big data via sampling a representative small collection of data. For the system approaches, the instructor will introduce Spark, one of the most popular big data computing software nowadays, to the students. Topics in Spark include the MapReduce model, Spark RDDs, DataFrames, DataSets, Spark SQL and Spark ML.

### 2. Course Intended Learning Outcomes (CILOs)

*(CILOs state what the student is expected to be able to do at the end of the course according to a given standard of performance.)*

| No. | CILOs[#] | Weighting* (if applicable) | Discovery-enriched curriculum related learning outcomes (please tick where appropriate) | | |
|---|---|---|---|---|---|
| | | | *A1* | *A2* | *A3* |
| 1. | Understand that the scalability issue lies at the core of making data science practical. | 10% | √ | √ | |
| 2. | Understand basic stream algorithms and sampling algorithms. Be able to prove the effectiveness of these algorithms. | 30% | √ | √ | |
| 3. | Implement data processing algorithms using Spark. | 30% | √ | √ | √ |
| 4. | Apply the algorithmic techniques and system techniques in solving scalability problems in real applications. | 30% | | √ | √ |
| | | 100% | | | |

** If weighting is assigned to CILOs, they should add up to 100%.*

*[#] Please specify the alignment of CILOs to the Gateway Education Programme Intended Learning outcomes (PILOs) in Section A of Annex.*

*A1: Attitude*
*Develop an attitude of discovery/innovation/creativity, as demonstrated by students possessing a strong sense of curiosity, asking questions actively, challenging assumptions or engaging in inquiry together with teachers.*

*A2: Ability*
*Develop the ability/skill needed to discover/innovate/create, as demonstrated by students possessing critical thinking skills to assess ideas, acquiring research skills, synthesizing knowledge across disciplines or applying academic knowledge to self-life problems.*

A3: *Accomplishments*
*Demonstrate accomplishment of discovery/innovation/creativity through producing /constructing creative works/new artefacts, effective solutions to real-life problems or new processes.*

## 3. Teaching and Learning Activities (TLAs)

*(TLAs designed to facilitate students' achievement of the CILOs.)*

| TLA | Brief Description | CILO No. | | | | Hours/week (if applicable) |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| Lectures | Learning through teaching is primarily based on lectures. Mini-lectures and small-group exercises will be used to facilitate conceptual understanding and applications of various methods tools and techniques. | √ | √ | √ | √ | 26 hours/semester |
| Laboratory Work | Tutorials and project consultations | | | √ | √ | 13 hours/semester |
| Course Project | The team-based projects provide students with the opportunities to familiarize and apply the tools learnt during the lectures through practical problem solving. | | | √ | √ | After class |

## 4. Assessment Tasks/Activities (ATs)

*(ATs are designed to assess how well the students achieve the CILOs.)*

| Assessment Tasks/Activities | CILO No. | | | | Weighting* | Remarks |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | |
| Continuous Assessment: 70% | | | | | | |
| Group projects | | | √ | √ | 30% | |
| Assignments | √ | √ | √ | √ | 40% | |
| Examination: 30% (duration: 2 hours) | | | | | | |
| Examination | √ | √ | √ | √ | 30% | |
| **The weightings should add up to 100%.* | | | | | 100% | |

For a student to pass the course, at least 30% of the maximum mark for the examination should be obtained.

## 5. Assessment Rubrics

*(Grading of student achievements is based on student performance in assessment tasks/activities with the following rubrics.)*

| Assessment Task | Criterion | Excellent (A+, A, A-) | Good (B+, B, B-) | Fair (C+, C, C-) | Marginal (D) | Failure (F) |
|---|---|---|---|---|---|---|
| 1. Group projects | The project is to evaluate the overall performance and the attitude of the students in understanding, utilizing, applying the methodologies, principles and skills. The teamwork and collaboration is also accessed. | High | Significant | Moderate | Basic | Not even reaching marginal levels |
| 2. Assignments | Assess students' understanding of computational methods and common techniques. | High | Significant | Moderate | Basic | Not even reaching marginal levels |
| 3. Examination | Examination questions are designed to assess student's level of achievement of the intended learning outcomes, with emphasis placed on understanding and correct application, mostly through clear explanation, and numerical calculation, of the various data processing techniques. | High | Significant | Moderate | Basic | Not even reaching marginal levels |

The midterm and tutorial exercises will be numerically-marked, while examination will be numerically-marked and grades-awarded accordingly.

**Part III Other Information** (more details can be provided separately in the teaching plan)

## 1. Keyword Syllabus

*(An indication of the key topics of the course.)*

Algorithmic approaches:

- Stream algorithms: heavy hitters, distinct element counting, sketching algorithms, matrix sketching, graph sketching
- Sampling algorithms: approximate counting, Chernoff bounds, Monte Carlo simulations, Markov Chain Monte Carlo, graph sampling

System approaches:

- Spark basics: MapReduce, RDD, DataFrames, DataSets
- Advanced features of Spark: Spark SQL, Spark Stream, Spark ML

## 2. Reading List

### 2.1. Compulsory Readings

*(Compulsory readings can include books, book chapters, or journal/magazine articles. There are also collections of e-books, e-journals available from the CityU Library.)*

| | |
|---|---|
| 1. | Lecture notes |

### 2.2. Additional Readings

*(Additional references for students to learn to expand their knowledge about the subject.)*