# Course Form for PKU Summer School International 2018

| | |
|---|---|
| **Course Title** | **Foundations of Big Data Systems**<br>**大数据系统基础** |
| **Teacher** | Prof. M. Tamer Özsu, University of Waterloo, Canada |
| **First day of classes** | July 9, 2018 |
| **Last day of classes** | July 15, 2018 |
| **Course Credit** | 2 credits |

**Course Description**

**Objective：**

The course addresses the foundations of modern big data systems. The focus is on data management infrastructure. The course will address the fundamental challenges and components of big data systems and approaches that have been developed to address them. The objective is that by the end of this course, students should have a good understanding of the foundations of these systems.

**Pre-requisites /Target audience**

Pre-requisites: database, data structure and algorithm
Target audience： senior undergraduate students, Master and PhD students

**Proceeding of the Course**

The course will review the foundational issues of big data systems. The focus is on data management infrastructure. This infrastructure is typically built on top of modern distributed/parallel computing platforms (e.g., MapReduce, Spark), run a distributed/parallel data management platform, employ main memory systems (both row stores for OLTP and column stores for analytics), and consist of multi-modal systems to handle different types of data coming from different data sources. This course will cover these foundational issues.

The topics that will be covered are the following:

- Fundamentals of distributed and parallel data management, focusing on data fragmentation, distributed query processing, distributed transactions, replication, and data integration;
- Main memory systems and column-based data representation;
- Big data analytics platforms (distributed storage systems, MapReduce, Spark, graph analytics, stream data management);
- NoSQL, NewSQL and Polystore Systems (Key-Value Stores, Document Stores, Graph Databases)

The course will consist of 20 hours of lectures, 4 hours per day , followed by students' presentations (8 hours).

**Assignments (essay or other forms)**

**Evaluation Details**

Students' presentation.

**Text Books and Reading Materials**

There is no textbook that covers everything, but students will be provided with preprints of chapters from the upcoming fourth edition of

- M. Tamer Özsu and Patrick Valduriez, Principles of Distributed Database Systems, Springer, forthcoming

In addition, there are some additional reading materials that are identified for each session below.

Slides for all the classes will be provided on the course webpage.

**Academic Integrity (If necessary)**

**CLASS SCHEDULE**
(Subject to adjustment)

| Session 1: Fundamentals of distributed and parallel data management | Date: 9 July 2018 |
|---|---|

**【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.)** This session will cover the classical distributed/parallel data management topics such as data partitioning and distribution, distributed query processing, distributed transaction processing

**【Questions】**

**【Readings, Websites or Video Clips】**

Chapters 2, 4, 5, 6 of the forthcoming 4th edition of "Principles of Distributed Database Systems". These chapters will be provided on the course web page

**【Assignments for this session (if any)】**

| Session 2： Main memory systems and column-based data representation | Date: 10 July 2018 |
|---|---|

**【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.)** An important aspect of big data systems is main memory processing and column-based storage for data analytics. This session will cover these issues

**【Questions】**

**【Readings, Websites or Video Clips】**

F. Faerber, A. Kemper. P.A. Larson, J. evandoski, T. Neumann, and A. Pavlo, "Main Memory Database Systems", Foundations and Trends in Databases, 8(1-2): 1-130, 2016.

**【Assignments for this session (if any)】**

| Session 3： Big data analytics platforms | Date:10 July 2018 |
|---|---|

**【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.)** This session will focus on the platforms that have been developed for big data analytics. The specific topics that will be considered are MapReduce, Spark, and stream processing.

**【Questions】**

**【Readings, Websites or Video Clips】**

Chapter 11 of the forthcoming 4th edition of "Principles of Distributed Database Systems". This chapter will be provided on the course web page.

Charu C. Aggarwal, Data Streams – Models and Algorithms, Springer, 2007.

**【Assignments for this session (if any)】**

| Session 4： Graph analytics & graph databases | Date: 11 July 2018 |
|---|---|

**【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.)** Graphs have emerged as an important representation in big data systems. In this session we consider both the graph analytics and graph database issues.

**【Questions】**

**【Readings, Websites or Video Clips】**

Chapters 11 and 12 of the forthcoming 4th edition of "Principles of Distributed Database Systems". These chapters will be provided on the course web page.

Da Yan, Yuanyuan Tian, and James Cheng, Systems for Big Graph Analytics, Springer, 2017.

Da Yan, Yingyi Bu, Yuanyuan Tian, and Amol Deshpande, Big Graph Analytics Platforms, Foundations and Trends in Databases, 7(1-2): 1-195, 2017.

**【Assignments for this session (if any)】**

| Session 5： NoSQL Systems | Date: 12 July 2018 |
|---|---|

**【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.)** We give an overview of NoSQL systems focusing on Key-Value Stores and Document Stores

**【Questions】**

**【Readings, Websites or Video Clips】**

Chapter 12 of the forthcoming 4th edition of "Principles of Distributed Database Systems". This chapter will be provided on the course web page.

**【Assignments for this session (if any)】**