Course Form for PKU Summer School International 2018

Course Title	**Data Management for Big Data Analytics**
	面向大数据分析的数据管理
Teacher	Prof. Leonid LIBKIN, University of Edinburgh
First day of classes	July 23, 2018
Last day of classes	July 31, 2018
Course Credit	2 credit

Course Description

Objective：

The course is about data management aspects of Big Data. Up to 80% of the big data effort is what is commonly known as data wrangling – preparing data for machine learning and data mining algorithms. In fact traditional databases remain the main tool of data analysts. The course aims to introduce students to challenges of big data, and prepare them to conducting research, in both academic and industrial settings, in the areas of querying and managing big data, and expose them to current research and development in connection with big data theory. This course will cover foundational issues in connection with three of four big V's in the typical characterization of big data, namely, Volume, Variety and Veracity.

Pre-requisites /Target audience

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

Proceeding of the Course

The course will review fundamental challenges introduced by querying big data, such as the need for revising the classical computational complexity theory in the context of big data. Regarding Volume, it will deal with the feasibility of computing exact query answers in big data within our available resources, and approximate query answering. For Variety, it will cover popular data models, including relational, XML, graph, and RDF models, and languages for them, as well as handling queries over data residing in multiple sources, focusing on both virtual and materialized integration, and efficient query answering. For Veracity, it will cover handling poor quality information, understanding current technologies and their deficiencies, correctness guarantees, and consistent query answering, and will look into how ontologies help produce better query answers. Students will be introduced to the study of several query languages including SQL for relational databases, Cypher for graph data, and SPARQL for RDF.

Specific topics will include:
Joins and conjunctive queries: evaluation and analysis
Scalable query answering

Approximation of queries
XML databases
Graph databases: path queries and patterns
Graph databases: querying property graphs
Querying RDF data
Incomplete information and correct query answering
Handling inconsistent data
Data integration
Data exchange
Ontology-mediated query answering

The course will consist of 28 hours of lectures, 4 hours per day , followed by students' presentations.

Assignments (essay or other forms)

Evaluation Details

Students' presentation.

Text Books and Reading Materials

As the material is largely new, there is no single textbook that presents it. Some aspects are covered in existing books, e.g. conjunctive queries in
1. Serge Abiteboul, Richard Hull, Victor Vianu, "Foundations of Databases", Addison-Wesley Publishing Company, 1995
or data exchange in
2. Marcelo Arenas, Pablo Barceló, Leonid Libkin, Filip Murlak, "Foundations of Data Exchange", Cambridge University Press New York, NY, USA ©2014

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

Academic Integrity (If necessary)

CLASS SCHEDULE
(Subject to adjustment)

Session 1: SQL as a data analytics tool	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) What is the most common tool used by data analysts? It is, as multiple surveys show, SQL - the main query language for commercial RDBMSs. We give a gentle reminder of what SQL is, and then point out some serious issues that arise when one relies on SQL queries for data analytics.
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 2： Conjunctive queries: evaluation	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) Conjunctive queries, also known as select-project-join queries, are the most fundamental queries used in database management systems. Since database design principles prescribe splitting data into multiple tables, such joins need to be taken to obtain useful information. Naïve evaluation of conjunctive queries however is a computationally expensive problem. In this session we look at ways of speeding up conjunctive query evaluation, and at their static analysis for efficient optimizations.
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 3： Scalable query answering	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) These days we deal with enormous data repositories, so large that even a linear time algorithm over them can take days or weeks. To answer queries, we need new notions of complexity. The key idea comes in the form of scale-independence (even if data is huge, the part relevant to the query is likely to be small). We study the concept and look at access information that lets us find scalable queries.
【Questions】

【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 4： Approximation of queries	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) If, due to the size of data or complexity of the query, it is infeasible to find exact query answer, we need to approximate query results. We look into approximations of joins and conjunctive queries by queries with complexity guarantees.
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 5： XML databases	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) We give an overview of XML navigational languages, and connect them with specification languages used in software and hardware verification. We also employ the connection to look at static analysis of XML queries.
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 6： Graph databases: path queries	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) Graph databases are becoming popular due to new applications such as social networks and the Semantic Web. We look at models of graph data and languages for them, based on path queries and graph patterns, and discuss the complexity issues that arise in the evaluation of such queries.
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 7: Graph databases: querying property graphs	Date:

【Description of the Session】In most products such as those by Oracle, SAP, Neo4j, the model that is predominantly used is of property graphs: in those, nodes and relationship carry sets of key-value pairs that can be queries. We look at languages that have been developed for such graphs, including both theoretical extensions of path queries, as well as Cypher, a pattern based language of Neo4j, and analogs of XPath for graphs.
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 8： Querying RDF data	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) RDF is the formal underlying the Semantic Web; it is essentially a form of graph data where labels and nodes can be mixed. We look at languages for RDF, such as SPARQL, from database perspective and show that they have very natural counterparts in the relational database world.
【Questions】

【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 9： Incomplete information	Date:
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) It is well known that standard relational languages such as SQL produce very counter-intuitive results when information is incomplete. We study formal models of correct answers to queries over incomplete data, and explain that SQL, as is currently implemented, differs from them in all possible ways. We also show how to fix query evaluation so that it would eliminate incorrect query answers.	
【Questions】	
【Readings, Websites or Video Clips】	
【Assignments for this session (if any)】	
Session 10： Inconsistent data and consistent query answering	Date:
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) Inconsistency arises when a database does not satisfy prescribed specification; often this a byproduct of merging several databases. If data cannot be cleaned, one needs to query inconsistent data. We introduce a model of such querying and study its computational costs.	
【Questions】	
【Readings, Websites or Video Clips】	
【Assignments for this session (if any)】	
Session 11： Data integration	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) In data integration, data needs to be pulled from various sources and queries. Often though it is infeasible to actually move data and restructure it under a new schema, in which case the querying process is completely virtual. We look at techniques for such virtual data integration based on query rewriting.
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 12： Data exchange	Date:

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) Data exchange is the problem of moving data between applications. These can rely on data structured according to different schemas, and thus one needs schema mappings to reconcile them. We look at building target instances based on schema mappings, answering queries over them, and analysis of metadata, i.e., mappings themselves.
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 12： Ontology-mediated query answering	Date:

【Description of the Session】Often data comes together with additional knowledge in the shape of an ontology. Using such an ontology can improve the quality of query answers. We look at some ontology languages that people use, and describe algorithms that use ontologies to facilitate query answering.
【Questions】

【Readings, Websites or Video Clips】
【Assignments for this session (if any)】