VLDB Database School (China)

15-19.08.2011

Xi'an China

Report on 2011 VLDB Database School

By
Zhanhuai Li
School of Computer Science
Northwestern Polytechnical University
Xi'an, China
lizhh@nwpu.edu.cn

1. Introduction:

The 2011 VLDB Database School took place in School of Computer Science of Northwestern Polytechnical University, Xi'an, China.
This year the main focus of the summer school was on Large-scale Data Analytics and Knowledge Discovery Based on Cloud Computing. Recently the cloud computing paradigm has attracted a great deal of attention in the information industry. Data management applications are potential candidates for deployment in the cloud. To work in this area, one must have the solid background of the traditional method as well as a good knowledge of the new sensitive technique used in the cloud. This summer school gave us an opportunity to explore the new horizon of deploying database management in the cloud. Prof. Amr El Abbadi from UCSB and Dr. Divesh Srivastava from AT&T Research Center were invited to be the instructors of the school. Both of them were far from excellent. They touched from data management in the cloud to information theory in the database systems. The lectures were well organized covering a theoretical basics, technologies, and perspectives.

2. Program:

Monday, August 15

The school started with the welcome address from organizer **Prof. Zhaihuai Li** from Northwestern Polytechinal University, China. He spoke about the idea behind the organizing the school and also gave an overview about one-week proceedings of the

school. This was followed by addressing speech form Prof. Zhiyong Peng, Vice-President of the School and Prof. Ge Yu, the vice chair of CCF-DBTC.
After the opening ceremony, Prof. Amr El Abbadi gave the lecture on managing data in the cloud. He talked about some knowledge of the concepts and terminology associated with database systems and distributed systems.

Figure 1: Opening Ceremony

Figure 2: Prof. Amr El Abbadi

Tuesday, August 16

On the second day of the school Prof. Amr El Abbadi described the first generation and second generation of data cloud management systems. He also discussed the design principles and design paradigms of 2^{ND} Gen. Systems.

Wednesday, August 17

In the morning, Prof. Amr El Abbadi talked about Multi-Tenancy and Data Analytics. Lastly, he discussed the limitations and opportunities of deploying data management issues on these successful and promising cloud paradigms.
In the afternoon, Dr. Divesh Srivastava gave the lecture on information theory for data

management. He talked about preliminary background in information theory.

Figure 3: Dr. Divesh Srivastava

Thursday, August 18

On the fourth day of the school Dr. Divesh Srivastava discussed the information theory applications in data anonymization and database design.

Friday, August 19

On the last day of the school Dr. Divesh Srivastava discussed the information theory applications in data integration and copy detection.

Closing summer school

Prof. Zhiyong Peng, vise-president of VLDB Database School, gave the closing remakes on the summer school.

Figure 4: Phots of 2011 VLDB School (China), Xi’an

Figure 5: Prof. Zhanhuai Li gives a certificate to Dr. Divesh Srivastava

Figure 6: Dr. Divesh Srivastava and VIP

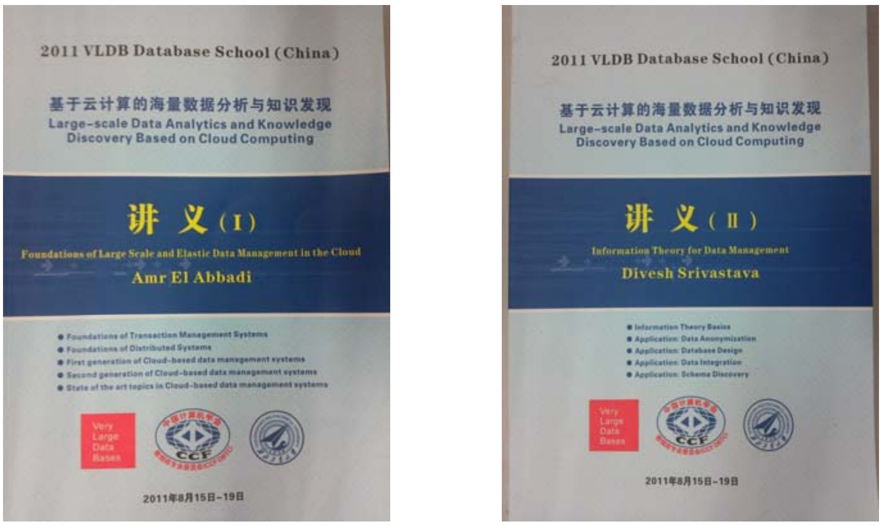

Figure 7: Lecture Notes

The program of the school is shown as follows.

August 14, 2011	
Registration (10:00 – 20:00)	
August 15, 2011 (Day 1)	
8:30 – 9:00	Opening Welcome Venue: **East Conference Room**
9:00 – 10:00	**Prof. Amr El Abbadi Title:** Session 1: Foundations of Transaction Management Systems
Coffee Break (Time: 10:00 – 10:20)	
10:20 – 11:50	Session 1: Foundations of Transaction Management Systems
Lunch (Time: 12:10 – 14:30, Venue: Hongyang Hotel)	
14:30 – 16:00	Session 2: Foundations of Distributed Systems
Coffee Break (Time: 16:00 – 16:20)	
16:20 – 17:50	Session 2: Foundations of Distributed Systems
Dinner (Time: 18:10 – 19:00, Venue: Hongyang Hotel)	
August 16, 2011 (Day 2)	
8:30 – 10:00	Session 3: First generation of Cloud-based data management systems
Coffee Break (Time: 10:00 – 10:20)	
10:20 – 11:50	Session 3: First generation of Cloud-based data management systems
Lunch (Time: 12:10 – 14:30, Venue: Hongyang Hotel)	
14:30 – 16:00	Session 4: Second generation of Cloud-based data management systems
Coffee Break (Time: 16:00 – 16:20)	
16:20 – 17:50	Session 4: Second generation of Cloud-based data management systems
Dinner (Time: 18:10 – 19:00, Venue: Hongyang Hotel)	
August 17, 2011 (Day 3)	
8:30 – 10:00	Session 5: State of the art topics in Cloud-based data management systems
Coffee Break (Time: 10:00 – 10:20)	
10:20– 11:50	Session 5: State of the art topics in Cloud-based data management systems
Lunch Time (Time: 12:10 – 14:30, Venue: Hongyang Hotel)	
14:30 – 16:00	**Prof. Divesh Srivastava Title: Information Theory for Data Management** Session 1: Information Theory Basics
Coffee Break (Time: 16:00 – 16:20)	
16:20 – 17:50	Session 1: Information Theory Basics
Banquet (Time:18:30 – 21:30, Venue: Le Garden Hotel)	
August 18, 2011 (Day 4)	
8:30 – 10:00	Venue: **International Conference Center** Session 2: Application: Data Anonymization
Coffee Break (Time: 10:00 – 10:20)	
10:20 – 11:50	Session 2: Application: Data Anonymization
Lunch Time (Time: 12:10 - 14:00, Venue: Hongyang Hotel)	
14:30 – 16:00	Session 3: Application: Database Design
Coffee Break (Time: 16:00 – 16:20)	

16:20 – 17:50	Session 3: Application: Database Design
Dinner Time (Time: 18:10 – 19:00, Venue: Hongyang Hotel)	
August 19, 2011 (Day 5)	
8:30 – 10:00	Session 4: Application: Data Integration
Coffee Break (Time: 10:00 – 10:20)	
10:20 – 11:50	Session 4: Application: Data Integration
Lunch Time (Time: 12:10 - 14:00, Venue: Hongyang Hotel)	
14:30 – 16:00	Session 5: Application: Schema Discovery
Coffee Break (Time: 16:00 – 16:20)	
16:20 – 17:50	Session 5: Application: Schema Discovery Closing
Dinner Time (Time: 18:10 – 19:20, Venue: Hongyang Hotel)	

3. Attendees:

The Summer school is targeted at young professionals working in the fields of data management. The recruitment was done by the organizer under the advices of the VLDB School and the CCF-DBTC.

The candidate sent his / her application with a short curriculum vitae including a list of publications and projects they attend, and a recommendation from his / her supervisor. The organization committee reviewed all applications and made a final decision depending on applicants' background, research achievements, and as well as the area distribution.

A total of 110 applications were submitted to the organization committee, from which 78 applications from 38 universities in China was finally accepted by the VLDB Database School (producing an acceptance rate of 70.9%). In addition, more than 20 students from local university were also admitted into the school, but supported by themselves. Most of attendees are PhD students. Others are junior faculty members and excellent master students.

4. Summary of 2011 VLDB Database School Expenses:

SUMMARY OF Expenses	
1. Sponsor from VLDB Endowment	$ 18,000
2. Sponsor from NWPU	$ 7,000
3. Registration Fee ($76 * 78)	$ 5,928
4. Tax for Registration Fee	$ -322
5. Honorarium for Instructors ($1000 * 2)	$ -2,000
6. Flights for Instructors	$ -5,462
7. Hotel for Instructors (9 days)	$ -1,938
8. Hotel for VIP (5 days)	$ -351
9. Hotel for Attendees (6 days)	$ -11,828

10. Attendee meals	$ -3,092
11. Banquet	$ -1,210
12. Refreshment	$ -667
13. Rental of Lecture Room	$ -1,246
14. Lecture Notes	$ -2,339
15. Transportation	$ -323
16. Communication	$ -150
17. **TOTAL**	**$ 0**

5. Acknowledgement

The VLDB Database School is supported by VLDB Endowment and Database Technical Committee of the China Computer Federation (CCF-DBTC). Thanks for their support.

We would like to thank our instructors, Prof. Amr El Abbadi and Dr. Divesh Srivastava, for their excellent lectures. They added a lot to our knowledge. We would also like to give many thanks to volunteers from the School of Computer Science, Northwestern Polytechnical University, for their hard work to make the 2011 VLDB Database School successful. Special thanks to Prof. Michael Franklin and Prof. Divy Agrawal for their help to organize the school.

6. Appendix:

The reports from the participants are as follows.

Feedback from East China Normal University

Professor Amr El Abbadi and Dr. Divesh Srivastava gave us tutorials on both solid background knowledge and cutting-edge technologies about cloud data management systems and information theories in data management. The tutorials were informative, systematic, in-depth, and interesting. They are inspiring for research and development work of students from East China Normal University.

Some teachers and students from ECNU are currently working on building a cluster-based data management system for unstructured data. The system focuses on storage and indexing for the data. Prof. El Abbadi's tutorials gave us a systematic introduction on transaction related technologies on how to build a such kind of system, which is very inspiring for our research.

Meanwhile, Dr. Srivastawa's tutorial on information theories in data management is very helpful for students and teachers working on web data management and information services. Our research focuses on effective and efficient information extraction from web data and analytical service providing based on web data. The information theory based approaches provide powerful tools for the research work.

There are very good interactions between instructors and students during the tutorials. Instructors posed quiz to collect feedback from students, while students may also be able to test ourselves.

We have two suggestions for future VLDB Summer School. First, it would be helpful if instructors can introduce the methodology to do research. For example, the instructors may introduce how to set up the problem for research, how to polish the idea to solve the problem, and finally how to write the research paper.

Second, it would be helpful to have a session for interactions. For example, after a tutorial, one student who is working on a related research problem may present his idea and work to the class. Then, the instructors may ask questions and give suggestions to his/her work, while all students may all gain benefits.

The VLDB Summer School 2011 is very successful. We'd like to thank Professor Amr El Abbadi and Dr. Divesh Srivastava, organizers and support from CCF TCDB. And we are looking forward for future VLDB Summer Schools.

Jinxian Wei

The just-past VLDB summer school offered an opportunity to learn about large-scale data analytics and knowledge discovery based on cloud computing. Both Professor Amr El Abbadi and Dr. Divesh Srivastava are respected professors. They gave us informative and engaging tutorials, which deserve to be chewed afterwards.

Throughout the five-day tutorials, the introduction and comparison of 1st and 2nd generation cloud systems impressed me most. Transaction management regains researchers’ attention after being left aside for several years along with the proliferation of key-value stores. Looking back the development of database community, it seems the history is being repeated all the time. The so-called 2nd generation systems add transactions to 1st generation ones, on the other hand, they try to minimize the distributed transactions in order to provide satisfactory performance and elasticity. In my opinion, there’s still a long way to go because limitations arise among the current design principles.

This summer school is a success. If there were any suggestions, I’d say paper writing techniques and industrial experience can be taken into consideration in the following years.

Feng Chen

It's very helpful for me through attending five-days VLDB Summer School at Xian.

Pro. El Abbadi and Dr. Srivastava are very gentle and patient for every question. By discussing with them, I expanded my thoughts. It’s grateful that some new ideas emerged by using the knowledge I learned from their lecture. I think information theory is helpful for identifying different-format semantic entities that have same meaning. In addition, we can find a new method that integrating information theory and NLP technology to find event tracking, event evolution, etc.

It’s also grateful for organization. The success of VLDB Summer School is the result of Northwestern Polytechnical University’s teacher and students’ effort. Thanks again! I look forward to the next gathering,

Yuming Lin

Dr. Srivastava impressed me with his talks on information theory for data management. Based on the fundamental information theory, Dr. Srivastava introduced four applications: data anonymization, database design, data integration and schema discovery. All of these have benefited us a great deal. I am most interested in database design based on entropy and mutual information. To determine whether or not a database is well-designed, we always estimate which level of norm form the database achieves. But Dr. Srivastava applied entropy and mutual information to evaluate the database, and he obtains a good effect with this method. I am inspired by this method, and we can adapt for generic use in our research.

Liang Zhang

Prof. Amr El Abbadi introduced several features of cloud-based data management systems, especially the idea of key value stores. The development of key value stores tend to satisfied the need of transaction, consistency, SQL etc. The design of MEGASTORE, G-STORE and the other new cloud-based data management systems take the key value store as data model or mix it with traditional data model of RDBM. I got much knowledge of key value storage database concept, including the basic ideas and the development of it in last few years.

Feedback from RenMin University

Yina Ye

This summer I have attended and completed all courses of VLDB Database School in Xi’an. The database school in Xi’an was very well organized which was very impressive to me. Moreover, the courses that the two instructors gave to us were extremely interesting, meaningful and insightful, which really benefited me a lot. Now let me share with you what I have learned from the two wonderful courses given by the two excellent instructors.

The first instructor whose name is Amr EI Abbadi gave us the course based on the topic of “Foundations of Large Scale and Elastic Data Management in the Cloud”. His course can be divided into five sessions.

Before all the sessions, he gave us a short introduction of the cloud computing. It included the points like why we need cloud computing now and the advantages and challenges that the cloud computing had. He gave a very clear elaboration of the difference between Scale and Elasticity, and he emphasized that in order to learn cloud computing well we need to have a good master of Database foundations and Distributed systems foundations first. That was why his course was organized by the way we could learn step by step.

In the first session, he talked about the foundations of Transaction Management Systems. This session included SERIALIZABILITY THEORY, CONCURRENCY CONTROL, RECOVERY, DISTRIBUTED DATABASES AND DISTRIBUTED COMMITMENT, which was rather easy for me since I had learned much of the knowledge on database classes in school.

The second session named Distributed Systems Foundations was a little hard for me, since most of the distributed systems knowledge he talked about was new to me, but I tried to get all the points. Overall, in this session he gave us many tools like quorums and vector clock which were very useful and used in many papers in the field of cloud computing. What is more, the theories he talked about in this session illustrated why database becomes the scalability bottleneck and thus cannot leverage elasticity when scaling in the cloud. The mechanism of partitioning data and committing in the distributed database system is very complicated and there are still difficult problems not solved.

In the third session, the instructor introduced the first generation of cloud data management systems including Google’s Bigtable, Yahoo’s PNUTS and Amazon’s Dynamo. They are designed for different commercial goals, therefore, they use different partitioning, fault-tolerance, consistency and replication approaches.

RDBMS scales well when limited to a single node, but meets with overwhelming complexity to scale on multiple server nodes, which results in a recent “NoSQL” movement. “NoSQL” has to use eventual consistency, which leads to a cost in programming model complexity. Therefore, the goal for the second generation is to build scalable, fault-tolerant, and consistent data management systems in the cloud that provides elasticity. The instructor introduced many typical systems whose design principles include separating system and application state, limiting interactions to a single node, decoupling ownership from data storage and limited distributed synchronization. These typical systems also take different partitioning approaches, including static and dynamic partitioning.

In the last session, the instructor mainly talked about multi-tenancy. Multi-tenancy requires elasticity in the database tier, and there are two tools for elasticity, that is, Migration and Autonomic Controller. The rest of the course is focused on the different approaches implementing these two tools.

The next instructor named Divesh Srivastava gave us the course based on the topic of “Information Theory for Data Management”. He is the head of the Database Research Department at AT&T Labs-Research, therefore, what he talked about was much more practical.

He first introduced the information-theoretic concepts to us, then he gave a “data-centric” perspective on the information theory, and at last he connected these concepts to the following four applications in data management. The information-theoretic concepts include Entropy, Conditional Entropy, Mutual Information, K-L distance and so on.

Application 1: Data Anonymization

For the need of Data Sharing, Data Retention and Usage, we require data anonymization. And we also need to make a good tradeoff between privacy and utility. The instructor illustrated different anonymization approaches for tabular and graph data and how we can use information theoretic measures to capture the loss of privacy.

Application 2: Database Design

In this section, the instructor showed how to apply information theoretic measures in relational database design. Good database design is essential for preserving data integrity. Information theoretic concepts are useful for measuring integrity constraints (Functional Dependency / Multivalued Dependency) and normal forms (BCNF / 4NF).

Application 3: Data Integration

Data in large organizations, governments is often siloed, it is collected at different times for different purposes, therefore data integration is needed to provide unified access to achieve higher utility. Multiple sources of data in the same domain exist on the web, different sources provide overlapping and complementary data, and using data integration will allow for comprehensive and high quality data. For schema matching, the instructor introduced the information theoretic approach in column alignment, opaque schema matching and finding foreign keys. He also talked about how to apply information theoretic measures into heterogeneity testing.

Application 4: Copy Detection

Copy detection is important in areas such as finding originality of rumor, finding manipulated data and finding truth on the web, and the instructor introduced how to apply information theoretic measures into document, software and database copy detection. At the same time, we shouldn't ignore the importance of the strategy's scalability and false positive/negative correctness.

I've learned a lot from this five days long systematic study in Xi'an. The instructors not only told us what to do, when to do, and how to do, but also why to do. Why there is a problem, how does it generate, based on the previous knowledge what tools do we have and what are the approaches to solve it. We study database, but it doesn't mean we should ignore knowledge in other fields of computer, instructor Divesh gave us a good example to apply information theoretic knowledge into several database related applications. Overall, the courses are helpful, the instructors are excellent, and the life in Xi'an is very comfortable, thanks to the host, sponsor and all the volunteers, this is really an unforgettable experience for me.

Xiaolu Zhang

I was so appreciate that I could have this hard-won opportunity to participate in the 2011 VLDB School, named "Large-scale Data Analytics and Knowledge Discovery Based on Cloud Computing". Prof. Amr El Abbadi and Dr. Divesh Srivastava are very knowledgeable and humorous, making these two courses very attractive. Through these courses, I learned a lot.

Prof. Amr El Abbadi mainly talked about Cloud-based data management systems and the foundations we need when we want to learn about it as a person who knows little about it. He spent 6 hours to introduce the foundations, including serializability theory, concurrency control, recovery, distributed databases, distributed commitment, distributed system, mutual exclusion and quorums, replication dictionary and log, global states and checkpoints, leader election, broadcast protocols, consensus and byzantine agreement, etc. He said that what we do now is to support the needs in our life. So during the next days, he used many examples to introduce the first and second generation Cloud-based data management systems, including TWITTER's problems and its solutions and lesions, GOOGLE storage system—Bigtable, GFS, Megastore, etc. Though I've read some papers about Bigtable and Megastore before, the content of the course shocked me, making me know that what I know about them is so little and which aspects should we go when we want to understand a new system that we never know before. During the courses, Prof. Amr El Abbadi highlighted the causes, application and the design principles of these systems. I think that this is what we most need to learn, because we not only need to learn the products, but the method

how to solve the problems.
Compared to Prof. Amr El Abbadi, Dr. Divesh Srivastava didn't arrange so much content, he manly introduced the information theory basics and several applications, including data anonymization, database design, data integration and copy detection. What related to my research is data integration, especially the schema-based matching and the instance-based matching. After the overview, he introduced three approaches: column alignment, opaque schema matching, finding foreign keys. Through the study, I understand data integration from another perspective, and know more comprehensive understanding of the data integration approach, I think these are very useful for my research in the future. In addition to data integration, I also learned some new knowledge, like data anonymization and copy detection.
Through the courses, I not only learned knowledge, but also improved my ability to listen to the report reported in English, which I think should be improved.
In a word, I learned very much from the courses, in addition to the knowledge related to my research, I also learn more about other aspects. Thanks again for this chance to take part in the summer school.

Jianxin Xue

VLDB Database School (China) 2011 was held from agu 15 to agu 19 in Northwestern Polytechnical University. Two professors were invited to give the lecture. Prof. Amr El Abbadi, processor of the Department of Computer Science at the University of California, Santa Barbara . Divesh Srivastava, the head of the Database Research Department at AT&T Labs-Research.
Prof. Amr El Abbadi gave a course about the Foundations of Large Scale and Elastic Data Management in the Cloud, including five courses. The first session is the Foundations of Transaction Management Systems, which is useful for my current work on CSQL.Then he introduce the foundation of distributed systems. After that he gave a course about the first generation of cloud-based data management system, the main feather is key-value store. The representative system including Google's Bigtable,Yahoo's Pnuts and Amazon's Dynamo.The second generation of cloud-based data management system differ from the first one is that it includes transaction in the cloud and elastic, the system include microsoft's SQL Azure,MIT's relational cloud,google's megastore and UCSB's G-Store.Finally he talk about multi-tenancy and data analytics. Traditional database systems focus on scale up for a single user .Today we should pay more attention on multitenancy and elastic. The multitenancy models inclue shared hardware, shared process and share table. Elastic in the database tire implication by migration and autonomic controller.
Prof. Divesh Srivastava gave topic about Information Theory for Data Management, including 5 courses too. Firstly he introduced the base of information theory, including Entropy, conditional entropy, mutual information, KL distance. In the following days he gave the lecture on Data anonymization, Data integration and Copy detection. They all based on the entropy and conditional entropy. The Data anonymization is used to hide information and protect the privacy. In the Data integration Divesh Srivastava gave some example from the paper .In the Opaque Matching it build complete, labeled graph GD for each database and computed the

H(X) for each label to match.

From these days study, I think Prof. Amr El Abbadi’s lecture is closed to my current work and I like the topic about the second generation of cloud-based data management system. In the following days I will focus on it.

Furong Li

It is of great honor for me to attend the VLDB Summer School this holiday. During the 7 days in Xi'an, I enjoyed the impressive courses as well as the beautiful scenery. As an undergraduate student, I cherish this opportunity and thank a lot to our lab. Next, I will give a shot summary of the courses and talk about what I have learnt from the courses.

In the first part of the courses, we have Prof. Amr El Abbadi to give lectures entitled Foundations of Large Scale and Elastic Data Management in the Cloud. In the first day, he talked about basic knowledge about database systems, especially transaction management. Then he gave lectures on distributed systems. We talked about message passing, log maintenance, recovery and hash. Though I have learnt about distributed systems and made use of HDFS before, this was the first time for me to have a deep look into the details of implementation of the distributed systems and the algorithms, such as Paxos and Snapshot.

Next, we moved to the first generation of Cloud-based data management systems. In this section, we had an overview of three key-value store systems: HBASE, PNUTS and Dynamo. From the diverse systems, we may come to a conclusion that different demands lead to different design goals. As a result, we see different consistency model, different replication policy and so on. Also, we can see that when designing the distributed systems, people borrowed a lot from the operating system area. So the knowledge of other fields will be helpful and it is a wise move for us to learn to grow the aspect of knowledge.

In the section of second generation of Cloud-based data management systems, we concerned about the update intensive workloads. To design both scalable and elastic platform, Abbadi proposed four design principles: Separate system and application state, Limit interactions to a single node, Decouple ownership from data storage and Limited distributed synchronization. These principles give me some inspiration, especially the last one. When you meet a stubborn problem, instead of solving it straightly, maybe you can it to be simpler.

In the last section, we talked about migration. We learnt about a system on live migration in shared-nothing databases for elastic cloud platforms. From their implementation, we can see there is still a lot of work to do in this area for the limitations in their system.

In the second part of the courses, we have Prof. Divesh Srivastava to give lectures entitled Information Theory for Data Management. Information theory is important in computer science, but I don't know much about it. At beginning, Divesh introduced the basic knowledge of information theory, including probability distribution, entropy, mutual Information, etc. Then he talked about four applications of information theory in data management area: Data anonymization, Database design, Data integration and

Copy detection. During the lectures, he used many examples which impressed us. These four applications are all easy to understand and very common in data management area, but the idea of information theory gives me a new way to look at them.

The VLDB Summer School gives us an opportunity to learn and to communicate. During the courses, we not only learnt knowledge in frontiers, but also learnt how to do research, how to find problems and how to solve problems, which are more important.

Min Zhang

The theme of VLDB summer school 2011 is “Large-scale Data Analytics and Knowledge Discovery Based on Cloud Computing”. The two main teachers that introduced this course are Prof.Amr El Abbadi from the University of California, Santa Barbara, and Dr. Divesh Srivastava from the Database Research Department at AT&T Labs-Research. Both of them are very kind, and gave us lectures in a very relaxed and humorous way.

The course given by Prof.Amr El Abbadi is calledd ” Foundations of Large Scale and Elastic Data Management in the Cloud”. In this course, we first reviewed some concepts of database and a broad overview of the key features of database management systems and distributed systems. And then we learned the novel computing infrastructures that have emerged as a compelling paradigm for scalable, elastic and reliable computing environments for building large-scale applications. Moreover, we have not only learned some theoretical knowledge, but also some applications based on these theories. For example, we explored the emerging approaches for building Internet scale persistent data stores (Google’s BigTable, Amazon’s Dynamo and so on).

The course given by Dr. Divesh Srivastava is named “Information Theory for Data Management”. In this course, we first learned the information theory basics, and learned some concepts, like entropy, K-L distance, conditional entropy and mutual information. Then we were give overview of some applications: Data anonymization, Database design, Data integration and Copy detection. In general, viewing information theory as a tool to express and quantify notions of information content and information transfer has been very successful as a way of understanding database design, enabling data integration, discovering and summarizing database schemas, performing data anonymization, etc.

In my opinion, 5 days is not enough for us to understand all the knowledge in the tutorials, and teachers also ignored some slides because of the limitation of time. What I have got from these lectures are some guidance of directions and some elicitation that can be used in my own research. What’ more, I can also exchange ideas with other students.

I also have some suggestions. I know the schedule is usually very tight, but I also wonder if we can have some time for our students to give short introductions about their research. Because I have found there are some teachers and doctors in this VLDB summer school, and maybe we can have a platform to exchange ideas like

Youth Forums.

Thanks for VLDB summer school, and I hope more and more students can take part in this school next year.

Lili Zhang

From August 15 to August 19, I was fortunate to attend 2011 VLDB Summer School organized in Northwestern Polytechnical University of Xi'an. In this event I learned a lot of latest knowledge and methods about data processing and database, knew many teachers and students within this field. After learning, my expertise had been expanded and improved and Prof. Amr El Abbadi and Dr. Divesh Srivastava, two speakers' rigorous scholarship attitude to their study made my attitude more correct. I understand that only a practical, rigorous and strive to work will make achievements.

The title of this Summer School is large-scale data analytics and knowledge discovery based on cloud computing. The course consists of two parts: one is instructed by Prof. Amr El Abbadi, whose title is foundations of large scale and elastic data management in the cloud and the other one is instructed by Dr. Divesh Srivastava, whose title is information theory for data management.

In the first part, we understand that cloud has three properties, those are commodity hardware, large scale and elasticity. With the cloud, we have many challenges in data increasing, data storing, data accessing and the failures. Thus, we must understand Transaction Management Systems and Distributed Systems. In Transaction Management Systems, we know the transactional concepts and for the concurrency problem database state is up-to-date at all times using On-line Transaction Processing. Each transaction is a correct mapping, so serial execution of transaction will be correct by Serializability Theory. There are lots of approaches to test the serializability. For concurrency controlling, we use Locking Protocol. Transactions execute in three phases, read phase, validation phase and write phase. For crash recovery, we use Logging Rules and the Crash-Recovery Algorithms. In Distributed Systems, we learn the Distributed System models and the concepts of lamppost logical clocks, vector clocks, quorums and distributed solution. Distributed checkpoint is the base of recovery. Each of the saved state is called a checkpoint and there are three kinds of the checkpoint. Distributed Snapshot Algorithm can detect the global state. For Leader Election, we can use Bully Algorithm and Ring Algorithm. For Broadcast Protocols, we learn the FIFO broadcast, causal broadcast, total order Protocol (ISIS) and the Paxos broadcast. For searching distributed data we know that Napster approach and Gnutella approach, and the location of the key of the data stored in distributed Hash Tables as nodes.

After the above, we study the first and second generation of Cloud Data Management Systems. To store data Google storage system uses Bigtable, which is used in different applications, such as Google Analytics, Google Earch&Map, and Personalized Search. And chubby is used for tablet management in Google file system, in Yahoo data is organized as hashed or ordered tablets using pnuts for the key-value storage. Dynamo provides primary-key only interface. Bigtable, pnuts and dynamo have different design goals.

In second generation of Cloud Data Management Systems, we care about transactions in cloud. There are two transaction processing benchmarks, Yahoo cloud serving benchmark (YCBC) and TPC-C benchmark. In Microsoft SQL Azure transforms SQL Server for cloud computing, which is simple one phase committing for replication. In MIT relational cloud is proposed to get the goals of scalability, elasticity and privacy. In Google megastore is used to please users, with the availability and scale. For dynamic partitioning, there are also many approaches, such as G-store in UCSB, database on S3 and consistency rationing in ETH, Deuteronomy in Microsoft.

In the second part, we understand Information Theory for Data Management, data anonymization, database design, data integration and schema discovery.

In Information Theory, we learn that tabular data can be represented as discrete distributions using random variables. We use randomization strategies for data anonymization. For data integrating, there are schema-based matching and instance-based matching, which have opaque schema matching. So plagiarism detection in documents, software and databases is necessary. To solve this problem we can use the strategies of LCS, Q-grams, COPS and winnowing for documents detection, use the text-based, tree-based and graph-based strategies for the challenges of the software copy detection. In Database, copying relationships can be complex. Solomon build Bayesian model to compute copy probability using data semantics. However, for global copying detection, those strategies may not work. We reason for each data item in a principled way to find the copying and adjust its copying probability. Based on mutual information we can quantify the loss of privacy.

To design a Relational Database, we should consider the schema, functional dependencies and normalization. In Functional-Dependency Theory, we may develop algorithm to generate lossless decompositions into BCNF and 3NF, then develop algorithm to test if a decomposition is dependency preserving. Then we use information theory to characterize how well-designed of databases.

For data integration, based on schema matching we have the approach to align columns using informative data values. Based on opaque schema matching, we build complete, labeled graph for each database with columns as nodes. And most works focus on inclusion dependencies to discovery foreign keys. We use randomness measure the likelihood finding the useful foreign-key and primary-key constraint. Information theoretic measures are useful for identifying columns with semantically heterogeneous values.

Five-day VLDB summer school life is very full, and it will be very helpful for my future research. I want to have the opportunity to participate this activity again. Thanks again for Northwestern Polytechnical University having given me this opportunity to learn.