Database Education at UC San Diego

Arun Kumar, Alin Deutsch, Amarnath Gupta,
Yannis Papakonstantinou, Babak Salimi, and Victor Vianu
University of California, San Diego
arunkk@eng.ucsd.edu, deutsch@cs.ucsd.edu, a1gupta@ucsd.edu
yannis@cs.ucsd.edu, bsalimi@uscd.edu, vianu@cs.ucsd.edu

1. INTRODUCTION

We are in the golden age of data-intensive computing. CS is now the largest major in most US universities. Data Science, ML/AI, and cloud computing have been growing rapidly. Many new data-centric job categories are taking shape in industry, e.g., data scientists, ML engineers, analytics engineers, and data associates. The DB/data management/data systems area is naturally a central part of all these transformations. Thus, the DB community must keep evolving and innovating to fulfill the need for DB education in all its facets, including its intersection with other areas such as ML, systems, HCI, various domain sciences, etc., as well as bridging the gap with practice and industry.

This article gives an overview of the DB curricula at UC San Diego and their rationales. Our Database Lab is one of the leading academic research groups in this area [2]. So, we have faculty resources to flesh out curricula that many other institutions may not have. Our research and teaching span all the major themes of theory, systems, languages, interfaces, and applications, as well as intersections with other data-oriented fields [5]. Areas of particular strength include database theory, data integration, semistructured and heterogeneous databases, hardware-conscious data processing, query processing and optimization, data exploration, data analytics, data systems for machine learning, causal inference, and responsible data science. Application areas of particular interest have included healthcare, Internet of Things, and social media.

As a result, our course repertoire in the DB area is extensive and eclectic. In this article we focus only on DB-related courses and exclude other prerequisites, e.g., Python for Data Science, as well as courses in nearby areas such as data mining, NLP, data visualization, and spatial computing. Note that UC San Diego follows the quarter system; so, each course is only 10 weeks long. We have over 18 distinct courses that are typically offered at least once, spread across 3 academic units: Computer Science and Engineering (CSE), Halicioglu Data Science Institute (HDSI), and Rady School of Management. The courses span all degree levels (BS, MS, and PhD), as well as professional and Online Masters degrees. In particular, as a strategic decision we created separate course series at HDSI tailored for Data Science majors instead of just cross-listing courses in CSE designed for Computer Science/Engineering majors. Key undergraduate courses (CSE 132A, DSC 100, and DSC 102) are offered twice or even thrice a year.

2. COURSES IN CSE

Four of five DB area faculty are affiliated with CSE. Naturally, CSE has the largest number of DB courses.

2.1 Undergraduate Level

The flagship database course, CSE 132A, is a required course for CS majors. It is typically offered at least twice a year, with enrollments surpassing 200. CSE 132B and CSE 132C are follow-up courses for students interested in application-oriented or systems-oriented aspects of RDBMSs. They are typically offered once a year and see enrollments in the order of 50. These courses are primarily aimed at students interested in becoming data engineers, DBMS developers, software engineers, data analysts, and DBAs.

CSE 132A: Database System Principles. It covers the basic concepts of databases, including data modeling, relational databases, query languages (SQL, QBE, JDBC, recursive queries), query optimization, database dependencies, schema design, and concurrency control. It includes hands-on assignments with RDBMSs. Prerequisites include the basics of logic and data structures.

CSE 132B: Database System Applications. It covers the design of databases, transactions, triggers, embedding SQL in general-purpose programming languages, and the architecture of database-powered websites. It also covers performance measurement, as well as the organization and use of index structures and materialized views. Prerequisite: CSE 132A.

CSE 132C: Database System Implementation. It covers the internals of an RDBMS such as data storage system, buffer management, indexing, sorting, relational operator implementations, query processing and optimization, parallel RDBMSs, and "Big Data" systems. It also includes two C++ programming projects to implement components in an RDBMS skeleton, BadgerDB from UW-Madison [1]. Prerequisites: CSE 132A, C++ programming, and basics of computer organization; optionally, an operating systems course.

CSE 135: Online Database Analytics Applications. This course is offered only sporadically nowadays. It covers data warehouse and data cube design, analytical SQL queries, online analytics applications, visualizations, data exploration, and performance tuning. Prerequisite: CSE 132A.

CSE 190: Post-Relational Data Models. It surveys a wide range of data models and high-level query languages that have achieved prominence with the advent of the Data Science revolution. These include graph database models and query languages in their various incarnations, ranging from XML to JSON, RDF and Semantic Web, and graph databases. The course distills the common ideas across these models and languages, connecting them to their common roots in object-oriented and SQL databases. Prerequisite: CSE 132A.

2.2 Graduate Level

These are offered once a year. The flagship course, CSE 232A, is popular with CSE's large MS pool and sees over 200 enrollments. The other courses are more specialized and see enrollments in the order of 20 to 50. CSE 233 and CSE 234 are one of a few courses of their respective kinds in the world, covering advanced topics.

CSE 232A: Database Systems Principles. It covers the principles of the internal implementation of SQL database systems: storage organization, query processing, transaction processing. It also covers selected advanced and modern functionalities, such as columnar storage, materialized views, and scalability. Prerequisite: CSE 132A or equivalent.

CSE 232B: Database Systems Implementation. It offers a hands-on approach to the principles of DBMS implementation via a project that takes students from specification of the grammar and semantics of the query language to implementing an evaluation engine, then building in optimizations. The project uses as vehicle the XML data model, but its lessons are universal across relational and post-relational models. Prerequisite: CSE 132A (or equivalent) and Java programming.

CSE 233: Database Theory. It covers the theory of relational databases, database dependency theory, deductive databases, incomplete information, and query languages, including connections to logic and complexity theory. Prerequisites: CSE 132A (or equivalent) and a basic course on computability and complexity.

CSE 234: Data Systems for Machine Learning. It covers data management and systems issues across the ML analytics lifecycle, including data sourcing, preparation, and organization for ML, programming models and systems for scalable ML training, systems for feature engineering and model selection, systems for inference and deployment, and ML platforms and feature stores. It offers a research project option instead of regular quizzes/exams. Students also review 8 to 10 cutting-edge research papers. This course is popular with students interested in becoming ML engineers or data scientists. Prerequisites: CSE 132A (or equivalent) and an OS course, or CSE 132C; an ML algorithms course.

2.3 Other Related Courses

Two other graduate courses overlap with DBMS topics but they are offered under the systems or ML areas.

CSE 223B: Distributed Computing and Systems. It covers distributed systems and networked servers, concurrent and event-driven architectures, remote procedure calls, load shedding, distributed naming/directory and storage services, replication and Byzantine fault tolerance, two-phase commit, consensus, CAP theorem, security in distributed systems, and blockchains.

CSE 255: Big Data Analytics with Spark. It covers scalability in data analysis using MapReduce and Spark, including minimizing bottlenecks in massively parallel computations, programming with PySpark, identifying computational tradeoffs in using Spark, performing data loading and cleaning using Spark and Parquet, modeling data with statistical and ML methods, and large-scale supervised and unsupervised ML using MLlib.

3. COURSES IN HDSI

UC San Diego launched HDSI in 2018, now its own academic unit, as part of a strategic decision to shape the future of Data Science education, research, and societal impact. As noted in a recent VLDB panel [6], we see the DB area as one of the key bridges between CS and Data Science. Thus, HDSI is also investing heavily in DB education but aimed largely at students interested in becoming data scientists and ML engineers, as well as data analysts and data engineers. HDSI majors are less likely (than CS majors) to become DBMS developers or software engineers. So, we re-thought the DB curricula from scratch for HDSI to de-emphasize some topics and add other relevant topics. Two of five DB area faculty are affiliated with HDSI. Researchers at the San Diego Supercomputer Center (SDSC) also regularly teach some HDSI courses.

3.1 Undergraduate Level

The two flagship courses, DSC 100 and DSC 102, are both required courses for HDSI majors. They are offered twice or sometimes thrice a year. Enrollments include 200 majors each year, with dozens more from other departments. DSC 104 is an elective for students interested in non-structured databases.

DSC 100: Introduction to Data Management. It introduces the storage and management of large-scale relational data, with an eye toward applications in Data Science. It covers the relational data model and some schema design aspects, relational algebra, SQL, some elements of query optimization, and some aspects of RDBMS-backed applications. This course is still evolving and we plan to add more discussion about the relationship between relational algebra and Pandas, including in the lower division course where Pandas is introduced. Prerequisites: Algorithms, data structures, and Python for Data Science.

DSC 102: Systems for Scalable Analytics. This is the first course of its kind in the world tailored for Data Science undergraduates. It offers a holistic bottom-up view of systems for scalable data-intensive computing. It covers the basics of computer organization and OS (only what is relevant for Data Science), memory hierarchy, data file formats, cloud computing, principles of scalable and parallel data processing, “Big Data” systems (MapReduce/Hadoop and Spark), and how all that matters for end-to-end workloads in Data Science. It includes Python programming assignments to analyze 40GB+ datasets using Dask on AWS and Spark on SDSC private cloud to perform data exploration, preparation, feature engineering, and ML model building. Prerequisites: DSC 100 and an ML algorithms course.

DSC 104: Beyond Relational Data Management. It introduces “NoSQL” data models, data formats, high-level query languages, and programming abstractions for semi-structured and graph-structured data, hierarchical and unrestricted graph DBMSs, array DBMSs, a comparison of expressive power of these data models, and parallel programming abstractions such as MapReduce and its descendants. Prerequisites: DSC 100.

3.2 Graduate and Other Related Courses

HDSI launched its MS and PhD programs in Data Science in Fall 2022. As with the UG courses, we rethought the graduate DB courses from scratch for HDSI instead of reusing CSE courses. A key reason was to offer viable pathways for students with non-CS UG majors (e.g., statistics, social sciences, or natural sciences) to enter Data Science careers after MS or pursue PhD-level research. Some of these courses are still being created. DSC 202 is a required course. DSC 204A is a conditional requirement that can be replaced with a graduate algorithms course by more theory-inclined students. CSE 234 (Section 2.2) and CSE 255 (Section 2.3) will be cross-listed and offered as electives.

DSC 202: Data Management for Data Science. It is the graduate version of a mix of DSC 100 and DSC 104, along with some more advanced topics. It covers the relational data model, relational algebra, basic SQL, “NoSQL” databases (document, key–value, graph, and column stores), and multidimensional data management, including data warehousing, OLAP queries, data cubes, and visualizing multidimensional data. Prerequisites: Similar to DSC 100.

DSC 204A: Scalable Data Systems. It is the graduate version of DSC 102. It covers the memory hierarchy, “Big Data” storage management, distributed scalable computing (cluster, cloud, and edge), parallel data processing at scale, dataflow programming models and systems (MapReduce/Hadoop and Spark), and their use for end-to-end ML analytics. Prerequisites: Similar to DSC 102; basics of computer organization and OS.

4. COURSES IN OTHER PROGRAMS

4.1 MAS in Data Science and Engineering

The Master of Advanced Science (MAS) is a professional degree offered by the Jacobs School of Engineering aimed at working professionals [3]. It offers graduate-level courses in a classroom-style environment but follow a hybrid model combining synchronous distance learning and one in-residence week each quarter. Data Science and Engineering (DSE) is one of three MAS programs, and it includes four DB-related courses. Apart from the two below that are custom-designed for MAS, CSE 255 content is offered as “Scalable Data Analysis” and DSC 104 content is offered as “DSE 250: Beyond Relational Data Models.”

DSE 201: Data Management Systems. It covers relational, hierarchical, and network data models, SQL and other query languages, DBMS architectures (including parallel, columnar, and array systems), and advanced SQL features, including user-defined functions, triggers, statistical functions, and support for spatial data.

DSE 203: Data Integration and ETL. It covers the fundamentals of data integration, including schema mapping and matching, entity disambiguation, ontology development and management, data provenance, and crowdsourcing and ML strategies for integration. It includes hands-on projects to integrate two or more datasets from an application domain of their choice, e.g., geospatial, healthcare, finance, bioinformatics, etc.

4.2 Online Masters in Data Science

UC San Diego launched its first Online Masters in 2022, focused on Data Science and created jointly by CSE and HDSI [4]. Like the MAS in DSE, OMDS is also aimed at working professionals interested in Data Science careers. But OMDS is more like a regular MS in its structure, rigor, and learning outcomes, albeit fully online. It also has more emphasis on statistics, deep learning, AI, and data ethics. It will feature a new DB course designed with a careful mix of content from DSC 100, DSC 102, and DSC 104. CSE 255 content will also be offered as DSC 232R.

DSC 208R: Data Management for Analytics. It covers principles, techniques, and tools for organizing, storing, querying, transforming, and using data for analytics and ML computations at scale. This includes the basics of data storage, acquisition, governance, organization, principles of the relational data model, relational algebra and its relationship to DataFrames, SQL, RDBMS features for faster querying and analytics, and basics of "NoSQL" systems. Major data quality issues and approaches to clean data, basics of cluster/cloud computing, MapReduce/Spark, and their application to scale feature engineering will also be covered. Methodologies to critically evaluate analytics results, including debugging and reasoning about bias and fairness in analytics pipelines will also be covered.

4.3 Rady School of Management

Relational databases and SQL are commonly taught to MBAs and other students in business schools as well. At UC San Diego, the Rady school also covers such topics in at least three analytics courses. A new course also covers "Big Data" systems such as Hadoop and Spark.

MGT 153: Business Analytics. It is designed to help a business manager use data to make good decisions in complex decision-making situations. It covers core business analytics concepts and skills, including Excel, relational databases, and SQL, principles of effective data visualizations and interactive data visualization (e.g., Tableau), and data preprocessing and regression analysis using data analytics programming (e.g., Python).

MGT 455: Customer Analytics. Many firms have extensive information about customers' choices and how they react to marketing campaigns but few have the expertise to efficiently act on such information. This course teaches a scientific approach to marketing with hands-on use of technologies such as databases, analytics, and computing systems to collect, analyze, and act on customer information.

MGTA 462: Big Data Technology and Business. It offers students the skills and knowledge to manage large-scale and complex data for business applications. Students will learn how to handle large volumes of data through database systems and to use big data technologies to perform scalable analytics.

MGTF 495: Special Topics in Business Analytics. It covers data modeling, relational database systems, SQL, text analytics, and elements of NLP using Python libraries. Class projects involve analyzing SEC 10Q reports from public company and combining this information with relational data about the company obtained from web sources.

5. CONCLUDING REMARKS

We hope this article on DB education at UC San Diego is helpful to organizations planning to reshape or expand their own DB education repertoire. We recognize that we are likely at the higher end of the spectrum on extensiveness. And yet, there are ever more emerging DB-related topics that could merit their own courses, e.g., ML for data management and RDBMSs, data ethics and responsible data management, knowledge graphs and knowledge management, scalable systems and interfaces for visual analytics, and DataOps and MLOps in this era of cloud, Web, and IoT. We believe all universities and in turn, wider society, will benefit from investing more in relevant DB-related education, research, and translation programs to better serve our modern data-driven world.

6. REFERENCES

[1] BadgerDB at UW-Madison. Online at `https://pages.cs.wisc.edu/~jignesh/cs564/projects/BadgerDB/BufMgr/docs`.

[2] CS Rankings for "Databases" Area. Online at `https://csrankings.org/#/index?mod&us`.

[3] MAS in DSE at UC San Diego. Online at `https://jacobsschool.ucsd.edu/mas/dse`.

[4] OMDS at UC San Diego. Online at `https://omds.ucsd.edu`.

[5] UC San Diego Database Lab. Online at `https://dbucsd.github.io`.

[6] Ives, Z., Gehrke, J., Giceva, J., Kumar, A., and Pottinger, R. VLDB Panel Summary: "The Future of Data(base) Education: Is the Cow Book Dead". *SIGMOD Record* (sep 2021).