Proposal for an MS in Data Architecture and Management

Dr. Kal Bugrara, Program Director IS/CSYE

Objective

We are proposing an MS degree in Data Architecture and Management where the focus will be on the skills needed by Data Engineers. This MS degree is to recognize **data engineers** as first-class citizens (instead of being secondary to Software Engineers and Data Scientists).

The Information Systems (IS) Program has a great variety of software engineering courses and has added Big Data and Data Science courses to its portfolio. In addition to the IS program, Northeastern has a Master of Science in Data Analytics, Master of Professional Studies in Analytics, Master of Science in Business Analytics and Master of Science in Data Science. These programs' courses deal with two aspects of data: collecting it and analyzing it but do not address the crucial activity that data scientists, data analysts, business analysts and many software engineers need to perform to make that data valuable – integrate it. That activity may be referred to as **data integration, data preparation, data curation, application integration and data engineering based on the integration use case and integration persona.** Our proposal for a Master of Science in Data Architecture and Management will focus on these activities.

In the past, data integration was primarily the concern in the business intelligence (BI), data warehouse (DW) and master data management (MDM) domains dealing only with structured data. Initially cloud applications, big data and data science activities were outside of BI and DW but that has changed as the data and data integration domains have expanded to all varieties of data.

Data systems engineering occurs because data is fragmented and usually scattered across many data sources but even if all the data one needed was in one place there is still an intensive need for integration because **information is data in context** and the context of data as collected is different than the many ways it needs to be transformed into useful information.

The next generation of data integration tools use machine learning to recognize logically complex interconnected patterns in vast data spaces automatically. The success of this line of work will have a radical impact on how to cost-effectively derive knowledge to push for social and economic change. We want data engineers to be ready for such a future.

The Nature of the Need

There is no question that data engineering is a heavy focus of business enterprises in all kinds of industries. It is said that 80 percent of data science projects are data engineering. It is also said that a typical enterprise has about 5000 databases. UHG, the health insurance company, for example, has about 30,000 databases scattered all over the place. These databases tend to be fragmented and dispersed in so many business units and geographic areas.

From a "siloed" to a holistic view of data, the proposed Data Architecture and Management Program is about taking gigantic amounts of data from diverse sources (sales, marketing, transactions, etc.) and turning this data into future business opportunities and/or avoiding risk situations. Businesses highly

regard such understanding since no one wants to be caught not making the right decision especially at the same time as the data they have is indicative of what should be done. However, assessing data is not always straightforward: it is critical to realize that while one pile of data might indicate certain things, when combined with other data, it very well may yield completely new information, cues, directives, etc. Data system engineering, in other words, can perform complex aggregations of a variety of vast arrays of data. In healthcare for example, data engineering of clinical data can help understand disease and produce new insights to the cure.

Scientists from all fields, especially math, physics, and biology, are now drawn to data science. They jump to these jobs for which statistical modeling skills are required. But the serious problem is that the majority of these scientists lack the data engineering skills required. So their focus stays inadequately on the math and statistics side rather than on the data-engineering side. This is inadequate since being able to deliver the data in context (their significant results) requires the engineering of big data systems; and such work can only be done by data engineers.

Technology in this area is rapidly maturing. And since data curation and engineering techniques designed to combine, process, and analyze data are becoming more mature as well as inexpensive, businesses are moving quickly to make an investment in such technology. The proposed MS in Data Architecture and Management program is designed to respond to such demands.

Our Data Architecture and Management Program will offer a multitude of courses in Data Engineering in addition to supplementary courses that are required to deliver the data results in a meaningful way to management.

We plan to cover data management, advanced data management, data warehousing and business intelligence, column data bases, data science engineering, and big-data engineering. On the Computer Systems Engineering/Software Engineering side, we offer advanced big-data programming using the powerful Scala language and a course on advanced data science as well as cloud computing. Multi-thread concurrent computing is also offered as it is important for synchronizing a huge set of servers working in parallel to do large scale analytics to make things run faster by 100's fold increase in speed. We cannot emphasize enough that, due to the high-level mathematical operations required to make these programs run, only software engineers have the capacity to work in these complicated areas. Only they can make the necessary mathematical algorithms execute quickly enough to get the finest results.

Moreover, our data engineers will be fluent in the application of data science **for the sake of building the actual data systems**. They will learn how machine learning algorithms work on top of statistical packages. If students choose to take the Algorithms class, students will learn the fundamentals of logical computing formulation and program construction as well as the mathematical modeling and analysis of algorithms -- an essential aspect of data science analytics. In the data science classes, students will learn how data pipe lines work in conjunction with clustering techniques, along with topic modeling and classification and logical regression techniques as well as Bayesian statistics. In the Engineering of Big-Data Systems classes, students will learn how configure and operate a Hadoop environment (large clusters of commodity hardware) and in the process they will learn how to integrate data from diverse sources, and to move and manage data through big-data platforms and pipelines (in-house or in the cloud). Data ingestion, the filtering and firing of millions of operations to run over large clusters of commodity hardware, is a data-engineering technique that we teach our students how to

perform through Scala, multi-threading, Spark programming, “map-reduce” techniques, in addition to data science, etc. We would leave our students in the Dark Ages of science were we to circumvent all of this complex technology; instead, we rigorously train them in the most up-to-date data techniques — enabling them to manipulate algorithms in a way that produces the very systems that data analytics is founded on.

In our planning of big-data projects class, students learn how to build the business case for big data integration projects and see them though an execution roadmap that involves understanding the architectures underpinning such gigantic platforms’ and the resourcing and cost issues. Funding models and ROI analysis are studied in this class.

Our user experience and web design will show students how to display the results in ways that make them make sense to management.

Program Courses

Please note that a number of courses that currently exist in IS and CSYE will be cross-listed, or preferably shifted to the new program, as indicated.

Core Courses

INFO 6105 - Data Science Engineering Methods and Tools (DATA 6105)

INFO 6210 - Data Management and Database Design (DATA 6210)

INFO 7370 - Designing Advanced Data Architectures for Business Intelligence (DATA 7370)

INFO 7250 - Engineering of Big-Data Systems (DATA 7250)

In cases where any of these courses currently may have prerequisites outside of the Data Architecture and Management program, these prerequisites will be modified to eliminate such dependencies.

Elective Courses

Students can choose up to 4 courses from the Data Architecture and Management electives below:

INFO 7290 - Data Warehousing and Business Intelligence (DATA 7290)

INFO 7255 - Advanced Big-Data Applications and Indexing Techniques (DATA 7255)

INFO 7390 - Advances in Data Sciences and Architecture (DATA 7390)

CSYE 7200 - Big-Data System Engineering Using Scala (DATA 7200)

CSYE 7245 - Big-Data Systems and Intelligence Analytics (DATA 7245)

CSYE 7250 - Big Data Architecture and Governance (DATA 7250)

Or from the Information Systems electives below

INFO 6255 – Software Quality Control Systems

INFO 6215 – Business Analysis and Information Engineering

INFO 6245 – Planning and Management of Information Systems

INFO 7325 – Introduction to Information Technology Auditing

INFO 6660 – Business Ethics and Intellectual Property for Engineers

INFO 7300 – Engineering of Cyber-secure Software Systems