June 16, 2020

Proposal
Stevens Data Cloud Project (SDC)
A Data Sandbox for Analytics and Data Science Students

"More Data Beats Clever Algorithms, but Better Data Beats More Data" – Peter Norvig (Google)

Motivation

Although a number of courses in our curriculum cover database and big data technologies and all students in all courses deal with data sets large and small, it is clear that we can do a better job of creating a data environment closer to what students will find on graduation. This scalable and flexible environment would support the development of professional data analysts and data scientists with the high level of expertise needed by industry.

For students SDC will provide a unifying thread, a sandbox for acquiring data management skills. SDC will be used in all courses replacing the current episodic, ad hoc, instructor-dependent instruction that characterizes our own and most other analytics curricula. Students will be able to access a curated set of data technologies - SQL and NoSQL database systems - and one or more very large "big data" sets for learning in specific classes and for use in home works and term projects. As students progress through the program they will be exposed to the techniques and technologies used by data analysts, data engineers and data scientists in all phases of the big data life cycle: data ingestion, data acquisition, data awareness, data wrangling, data stream management, data security, data process optimization, machine learning and deep learning and data visualization.

For faculty and PhD researchers, SDC will provide the same technologies and data sets plus skilled student research assistants able to support research on very large, perhaps proprietary, data sets.

Finally, employers, who have already embarked on data on the cloud journey, will gain a source of well-prepared graduates able to contribute on day one to their data science teams.

Architecture

SDC has three synergistic components that will be available to students and researchers.

1. Database systems linked to particular courses:

   MIS 630A Database Management: a relational database such as MySQL
   Topics: Design of transaction databases, SQL queries

   MIS 636 Data Warehousing and Business Intelligence: MongoDB, Cassandra, SQL Workbench
   Topics: ETL, NoSQL, document databases, distributed databases

   BIA 678 Big Data Technologies: Hadoop HSDL, Apache Kafka
   Topics: Distributed data bases, Spark, streaming databases

All BI&A and other Stevens classes would access the cloud-based environment for data analytics: machine learning, deep learning, visualization and so on. Students will develop projects using GitHub

2. A cloud-based environment similar to what students will use in industry on graduation. Currently, we envision the use of AWS for this purpose.

3. Research level databases in at least one area critical to our own research.

   An example, though hopefully its disproportionate value will be temporary, see https://ieee-dataport.org/open-access/free-dataset-newsmessage-boardsblogs-about-coronavirus-4-month-data-52m-posts . This data set, taken from news feeds about Coronavirus-19, has resulted in many thousand accesses/uses, and considerable visibility for its originators[1].

   Another example would be to use data.gov such as health care data from https://catalog.data.gov/dataset?_organization_limit=0&organization=hhs-gov#topic=health_navigation. This data set allows for analysis of scenarios such as patient survey's, outpatient surgrey surveys and much more.

Development

SDC will be developed by faculty and student team over a period of several years.

Phase 1 – Proof of Concept (Summer 2020)
Secure initial funding for student help. Obtain student access to AWS, install major data bases (MySQL, MongoDB), HDFS, Kafka, Data Connectors. Develop proposal for phase 2, which will include funding for development and funding. Use IEEE COVID database for pilot implementation.

Phase 2 – Development (Fall 2020 – spring 2021)
Secure funding for full implementation. Develop training programs for students and faculty. First use by courses MIS 630, MIS 636, BIA 678. Enhance training programs, develop more facilities.

Phase 3 – Full Deployment for both internal and external use (2021-2022)

Funding

Student and faculty resources will be required to develop and maintain SDC. Funding will be sought from both school and industry sources.

Concluding Thoughts

Often our courses feature "Team" projects that come close to replicating a static, and fairly flat, version of this using one of the many excellent public data sets (e.g. Kaggle, IEEE-Dataport). Sometimes we even get the loan of a data set from a corporation as well. However, we seldom

[1] . One of our faculty members, David Belanger, leads the IEEE-Dataport project that underlies this, and more than 1000 other available data sets.

replicate the dynamic nature of the data, its semantics, and its structure. The motivation for SDC is that great data analysts/scientists/engineers become, over time, intimately familiar with a few critical data sets, and learn how to take advantage of the data itself, including its domain, syntax, and semantics. Providing guidance in this direction to our students, would be a differentiator for our program.

There is a very important secondary motivation – the creation of research level databases in at least one area critical to our own research. As an example, though hopefully its disproportionate value will be temporary, see https://ieee-dataport.org/open-access/free-dataset-newsmessage-boardsblogs-about-coronavirus-4-month-data-52m-posts . This data set, taken from news feeds about Coronavirus-19, has resulted in many thousand accesses/uses, and considerable visibility for its originators. It will shortly be joined by other related data sets. One of our faculty ,member, David Belanger, leads the IEEE-Dataport project that underlies this, and more than 1000 other available data sets. Having great data is one of the fastest ways to recognition in data science. One of the best quotes on this comes from Peter Norvig (Google) – "More Data Beats Clever Algorithms, but Better Data Beats More Data". Of course, this needs to be combined with expertise on the data itself.

Why it might be possible

Some things have changed fairly dramatically over the past few years. Access to data, powerful tools to manage and analyze the data, and very powerful resources for students/researchers to use in their work, has changed the game in terms of education in data science and engineering. Over the last few years, in one of my courses – BIA678 (Big Data Technologies), we have moved to an environment where students can, and are expected to, access fairly large data sets (millions or billions of rows, or bytes of text) with significantly interesting semantics. In addition, they are exposed, and often work on pipelines of data using powerful data streaming tools. They do this using a combination of laptop tools (e.g. Cloudera, Tableau), and AWSEducate (EMR). These tool sets are available to them generally without cost. So, we have the infrastructure to do what we need to do in a way that would not have been workable a few years ago. There are still some things missing. Among them the experiences of integrating a variety of different data sets to solve problem, and working with data that is evolving and semantically/structurally complex. There appears to be at least some interest in actively helping from members of the BIA Advisory Board Members.

What it might it look like

There are a few layers of this idea in terms of ease/time/practicality of implementation, likelihood of gaining appropriate resources (including human, machine, and data) at an affordable cost, and time. In rough order of level of difficulty they are:

- Choose one or two domains with an accessible, reasonably large, and interesting set of data. Create a data base for each of them, in an appropriate format. Ideally these data sets would be evolving so that we, e.g. students, would be able to add content to them over time, understand how that content might be changing (e.g. drift), and run various analytic studies on them. Choice of the domain is critical both in terms of available data, domain expertise, and ongoing interest level. Possible areas might be: supply chains, networking, finance, marketing, etc. Sources could be primary, perhaps possible in a few cases; publicly available data; new feeds, partnerships, etc.

- Second level. Since, to be really interesting, data bases should represent the integration of a few sources, it would be useful to create more that one source for any given data subject.. This could mean multiple databases, or an integrated database, but would allow for analysis on the integrated data.
- Third level. Modern data analysis results are increasingly time sensitive. A source of streaming data that would allow the creation of data pipelines would be very nice.

What would it take to do it.

- Appropriate sources of interesting data, in relevant domains.
- A resource to store, manage, manipulate, and analyze such data. Depending on size and complexity. Leveraging our current use of AWSEducate could be a start.
- Human resources to build and maintain appropriate software so that it is moderately easy for students to access and use the facility.

A Few Potential Barriers.

- Although AWSEducate has been idea for classroom structured use (i.e. both without cost to the student, reliable, feature rich, and fairly easy to use) this capability would cross semester boundaries, hence would require agreement on a somewhat altered model.
- Creation, but particularly maintenance, of the data and systems involved requires ongoing commitment.
- It requires inspired insight into appropriate domains, and negotiation of data in those domains.