The Data Engineer

Mike Tamir
Chief Science Officer
Galvanize

Steven Miller
Global Leader Academic Programs
IBM Analytics

Alessandro Gagliardi
Lead Faculty
Galvanize

Businesses are quickly realizing that data scientists can only go so far without the team in place to support their day-to-day work, but more importantly to operationalize their work.

The 'sexy' new role of data scientist[1] has driven tremendous action from universities. New "data science" programs (broadly construed) at the undergraduate, graduate, and professional education level are now being regularly announced. Schools typically respond by building upon an multiple existing programs (comp sci, math, statistics, business) and filling in gaps as needed to quickly create new certificate and degree programs rooted in their current strengths & capabilities.

The result is that it is very hard to compare programs, and even harder to pin down exactly what constitutes data science. Is data science simply a field of applied machine learning fundamentals? Or is a broader definition needed to take full advantage of modern data rich applications?

Data science is better thought of as a broad field with numerous sub-fields, not unlike physics which has 5 major fields (applied, astro; atomic, molecular, and optical; condensed matter; particle) which in turn have 5 to 26 subfields[2]. While physics is a mature discipline, data science is nascent. We do expect the major fields & initial subfields to emerge quickly as we strive to understand data science.

[1] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/
[2] https://en.wikipedia.org/wiki/Template:Subfields_of_physics

Since data science crosses numerous existing fields we are already seeing a variety of programs with different focus.

- Applied Data Science fits the scope of the majority of the initial programs such as Case Western's B.S. Degree[3]
- Data science specializations within existing fields such as medical & bioinformatics, and actuarial science. The National Institutes of Health for example has created a new group focused on data science[4].
- Computational math, science and engineering, often taking root in high performance computing centers,[5] are being created[6] to meet the emerging need.
- Computer science led labs such as the Berkeley AMPlab[7] (Algorithms, Machines, People)

A major field of data science that is receiving little notice is Data engineering. Data engineering is at least a distinct sub-field of data science if not it's own field altogether. Let's explore why this is the case.

Those on the leading edge are not only analyzing data, but are increasingly implementing solutions that change how the business is run – in software by machines -- recommendation engines, fraud detection, real time pricing & bidding, the list of possibilities is endless.

These solutions often combine data from multiple sources and must run in near-real time at scale supporting thousands if not millions of simultaneous users. And must be secure while protecting individual privacy. These are not the skills of the data science modeler, but are the skills of the data engineer.

A few have pigeon holed the data engineer as data wranglers or plumbers who clean up data and make it ready to analyze. That is a key part of the work, but is only a component.

The data engineers work includes a broad range of knowledge and skill:

[3] http://datascience.case.edu/BS-degree
[4] https://datascience.nih.gov/
[5] https://icme.stanford.edu/
[6] https://cmse.natsci.msu.edu/
[7] https://amplab.cs.berkeley.edu/

- Extract, clean, and integrate data (wrangling)
- Bridge between the data science models and production systems
- Implement machine learning & computational algorithms at scale
- Put the right data system to work for the job at hand. Meaning they need a deep understanding of transactional ACID databases along with a growing variety of NoSQL databases including JSON document, graph, column stores, and partitioned row.
- Demonstrate a deep understanding of distributed computing and database considerations for consistency, scalability, and security.
- Protect customer privacy and anonymity.

While educators are fast at work creating data science programs, very few are focused on data engineering. Yet, data engineer job openings actually outnumber data scientist job openings. As December 10th 2015, LinkedIn has openings for 52,870 data engineers[8] vs. 25,070 data scientists[9].

Where will the skilled data engineers of the future come from? Certainly not graduates from the typical undergraduate computer science or information systems program. The typical undergraduate program offers one course in database, but usually only as an elective, and typically focused on older relational (RDBMS) thinking and system administration.

Master's programs often provide a few more options, but not much. Courses are typically offered as electives and only cover a subset of the skills needed by a data engineer. Big data master's programs that aren't just about the data scientist are emerging, but they are comparatively rare (UCSD[10] or Dundee[11] for example).

Major employers in the bay area have responded by supporting a stop-gap solution the data engineering recruiter fee based programs such as Insight's[12] which focus largely on the data wrangling aspects of the data engineer .

[8] https://www.linkedin.com/job/data-engineer-jobs

[9] https://www.linkedin.com/job/data-scientist-jobs

[10] http://www.jacobsschool.ucsd.edu/mas/dse/

[11] http://www.dundee.ac.uk/study/pg/dataengineering/

[12] http://insightdataengineering.com/

In the Galvanize data engineering[13] course, we cover the state of the art in the technology that makes data science possible with so-called "big data". "Big data" is frequently described in terms of the three traditional "V"s: variety, volume, and velocity and with all three, there is certainly the case that–to quote Philip W. Anderson–"more is different." In each case there is the situation where traditional tools simply won't work, both on a technological level and on a theoretical level. At the same time opportunities open that would never have been possible in the old world of "small data".

One of the things that makes living in the digital age so interesting is that we have access to so many different forms of data and have the capacity to cross-reference and evaluate them. Traditional science tends to focus on the data emitted by one sort of measuring device, but data science often contends with data from many different sources and of many different types. How do you predict stock prices from Twitter status updates? New approaches to how you use and combine different sources of data require a diverse set of skills that are taught in our program.

"Big data" has been described by some as "more data than can fit in memory." This definition, while limited, is operationally useful. As soon as big data can no longer fit in memory, traditional tools, from Excel to R, simply stop working. Python can be made to work, but it requires a fundamental shift in how one approaches the problem. The same goes for other languages. Similarly, traditional statistical analysis begins to break down. As N reaches into the millions, p frequently shrinks to zero, providing ample opportunity for spurious correlations to lead to faulty conclusions. Both of these problems (the technological and the theoretical) can be solved by sub-sampling, but that ignores the fact that new opportunities arise that may be far more effective.

The appearance of large volumes of data makes permutation analysis and cross-validation easy, making possible analyses that are both more ecologically valid and easier to interpret. However, in order to do this, we need tools to allow us to work with huge amounts of data that will not fit on your laptop.

[13] http://www.galvanizeu.com/program

Many companies today, particularly web companies—but increasingly other industries as well—have access to huge amounts of data that they do not know what to do with and cannot make use of. Simply having terabytes (or petabytes) of access logs in a Hadoop cluster does not make a data-driven organization. Those data are useful only if they can be made accessible in a meaningful way. Enter the data engineer. A data engineer makes it possible to take huge amounts of data and translate it into insights. The important phrase here is "makes it possible". To many data scientists and most data analysts (and nearly all managers), huge volumes of data are, in themselves, quite useless. A savvy data engineer can provide an interface to those data that make them useful.

High velocity in data also provides a qualitative shift in data science, but again, only if the engineering is in place to make it possible. Some traditional approaches to data science are already able to make use of huge amounts of data, but they tend to do so as though it were a one-time event. This is fine if you are trying to understand what happened last year, but it is not useful if you are trying to make decisions about what's happening right now. Data engineering allows us to tap into the stream of data *as it's happening* and do something about it. In some cases, that simply means providing a fault-tolerant scalable architecture that can remain responsible with millions of users try to access a service at the same time (something that every successful web company must deal with). But to take it a step further, we see the potential of having systems learn and adapt to their users in a way that was not previously conceivable. This is already happening at large companies like Google and Facebook, but the technology is available to all, if only there were the people trained to make it possible. Enter the data engineer.

Data Engineering Recap

Data engineering is not simply about maintaining a repository for huge volumes of data, it is about creating possibilities for everyone from developers to data scientists to executives. The technologies are evolving quickly and a good data engineer must always be learning, but certain fundamental patterns have been established which are guiding the way forward. In our program we focus both on the state-of-the-art as well as the principles behind them so that when the next technology comes out, we will be well equipped to adopt it.