The Full-Stack Data Scientist

CKM Analytix

October 2019

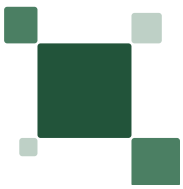

Table of Contents

Executive Summary

Market views on what constitutes data science and the data scientist continue to evolve, with frequent confusion over how to grow and generate business value from such teams.

Companies have established data science teams with and assigned data scientist titles to individuals with a vast array of skills and experience, which often leads to confusion over what skillsets are required to be an effective data scientist. If a data scientist is to be effective at delivering tangible business impact then such an individual will need more than pure technical skills—they need to be a full-stack data scientist.

In this paper we review the core skills of a full-stack data scientist and the importance of those skills in driving tangible improvements for the stakeholders of any data-driven effort.

Such skills include:

- Identifying/understanding the business problem (abstract level)
- Understanding the business/operation (detailed SME level)
- Identifying data sources
- Data engineering
- ETL
- Analysis/Modeling
- Story communication
- User empathy
- Productionizing output
- Having an impact

Underlying all these areas is a foundational focus on continually learning and adapting new skills, and an ability to quickly learn technical and domain-specific subject matter core to developing and executing analyses that will have an impact in any given industry. The full-stack data scientist can quickly learn when they need to learn when a situation says they need to know it. The full-stack data scientist never uses "I don't know" as an excuse, but rather thrives on this as an opportunity.

Introduction

The ultimate measure of success of a data science initiative in business is this: Did it have a positive and measurable impact for stakeholders?

Put another way, doing good data science requires a lot more than just good data science. That bar is easy to define but often hard to achieve.

Companies large and small continually evolve their thinking on applying data science to their business activities, building data science teams, and the value returned from investments in both. Underlying all this is the ongoing tricky question of "what is a data scientist?"

Way back in 2013 I wrote a piece called "The Data Scientist—Illusive or Elusive?" at a time when the market was just getting up to speed with the broad advancements falling under the umbrella of data science[1]. The widely cited "Data Scientist: The Sexiest Job of the 21st Century"[2] article had recently appeared in the *Harvard Business Review* and companies were scrambling to hire data scientists and promote their data science efforts.

That original piece was an attempt to add some descriptive specifics to the skills necessary to succeed in data science. A lot has evolved since 2013, but the basic tenants of skills outlined in back then have stood the test of time.

Fast forward 6+ years and in many ways the market is still broadly confused about what data scientists should be doing, how to integrate data science teams into business operations, and what constitutes a strong data scientist. With more years of experience behind us it's increasingly clear that the future of the data scientist revolves around the skillsets associated with the full-stack data scientist.

In this article we define the skills of a full-stack data scientist, outline why that complement of skills is key to successful data science initiatives, and address some of the broader title inflation that only adds to the confusions around the skills required for developing future data science leaders.

---

[1] "The Data Scientist: Elusive or Illusive?" https://ckmanalytix.com/blog/post/2/

[2] Davenport, T.H. and Patil, D.J. (October 2012) Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*.

What is a full-stack data scientist?

Put simply, a full-stack data scientist is one who takes a data-driven concept from identification/ideation through to execution that results in some tangible, measurable and impactful improvement. There is a heavy emphasis on being able to drive an organization to do something, not just analyze something. In the technical sense "full-stack" in the data science realm has a lot of parallels with descriptions of full-stack developers—i.e., one that can handle all aspects of the technical development process.

However, "full-stack" in a data science context also includes several additional requirements specifically focused on ensuring that this technical prowess works towards an end that includes tangible improvements for stakeholders. For example, one could have a full-stack of technical skillsets and produce a sophisticated ensembled machine learning model for a business process. Technically and academically the asset produced may be very impressive, but if that asset fails to have tangible and positive improvements for stakeholders then it will ultimately fail to be of much use to the business. The full-stack data scientist understands that success means improving the business and adapts their activities to always focus on achieving that goal.

But what about 'specialists' who only focus on one element of the stack?

It's no secret that a lot of the realities of doing data science are less than headline grabbing excitement. Much has been written about the fact that the majority, if not the vast majority, of most data projects is still spent on foundational tasks like collecting and cleaning data. [3]

Few data scientists find this part of an engagement to be the most exciting and would happily let others solve these problems for them. The urge is high to focus only on selected parts of the expertise required to be a full-stack data scientist—dare we call them the 'fun' parts. To do this some will quickly try to position themselves as 'specialists' in an attempt to avoid the 'less fun' parts of data science.

Is such a person truly a 'specialist?' The question boils down to this: If you are claiming to be a specialist are you truly moving the market needle in your space or are you simply trying to

[3] For example this article was among the first mainstream discussions on this topic: https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

carve out a world where everything else in the full-stack could be done by other people? The significant majority of those claiming they are 'specialists,' and thus don't do everything in the full-stack, fall into that later description.

'Moving the needle' here is a very high bar. Creating net new assets for the data science community (e.g., publishing significant new approaches to machine learning that are then adopted by the community) would be 'moving the needle.' Simply deploying tools and assets others created is a valuable skill but isn't 'moving the needle' enough to claim being a specialist. Any decent full-stack data scientist can also similarly develop advanced models by leveraging a broad suite of published utilities.

From a career development standpoint it would be exceedingly rare that one could truly meet the high bar for being a specialist without first being a solid full-stack data scientist. This shouldn't be surprising as it's the profile that most professions follow—specialization follows first establishing one's broad general credentials—but within data science there are some that would like to carve out one area right from the start (e.g., machine learning for NLP) and focus on nothing else. That approach wouldn't work for most professions and longer term it won't work for data science either.

The risk for those trying to resist the push towards full-stack data science is that all but the most expert specialists will always be trumped by the ever-growing body of those expanding their skillsets across full-stack data science.

The role of title inflation on all this

You'll notice that 'analysis' or 'model building' is actually only one of a long list of skills in the repertoire of the full-stack data scientist. Historically those that focus just on analysis and models were titled as business analysts, data analysts, BI engineers or similar. As talk of data science picked up momentum across the business community a significant degree of title inflation took off, further adding confusion to what it meant to be a data scientist.

Seemingly overnight companies claimed to have thousands of data scientists on staff. Closer inspection reveals many such claims to simply be re-branding exercises with longstanding armies of data analysts suddenly finding themselves now being called data scientists. While some of the tools may have evolved and improved, the overall remit of such teams remained broadly unchanged.

Much has been written about such title inflation and the challenges this often creates for companies from a recruitment standpoint.[4] Entry-level roles are often much more title sensitive and thus simply handing out a more senior sounding title proves an effective and free recruitment tactic for companies looking to scoop up talent.

Many companies have cited their frustration with this, with some eventually just deciding to follow a trend of rebranding their 'analysts' as 'data scientists' to keep pace with title inflation elsewhere—then rebranding their existing data scientists as something else (e.g., applied scientists or research scientists). We'll leave a deeper discussion of titles and the broad degree of title inflation for another day, but such movements do only further add to market confusion around skillsets.

For the purpose of this paper we'll use the phrase full-stack data scientist to describe senior data science roles, albeit with the understanding that different companies may assign different titles to such roles. The point of the discussion here is the skillsets involved.

Skills of a Full-stack Data Scientist

To accomplish the remit of "executing on a data science business initiative from beginning to end" a full-stack data scientist should be able to comfortable handle the following aspects of such an initiative, including:

- Identifying/understanding the business problem (abstract level)
- Understanding the business/operation (detailed SME level)
- Identifying data sources
- Data engineering
- ETL
- Analysis/Modeling
- Story communication
- User empathy
- Productionizing output
- Having an impact

[4] For example https://eng.lyft.com/whats-in-a-name-ce42f419d16c

Nobody can be an expert in every one of these areas, but one can develop a strong set of skills and experience in these areas along with an ability to learn what you need to know at that moment to get the job done. This concept will appear several times in the sections that follow, but that's because being able to adapt to the situation at any given point in time—even if that situation extends beyond previous experience to date—is the ultimate differentiator between an expert and a novice in any field.

On specific projects, subject matter experts (SMEs) also play a key role in helping the full-stack data scientist quickly turn data into action—for example, system administrators for a key data source will be key collaborators in gaining access to critical data for the project.

A more detailed discussion of each of these core skillsets follows below:

Identifying and Understanding the Business Problem (abstract level)

Launching into a bunch of data with the intent of 'finding something interesting' is rarely a good plan. Such efforts—frequently called fishing expeditions—may occasionally turn up something useful, but more often end with a frustrating back and forth between data science teams and business stakeholders as both sides try to find something that is both interesting and useful to the business. For the best chances of success, a data science effort needs to start with a defined problem to solve.

At some level nearly all data science efforts could be broken down as an effort to impact one or more of the 3 core forces of service operations:

- Improving service quality
- Identifying and mitigating risks
- Maximizing efficiency

We've previously written about these three forces of operations. [5] A full-stack data scientist can work with business stakeholders to define what success looks like for the effort as defined around such core business value levers. The data scientist can also help prioritize efforts when, as is not uncommon, initial planning identifies may possible paths forward.

Good technical assets (e.g., a great machine learning model) are key to achieving that success, but success is ultimately measured by having positive impact on the core question at hand

[5] "The 3 Natural Forces of Service Operations" https://ckmanalytix.com/blog/post/34/

(e.g., applying that model to have a measurable improvement on the efficiency of the underlying operation).

Throughout the course of working towards that defined objective, a good full-stack data scientist will typically identify many other interesting opportunities that can increase the depth or breadth of the scope of work. The uncovering of such previous unknown unknowns is a key benefit of data-driven approaches, but it's important that the core mission focused on a defined business problem remains in sync with the business to ensure that stakeholders get the value they need when they need it.

Understanding the In-Scope Business/Operation (detailed SME level)

Data scientists will frequently find themselves in a position of having to develop and present findings to stakeholders that are far more experienced in the particular subject matter at hand. The ability to incorporate relevant subject matter expertise into the data science work is critical not only to the generation of useful analyses but also to establishing and maintaining credibility with such stakeholders.

The full-stack data scientist doesn't need to be, and most likely is not, the subject matter expert (SME) for the in-scope business subject but they do need to know enough to best leverage the expertise of available SMEs and to maintain credibility with business stakeholders. Maintaining that credibility is key to a strong working relationship between the data scientists and the business stakeholders.

For example, a full stack data scientist being asked to work with a laboratory research team should invest some effort beforehand to bring themselves up to speed with what happens in the lab and the basic scientific principles behind the research work there. This is not all that unlike a politician preparing for a debate. Nobody expects them to be the world expert on every topic, but if asked questions about a topical matter they need to be able to speak intelligently about it. A solid full-stack data scientist briefs themselves accordingly prior to any meeting in a similar way.

The risk of getting this wrong is that the eventual data analysis, results and proposed actions will say or propose things that don't pass a smell test by the SMEs. When that happens, excitement over the possibilities of using data science to help quickly turns into annoyance, frustration and lost trust in the team. Once that trust is lost, it's hard to gain back.

Naturally, all of this relies heavily on the underling willingness to venture into unknown or less familiar territory that was mentioned as the ultimate core skill.

Identifying Data Sources and Data Engineering / ETL

Data is, of course, the raw material behind all the work of the data scientist.

In the ideal world a company has a data lake complete with every bit of raw data generated by every possible data source. The data scientist simply needs to reference the meticulously maintained data dictionary and tap into an API for instant high-speed access to anything they could possibly need, complete with detailed documentation.

OK, back to reality. We've never seen any large company that has a complete end-to-end data lake like this. While some companies have made great strides in creating centralized data repositories these efforts consistently struggle to achieve the state described above. That doesn't mean such efforts were wasted—even incomplete data lakes still streamline efforts to access data—but it does mean that a data scientist can't avoid the need to sometimes identify, establish and ingest data feeds from new data sources.

The full-stack data scientist is able to take charge of such efforts to probe and understand where digital data is created, where its stored and how to get access to it. Some of these data sources may be quite obvious while others are more tangential. Understanding the potential signal(s) generated by each and incorporating that into a broader analytical effort is key to generating the most complete and accurate analysis possible.

Once the data is in hand, it can't just blindly be used for analysis. Traditional pipelines talk about Extract-Transform-Load (ETL). There's one more essential step up front and that is 'Understand.'

Understand what each field in the data means, or is intended to mean, and understand how data is populated in that field. Skipping this step can get you burned badly by, for example, assuming something in the dataset means one thing when it's actually recording something completely different. This concept is discussed in much more detail in another recent article "Data Isn't Just Data."[6] Getting this understanding right requires an inquisitive mindset more

[6] "Data Isn't Just Data" https://ckmanalytix.com/blog/post/30/

than anything else and an unrelenting push to understand why and how things appear within the dataset.

As for the rest of the skills required for ETL, there's no specific right or wrong way to do this provided you understand the implications of any adjustments to the original raw data. From a technical skills standpoint one needs to have a solid understanding of databases, APIs, networking, data pipeline technologies and a good understanding of fitting the right tools to the right job (e.g., a giant Kafka cluster probably isn't needed to ingest some flat files from a server).

In terms of specific skills, one need to know what you need to know to get the job done—sounds a bit silly perhaps, but that's the truth. It's impossible to train for every possible scenario and thus any sort of checklist for skills is just unrealistic. Rather, the true skill here is the ability to quickly learn and adapt to whatever situation you find. Clearly over time more experience makes that process easier, but that ability to be confidently go into the unknown beats just about any inventory of established skills provided you can couple that ambition with a strong sense-check-sense to know when to take a step back if things aren't going as planned.

The skills required here are often procedural as much as they are technical, particularly within larger organizations. Accessing new sources of data and getting that data to move from A to B requires one to quickly get up to speed with local procedures, governance practices and security protocols.

Being charismatic and quickly making friends across the organization doesn't hurt either. If people see you and your team as returning strong dividends in the form of valuable insights from the data then, unsurprisingly, future data access requests tend to be much easier!

Generating Models and Conducting the Analyses

When many people talk about 'doing data science' they often mean building models or running the analyses. Those falsely believing that they can be 'specialists' fall into the trap of thinking they can just build models and let others worry about everything else.

As has been established so far, that's typically not a viable strategy either for successful data science efforts or longer-term career planning. That said, building high quality models and running impactful analyses is of course a core skillset for the full-stack data scientist.

If you're looking for a checklist of the core algorithms, packages, approaches and skills you need here to tick the box in this area you're about to be sorely disappointed. In fact, if someone tries to offer that level of specificity in defining what a good data scientist should know then you should probably be quite suspicious of such advice.

Rather, I'll simply just say this: You need to know what you need to know to get the job done. A full-stack data scientist that doesn't already have X skill never uses that as an excuse, rather they use the situation as an opportunity to quickly expand their skillset and apply new skills to the problem at hand.

There are plenty of high quality free resources available for filling those gaps as they are identified. Many early data scientists focus too heavily purely on technical skills. Maintaining a broader focus on the skillsets described here and developing an ability to quickly adapt to needs as they develop will set one up for sustainable long term success.

This is not all that unlike traditional science. Nobody can possibly know everything. However, a good scientist understands the broad art of the possible, is not afraid to push boundaries of their own knowledge, and can quickly learn new techniques and approaches when required to take the work in front of them to the next level.

To achieve this, the full-stack data scientist:

1. Has a solid understanding of what's happening across the market in terms of the tools, new packages and approaches (as is the case with traditional science fields if you're not spending at least 1 hour per day just reading and staying up to tune with the field you're probably behind in this area)

2. Can quickly imagine the art-of-the-possible by leveraging their knowledge of the field to identify creative approaches to the challenges before them—even if those approaches extend beyond the realm of their current experience or skills

3. Quickly close the gaps in their own skills to execute on the plan(s) developed

Story Communication and User Empathy

Good communication in data science often means taking a technical message and making it simple. That doesn't mean you need to 'dumb down' the message—it means you need to clearly communicate what matters to stakeholders.

Data scientists sometimes have a tendency to focus on the technical aspects of their approach. I once observed a data-driven senior executive tell a data science team "I know you're smart. I trust that you've got all the math right. What I need to know is what are your results telling me about how to run my operation differently?"

That's a good message to always keep in the back of your mind. Be prepared to back up the analysis with the mathematics if needed, but it's unlikely to be the thing to lead with. Communicate the message that your stakeholders need to hear, which is the message that is going to help them have an impact.

Successfully telling a good story, in data science or elsewhere, has a lot to do with the ability of the storyteller to empathize with their audience. What drives this person and their business? What are the things keeping them up at night? What generates value for them? Am I speaking to these things or am I just speaking to things that I myself find interesting?

Having empathy in a technical or business setting generally means you need to prepare extensively for future interactions. You're not just going to magically know the answers to the questions above, rather you need to research and explore these issues before key interactions. Someone with good empathy can read a person's verbal and non-verbal communication during a discussion but has also extensively prepared for that discussion. Winging it is not a great strategy here.

Having an Impact / Productionizing Output

So what? It's a question that data scientists often don't ask themselves enough when working with stakeholders. There's a time for general tinkering experimentation, but success in a business environment means having a measurable positive impact on the operation.

As discussed earlier, the target of the 'so what' needs to be discussed up front before a project begins. For example, decreasing the time it takes to complete a business process without impacting user experience or taking unnecessary risks.

In addition to analysis focused on the 'so what,' a full-stack data scientist also thinks about and can execute on a way to productionize their data science. In other words, how will stakeholders use the output of data science to actually have a positive impact?

Occasionally analysis can be run once and a single static review of the resulting output is sufficient to have the desired impact. More commonly the analysis needs to be seamlessly incorporated into day-to-day business practices to alter what humans and machines are doing and how they're doing it.

The full-stack data scientist is experienced at such analysis product creation for stakeholders and can produce, at a minimum, a working prototype of productionized analysis. Accomplishing that requires a general understanding of basic web frameworks, data pipeline automation and other back-end and front-end technologies. For example, transforming initial work in Jupyter Notebooks into an automated workflow and interactive front-end. Building such a prototype would require a basic working understanding of model-view-controller architectures to take static code and made it interactive.

The ability to build working prototypes leads to a far more productive relationship with application development teams. The days of simply handing off a requirements document to "the developers" and waiting for a product to arrive are over.

The full-stack data scientist also understands relevant security frameworks on the applications being used and is able to optimize prototype algorithms for production scale. On scaling, the full-stack data scientist thinks a lot like a traditional product engineer—figure out what you want something to be first, then figure out how to do it at scale. Starting the other way around tends to unnaturally restrict ones thinking to only the limits of previous production-scale achievements.

Such work often requires close interaction with engineering and application development teams, particularly if existing systems are to be altered or upgraded. The full-stack data scientist can work closely with such engineers. Similar to the earlier comments on working

with SMEs the objective is not to replace the engineers but rather know enough about the product architecture (or quickly learn such information) as to best seamlessly integrate the data science work in to data product work. The engineering teams should likewise make efforts to understand what the analysis aspect of products is doing and why those algorithms are important. Such knowledge is key to basic sense-checking of deliverables.

Wrapping Up

Bringing all the above together is hardly trivial and requires significant dedication and diligence to develop and maintain these skills. Every one of the essential full-stack data science skills identified is constantly evolving. It's simply not possible to be an expert at every single thing and even if you were today you'd quickly become stale without constant learning.

Thus, as has been drilled home several times in this piece, the key overarching skill is agility and an ability to quickly learn and apply new things. Doing so head-on will make one well placed to get and stay competitive in the exciting and ever-evolving world of data science.

About the Author

Dr. Nicholas Hartman is the Chief Innovation Officer CKM Analytix where he leads efforts to develop new products and approaches for applying advanced data science towards improving complex business operations. He combines a rigorous background in hard science with over a decade of operations management expertise to develop and deploy new ways to give data a voice to drive continuous improvement. Hartman is also a prolific writer and speaker on the ever-evolving field of data science.

Dr. Hartman holds a Ph.D. from the University of Cambridge, which he attended as a Marshall Scholar, and a B.S. from Penn State University where he recently received a 2018 Outstanding Alumni Award from the Eberly College of Science.