The State of Data Engineering

Executive Summary

Today, there are 6,500 people on LinkedIn who call themselves data engineers. In San Francisco alone, there are 6,600 job listings for this same title. The number of data engineers has doubled in the past year, but engineering leaders still find themselves faced with a significant shortage of data engineering talent.

The need for data talent is born from a fundamental shift: tech companies are now data companies. Uber, AirBnB, Spotify–these companies build data products, and as a result, are scrambling to hire (and hold onto) the people that build and maintain data systems. Josh Wills, Data Engineer at Slack, half-joked, half-pleaded at DataEngConf 2016, "Please don't hire my data engineers, they are all here now." Even Slack, one of the hottest tech companies in the valley, is worried about holding onto this valuable talent.

This is a challenging environment for engineering leaders. You are ultimately responsible for the success of your product, but to achieve that goal, you need to fight with hundreds of companies over the same set of talent. This report is your guide to understanding those highly sought-after individuals, and exactly why this shortage exists. In our research, we set out to discover:

- The number of data engineers in the market today
- Their backgrounds and core skills—information that is particularly valuable for leaders thinking about how to transition software engineers into data engineering roles.
- Employment information that can help you make the case for investing in this often expensive skill set.

Answers to these questions are paired with input from engineering leaders at Stripe, MIT, Looker, and more; who share their strategies for finding and retaining talent, developing data engineering talent in-house, and prioritizing a data engineering team's projects. This report presents a clear snapshot of the current state of data engineering.

Key Findings

Population

6,500 people call themselves "data engineers" on Linkedin.

Growth

The number of data engineers more than doubled from 2013-2015.

Whereabouts

50% of data engineers are located in the United States.

Previous Titles

42% of data engineers graduated from a Software Engineering role.

Industry

The Information Technology and Services industry employs the largest number of data Engineers.

Skills

The top five skills listed by data engineers are: SQL, Java, Python, Hadoop, and Linux. R isn't in the top 20.

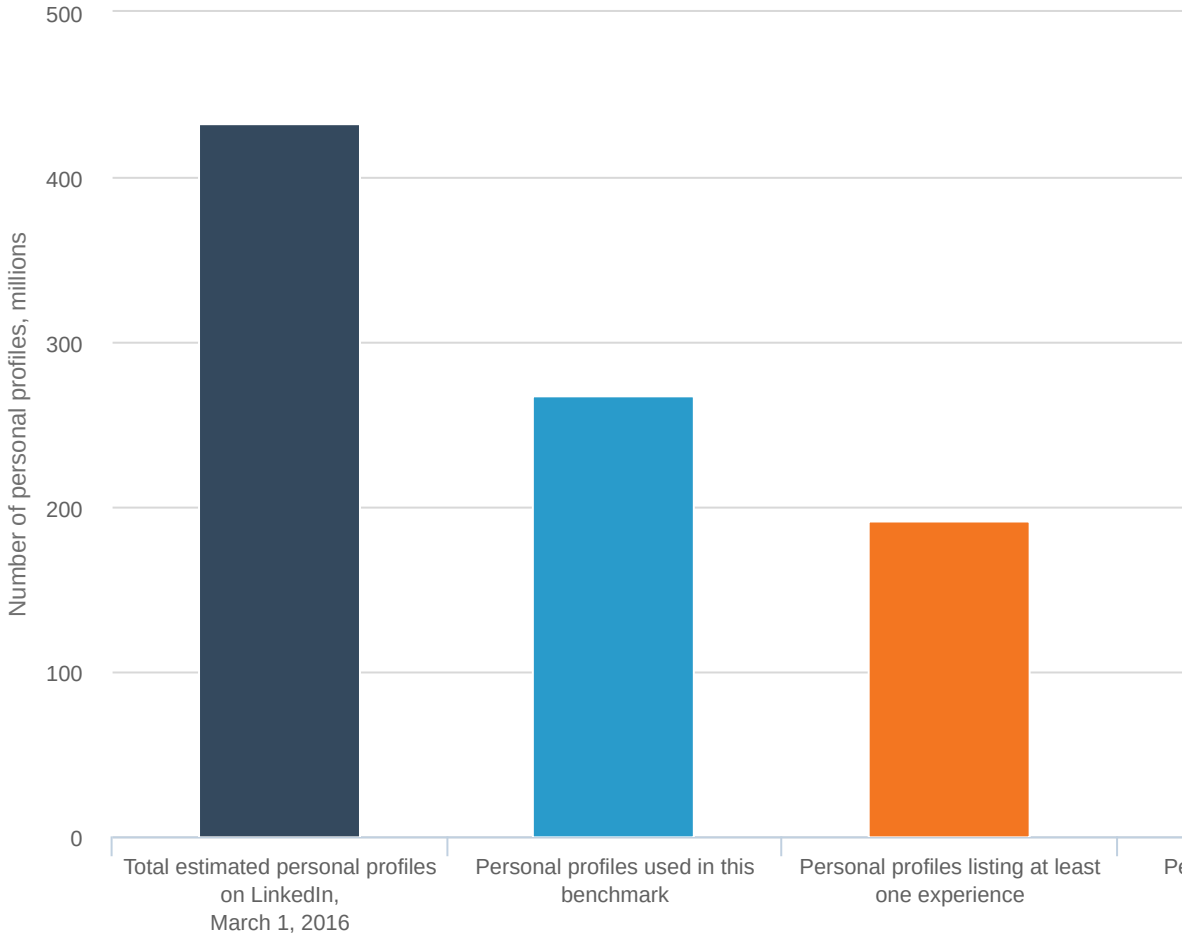
Methodology

This report is based on self-reported information from Linkedin, including all publicly visible personal and company profiles, skills, and professional experiences. The data is current as of March 2016.

We identified data engineers based on their professional headline and current title, and only included data engineers associated with identifiable companies

A summary of the dataset is provided in the chart below.

SUMMARY OF THE LINKEDIN DATASET USED IN THIS BENCHMARK



For the data engineers we identified, we analyzed:

- 30k professional experiences
- 82k experiences
- 3.4k companies

Analysis Tools

The analysis was carried out in Python, SQL, and Jupyter. Python packages charts and python-highcharts were used to create interactive visualizations in HighCharts and HighMaps. Data was stored and processed using Amazon Redshift.

Total Number of Data Engineers

How many data engineers are there?

It's easy to understand why data engineers are in such high demand; we're currently in the development phase of "the big data stack." There isn't consensus yet on how the stack will mature, and difficult technological problems arise at every turn. Because of this, it requires serious software engineering chops to build and deploy this technology today, and there just aren't a lot of people with these skills. Additionally, because these individuals are building the data infrastructure that companies like Uber, Spotify, and Slack rely on to deliver their products, the role couldn't be more critical.

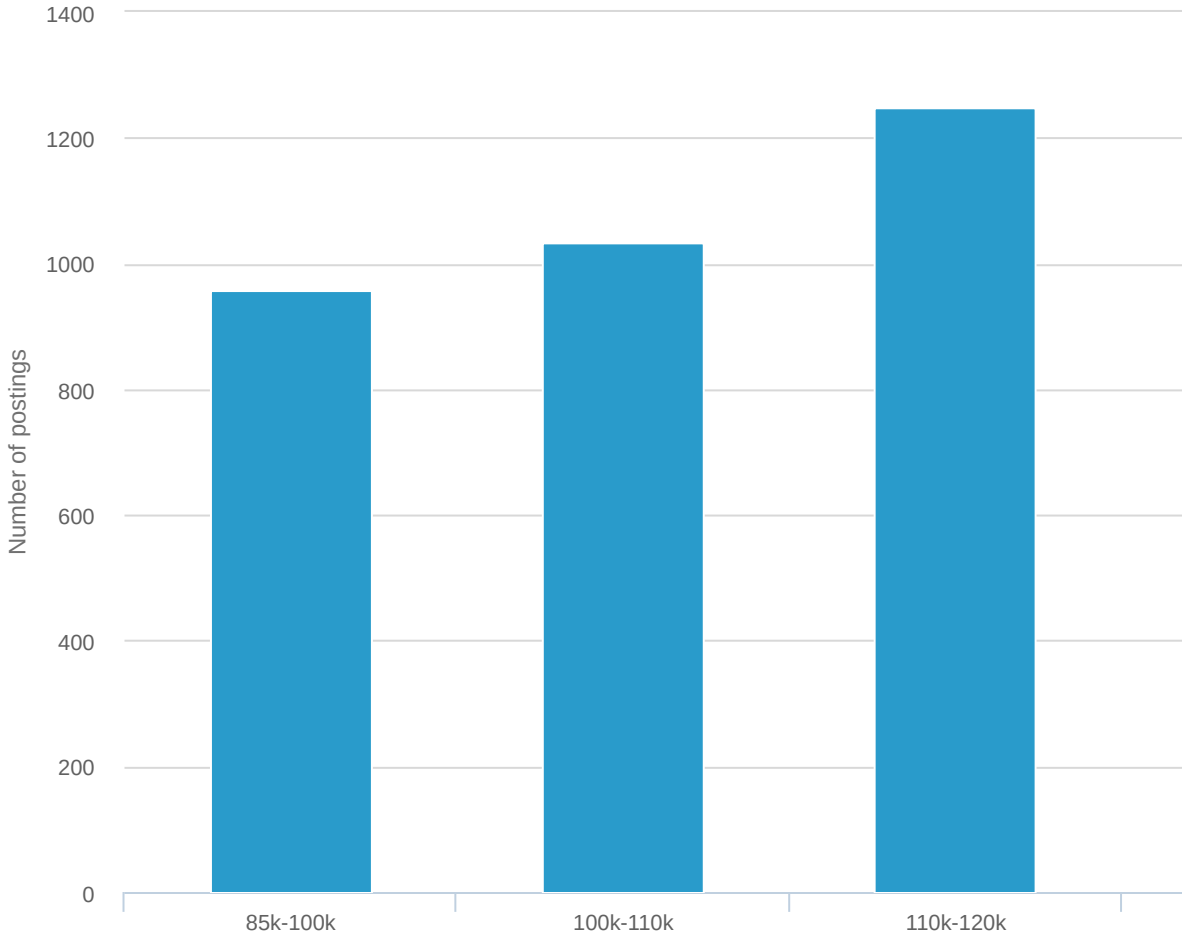
We found a grand total of 6,500 people who call themselves "data engineers" on Linkedin.

We have no doubt that plenty of folks are doing the work of the data engineer who aren't using this title, but in this report we focus specifically on people who self-report having this title. There's plenty of potential fuzziness around the definition of "data engineer" (all software engineers work with data in some fashion!) and we don't think there's a perfect answer. We felt it was best to let the practitioners speak for themselves.

6,500 is not a big number. In fact, we were a little surprised at just how small it is. For comparison: as of this writing, there are 6,600 data engineering job postings on Indeed. And that's just in the San Francisco Bay area.

Salary data also confirms that data engineers are in demand. Anecdotally, top data engineering positions at tech giants like Facebook, Amazon, and Google can exceed $500k. Indeed's data shows a more modest distribution, but salaries well into the six figures none-the-less:

NUMBER OF DATA ENGINEER JOB POSTINGS IN SAN FRANCISCO VS. SALARY RANGE



Expert Insight:

The Demand for Data Engineering Talent

Jonathan Coveney, Data Engineer at Stripe

For nearly a decade, Jonathan Coveney has been deep in the data world, building data systems at Twitter, Spotify, and even doing a stint at The Apache Software Foundation before joining Stripe. From his perspective, there are three important trends driving up the demand for data engineering talent:

- New sophistication in how companies think about data and the people who manage it. "There's a growing sense today that data isn't just a byproduct," Jonathan says, "but rather it's a core part of what a company does." In the past, a

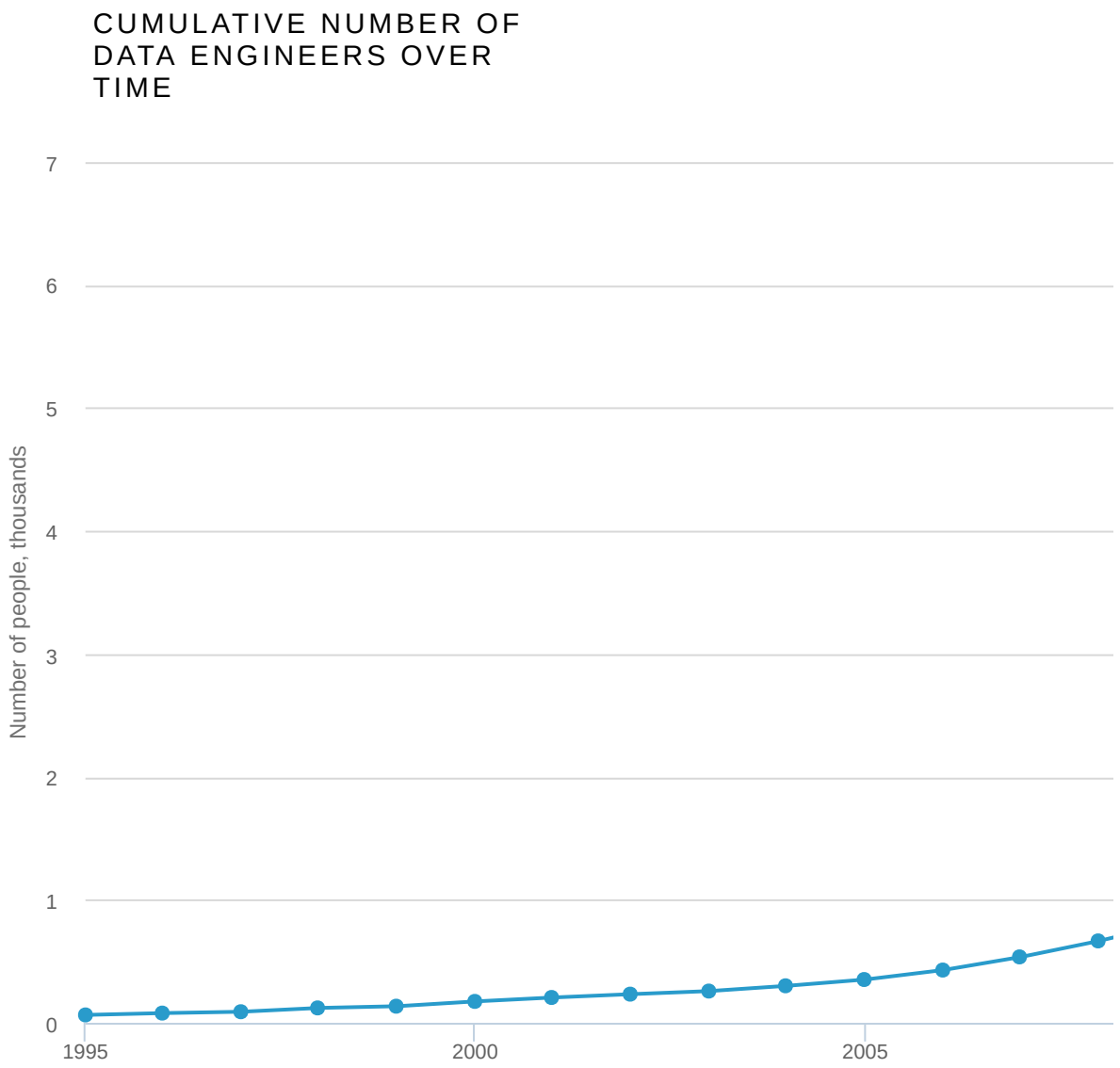
company might run a one-off analysis of a database log; today, data engineers own the analysis and organization of that data. They identify with data processes, which leads to a more nuanced understanding of their data architecture.

- Push towards machine learning. Thanks to machine learning advances, access to proprietary data has become a major competitive advantage for firms in almost all industries. Collecting and making this data available has therefore become a key strategic function.

- Companies building data products. There's some overlap here with machine learning, but Jonathan uses the example of maps to describe the difference: "The machine learning aspects of maps include things like traffic detection and routing, but the infrastructure of maps relies on managing and organizing massive volumes of data–that's data engineering." Today's tech companies need to play well in both areas.

How has the number of data engineers changed over time?

Linkedin profiles show an individual's self-reported employment history as a list of titles with start and end dates. This information allows us to construct a timeline of the job market. Take a look at the chart below; it's hard to overstate just how quickly this space is growing.



The number of data engineers more than doubled from 2013-2015. And based on the job posting data from earlier, this growth isn't about to slow down.

For comparison, there are currently about 2x the number of data scientists (roughly 11,400), but the growth rate of data engineers is much faster than anything the data scientist job market ever experienced: In this same period, the number of data scientists grew by a little over 50%.

This is particularly interesting when you consider the saturated press around data scientist hiring. The feature-length article on data engineers has yet to be published.

Where do data engineers come from?

The rapid influx of data engineers begs an obvious question: who are these people? What did they do beforehand? We looked at the data, asking the specific question "What was the job title held by this person just prior to them taking their first role as a data engineer?" This is instructive in that it tells us about the DNA of data engineers.

We had a few theories of what we would find when we looked at this question:

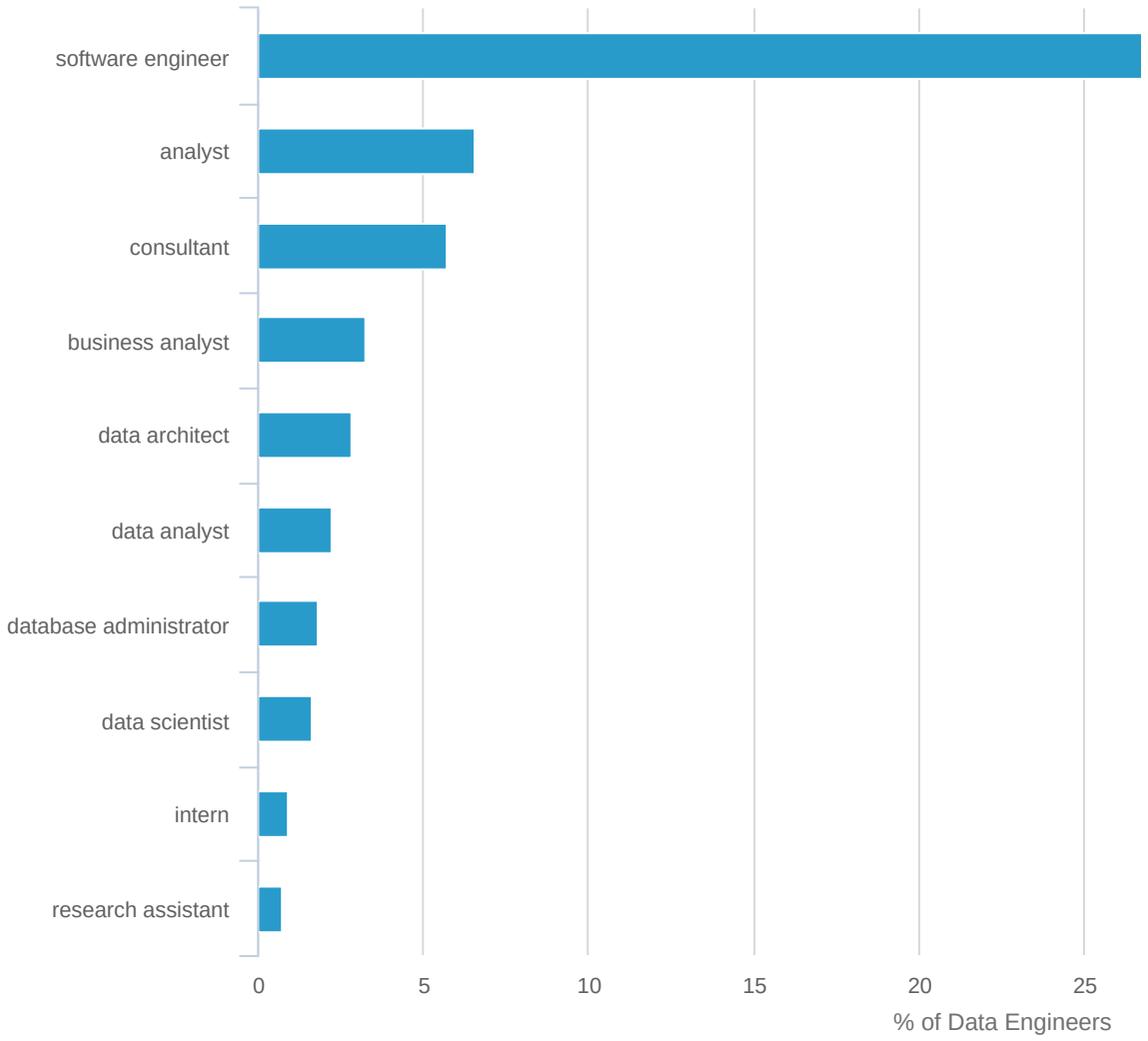
- Data engineers bridge the boundary between software engineering and data science in that they create the production code that allows data science to scale reliably. We expected to see both software engineering and data science represented in the data.

- Because so much of data engineering is about scale, data engineers bridge the gap between software engineering and devops. Because of this we anticipated seeing some devops-specific titles.

- Database administrators have historically played a very similar role within companies. We anticipated seeing some DBAs who have transitioned to this more modern role.

We found that our three hypotheses played out to some extent, but one thing was very clear: data engineers share most of their DNA with software engineers. Here are their top ten prior job titles:

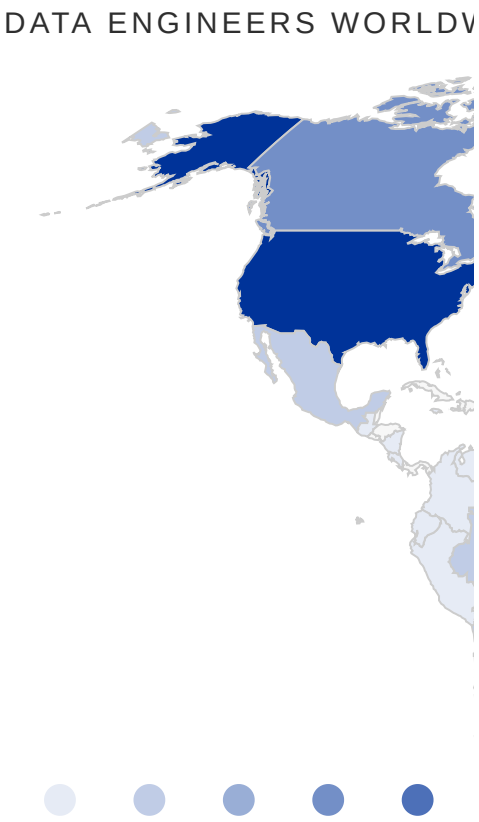
DATA ENGINEERS BY PRIOR ROLE, TOP 10



This makes intuitive sense: data engineering is a subspecialty of software engineering. The two fields share methodology and tools. While individuals from other disciplines do transition into the role, the most common path starts at the more general "Software Engineer" title, and progresses to the more specialized "Data Engineer".

Where are data engineers located?

50% of all data engineers live in the US. This isn't entirely surprising, as the term itself and much of the foundational technology comes from technology companies and universities in America.

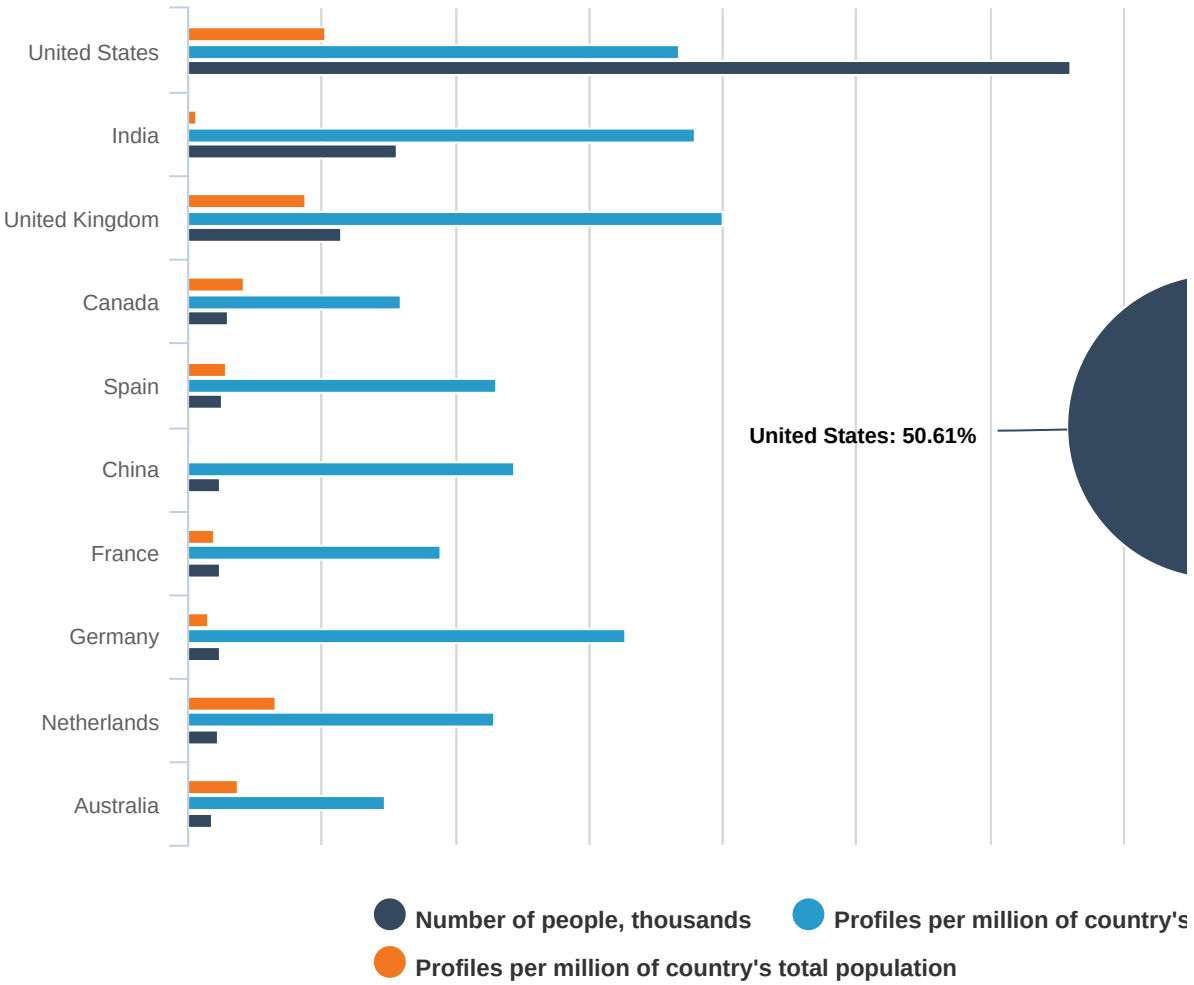


This is interesting particularly because it validates conventional wisdom within the data engineering field. Most of the space's technology has either come out of a small set of universities—most especially Berkeley—or from the software engineering teams of the biggest internet companies in the world. Google, Facebook, Linkedin, and Amazon were struggling with big data and had resources to throw at the problem long before the rest of the industry. Not only have they invented much of the technology, they've also acted as training grounds for talent.

However, this chart is slightly misleading. While the US has the most data engineers by far, they also have the most profiles in the world: nearly 4x that of the next country, India.

To normalize the data, we broke out the top ten countries from the chart above and looked at how their data engineer population relates to the number of LinkedIn profiles from that country, as well as the population as a whole.

DATA ENGINEERS PER COUNTRY



Missing from this list is Israel, which in our previous benchmark, ranked highest in terms of data scientists per million of their population. As we mentioned, Israel has long been known as a startup nation with a strong tech presence in "Silicon Wadi." It's surprising that this doesn't translate to a higher density of data engineering talent.

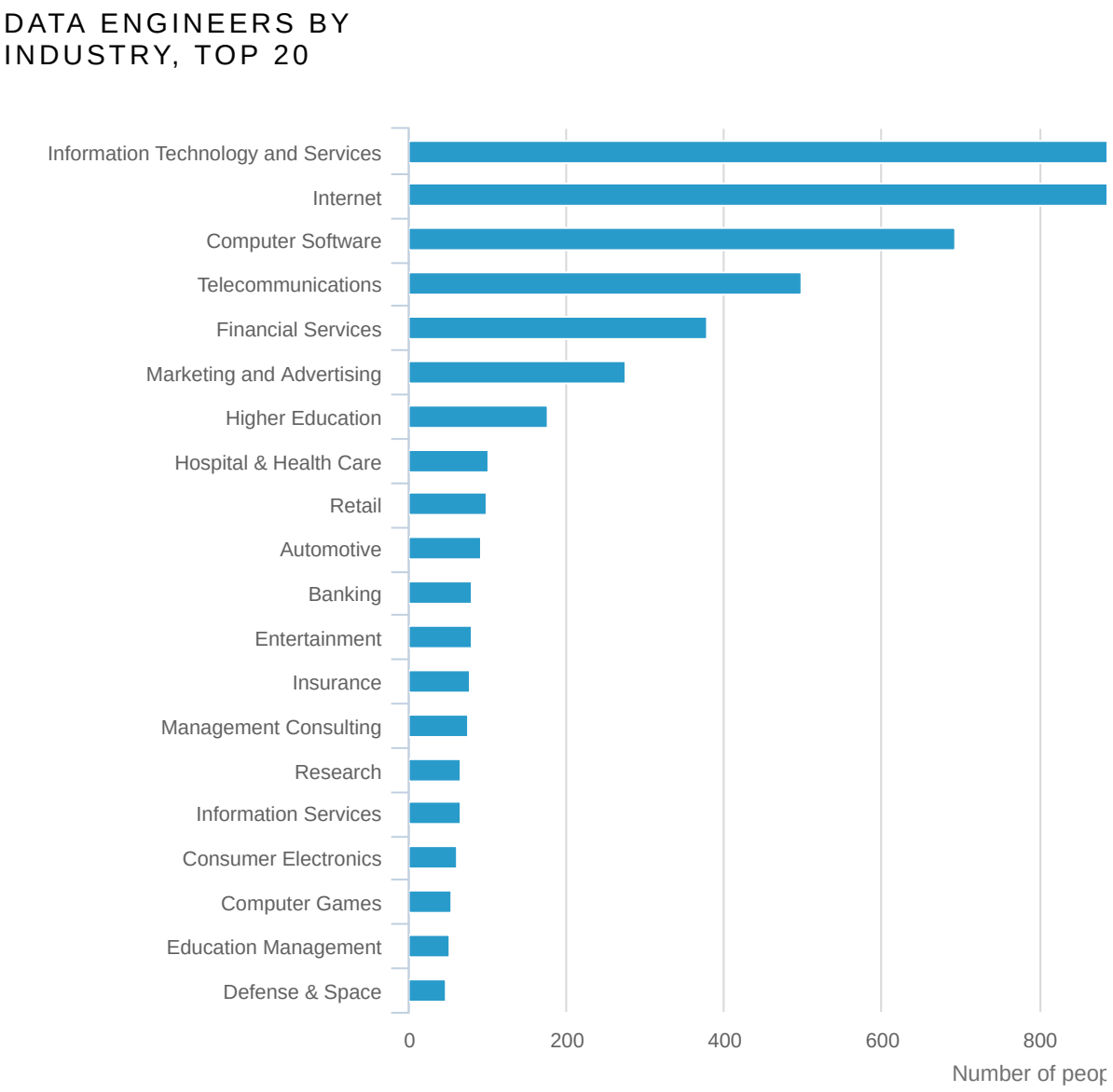
Top Employers of Data Engineers

What industries employ the most data engineers?

Companies that experience challenges related to scaling the storage, transmission, and processing of data are those in need of data engineering talent. These challenges arise mostly

within tech companies, but what about industries like telecom, biotech, and insurance? Don't these industries need data scaling help as well?

When we looked at where these data engineers are working, we found that a wide range of industries require a data role.



Telecom and financial services are up towards the top, as we expected, but the petabytes of DNA being sequenced in biotech today don't seem to be pushing it towards the top of the list.

The takeaway from this chart shouldn't be that other industries don't need or don't employ people who function as data engineers. Rather, the title "Data Engineer" has been popularized within a certain industry—internet tech—and the usage of this particular title is still nascent. The technology, process, and mindset within this space is beginning to spread to other industries.

Expert Insight:

Data Engineering in Healthcare

Alyssa Kwan, Data Engineer at Clara Lending

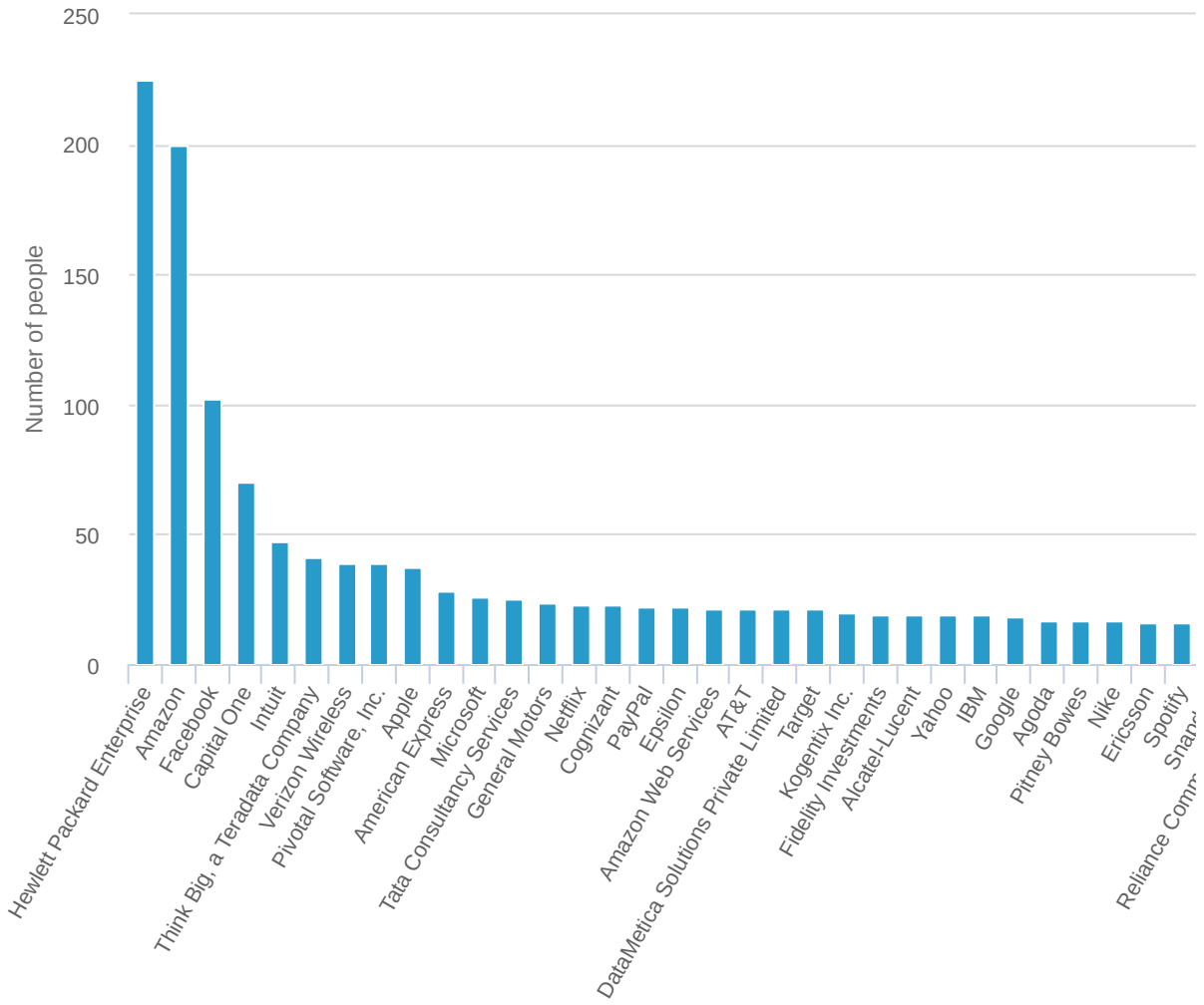
Alyssa Kwan has built data infrastructure in a variety of industries including Financial Services at Citizen's Bank and Clara Lending, Marketing and Advertising at Nanigans, and healthcare at Clover Health. She's seen first-hand how much the work of data engineering changes depending on the industry one is working in. "For most tech companies, the primary concerns in data engineering are around issues of scale. But in industries like healthcare, your primary concern is complexity. You need detailed models that conform to both compliance and auditability requirements. And then there's the human perspective to that industry as well–you can't risk messing up someone's healthcare."

This has a big impact on the specific data engineering skills companies hire for. "If you're doing web analytics, in the case of marketing and advertising platforms, you want a developer experienced working with off-the-shelf tools, likely a DBA type person. The model for this role is more set." But in more complex industries like Financial Services and Healthcare? "You're looking for someone that's more of a classic computer science person. There are very few organizations in protected industries that have done this work or are talking about it. So you want someone that's interested in the space, a generalist that is comfortable with the amount of complexity."

What companies employ the most data engineers?

The popularity of data engineers in tech becomes even more clear when looking at companies employing these data engineers. Within the top ten companies there are only two companies not specifically in technology or data: a telecom company (Verizon) and a financial institution (Capital One).

DATA ENGINEERS PER COMPANY, TOP 50



It's interesting to pick out companies who employ a disproportionate number of data engineers. For example, Spotify (1600+ employees) is far smaller than Pitney Bowes (16k employees), but employs roughly the same number of data engineers.

The data clearly shows that some of today's tech "unicorns" value the data engineer role very highly. And, considering that there are 6,600 companies in San Francisco currently looking to hire a data engineer, it doesn't seem like this is about to change in the very near future.

Expert Advice:

Recruiting & Retaining Data Engineers

Maggie Xiong, Director of Data Engineering at The Huffington Post

Maggie has been working with data for over a decade. In that time, she's learned some important lessons about how to recruit and retain the engineering talent needed to build data technology. She points to three factors:

- Build an enviable team. "I hired top-notch architects on my team early on, and that made recruiting easier. Smart data engineers want to work with people they can learn from."

- Remove the wall between data engineering and data science. "A lot of companies separate the data platform team from the data science team. I took the job at Huffington Post because they gave me the authorization to remove that wall. For algorithms-minded folks, they want to see their work in production. They're not thinking about monitoring, recovery scenarios, documentation. A data engineer brings that rigor to data science that makes both groups successful. When teams are working separately, a lot of each of their work never sees the light of day."

- Create an environment of creativity. "Big data people on both sides of the equation – systems and algorithms – tend to be highly creative. New technologies and approaches are coming out every month, people want the space to explore and learn."

Skillset of Data Engineers

We've gotten to know a lot about data engineers at this point, but what, exactly, do they do?

Earlier in the report we editorialized on this topic. The common understanding of the role of the data engineer is two-fold:

- Make data available to consumers throughout the business

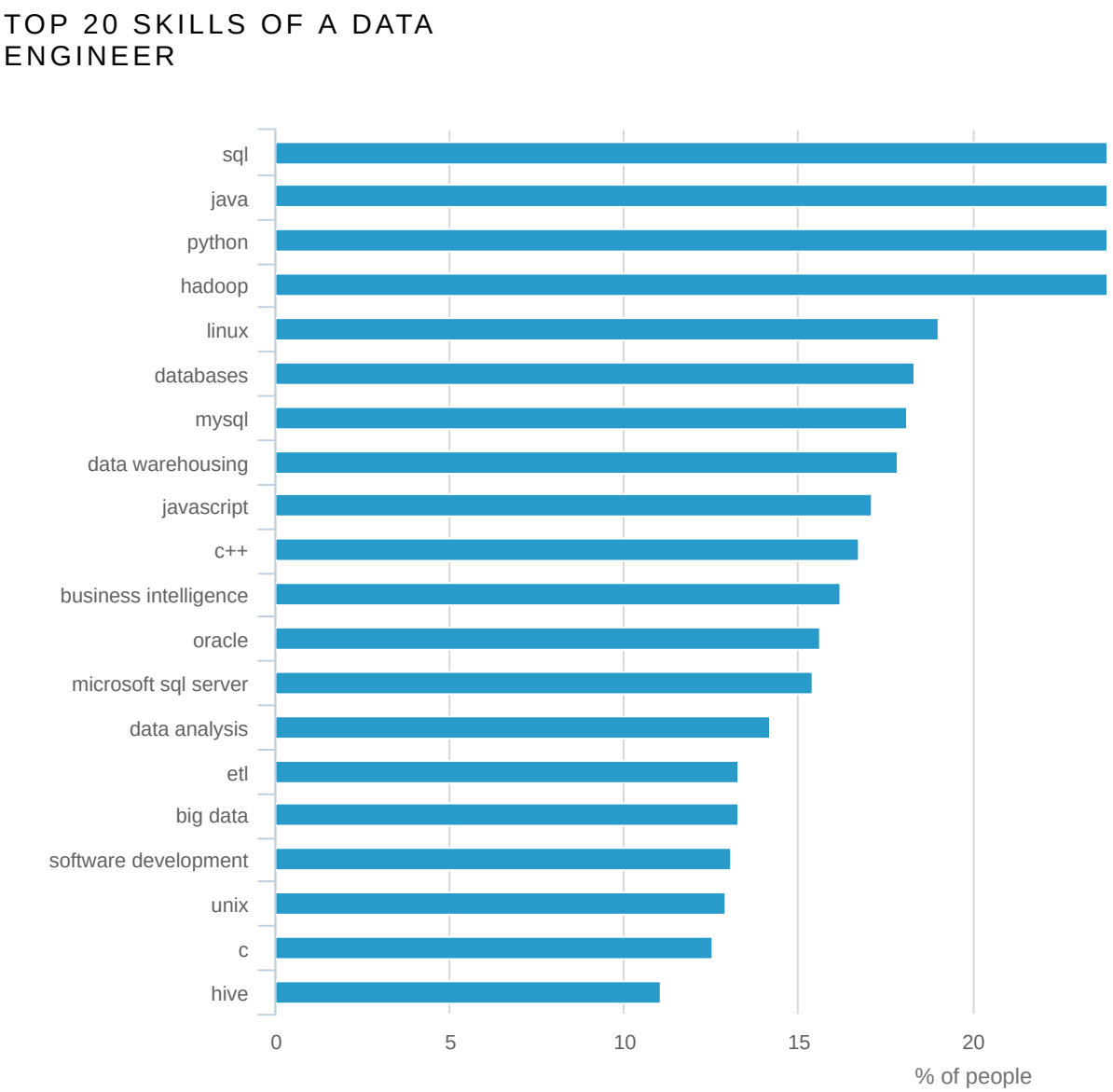
- "Production-ize" algorithms that can be turned into data products.

While this seems like a fair assessment of the role, we'd prefer to let data engineers speak for themselves. Fortunately, Linkedin profiles have an entire section devoted to skills, which can

say a lot about a person's role. While this section is often the least well-maintained of the sections of a profile, we're confident that you'll find the conclusions that can be drawn from this data extremely interesting.

What are the primary skills of a data engineer?

The skillset of a data engineer obviously trends heavily towards data, while keeping some of the core software skills that many developed in prior roles. Take a look at the top 20:



There are three specific things that we find notable on this list:

- SQL, a declarative language that most software engineers think of as little more than something to wrap up in an ORM, is the most common skill for data engineers. This is big.

For years, SQL was a bit of an ugly duckling within data tech with the ascendancy of NoSQL approaches. However, SQL isn't going anywhere—in fact, it's enjoying a renewal as SQL-based interfaces for unstructured data (e.g. Impala, Drill, Hive, and Presto) increase in popularity.

- Java is by far the most popular programming language on the list. This makes complete sense: since the original founding of Hadoop in the mid-2000's, the JVM has been at the heart of data processing.

- Python is extremely common, but R doesn't even make the list. Python is frequently used both for data engineering workloads as well as analytics workloads, whereas R is specifically used for analytics. Within the data science community, both of these skills are roughly equal in weight. The difference in these two populations is striking.

Beyond that, there's a tight focus on the highly-relevant technical skills needed to work with data.

Expert Insight:

Know the Type of Data Engineer You Need

Mike Xu, Data Architect at Looker

Part of Mike Xu's job at Looker involves developer relations, which means he gets to hear quite a bit about what makes data engineers both happy, and not so happy. One of the most common complaints? Companies don't understand the nuances between data engineering roles.

For the most part, Mike sees data engineers fall into four buckets:

- Warehouses: Focuses on optimizing warehouses for analytics, writing and managing data transformations

- Tools: Highly skilled at using a specific tool in the big data toolbox

- Architecture: Talented end-to-end thinkers, planning everything from data collection, to how teams will use the data

- Ops: Focuses on building database and data tool instances, and managing permissions and security

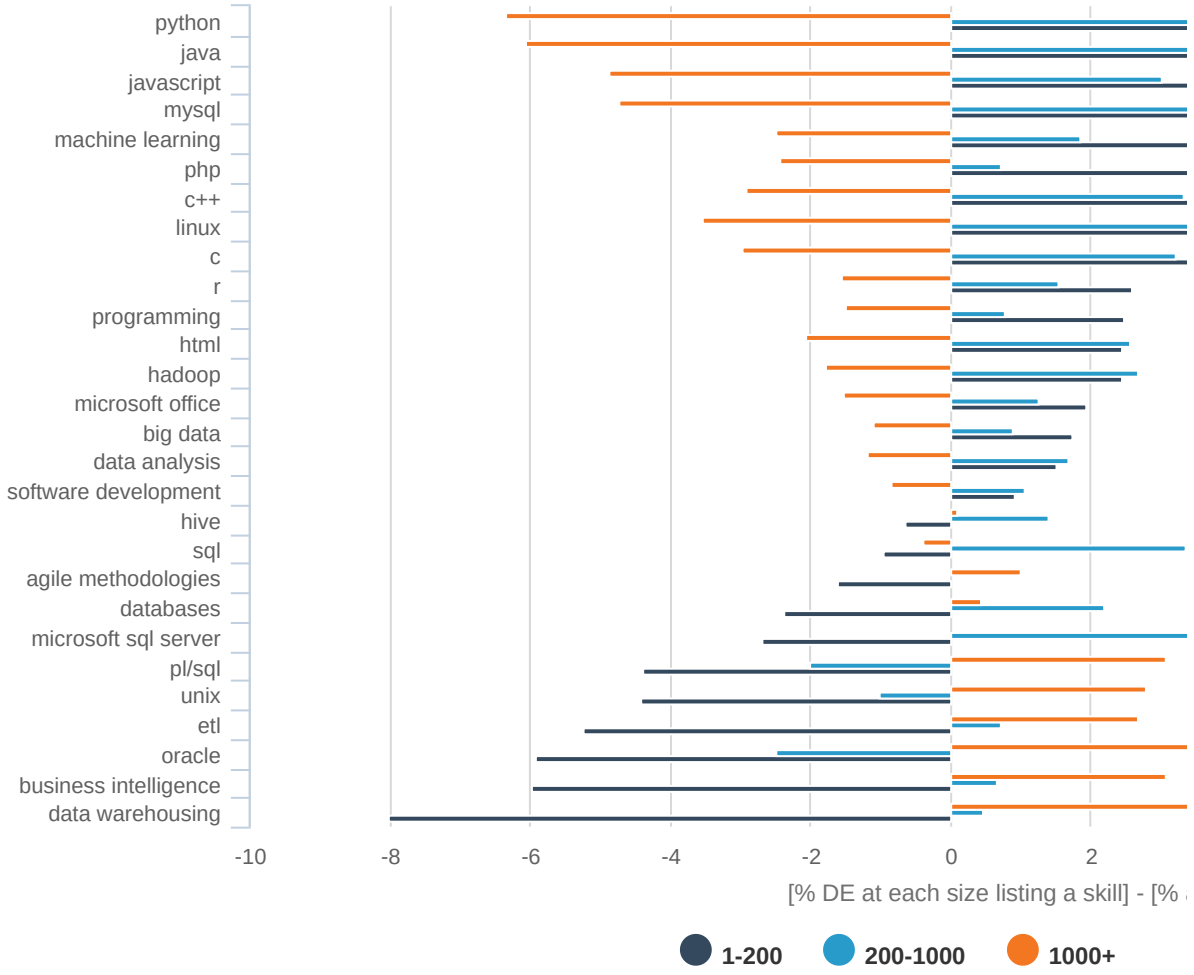
This distinction is particularly important during the hiring process. "A lot of data engineering candidates I talk to are frustrated by the lack of insight into the projects they would be working on. Without that information, they're unable to evaluate if the role aligns with the nuance of their skill set." If you want to attract data engineering talent in this market, make it clear what type of data engineer you need.

How do the skills of data engineers change with company size?

It's a core function of data engineers to deal with the scalability challenges that arise with increases in dataset size. As such, we thought it would be instructive to look at how skills changed with company size, given that larger companies will often have more data.

The chart below shows the relative difference in prevalence of skills based on the size of the company employing the data engineer. Skills at the top are more prevalent with data engineers at small companies; skills at the bottom are more prevalent in companies with 1,000 or more employees.

SKILL DIFFERENCES OF DATA ENGINEERS ACROSS COMPANY SIZES



We anticipated that as company size increased, so would the focus on scaling-related skill. However, that's not the story the data told. Instead, data engineers at larger companies tend to be more focused on "enterprise" skills like ETL, BI, and data warehousing, whereas data engineers at smaller companies focus more on core technologies.

Expert Insight:

Data Engineering in the Enterprise vs. at Startups

Will Smith, Principal Data Engineer/Architect at MIT

Will Smith has built data technology at companies like Nokia and Warner Bros Games, and in his experience, the skill set of a data engineer has less to do with whether

they're working for an enterprise company or startup, and much more about how they answer one question: is their data schema-on-write, or schema-on-read?

"Enterprise companies are experienced working with Informatica, Oracle, SAP–the business intelligence (BI) side of data engineering," Will says. "In these environments, there's a team that has defined a schema of what the data will look like: schema-on-write. BI teams know exactly how the data will look in the warehouse."

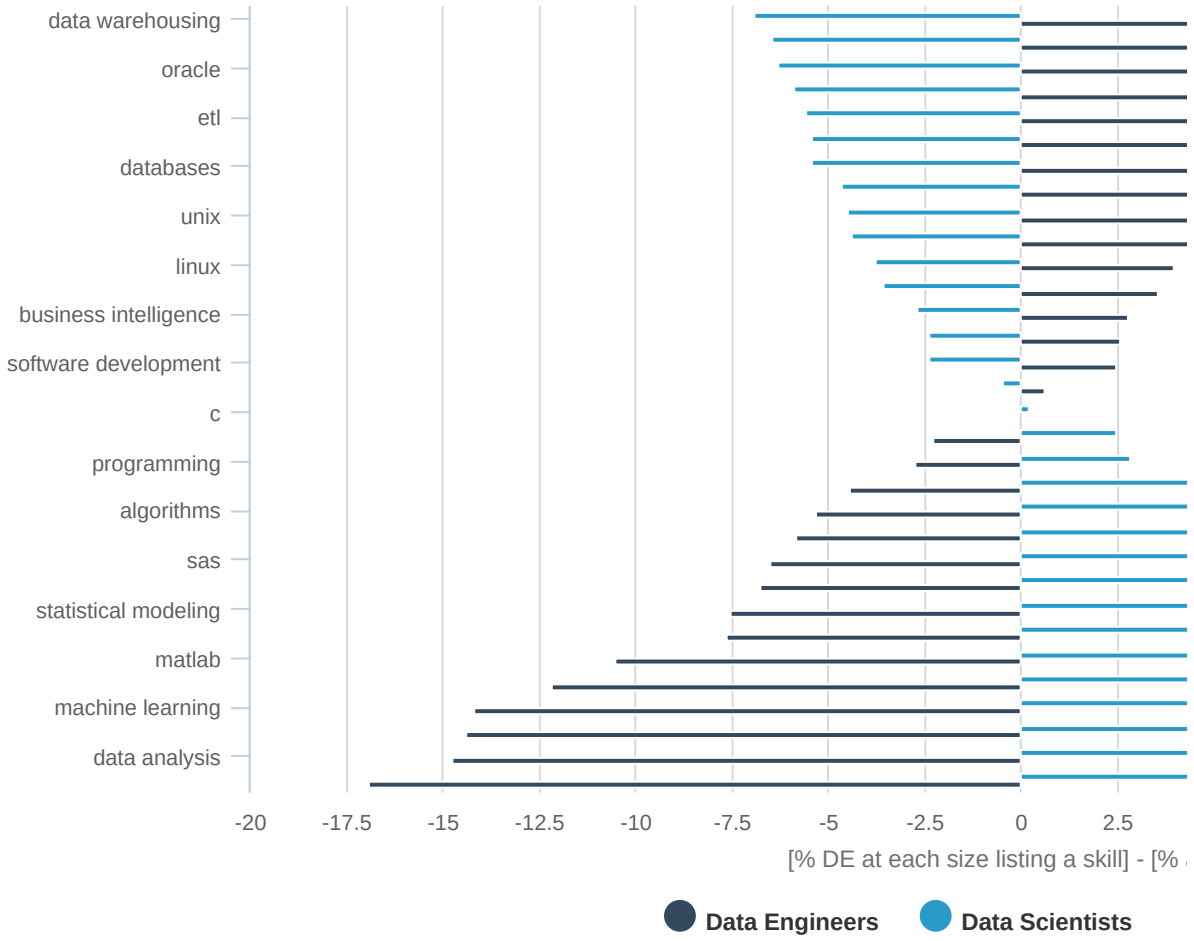
But companies building data technology today are functioning in a schema-on-read environment. Will shares this example: "Imagine you have terabytes of log data from ad impressions in JSON. The data engineer has no idea what they will find in that data. The skillset now requires the developer to do data discovery and develop code, rather than just using straight SQL. This is a very different skill set than is needed in the schema-on-write environment."

Will saw this first-hand while working on Nokia's big data initiatives in 2011. While working with enterprise-size data, the design pattern they used (learned from ex-Yahoo developers) was built for schema-on-read, an approach more typically adopted by smaller companies. "Building this way is how today's data engineers get economies of scale. They can ingest data from any source. Legacy BI systems can't do that. You point them at a data lake and, because of schema-on-write, they can't do anything with it because they don't know what's in there and we don't know what we don't know about the data."

How do the skills of data engineers and data scientists compare?

With this dataset, we're able to compare the skills of data engineers vs. those of data scientists. And the data paints a very clear difference between the two roles. Think of data engineers and data scientists appearing at opposites sides of a spectrum. This chart shows where skills are on that spectrum, with the the top representing skills more prevalent on data engineer profiles, and the bottom highlighting skills reported mostly by data scientists.

SKILL DIFFERENCES BETWEEN DATA ENGINEERS AND DATA SCIENTISTS



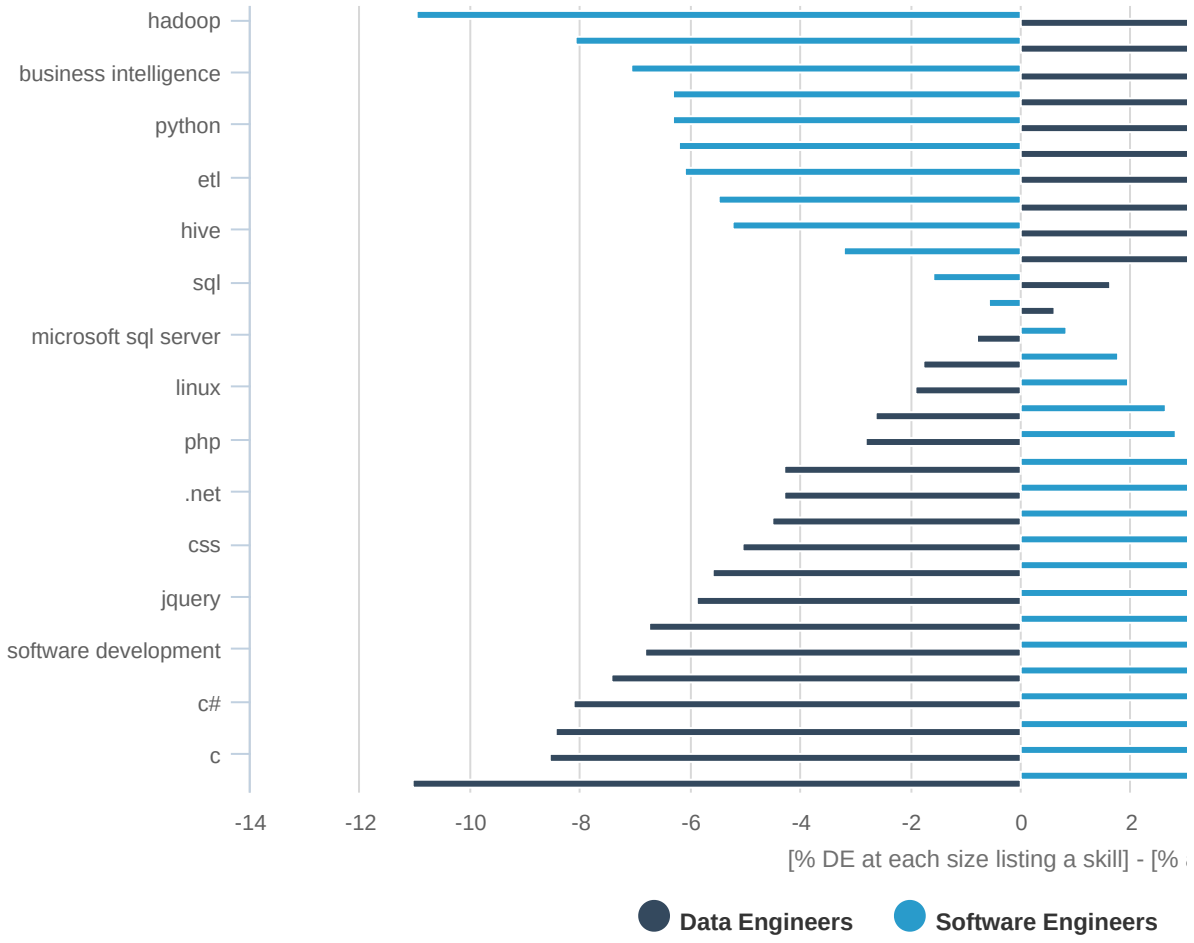
Data engineers focus on making data available and processing it in production environments, which explains why data warehousing appears at the very top, followed closely by Java, the language often used to productionize algorithms.

Data scientists, on the other hand, focus at the top of the stack. As we alluded to earlier, usage of R is a huge difference between these two groups, but that's followed very closely by other analytical skills like machine learning, stats, and modeling.

How do the skills of data engineers and software engineers compare?

The difference between data engineers and data scientists is clear, but what about data engineers and their software engineering compatriots? After all, as we showed earlier, a plurality of data engineers come from a software engineering background.

SKILL DIFFERENCES BETWEEN DATA ENGINEERS AND SOFTWARE ENGINEERS



The skills that are the most data engineer-centric are Hadoop, data warehousing, and BI—exactly what you would expect. And conversely, almost all of the skills listed on the software engineer end of the spectrum are focused on front end web development. The two biggest exceptions to that, C and C++, are both languages not commonly used in the modern big data stack.

While many data engineers may come from a software engineer role, they haven't simply changed to a trendy new job title for a pay raise; they've had to differentiate themselves by learning new skills along the way.

Expert Insight:

Building Better Relationships Between Data Engineers and Data Scientists

Ryan Orban, Chief Technology Officer at Galvanize

"Think about the relationship between designers and front-end developers," Ryan Orban says. "One comes up with the ideas, the other implements. And it can cause a lot of tension." The relationship between data engineers and data scientists is similar in Ryan's opinion, and so is the solution to easing that tension. "Just like designers are always told to learn to write some code, and front-end developers are told to get more comfortable building mockups, I encourage data scientists and data engineers to learn some subset of their counterparts' skills."

So just how deep should a data engineer go into the data science world? "Data engineers should have a basic understanding of machine learning," Ryan says. "They don't need to know all the mathematical theory, but they should be able to judge efficacy and accuracy. Conversely, data scientists need to understand infrastructure, how things scale, and have a rudimentary understanding of production level languages."

This shift toward going deeper into the expertise of the other's discipline is happening in other areas as well. Companies used to hire data scientists that would report into marketing, product, or business analytics, while data engineers would report into the broader engineering function. This creates a misalignment of goals. Ryan says this trend is changing: "The concept of the ‘data team' a group of both data scientists and data engineers, is growing in popularity. Just this simple change does a lot to improve the relationship between these two teams."

The Future of Data Engineering

Asim Jalis
Principal Data Engineer and Lead Instructor, Galvanize

As software continues to eat the world, businesses looking to be a part of that revolution will need to hire data engineers. The companies today that are already employing data engineers

have realized data's potential as a strategic asset, and as others follow suit, demand for this skill set will only increase. During this talent shortage, many will begin looking for software developers to step into this role. However, there are a few good reasons to be cautious about that role change:

1. Software developers are skilled in their approach to special cases, building algorithms and writing elaborate code to handle outlying problems, but big data development requires a more holistic understanding and concern with the data architecture in its entirety.

2. Beyond the nuanced differences in the algorithms the two roles use, their style of programming is also different. While traditional programs are sequential and keep their current state, data programs are massively parallel and distribute their state across hundreds or thousands of machines.

3. Data engineers need to be familiar with how to integrate machine learning algorithms into their applications.

Software developers are certainly capable of transferring into data engineers. However, those individuals must develop an understanding of new mental models, as well as new ways of thinking, before they can work effectively with big data. At Galvanize, we're seeing more and more companies make an effort to promote these role changes by sending their software developers through our data engineering program.

The practice of data engineering will continue to specialize over the coming years, and with it, an increase in the capabilities of what companies can build and accomplish with their own data. I couldn't be more excited to see how these trends play out over the coming years.

The Learning Community For Technology

Galvanize brings together education, networking and workspace in 9 state-of-the art campuses across the U.S. This unique community cultivates collaboration and innovation, and propels students, startups and entrepreneurs towards tech-industry success.

LEARN MORE ➔

ETL Service Built for Developers

Stitch connects to your first-party data sources – from databases like MongoDB and MySQL, to SaaS tools like Salesforce and Facebook Ads – and streams that data to your warehouse. Create your account and get 5 million rows for free.

GET STARTED ➔

PRODUCT

Extensibility

Orchestration

Security

Performance

Embedding

Pricing

Compare ETL tools

INTEGRATIONS

Sources

Destinations

Analysis Tools

LEARN

Documentation

Community

Blog

Changelog

Resources

GitHub

Status

COMPANY

About

Careers

Contact

Customers

Partners

SOCIAL

Twitter

LinkedIn

Facebook

Instagram

© 2019 Stitch, Inc.

Terms of Use

Privacy Policy