CloudAcademy

Cloud Academy December 2018 Data Report

DATA ENGINEERING EDITION:

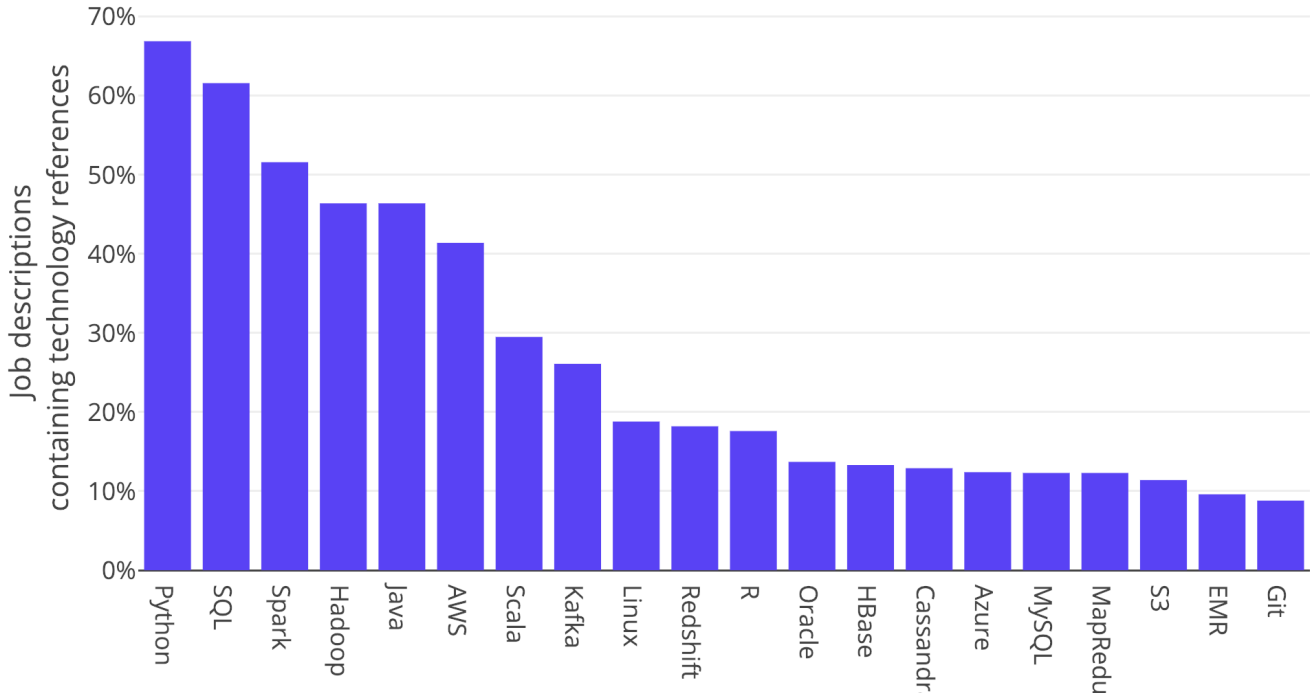
Introduction

In this month's data report, we'll start to explore some data that helps us unpack the world of data engineering and the role of the data engineer. We wanted to examine the role because there's a sense in the market that many companies hire data scientists but lack the appropriate infrastructure to allow them to do their jobs. Almost 50% of data scientists surveyed in 2017 by Kaggle reference dirty data as a barrier to their work. Across every vertical, Cloud Academy sees enterprise organizations accelerating their investment in big data processing and analytics — including in highly regulated industries like finance and healthcare. In short, this is an important role with important implications for the companies investing in the space.

What is a Data Engineer?

Data Engineers are responsible for big data pipelines. They build, test, manage, and prepare the databases and the scalable data processing systems that data scientists need to do their jobs and that businesses, more generally, need to scale. The data engineering job function increasingly serves to link the work of developers to that of data scientists and vice-versa. You can read more about the role of the data engineer on the Cloud Roster.

The Top 20 Technical Skills in Demand for Data Engineers

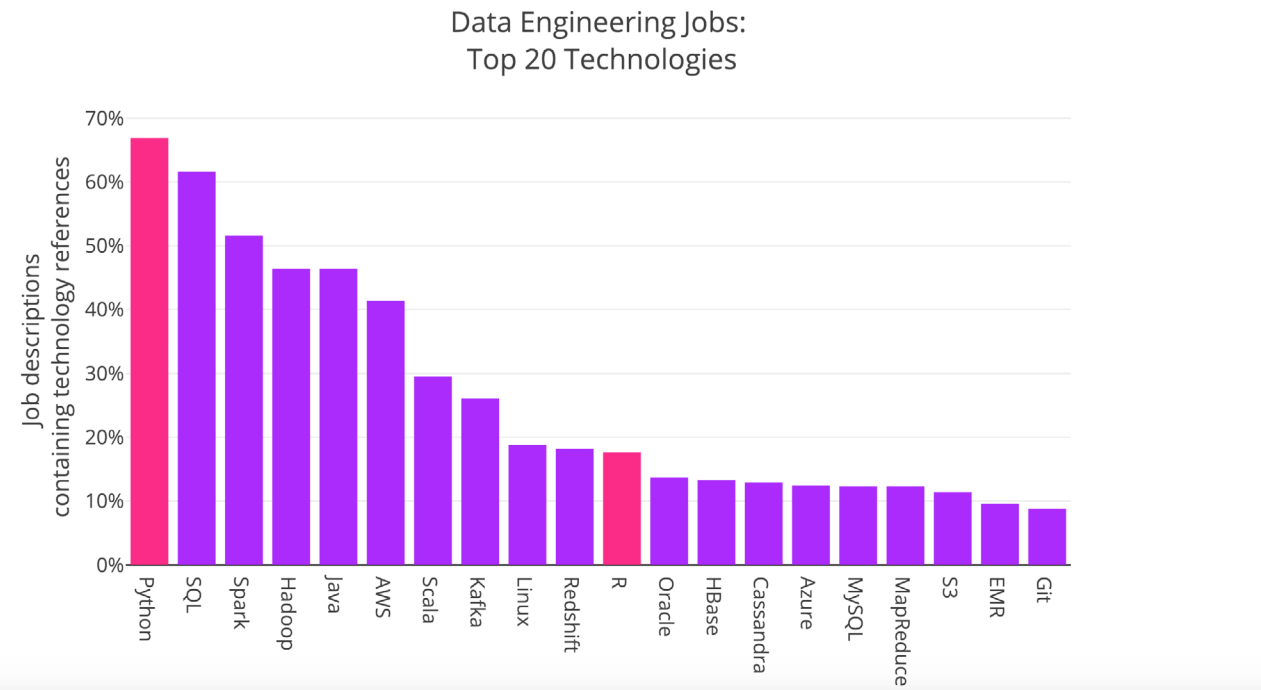


What are the top technologies that are in demand for data engineers? As of December 2018, the top skills in demand and the percentages of job posts that mention them are:

Rank	Skill	% of Posts	Rank	Skill	% of Posts
1	Python	67%	11	R	18%
2	SQL	62%	12	Oracle	14%
3	Spark	52%	13	HBase	13%
4	Hadoop	46%	14	Cassandra	13%
5	Java	46%	15	Azure	12%
6	AWS	41%	16	MapReduce	12%
7	Scala	30%	17	MySQL	12%
8	Kafka	26%	18	S3	11%
9	Linux	19%	19	EMR	10%
10	RedShift	18%	20	C++	9%

Source / Latest Data: Cloud Catalog

Python vs. R: Python is the language of data science, not R



R has the reputation as being primarily a programming language of academics and mathematicians. Public resources are conflicting on its popularity. The Tiobe Index shows an overall downward trend in search engine requests for R popularity, but the methodology doesn't take job function into consideration, which we believe is important. (Scroll to the bottom of this post for more information about our methodology).

The Kaggle 2017 report surveyed over 16,000 data professionals. It found that, while Python may be the most common language overall, statisticians and data science professionals were more likely to report using R at work than other roles. Among data scientists, 86.5% reported using Python and 70.6% reported using R at work.

Among engineers, 81.1% reported using Python and 46% responded that they use R at work.Our analysis indicates significantly less demand for data engineers proficient in R compared to those proficient in Python. **Of the data engineer job posts in our data warehouse, 65.6% mentioned Python — a high degree of skill relevancy. Just 17.6% of data engineer job posts mentioned R as relevant to the role.**

Python is known to be an intuitive language that's used across multiple domains in computer science. It's easy to work with, and the data science community has put the work in to create the plumbing it needs to solve complex computational problems.

It could also be that more companies are moving data projects and products into production. R is not a general purpose programming language like Python.

Scala vs Python vs Java for Apache Spark Jobs: What's the leading programming language to use with Spark?

Apache Spark is the leading data processing engine and one of the most active open source projects around. Spark supports three production-ready languages: Scala, Python, and Java. We set out to answer the question: between Scala, Python, and Java, which language is most often used with Spark?

First, a quick summary of the languages we're discussing.

Language	Verbosity	Type Checking	Ease of Use	Friendliness with Spark
Java	High	Static	Medium	Friendlier with Java 8, somewhat painful Spark API
Python	Low	Dynamic	Easy	Friendly Pyspark library, but slower than Scala for certain workloads
Scala	Medium	Static	Medium	Very friendly, but higher learning curve than Python

Next, let's look at the more specific arguments we've heard for using a particular language with Spark.

For Scala:

- *Spark itself is written in Scala, which makes using Scala faster than Python (3x – 10x faster, at least for Spark 1.x) on the cluster **
- *Since Spark is written in Scala, using Scala forces developers building applications on Spark to more deeply understand the code*
- *Spark ships new features to the Scala API first*
- *Scala isn't that much harder than Python, but there's a learning curve*

For Python:

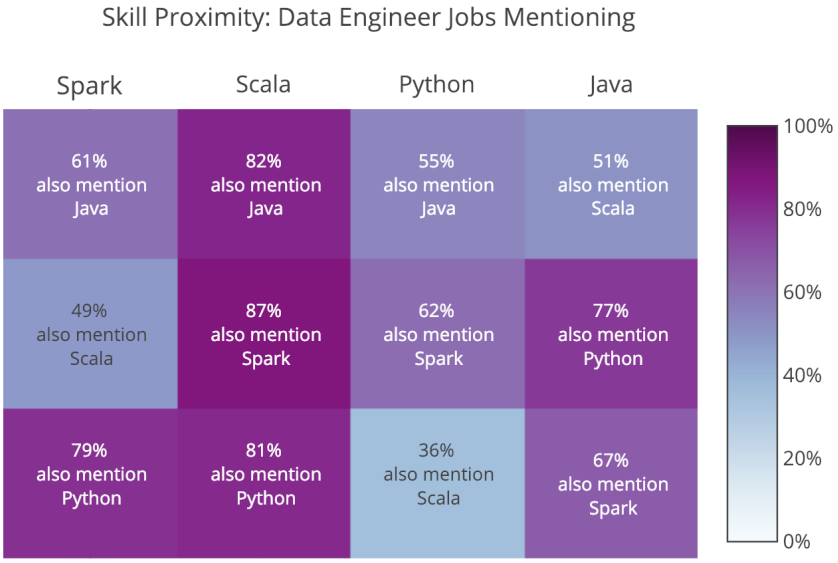
- *Python is generally a lot easier for data science teams to use*
- *Far more developers know Python, so projects aren't disrupted when teams change*
- *Spark's Java API is really verbose; Pyspark is much simpler*

For Java:

- *There are 9 million Java developers in the world today*
- *Java has been measured to be faster than Pyspark for complex processing jobs **

* Here's a write-up on performance benchmarking with different types of Spark jobs. The author's conclusion is that "if performance matters use either native UDFs, Scala or Java."

Let's assume performance matters. If we were to force a consensus on the performance with Spark debates, then from most performant language to least, we would have (1) Scala, (2) Java, and (3) Python. Now let's look at what's happening in the data engineering market. The chart below is a measure of skill demand proximity as measured by Cloud Roster over the last quarter.



Stack rank the demand for programming language skills among professionals with Apache Spark (column one) and you get the exact inverse of our performance ranking. **Of data engineer job posts that mention Spark: 79% also mentioned Python, 61% also mentioned Java, and 49% also mentioned Scala.**

In other words, the data tells us that Python is much more likely to be used with Spark than Scala. The difference cannot reasonably be explained by unimportant Spark clusters where performance does not matter. If we are to believe that the performance differences between Scala and Python are meaningful, it is much more likely that there exist companies that need to examine their technology decision criteria; are companies making an "it fits" or "best-fit" decision? It stands to reason that there are almost certainly cases out there where companies are settling for delivery timeliness at the expense of performance.

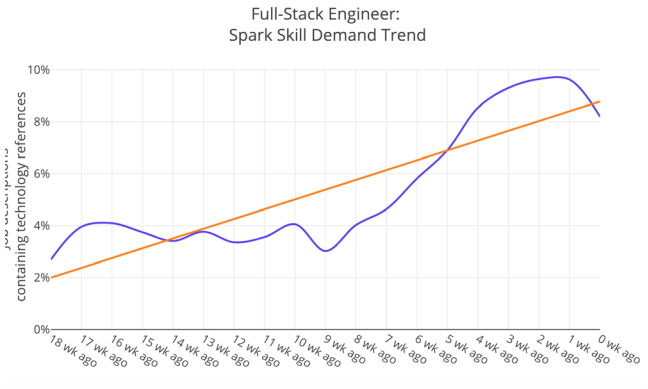
> " Are companies making an "it fits" or "best-fit" decision?

It is worth pointing out the high degree of skill proximity between Scala and Spark. Of the total number of Scala engineers being sought, 88% are also being sought for Spark expertise. In other words, it appears that where Scala is being used, it's being used in connection with Spark data processing and perhaps not many other places. The same cannot be said for Python. Of data engineer posts that mention Python, we can assume that for 38% of the time it's for something other than Spark processing jobs (1 – 0.62).

For organizations, our advice is to make sure to have a measurable pulse of your digital skills portfolio. Use training to align those skills to support a roadmap that prioritizes doing things the right way. For the many Java developers out there, it's a good idea to invest in learning an additional language. Last year, StackOverflow projected that Python would overtake Java in terms of popularity in 2018. Finally, for those who know any three of the languages, we suggest investing some time familiarizing yourself with Apache Spark if you haven't already. It's here to stay.

The data engineer toolset is bleeding into other areas of computer science

A few years ago, the common refrain was that "every company needed to become a technology company." In almost all cases, that was and continues to be true. As we look forward to 2019, every company needs to be thinking about how to become a better data company, from data stewardship and security to infrastructure and analytics. The tools have arrived, and we are seeing early signs in our data that roles outside of data engineering and science are being impacted.



The numbers are still relatively small, but we can see that Spark, typically the domain of just data engineering, is growing in importance for Full-Stack Engineers. While Spark does not yet appear in the top 20 skills for Full-Stack Developers, we expect it might break into the top 30 during 2019. As customer and employee experiences begin to rely on the data that end users give to companies, applications will increasingly need to be able to quickly capitalize on it in meaningful ways. We suspect that this is an initial trendline in what's to come.

Key Takeaways

- Python is the language of data engineering, not R. Python is almost four times more in demand than R for data engineers.
- Python appears more popular than Scala for Spark processing jobs, despite Scala having tangible benefits
- Python is most in demand for Spark, followed by Java in second, and Scala in third — despite some benchmarks showing the exact inverse ranking by performance
- Early signals in our data indicate that roles outside data engineering will need to get familiar with data engineering tools

Training Resources

If you have a team that is building a big data pipeline or getting into data science, you might be interested in these training resources.

Learning Paths:

- Zero to Deep Learning Bootcamp One – Introduction to Data Science and Machine Learning
- Building a Solution Using Artificial Intelligence and IOT
- Introduction to Azure Machine Learning
- Using Azure AI Services to Build Customer Solutions
- Applying Machine Learning and AI Services on AWS
- Machine Learning on Google Cloud Platform
- Introduction to and Optimizing Google Big Query
- Big Data – Specialty Certification Preparation for AWS
- Big Data Analytics on Azure
- Data Engineer – Professional Certification Preparation for Google

Hands-on Labs:

- Building a Data Pipeline in DC/OS
- Analyzing IoT Data Using Azure Stream Analytics
- Getting Started with Amazon Redshift
- Query encrypted Amazon S3 data with Amazon Athena

About the Data in this Report

Cloud Academy collects and analyzes upwards of 3,000 job descriptions each day for several cloud job roles based in the United States (we do have plans to expand). The data is then de-duplicated and analyzed. The technical job roles for which we collect job postings are: Cloud Architect, DevOps Engineer, Full-Stack Engineer, Network Architect, Security Engineer, QA Engineer, and Data Engineer. This data report leverages our job posts data warehouse and references measurements like **Skill Relevancy** (this is the association of a given technical skill to a job role based on the existence of a technology's mention in a job post) and **Skill Proximity** (this is the frequency with which two or more skills appear within a category of job posts during a particular timeframe).

WRITTEN BY

Alex Brower

Alex is responsible for go-to-market research, operations, and analysis at Cloud Academy.

Cloud Academy is the leading enterprise training platform that accelerates teams and digital transformation.

Companies trust Cloud Academy to deliver multimodal training on the leading clouds (AWS, Azure, Google Cloud Platform), on the essential methodologies needed to operate on and between clouds (DevOps, Security), and on the capabilities that are unlocked by the cloud (machine learning, IoT).

From the fundamentals to advanced scenario training, Cloud Academy empowers organizations with the knowledge, critical thinking, and hands-on experience needed to adopt, operate, and optimize the multi-cloud.

cloudacademy.com