O'REILLY®

The Big Data Market

A Data-Driven Analysis of Companies Using Hadoop, Spark, Data Science & Machine Learning

Aman Naimat

Strata+ Hadoop WORLD

Make Data Work
strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World helps you put big data, cutting-edge data science, and new business fundamentals to work.

- Learn new business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # D2044

The Big Data Market

A Data-Driven Analysis of Companies Using Hadoop, Spark, and Data Science

Aman Naimat

O'REILLY

THE BIG DATA MARKET

by Aman Naimat

Editors: Marie Beaugureau, Ben Lorica
Designer: Ellie Volckhausen
Production Editor: Shiny Kalapurakkel

Printed in Canada.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (http://safaribooksonline.com). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

ISBN: 978-1-491-95991-6

Table of Contents

THE BIG DATA MARKET

THIS REPORT IS A DATA-DRIVEN STUDY OF THE COMPLETE BIG DATA MARKET

USING A ONE-OF-A-KIND ANALYSIS of billions of documents, this report shows who is—and isn't—using big data tools and techniques like Hadoop, Spark, and machine learning.

THIS REPORT IS A DATA-DRIVEN STUDY of the complete big data market. It is derived from live data triangulated across the entire business world—websites, meetups, hiring patterns, business relationships, blogs, press, forums, SEC filings...everything—using data crawlers and proprietary natural-language parsing technology developed by Spiderbook. This bottoms-up data methodology is in sharp contrast to traditional approaches dependent on anecdotal evidence derived from small-sample user and analyst client surveys.

But as revolutionary as big data is, our analysis of more than 500,000 of the largest companies in the world reveals that a very small percentage of them have embraced big data methodologies in reality. One could argue that big data is still very much in the early adoption phase of the technology adoption model.

One could argue that big data is still very much in the early adoption phase of the technology adoption model.

The numbers, perhaps studied for the first time looking at actual data, suggest that there remains a lot of room for growth in the big data market, and newer technologies like Spark are overtaking Hadoop's MapReduce, the current reigning patriarch of the open source big data movement.

This report first covers the flagship of the big data technologies, Hadoop. Next we look into the use cases for big data technologies, highlighting some surprising results about budgets and spending on big data and data science. We follow that up with an examination of fast data, using Spark, Kafka, and Storm as indicators of fast data projects. Finally we culminate the report by providing characteristics of big data users—the engineers and data scientists who work with these technologies.

Big Data in the Real World

OUR RESULTS ARE BASED ON Spiderbook's automated analysis of billions of publicly-available documents, including all press releases, forums, job postings, blogs, tweets, patents, and proprietary databases that we have licensed. We use these documents to train our artificial intelligence engine, which reads the entire business Internet to understand these signals. The result is a remarkably accurate, near-real-time snapshot of the technologies in use at more than a half-million companies.

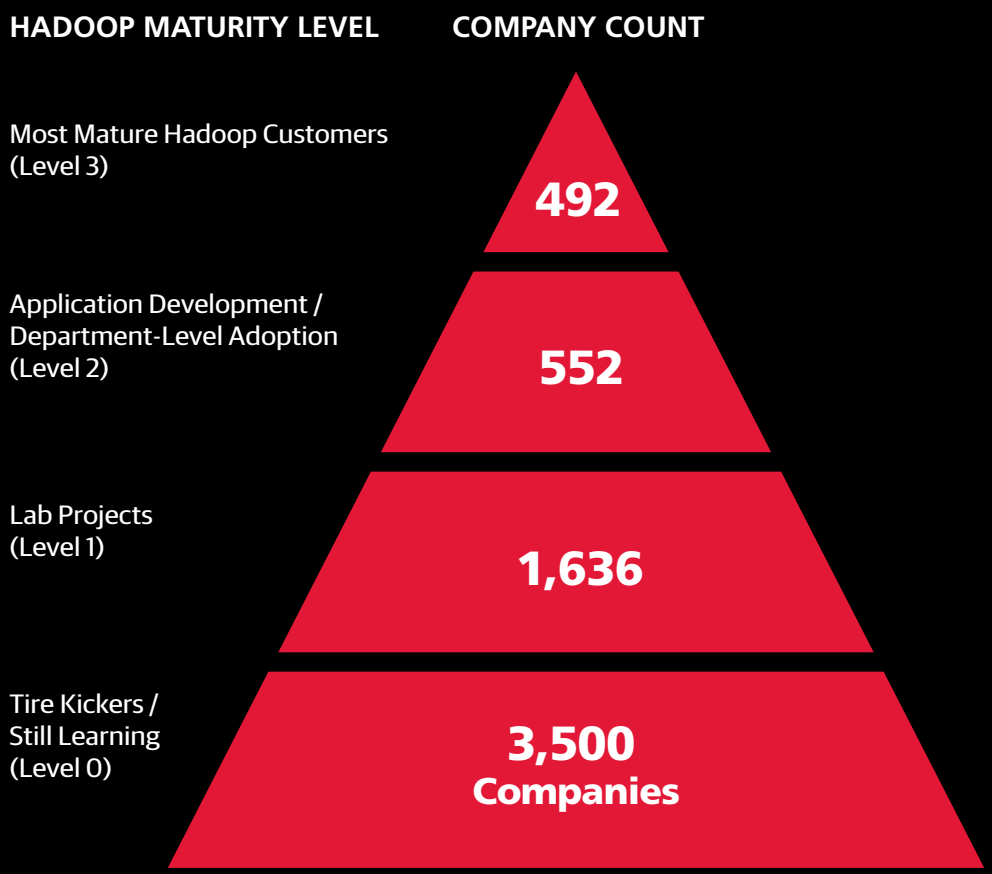
What types of trends are we looking for? For instance, we look at the skills held by employees at every company in our analysis to find out who is using various tools and platforms; for example, who is hiring folks with skills in Apache Spark; and which companies employ data scientists, and how many. In addition, we also use natural-language processing to understand business relationships between companies and vendors in the big data space, along with who is working on which use cases.

Hadoop Usage and Big Data Adoption

OVERALL, WE FOUND ONLY 2,680 COMPANIES that are using Hadoop at any level of maturity. Of those, 1,636 are at the lowest level of big data maturity: these companies are just getting started or working on a "lab" project. Another 552 are at the second level, where they've been using Hadoop and have a big data project within their companies at a small scale (at a department level or within a small startup). And just 492 are at the most advanced level, with evidence of major deployments, production-ready pipelines, and experienced Hadoop developers. Level 0 companies are still trying to learn about these technologies, attending meetups and conferences, and listening to webinars, but not actively working on any big data projects.

HADOOP MATURITY LEVEL

NUMBER OF U.S. COMPANIES USING HADOOP BY MATURITY OF HADOOP PROJECTS



Bigger Companies Are More into Big Data

Surprisingly, larger enterprises (those with more than 5,000 employees) are adopting big data technologies such as Hadoop much faster than smaller companies. You'd think smaller, younger companies would be nimbler and quicker to embrace new technologies, but when it comes to big data, the opposite is the case.

We found more than 300 large companies that have made serious investments in Hadoop. By contrast, there are only another 300 companies with 5,000 or fewer employees that are mature Hadoop users.

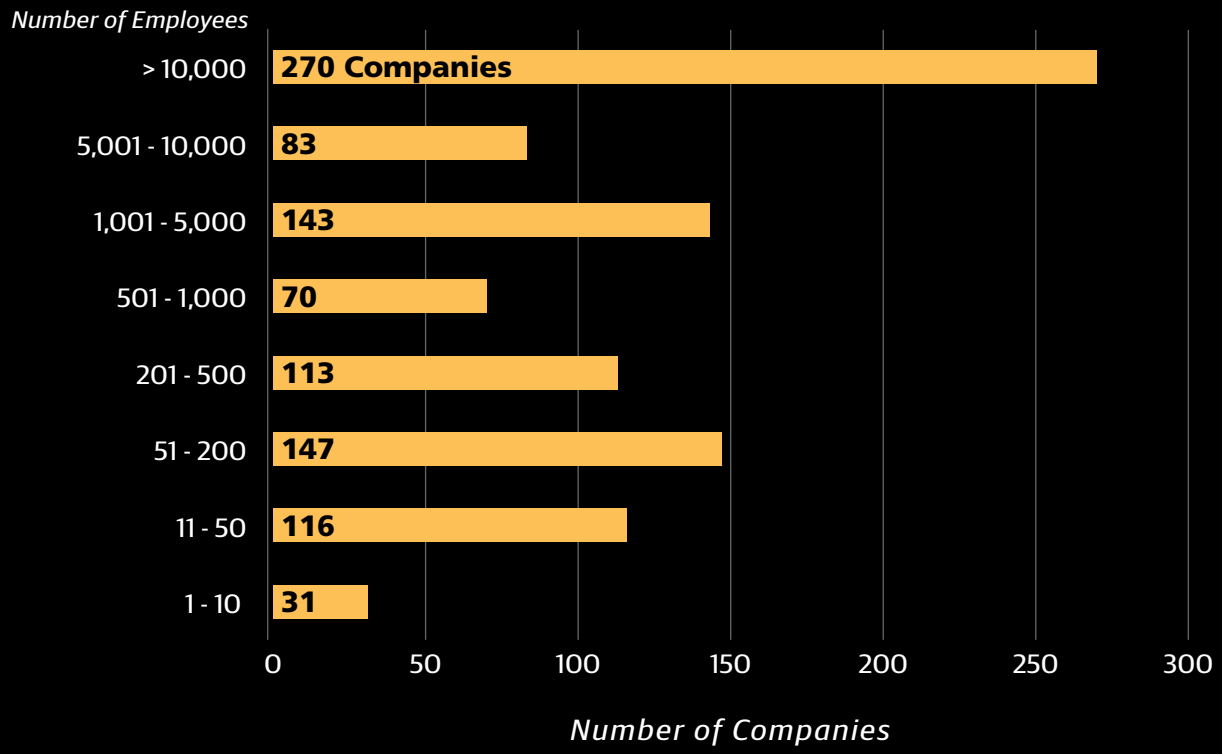
Because there are ten times more small companies, this means that in smaller companies, Hadoop has less than one-tenth the penetration that it has in the large-company set.

Most of the smaller companies adopting Hadoop are high-tech, data-oriented companies themselves. We don't know why smaller enterprises are lagging. Is it because they can't afford Hadoop and related technologies or is it because they can't pay the high salaries commanded by data scientists and data engineers? Or perhaps they just don't have as much data?

HADOOP ADOPTION BY COMPANY SIZE

NUMBER OF U.S. COMPANIES DEPLOYING HADOOP BY COMPANY SIZE

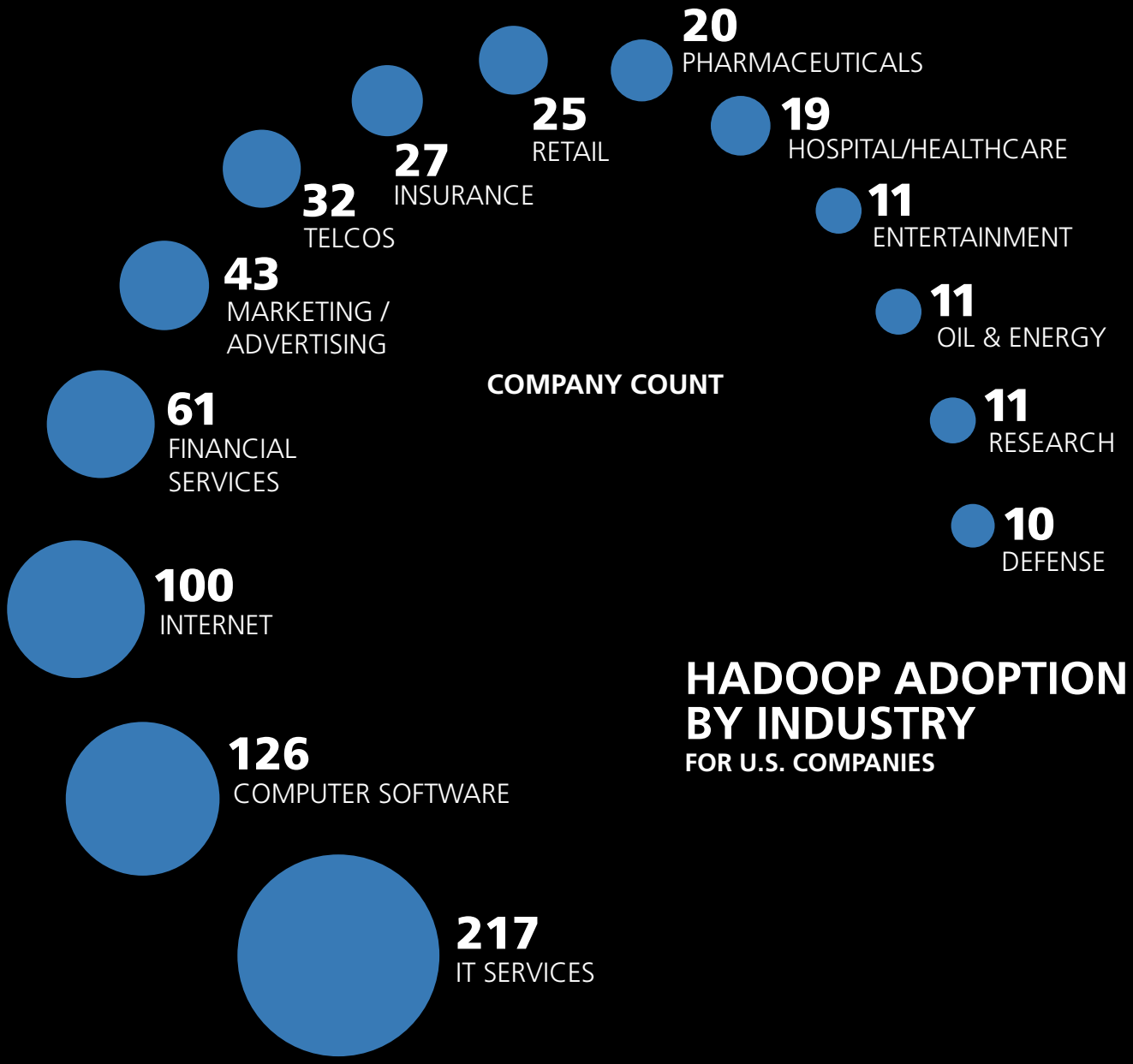
(NUMBER OF EMPLOYEES)



Oil and Pharma Lag; Financial Services Lead

Oil and gas companies as well as pharmaceutical manufacturers typically have enormous datasets, yet our analysis finds that they are not adopting Hadoop in great numbers. However, financial services companies are—even though this sector is not typically regarded as a fast adopter of new technology.

Perhaps the financial sector has been influenced by the early lead of companies like American Express. Or, perhaps these companies are migrating directly from mainframes to Hadoop, skipping generations of technologies in between. Startups like Paxata and Syncsort have emerged to help companies do just this.



HADOOP ADOPTION BY INDUSTRY

FOR U.S. COMPANIES

Use Cases for Big Data Technologies

THERE ARE MANY VISTA POINTS from which you can measure what companies are doing with Hadoop and Spark, and how much money they are spending. We could count projects, number of people deployed to the projects, number of companies doing the projects, how many use cases, and so on. Some of these measures are difficult to find from the outside (or perhaps even from the inside). We decided to triangulate how much money is spent on each of the use cases for Hadoop and Spark.

The biggest costs associated with Hadoop projects is human capital, so we measured the number of people working on different use cases across the approximately 2,700 companies actually using Hadoop. We also took into account the level of maturity of their use case: is it in production, do they have a director/VP of big data, and so forth. The missing pieces here are the infrastructure costs around machines/hardware/vendor support, but one can assume those are highly correlated to the number of people working on the projects.

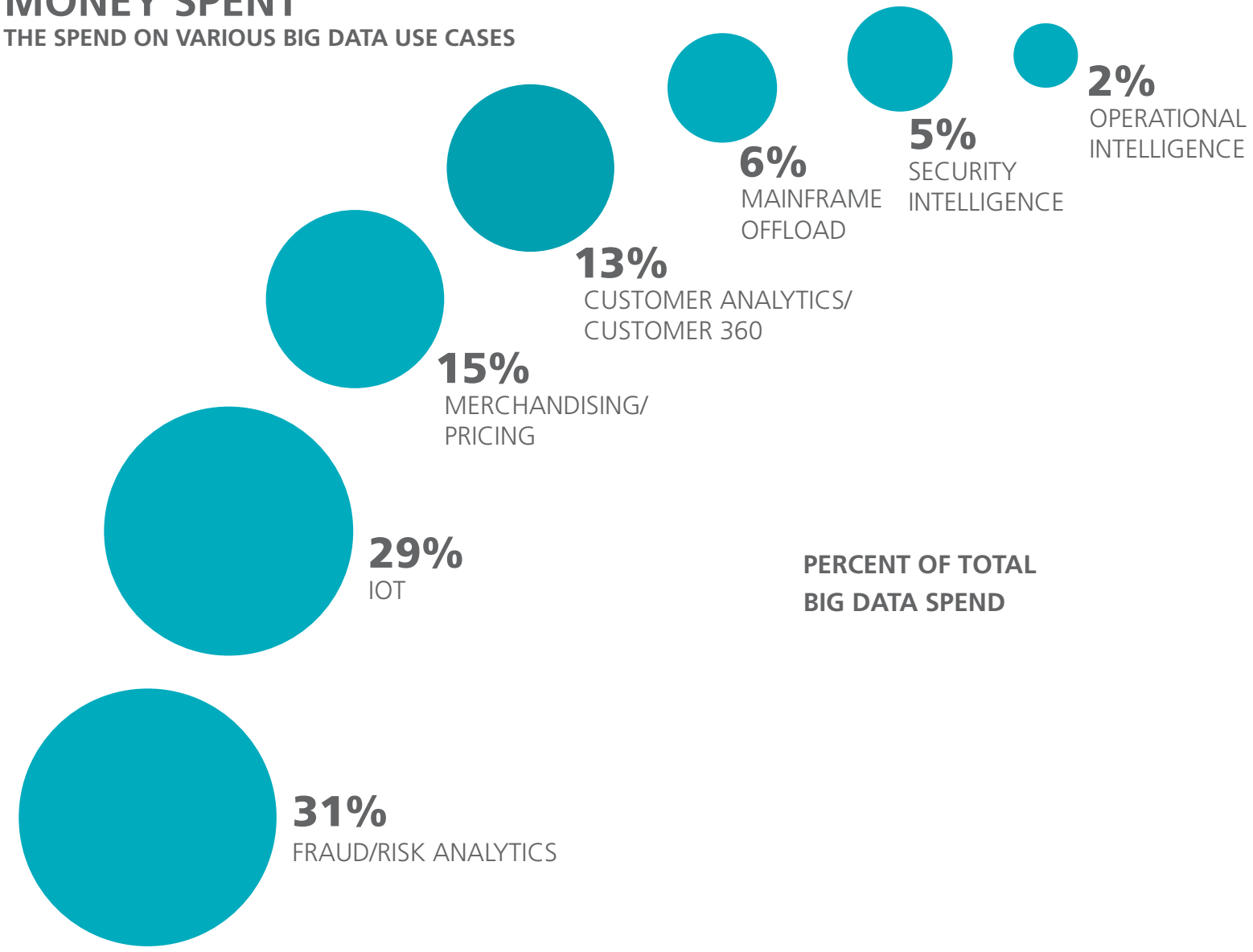
The biggest costs associated with Hadoop projects is human capital.

Fraud, Security Intelligence, and Risk Management Use Cases Have the Biggest Budgets

The results of our analysis are a stark contrast to what we generally hear as the most common Hadoop use cases. Although there might be more companies working on customer analytics and customer data-related projects with Hadoop, our data suggests that one-third of the money—and people—budgeted for big data projects are dedicated to risk analytics, fraud, and security intelligence. One can imagine that such work is not publicized as much as the 360-Degree Customer View or company analytics use cases, but it looks like the budgets for fraud and risk analytics are much higher than other use cases. Also surprising is how much budget is being invested into Internet of Things (IoT)–related use cases for Hadoop and Spark, because although there is talk about streaming data from IoT in the big data circles of Silicon Valley, it is not considered the bread-and-butter use case.

MONEY SPENT

THE SPEND ON VARIOUS BIG DATA USE CASES



Fast Data Is Moving Fast

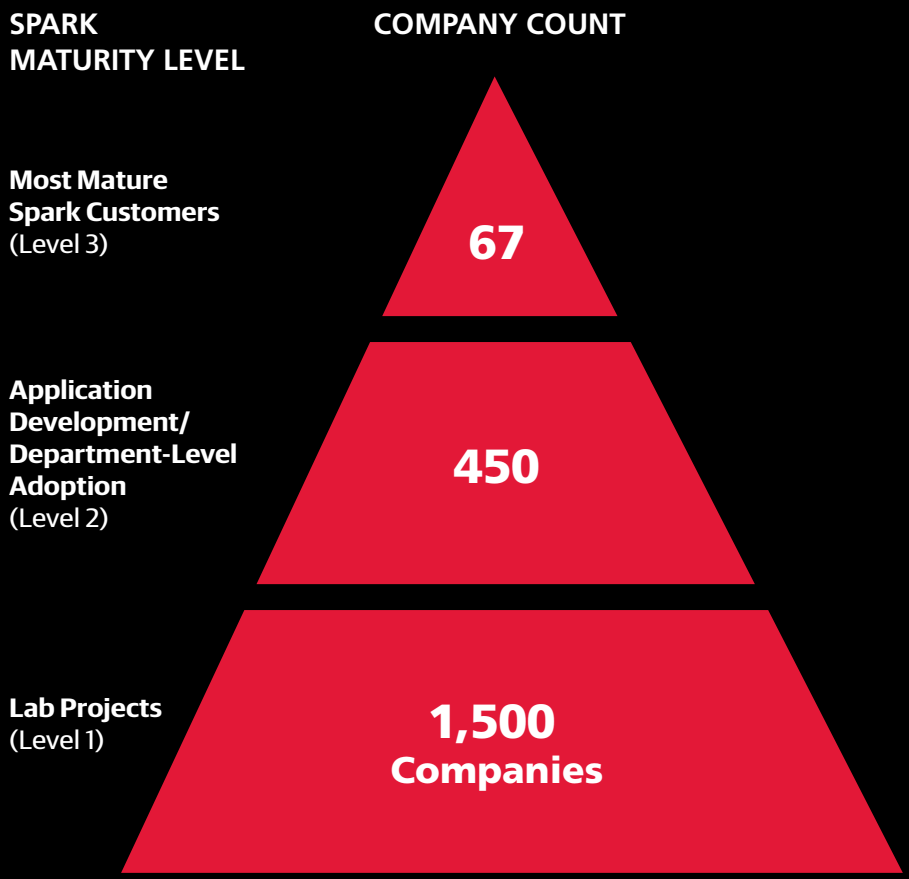
FOLLOWING OUR ANALYSIS OF HADOOP, which served as a proxy for big data in general, we will now break down the market dynamics of streaming and real-time systems like Spark, Storm, and Kafka, which we can use to identify fast data projects. There are major commercial technologies in this space like VoltDB, Amazon Kinesis, Data Torrent, TIBCO, and others, but we focused exclusively on the adoption of open source fast data technologies.

The Apache Spark market is reaching Hadoop-like customer adoption

We found upward of 2,000 companies with different levels of adoption of Apache Spark and related streaming technologies. What is surprising is that the Apache Spark market is reaching Hadoop-like customer adoption so quickly. However, although there are more than 500 companies with production-level Hadoop maturity, we found only 67 companies with that level of Apache Spark maturity.

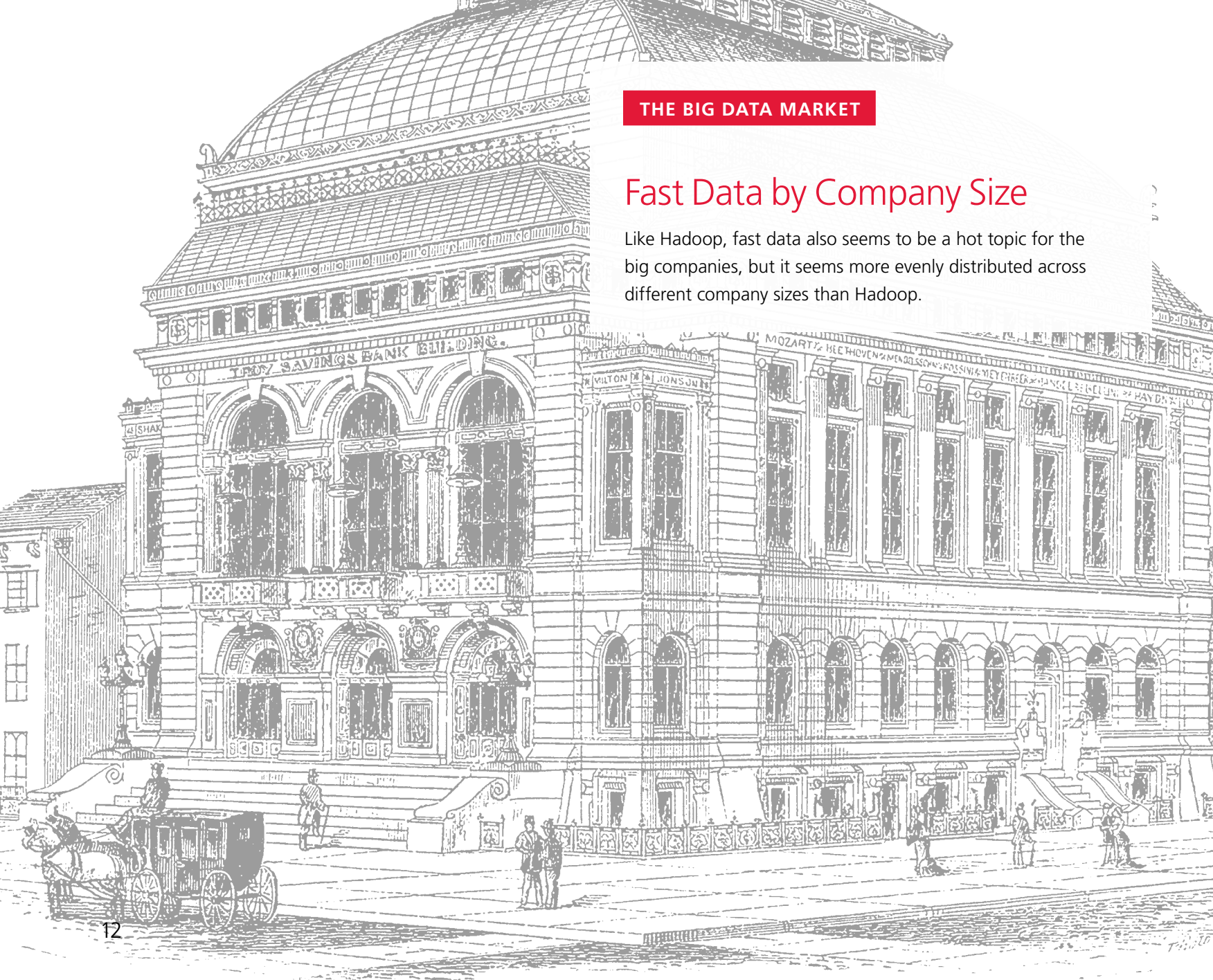
COMPANIES USING FAST DATA

NUMBER OF U.S. COMPANIES USING SPARK BY MATURITY OF SPARK PROJECTS



Fast Data by Company Size

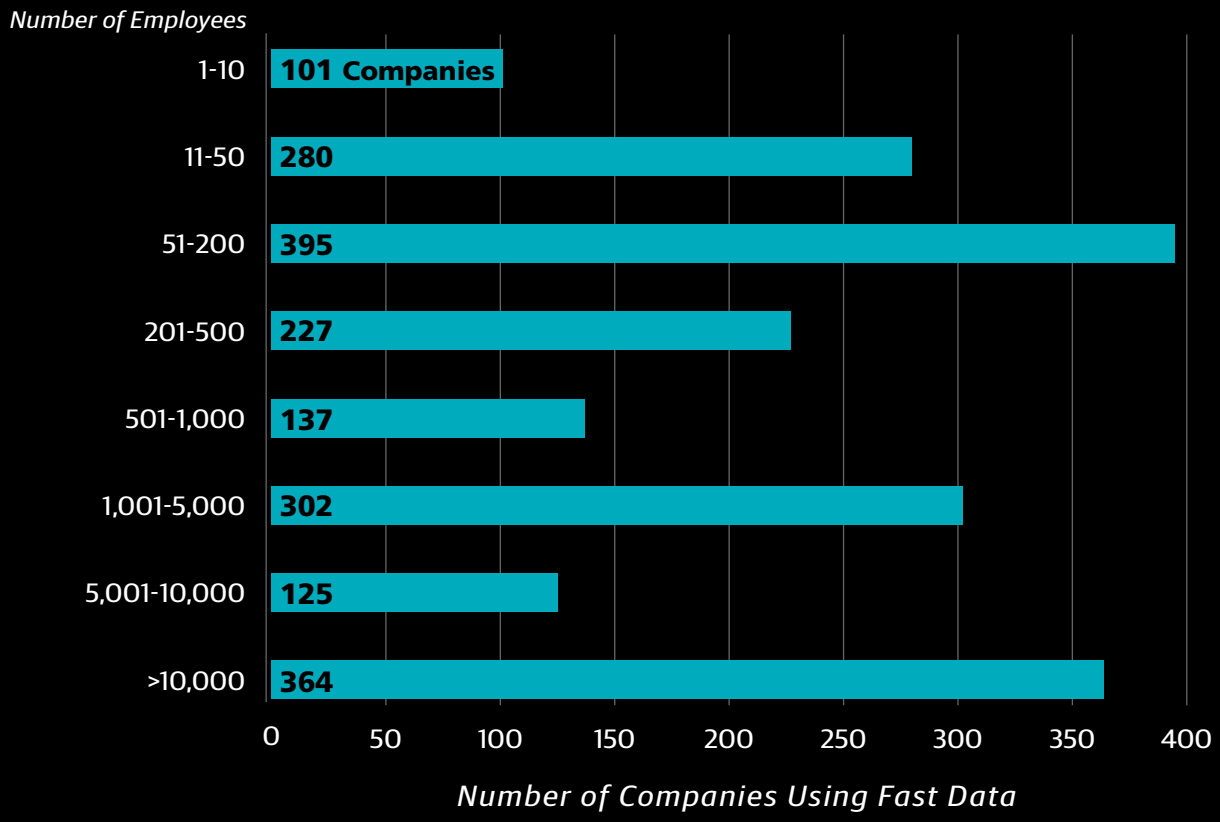
Like Hadoop, fast data also seems to be a hot topic for the big companies, but it seems more evenly distributed across different company sizes than Hadoop.



FAST DATA ADOPTION BY COMPANY SIZE

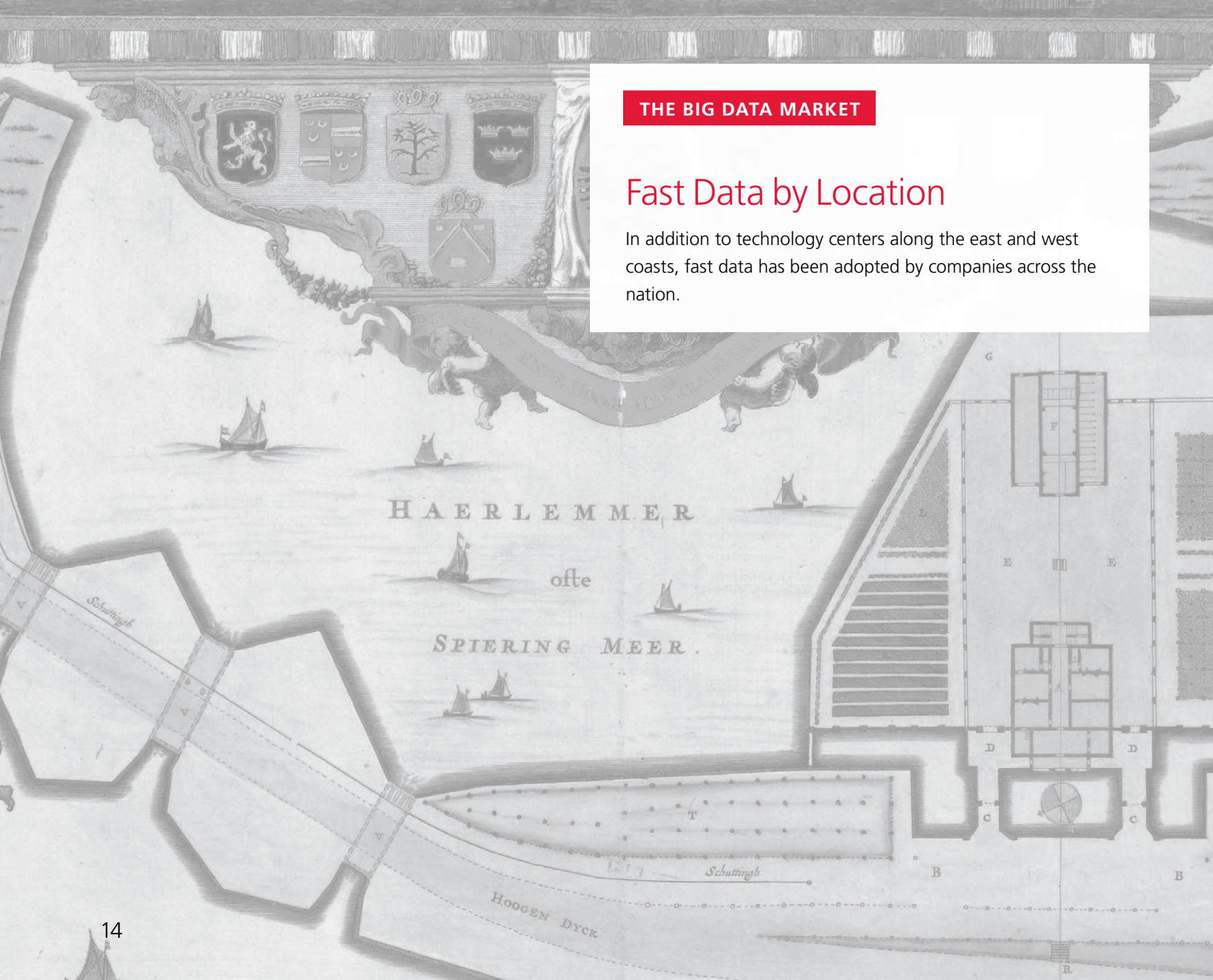
NUMBER OF U.S. COMPANIES DEPLOYING FAST DATA BY COMPANY SIZE

(NUMBER OF EMPLOYEES)

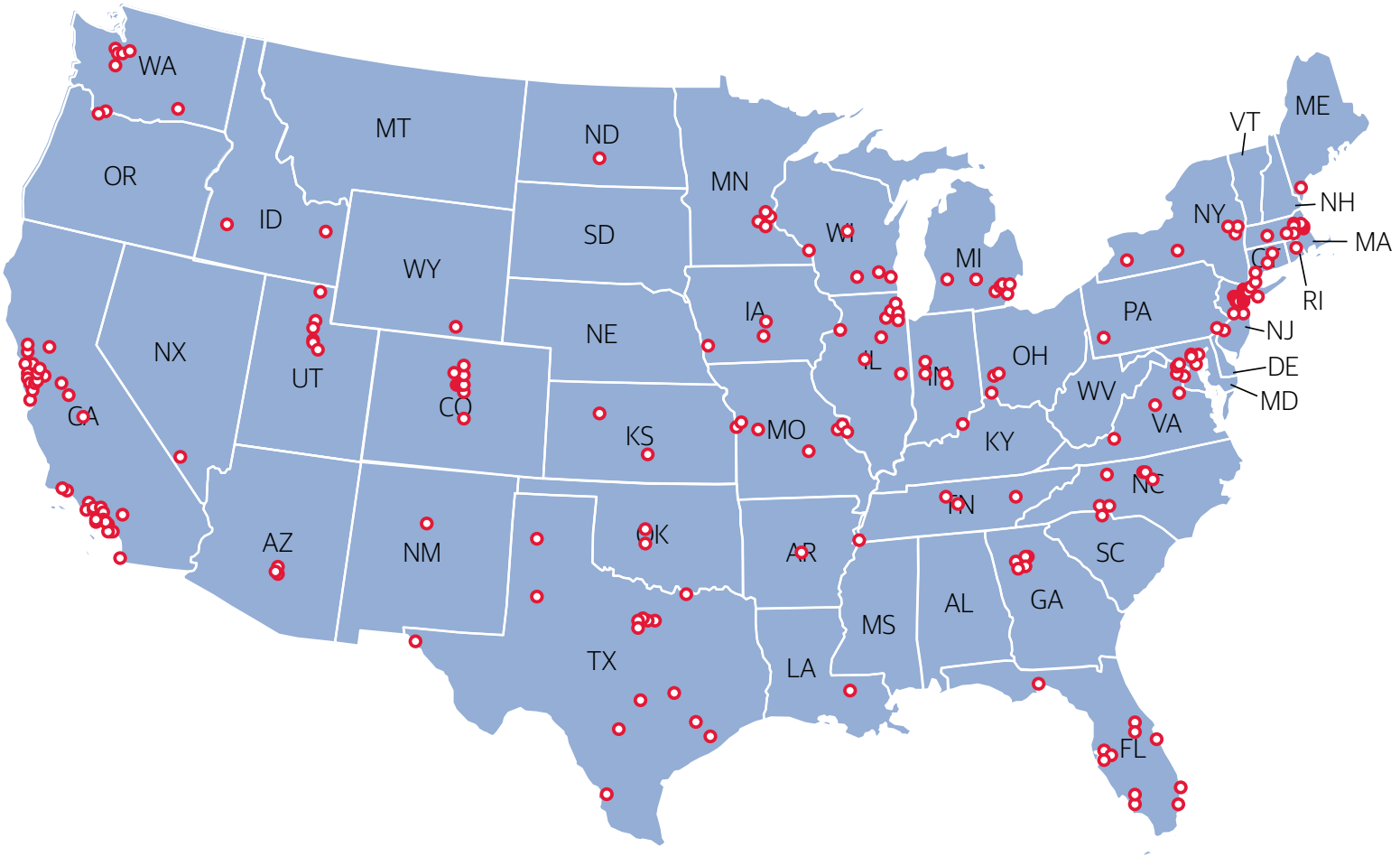


Fast Data by Location

In addition to technology centers along the east and west coasts, fast data has been adopted by companies across the nation.



LOCATIONS OF U.S. COMPANIES ADOPTING FAST DATA

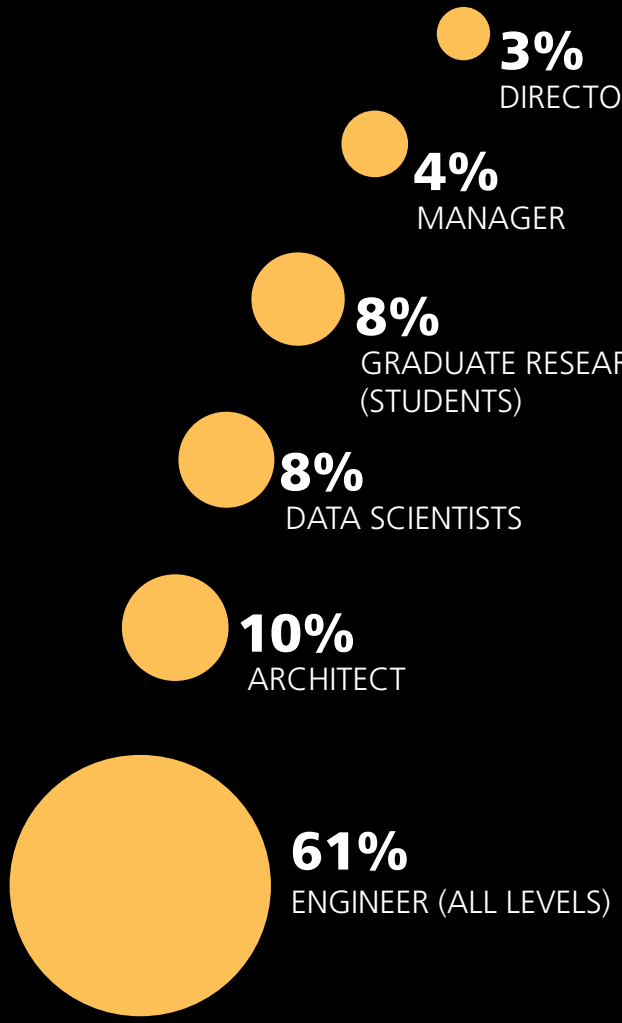


Fast Data by Job Role and Degrees

Unsurprisingly, the leading job role for fast data technology users (i.e., those using Spark, Storm, and/or Kafka) is that of engineer. One notable fact is that there are a ton of students doing research using fast data technologies. Perhaps fast data vendors should think about providing their expertise and technologies to researchers/students who are starting to apply this technology.

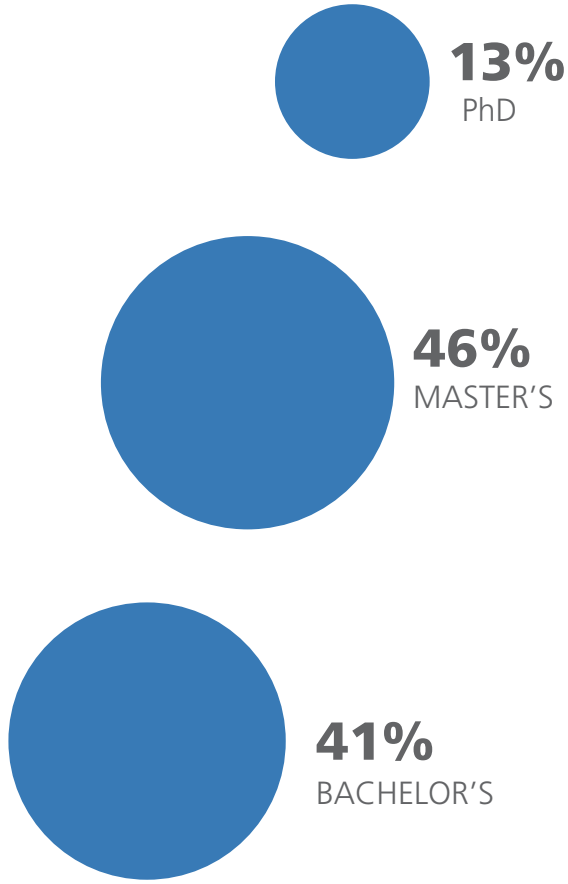
According to United States Department of Education Statistics, in 2014 only 23 percent of people with higher -education degrees have master's or doctoral degrees. But 59 percent of fast data users hold degrees above a bachelor's. We merged our data with third-party contact data providers like ZoomInfo to understand education qualifications and job history of these engineers and data scientists.

JOB ROLE
THE ROLES USING FAST DATA



PERCENT OF USERS

DEGREE
LEVELS



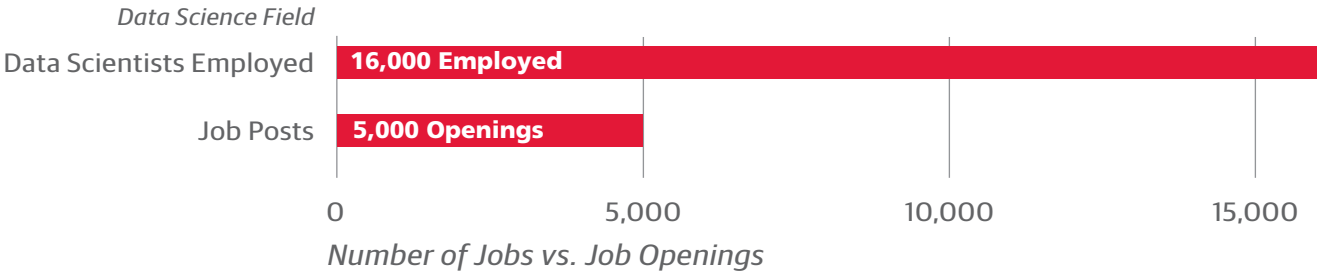
PERCENT OF USERS

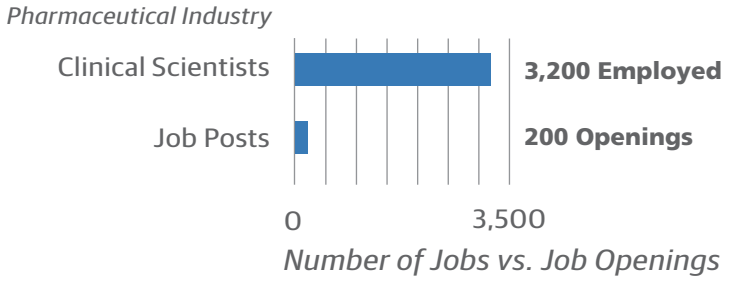
THE BIG DATA MARKET

The Need for Data Scientists Is Exploding

A LARGE NUMBER OF USE CASES FOR HADOOP and Spark are in and around data science. A quick search for "Data Scientists" on LinkedIn shows that there are 16,000 employed in the US. What is shocking is that there are a whopping 5,000 open job posts for people with data-science skills. One can appreciate the demanding difference if you compare this to clinical scientists at pharmaceutical companies. There are 3,200 clinical scientists in the U.S., according to LinkedIn, but a crawl of job posts shows only 200 vacancies for that position across American companies.

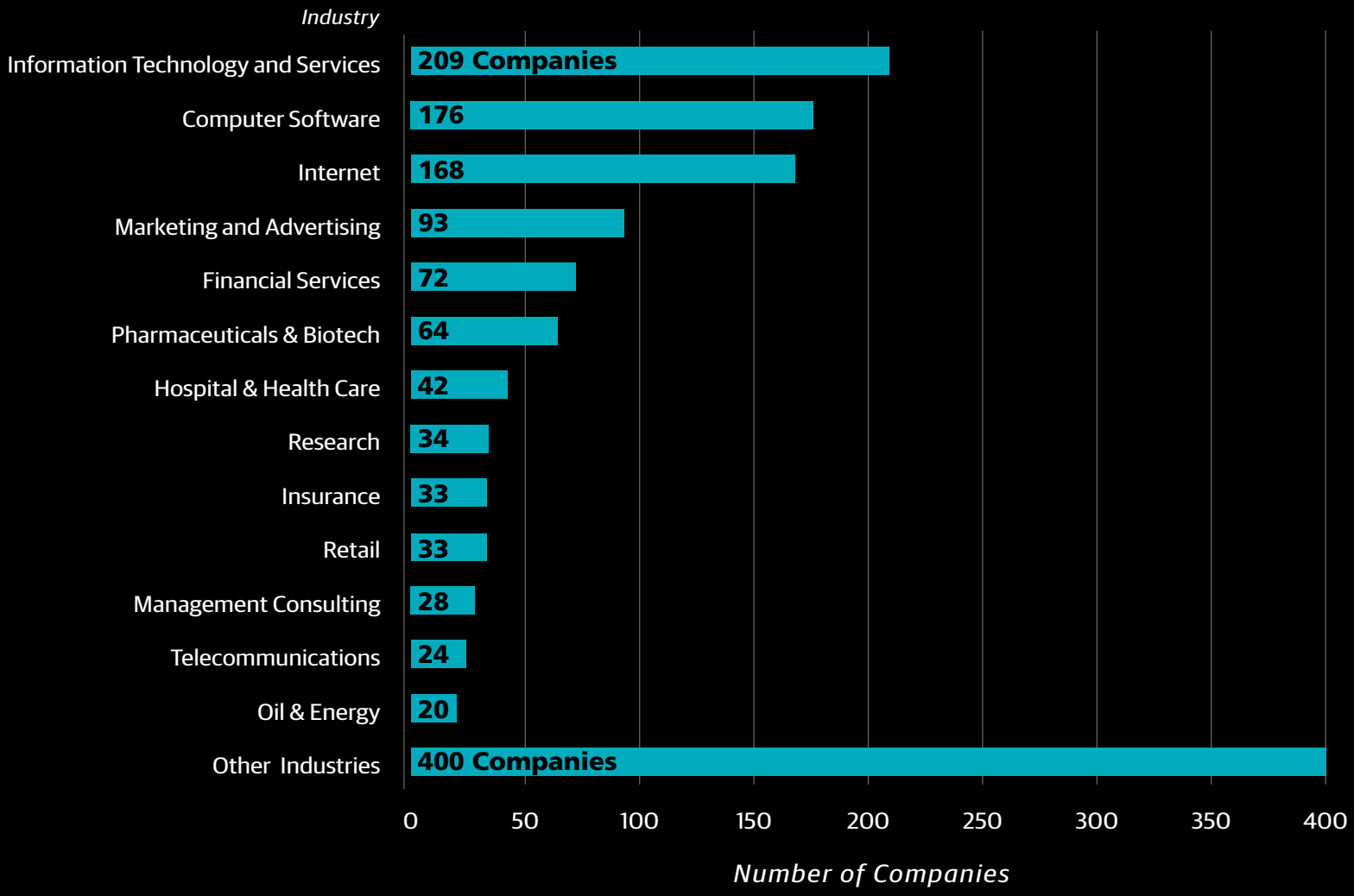
U.S. DEMAND FOR DATA SCIENTISTS VS. DEMAND FOR CLINICAL SCIENTISTS





DATA SCIENCE ADOPTION BY INDUSTRY

NUMBER OF U.S. COMPANIES WITH THE MOST ADOPTION OF DATA SCIENCE, BY INDUSTRY

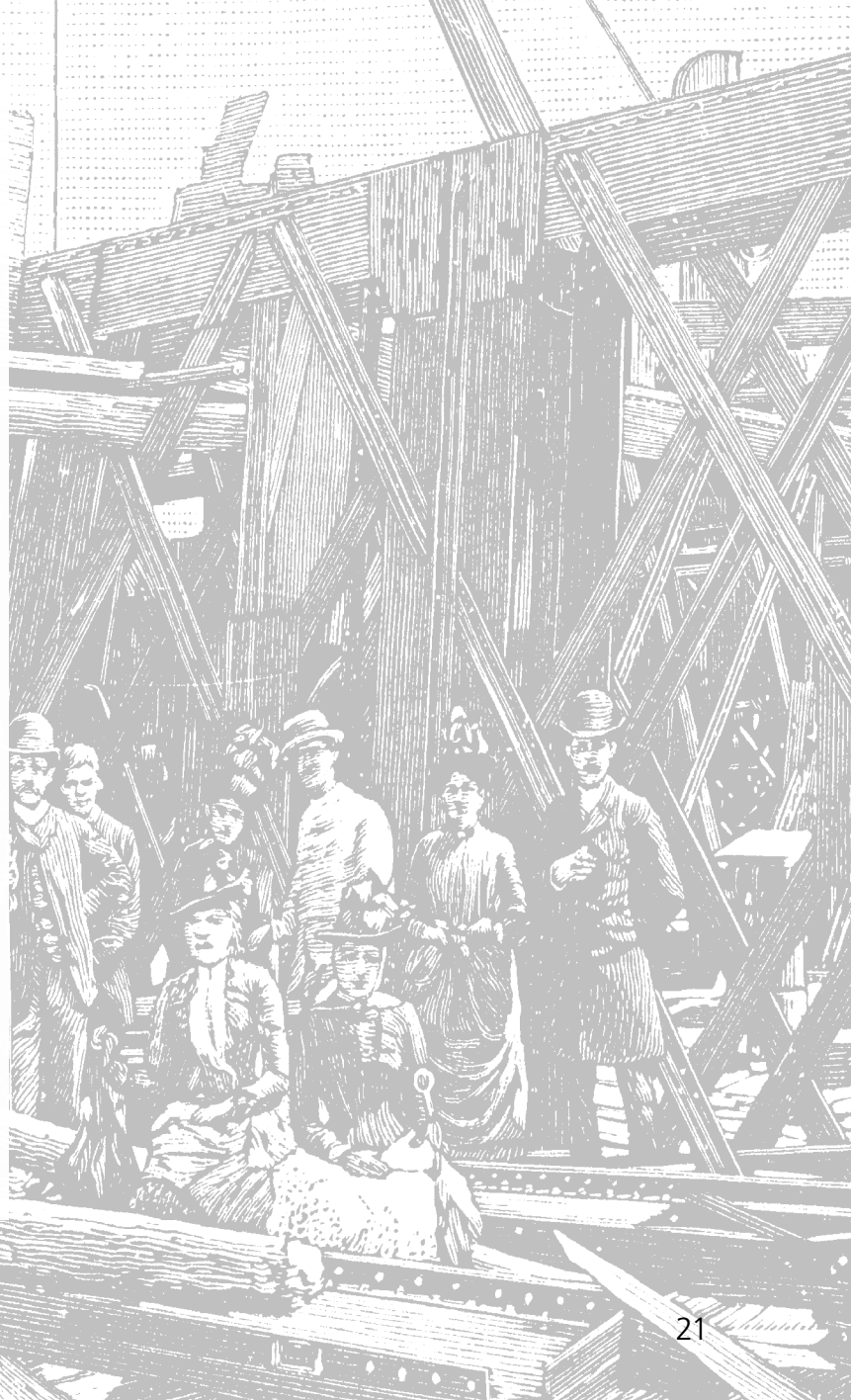


Data Science in Industry

Commercial industries are using data science and data scientists in the U.S. After high-tech/Internet companies, marketing and advertising companies are leading in data science spending, followed by financial services. We found that financial services companies employ more data scientists per company, but there are fewer financial services firms than smaller marketing and advertising companies using data science.

Diversity Problems in the Big Data World

Because we had fairly high coverage of all people working in big data and data science, we used a baby-name database to classify the gender of all data scientists and decision makers. Although not perfect, we have found this method to be 92 percent accurate in the past. We found that only 6.3 percent of the names of big data users and decision makers are female. It's clearly a male-dominated market. This is even more bleak than the general software industry, in which females represent 16 percent of the market



The Future of the Big Data Market

SPIDERBOOK TODAY CANNOT UNDERSTAND trends over time, because we have only recently begun gathering all data for big data adoption across businesses. We have not run this across time to know how this market has besen progressing. But, we can study the trends superficially by looking at Google search.

Even though our data goes deeper into behavioral signals like spending patterns, users, visiting meetups, or business deals, we can look at Google search trends to see how the market is progressing. The chart that follows from Google trends Apache Spark versus Hadoop versus Data Science corroborates our findings that Apache Spark is quickly catching up to Hadoop and the exploding nature of data science.