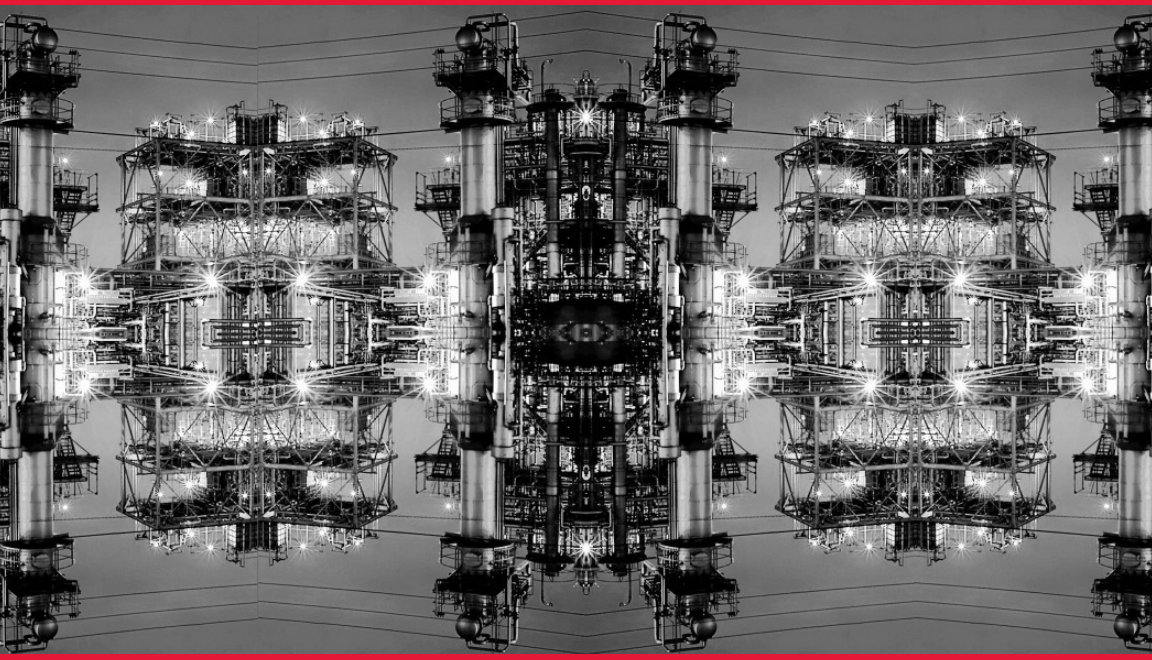
O'REILLY®

Data Science for Modern Manufacturing

Global Trends: Big Data Analytics for the Industrial Internet of Things



Li Ping Chu

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

Data Science for Modern Manufacturing

Global Trends: Big Data Analytics for the Industrial Internet of Things

Li Ping Chu

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Data Science for Modern Manufacturing
by Li Ping Chu

Printed in the United States of America.

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*.

Editor: Shannon Cutt
Production Editor: Kristen Brown
Copyeditor: Octal Publishing, Inc.
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

July 2016: First Edition

Revision History for the First Edition
2016-06-10: First Release

978-1-491-95896-4

[LSI]

Table of Contents

Data Science for Modern Manufacturing **1**
Preface 1
Introduction 2
Industrial Internet 3
(Industrial) Internet of Things 11
Big Data and Analytics 15
Machine Learning 22
Autonomous Robots, Augmented Reality, and More 25
Challenges 27
Conclusion 29

Data Science for Modern Manufacturing

Preface

When I was approached about the opportunity to write this report, I was told that O'Reilly was looking for someone with a technical background, experience in writing, and the ability to communicate in Mandarin Chinese to put something together that included the topics of Big Data, Manufacturing, Internet of Things, *Made In China 2025*, *Industrie 4.0*, and Industrial Internet. I told them, "No problem!" and then set off to do some research. What I found was that there is no shortage of information available—there are literally hundreds, if not thousands, of articles and reports that on these topics—but there aren't a lot of straightforward answers. I began to imagine how incredibly frustrating it would be if I were a decision maker for a manufacturing company and I knew that we needed to act fast to kick off an *Industrial Internet* project but couldn't be certain about the quality of information out there.

Therefore, the purpose of this report is to deliver to you the fundamentals of the Industrial Internet—particularly if you're in the business of "making stuff." With cutting edge technology, it's impossible to write a text that will be definitive, but I attempted to compile as much of the relevant information in one place to help you cut through some of the jargon and marketing hype. In this report, you will learn about what Industrial Internet is, what governments are doing to promote Industrial Internet, the technologies that are the backbone of the digital revolution in industry, and the challenges and problems that you should consider. We will also closely examine

the *Industrial Internet of Things* (IIoT) and the role of *Big Data Analytics* in all of this. We've also had numerous experts in the industry from around the globe weigh in and share their thoughts and opinions. We hope that after reading this report, you will feel properly equipped to have an informed and meaningful conversation on these topics.

Introduction

The world's leading nations are standing at the precipice of the next great manufacturing revolution and their success or failure at overhauling the way goods are produced will likely determine where they stand in the global economy for the next several decades. Despite the uncertain economic outlook as of this writing, the ranks of the world's middle-income families are still slated to balloon to 3.2 billion in 2020 and 4.9 billion in 2030 (from 1.8 billion in 2009).[1] With this newfound buying power comes massive increased demand for high-quality consumer goods at a reasonable cost. To meet this demand will require an equivalent increase in output and efficiency from manufacturers, and this increased output is going to come from breakthroughs in Information Technology—in particular the *Internet of Things* (IoT) and Big Data Analytics.

However, the expanding market is not the only factor driving companies to modernize their production facilities. Increasingly, top manufacturing nations are seeing factories move to countries where wages are lower. Companies that have located their manufacturing in industrial powerhouses like Germany and China are feeling the pinch as labor costs rise. For the time being, Chinese workers can still claim to be far more efficient than their counterparts in India and Vietnam, and Germany will remain the European export leader for the foreseeable future due to its highly specialized industries (in particular auto and machinery), but neither of them are content to rest on their laurels.

Furthermore, China posted GDP growth of only 6.9 percent for 2015, which is its weakest growth rate in 25 years. Economic projections for 2016 and beyond suggest that the once gaudy economic expansion of the previous decades is tapering off as the Chinese

1 "The Emerging Middle Class in Developing Countries" by Homi Kharas.

economy matures. This phenomenon is being referred to as the "New Normal" by China's policy makers who are looking for ways to secure a sustainable rate of economic growth for the future. Germany has scaled back its forecast for GDP growth to 1.7 percent for 2016 in the wake of slowing demand from emerging markets. Both nations are highly dependent on manufacturing exports as a component of their economies (22.6 percent for China and 45.7 percent for Germany as of 2014)[2] and are therefore more vulnerable to downturns in the economies of their trade partners. By using smart technologies, these export goliaths are hoping to optimize their supply chains and, in turn, minimize the effect fluctuations in the global markets have on their local economies.

To this end, the governments of Germany and China have both drawn up extremely ambitious plans to bring their manufacturing sectors into the 21st century. Germany has dubbed its plan Industrie 4.0 in reference to the fourth major industrial revolution. Taking a page from Germany's book, the Chinese have come up with Made in China 2025, which—in typical Chinese fashion—is further reaching and even more expansive in its aims. This report will present you with a comprehensive look at both of these initiatives and closely examine the technologies that will be underpinning them as well as the challenges ahead.

Industrial Internet

Before we can really begin to understand the details of the German and Chinese plans, we need to define *Industrial Internet*. In the report "Industrial Internet" (O'Reilly, 2013), Jon Bruner states:

> The Industrial Internet is the union of software and big machines—what you might think of as the enterprise Internet of Things, operating under the demanding requirements of systems that have lives and expensive equipment at stake. It promises to bring the key characteristics of the Web—modularity, abstraction, software, above the level of a single device—to demanding physical settings, letting innovators break down big problems, solve them in small pieces, and then stitch together their solutions.

Another way to wrap your mind around this concept is to first imagine a company with several manufacturing centers. Now imag-

2 Exports of goods and services (percent of GDP).

ine all of the information systems, employees, and machines (from assembly line robots to forklifts), tools, and monitoring systems (cameras and sensors) in the company as nodes on a network, which are in turn connected to the Internet. Each of these nodes is constantly producing and receiving data on the current situation in the plant and the Internet at large. As conditions change, the individual nodes respond accordingly.

To better illustrate how this would work, let's run through a hypothetical scenario for a make-believe manufacturer of selfie sticks. This particular company (which we will call Vanity Products Unlimited, or Vanity for short) is the largest manufacturer of selfie sticks in the world. Demand is high, but its plants usually run at around 70 percent capacity during the nonpeak season.

In our scenario, news has hit that the second largest producer of selfie sticks has suffered a plant fire. Although no injuries or fatalities were reported, it will be a minimum of three months before it will be back online. Vanity's systems, which are always monitoring the market for relevant news about the current marketplace, detect the event and make a number of calculations about the unmet demand that will result from the incident. These calculations will be based on a number of variables, including historical data, current inventory stocks, market demand, and so on. With minimal human interaction, the system places orders for parts and raw materials, schedules additional personnel for plant shifts, and starts up additional production lines at the facilities to increase output. The system also makes appropriate logistical arrangements to ensure that the products get to the locations where demand is highest—balancing delivery time against cost—to take advantage of the sudden shortfall in product and maximize profits.

This is just one hypothetical example, but it gives you an idea of the potential of how intelligent, interconnected systems combined with inexpensive sensors will be crucial to the future of business. The truth is, the potential for the Industrial Internet is nearly infinite and will only increase as more information and experience is acquired over time, revealing patterns and trends in the oceans of data that are being created. Although this example is focused on a manufacturing business, the Industrial Internet will touch all sectors, from medical care to petroleum production. With so many different industries and so much technology, who is going to ensure

that all of this hardware and software from various different vendors is going to be compatible? Enter the Industrial Internet Consortium.

The Industrial Internet Consortium

The Industrial Internet Consortium (IIC) is a not-for-profit partnership established in March 2014 with the stated goal to, in its words, "accelerate the growth of the Industrial Internet by identifying, assembling, and promoting best practices." Its membership is international and consists of companies of all sizes, universities, researchers, academics and government organizations.

The consortium concentrates on three key areas, technology, testbeds, and security, to address issues regarding *interoperability*, *connectivity*, and *security*. It is extremely important to note that unlike most other technology consortiums (such as IEEE or W3C), the IIC has not been founded on the principle on creating standards. Rather, it is a facilitator of testbeds and dialog between the disparate member organizations, with the intention of giving them a common place to work together. The hope is that this cooperation will organically create common standards and reference designs that will be adopted across the various industries and verticals.

It should be noted that the IIC is not the only game in town when it comes to standards groups. Well before the IIC was formed, there was the Internet Protocol for Smart Objects (IPSO) Alliance, whose goal is to establish the Internet Protocol as the generally agreed upon protocol for IoT for the energy, consumer, healthcare, and industry areas. Other names that you might hear associated with IoT standardization are the AllSeen Alliance, the Thread Group and the Open Connectivity Foundation. However, the IIC, unlike these groups is solely focused on the *industrial sector*. On the other hand, over in Europe you have the Plattform Industrie 4.0 (the committee led by the German government to carry out the recommendations laid out by the Industrie 4.0 Working Group in 2013) and Mantis, a cooperative between various universities in the EU and private industry whose goal, in its words, is to "develop a Cyber Physical System-based Proactive Maintenance Service Platform Architecture, enabling Collaborative Maintenance Ecosystems."

Industrie 4.0

Industrie 4.0 is the German initiative to implement the technology and philosophies of the Industrial Internet. The first mention of the term Industrie 4.0 was at the Hannover Fair in October 2011, and subsequently it was adopted as part of the broader High-Tech Strategy 2020 in November of the same year. This was soon followed up by the formation of the Industrie 4.0 Working Group in January 2012.

The working group consisted of leading academics, researchers, and experts in a number of fields, such as information and communication technologies, production research and user industries, for the purpose of determining strategies and making recommendations on how to move forward. The responsibility of coordination and oversight went to the German Academy of Science and Engineering (Acatech) and was chaired by Dr. Siegfried Dais, who at the time was the deputy chairman of management at Bosch Industries, and Henning Kagermann of the Academy. The final draft of the findings and recommendations by the committee were presented at the 2013 Hannover Fair.

The report led to the creation of the Plattform Industrie 4.0, which is the alliance for the coordination of Germany's industrial digitization efforts. It is led by the German Ministry of Economic Affairs and Energy and the Ministry Education and Research. The government's hands-on role and their investment of €200 million for research demonstrates its deep commitment to the success of the initiative.

The Plattform itself has many striking similarities to the IIC. For starters, it was founded with the goal of bringing together leaders in academia, industry, and government for the purpose of tackling the major issues involving the implementation of Industrial Internet practices. Also similar to the IIC, the platform has a number of working groups dedicated to trying to solve the technical issues around Industrial Internet. In the case of Plattform Industrie 4.0, the working groups are broken down according to their concentration in the following areas:

- Reference architectures, standards, and norms
- Research and innovation
- Networked systems

- Legal framework
- Work, education, and training

However, unlike the IIC, Plattform Industrie 4.0 is purposely taking a less proactive role in helping to create standards. It has instead chosen to focus on a more advisory role, making recommendations and bringing together the various stakeholders together to discuss the issues and coordinate efforts. The committee then supports the projects to which the discussions give rise. However, the actual implementation is still left to the outside organizations.

The reach of Industrie 4.0 will not be confined to Germany, either. German Chancellor Angela Merkel is using her extensive influence on the other EU member states to begin adopting the Industrial Internet ethos. She has even gone on to state that her nation would actively cooperate with China to link the Industrie 4.0 and Made In China 2025 strategies. For Germany, as one of the leading suppliers of industrial machinery in the world, encouraging her trade partners to upgrade their manufacturing systems represents a massive business opportunity.

Also very much worth mentioning is the interoperability agreement between the IIC and the Plattform. In March 2016, the two groups agreed to a "clear roadmap to ensure future interoperability" with the possibility for more direct collaboration on test beds and standards development down the road.[3] It is likely that the two sides bent to the demands of many of the major players such as Bosch, Cisco, and Siemens, who are on the steering committee of both and likely felt they were duplicating their efforts. The consensus is that this will lead to less incompatibility between standards and more adoption of Industrial Internet overall.

Made in China 2025

Made in China 2025 is China's answer to Germany's Industrie 4.0, but it is even broader in its ambitions. The initiative was officially unveiled in May 2015 after it was first announced at the Lianghui Meeting earlier in March of the same year. It is the product of two and a half years of work by the Chinese Ministry of Industry and

3 You can find the full press release here: *http://www.iiconsortium.org/press-room/03-02-16.htm*.

Information Technology (MIIT) with input from experts from the China Academy of Engineering. In many ways, Made in China 2025 is meant to replace the *Strategic Emerging Industries* approach of the previous administration (led by Hu Jintao and Wen Jiabao) to advance the country's objective to become a global innovator.

So what exactly is the goal of this new strategy being adopted by China? In the words of Premier Li Keqiang, it is to:

> …seek innovation-driven development, apply smart technologies, strengthen foundations, pursue green development and redouble our efforts to upgrade China from a manufacturer of quantity to one of quality.

Analyzing the text of the policy, we can break it down into several components:

Innovation

Global investors have been bearish about China's economic situation over the past year with the country's GDP growth slowing to around 7 percent. However, the Chinese government seems to view the situation differently. It has embraced it as the "New Normal" and has set an annual growth target of 6 to 7 percent through the year 2020. Although some second-tier cities that rely on industries like steel production are experiencing recession, highly skilled workers are in huge demand in cities like Shanghai which are still seeing double-digit growth. Made in China 2025 is one of Beijing's tools for encouraging the expansion of the "new economy" as it begins to deemphasize construction, heavy industry, and commodity production that were largely state-owned or assisted. Investing in R&D, creating industry standards, and amassing IP are crucial to the strategy of transforming the country from a producer of raw materials and assembler of goods to a global innovation leader.

Quality over quantity

China accounts for 25 percent of all global production and is the leading manufacturer of mobile phones, air conditioners, and shoes, among other products. Top brands known for their excellence, such as Apple, Samsung, Nike, and Toyota, all have massive production facilities in the country or outsource to manufacturers in China. Despite this, the term "Made in China" is still largely synonymous with cheap, poorly made, and disposable products. This is not only a result of the many years of pro-

ducing low-end goods, but also because a significant proportion of China's exports are still not manufactured with quality in mind. Therefore, one of the the key components to Made in China 2025 is the emphasis on improving quality across the board so that local brands can flourish outside of the country. This is not unlike how Korean brands like Samsung transitioned in the mid-2000s from being a manufacturer of commodity household appliances to competing with Japanese and American brands like Apple and Sony at the upper end of the consumer electronics sector based on the merits of their build and design standards.

Green development

Even though China has, in the past, endured a fair bit of criticism for its perceived lack of environmental policy, Made in China 2025 demonstrates that Beijing is serious about cleaning up its act. In fact, the word "green" appears 46 times in the document text. The end goal here is to not only create products that are environmentally friendlier, but to build an industrial chain that is green at all levels. Their motivation for this couldn't be more straightforward: China understands that creating new green factories and retrofitting existing facilities to be cleaner and more energy-efficient not only has a positive environmental impact, it's an opportunity for economic growth.

Apply smart technologies

The crux of Made in China 2025 comes down to modernizing the manufacturing sector by using smart technologies. This is to say that the state intends to realize the preceding points by applying the tenets of the Industrial Internet. It has been noted by many that China has fallen behind in bringing some of its older factories up to current specifications and many processes that should be automated are still being done en masse by human hands. The slowdown in Europe and the US gives China the opportunity to catch up with heavy investment in high-tech tooling, robotics, networks, and computer systems. On the other end, the leading manufacturers in the nation, who have consistently made investments to keep their facilities up to date, will be given additional support from the state. China is the largest market in the world for industrial automation and robots, and with Made in China 2025 it will become even big-

ger, while simultaneously expanding the role of IoT and Big Data Analytics.

The state's plan to ensure the success of Made in China 2025 comes in the form of subsidies and incentives, policy reform, and financing for projects to further these goals. To encourage invention, Beijing will tighten up intellectual property rights protection—particularly for small and medium-sized enterprises. It will also create manufacturing innovation centers throughout the nation, with the goal of having 15 up and running by 2020, and 40 by 2025.

It is also important to note that Made In China 2025 dovetails with another major initiative known as *Internet Plus*. As you can probably tell from the proposal's name, the idea is to upgrade the Internet in China. This will involve improvements to the network infrastructure and expansion of broadband availability as well as the integration of mobile Internet, cloud computing, Big Data, and IoT. It also entails converging consumer IoT with IIoT technology in a variety of sectors, including medicine, government, and finance. In terms of the manufacturing area, the idea is that consumer IOT will enable companies to gather data, monitor, and remotely control machines and devices among many other functions. The data gathered informs the manufacturer of the performance of their products under real-world scenarios, which can be fed into their systems for analysis. This in turn improves efficiency and quality throughout the entire supply chain.

In the opinion of Wang YaBing, senior consultant for Baifendian—one of China's leading Big Data consulting companies—Made in China 2025 represents an opportunity for China to drag some of its industries into the modern age:

> These types of more traditional manufacturers, who have mostly not utilized the Internet, will be able to take advantage of these new models to make massive advances in a number of areas which include logistics, production, and sales and marketing.

Mr. Wang has said that at Baifendian they are already in the process of implementing several Industrial Internet projects with companies in the pharmaceutical and accessories industries. Their technology is also at the heart of a system to monitor and detect failure for TV broadcast equipment.

He also added that, although China is putting effort into trying to help small and medium-sized manufacturers compete, it's likely that

the nation will see the rise of mega corporations, as the technology gap increases between the larger firms and small and medium-sized companies:

> For example, (the state government) is assisting with implementation of many modernization projects. Many manufacturing and Internet technology companies are teaming up to innovate in the design of new products—and in this area the Chinese Government is quite supportive…However, these are the top thousand companies in the nation with experienced staff and resources to execute projects of this type. However, small and medium sized manufacturers—such as companies in the metals and the fast-moving consumer goods (FMCG) industries—are going to face a number of tough challenges catching up. And, I think it's likely that many of these companies will be absorbed into larger firms and it's very possible that China will experience a situation similar to Japan where you have a number of mega-sized companies dominating the market in the future.

(Industrial) Internet of Things

This *Internet of Things (IoT)* is garnering a massive amount of attention in the media. When applied to the industrial area, it is commonly referred to at the *Industrial Internet of Things*, or IIoT. In this section, we will examine how IIoT tech is going to revolutionize how manufacturing gathers and processes data. We will also examine a real-world example of a company that's taken the plunge and realized real gains as a result of implementing IIoT.

Before we dig deeper, we should probably take a moment to clarify what IIoT actually is. The truth is there is no consensus on the exact definition of the term. In many publications, IIoT and Industrial Internet are one and the same—it is the whole package from the Big Data Analytics systems, to the automation and robotics, to sensors and monitoring devices, and beyond. On the other hand, others consider IIoT to be limited to the embedded systems, controls, sensors, and monitors on industrial equipment and the software systems powering them. For this purposes of this report, we will go with this definition.

The standard definition of the IoT is "the network of physical objects—devices, vehicles, buildings, and other items—embedded with electronics, software, sensors, and network connectivity that

enables these objects to collect and exchange data."[4] This definition omits any mention of the backend systems involved in processing the data. At times you'll even see the term Industrie 4.0 used interchangeably with Industrial Internet. Whenever industry buzzwords take off, there will always be discrepancies in their interpretation and a variety of interests that want to control the terminology.

If you are even a casual observer of technological trends, you know how hyped the IoT has been in recent years. The number of Internet-connected devices will be more than 20 billion within 5 years by reasonable estimates, and adoption will only accelerate from there.[5] Although it is important to distinguish the industrial side of IoT tech from the consumer and commercial sides, there will be a symbiotic relationship between all of them, and successful implementation of all three will be necessary for corporations with their eyes on the future.

When specifically talking about IIoT in manufacturing, the focus tends to be equipping assets with sensors and/or networking capabilities. These assets include everything from the parts in inventory, to the machines on the production line, to the facilities themselves. By monitoring the state of machines and the products *as they are being manufactured*, a much higher resolution picture of what is happening at any particular moment can be assessed in ways that would have simply been impossible in the past.

It should also be noted that even though a lot has been made of attaching relatively inexpensive sensors and network adapters to existing and new industrial machines as the breakthrough innovation in IIoT, the fact is that nearly all industrial machines already come equipped with a plethora of sensors and a large percentage also have some type of networking capability. Although this sounds like it would save a great deal of effort, the truth is that it more often creates its own unique set of challenges. For decades, manufacturers of industrial equipment have developed data formats, protocols, and even networking hardware that are industry specific or even proprietary. So, whereas the IT industry has now evolved to the point where commodity hardware and protocols have become the de facto standard for nearly all except the most specialized applications, the

4 *https://en.wikipedia.org/wiki/Internet_of_Things*

5 *http://www.gartner.com/newsroom/id/3165317*

manufacturing world remains extremely fragmented. Regarding the variety of machines, protocols, and data formats in the industrial space, Nathan Oostendorp, CTO and cofounder of Sight Machine, explains:

> As far as the number of types of machines, tens of thousands to hundreds of thousands is pretty conservative. There's a lot of automation that is built to spec. With transport mechanisms and protocols, that's covered by a couple dozen cases. But in terms of what the data says and how it needs to be mapped, that is really a massive variety problem and I don't think it's something where you're going to be able to compile the list of all known machines and walk into a plant completely cold and have a turnkey solution that takes all of its data automatically and gives you a perfect digital picture of what's going on. There's always going to have to be some amount of modeling that goes on, and some information that comes from the process itself that will inform how you're going to report on that data.

A Platform Built for Manufacturing

Despite these hurdles, Sight Machine has seen real-world success with its product, which they tout as a "Big Data Analytics platform that was built exclusively for the manufacturing industry." One such case Oostendorp went on to describe involved a major automobile manufacturer that was able to reduce the number of defective parts being produced by applying *Industrial Internet principles*:

> One of the Big Three automakers was really interested in understanding the root cause of quality problems. It was really like an internal supply chain–type problem. A part would be cast at one plant, machined at another plant, assembled in a third plant. At every stage different serialization schemes were being used. Different tests were being performed, but the data wasn't necessarily being fed to the process upstream in order for them to improve their processes. So when they instrumented this process with Sight Machine, it allowed all the plants to look at the same data and to understand how things that were happening in the casting process or the machining process had an impact all the way down to when the pieces were finally fit and what sort of defect states were in that. This solved a really big problem that had been plaguing this company. With Sight Machine, they detected and eliminated 5 times the defects they suspected they were suffering.

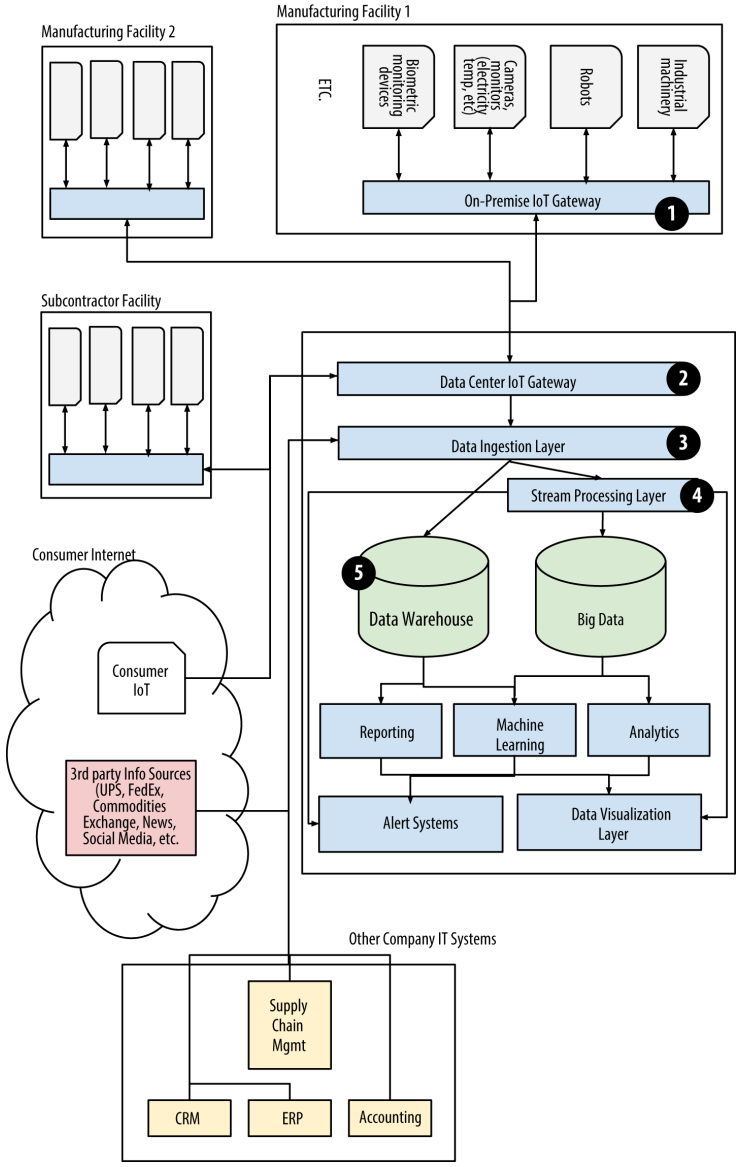
Most initial Industrial Internet initiatives are focusing on what many consider the low-hanging fruit for their initial trials. A couple of goals that you'll see mentioned repeatedly are the concepts of doing

predictive maintenance and anomaly detection. This tends to be because these are the problems that are the easiest to tackle and obtain a high return on investment. We will discuss these two particular use cases in further detail later on.

But the fact remains that the vast majority of legacy equipment and machines are not IIoT-ready. Matthew McNeely, founder and head designer and engineer at Nimble Industry—a company that has developed a product that lets industrial equipment manufacturers equip their new and existing machines with IIoT capabilities—puts it succinctly by saying:

> Not until industrial equipment manufacturers begin embedding modern distributed computing intelligence into their products will the full potential of the Industrial Internet be realized.

Big Data and Analytics



1. The duty of the on-premise gateway is primarily to handle security and authentication between the facility edge nodes and the remote IoT gateway. It also brokers the transmission of data

between the nodes themselves and remote data and control sources. Some may also handle data normalization and will perform some amount of analytics processing—often referred to as edge processing. While not pictured, the gateway sits behind a firewall and is not directly connected to the Internet.

2. The data center IoT gateway is responsible for two-way communication between remote devices and gateways. It handles connectivity, security, and authentication. While not pictured, the gateway sits behind the firewall and is not directly connected to the Internet-at-large.
3. The data ingestion layer handles the influx of data from the variety of sources and distributes it to the repositories.
4. The stream processor is responsible for doing real-time analytics on the data coming in. It does not hold onto the data for extended periods of time.
5. While the Big Data system will hold onto the data, many enterprises will still elect to collect and hold onto their data in a Data Warehouse for security and permanence because Big Data is often not considered a reliable long-term storage solution.

Big Data is certainly in no way a new concept, and ever since companies began collecting large amounts of information, they have been trying to find ways to analyze it to gain insights into their operations. But, steady developments in technology over the past few decades are now making it possible for these same companies to pull in a greater variety of data from a far more diverse number of sources, store this massive volume of information, and then do deep, impactful analysis on it—often in real time. These developments in technology include the following:

Internet infrastructure
: The universality of Internet infrastructure—in both wired and wireless forms—coupled with the expansion of available bandwidth makes it possible to transport large volumes of data efficiently and economically.

Data storage
: The simultaneous massive increase in the capacity of data storage and decrease in price combined with ever faster data access and write times makes holding onto enormous amounts of digital information viable. Further, considering how inexpensive

RAM has become, it is becoming possible to store an entire working set of data *in memory*, making in-memory computing realistic and affordable.

Computation power
: The exponential growth of computational power in commodity CPUs, which is then further multiplied by distributed computing make it possible to model very large data sets without the need for what, in the past, would have required custom-built supercomputers.

The term Big Data is not unlike IoT in that it's a buzzword without an official definition, but it is generally felt that Big Data is characterized by the three V's: velocity, variety, and volume. This means that Big Data systems must be capable to dealing with a large amount of disparate data types coming in at great speed.

The earliest adopters of Big Data tended to be in the fields of science and technology, and finance and marketing. The reason for this is that individuals and organizations in these areas have use cases that are obvious and their data is, for the most part, better structured and available in digital form; this frequently isn't the case for companies in the manufacturing sector. So even though manufacturers have just as much to gain, if not more, from what Big Data offers, implementation in this environment poses a unique challenge. It is for this reason that IIoT, for manufacturing firms, has provided the missing piece to the puzzle.

Hardware

One of things that will prove to be a great relief to many IT managers and CTOs is the across-the-board standardization of the hardware that is necessary to power Big Data systems. All of the experts that were interviewed for this report stated that they either directly used commodity hardware (in the form of x86-based processors, HDD/SSD, and memory) either in their own facilities or via a cloud service.

All of the implementations of Big Data systems that were examined for this report were fully capable of scaling both horizontally (by adding a new machine to the cluster) or vertically (by upgrading the hardware within the servers in the cluster). The benefits of this cannot be understated—the reliance on off-the-shelf components greatly reduces cost and complexity. It also means that many compa-

nies can begin testing out Big Data processing with little-to-no initial investment. Apache Hadoop and Apache Spark are two open source solutions most popular among data scientists that can be run on a single machine and be scaled up from there.

Platforms

For this report we examined several distinct approaches to implementing Big Data Analytics in a manufacturing setting. This is in no way a complete list, (notable omissions from this list include Sight Machine's platform, which we talked about in the previous section), but it should give you a reasonably good understanding of what kinds of options are out there and the pros and cons of each type of these implementations.

Apache Hadoop, Apache Spark

The biggest names in open source Big Data are Apache Hadoop and Apache Spark, and, although there are other open source solutions, none of them have the same loyalty and install base. These projects also have the benefit of having a large number of tools as well as reporting and analytics solutions available as a result of their popularity. For organizations that want a more feature-rich version of Hadoop, an entire industry has grown up around creating commercial distributions of the widely used platform. These companies generally offer support and have connections to consulting services to assist with implementation and maintenance.

Building a custom solution for your organization means that you can set up the system to suit its specific needs. As mentioned before, getting started with Hadoop can be done with minimal initial expense. And, if a company opts not to host its cluster in its own data center, there are cloud providers that can get that company started with Hadoop/Spark simply by signing up—most notably Amazon Web Services' (AWS) Elastic MapReduce.

The obvious benefit of running your own Hadoop cluster is that all of your data stays on your own machines, within your control. But this means you are also responsible for doing all of your own security, upgrades, and for building the network infrastructure; not to mention that you'll need the right staff to properly manage the cluster. On the other hand, if you go with a cloud-based host, you are entrusting a third party with your data. Although in general this

shouldn't pose an issue,[6] depending on the sensitivity of the information, it might be a deal breaker for some.

Then, there is the final issue of Hadoop not being purpose-built for industrial manufacturing, from the start. Hadoop is the result of Yahoo's attempts to build a better, more scalable search engine. So, even though it has proven to be extremely versatile in a variety of scenarios, at the end of the day a company in the business of making sneakers has very different needs from a web marketing company. At this time, none of the commercial Hadoop distributions have features that are specifically for the Industrial Internet (although this will likely change soon). This means that making Hadoop work for manufacturing at this time will almost certainly require a lot of custom development, which equates to additional investment in both time and money. This doesn't mean that Hadoop can't work as part of a solution from other vendors, however. As we will see, even if you don't use it as the primary backend for your Big Data deployment, you can stream data from a variety of sources, into a cluster, so you can still take advantage of Hadoop's feature set for various purposes.

AWS Big Data/IoT

As mentioned previously, AWS is a suite of cloud services that supports both Hadoop and Spark. But looking beyond these tools, AWS also have a plethora of other solutions for companies that are looking to develop a Big Data project. Going into detail and comparing the pros and cons of each would take up a report unto itself, so we will only touch upon some of the more noteworthy components.

Redshift
: This is AWS's data warehousing solution that accommodates fairly pain-free storage of petabytes of your company's data. Beyond the clear benefit of simply being able to turn on the service and get started for relatively little cost, Redshift uses standard SQL for querying. This means many existing business intelligence, analytics and reporting tools are compatible with Redshift. It also means your business intelligence staff should be able to pull the data from the warehouse by using a query lan-

6 For reference, you can look at AWS's FAQ on Data Privacy at *https://aws.amazon.com/compliance/data-privacy-faq/*.

guage and technology they are already familiar with. For many companies, a solution like Redshift will be capable of fulfilling the majority of their Big Data needs.

AWS IoT
: This is a platform that allows for the bi-directional (push and pull) of data from a variety of IoT devices and applications—including industrial hardware. In addition to being easy to set up and develop for, this component is touted as being highly secure. The drawback, however, is that with this heightened level of security, many embedded devices and Programmable Logic Controllers won't be able to directly connect to the service, necessitating some type of proxy to handle the authentication on their behalf and adding a layer of complexity. Still, the added attention to security should be considered a plus for the nascent platform.

SQS
: SQS is AWS's message queuing service, which is better suited for handling the stream of information that comes from sensors that are constantly monitoring the machine state known as *time series data*. It is designed to be used with AWS data storage solutions such as DynamoDB (AWS's NoSQL DB) and Redshift. From there the data can be moved to an Elastic MapReduce instance, using the Data Pipeline service.

It should also be brought up that one of the major benefits of building out your solution with offerings from AWS is that they can be mixed and matched as you please. So, if you want to use AWS IoT in concert with an AWS Elastic MapReduce, it will be fairly straightforward to implement. AWS services are billed based on usage, so you only pay for what you use. But this also means that if you aren't careful with how you implement your solution using the services, you might end up with sticker shock when it comes time to pay the bill.

One of the interview subjects for this report says that his organization—a major manufacturer of printers and photocopiers—created a solution to monitor and manage its leased assets almost entirely using AWS services. According to him, without AWS, the project simply could not have been completed from a budget and project scope standpoint. He also expressed his satisfaction with the service's performance and pointed out that his organization, although initially skeptical of having Amazon host its data, managed to work out

special contract terms that put the executives at ease by not allowing AWS direct, physical access to the data.

GE Predix

One of the newest offerings in this space is GE's platform for Industrial Internet named Predix. As the originator of the term *Industrial Internet*, the one-and-a-quarter century old industrial behemoth has a great deal of skin in the game and has invested heavily in every facet of this emerging tech. Predix touts itself as a Platform as a Service (PaaS) and, although there is discussion about allowing customers to host their own instances of Predix, GE strongly encourages users to go with its cloud-based, hosted solution[7]. Much like AWS, Predix uses a usage based payment scheme, so your company only pays for what it uses. And, although Predix is underpinned by open source technology from Cloud Foundry, there should be no doubt that the platform is built strictly for use with the Industrial Internet.

At its core, Predix is designed to ingest, store, and process machine data. Furthermore it provides various packages for building out analytics and even has SDKs for companies interested in developing mobile apps for monitoring and management of their assets.

One very important aspect of Predix is that any enterprise that wants to build a solution using it can do so whether or not they operate GE industrial machinery. And out of the box it should provide manufacturing more of the tools that they need to be successful.

In the words of Gytis Barzdukas, head of product management for Predix, "(Predix) has things like machine data, asset data and time series data storage. And analytics services, analytics runtime, and analytics catalog that have been designed to work with industrial assets and are therefore a tier above what you get from open source solutions based around databases—they are really targeted at industry."

7 This technical decision, according to Mr. Barzdukas, is due to the cloud's ability to access additional processing power on demand when performing computation-heavy operations such as analytics. He added that in 2016 the company will be rolling out a "hybrid model" which pushes some of the computation to the edge devices but will still require a cloud-based instance.

One caveat to all of this is that, as of the time that this report was written, Predix does not offer Hadoop as part of its platform, which is a bit of a surprise due to its popularity among data scientists. That said, Barzdukas did go on record as saying that the platform will "support Big Data technology like Hadoop in the future."

Siemens Sinalytics

Siemens' Industrial Internet platform is known as Sinalytics and, while it offers similar features as Predix, it uses a very different business model. For Siemens, "Sinalytics is used to deliver services by Siemens, to our customers, whereas Predix is basically the market that's directly (facing) the outside world." According to Matthias Goldstein, VP of Digitalization at Siemens Corporate Technology. So, unless you have a contract for Siemens equipment, you won't have access to the platform—at least for the time being.

The company really sees Sinalytics as a value add for their customers and it has already been used successfully in the field for doing monitoring, predictive maintenance, and anomaly detection, e.g., for the Munich-based giant's rail and energy projects. Although my interview subjects were not at liberty to discuss the names of some of Sinalytics' manufacturing customers, they did say that they have clients in the pharmaceutical and food production businesses that are already putting the product through its paces.

Key differences between Sinalytics and Predix include the aforementioned business models, but also Sinalytics flexibility with regard to its deployment schemes. Says Goldstein, "We have different deployment models. On-premises, hybrid, cloud so we have a more decentralized approach…[Sinalytics] uses a more open, flexible, customizable way to deliver data analytics matching very different needs in the industries we serve."

Machine Learning

The field of artificial intelligence has been around for decades, and the world has seen massive advances in what is considered *deep learning* (e.g., IBM's Deep Blue and Google's AlphaGo), but it's only within the past decade that we've seen practical applications of machine learning in an enterprise setting. In the past few years, there has been an explosion in the number of products available that integrate machine learning within a business intelligence platform.

In a manufacturing setting, machine learning is used mostly for finding patterns in industrial data for the purposes of *anomaly detection* and *predictive maintenance*. Anomaly detection is certainly not specific to manufacturing, but it is used differently when applied to manufacturing-specific problems.

Anomaly Detection

In looking for abnormalities, the first step is to establish what is *normal*. Organizations that already have historic data have a leg up in this area because this data can be fed into most machine learning systems to help establish the necessary baselines. Unfortunately, if an organization lacks *existing* data, the system will need to observe the data *over a period of time* before it can be confident about what to expect. This period of time can change depending on the enterprise, and whether activity varies greatly from season to season, for example.

Manufacturers can benefit from anomaly detection in a number of ways; a prime example is by using it to discover defective products early in the production pipeline. Early anomaly detection can give machine operators advance warning of issues downstream in the manufacturing process so that these issues can be resolved quickly and without shutting down the production line.

Predictive Maintenance

Predictive maintenance is a subset of anomaly detection that focuses on determining the mechanical status of a machine—for example, whether a machine is approaching its maintenance window or if failure is imminent. By comparing current sensor readings to historic data, the system can use predictive maintenance to detect issues early on, letting the company handle repairs at a time when overall impact to the system is minimal. This level of prediction can prevent costly and unplanned maintenance as well as lost earnings that might otherwise arise and affect service agreements.

Applications in Machine Learning

Both GE's Predix and Siemen's Sinalytics incorporate machine learning algorithms in their platforms, and Amazon AWS Machine Learning and Microsoft Azure Machine Learning are both commercially available services for companies that already have Big Data

implementations and would like to add machine learning capabilities. There are also smaller companies that are bringing machine learning to industrial sector clients, such as Anodot and Plat.one.

Current machine learning environments are also far more user friendly than ever before. Most modern machine learning tools are rules based and even have GUIs to help build models. Many of these models can be built by business intelligence staff and data scientists who have knowledge of how to do some scripting, and they can be deployed on-the-fly, without custom code.

More advanced machine learning features include asset simulation, in which industrial machines and facilities are modeled in software, to simulate a variety of scenarios. This capability will let industrial enterprises find ways to optimize all of the variables in their assets to maximize efficiency for any situation. In GE's Predix, this feature is called the "Digital Twin," and although it has yet to model any manufacturing assets using the tool, it claims than nearly any kind of machine can be simulated using this software.

Natural-Language Processing

One of the biggest challenges in analyzing data from industrial machinery is finding the *meaning* in the data (data such as error codes and sensor readings). Data formats are often buried deep in service manuals—meaning that much of this information needs to be mapped into systems *manually*, before it can communicate any meaning to the actual systems. Steven Gustafson, leader of the Knowledge Discovery Lab at GE, explains:

> [In a factory,] we have many different kinds of machines provided by many different manufacturers. They're usually connected to control systems in basic ways just for alarming, safe shutdown, and other safety features. And now we want to have a whole plant view of what's going on, so we can do optimization. Machine learning is already having a big impact, and the main way is on the data side.
>
> So, we need to do a lot of work to get data structured, and that could be from looking at using natural language processing, and extracting the learnings from plant failures, machine failures, or from other issues, and getting them out of reports. …Because, if you took a plant that might have dozens of different kinds of systems that are generating alarms, those alarms usually come with a numeric format, with a string, that is a description of the problem. And, surprisingly, a lot of the natural language processing work involves going through and normalizing all of that alarm informa-

> tion, so that when it flows back in, it is in a digitized form—I like to call it a "computable form"—then we can do automated inference reasoning on it.

Autonomous Robots, Augmented Reality, and More

One of the most important things to stress is that Industrial Internet is not so much new technology as it is the *implementation* of a number of technologies that are now coming into maturity. Without a doubt, Big Data and IIoT are the most important of these emerging technologies, but they are part of a larger ecosystem that will shape the future of industry.

We have explored IIoT and Big Data in depth, but following are four other technologies that will reshape the manufacturing landscape in the coming years.

Autonomous Robots

The systems controlling future generations of robots are going to have complex processors and AI algorithms on board. They will be among the many edge nodes sharing data and cooperating with other machines and humans in concert. Currently, about 10 percent of the world's labor is done using robots. According to estimates, this percentage will jump to 25 percent by 2025. This increase is being driven by the increasing cost of manual labor and the decreasing cost of robotic equipment. More important, robots have advanced to the point at which they have the dexterity to compete with human hands in tasks for which they were previously too clumsy.

Not only does implementing robots make manufacturers more competitive, it prepares them for a future in which the labor force will shrink. Both China and Germany will be hit hard by a combination of several decades of low birth rates and a large number of older citizens leaving the workforce. In the case of China, the population of working age adults is expected to drop four percent (from 1 billion to 960 million) by 2030. The use of robots will mean that human labor can be reappropriated to do tasks that require greater cognitive capacity and less repetitive movement.

Simulation

Future factories will be able to do simulation runs for new product lines before they actually make the changes to the machine tooling and settings. This will reduce costs by providing a way to work bugs out of the software well before a single product enters the physical world, resulting in reduced time to bring a product from the design phase to retailers' inventories. An example of this technology at work is the aforementioned GE Digital Twin.

Additive Manufacturing

3D printing and rapid prototyping technology are already essential to the design phase of products, and we are seeing companies add value through product customization to increase profits (for example, the myriad options on today's automobiles). This means that the industrial machines of the future need to be dynamic to assemble these increasingly complex products with their many variations. Programmable Logic Controllers (PLCs), with their relatively static ladder logic, will be replaced by machines capable of receiving special instructions for each item being assembled on a line, and adapting to what and how it needs to execute its tasks, based on the requested options and customizations.

Augmented Reality

When sensors and data become omnipresent in manufacturing centers, implementing Augmented Reality (AR) will no longer seem like a pipe dream. In the future, an engineer will simply glance at any machine on a factory floor and see its diagnostic sensor readings (such as temperature, telemetry, wear and tear), the service history, and the manuals and schematics. Assembly floor workers will be able to look at the item that they are working on at a particular moment and see what model the item is, what options it has, and what tools and parts they will need to complete the job. When a worker asks a question out loud, the system will promptly respond with the answer. AR will assist humans as they work side by side with automatons to bring about larger productivity gains.

Challenges

Any enterprise embarking on a major IT project is going to experience some pain during the process, but Industrial Internet projects can be especially daunting considering all of the parts of your organization that will be affected. Despite this, the gains from increased automation, monitoring, and data analysis far outweigh the cost and effort for manufacturers who want to stay competitive. To avoid making potentially fatal errors, enterprises need to be aware of the challenges that lie ahead and plan accordingly. Aside from dealing with issues related to the lack of standardization, budget, and organization, here are some of the major challenges you should expect encounter as you begin incorporating an Industrial Internet project:

Security

Easily one of the top concerns of enterprises when considering an IT project is, "Is it secure?" All the benefits of developing an Industrial Internet solution are worthless if they put a company at added risk of cyber attacks and espionage. One of the most valuable assets of any organization is its data, so it makes sense to be overly cautious when approaching this problem. According to Urko Zurutuza, coordinator of the Telematics Research Group at Mondragon University:

> When factories start connecting IT networks to OT [operational technology] it can introduce problems, because these networks were completely isolated before. The OT networks have lots of old OSs running that work well for that process, but they are not reliable for communicating with other networks. So, that means some malware or virus that comes in through the IT network can spread to the other part. And that's a very dangerous issue.

Fortunately, with the explosion of Industrial Internet projects has come a commensurate increase in companies offering products and services for this specific market, and to meet this very challenge. Cisco, a market leader in the field of networking, has long been involved with selling products and services for securing industrial networks and is one of the founding members of the IIC. Infineon also manufactures and markets products aimed at protecting industrial networks. Windriver, a subsidiary of Intel, is also very active in this space and has developed a product called Intelligent Device Platform XT for the security and management of IIoT assets. And, GE has spun off its industrial network division into its own company called Wurldtech, which offers secure devices (marketed under

the OpShield name) as well as consulting services and security auditing.

Data Integration

A massive task that should in no way be overlooked is the amount of *data integration* that will be necessary to get the most out of any Industrial Internet project. As a first step, most manufacturers will concentrate on creating a secure and stable environment so that they can begin pulling this precious information from their facilities and assets. But, being able to analyze and visualize this data is only the beginning. The true value of this new wealth of data can be realized only when it can be *correlated* with the other data within an organization from the CRM, ERP, supply chain, and operations systems. Looking beyond the enterprise itself, integration between customers and suppliers, and contractors and subcontractors will create new insights and streamline many processes. However, integration is not only an IT problem: it's an organizational one. As Stephen Gustafson explains:

> In the past, data was a high-value asset within organizations. Folks would find out how to get value out of it and it wasn't always shared as broadly as it should be because it was so powerful. And so we really have to push this culture change of making data available to everybody who needs access to it. But that's enough because I can send you all the data sets that I have but you wouldn't know what they are. And even if they had some kind of meaningful labels, you still wouldn't know what the context is. And so one of the big challenges is to have this across the company first, and then across industry semantics on the data environment.

So, this challenge is two-fold. The first part involves aligning the goals of all of the stakeholders, at which point organizations can begin to go about the monumental task of tackling the integration problem.

Staff

It has been said that "good help is hard to find," but when you're dealing with emerging technologies, it can appear almost impossible to recruit the talent necessary to get these complex Industrial Internet projects off the ground successfully and running smoothly. It seems like most recent graduates in the fields of computer science and statistics are drawn to the prospect of fast money from tech

startups. Yet, as it becomes clear that there is an equally prosperous alternative path, we will likely see a new generation of young talent that is interested in working on large-scale enterprise systems.

To ensure that they have the right staff to successfully execute their Big Data initiatives, 49 percent of large industrial companies are creating positions for chief analytics officers, and 50 percent are forming specific groups within their companies; 63 percent are stepping up their recruitment efforts, and 54 percent of these enterprises plan to team up with various consulting firms and vendors to help meet their demands.[8]

In the near term, there is likely to be a shortage of talent for Industrial Internet projects. However, the upshot is that the dearth of capable staff is likely a temporary phenomenon, with a steady rise in experienced labor as the Industrial Internet transitions from its infancy to maturity.

Conclusion

Much like the emergence of the Internet completely revolutionized the way the world communicates, so too will the Industrial Internet transform the operations of companies that depend on large machinery. The Industrial Internet is an ever-changing landscape, and it seems like there are new developments every few weeks, if not days. Staying on top of what's happening can seem like a full time job unto itself. Undoubtedly, in the time between when this article was written and when it is published, major new developments will be announced. It's both an exciting and daunting time to be in the manufacturing business. I hope that this report was insightful and has provided you with the information and inspiration to make this data-driven future a reality.

8 *http://bit.ly/1PIzxE6*

About the Author

Li Ping Chu is a veteran software developer of the Silicon Valley tech boom. With 15 years of working experience ranging from five-person startups to consulting for major financial firms like Charles Schwab, and major e-tailers like The Gap and Williams-Sonoma, he has been involved with projects of all kinds and all sizes. He is currently located in Taipei where he most recently helped build an analytics engine for a local mobile gaming company. He loves dogs and tolerates cats.