O'REILLY®

Speeding from Data to Insight in Financial Services

Best Practices for Getting Actionable Insight from Data Early and Often

Alice LaPlante

Speeding from Data to Insight in Financial Services

Best Practices for Getting Actionable Insight from Data Early and Often

Alice LaPlante

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

Speeding from Data to Insight in Financial Services
by Alice LaPlante

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://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Rachel Roumeliotis
Production Editor: Nan Barber
Copyeditor: Octal Publishing, Inc.
Proofreader: Charles Roumeliotis

Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

May 2018: First Edition

Revision History for the First Edition
2018-04-25: First Release

This work is part of a collaboration between O'Reilly and Kyligence. See our *statement of editorial independence*.

978-1-492-03310-3

[LSI]

Table of Contents

Speeding from Data to Insight in Financial Services

Introduction: The Data Deluge Is a Blessing—and a Curse

Today's markets are increasingly unpredictable. Across all industries, only half of top-performing companies maintain their leadership over 10 years, according to McKinsey. Businesses competing in industries that are in the midst of digital transformation face even more volatility as *digital-native* firms come out of nowhere to displace them. Such firms have the advantage of no aging legacy processes or infrastructure to hold back their fresh ideas and Agile strategies.

Financial services firms have some of the highest stakes. On top of market uncertainty, they must combat fraud and deploy robust security measures to meet a growing set of regulations—all while combating digital-native upstarts that are redefining their industry.

What can help? Data. More specifically, mining data for actionable insights that helps with critical decision making.

Research shows that even the largest, most entrenched incumbents, if they make wise investments in digital technologies—especially data analytics—are as likely to steal revenues from traditional players as digital natives. Indeed, frequently the fast-moving incumbents are the ones creating life-threatening competition to "slow movers".

But data today is both a blessing and a curse: a blessing because making wise use of data is the number one competitive advantage a business can possess today; a curse because there's too much of it, and it's difficult to access.

Estimates of just how much digital data exists in the world vary considerably. But most experts agree that it's more than we humans can easily grasp—and that it's growing at astonishing speeds. Ninety percent of all data has been created in the

past two years. And it's expanding by 2.5 quintillion bytes a day. These numbers are almost too large to visualize. To make them more concrete, understand that a quintillion seconds is 32 billion years, approximately twice the age of the universe.

Less than one percent of all this data is being captured and stored for analysis. This means that a treasure trove of data is out there that's worth investigating.

Naturally, businesses are attempting to exploit this data. They're spending big bucks to do so. According to IDC, worldwide revenues for big data and business analytics will grow from $130.1 billion in 2016 to more than $203 billion in 2020, at a compound annual growth rate (CAGR) of 11.7%, as shown in Figure 1-1.

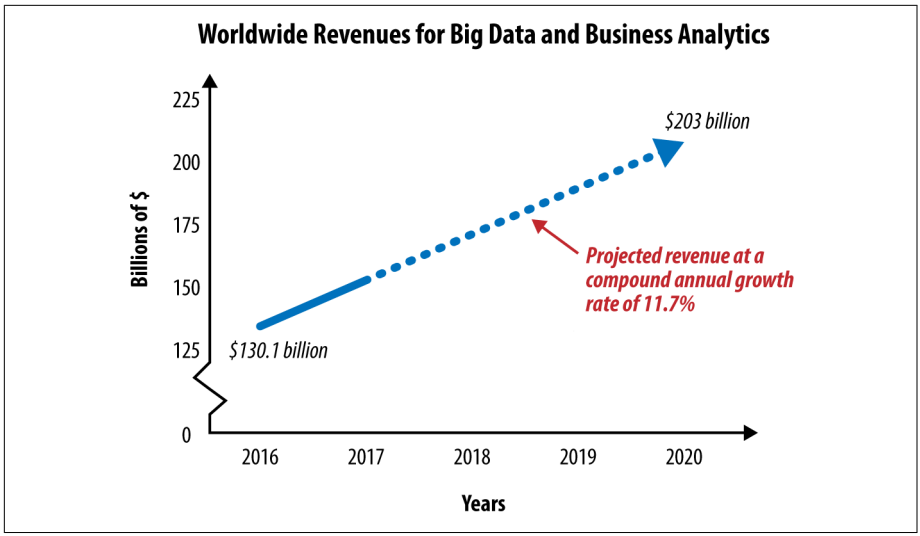

Figure 1-1. Worldwide Revenues for Big Data and Business Analytics

So, how's it going so far?

Not so good. Although more than 85% of enterprises say they've jumpstarted programs to create data-driven cultures, only 37% report success.

That's because simply being able to access data is already out of date. Instead, businesses need actionable *insight* from data, early and often.

This is most difficult for established businesses. Whereas digital natives are built from the ground up to take advantage of big data analytics, legacy companies have to do the hard work of overhauling or changing their existing IT environments. And transitioning to a data-driven era is not a simple proposition. Numerous companies have made significant investments in big data technology but have yet to realize the returns they were promised. Many of them are struggling to transform themselves culturally—a major requisite to becoming data

driven—but lack the technical infrastructure and expertise, so they're are stuck with archaic business processes that prevent them from capturing value from data analytics.

Many companies are turning to open source solutions. Today, open source big data technologies like Apache Hadoop and Spark are quite established in even the most conservative financial services organizations. One survey found that nearly 60% of enterprises expect to have Hadoop clusters up and running within 12 months. Forrester reports that use of Hadoop is rising approximately 33% annually.

In this report, we talk about the challenges raised by the data deluge for financial services companies and the barriers—technical and otherwise—that they face in making the most of data analytics tools. We examine the current data analytics tools landscape and introduce Apache Kylin, an open source technology that originated at eBay, as addressing many of these challenges. And we provide stories and best practices from market leaders in financial services that are deriving true value from their data troves through the use of Kyligence, an enterprise-class version of Kylin.

Big Data Trends in Financial Services Companies Today

Financial services firms—especially banks—are smack in the middle of the data deluge. In addition to being the industry with the largest investment in big data and business analytics solutions (nearly $17 billion in 2016), financial services will also see the fastest spending growth on big data analytics tools in coming years, according to IDC.

What are financial services firms doing with these tools? They're using data for everything from fraud detection to risk management, to enhancing the customer experience. They're also coming up with new products and services and new business models to sell them. Clearly, big data has turned the corner from the proof-of-concept (PoC) stage to full-scale deployment in the financial services industry.

But the scale of these investments is putting pressure on financial services CIOs (chief information officers), CDOs (chief data officers), and other executive leaders of big data initiatives to show results. And, according to a recent McKinsey survey, measurable results are not easily forthcoming.

Financial services firms face three challenges in particular:

A major technological and cultural shift is necessary to be data driven
: Because most leading financial services firms are not digital natives, they have legacy IT infrastructures—and legacy attitudes—that need to be transformed. Big data requires big changes, and without significant collaboration and consensus, large investments, and commitment across the enterprise, progress can be very slow.

Security remains a major concern
: It has been effectively argued that for financial services firms, information is as valuable as the currency they handle. They need to manage it as such. To "democratize" data and put it into the hands of business users to make better decisions opens them up to greater risk. They must somehow mitigate this risk.

Data talent is scarce
: This is a global problem, with financial services firms in Asia competing with banks in North America and Europe for skilled data professionals. This will ease as time passes, of course, but recruiting top talent today requires considerable effort and resources.

However, although most big data initiatives in financial services are young, the industry is still on the leading edge of what's being done with data. Financial services firms are already using data to monitor customer and market trends. They already have the capability to deliver personalized messages and to customize customer service to individual customers. They've recognized the importance of the move toward mobile and are putting significant investments into mobile apps that make it easy for customers to conduct business from anywhere. They're using data to fight fraud, to take advantage of artificial intelligence (AI) to make important financial decisions, and are aggressively finding ways to monetize data by offering new products and services.

As such, financial services firms are also on the front lines of the current challenges and limitations of big data technology. In the next section, we're going to look at one such challenge: *data lakes*.

The Problems with Data Lakes

Data lakes are hot right now. Gartner reports that inquiries it received about them increased a full 21% between 2016 and 2017. But at the same time that financial services firms are investing in massive centralized data lakes to store their digital jewels—their enterprise data—many are expressing doubts about whether data lakes are the best strategy for dealing with the data deluge.

Data lakes are enterprise-wide data management platforms that store disparate types of data in their native formats. The data resides in the data lake until someone in the organization—a data scientist, analyst, or business user—queries it. The grand idea is that consolidating enterprise data, both structured and unstructured, eliminates data silos. Users throughout the organization are thus encouraged to share and use information to glean valuable insights that they can then use to make critical business decisions.

But are data lakes delivering? Some observers say no.

Andrew White, a vice president and distinguished analyst at Gartner, said in a 2014 statement, "The need for increased agility and accessibility for data analysis is the primary driver for data lakes. Nevertheless, while it is certainly true that data lakes can provide value to various parts of the organization, the proposition of enterprise-wide data management has yet to be realized."

One of the most significant challenges is that substantial expertise and skill are required to make use of data stored in a data lake. Business users frequently don't understand the *metadata* (the information about the data) stored in the data lake, and lack context as to where the data came from, what has been done to it, and its quality. Moreover, the tools for accessing the data are notoriously difficult for nontechnical users to master. SQL is popular tool among analysts, but tools to utilize SQL in the Hadoop world are scarce. Those that do exist are somewhat clumsy. The promise that ordinary business users could simply tap into the data lake for valuable insights has therefore fallen flat.

Another problem is that big data is proving to be too—well…*big*—for data lakes. Although the idea behind data lakes is to capture much more—and much more varied—data than could be stored in a structured data warehouse or relational database management system so as to not constrain the potential for future analyses, in fact performance is a huge problem. Existing data analytics tools simply cannot perform at required speeds against large data lakes as they can against purpose-built databases that are optimized for particular types of queries. Queries against these massive data stores take hours and hours, if not days or weeks, to be returned.

Bottom line: storing all data in a centralized location in the hopes that it will someday be useful has not been working for many, if not most, businesses. A new approach is needed to get value from the data stores that enterprises are accumulating.

Bridging the Data Lake Insight Gap with Apache Kylin

eBay ran into many of the data lake–related big data issues back in 2013 that are now plaguing financial services firms. In fact, arguably it didn't have a data lake, but an information store so vast that its internal data team called it a "data ocean."

With more than 800 million active auction listings, eBay generates a lot of data. Every minute of every day more data is added to its already-tremendous stores of information, which are retained in its Hadoop data ocean. Making use of this information has always been a top priority for the company.

Even five years ago, a plethora of Hadoop frontend tools existed to improve upon basic Hadoop File System (HDFS) and MapReduce functionality. However, eBay was attempting to analyze data of 10 billion rows from multiple perspectives—and to do it extremely rapidly. And, although it had a team of Hadoop-knowledgeable big data scientists, most of its data team were accustomed to standard SQL queries and were frustrated by the existing tools. They attempted to work around the challenges by exporting data out of Hadoop into Online Analytical Processing (OLAP) and other SQL query-based systems, but that added steps to a process that was already too slow for the business.

"We needed near real-time decisions on these extremely large datasets. Without them, we couldn't respond fast enough," said Debashis Saha, then–vice president of eBay's commerce platform infrastructure, in a 2016 interview with InformationWeek.

Finally, after searching for a tool that could help it make sense of all its data, eBay concluded that the only answer was to design and develop its own solution. Thus, Kylin was born.

Kylin takes advantage of Hadoop's ability to scale-out to many thousands of nodes on a cluster. It then uses MapReduce's distributed processing capabilities to boost performance. Best of all from the perspective of business users, it processes SQL queries from popular data visualization systems like Tableau, returning them in standard ANSI format.

Kylin also inserted OLAP back into the analytics equation. OLAP was not new. Assembling data cubes that could be analyzed from different perspectives was done even before Hadoop came along. But Kylin enabled cube building on a truly unprecedented scale. Kylin then builds "smart indexes" on that same scale. Because indexes are prebuilt, users can analyze even the largest Hadoop data lakes.

Storing precalculated results to serve analysis queries has been done for decades. But when data grows as big as it was at eBay, precalculation was simply not possible using the existing tools, even with the most advanced hardware. However,

Hadoop's distributed computing capabilities enabled Kylin to perform these calculations in parallel and merge the final results, slashing processing time to previously unheard-of levels.

Ultimately, Kylin is distinguished for its extreme performance and high concurrency, while simultaneously offering traditional modeling features so that business users can adapt to it easily. The problem Kylin solves that other tools have not yet been able to address: high performance on the largest datasets; ability to scale and retain these high-performance levels even for the largest datasets; and ease of use for business analysts accustomed to SQL and other mainstream tools.

By October 2014, Saha's team had advanced sufficiently on development of Kylin to propose the initiative to the Apache Software Foundation as an open source project. In November 2015, Kylin exited incubation and became a top-level project with more than 30 developers.

The Overall Analytics Tools Landscape: What Else Is Out There to Help?

Whether to go with an open source or proprietary big data solution is one of the first questions financial services firms with large datasets face when they look for data analytics tools. Here are the pros and cons of each.

Commercial Big Data Solutions

The main advantage of commercial solutions is that they are well established. Most have been around the block numerous times, and offer solutions that focus on well-defined business processes such as modeling default risk for mortgage portfolios. They are proven, stable, and reliable.

They also offer enterprise support rather than the loosely connected network of volunteer open source developers that most open source customers must depend on (more on that in a few moments).

However, commercial software has some distinct disadvantages, as well. For starters, it is generally delivered out of the box with prepackaged, "black-box" algorithms, which, although extensively tested and documented, cannot be inspected by customers. This contrasts sharply with open source solutions that offer transparent access to source code.

Proprietary also means being locked in. Customers become dependent on a vendor for products and services, unable to use another vendor without incurring substantial "technical debt," or the high costs that would be incurred if they switched.

Proprietary solutions each have their own distinctive interfaces. This means that users are forced to adapt to a specific way of working, which can be radically different from what they are used to. Hiring specialists who have experience with a particular tool can be difficult—and pricey—due to the tight technology labor market. The alternative is training existing or new employees, but there's frequently a steep learning curve to these interfaces.

Finally, these proprietary solutions suffer from many of the performance issues found in legacy big data solutions. They simply can't scale to keep up with the volume of data being generated and stored by financial services firms.

For these reasons and more, many financial services firms are now considering open source big data solutions.

Open Source Big Data Solutions

The obvious advantage of open source software is that it is cost-free. This lowers the risk of deploying it considerably.

The second advantage? No vendor lock in. Open source is open.

Another advantage is that open source products tend to be more on the technical edge than commercial products. A global community is contributing to the project, constantly innovating and incorporating the latest advances to it, which then are released to the public at more rapid intervals than commercial software upgrades.

Open source is also more transparent and easier to evaluate because the source code is available. Financial services firms additionally have found that they can request special features to meet their specific requirements and collaborate closely with the open source community to implement those features in the product—something that is rarer in the commercial world.

One disadvantage is support. Such support typically consists of comprehensive frequently asked question (FAQ) databases, support hotlines, newsletters, and even professional training courses and certifications. When financial services firms build business-critical big data systems, they want to know that they can depend on qualified professional experts to help if there are problems. In the open source world, they depend on the community, and the vast majority of community members are volunteers.

Another disadvantage of open source can be stability. Because open source solutions tend to be on the leading edge of innovation, using them in large-scale projects is often pioneering work—work no other financial services firm has attempted.

Best of Both Worlds

One scenario that combines the best of both commercial and open source worlds is when a company makes an enterprise-grade product from open source software and then releases it under a license that guarantees both support and related Service-Level Agreements (SLAs). In doing so, an open source vendor provides quality controls and safeguards along with enterprise-class service and support, to lower the risk of deployment.

As open source becomes more common in enterprise IT departments, this model is increasingly popular. In fact, there has actually been a decrease in software being released under traditional proprietary licenses in recent years, due to a surge in software released under an open source framework.

Existing Tool Landscape

There's no lack of big data analytics tools in both commercial and open source categories to help financial services firms attempt to make sense of all the data they are accumulating in their data lakes. But it can be difficult to differentiate between the many products classified as big data analytics software, as the claimed functionality, features, and capabilities are often very similar. What sets solutions like Kyligence—based upon Kylin—apart is the interplay of extreme high performance, industry-standard SQL interface, and open architecture, which allows you to easily integrate it with legacy systems.

First, a definition: big data analytics products are software tools that enable organizations to run analytics applications—descriptive, predictive, and prescriptive—on big data computing platforms. These platforms usually are systems capable of parallel processing by running on clusters of commodity servers, possess scalable distributed storage, and support databases such as NoSQL. The tools are designed to enable users to rapidly analyze large amounts of data, often in real time.

Big data analytics tools generally are capable of supporting a broad range of data types, from highly structured, to semi-structured, to completely unstructured. Teradata, SAP HANA, Greenplum, IBM Cognos, and HP Vertica are just some of the commercial solutions available. These large, established proprietary data analytics platforms have been around for a while, and have both advantages and drawbacks. On the one hand, they are proven and provide largely stable and reliable solutions. On the other hand, companies that invest in them risk vendor

lock-in, steep learning curves for users, and—increasingly—performance constraints for companies that possess large datasets.

Startups such as AtScale, Kyvos, Dremio, and Jethro have recently joined the pack. They do offer more flexibility and less risk of vendor lock-in by giving users a choice of their favorite frontend data querying and visualization tools. However, they also share the performance constraints of legacy proprietary tools when attempting to cope with extremely large datasets.

This is where Kyligence shines. Already proven in some of the largest enterprises in the world, it specifically was designed to work with the largest datasets on the planet, while giving easy access to business users and analysts. Its coupling of performance and ease of use, along with the fact that its open architecture allows it to be integrated with legacy systems, puts Kyligence at the top of the list of both proprietary legacy, upstart digital native, and open source competitors.

Financial Services Case Studies

A number of high-profile financial services firms around the world have implemented Kyligence, an enterprise-class version of Kylin released in August 2016. Many of its first enterprise users were in Asia. Here are the big data journeys of three of them: China Construction Bank Corporation; China UnionPay; and China Pacific Insurance Group Company Limited.

China Construction Bank

China Construction Bank Corporation, headquartered in Beijing, is a leading commercial bank in China. Its predecessor, China Construction Bank, was established in October 1954. It was listed on the Hong Kong Stock Exchange in October 2005 and the Shanghai Stock Exchange in September 2007. At the end of 2016, the market capitalization of China Construction Bank reached $192 billion, making it the fifth-ranked bank among all listed banks in the world, and number 28 in the Fortune Global 100.

With almost 15,000 branches and 365,000 employees, the bank provides financial services to hundreds of millions of personal and corporate customers and cooperates closely with leading enterprises in strategic industries in the Chinese economy as well as numerous high-end customers. The bank has commercial banking branches and subsidiaries in 29 countries and regions with 251 overseas entities, and its subsidiaries cover asset management, financial leasing, trust, life insurance, property and casualty insurance, investment bank, futures, and pension services.

Zhi Zhu, vice senior manager of IT for China Construction Bank, says that the organization faced three key challenges as it began its big data journey: the emer-

gence of new business models; performance problems with its legacy data systems; and the rapid pace of innovation in the big data technology sector:

Emergence of new business models
: New technology tends to create new opportunities as well as challenges that require new business models. One example of this that has worked out well for China Construction Bank is risk management, says Zhu. "In the past, we would conduct risk management by purchasing lists and solutions from outside vendors, usually international organizations," he says. "With the accumulation of data, we realized we could now explore patterns on datasets using technology in ways that we couldn't before." As a result, China Construction had to reinvent how it approached risk management—but it also had to find analytics solutions capable of crunching the enormous amounts of data it now routinely collected and stored. By doing so, "we turned risk management from an expense to a revenue-generating business operation," he says.

Big data performance
: Traditional methods of data analysis take too long in the new big data environment. This raises significant challenges for China Construction Bank data engineers, who have to ship products rapidly on a tight schedule. "For companies like ours, with a massive amount of data and a legacy data analysis system in place, this is very challenging," says Zhu.

Rapid technological innovation
: The development cycles of new technologies are accelerating—and big data technology is no exception, says Zhu. "This pace quickly renders legacy applications and systems outdated or obsolete," he says. "However, we can't just outright abandon our legacy systems because so much has been invested in them already." This indeed is a problem for many companies, says Zhu, who was looking for new solutions that could be seamlessly integrated or migrated onto existing legacy applications, so the bank could continue to get mileage out of its previous investments. After all, innovation can hurt the bank's ability to ensure business continuity. "We've also found that new big data technology usually cannot cover all the business needs that our legacy systems already handle," says Zhu. "This is a dilemma."

Migrating to new technologies is probably the most severe challenge, says Zhu. "While legacy technology is costly, migration can be even more expensive," he says. Migration can also cause headaches in terms of maintaining business continuity. For example, over the past decade, a lot of the bank's business applications run on Teradata's platform. Although expensive and complex, the Teradata solution is stable when compared to some of the emerging big data technologies. "It's tough for us to control costs, evaluate the stability of new solutions, and

maintain business unit continuity without disruption, all while migrating applications to new big data platforms," says Zhu.

Even for cases in which costs can be significantly reduced with newer big data technology, some business challenges specific to the financial services industry can still be addressed only with legacy systems, says Zhu.

For example, many emerging solutions can't efficiently provide access to different kinds of financial or accounting information—an essential function in today's security-conscious era. Banks like China Construction Bank also need to be able to flexibly organize and manage the multiple viewpoints of multiple teams within its complex organizational structure—a typical requirement in a bank of its size and scope.

Prior to adopting the Apache Kylin solution, China Construction Bank was using a combination of Greenplum, Teradata, Oracle, and solutions in the Hadoop ecosystem like HDFS, Hive, and HBase. "Our data was first stored in a buffer area to be processed for proper file sharing," says Zhu. "It was then cached and went through some lightweight computing using Greenplum."

Some of the computed results would go directly to Oracle for high-concurrency access to support data analysis by different application components, but for integrated calculation and computing, data was transferred to the bank's Teradata platform. Finally, all of the processed data was moved to a shared-access area for all users to access the data they needed. Unstructured data was processed using HDFS/Hive, and the results were stored in HBase to be analyzed by different applications.

But although reliable, this setup wasn't meeting the bank's need for the real-time, concurrent access to data driven by the rapid rise of mobile applications.

"After a careful analysis, our technical team found that more than 90% of our businesses rely on descriptive statistics and 80% of those statistics require high concurrency," says Zhu. "To meet this challenge, we needed a new technology that had low implementation and deployment costs." The bank first tried to develop a solution internally, but couldn't satisfy all of its needs. "During our evaluation process, we found Apache Kylin, which matched all our requirements," he says.

One of Kylin's advantages is its consistent adherence to and implementation of Kimball Multi-Dimensional Data Modeling Theory, says Zhu. Additionally, Kylin maintains a standard ANSI SQL interface, which avoids the implementation complexities of MapReduce by making querying available to business analysts without learning a new tool.

"It's also fast, delivering query results within subseconds," says Zhu. The bank deployed Apache Kylin in the shared access area in its architecture and estab-

lished data streaming channels similar to Kafka (see Figure 1-2). "This implementation gave us a near-real-time data warehouse solution that delivers good performance on both computing and data access," says Zhu.

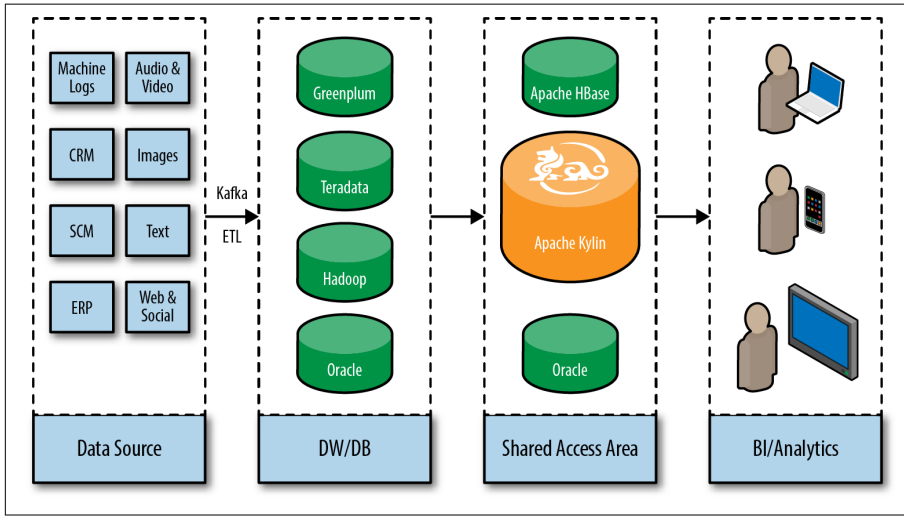

Figure 1-2. Apache Kylin for shared access and unified OLAP services

Today, all employees at China Construction Bank can subscribe to and access data using mobile devices at near real-time speeds. Kylin interfaces well with the legacy systems built on IBM Cognos, Oracle, and Greenplum. All this not only meets the bank's performance requirements, it does so in an astonishingly cost-effective way because the deployment cost of Kylin was low. When compared to the tens of millions of dollars it would have taken to implement a new commercial big data analytics system, "this is amazing," says Zhu. "And the number of data engineers we needed to allocate for deployment and implementation was reduced from 60 to just 10. Our ROI was very high."

Zhu continues, "With Kyligence's help, employees from all levels and functions of our company—from branch offices, financial planning, and partnership channels, to risk management, credit management, and data management—can now simultaneously access, analyze, and compute data depending on their needs. This significantly shortens the feedback loop on the development cycle of new products and businesses."

In the future, China Construction Bank is likely going to migrate its entire data system from the Teradata and Greenplum commercial platforms to the open platform ecosystem. "We will fully embrace Hadoop 3.0 and Kyligence's industrial best practices to further reduce deployment and implementation costs while increasing our use of artificial intelligence," says Zhu. On the data governance side, the bank will look to standardize its computation results.

"And finally, we hope to free up more of our technical talent from deployment and implementation tasks, and train them to become analysts and data scientists, where they will add considerable value to the business," Zhu says.

China UnionPay

China UnionPay is a Chinese financial services corporation headquartered in Shanghai. Founded in 2002, China UnionPay is the clearinghouse for China's banking card industry—the equivalent of MasterCard and Visa—that operates under the approval of the People's Bank of China (the country's central bank). It is the only interbank network in China that links all the ATMs of all banks throughout the country. It is also an electronic funds transfer at point of sale (EFTPOS) network, and the largest card payment organization—debit and credit cards combined—in the world, including MasterCard and Visa.

Yingzhuo Wang is the deputy general manager in data service department at China UnionPay. He says that the company faced many challenges during its transition from managing its data through a traditional standalone relational database management to a true unified "big data approach."

"In our previous infrastructure, the architecture looked like isolated stacks—almost like chimneys," Wang says. "Transforming that to a unified big data platform to serve the entire company's IT needs impacted everything we did, from architectural design, to system endpoints, to our IT management operations and processes."

One of the biggest challenges Wang faced in his work at China UnionPay when searching for big data solutions was the rapid pace of technology development in the big data space. "Big data technologies advance very quickly, so you often see systems and solutions deployed in production with a lot of 'battle scars,'" says Wang. That's not surprising. Compared to more mature solutions, newer big data technologies are volatile when it comes to reliability and stability, he says. "Finding the right technology that would reliably provide the services we needed while taking advantage of new innovations and features was difficult," he says.

Plus, as Wang points out, the learning curve for staff was very high. "The big data ecosystem has a lot of variety and is very vibrant and constantly changing, so making sure we were deploying the right tools for the right applications was quite challenging."

Very specific big data requirements

Wang's work at UnionPay had some very specific requirements when searching for big data solutions.

Take data backup. Previously, in its legacy environment, the bank had used straightforward SQL commands like INSERT or LOAD to move data to data

warehouses for replica and backup. In a big data environment, however, data must first be written into HDFS and then transformed into Hive or Impala's format, "so the whole process had to completely change," says Wang.

For extracting data, the bank often needed to do so by full columns under certain conditions, and in its previous infrastructure setup, it could accelerate those queries by adding indexes. However, in a big data environment, the indexing approach isn't the optimal way to go.

"The one solution we'd seen that accelerated these types of queries well was HBase, but the implementation wasn't as simple as just setting additional indexes," explains Wang.

Finally, there was data querying. "The most common way to query a database is using SQL, but most big data technologies expose features with their own application programming interfaces (APIs), and support for SQL is still evolving," he says. "We thus had to do a lot of work building our own endpoints to support existing data query behaviors using SQL as part of our transition into a big data environment."

Then, Wang's team found the Kyligence solution, based on Apache Kylin.

Previously, UnionPay had been using Cognos widely and "with decent success," says Wang. "Our team liked Cognos' ability to do multidimensional data analysis and relied on it quite a bit." However, as growth of the bank's data accelerated, it became obvious that Cognos couldn't handle the new load. "It was becoming a bottleneck," he says. "So, a big part of implementing our big data strategy and constructing a unified data platform was to find an alternative to Cognos—a solution that could do the same kind of multidimensional analysis and take advantage of all the powerful capabilities of a big data platform without forcing our users to change their behaviors."

UnionPay chose Kyligence for four reasons:

Seamless integration
: Kyligence seamlessly integrates with all the other tools in the big data ecosystem, which made it a perfect complement to many of the other big data technologies that the bank wanted to use.

Constant improvement and optimization
: "The Kyligence development team is top-notch and very receptive to the customer's needs and willing to incorporate our needs into its products," says Wang.

Co-development
: UnionPay often had requirements for special features in its big data environment. Getting what it needed simply wasn't possible with Cognos, because it is a commercial, proprietary system with its own roadmap. "It is much easier

to co-develop with the Kyligence team because the core product, Kylin, is open source," says Wang.

Supported by open source
: Because its core code base is open source and its enterprise support very timely and professional, working with Kyligence gave UnionPay both a lot of visibility into the product's code and access to a professional team for enterprise-level support. "This was the best of both worlds—enterprise and open source," says Wang.

Today, the interface layer (Mizar) presents a unified interface for all of UnionPay's applications to interact with its data. The "negotiator" (Dubhe/Megrez) is responsible for managing system resources in an ongoing basis, constructing execution policies, and adjusting workloads. The monitoring piece (Alioth/Phecda) is in charge of monitoring the status of task execution and auditing system security. The core service (Phecda/Merak) executes the bank's security policy and access control as well as constructs the necessary environment to execute the tasks. Tornado is the engine that drives the data-retrieval process from the different data sources to the core service, which goes to the application layer, and Kyligence ties it all together with multidimensional analytics performed using standard SQL commands (see Figure 1-3).

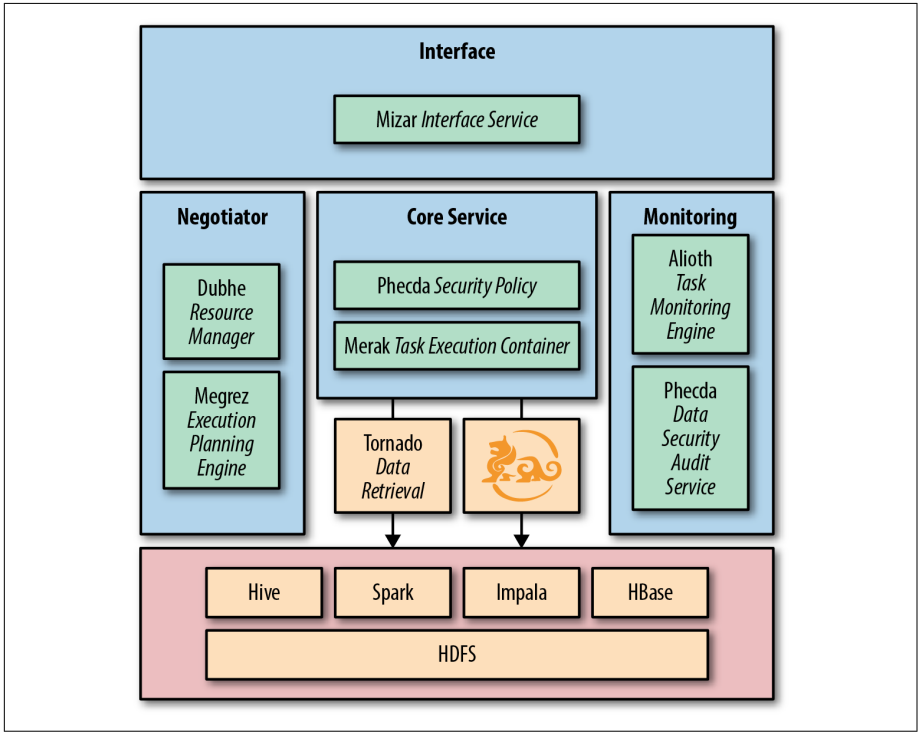

Figure 1-3. UnionPay's current big data infrastructure

Kyligence adeptly resolved access-control issues that UnionPay previously had with Cognos. "We no longer have to do access control based on different branch offices, and can implement granular access control across different offices, departments, and teams," says Wang.

Kyligence also supports fast multidimensional data queries on large-scale datasets. For single-dimensional queries, results are returned within seconds.

Finally, the bank found its cube rebuild capacity greatly improved, says Wang. "On a dataset of 100 million rows, rebuilding and refreshing a 32-dimensional cube on a 64G cluster with 10 nodes now takes only about two hours," he says.

With the new big data infrastructure, members of the UnionPay sales team can quickly pull relevant, up-to-date data and analysis to support their business activities, while greatly reducing the workload that each team member has to undertake to extract that data on a daily basis. "The speed and agility in which Kyligence allows our business team to access data empowers them with the right information to quickly make sound business decisions," he says.

The new Kyligence-based data environment has also been instrumental in supporting development of new products and business models, says Wang. By making so many different types of internal data available and accessible via a unified set of entry points, different business units can quickly gather information from multidimensional datasets and rapidly iterate and innovate.

Currently, UnionPay is engaged in expanding its big data environment and building an enterprise-grade unified data platform and process, including storage, security, and production use. It plans to expand the use of big data technologies internally. Gradually it will apply new technologies to its mission-critical applications while elevating its ability to more effectively use and process data.

"We're also exploring the possibilities of merging multiple data sources together," says Wang. "Right now, data security is an increasingly important issue, but integrating different data sources together is also an important practice and trend in the big data era." The bank is investigating how to adopt this practice in a way that is proper legally and won't compromise data security.

He has some words of advice for other financial services firms looking to build a big data environment.

"Building a robust big data platform is a system-level project that requires a complete architecture-level approach and a holistic design," Wang says, cautioning that it also needs proper evaluation and support in terms of resources and investment. Half measures won't do.

Additionally, big data isn't just a technical innovation, but also a major shift in terms of architectural design and philosophy, he says. To get the most out of big data innovation, everyone in the company must be on the same page.

Finally, Wang says, it's important to remember that the real value comes from the data, whereas the technology is there to help companies derive the most value out of that data.

"Thus, you should avoid any pure technical considerations," he says. "Instead, focus on how a piece of technology can support the security, integration, application, and discovery of data, as well as how these elements can support the development of new products and services for your customers. In short, the true value of any technology lies in how well it serves your business purposes."

China Pacific Insurance (Group) Company Limited

Shanghai-based China Pacific Insurance (Group) Company Limited (also known as CPIC), is a Chinese insurance company established on the foundation of the former China Pacific Insurance Corporation, which was founded in 1991. As one of the largest insurance companies in China, Pacific Insurance provides integrated insurance services, including life insurance, property insurance, and reinsurance. The company's property insurance products include car insurance, insurance of family properties, liability insurance, investment insurance, and accident insurance, while its life insurance products include endowment insurance and health insurance.

When it started its big data journey, the three biggest challenges facing Pacific Insurance were data silos, lack of technical talent, and data security, says Mincheng Wu, deputy general manager of IT at CPIC.

Data silos
: "Previously, we did a lot of work consolidating our data into data warehouses and data marts, but from a big data application angle, there were lots of deficiencies in terms of the comprehensiveness of the consolidation," says Wu. "This directly impacted the user experience when using the data in business-critical applications."

Lack of technical talent
: Like everyone else trying to participate in the big data world, Pacific Insurance was limited by the severe dearth of engineering and development talent in the field.

Data security
: As it consolidated its data into a single data lake, the requirements for security became much more stringent—especially for the very sensitive customer data. "We found it challenging to make sure we applied third-party customer data in ways that were both safe and lawful," says Wu.

Before implementing Kyligence as its data analytics solution, Pacific Insurance had previously used the tools that came packaged with traditional BI tools to

construct cubes. However, this approach led to severe performance issues as its volume of data grew.

"We chose Kyligence to address performance issues with traditional business intelligence and analytics workload solutions," says Wu. "As a large insurance company, the core competitiveness of our products rests on how effectively we can understand and use data to help our customers, so this workload is critical to our success."

Kyligence's enterprise-grade analytics platform turned out to be a good fit to meet Pacific Insurance's demand for executing high-concurrency, high-dimensional data analytics on a distributed big data platform. The company's current infrastructure is a distributed Hadoop environment, with Kyligence in the middle to provide precalculated cubing. Kyligence was easily integrated to business intelligence (BI) tools on the frontend for Pacific Insurance users.

"Previously, when we were using traditional BI tools to construct cubes, the size of a single cube was a limiting factor," says Wu. That's because, if the number of dimensions becomes too big, it reduces query speed and the efficiency of the computation as it tries to generate reports.

"After we deployed Kyligence, we were able to execute precalculation on top of our Hadoop platform for cubes with many more dimensions, which significantly accelerated the query speed to generate reports," says Wu. "This performance boost also allowed us to more effectively take advantage of the scalability of a big data platform."

Right now, Kyligence's core value is that it accelerates users' data query speed, which significantly increases their efficiency as it enables them to mine insight from more data in a more granular way. "This new capability also improved our user experience," says Wu.

Looking forward, in 2018, Pacific Insurance is making AI its top strategic priority. "We are planning a series of enterprise-level AI products to develop," says Wu. "Based on our big data efforts, we will begin building an AI-powered technology platform, train a team of AI experts, and increase our capacity for production-level AI application and cutting-edge research through strategic outside partnerships."

Best Practices for Getting Insights from Data Faster

As these case studies show, the financial services industry is leading the pack when it comes to emphasizing and investing in big data, especially on data analytics capabilities. Here are some best practices that all financial services institutions that are transitioning into the big data analytics area should consider:

Embrace distributed systems for data warehousing

To meet the challenge of big data storage, distributed data systems like Hadoop have advanced and matured over the past decade. Traditional data warehousing technology simply cannot meet the demand of the "3Vs"—volume, velocity, and variety—that define big data operations. The performance, reliability, and increasing adoption of distributed computing engines like Spark, Hive, and Kylin, which are all part of the Hadoop ecosystem, have become the de facto standard for big data operations that can scale and grow as financial services firms' data stores continue to increase.

Buy new technology that is cost-effective, not costly

Traditional data warehousing and business intelligence tools are expensive to both buy and maintain. This is not the case for new distributed data technologies built on top of and within the Hadoop ecosystem. The architecture and hardware requirement of Hadoop are predicated on the availability of cheap x86 servers, making horizontal scaling of data storage easy and affordable. This design can help financial services companies decrease their IT expenditure and achieve the flexibility and capacity needed to grow.

Don't compromise on choosing the right data analytics platform.

It's no longer enough to scale the storage capacity as data grows. Adopting high-performing data analytics capabilities to mine valuable business intelligence has also become mission critical in the financial services industry. Traditional data analytics tools have failed to meet this demand in terms of capacity, concurrency, and analytics performance. After storing terabytes, if not petabytes, worth of data on a Hadoop-powered system, spend time to choose the right data analytics engine that is compatible with the tools users are already familiar with so that they can unearth the insights required to grow the business and better satisfy customers.

Don't forget about integration up and down the stack

Unlike internet companies and new tech startups, which can build their technology infrastructure anew, most large financial services companies have already invested in IT infrastructures that are out of date, but they don't have the luxury to start from scratch. Striking a balance between embracing new technologies with open arms while still getting some mileage out of existing investment in legacy infrastructure is a difficult question for every CIO to answer. For example, after a company moves its data to Hadoop, the variety of data analytics tools that sit on top of Hadoop could dramatically increase the learning curve for BI analysts. Choose a platform that integrates well with legacy systems and tools that data analysts are already skilled in using—tools like Tableau, Power BI, or Excel—so that they can get up and running quickly to begin mining the business insights required to stay competitive.

About Kyligence

In resource-intensive systems, queries compete for resources. When the workload is large, it can take a long time—hours or sometimes even days—to get a response. Although SQL on Hadoop is improving, it is still common to wait many minutes or even hours for a query to return, especially when a dataset is large.

To solve this problem, Kyligence built a unified analytics platform that simplifies big data analytics for business users, analysts, and engineers. Kyligence offers enterprise and cloud versions of Apache Kylin focused on speeding up mission-critical analytics at web scale. It seamlessly supports public clouds such as AWS, Azure, and Google. Kyligence also offers self-service data access and seamless integration with BI tools, with no need for user to have programming skills. A native OLAP solution on Hadoop, Kyligence interacts with clusters via standard APIs.

Kyligence is based on the open source Apache Kylin. The company was founded in March 2016 by the creators of Apache Kylin and has dual headquarters in Santa Clara, California, and Shanghai, China.

Kyligence enables subsecond SQL query latency on petabyte-scale dataset, provides high concurrency at internet scale, and empowers analysts to design BI on Hadoop with industry-standard data warehouse and business intelligence methodologies.

Kyligence also offers the following:

Native SQL support on both Hadoop both on-premises and in the cloud
: Many big data analytics tools have their own query languages or proprietary storage engines. But analysts can find it difficult to learn new query languages or to move data out of HDFS/BLOB storage to different platforms. With Kyligence's native SQL support and ODBC drivers, customers can use SQL interface and their favorite BI tools, no matter what size the dataset.

Speed up mission critical query
: The amount of time it takes for a query to be returned is the most important metric in big data analytics. Performance will deteriorate if the cluster resource cannot scale out when the original data grows by factors of 10. Kyligence solves this problem by providing precalculated cubes that are processed in parallel on a distributed environment.

Batch and streaming OLAP
: Kyligence's platform can consume from batch data sources like Hive, Spark, SQL, and other RDMBs. It also can consume streaming data from Kafka. Businesss can simply interact with Kyligence using ANSI SQL to easily achieve both historic and near-real-time reporting.

Elastic architecture

On very large datasets of gigabytes, terabytes, and even larger, Hadoop provides an elastic infrastructure for batch processing. Kyligence then provides an equally elastic interactive analytics technology to enable scale-out solutions.

In Conclusion

Leading financial services firms are exploiting the vast reservoirs of data that they increasingly capture and store—not only to improve operations but also to attract and retain customers, and even to launch entirely new revenue-generating products and services. Even established financial services firms are finding that they can move swiftly, decisively, and with agility when they use the right big data analytics tools.

To take advantage of the new ground broken using big data, however, leading financial services firms need to continue on their journeys, and carefully consider which data analytics tool is best to keep them on top of their games. Apache Kylin, the open source analytics software developed at eBay to solve the same challenges that financial services firms face today, should be a leading candidate. Kyligence, which offers an enterprise-grade version of Kylin, is already in use at leading banks, insurance companies, and brokerages, and has proven its worth in terms of performance, stability, and scalability.

About the Author

Alice LaPlante is an award-winning writer who has been writing about technology and the business of technology for more than 20 years. Author of seven books, including *Playing for Profit: How Digital Entertainment Is Making Big Business out of Child's Play*, LaPlante has contributed to *InfoWorld*, *ComputerWorld*, *InformationWeek*, *Discover*, *Bloomberg Businessweek*, and other national business and technology publications.