Big Data Cases in Banking and Securities

A report from the front lines

Sponsored by

Jennifer Costley
Peter Lankford
30 May 2014

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

About the study sponsor

Today the financial services industry depends on innovation more than ever to run its business. Intel based technology for clients, servers, storage, and networking is the foundation for the new and open infrastructure required to deliver this innovation. As Intel continues to deliver value in each new generation of its products, it is important to understand the business requirements and challenges of the key end users of its technology. For this reason Intel is proud to sponsor this research by STAC and to drive thought leadership in the fast growing area of 'Big Data'. Silicon technology, complemented by software products and services from Intel, provide an ideal foundation on which to build solutions for the current and future requirements of data intensive computing for financial services.

About STAC

STAC® is a technology-research firm that facilitates the STAC Benchmark™ Council (www.STACresearch.com/council), an organization of leading financial institutions and technology vendors that specifies standard ways to assess technologies used in finance. The Council is active in an expanding range of big-data, big-compute, and low-latency workloads. STAC helps user firms relate the performance of new technologies to that of their existing systems by supplying them with STAC Benchmark reports as well as standards-based STAC Test Harnesses™ for rapid execution of STAC Benchmarks in their own labs. Some STAC Benchmark results from vendor-driven projects are made available to the public, while those in the STAC Vault™ are reserved for qualified members of the Council.

About the authors

Jennifer Costley is a scientifically-trained technologist with broad multi-disciplinary experience in enterprise architecture, software development, line management, and infrastructure operations. Following 31 years of technology leadership in organizations like Credit Suisse, Bankers Trust, and Double Click, Jennifer now consults to companies, non-profit organizations, and individuals in areas related to data, governance, and sustainability. This includes roles with the IEEE and the STAC Benchmark Council. Jennifer has a PhD in Chemical Physics from Columbia University and a Bachelors in Physics and Chemistry from Brandeis University.

Peter Lankford is founder and director of STAC, which facilitates the STAC Benchmark Council. In this role, Peter has helped the finance industry create benchmark standards in areas such as data-bound time-series analytics, compute-bound risk simulations, and I/O-bound transformation and distribution of high-speed data. Prior to STAC, Peter was SVP of the the $240M market data technology business at Reuters and held management positions at Citibank, First Chicago, and operating-system maker IGC. Peter has an MBA, Masters in International Relations, and Bachelors in Chemistry from the University of Chicago.

Executive Summary

Investment banking and retail banking often appear near the top of the list of industries investing in "big data" technology. Yet information about how banks are using that technology is sparse.

STAC® recently worked with several of the largest global banks to identify concrete use cases that pose big data challenges. Our objective was to study cases that were specific to banking rather than cases such as web log analytics and intrusion detection, where requirements tend to be fairly common across industries and are better understood. By interviewing staff with direct knowledge of the cases, we were able to characterize the workloads involved and understand the business problems that arise with traditional technologies. We were also able to learn about the advantages and challenges of the new approaches that banks were taking to these workloads. This paper summarizes some of these findings.

The primary purpose of these discussions was to lay the groundwork for technology benchmark standards that can be applied to big data problems. The STAC Benchmark Council (to which these banks belong) develops benchmark specifications for technologies used in strategic business functions. These specifications become a common yardstick by which user firms and vendors can understand the capabilities of competing solution stacks. We're grateful to Intel (also a member of the Council) for providing the seed funding to accelerate the benchmark development process in big data. The production of benchmark specifications covering performance, scaling, resource efficiency, resilience, security and entitlements, and other key business metrics is now underway. Meanwhile, we decided to share our learnings outside the Council through this paper, both to raise the level of awareness in the industry and to attract more participants to the project.

When soliciting cases from the banks, we defined "big data" as a problem, not a technology. Specifically, we defined a big data workload as one that is **too difficult or expensive to handle using traditional technologies, largely due to data scale or complexity**. By focusing on problems rather than solutions, we did not restrict the investigation to particular technologies and were able to include some cases that banks had not yet solved. This definition was also flexible enough to cover old workloads as well as new ones. A workload could be "too difficult or expensive to handle using traditional technologies" either because the workload is uncommonly cumbersome or because new technologies render traditional technologies difficult or expensive by comparison, even for existing problems. As we will describe, we found both kinds of cases.

Our study was qualitative, not quantitative. Because the purpose was to characterize some important workloads (rather than, say, estimate market sizes), we dove deep into a few cases with select firms rather than canvassing numerous organizations with high-level questions. If our research into big data use cases is currently in the "stamp collecting" phase, then this paper describes our first stamp collection.[1]

While the number of cases is small, our long experience in the industry suggests they are worth noting. The banks we interviewed tend to be technology leaders, often building their own solutions from the best tools available rather than looking for pre-built applications to buy. The solution patterns they establish tend to filter into the broader industry over time, both to other banks and to application vendors.

[1] "All science is either physics or stamp collecting." - Ernest Rutherford

Looking across these cases, a few themes emerged. Here are the highlights:

- It's not just hype. Investment and retail banks have moved to new technologies for big data problems in important business functions, and usage is growing.
- It's not just ETL. The analytic complexity of the workloads we studied ran the gamut, from basic transformations to machine learning.
- About half the cases we encountered were about doing new things, while half were about doing existing things faster or more cheaply. We're sure there was selection bias against reporting "new things," which tend to be competitively sensitive and thus not readily discussed.
- Of the famous "three V's" (volume, variety, velocity), the strongest driver was volume. About half the cases we encountered involved a petabyte or more of data. Variety took second place. While some of the cases involved natural language text (e.g., email, social media), most of the content in the workloads we studied was more structured, and none of it was completely unstructured (e.g., video, images). Nevertheless, some of the cases involving highly structured formats still presented great variety at a semantic level. What is often called "velocity" was a driver in only one of our cases. This is partly because the retail banks don't deal with much velocity yet, while the investment banks have dealt with high velocity for so long that their current solutions for those problems are generally superior to emerging big data technologies.
- Hadoop has pole position in these cases. Despite the technology-agnostic way that we solicited cases, Hadoop turned out to be prominent in most of them. Other technologies were often involved, but Hadoop was at the center.
- About half the banks have built or are building multi-tenant analytics platforms using new, big data technologies. Some are departmental in scope, while some span the entire corporation, across retail, commercial, and investment banking. This is an area rich with possibilities and challenges.
- Most of the banks felt that expansion of big data technologies into a broader range of use cases was constrained by insufficient functionality in critical areas such as entitlements.

This paper describes these themes in more detail. Specifics of the workloads and the proposed big data benchmark specifications are restricted to members of the STAC Benchmark Council. Membership is available to both vendors and users of technology. We welcome the involvement of all interested parties.[2]

[2] See www.STACresearch.com/bigsig

Background

The STAC Benchmark™ Council consists of leaders from over 200 financial institutions and 50 vendor organizations who develop technology benchmark standards based on workloads that are strategically important to user firms. Examples of Council efforts in recent years include STAC-M3™, the benchmark suite for analytics on large stores of historical market data (the lifeblood of trading), and STAC-A2™, benchmarks based on market risk management (the kind of workloads that drive hundreds of thousands of processor cores across Wall Street).

With more and more of their engineering challenges falling under the umbrella of "big data", a number of banks and vendors in the Council formed the STAC Big Data Special Interest Group (SIG) in March 2013. Its aim was to provide a forum to discuss challenges and solutions in the area of big data and ultimately to develop technology benchmarks for big data workloads.

Few areas cry out for good technology benchmarks more than the Wild West of big data. Solution designers face dozens of new software and hardware products. They must understand which products and design patterns are suited to which use cases. And they must determine whether these products deliver not only the transformative capabilities that they promise, but also the boring-yet-critical functionality taken for granted in traditional architectures. The need for rigorous big data benchmark standards is especially strong in finance, where the opportunity to turn information into money is huge, but the cost, quality, and security constraints grow stronger by the day.

As a founding member of the SIG, Intel heard this cry and graciously provided seed funding to accelerate the process of defining the workloads, developing the benchmarks, and sharing the learnings.

The workloads

In the first quarter of 2014, we studied 16 projects at 10 of the top global investment and retail banks, interviewing individuals with direct knowledge of the workloads. From these interviews, we documented 10 workloads and two "meta-workloads" that met our big data definition.[3]

As explained above, we defined big data as a workload that is too difficult or expensive to handle using traditional technologies, largely due to data scale or complexity.[4] In most of the cases, the firms were handling these workloads partly or completely via new technologies. In a few cases, the banks were still using traditional technologies but were actively seeking new approaches. In one case, a bank described a workload that might be a good candidate for big data technologies if it weren't for certain limitations it perceived in those products.

Table 1 on the following page briefly summarizes the workloads. The columns to the right of a given workload indicate the type of banking involved in the interviews in this study. Some workloads that arose in just one category in our interviews may also apply to the other category or to other forms of banking not represented in this table. Roughly half the workloads arose in both retail banking and investment banking contexts (in fact, some of them such as enterprise credit risk reporting also involve other forms of banking, such as commercial banking). One of them was clearly specific to retail (card fraud detection), while several were specific to investment banking (mostly related to trading). The two IT-related workloads are not necessarily industry specific, but we included them because of the centrality of IT to investment and retail banking. We imagine that other IT-intensive, heavily regulated industries have similar use cases.

[3] These do not sum to 16 because multiple interviews concerned some of the same workloads.
[4] This definition emerged from the initial meeting of the SIG.

Workloads encountered in this study		
Workload	**R**	**I**
Card fraud detection. Using statistical models on credit and debit card transaction data to detect fraudulent activities while minimizing false positives.	✓	
Securities fraud early warning. Detecting potential securities fraud by institutional clients to support legal requirements for due diligence of transactions, by analyzing data drawn from hundreds of internal systems.		✓
Enterprise credit risk reporting. Providing a consolidated view of exposure to different kinds of credit risk across retail and wholesale products, based on loan ledgers, trading positions, economic history, and other sources.	✓	✓
Tick analytics. Providing simple analytics on time-stamped order book data collected from real-time exchange feeds.		✓
Social analytics for trading. Using historical social media content to develop indicator histories (e.g., sentiment) for use in trading algorithms. Creating indicators from realtime social feeds to enable use of those algorithms.		✓
Trade visibility. Supporting brokerage functions (customer service, regulatory reporting, surveillance, advisory, prospecting) that need to interrogate deep histories of customer transactions across all assets and businesses.		✓
Archival of audit trails. Providing storage and retrieval of transactions for many years, as required by regulatios such as FINRA's Order Audit Trail System (OATS).		✓
Customer data transformation. Transforming inbound data feeds from custodians and other sources and loading into a centralized data warehouse.	✓	
IT policy compliance analytics. Identifying deviations from policy by analyzing tens of millions of daily event records and config files related to thousands of regulatory-relevant applications and their associated infrastructure.	✓	✓
IT operations analytics. Using event logs, config files, emails, and utilization data to improve planning, performance, and resource management and to identify security risks.	✓	✓
Certain mainframe workloads (a meta-workload). Tasks that use VSAM file structures that translate easily to key-value formats.	✓	✓
Analytics Platform as a Service (a meta-workload). Providing advanced analytics capabilities as a service for multiple workloads across multiple lines of business.	✓	✓

R = Retail banking & wealth management. I = Investment banking & brokerage.

Figure 1

Business drivers

The business forces driving banks to seek big data technologies to handle the workloads above will not surprise those familiar with the recent focus of the banking industry.

Almost half the projects we encountered were driven by regulatory or legal requirements. The impact of regulation cuts across all functions within a bank and reaches deep into the organization. Risk regulation under Basel III (and other regimes) requires more comprehensiveness and accuracy in the calculation of exposures across asset classes and counterparties. Legal settlements have specific requirements to improve transaction oversight and avoid the recurrence of sanctions. Even IT organizations in banks are not exempt: many of their processes fall under the SEC's Regulation SCI, Fed IT policy requirements, and the ongoing impact of Sarbanes-Oxley (SOX) certification requirements. Our research showed that such requirements are driving banks to combine data from far more internal systems than they are accustomed to integrating, as well as to create more powerful analytics than are readily accessible in their legacy systems.

About a quarter of the projects were driven by a desire to improve agility, in three ways. The first was to accelerate the response times for analytics. In business functions such as fraud detection and automated trading, the speed with which a firm can test new ideas (i.e., algorithms) has a direct impact on how quickly it can react to changing conditions in the external world. (In fact, one firm credited their big data platform with enabling an arrest of a fraudulent card user on the day following the first incident!) A second driver was to overcome rigidities in data integration. For example, one bank said that its RDBMS-oriented process for loading data into a data warehouse required many person hours and much calendar time just to incorporate a new attribute that appeared in source data. Finally, some banks sought agility through increased automation of process bottlenecks. In one case, automating SOX compliance analysis enabled the bank to prioritize among numerous potential remediation areas by performing what-if analyses of their impacts.

The remaining quarter of the projects were driven primarily in pursuit of IT cost savings. Cost containment continues to be a key strategic imperative for banks, not only to boost efficiency but also to conserve capital. The amount of a bank's capital directly affects how much risk it can take on, which in turn affects its potential revenue. More stringent capital adequacy requirements from the regulations above have only made that equation more difficult. Banks told us that through deployment of new architectures to handle workloads in this study, they have reduced costs for software, hardware, and labor—and they aggressively intend to extend those savings. For example, firms cited the license costs of relational databases, data warehouse appliances, and analytic databases as a key target for reduction. A significant number were also using distributed file systems and low-cost disk storage to replace traditional SANs and tape backups. And some firms were on a clear path to offload workloads from their mainframes, in order to reduce cost allocations for both mips and storage.

Irrespective of their initial business driver, nearly all of the projects enjoyed IT cost savings. We frequently heard banks say: "The project paid for itself quickly." This is an example of the virtuous circle that attends many disruptive technologies. As illustrated in Figure 2, a business unit that adopts big data technology for one reason soon finds itself able to realize additional benefits. In several cases that we studied, no matter where the firm started on the circle, it found itself able to move in either direction once it had achieved its original goals. For example, a bank that used big data technology in order to revolutionize the comprehensiveness of its IT policy compliance analysis to meet expanding regulatory requirements also reduced costs significantly compared to its previous process.

Virtuous circle: business drivers and side benefits

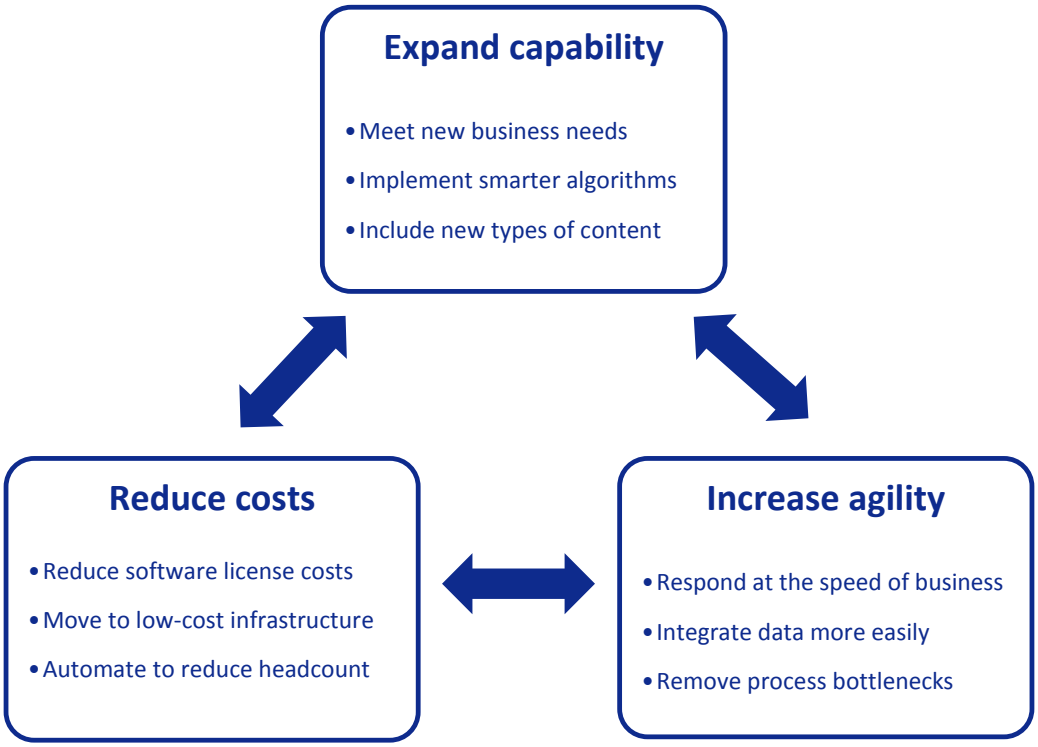

Figure 2

Nature of the workloads

Figure 3 plots the workloads we studied against three dimensions that matter to a solution designer:

1. The horizontal axis represents the "structuredness" of the data involved. In casual conversations (and the tech media) this is often presented as a Boolean variable: content is either structured or it's not. But the reality is more complex. For example, many people consider the body of an email to be unstructured, since it has no rules for what goes where. However, some banks think of email messages as structured data because they are readily parsed into identifiable concepts that can be searched and otherwise operated upon. In the end, most banks view "structuredness" as a spectrum that ranges from highly structured content like transaction records (which tend to have pre-determined layouts and clearly defined field values) to highly unstructured content such as video (where pretty much the only element of structure is time).
2. The vertical axis corresponds to "analytic complexity", which ranges from simple transformations at the low end of the scale to machine learning at the high end. While vague, analytic complexity signals the sophistication of the personnel and tools required. The more complex cases require data scientists and advanced statistical tools, while the less complex are readily handled by administrators wielding scripts. Analytic complexity is also a proxy for the data-processing intensity of the workload—that is, how much it exercises compute and I/O. However, it's an imperfect proxy, since the extent to which a given workload is bottlenecked on processors, networks, or storage will depend a great deal on the architecture that carries it. For example,

traversing a graph is efficient in a well-tuned graph database but can be horribly burdensome to a relational database.

3. The thickness of each ellipse crudely represents the size of the data involved in the workload. Data sets less than a petabyte have thin borders, while those of a petabyte or greater have heavy borders. Note that a petabyte is an arbitrary dividing line. Less than a petabyte of data might still be considered quite high volume.

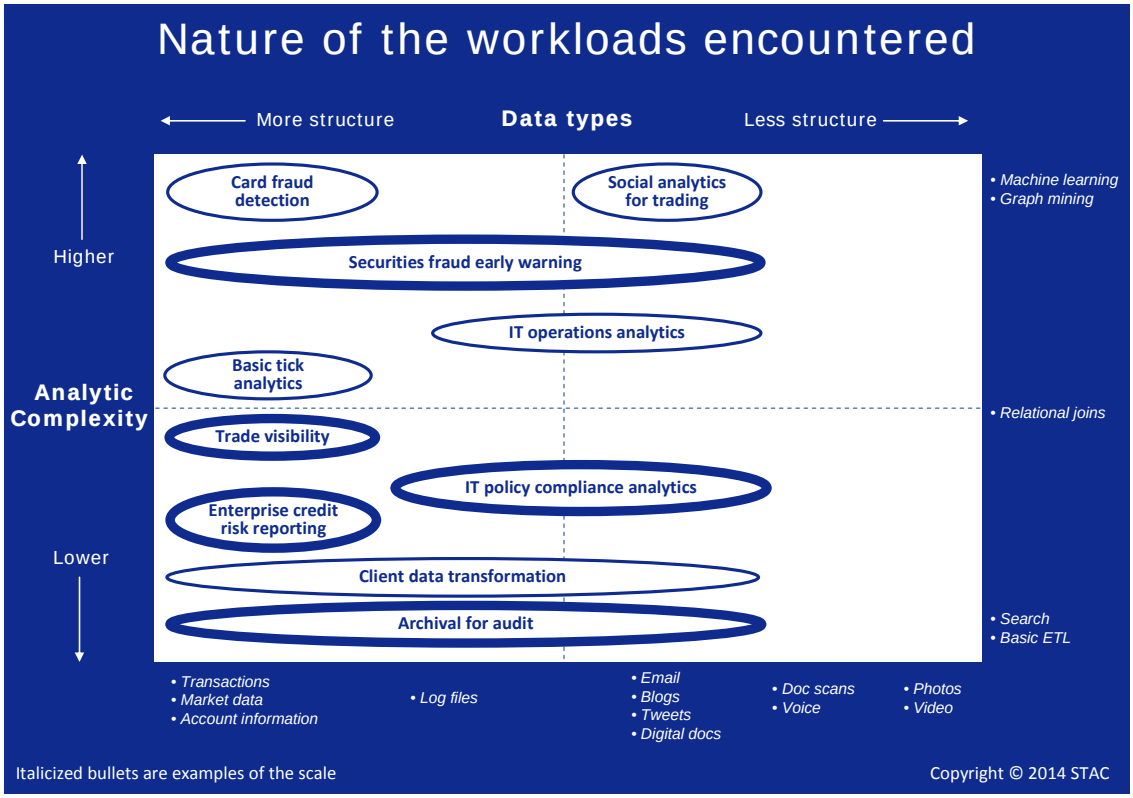

Figure 3

It is evident from the chart that about half of the workloads were a petabyte or greater. One can also note that although there was a range of "structuredness" in the content covered by these workloads, it was biased toward the structured end and was utterly unrepresented in the completely unstructured end (for example, none of our cases involved images or video). In fact, the diagram doesn't entirely show this bias. Although roughly half of the workloads required integration of data from different systems (i.e., spanning the silos), those data were fairly structured. Nevertheless, the data sometimes varied greatly at the semantic level (e.g., different departments using the same field for a different purpose or the same purpose but using a different vocabulary, etc.). This sort of variability was a key challenge for integration in some cases.

In terms of analytic complexity, our cases were about evenly split between the relatively simple and the relatively complex. Significantly, a key driver of the projects concerning complex analytic workloads was improving the quality of those analytics. This sometimes meant analyzing an entire dataset instead of statistical samples. In other cases, it was about supplementing traditional data with new kinds of content (e.g., email, social media). And sometimes the key to quality was accelerating the analytics. Decreasing the time-to-answer enabled more iterations, which in turn led to better models.

Framing these in the Three V's

It's worth noting that Figure 3 implicitly captures two of the famous "Three V's" popularized by IT analysts and vendors (volume, variety, and velocity).[5] As explained, the thickness of an ellipse indicates volume. And its width indicates variety in the data structures. So what about the third V, velocity? Why doesn't it feature in our conceptual map? Simply because it only arose in one case.

In big data circles, the velocity challenge is variously defined as:

1. Keeping up with a high rate of data flowing into the system. That is, not losing data.
2. Quickly making data available for queries as it streams in. That is, not making users wait until tomorrow to query what happened 6 minutes ago.
3. Performing event-driven ("streaming") analytics. That is, enabling algorithms to respond in real time to incoming data. (Note that this "push" model is different from ensuring up-to-date data for the "pull" model of velocity type 2.)

Just one of our cases was driven to new technology in part by a concern with velocity: social analytics for trading. The event-driven pattern in this workload (as opposed to its historical analysis pattern) is all about maximizing the speed of semantic analysis on incoming textual data (velocity type 3). Too much latency between a market moving tweet and an action by a trading algorithm can turn profit to loss.

Aside from this single case, none of the cases offered up by the banks were driven by a concern with velocity. While some of the banks forecast an eventual requirement to handle high velocity in fraud or customer-service scenarios, they did not have that requirement today. And although it's true that any solution for basic tick analytics is sensitive to velocity type 1 (often ingesting data from the world's busiest exchanges at millions of updates per second), keeping up with ingest rates is not what motivated banks in this study to use new technologies for tick data. Traditional "tick databases" are more than capable of handling today's ingest rates. The motivation was to reduce cost.

In our estimation, the tick data example illustrates the general reason why data velocity did not surface as a driver of big data technology adoption in investment banks. For many years, these firms have dealt with higher message rates than most firms in other industries, so their "traditional technologies" are not challenged.[6] They have built their own software or purchased highly tuned applications and middleware from specialist vendors . Big data software from Silicon Valley, as innovative as it is, does not increase capability along this dimension at this time.[7]

"Certain mainframe workloads"

The workloads in Figure 3 are defined in terms of their inputs, outputs, and functions, irrespective of the technology used to support them. Table 1, however, includes one entry that is defined in terms of

[5] The earliest reference we have found was from the Meta Group in 2001. See: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[6] This underscores the contextual nature of our definition of big data. While an electrical utility contemplating how to deal with an influx of 100,000 smart meter updates per second is likely to consider that a big data problem, a bank accustomed to consuming and acting on 6 million updates per second in the equities and options market is not.

[7] This is not to say that big data technologies will never enter the high-velocity niches in finance. As these technologies mature in other industries that have velocity issues (perhaps those dealing with the "Internet of Things"), banks could start to adopt them if they reduce costs or offer some other compelling benefits.

the technology used to support it: "Certain mainframe workloads". While this category no doubt overlaps with other workloads in the list (Enterprise Credit Risk Reporting was one example we encountered), we gave it its own designation because the customers who brought it up seemed fairly energetic about moving these workloads to Hadoop.

The datasets in these workloads may be of modest size (hundreds of TB), but the amount of processing tends to be large. What they have in common is that they utilize certain VSAM file structures that easily translate to Hadoop key-value format.

The banks said that mainframe resources are expensive relative to big data alternatives. Part of this is the cost of mainframe processing time and storage, which is due both to the acquisition cost of the machinery and software and to the costs of operating the system. The other part was the cost (and risk) of relying on COBOL programming: finding good talent for this legacy skillset is getting harder.

Nature of the solutions

While the main point of this project was to learn about big data workloads (which are independent of the technology used to support them at any point in time), we also asked interviewees about the ways they had solved these big data challenges, if indeed they had. It turns out that about two thirds of the workloads in Table 1 were being handled in production by big data technologies at the time of our interviews.

Hadoop played a central role in most deployments and was also the focus for those cases where the solution had not yet been set. Other technologies were often involved (e.g., NoSQL databases), but Hadoop was at the center. The major exceptions were a tick analytics deployment that used a document database and an IT policy analytics deployment based on a graph database. Apache Spark and other distributed, in-memory technologies also appeared to be gaining mindshare with interviewees (perhaps market share), particularly in time-sensitive areas like social analytics for trading. On the front end of the solution stack, several banks mentioned R and SAS on Hadoop as important components, while Python and a range of visualization tools also came into play.

The role of corporate IT

Because a great deal of big data software is open source, it is relatively easy for any technologist within a bank to download a product and experiment with it. By lowering barriers to entry, this makes it tempting for business units to build their own big data solutions without relying on central IT groups. Not surprisingly, several of the deployments we encountered were at a business-unit level. Nevertheless, several of them involved a shared infrastructure managed by corporate IT.

Figure 4 offers a way of thinking about the role of corporate IT departments in terms of the service layers of any analytics proposition. In about half the cases we encountered, corporate IT served as a center of excellence—advising line-of-business (LOB) technologists on which technologies to use and how to set them up—or had taken on responsibility for managing a number of LOB-specific clusters (these are the first two columns in the diagram).

In the other half, the banks had built or were in the process of building a new centralized platform to provide analytic services to several lines of business, which we dubbed an "analytics platform as a service". In some cases, these platforms had a corporate-wide mandate across retail, commercial, and investment banking. The firms said that the benefits of this model included cost efficiency and the potential to exploit shared data. However, few of them, if any, had built shared data pools yet (i.e., an "enterprise data hub"), despite a long-run ambition to do so. While data from different departments

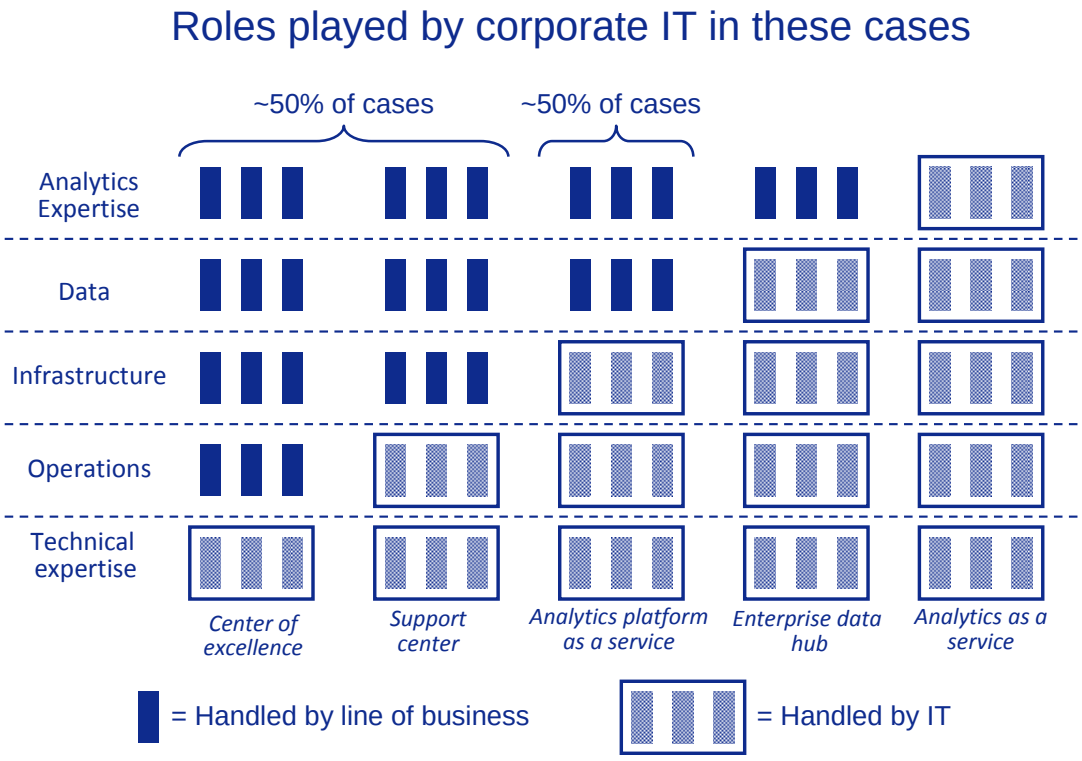

Figure 4

might be brought together, the resulting data stores were dedicated to a particular business purpose (e.g., using commercial bank information for the benefit of the retail bank). They did not constitute a so-called "data lake", from which any line of business could drink. Nor had any bank achieved end-to-end "analytics as a service", though some had an ambition to provide the necessary analytics expertise.

Analytics platform as a service

Centralized analytics are not new. Many firms have enterprise data warehouses and BI tools that serve multiple businesses. What is new is the kind of analytics involved, an increased desire to leverage data across traditionally siloed businesses, and a desire to shift the IT cost curve radically.

We considered centralized analytics a "meta-workload": i.e., a workload consisting of multiple underlying workloads from multiple lines of business. That is, centralized analytics platforms are "multi-tenant." Managing multi-tenant workloads introduces technical challenges beyond those imposed by each workload on its own, which makes this case worthy of its own discussion.

The sub-workloads on these platforms spanned both interactive analytics and batch jobs. They included several of the workloads profiled elsewhere in this study, plus many others, such as:

- Credit card prospecting and marketing
- Consumer credit-risk analysis (incorporating commercial bank information)
- Marketing equities research to institutional investors
- Anti money laundering
- General ETL offload for multiple use cases. (This does not really fit the definition of analytics but leverages the same infrastructure.)

Despite the variety of workloads, the new analytics platforms we studied had a few things in common. The petabytes of data they housed were skewed toward the structured end of the spectrum, just as we saw for the individual workloads in Figure 1. The users of the platform were internal to the bank (that is, direct use of the service had not yet been extended to customers). And the core of the platform was a Hadoop cluster, from tens to low hundreds of nodes. This was usually integrated with an existing data warehouse, and often supplemented with NoSQL databases and a variety of analytic tools, including existing BI tools.

The hurdles that they faced along this road were common to any major implementation of a shared service:

- Disaster recovery and failover planning. This includes addressing cross-border regulations, as well as differentiating critical "golden source" data, which has high availability requirements, from "exploratory" data whose loss could be tolerated.
- Transition and migration costs (cost of hiring and developing staff).
- Governance: Establishing rules for who gets what resources when.
- Providing quick response times to accommodate new workloads, and offering multiple configuration options to meet workload requirements.

In one interview, an internal user unfavorably compared the flexibility and agility of the firm's centralized Hadoop service to that of a particular cloud-based Hadoop provider. While business groups were mostly prohibited from using such outsourced alternatives (which is another story), those cloud services provided an external benchmark for cost and service levels that challenged internal IT.

Key Challenges/Gaps in Big Data Technology

As a final point of inquiry, we asked interviewees about key challenges facing big data technologies in their institutions, either for the use case in question or with respect to expanding the technologies to additional use cases. A few things came up more than once:

- **Security and entitlements.** Carefully and reliably controlling who can access what data is fundamental to financial services, from investment banking and brokerage to retail banking. In some cases, inadequate support for access control is holding back adoption of big data products. It almost certainly precludes the creation of cross-business data lakes. The required functionality includes high-performance encryption of data at rest and in-flight, as well as fine-grained entitlements control that can be managed via standard corporate systems.

- **Multi-tenancy.** Multitenancy is the practice of operating multiple independent applications in a shared environment where they are competing for underlying resources, e.g. operation within a Hadoop cluster. Under these conditions, the ability to share resources and data across multiple businesses while maintaining the required level of service to each is a key concern. Poorly behaved applications can be "noisy neighbors," consuming so much resource that they degrade the performance of other apps in the cluster.

- **Product and vendor maturity.** Most of the firms we interviewed accepted the high rate of change in open source big data products as a cost of doing business. And they viewed the relatively small size of the related vendors as a risk worth taking. However, some of them felt these issues limited the footprint of big data technologies in their organizations. In particular, product changes that require modifications to applications can be problematic in highly

regulated sectors, where application owners must demonstrate that modified application logic continues to satisfy regulations.

- **Interop limitations.** Most of the big data solutions in our study required integration with existing systems. The banks seemed fairly happy with interoperability on the back end but sometimes felt that front-end interop was limited (e.g., popular BI tools suffered reduced functionality on Hadoop). Others expressed dissatisfaction with the level of performance they could achieve when integrating multiple big data infrastructures (e.g., Hadoop and NoSQL databases). Each delivered great performance on its own; but when combined, the whole was less than the sum of the parts.

Conclusion

It bears repeating that we do not view this set of use cases as representative of all big data use cases in retail and investment banking. On the contrary, we think they only scratch the surface, since we came across many additional cases that we did not have time to profile. However, even our limited sample points to some fairly robust conclusions. Clearly there are many workloads that banks feel are too difficult or expensive to handle using traditional technologies due to scale or complexity. The potential of new technologies to expand capability, improve agility, and reduce costs in such cases is substantial. But these technologies must overcome specific hurdles in order to broaden their acceptability to banks.

In the interplay that will ensue between customer needs and product evolution, we believe that independent, technology-agnostic benchmark standards will be an important catalyst. They both accelerate technology selection at user firms and shorten the sales cycle for vendors. By demonstrating the strengths and weaknesses of any technology stack, such standards also help ensure that deployments are successful and that products succeed in the market based on their merits. And by providing vendors with a set of relevant workloads for product developers and a rallying point for product marketers, multi-customer benchmarks propel the entire industry forward at a faster rate.

We invite both technology users and technology providers to join the STAC Benchmark Council's project to produce these standards for big data.[8]

[8] www.STACresearch.com/bigsig