O'REILLY®

Data Infrastructure for Next-Gen Finance

Tools for Cloud Migration, Customer Event Hubs, Governance & Security

Jane Roberts

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

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

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

Job # D2044

Data Infrastructure for Next-Gen Finance

Tools for Cloud Migration, Customer Event Hubs, Governance & Security

Jane Roberts

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

Data Infrastructure for Next-Gen Finance
by Jane Roberts

Printed in the United States of America.

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

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

Editor: Nicole Tache
Production Editor: Kristen Brown
Copyeditor: Octal Publishing, Inc.
Interior Designer: David Futato
Cover Designer: Karen Montgomery

June 2016: First Edition

Revision History for the First Edition
2016-06-09: First Release

978-1-491-95966-4

[LSI]

Table of Contents

Preface

This report focuses on data infrastructure, engineering, governance, and security in the changing financial industry. Information in this report is based on the 2015 Strata + Hadoop World conference sessions hosted by leaders in the software and financial industries, including Cloudera, Intel, FINRA, and MasterCard.

If there is an underlying theme in this report, it is the big yellow elephant called Hadoop—the open source framework that makes processing large data sets possible. The report addresses the challenges and complications of governing and securing the wild and unwieldy world of big data while also exploring the innovative possibilities that big data offers, such as customer event hubs. Find out, too, how the experts avoid a security breach and what it takes to get your cluster ready for a Payment Card Industry (PCI) audit.

CHAPTER 1

Cloud Migration: From Data Center to Hadoop in the Cloud

Jaipaul Agonus
FINRA

How do you move a large portfolio of more than 400 batch analytical programs from a proprietary database appliance architecture to the Hadoop ecosystem in the cloud?

During a session at Strata + Hadoop World New York 2015, Jaipaul Agonus, the technology director in the market regulation department of FINRA (Financial Industry Regulatory Authority) described this real-world case study of how one organization used Hive, Amazon Elastic MapReduce (Amazon EMR) and Amazon Simple Storage Service (S3) to move a surveillance application to the cloud. This application consists of hundreds of thousands of lines of code and processes 30 billion or more transactions every day.

FINRA is often called "Wall Street's watch dogs." It is an independent, not-for-profit organization authorized by Congress to protect United States investors by ensuring that the securities industry operates fairly and honestly through effective and efficient regulation. FINRA's goal is to maintain the integrity of the market by governing the activities of every broker doing business in the US. That's more than 3,940 securities firms with approximately 641,000 brokers.

How does it do it? It runs surveillance algorithms on approximately 75 billion transactions daily to identify violation activities such as

market manipulation, compliance breaches, and insider trading. In 2015, FINRA expelled 31 firms, suspended 736 brokers, barred 496 brokers, fined firms more than $95 million, and ordered $96 million in restitution to harmed investors.

The Balancing Act of FINRA's Legacy Architecture

Before Hadoop, Massively Parallel Processing (MPP) methodologies were used to solve big data problems. As a result, FINRA's legacy applications, which were first created in 2007, relied heavily on MPP appliances.

MPP tackles big data by partitioning the data across multiple nodes. Each node has its own local memory and processor, and the distributed nodes are handled by a sophisticated centralized SQL engine, which is essentially the brain of the appliance.

According to Agonus, FINRA's architects originally tried to design a system in which they could find a balance between cost, performance, and flexibility. As such, it used two main MPP appliance vendors. "The first appliance was rather expensive because it had specialized hardware due to their SQL engines; the second appliance, a little less expensive because they had commodity hardware in the mix," he said.

FINRA kept a year's worth of data in the first appliance, including analytics that relied on a limited dataset and channel, and a year's worth of data in the second appliance—data that can run for a longer period of time and that needs a longer date range. After a year, this data was eventually stored offline.

Legacy Architecture Pain Points: Silos, High Costs, Lack of Elasticity

Due to FINRA's tiered storage design, data was physically distributed across appliances, including MPP appliances, Network-Attached Storage (NAS), and tapes; therefore, there was no one place in its system where it could run all its analytics across the data. This affected accessibility and efficiency. For example, to rerun old data, FINRA had to do the following:

- To rerun data that was more than a month old, it had to rewire analytics to be run against appliance number two.
- To rerun data that was more than a year old, it had to call up tapes from the offline storage, clear up space in the appliances for the data, restore it, and revalidate it.

The legacy hardware was expensive and was highly tuned for CPU, storage, and network performance. Additionally, it required costly proprietary software, forcing FINRA to spend millions annually, which indirectly resulted in a vendor lock-in.

Because FINRA was bound by the hardware in the appliances, scaling was difficult. To gauge storage requirements, it essentially needed to predict the future growth of data in the financial markets. "If we don't plan well, we could either end up buying more or less capacity than we need, both causing us problems," said Agonus.

The Hadoop Ecosystem in the Cloud

Many factors were driving FINRA to migrate to the cloud—the difficulty of analyzing siloed data, the high cost of hardware appliances and proprietary software, and the lack of elasticity. When FINRA's leaders started investigating Hadoop, they quickly realized that many of their pain points could be resolved. Here's what they did and how they did it.

FINRA's cloud-based Hadoop ecosystem is made up of the following three tools:

Hive
: This is the de facto standard for SQL-on-Hadoop. It's a component of Hortonworks Data Platform (HDP) and provides a SQL-like interface.

Amazon EMR
: This is a managed Hadoop framework that brings elasticity to Hadoop clusters in the cloud.

Amazon S3
: This is Amazon's storage service with practically infinite storage (up to five terabytes).

SQL and Hive

FINRA couldn't abandon SQL because it already had invested heavily in SQL-based applications running on MPP appliances. It had hundreds of thousands of lines of legacy SQL code that had been developed and iterated over the years. And it had a workforce with strong SQL skills. "Giving up on SQL would also mean that we are missing out on all the talent that we've attracted and strengthened over the years," said Agonus.

As for Hive, users have multiple execution engines:

MapReduce
: This is a mature and reliable batch-processing platform that scales well for terabytes of data. It does not perform well enough for small data or iterative calculations with long data pipelines.

Tez
: Tez aims to balance performance and throughput by streaming the data from one process to another without actually using HDFS. It translates complex SQL statements into optimized, purpose-built data processing graphs.

Spark
: This takes advantage of fast in-memory computing by fitting all intermediate data into memory and spilling back to disk only when necessary.

Amazon EMR

Elastic MapReduce makes easy work of deploying and managing Hadoop clusters in the cloud. "It basically reduces the complexity of the time-consuming set up, management, and tuning of the Hadoop clusters and provides you a resizable cluster of Amazon's EC2 instances," said Agonus. EC2 instances are essentially virtual Linux servers that offer different combinations of CPU, memory, storage, and networking capacity in various pricing models.

Amazon S3

S3 is a cost-effective solution that handles storage. Because one of FINRA's architecture goals was to separate the storage and compute resources so that it could scale them independently, S3 met its requirements. "And since you have the source data set available in

S3, you can run multiple clusters against that same data set without overloading your HDFS nodes," said Agonus.

All input and output data now resides in S3, which acts like HDFS. The cluster is accessible only for the duration of the job. S3 also fits Hadoop's file system requirements and works well as a storage layer for EMR.

Capabilities of a Cloud-Based Architecture

With the right architecture in place, FINRA found it had new capabilities that allowed it to operate in an isolated virtual network (VPC, or *virtual private cloud*). "Every surveillance has a profile associated with it that lets these services know about the instance type and the size needed for the job to complete," said Agonus.

The new architecture also made it possible for FINRA to store intermediate datasets; that is, the data produced and transferred between the two stages of a MapReduce computation—map and reduce. The cluster brings in the required data from S3 through Hive's external tables and then sends it to the local HDFS for further storing and processing. When the processing is complete, the output data is written back to S3.

Lessons Learned and Best Practices

What worked? What didn't? And where should you focus your efforts if you are planning on migrating to the cloud? According to Agonus, your primary objective in designing your Hive analytics would be to focus on direct data access and maximizing your resource utilization. Following are some key lessons learned from the FINRA team.

Secure the financial data
: The audience asked how FINRA secured the financial data. "That's the very first step that we took," said Agonus. FINRA has an app security group that performed a full analysis on the cloud, which was a combined effort with the cloud vendor. They also used encryption on their datacenter. This is part of Amazon's core, he explained. "Everything that is encrypted stays encrypted," he said. "Amazon's security infrastructure is far more extensive than anything we could build in-house."

Conserve resources by processing necessary data
: Because Hive analytics enable direct data access, you need only partition the data you require. FINRA partitions its trade dataset based on a trade date. It then process only the data that it needs. As a result, it don't waste resources trying to scan millions upon millions of rows.

Prep enhances join performance
: According to Agonus, bucketing and sorting data ahead of time enhances join performance and reduces the I/O scan significantly. "Joins also work much faster because the buckets are aligned against each other and a merge sort is applied on them," he said.

Tune the cluster to maximize resource utilization
: Agonus emphasized the ease of making adjustments to your configurations in the cloud. Tuning the Hive configurations in your cluster lets you maximize resource utilization. Because Hive consumes data in chunks, he says, "You can adjust minimum/maximum splits to increase or decrease the number of mappers or reducers to take full advantage of all the containers available in your cluster." Furthermore, he suggests, you can measure and profile your clusters from the beginning and adjust them continuously as your data size changes or the execution framework changes.

Achieve flexibility with Hive UDFs when SQL falls short
: Agonus stressed that SQL was a perfect fit for FINRA's application; however, during the migration process, FINRA found two shortcomings with Hive, which it overcame by using Hive user defined functions (UDFs).

  The first shortcoming involved Hive SQL functionality compared to other SQL appliances. For example, he said, "The Windows functions in Netezza allow you to ignore nulls during the implementation of PostgreSQL, but Hive does not." To get around that, FINRA wrote a Java UDF that can do the same thing.

  Similarly, it discovered that Hive did not have the date formatting functions it was used to in Oracle and other appliances. "So we wrote multiple Java UDFs that can convert formats in the way we like," said Agonus. He also reported that Hive 1.2 supports date conversion functions well.

The second shortcoming involved procedural tasks. For example, he said, "If you need to de-dupe a dataset by identifying completely unique pairs based on the time sequence in which you receive them, SQL does not offer a straightforward way to solve that." However, he suggested writing a Java or Python UDF to resolve that outside of SQL and bring it back into SQL.

Choose an optimized storage format and compression type
: A key component of operating efficiently is data compression. According to Agonus, the primary benefit of compressing data is the space you save on the disk; however, in terms of compression algorithms, there is a bit of a balancing act between compression ratio and compression performance. Therefore, Hadoop provides support for several compression algorithms, including gzip, bzip2, Snappy, LZ4 and others. The abundance of options, though, can make it difficult for users to select the right ones for their MapReduce jobs.

  Some are designed to be very fast, but might not offer other features that you need. "For example," says Agonus, "Snappy is one of the fastest available, but it doesn't offer much in space savings comparatively." Others offer great space savings, but they're not as fast and might not allow Hadoop to split and distribute the workload. According to Agonus, Gzip compression offers the most space saving, but it is also among the slowest and is not splittable. Agonus advises choosing a type that best fits your use case.

Run migrated processes for comparison
: One of the main mitigation strategies FINRA used during the migration was to conduct an apples-to-apples comparison of migrated processes with its legacy output. "We would run our migrated process for an extensive period of time, sometimes for six whole months, and compare that to the output in legacy data that were produced for the same date range," said Agonus. "This proved very effective in identifying issues instantly." FINRA also partnered with Hadoop and cloud vendors who could look at any core issues and provide it with an immediate patch.

Benefits Reaped

With FINRA's new cloud-based architecture, it no longer had to project market growth or spend money upfront on heavy appliances

based on projections. Nor did it need to invest in a powerful appliance to be shared across all processes. Additionally, FINRA's more dynamic infrastructure allowed it to improve efficiencies, running both faster and more easily. Due to the ease of making configuration changes, it was also able to utilize its resources according to its needs.

FINRA was also able to mine data and do machine learning on data in a far more enhanced manner. It was also able to decrease its emphasis on software procurement and license management because the cloud vendor performs much of the heavy lifting in those areas.

Scalability also improved dramatically. "If it's a market-heavy day, we can decide that very morning that we need bigger clusters and apply that change quickly without any core deployments," said Agonus. For example, one process consumes up to five terabytes of data, whereas others can run on three to six months worth of data. Lastly, FINRA can now reprocess data immediately without the need to summon tapes, restore them, revalidate them, and rerun them.

CHAPTER 2

Preventing a Big Data Security Breach: The Hadoop Security Maturity Model

Nick Curcuru
MasterCard

Ritu Kama
Intel

Sam Heywood
Cloudera

Hadoop is widely used thanks to its ability to handle volume, velocity, and a variety of data. However, this flexibility and scale presents challenges for securing and governing data. In a talk at Strata + Hadoop World New York 2015, experts from MasterCard, Intel, and Cloudera shared what it takes to get your cluster PCI-compliance ready. In this section, we will recap the security gaps and challenges in Hadoop, the four stages of the Hadoop security maturity model, compliance-ready security controls, and MasterCard's journey to secure their big data.

Hadoop Security Gaps and Challenges

According to Ritu Kama, director of product management for big data at Intel, the security challenges that come with Hadoop are based on the fact that it wasn't designed with security in mind; therefore, there are security gaps within the framework. If you're a business manager, for example, and you're thinking of creating a *data lake* because you'd like to have all your data in a single location and be able to analyze it holistically, here are some of the security questions and challenges Kama says you will need to address:

- Who's going to have access to the data?
- What can they do with the data?
- How is your framework going to comply with existing security and compliance controls?

Kama says one of the reasons that big goals and vague projects like data lakes fail is because they don't end up meeting the security requirements, either from a compliance standpoint or from an IT operations and information security perspective.

Perimeter security is just a start. "It's no longer sufficient to simply build a firewall around your cluster," says Kama. Instead, you now need to think about many pillars of security. You need to address all of the network security requirements as well as authentication, authorization, and role-based access control.

You also need visibility into what's going on in your data so that you can see how it's being used at any given moment, in the past or present. Audit control and audit trails are therefore extremely important pillars of security. They are the only way you can figure out who logged in, who did what, and what's going on with the cluster.

Next, and above all else, you clearly need to protect the data. "Once the data is on the disk, it's vulnerable," said Kama. "So is that data on the disk encrypted? Is it just lying there? What happens when somebody gets access to that disk and walks away with it?" These are among the security issues and challenges that the enterprise must confront.

The Hadoop Security Maturity Model

According to Sam Heywood, director of product management at Cloudera, "The level of security you need is dependent upon where you are in the process of adoption of Hadoop, for example whether you are testing the technology or actually running multiple workloads with multiple audiences."

The following describes the stages of adoption referred to as the *Hadoop security maturity model*, as developed by Cloudera and adopted by MasterCard and Intel. Follow these steps to get your cluster compliance ready.

Stage 1: Proof of Concept (High Vulnerability)

Most organizations begin with a proof of concept. At this stage, only a few people will be working with a single dataset; these are people who are trying to understand the performance and insights Hadoop is capable of providing. In this scenario, anyone who has access to the cluster has access to all of the data, but the number of users is quite limited. "At this point, security is often just a firewall, network segregation or what's called 'air gapping' the cluster," said Heywood.

Stage 2: Live Data with Real Users (Ensuring Basic Security Controls)

The next stage involves live datasets. This is when you need to ensure that you have the security controls in place for strong authentication, authorization, and auditing. According to Heywood, this security will control who can log into the cluster, what the dataset will include, and provide an understanding of everything else that is going on with the data.

Stage 3: Multiple Workloads (Data Is Managed, Secure, and Protected)

Now you're ready to run multiple workloads. At this stage, you are running a multitenant operation and therefore you need to lock down access to the data; authorization and security controls are even more important at this juncture.

The goal at this stage, explained Heywood, is to be able to have all of the data in a single enterprise data hub and allow access to the

appropriate audiences, while simultaneously ensuring that there are no incidental access issues for people who shouldn't have access to given datasets. Additionally, you will need to provide a comprehensive audit.

Stage 4: In Production at Scale (Fully Compliance Ready)

After you have a production use case in place, you can begin moving over other datasets. You are finally ready to run at scale and in production with sensitive datasets that might fall under different types of regulations. According to Heywood, this is when you need to run a fully compliance-ready stack, which includes the following:

- Encryption and key management in place
- Full separation of duties
- Separate sets of administrators to configure the parameter, control the authorization layer, and conduct an ongoing audit
- A separate class of users who are managing the keys tied to the encryption

That's an overview of what it takes to run a fully compliant data hub. Now let's take a look at the tools to get there.

Compliance-Ready Security Controls

The following describes the journey to full compliance with a focus on the tools to configure a data hub.

Cloudera Manager (Authentication)

Using Cloudera Manager, you can configure all the Kerberos authentication within your environment. However, very few people know how to configure and deploy a Kerberos cluster. As a result, MasterCard automated the configuration process, burying it behind a point-and-click interface.

Apache Sentry (Access Permissions)

After you have authenticated the user, how do you control what they have access to? This is where Apache Sentry comes in. According to Heywood, Apache Sentry provides role-based access control that's

uniformly enforced across all of the Hadoop access paths. You specify your policies and the policies grant a given set of permissions. Then, you link Active Directory groups to those Sentry roles or policies, and users in those groups get access to those datasets.

According to Heywood, what makes this method powerful is that the roles provide an abstraction for managing the permissions. When multiple groups need identical sets of access within a cluster, you don't need to configure the permissions for each of those groups independently. "Instead, you can set the access policy and relate all those groups back to that one policy; this creates a much more scalable, maintainable way to manage permissions," Heywood said.

Cloudera Navigator (Visibility)

"Even if you have strong authentication and authorization in place, you need to understand what's happening within the cluster," said Heywood. For example:

- What are users doing?
- Who's accessing a given asset?
- How are those assets being copied around?
- Where is sensitive data being copied to in new files?

This falls under the visibility pillar. Cloudera Navigator provides a comprehensive audit of all the activities that are taking place within the cluster and is the tool of choice for MasterCard.

HDFS Encryption (Protection)

The compliant data protection is encryption. "With HDFS encryption, you can encrypt every single thing within the cluster for HDFS and HBase," said Heywood. This encryption is end-to-end, meaning that data is encrypted both at-rest and in-flight; encryption and decryption can be done only by the client. "HDFS and HDFS administrators never have access to sensitive key material or unencrypted plain text, further enhancing security," said Heywood.

Cloudera RecordService (Synchronization)

Cloudera has recently introduced RecordService, a tool that delivers fine-grained access control to tools such as Spark. According to

Heywood, "If you had a single dataset and you had to support multiple audiences with different views on that dataset through MapReduce or Spark, typically what you had to do is create multiple copies of the files that were simply limited to what the given audience should have, and then give the audience access to that one file." This would normally be very difficult primarily due to synchronization issues and keeping copies up to date. RecordService eliminates those problems. "You configure the policy in Sentry, it's enforced with RecordService, and no matter which way the user is trying to get to the data, we will uniformly and consistently provide the level of access control that's appropriate for the policy," said Heywood.

MasterCard's Journey

What does it take from an organizational perspective to combine the technology with the internal processes to become a PCI-ready enterprise data hub? To help answer that question, Nick Curcuru, principal of big data analytics at MasterCard Advisors, told his company's story.

MasterCard has the second-largest database in the world with more than 15 to 20 petabytes of data, and doubling annually. Additionally, MasterCard applies 1.8 million rules every time you swipe your card. "We process it in milliseconds," said Curcuru. "And we do that in over 220 countries, 38,000 banks, and several million merchants and users of the system." Big data is not a new concept for this company. By 2012, Hadoop was a platform that could begin to keep up with its needs and requirements.

During MasterCard's first pilot of Hadoop, it was still just trying to determine what it could do with Hadoop and whether Hadoop was capable of providing value. It was immediately apparent to the analysts that it was definitely something MasterCard wanted to work with because it would allow it to bring together all of its datasets rapidly and in real time. But MasterCard's first and most immediate question was how to secure it. Because there was no security built into the platform at the time, MasterCard decided to partner with Cloudera. It began building use cases and going mainstream by the end of 2012, as it was able to create more and more security. MasterCard achieved wide adoption of Hadoop technologies by 2013 and became a PCI-certified organization in 2014.

What follows is an account of what MasterCard learned along the way, and the advice it offers to others attempting to create their own compliance-ready enterprise data hub.

Looking for Lineage

MasterCard knew that lineage and an audit trail were imperatives. "We are about security in everything we do," said Curcuru. "From our CEO who tells us that we are stewards of people's digital personas, all the way down to individuals who are actually accessing that data."

MasterCard's tools of choice? Cloudera Navigator, which is native in Hadoop, Teradata, Oracle, and SAS. The company advises using native tools first and then looking outside for additional support. After you have the tools in place, what do you need to do to become PCI certified. "You have to create a repeatable process in three areas: people, process, and technology," said Curcuru.

Educating the People

"Most people focus on technology," said Curcuru, but technology is only about 10 percent or 15 percent of what it takes. "People and process play a much bigger role, largely due to the education needed across the board. You're going to have to educate people how the cluster gets secured. Not only your IT department, but everyone from marketing to auditors to the executive team must be trained on Hadoop's security capabilities."

The following sections describe the areas where MasterCard found people needed the most training.

Segregation of Duties

Segregation of duties matters when you're conducting an audit or securing a cluster. If there is a breach, you can identify whether it happened by someone with a valid ID. "A lot of the breaches are happening now with valid credentials," said Curcuru. Gone are the days when one database administrator can have access to your entire system. Similarly, he explained, if your analysts only have access to a certain piece of data and you have a breach, you can isolate where

that breach happened. Auditors, he emphasized, will ask about segregation of duties and who can access what.

Documentation

Documentation is critical. Like it or not, if there is a breach, auditors and attorneys are all going to be asking for it. "This is where automated lineage comes into play," said Curcuru. "You don't want to be creating documentation manually."

Awareness Training

Continual knowledge transfer and awareness training are also absolute necessities. For example, warns Curcuru, be aware of the breach that memory sticks create. "Don't leave those sticks in your hotel room. That type of awareness training, that type of knowledge transfer, has to exist," he said.

Strong Authentication

Curcuru advocates multifactor authentication, which is a security system that requires more than one method of authentication from independent categories of credentials to verify the user's identity for a login or other transaction. "It's not just a domain password and an ID," said Curcuru. "You've got to have something that maybe only you know or the user only knows along with that ID and password. And it can be tokenized."

Security Logging and Alerts

Hadoop provides the ability to do security logging in real time, which lets you track the behaviors of people who are accessing your data. If analysts do something out of the norm, such as try to get into databases they don't have access to, you can create proactive alerts in real time.

Continuous Penetration Testing

MasterCard conducts continuous penetration testing, which involves testing a system, network, or application to find vulnerabilities that hackers might exploit. And, according to Curcuru, it's definitely something that everyone from auditors and your legal team to internal operations will ask about.

Native Data Encryption

Curcuru encourages organizations to look at how they're handling native data encryption because not all encryption is the same, nor is how you use it. "How the business wants to be able to use the data makes the use of encryption different, as well," he said.

Embedding Security in Metadata

Embedding security in metadata is particularly important because it makes it possible for the security to follow the metadata through the cluster. Curcuru compares it to taking his daughter to Disney World. "She is with me. I don't just don't let her go run willy-nilly." Similarly, embedding security in metadata lets you to see if the data is crossing multiple machines, when it's in motion or at rest, and whether there are multiple copies of it. Embedded security in metadata also makes it easier to explain to an auditor what's going on with your data.

Key Management

Do not store keys with the data. "Key management is important, as well as access to those keys," said Curcuru. "If you have access to the data, you should not have access to the keys." Even you are an organization of two or three people, those two roles and responsibilities should be separated.

Keep a Separate Lake of Anonymized Data

MasterCard keeps a data lake separate from a discovery area, and restricts access to it. Data is anonymized and restricted by using role-based access, even if it's in the sandbox for exploration. For example, an analyst might know that a specific amount of money was spent at a specific store, but she does not know who spent that money.

Top Tips for Securing Data in the Enterprise

As a member of the MasterCard Advisors, Curcuru is often approached by companies seeking guidance on what they should do with their sensitive data. His top tips are as follows:

Assess your maturity level across people, process, and technology
: Curcuru urges people to begin here and be honest with themselves across dimensions. For example, your technology capabilities might be in place, but you might still need a lot of educational awareness. This is where MasterCard's advisors can help, he says.

Review your data strategy
: Machines and technology are not going to ask you ethical questions, said Curcuru, so it's important to get clear about the data you want to expose so that you can "be a good steward of your information."

Identify your education and training needs
: Curcuru encourages continuous knowledge transfer and awareness training.

CHAPTER 3

Big Data Governance: Practicalities and Realities

Steven Totman and Mark Donsky
Cloudera

Kristi Cunningham
Capital One

Nick Curcuru
MasterCard

Ben Harden
CapTech Consulting

According to data governance practitioner and Data Management Group founder John Adler, data governance is simply a mechanism to manage data risk. Because big data involves large amounts of unstructured data, there is no formula to ensure effective data governance in the big data sphere. And without it, companies can face a world of trouble.

In a panel at Strata + Hadoop World New York 2015, experts from Cloudera, Capital One, MasterCard, and CapTech Consulting met to discuss what it takes to implement a big data governance strategy.

The Importance of Big Data Governance

One of the defining differences between traditional data governance and big data governance is that big data governance is more concerned with how data is used and the context of that usage. This means trying to understand not only what the information is, but how to use it, including privacy concerns and ethical issues around the data.

According to Kristi Cunningham, Capital One's vice president of enterprise data management (whose statements are a reflection of her broader experience and views, and not directly those of Capital One), Capital One is focused on applying the same principles to all data, whether it is little data, big data, hosted on-premises or in the cloud. The focus is on understanding the data, knowing where it is, accessing it, and trusting its quality.

According to Mark Donsky, who leads data management solutions at Cloudera, what makes big data governance challenging is that "Everything that makes Hadoop particularly powerful is exactly what makes it difficult to govern." It's not as straightforward as data in a traditional enterprise application because not only is there a lot more data, but lots of different data types, users, and compute engines. And sensitive data might be scattered in multiple formats across the system.

Steve Totman, financial services industry lead for Cloudera's Field Technology Office, added, "Without governance and security, the data lake can very quickly become a data swamp." The only way to prevent it from becoming a swamp is governance and security.

The panel went on to discuss the challenges and solutions they were experiencing in the financial sector. One of the main challenges seems to be finding a balance between empowering users to access and use data while also avoiding a security breach. This leads to the issue of accountability, which invokes questions about where data is coming from, where it's going, how it's going to be used, and by whom.

What Is Driving Big Data Governance?

A lot of issues are at play when it comes to making data both available and safe. Finding a happy medium between these seemingly

contradictory needs is no small feat and is the force behind much of what is driving big data governance today. Some factors driving big data governance today, according to the panel, include the following:

More rigorous regulations
: One of the complications of governing data usage is that various pieces of data might seem innocuous in isolation, but when these pieces are brought together, they become identifiable. Therefore, regulations are largely focused on how data is being used. These regulations vary depending on the data's country of origin.

Individual interest
: Individuals are becoming increasingly interested in how organizations are using their data and are expressing their concerns and limiting their permission on how it can be used.

Data catalogs
: Another challenge for the enterprise is to find a way to let analysts know what data is available to them. Thus, there is a movement toward cataloging data so that it can be shared across the organization.

Security
: One of the big drivers of big data governance is trying to avoid a security breach. Although some believe that creating a central hub where a company's data resides in a single location might make it the most secure, there's also an awareness that having it in one location makes it a target for a breach.

Usage—to empower or to protect?
: One of the questions the panelists say they are grappling with is how to empower users versus constrain them. That is, how to make data available rather than prevent people from doing their jobs? The challenge is protecting people from using data incorrectly.

Accountability
: There's a producer-consumer model in which accountability is a key part of governance so that one knows where the data is sourced. Governance, then, is not just about knowing who is using the data and how they are using it, but also knowing where the data is coming from.

Lineage: Tools, People, and Metadata

One of the biggest concerns in the session as expressed by the panel as well as the audience was the problem of lineage and the effort to solve it using a toolset.

Data lineage is generally defined as a life cycle that tracks data's origins and movements, as well as what happens to it over time. Knowing data's lineage helps teams analyze how their information is being used for specific purposes. For example, if there's a security breach, you can forensically track where the data moves and who touches it along the way.

Totman said figuring out data lineage "is like going to the alley behind a bunch of restaurants and digging through the trash to figure out what people were eating." His point, he said, is that "metadata is not always a natural consequence. Instead, it's often a byproduct."

In terms of tools, the panel agreed that automating lineage is necessary. According to Ben Harden, big data practice lead at CapTech Consulting, the challenge is that "it's difficult to holistically know how data is being used at the HDFS level and to be able to capture the movement of that data." Why? Because much of the data is unstructured; therefore understanding how data is connected and how to track it can be very difficult.

Cunningham acknowledged that lineage is both critical and difficult, but noted that there are currently a number of automated toolsets available without which lineage would be impossible. Capital One has automated 80 percent of lineage in their legacy environment all the way from the source to the consumption of that data. Cunningham noted that an additional challenge they are currently struggling with is figuring out how to make the lineage information consumable and usable.

Donsky noted that Cloudera has the only PCI-certified Hadoop distribution thanks in large part to Cloudera Navigator, which automatically collects lineage for every transformation that takes place inside the Hadoop cluster. This means that anything the end users do will automatically be collected. He added that one of the key aspects of lineage is that it cannot be opt in. "Lineage has to be something that is a turnkey solution; that is, an actual consequence of interacting with the system." For example, if someone creates a

new dataset, governance artifacts are automatically generated and made available.

Curcuru added, "When you talk about data itself, it's not how it's used; it's the attributes you attach to it." The attributes, he explains, let you understand what the data is and how it was put together.

ROI and the Business Case for Big Data Governance

Although ROI depends on your industry, Curcuru suggests the way to measure it in the financial sector is to consider what a big data security breach would cost you; thus, your investment is actually related to cost avoidance. As a result, it's not about ROI but instead about the net present value of what a breach would cost you. He suggests that the business ask itself what it is willing to pay for a security breach, what it is willing to pay to avoid one and then to do some research to find out the average cost per industry to build a business case.

Another approach to building a big data governance business case is to recognize that data is an asset and to therefore treat it as such. This means ensuring that the enterprise understands how to maintain it, govern it, and fund it. Cunningham advises starting small. "Try to use tangible examples of where you've had issues and how this could have been remediated or avoided if you'd done differently."

Although each business case might need to be industry-specific, the following general questions might also aid in making the business case:

- What's the cost of analyses not being performed because the user can't find datasets or trust them?
- What's the cost of performing the wrong ETL workloads and not bringing them over to Hadoop, and potentially spending 30 times the amount by having it in the data warehouse?

"Governance is the foundation of a much broader set of data management capabilities," emphasized Donsky. Whether it's data stewardship or data curation, if people can't find the datasets, he asks,

what's the real benefit of big data? "What's the business cost of not doing governance properly?"

Ownership, Stewardship, and Curation

Totman reminded the audience that the term *steward* does not mean ownership. "It means you're taking care of someone else," he said. "There is an eventual data owner. You must identify them. You must track them and hold them accountable."

According to Curcuru, data governance is led by general counsel at MasterCard. "Our products and services are at the table as well as our operations and technology team."

For many clients, however, IT owns the data governance function. According to Harden, that's a mistake. "The business needs to be the part of the organization that's owning the data," he said. "They understand the data, they should be stewarding that data. IT can provide the services to be able to do data lineage, but the people and process should come from the business side."

The risk management department at Capital One currently owns data governance. However, according to Cunningham, "I don't think there is a right answer. I think it's going to be whoever is the biggest advocate for it and is going to champion it across the organization, because that's what it takes. And then to bring funding with it."

In addition to being responsible for compliance, there's also the issue of stewardship and curation that involves making data available to end users which, according to Donsky, is only successful when there's a good understanding of the business context.

The Future of Data Governance

Panelists predict the future of data governance in the financial sector will revolve around ethics, machine learning, data quality management, and data access.

Ethics

According to Curcuru, the future of data governance involves consent. "It is the stewardship of how the data is used," he said. "Data governance used to always be in IT. It used to be ETL and taxonomy, but the questions being asked in boardrooms these days involve eth-

ics, i.e., 'We have this data, but should we use it?'" These questions are being asked not just due to liability issues, but also because of damage to the brand that can be caused by the misuse of data.

Does data governance need to move from descriptive to prescriptive? Apparently the answer is yes, no, and it depends. Curcuru emphasized that the answer boils down to the question of how one wants to use the data. In other words, how does what you're doing with the data relate back to the business use case, and how was that data originally intended to be used?

Harden agreed that certain controls need to be in place "so that we protect ourselves from ourselves." But, he added, "We also have a set of tools and processes and data available to us that we've never had before. We want to be able to innovate with that and use that data to solve new problems."

The underlying question that Curcuru says he comes back to is "How do I make sure that I can protect and be stewards of that person who is giving me the privilege of handling their data?" Cunningham agrees that big data governances need to be prescriptive as well as flexible.

Donsky suggested that it's possible to set up prescriptive governance up front, but that governance needs to allow an element of agility to enable users to do whatever they want within safe environments while also making sure that the enterprise is honoring its customers' trust.

"The real benefit of Hadoop," says Donksy, "is that you can store petabytes worth of data in the original fidelity." A lot of the data wrangling, data preparation, and data consumption is done with more agility in Hadoop, and the net result, he says, is the foundation of governance. He advises, however, that it's imperative to think about governance at the beginning of a project rather than trying to inject it after the fact.

Machine Learning

Harden suggests that data governance might soon incorporate machine learning techniques. Imagine, he said, being able to take all the information about how data is used in Hadoop, publishing a model of how people are consuming the data, and then building governance capabilities based on that. From there, the system can

learn to automatically understand what kind of data is coming into your cluster.

Automated big data governance is certainly the goal, said Harden. The answer of whether it is possible depends on whether anyone is willing to make the initial investments. There are a number of players and tools that are moving in that direction, he says, including Alation, Collibra, FINRA's open source governance tool Herd, and LinkedIn's open source data discovery and lineage portal, WhereHows.

According to Harden, Alation combines machine learning with human insight to automatically capture information about what the data describes, where the data comes from, who's using it, and how it's used. Collibra focuses on automating data management processes for data stewards, managers, and users. And FINRA's Herd provides a unified data catalog that helps separate compute from storage in the cloud. Harden explained that Herd also helps track lineage, manage clusters, and automate processing jobs. Herd tracks and catalogs data in a data repository accessible via web service APIs. The repository captures audit and data lineage information to fulfill the requirements of data-driven and highly regulated business environments. LinkedIn's WhereHows is a frontend system for capturing data on ingest. According to Harden, it provides a central repository and portal that integrates with data processing environments to extract coarse and fine-grained metadata.

Data Quality Management

Cunningham envisions a world where good data quality management and metadata management are incorporated into enterprise operations. According to Harden, "Quality is a component of governance. Having a program in place will drive higher data quality."

Data Access

Cunningham and Donsky further envision a future in which there are tools that enable and support making data accessible and easier to understand, providing a broader view of serving the whole class of users on top of the foundation that is data governance. Harden adds that "The tools are only as good as the people and processes in place."

CHAPTER 4

The Goal and Architecture of a Customer Event Hub

Arvind Prabhakar
StreamSets

Modern data infrastructures operate on vast volumes of data generated continuously and by independent channels. Enterprises such as consumer banks, which have many such channels, are beginning to implement a single view of customers that can power all points of customer contact.

In a session at Strata + Hadoop World New York 2015, Arvind Prabhakar, CTO at data integration company StreamSets, presented an architectural approach for implementing a customer event hub. He also discussed the key challenges and solutions to overcome them.

What Is a Customer Event Hub?

The Customer Event Hub (CEH) makes it possible for organizations to combine data from disparate sources in order to create a single view of customer information. This centralized information can be used across departments and systems to gain a greater understanding of the customer. "It's the next logical step from what has traditionally been called a 360 degree customer view in the enterprise," said Prabhakar. "But it differs greatly from the 360 degree in that it is bi-directional and allows for an interactive experience for the

customer," he said. The goal is to enhance customer experience and provide targeted, personalized customer service.

360-Degree Customer View versus Customer Event Hub

In the 360-degree customer view, a customer is surrounded by an ever-increasing set of channels of interaction; the view is an augmentation of all of these channels, all the data, all interactions that are happening with one particular customer across all these different channels. The 360 view brings the data together to create a single view for consumption.

The purpose and advantage of having a 360 view, explained Prabhakar, is that it gives you a consistent understanding of the customer and helps you build relevant functionality. The problem is that these various channels are often implemented as silos and therefore they are isolated from one another, which creates a fragmented user experience. The CEH collapses all these channels into a single omnichannel.

"The key difference between a CEH and a 360-degree customer view is the interactivity," said Prabhakar. "A 360-degree view is for consumption by the enterprise, whereas a CEH is a bi-directional channel" that allows for an interactive experience for the customer, as well; it gives the customer a consistent view of the enterprise, which is critical in establishing relationships with your customers.

A Customer Event Hub in Action

For example, describes Prabhakar, a high-value banking customer is trying to transfer money online but cannot do it. As a result, the customer calls the bank's technical support line. Unfortunately, this leads to even greater frustration.

Prabhakar suggests instead that financial institutions consider the possibilities of a call center response application that understands the needs of the caller. If the system knew, for example, what the customer wanted, it could route the caller to a much more immediate answer and result in a much more satisfying experience. "That's the kind of use you can get from a Customer Event Hub," he said.

Key Advantages for Your Business

According to Prabhakar, "All enterprises need to operate a CEH; it's imperative for business agility as well as competitive advantage." Some of the benefits of operating a CEH include:

Enhanced customer service and real-time personalization
: "We all want the services and channels we engage with to be aware of who we are, what we like, and to respond accordingly," said Prabhakar. "But there's often a lag between when we exhibit certain behaviors and when the systems pick them up." A CEH provides a way for enterprises to bridge that gap.

Innovative event-driven applications
: As we're increasingly finding new ways of engaging and working with the social channels, the CEH gives you the capability of building the next-generation infrastructure for new applications.

Security
: Security is enhanced because the CEH lets you track up-to-the-minute activity on all your users and customers interacting with your enterprise.

Increased operational efficiency
: With the CEH, you can eliminate the losses that are the result of a mismanaged application, mismanaged effort, or mismanaged expenses. This lowers the operational costs, which also means you increase the operational efficiency of the enterprise.

Now that we understand the purpose and benefits of CEHs, let's take a look at how to build one.

Architecture of a CEH

At a high level, there are three processes that go into the working of a CEH:

- Capturing and integrating events coming from all channels.
- Normalizing, sanitizing and standardizing events, including addressing regulatory compliance concerns.
- Delivering data for consumption through various feeds and end-consuming applications.

Capturing and Integrating Events

According to Prabhakar, the first phase of enablement involves pulling together or capturing all the interaction channels and then integrating them. This means creating an event consolidation framework, often referred to as the *event fire hose*. This is how you bring the events into the CEH.

What kind of data and events are in the fire hose? Social media, structured and unstructured data, electronic files, binary files, teller notebooks, and so on—in other words, an ever-exploding and always expanding set of formats, both human- and machine-generated.

Naturally, due to the diversity of formats, you're not going to have a uniform level of control over all of this data. "Your capability of running an application across all these channels will be limited by not being natively tied to those channels," said Prabhakar. And this is what the CEH solves.

Sanitizing and Standardizing Events

Next, you need to sanitize and standardize the data. According to Prabhakar, "The goal is to create a consistent understanding of those events for your consuming endpoints." An additional goal, of course, is to meet compliance and regulatory requirements. Ultimately though, standardization makes it possible for you to thread a story together across these channels and events.

Prabhakar explained that standardizing the data and preparing it for consumption primarily involves attaching metadata to every event. This process generally involves threading a handling mechanism around each event so that anybody can identify it, parse it out, and take action around it.

Delivering Data for Consumption

With the CEH, you can deliver data to various feeds and applications. According to Prabhakar, "If you're delivering the data to an HBase cluster, chances are your online web applications could directly reference them and you can have it deliver these events in a very low latency manner." Thus, you can access the data online across your enterprise. Prabhakar explained that you can also send this data into batch or offline processing stores.

In the earlier customer experience example, the call center application magically knew that a valuable customer had been trying to do something on the company's website. It knows because the data has been delivered to another channel to produce a more meaningful user engagement.

Sounds relatively straightforward, doesn't it? If so, why isn't everyone building one?

Drift: The Key Challenge in Implementing a High-Level Architecture

Why aren't CEHs very common yet? Prabhakar explains that, at a high level, it boils down to one word: drift. "Drift is the manifestation of change in the data landscape," he said. Drift can be defined as the accumulation of unanticipated changes that occur in data streams and can corrupt data quality and pipeline reliability. This results in unreliable real-time analysis, which ultimately means that bad data can lead to bad decisions that affect the entire business. Drift can be categorized into three distinct types:

Infrastructure drift
: This refers to the hardware and software and everything related to them such as physical layouts, topologies, data centers, and deployments; all of which are in a constant state of flux.

Structural drift
: Prabhakar explained that flexibility is usually a positive structural attribute; therefore, formats such as JSON are popular in part because they are flexible. The drawback, however, is the very thing that makes them attractive: they can change without notice. This means that if you have events in JSON format, they might change.

Semantic drift
: The most subtle and perhaps most dangerous kind of drift, says Prabhakar, is the semantic drift. Semantic drift refers to data that you're consuming that has either changed its meaning or for which the consuming applications must change their interpretation of it. According to Prabhakar in a 2015 blog post written for elastic.co, "When semantic changes are missed or ignored—as is common—data quality erodes over time with the

accumulation of dropped records, null values, and changing meaning of values."

According to Prabhakar, this infrastructural drift becomes a monumental challenge to overcome in order to be able to build a CEH. Why? Because drift means change and everything from the applications to the data streams are in a constant state of change and evolution.

So, how do you deal with this? One way is to write code, he says, or your own topologies or producers and consumers. Unfortunately, as Prabhakar points out, "Those will get brutally tied to what you know today, and they would not be resilient enough to accommodate the changes that are coming in."

Ingestion Infrastructures to Combat Drift

CEHs act as the "front door" for an event pipeline. Reliable and high-quality data ingestion is a critical component of any analytics pipeline; therefore, what you need is an ingestion infrastructure to address the problem of drift.

One such infrastructure, according to Prabhakar, is the StreamSets Data Collector. In Prabhakar's 2015 blog post for elastic.co, he writes, "StreamSets Data Collector provides an enhanced data ingestion process to ensure that data streaming into Elasticsearch is pristine, and remains so on a continuous basis." It provides an open source Apache-licensed ingestion infrastructure that helps you to build continuously curated ingestion pipelines, and improves upon legacy ETL and hand-coded solutions.

Microsoft Azure also offers an event hub ingress service—Azure Event Hubs—which is a highly scalable data ingress service. Additional ingestion and streaming tools include Flume, Chukwa, Scoop, and others.

About the Author

Jane Roberts is an award-winning technical writer with over 25 years' experience writing documentation, including training materials, marketing collateral, technical manuals, blogs, white papers, case studies, style guides, big data content, and web content.

Jane is also a professional artist.