StreamSets

DataOps

The Authoritative First Edition

WHITE PAPER

Table of Contents

Foreword (by John Schmidt)

DataOps is one of three technology breakthroughs to let people use data as easily as they plan a trip around the world or buy a pair of shoes on the web. Before I explain, let's step back first.

Before the internet, the effort to buy a pair of shoes from home involved waiting for the Sears catalog to be delivered, connecting to the call-center with your rotary dial phone, describing the item you want to buy, and waiting for the package to be delivered. The process to pay for the shoes was even longer since it involved receiving a paper invoice by mail (USPS), sending a paper check to Sears (USPS), moving the paper around internal departments at Sears until it all arrived at Accounts Receivables, and eventually entered the bank's check clearing process (by paper).

In the modern internet age, the order-to-cash process has reduced from several months to a few seconds (basically instantaneous)! There were basically three technology breakthroughs that enabled this dramatic result; the internet protocol, the browser (universal user interface) and online electronic payments. Other developments like the digitization of catalogs, Amazon's 1-click order process, and delivery tracking on cell phones helped, but it was the three inventions of the internet, browser and online payments that make the breakthrough a reality.

I've been wondering for years when IT professionals would invent the same capabilities for finding, delivering and using data as quickly as we can buy a pair of Nike shoes or a PlayStation. What three innovations will emerge to make data easy enough for your mother to use, rather than a pain? Well, we don't yet have all three inventions for data yet, but we have two of them!

The first breakthrough was Big Data, which emerged around 2005. Big Data enabled advanced analytics by making it possible to gather massive amounts of data in any format, without the strict rules of historical database structure, and then save it in a distributed storage/compute platform running on commodity hardware and open source software, at less cost than traditional data warehouses. The first challenge of managing the exponential growth in data scale, variety and speed of has been solved.

The second breakthrough is DataOps which was first mentioned in 2014. It is the alignment of people, automated technologies, and business agility to enable an order of magnitude improvement in the quality and reduced cycle time of data analytics. DataOps expedites the flow of data for effective operations on both traditional and big data, by leveraging self-service capabilities to bypass traditional methods of engineering customized programs. DataOps has demonstrated its capabilities and effectiveness in multiple industries on a global basis to give us the confidence to label this publication as **The Authoritative First Edition of DataOps**.

The third technology breakthrough has not yet arrived, but elements are starting to emerge. The idea is to make it easy to find and assess the value of the wide variety of information in today's complex data landscape using sophisticated tools such as natural language metadata, automatic classification of data in written or spoken form, artificial intelligence to connect with real-world objects or processes, showing data as 3D holograms, and so on. But let's get back to DataOps which is the latest innovation!

DataOps – A Data Management Breakthrough!

DataOps is not named randomly. It builds in the use of DevOps, which is a widely adopted and well-understood practice that accelerates software development by leveraging automation and monitoring to enable agile collaboration across application designers, operations staff and business users. While DataOps does have some capabilities like DevOps, it is a more sophisticated capability, and the comparison downplays its importance and distinction. DataOps is a paradigm shift: It is a fundamental change in the basic concepts and practices of data delivery and completely challenges the usual and accepted way of integrating data.

In today's world, there isn't a single IT organization that can control and engineer all aspects of the data that their enterprise needs. Data is now the connective tissue upon which complex enterprise logic is being built, spanning numerous applications. DataOps enables faster delivery of existing and new data services and products in the face of changing environments, requirements, infrastructures and semantics while preventing data threats. DataOps enables applications to be good citizens of the ecosystem by respecting the implied contracts between them despite unexpected data drift that emerges from changing technologies.

Above, I claimed DataOps is a paradigm shift. For a current example of what I mean, consider what Tesla has realized. Cars have been around for 100 years and car technology has obviously evolved, bit-by-bit, since the Model T. But then Tesla came along. Recently I drove a Tesla from Florida to Toronto. In the 3,000-mile round-trip journey, I didn't buy any gas, the electricity was free, the car accelerated faster than any others on the road, the "engine" was silent, and it drove itself most of the time.

Furthermore, you never need to add oil or antifreeze (it doesn't need either) and it doesn't need any maintenance other than filling the tires with air and windshield wiper fluid. In some cases, you can enhance the performance of your car in your driveway with an online payment; the 60KW Tesla S can be upgraded to a 75KW battery through the car's computer screen. This is possible because the car already had the larger battery installed at the factory and it just needed to be activated by a software license. Tesla is more of a computer with wheels than a car. In short, a paradigm shift.

In the same way that Tesla has flipped the polluting, high-maintenance, manually-controlled automobile into a clean, friction-free driving "computer-mobile", **DataOps transforms stodgy, centralized Business Intelligence "dashboards and reports" into real-time and democratized analytics capability that unlocks the huge potential of all your data. DataOps transforms the traditional approach of designing and building custom data movement software into self-service capabilities that people simply operate.** That looks like a paradigm shift to me.

Why Care About DataOps?

> ***"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."*** **— Albert Einstein**

DataOps is an innovative and powerful capability in the world of information management. It is a rather new practice being adopted by leading organizations. In just a few years it will become the default method for managing data needed by enterprises to support their digital transformation programs and modern data analytics practices.

As per Einstein's quote, to take advantage of this capability, you need to change your thinking. The very needs of analytics, business intelligence and warehousing have changed. Technologies like BI & EDWI solve for descriptive and diagnostic analytics, but predictive and prescriptive analytics have changed the game. Leaders across all industries have realized that their long-term viability and organizational existence in many cases depend on the ability to do a LOT more with data. New roles like the Chief Data Officer are exploring forward-looking analytics, smart products and services and perhaps even productizing data in support of a completely innovative business model. Data is no longer an intangible asset, it is a vital pillar of corporate strategy.

This new world demands agility, both for developing data management practices and for operating them in the real, ever-changing world. DataOps is a new capability that aligns with this need for agility and changes to your methods. Specifically, stop developing brittle data integration software that has to be rebuilt whenever something changes and stop architecting complex data models that take months to create and can't keep up with the rapid changes in today's world.

Instead, DataOps uses smart tools to discover data, detect changes, flow it where it is needed, and monitor operations. In short, focus on the outcome of delivering value (driving the car) rather than building or worrying about perfecting every underlying component (customizing and tuning the engine). The diagram below summarizes why you should care about DataOps.

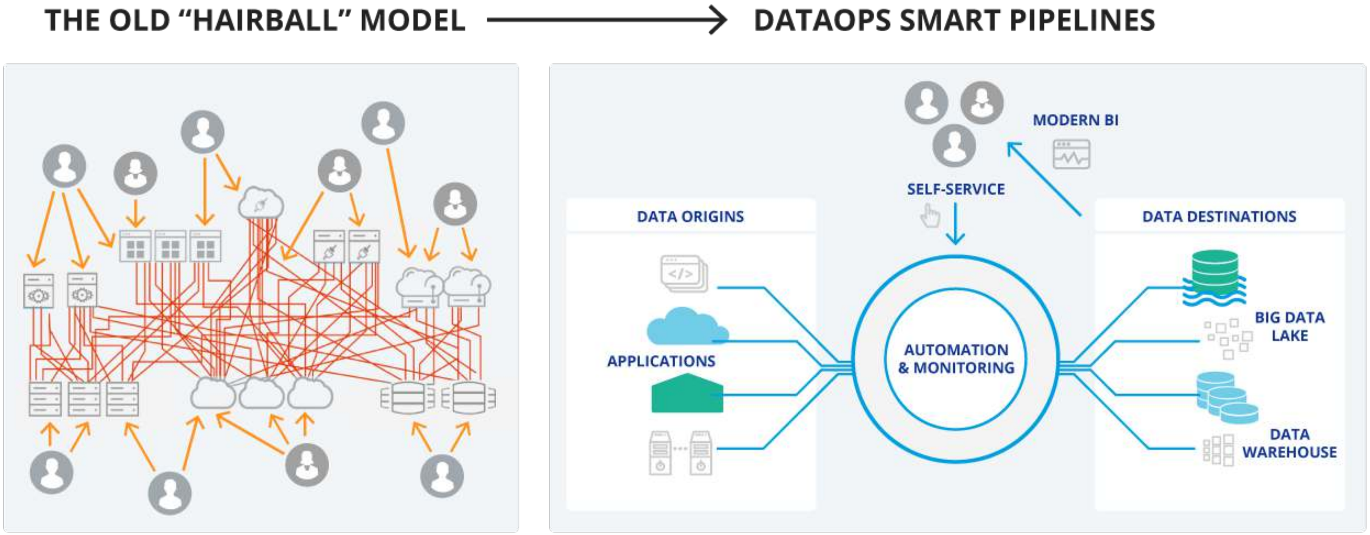


Figure 1. Key DataOps Capabilities vs. Traditional Methods

The image on the left is the Integration Hairball. It shows data flows between a company's application systems — both in-house and in the cloud — through a complex series of integrations, each of which were custom built for a fixed data flow by different people using a variety of technologies.

For example, for a Fortune 500 retailer, my ICC team implemented a metadata repository and a process to capture and maintain the integration points going forward. Within 3 years, we had 5,000 integration points captured in the repository at which point I left the company. Seven years later I spoke with the manager of the team and asked him how many integration points they had in the metadata. The answer was 45,000!

The fact that they had an actual number is positive since it means they had standardized processes to capture integration data facts. If you were to compare this to other companies, you will find that only a small percentage could give you an accurate number. The reality is that they are building integration points ad-hoc as needed to support individual projects, without a master plan or holistic view for all the data flows. The result is high complexity which is consuming a huge portion of their budget. A rough estimate is that 20% or more for typical IT budgets are consumed by the integration hairball. An even larger problem is that the complexity of these custom integrations is adding up to 25% to every effort; about 3 months of a one-year project is wasted due to the hairball complexity.

You don't want an Integration Hairball — although it has been, and still is, the reality in many enterprises. See the upcoming section on DataOps in a Nutshell in this paper to eliminate the Hairball!

A Brief History of Data Evolution

To build on why you need to change your thinking, a brief history of computer technology, how priorities have changed, and why DataOps is critical is justified. Starting with the introduction of electronic computers in the 1950's, there are four areas of computer evolution: 1) hardware, 2) software, 3) the internet and 4) data.

Hardware Era

Hardware became interesting in the 1950's with the transistor to replace vacuum tubes, new magnetic core memory and creation of integrated circuits (IC). Solutions based on mainframe computers evolved to minicomputers, Personal Computers (PC), smartphones and more recently the Internet of Things (IoT) and even quantum processors. Technology continues to advance with no end in sight, but my point is that hardware is basically a commodity. Computer speed and power has increased, and hardware still demands management attention and investments even with dramatic reductions in cost. That said, software, the Internet, cloud and data became the center of attention for innovation and competitive.

Software Era

The early computer hardware used 1st generation (machine level) and 2nd generation (assembly) programming. It wasn't until 3rd generation languages like Fortran, Cobol and Basic caught the attention of programmers in the 1960's that software began to demand efforts from IT organizations. Organizations could now write programs for airline reservations, bank account processing, accounting systems, and replace manual activities like connecting telephone calls and printing sales invoices.

By the 1990's, software companies, like Microsoft, Oracle, IBM, SAP, and eventually Google, were computer leaders and by the 2000's, 80+% of CIO's were promoting the principle of "buy rather than build". Big enterprises ended up with hundreds or thousands of applications (one bank I worked at had 18,000 applications) which kept their programmers busy developing data integrations. The software application fragmentation in turn generated a wave of software vendors specializing in integration tools and ultimately fueling the migration to Cloud software and the need for DataOps.

Internet & Cloud Era

Since the internet, the world has undergone a sea change thanks to the power of the World Wide Web. The internet is now influencing not just for how individuals learn, work and communicate, but also how organizations use and manage computers and information technology. The Cloud Era is lead by vendors who are renting software tools and complete application systems rather than selling them. The ease and exponential growth for business units to acquire solutions drives the need for DataOps due to the dramatic increase in the volume, variety and number of data sources. Traditional methods of extracting, transforming and integrating data can't adapt to the flood of internet-generated data.

Data Era

Data is now leading technological innovation more than hardware, software or the internet. During the earlier eras, data was combined with hardware and software and was basically an integral component of a system to serve a business function. But data now comes from more than just application systems — it also comes from IOT and similar devices (cell phones, digital cameras, health monitoring devices, home security systems, etc.) and a range of cloud applications.

Data started to change from a by-product of business activities to a foundational driver for business innovation during the Information Economy, a concept introduced in The Third Wave by Alvin Toffler in 1980. Examples of organizations that are leveraging this new era are Google, Facebook and other social media companies which treat data as their product/service and the source of financial results. Other data-driven businesses include the world's biggest taxi service, Uber, which has no taxis, and the world's biggest hotel chain, Airbnb, which has no hotels. Being data-driven is not reserved for just digital native companies but is incumbent on players in every industry — e.g. health care, education, manufacturing and government.

Data is creating a "virtual reality" and taking on a life of its own.

DataOps in a Nutshell

What does DataOps do that addresses the emerging era of data dominance, and how does it do it? DataOps expedites the on-boarding of new and uncharted data and flowing the data to effective operations within an enterprise and its partners, customers and stakeholders, all the while preventing data loss and security threats. Unlike traditional point solutions, DataOps uses "smart" capabilities of automation and monitoring including:

1. Self-service tools for professionals to find, move, consolidate and annotate data.

2. Discovery of the technical blueprint of data sources like structured relational stores, semi structured NoSQL stores and unstructured binary data.

3. Automation creation of data processing jobs without needing to specify schema and structure in advance.

4. Expedited handling of data errors and exceptions during data flow processing.

5. Discovery and automatically handling data source changes. For example, if a data source has a new schema with new fields since the last flow, automatically include the changes and continue flowing the data.

DataOps also leveraging extensive monitoring of data-in-motion including capturing operational events, timing and volume, generating reports and statistics that provide global visibility of the entire and interconnected system, and notifying operators of significant events, errors or deviations from the norm. Monitoring is especially important now because the data landscape is more fluid and continues to evolve dynamically. With more and more actors in the modern data supply chain, the data infrastructure is no longer a static plan that can be crafted once and executed — it is now a constantly emerging picture that shifts to align with business imperatives at clock speed.

After 2000 years since the abacus, there is a new dominant driver for the effective use of computers — the data itself. The skills and competencies you used in the past to manage and control the effective use of data are not sufficient. In many situations what you learned or practiced 10 years ago is worse than irrelevant, it is wrong!

In today's world, there isn't a single IT organization that can control and engineer all aspects of the data that their enterprise needs.

Across companies of all sizes, the data changes, systems and applications are evolving, and infrastructure changes with the emergence of new platforms. The net result is that you can't control the actual data because you don't control all the systems and variations of sources and infrastructure. Because of this, you will continually suffer from Data Drift.

> ***Data Drift**: The unpredictable, unannounced and unending mutation of data characteristics caused by the operation, maintenance and modernization of the systems that produce the data.*

In addition to Data Drift, IT professionals, and those in the business who are using self-procured cloud native services because they don't want to deal with IT, should care about DataOps because their biggest systems are not the core application systems that they operated during the prior decades. Rather their biggest system, and therefore the largest cost, is the collection and sharing of data regardless of source (internal enterprise, external partners, customers, public and other) and regardless of technology. The biggest "system" is in fact the continuous and rapid flow of data. And because the holistic flow of data is generally unmanaged, it presents the greatest opportunity for efficiency improvements and business value.

Why do you need a professional DataOps team now? You haven't had DataOps to this point, so what has changed? There are three forces which are pushing data behavior into news concepts that demands new thinking and new capability. These three forces have united to create a "perfect storm" in data management; a combination of events which are not individually dangerous but occurring together produce a disastrous outcome. The three forces that are creating the data storm are outlined below.

1. **The first force** is the four V's of big data (volume, variety, velocity, and veracity) which is like a hurricane making it hard to find, deliver, and access data. Once you feel you have data under control, it changes. The quantity, speed and endless variety of data (unstructured, structured, batch, real-time, streaming, cloud, IOT…) feels like chaos of a hurricane. It all must be rationally defined to be trusted, make sense, be truthful and be protected from people who may damage it or steal it. This is a scale of complexity that didn't exist in earlier years.
2. **The second force** is an unceasing wave of technology change. Data management technology is endlessly being improved to find data in new devices and structures. It needs to be transformed, delivered to where is needed, and cataloged, analyzed, monitored, secured, compressed, archived, and the list goes on and on. In its totality, it feels like a tidal wave — a tsunami of technology.
3. **The third force** is that data is more valuable than ever. Data is now independent from applications and must be managed explicitly in all states and forms in order for the enterprise to operate its business critical requirements. Data is highly valuable since it is no longer just facts about your business or operations, analytics are predictive and prescriptive and, in many respects, data "is" your business. The third force is that data is an asset that needs to be governed and secured and at the same time needs to democratized and used widely. As one CIO said, the biggest cost of large data volumes is not the storage, it is management's time to talk about it.

To sum up this section, data is the new oil — the most valuable asset in many companies. No longer is data confined to a data store like an EDW that gets processed and lifecycle managed by career professionals. It is now the foundation on which modern enterprises build their business-critical logic. DataOps is a new capability that builds on a combination of new technologies, information management practices, and collaboration across functions across your enterprise.

The Astonishing Speed of DataOps Practices

The following scenarios shows how traditional methods for a specific data analytics solution require 8 months, while DataOps delivers the same result in 2 weeks. If this sounds unbelievable, read on.

You are a lead architect tasked with building an architecture for a new data science application that will apply machine learning for predicting customer purchasing patterns and in real-time recommend items that a customer may buy. The Data Scientist is clear that the more data she gets, the better the models and predictions will be.

You learn from the Data Scientist that he expects live feeds (real-time) from the company's ecommerce website, which also is accessed by a mobile app, historic data from several data-marts /warehouses around the organization (customer profile, transaction histories, preferences and habits data, product data, service history data, etc.), and data from external APIs such as such weather and crime statistics near customer location, and finally, promotional data from 3rd parties looking to promote similar products.

After several discussions with the Data Scientist, several days to collect ideas from other architects and scanning the enterprise portfolio of applications and data stores, you build a list of 15 high-priority internal data sources, 5 additional data sources that seem likely, and 10 key websites with public or social media data that is relevant. The data sources run the spectrum from bulk loads of historic data, to real-time data streams, to API calls from outside sources. They contain structured, semi-structured and unstructured data and some of the data include sensitive PII that, if leaked, could be a major liability for the company.

This example below shows how Scenario A consumes 8 months to build the solution while Scenario B delivers the same result in 2 weeks based on a DataOps COE using technology from StreamSets.

Scenario A

The lead architect scans the needs and realizes he needs:

- Developers with the right skills to handle the unique characteristics of all of these unique feeds. They should understand the mechanics of getting data out of core applications, a mobile app, as well as semantics of reading out of large data warehouses.
- Operations Engineers who need to understand how to bring up and maintain the execution engines that run these varied and dynamic workloads; when doing massive batch loads, during peak shopping hours that need a lot of capacity, and at other times. These folks also need to make sure they don't build up unnecessarily large infrastructures that are idling 50% of the time.
- Operations Leaders that are monitoring every aspect of this architecture. They need to know what data is flowing for normal operations and when the flows deviate from the norm.
- Security engineers who need to ensure that no sensitive data flows into areas that are classified as being insecure and that those without the right clearances are unable to see secure data.

The lead architect also has to contend with:

- Developers are not equal — engineers that have built mobile applications that are transmitting data back to the data center don't necessarily understand the semantics of reading large volumes of data from the central warehouse.
- Developers will have to resort to using different tools/execution engines to solve each of these different problems.

- Operations Engineers have to spend a lot of time understanding the technology and operational characteristics of each execution engine and develop different automation strategies for each of them. Not all the tools used will scale elastically; as a result, they will need to size environments for peak load — resulting in potentially large infrastructure expenditures and many more machines to manage — or significant architecture and design innovations to work around capacity variations.
- Operations Engineers will also have to setup monitoring dashboards for every tool they use, and set up alerts unique to each tool — which also consumes architecture time or additional implementation steps.
- Security Engineers will need to examine every data flow point,check to see what PII exists, and define policies to protect the data and communicate violations to developers who need to go enforce those policies. Developers will do this for every tool/dataflow they've developed based on the underlying technology.
- If changes occur to the datasets (i.e. new PII fields show up), if and when the change is detected, both Security Engineers and Developers will scramble to figure out how to protect the new data, and what if anything is at risk since they didn't detect the new pattern initially.

In summary, in Scenario A the lead architect needs to plan at least 5 projects (or more depending on how many different technologies are involved) with each team needing an average of 5 staff. The typical cycle-time of each project is about 3 months, but they can’t start at the same time due to the availability of key staff or needing to work around system or operational constraints, so let's assume that all 5 projects finish at the end of 6 months. At that time the final integration test across the 20 application, capacity validations, corrections and bug-fixes, and production deployment consume another 2 months. Voila –after 8 months the Data Scientist can start his analysis.

Scenario B

When using the DataOps COE based on StreamSets technology, this project gets much simpler:

- **Any Developer**, without specialized skills, can create pipelines that get data from any data warehouse, external APIs or execute directly on big data clusters. When they connect to the source systems, they don't need to understand the underlying schema of the data or the particulars of extracting data from that source. These automatic pipelines allow the data engineers to express their intent in an easy manner and do the heavy lifting automatically. For example, format changes, filtering by value, aggregate functions and more can be applied without the data engineer knowing the full schema of the data, and will continue to work even when the schemas change at runtime as long as the intended requirements are satisfied.

- The **Operations Engineer** can just build up a Kubernetes based infrastructure that they already use for the other microservices based applications the company uses. StreamSets will run its execution engine on this infrastructure and the engineer can choose not to get into too many details of how the execution environment works. System level monitoring they've setup for any and all Kubernetes applications will easily be automated and replicated for the StreamSets execution environment. Overall, they get tremendous efficiencies of scale.

- The **Operations Engineer** also pulls up the StreamSets dashboards and monitors runtime execution of all the dataflows. They get alerted for issues no matter where in the architecture the problem arises. And finally they can monitor the overall topology (living map) and share it with the **Data Architect**. This living map represents the current state of the architecture exactly how the architect originally imagined it.

- The **Security Engineer** sets up centralized policies for handling secure data and ensures that each underlying pipeline (no matter who developed them) protect the data. They can see an audit report of all PII flowing through the environments and prove that they have taken all the steps to protect it.

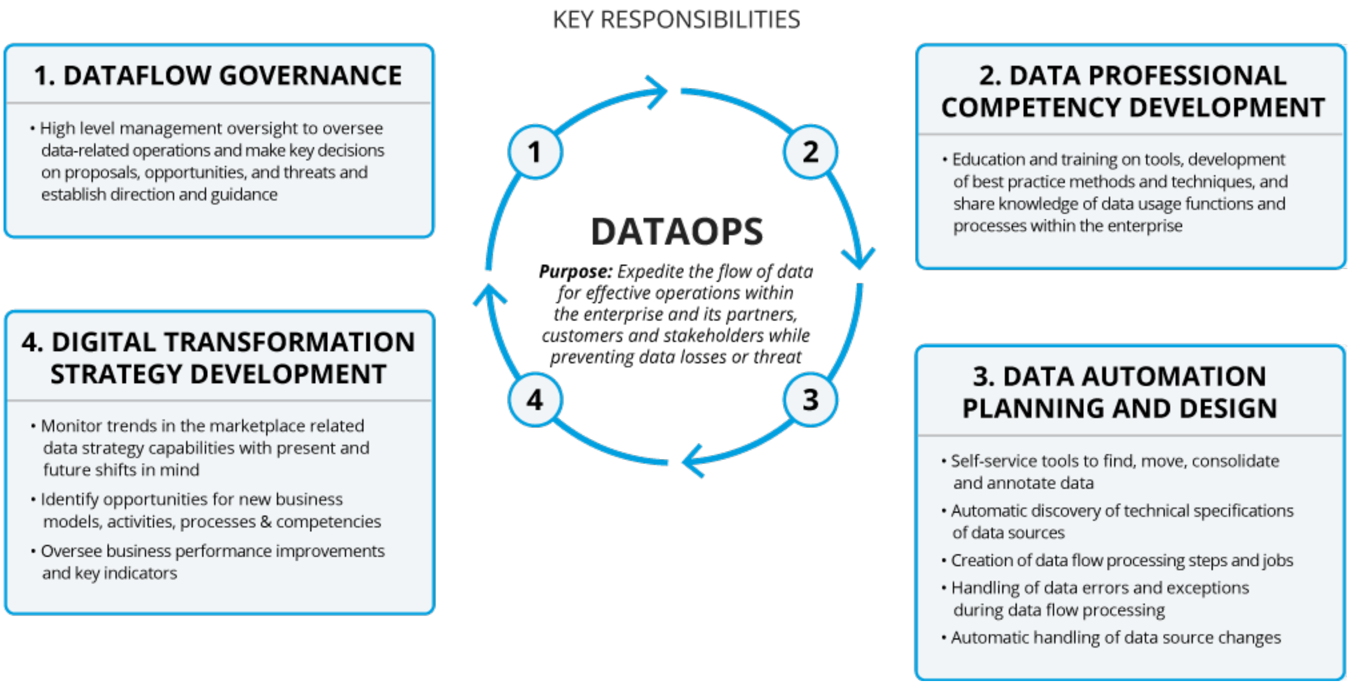
In summary, Scenario B is able to complete the solution with just 5 specific individuals. Because of StreamSets technology, the team is able to support all 20 internal and 10 external data sources and requires less effort due to the level of automation. Voila – within 2 weeks the Data Scientist can start his analysis.

DataOps Playbook Primer

Now that we've demonstrated the power of a DataOps approach, let's build on the earlier description of **DataOps in a Nutshell,** by detailing and defining a) its purpose and responsibilities, b) the services it provides, c) the problems and challenges it resolves and d) the required practices and skills compared to traditional methods.

DataOps Purpose and Responsibilities

In today's world, there isn't a single IT organization that can control and engineer all aspects of the data that their enterprise needs. Data is now the connective tissue that complex enterprise logic is being built on, spanning numerous applications. DataOps delivers flexible data flows that maintain the connectivity between systems in the face of changing environments, requirements, infrastructures and semantics, all while preventing data losses and threats.



KEY RESPONSIBILITIES

1. DATAFLOW GOVERNANCE

- High level management oversight to oversee data-related operations and make key decisions on proposals, opportunities, and threats and establish direction and guidance

2. DATA PROFESSIONAL COMPETENCY DEVELOPMENT

- Education and training on tools, development of best practice methods and techniques, and share knowledge of data usage functions and processes within the enterprise

DATAOPS

Purpose: *Expedite the flow of data for effective operations within the enterprise and its partners, customers and stakeholders while preventing data losses or threat*

4. DIGITAL TRANSFORMATION STRATEGY DEVELOPMENT

- Monitor trends in the marketplace related data strategy capabilities with present and future shifts in mind
- Identify opportunities for new business models, activities, processes & competencies
- Oversee business performance improvements and key indicators

3. DATA AUTOMATION PLANNING AND DESIGN

- Self-service tools to find, move, consolidate and annotate data
- Automatic discovery of technical specifications of data sources
- Creation of data flow processing steps and jobs
- Handling of data errors and exceptions during data flow processing
- Automatic handling of data source changes

Figure 2. DataOps Purpose and Responsibilities

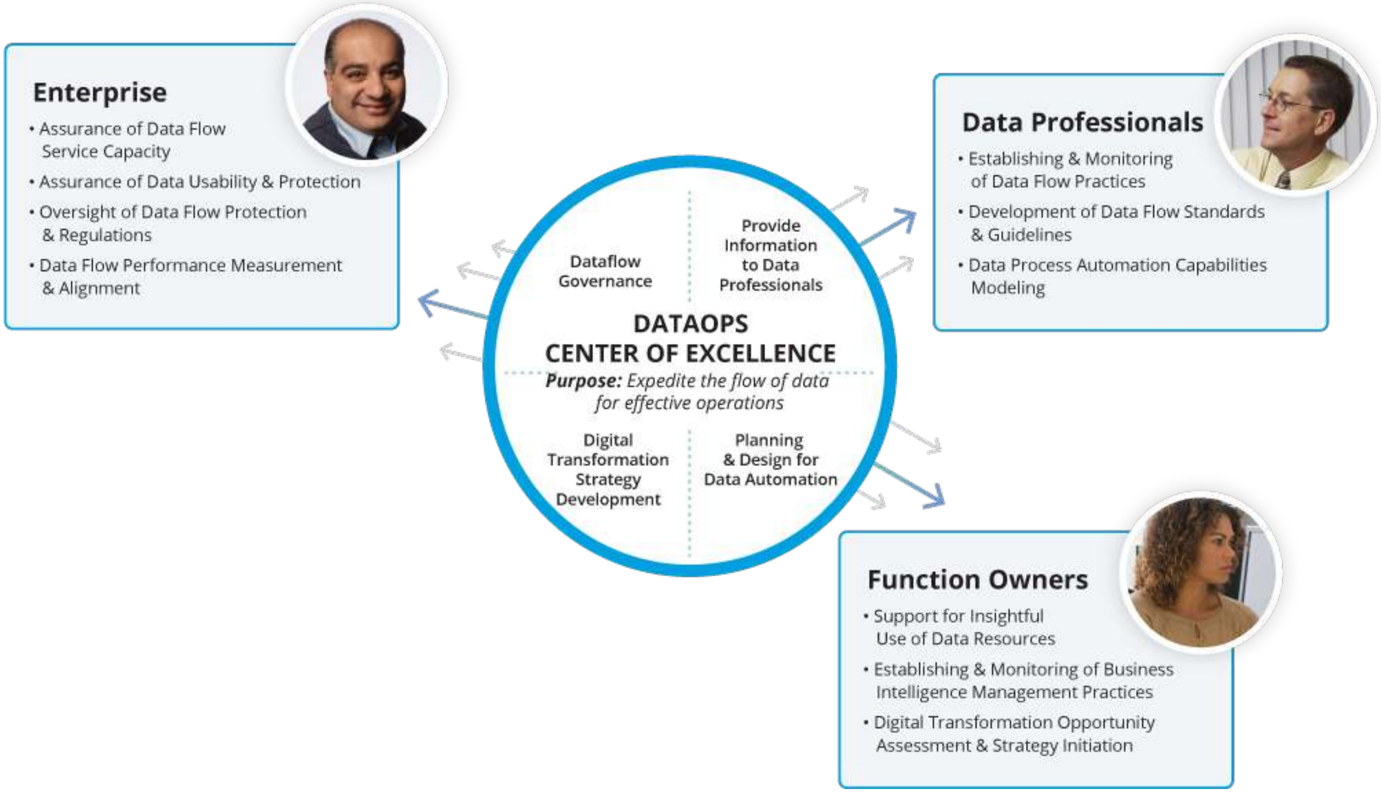
DataOps does not change the application behavior, it changes the dataflows to ensure that all application's implied contracts are respected in the face of drift. It accomplishes this through four key responsible areas as outlined in Figure 2 above:

1. **Operational Governance**: High level management oversight to oversee data-related operations, to make key decisions on proposals, opportunities, and threats, and to establish direction and guidance.
2. **Provide Information to Data Users**: Education and training on tools, development of best practice methods and techniques, and knowledge sharing of data usage functions and processes within the enterprise.
3. **Data Strategy Development**: Support digital transformation programs and modern analytics in several areas:
    - Monitor trends in the marketplace related to data strategy capabilities with present and future shifts in mind,
    - Identify opportunities for data-driven business models, activities, processes and competencies, and
    - Oversee efforts for business performance improvements and developing new key performance indicators.

4. **Planning and Design**: Ensure selection of appropriate tools that are built for DataOps, and that allow for data security by design and not as an afterthought. Disallow tooling that is opaque and doesn't allow for a high degree of automation and monitoring, Enhance acquired tools to support automation and self-service capabilities for data users.

DataOps Services

In addition to its responsibilities, we can describe the DataOps Center of Excellence as shown in Figure 3 in terms of the services it provides to a) data professionals, b) function owners for business units or operational teams, and c) the enterprise from a holistic perspective.



Enterprise
• Assurance of Data Flow Service Capacity
• Assurance of Data Usability & Protection
• Oversight of Data Flow Protection & Regulations
• Data Flow Performance Measurement & Alignment
Dataflow Governance
Provide Information to Data Professionals
DATAOPS CENTER OF EXCELLENCE
Purpose: Expedite the flow of data for effective operations
Digital Transformation Strategy Development
Planning & Design for Data Automation
Data Professionals
• Establishing & Monitoring of Data Flow Practices
• Development of Data Flow Standards & Guidelines
• Data Process Automation Capabilities Modeling
Function Owners
• Support for Insightful Use of Data Resources
• Establishing & Monitoring of Business Intelligence Management Practices
• Digital Transformation Opportunity Assessment & Strategy Initiation

Figure 3. DataOps Services and Supported Stakeholders

The primary DataOps services to support the enterprise are:

- **Assurance of Data Flow Service Capacity**: Ensures data movement and exchanges between sources and destinations, effective operations of data pipelines and control systems, and the ability to support data volumes peaks and preparation for future demands.

- **Assurance of Data Usability and Protection**: Ensures data usability and approved access for enterprise stakeholders and for establishing the relationship between data owners and data consumers and analysts (e.g. data scientists).

- **Oversight of Data Flow Protection and Regulations**: Guidance for measuring, analyzing and reporting data delivery operations, quality, security, and compliance rules.

- **Data Flow Performance Measurement and Alignment**: Monitor the data flow operations and measure volume, speed and trends including usage and availability across enterprise users.

The primary DataOps services to support business functions and their leaders are:

- **Support for the Productive Use of Data Resources**: Ensures capturing, collecting, and publishing data about data and data processes to support data analysis and strategic planning functions. Use cases include business definitions, data source and owner information, and metadata related to data capture, update, change history, and distribution.

- **Establishing & Monitoring of Data Analytics Management Practices**: Implementation and oversight of BI policy, standards and procedures, resolving analytics and reporting issues, discovering new meaningful patterns in data, communicating business insights to relevant planning functions, oversight of sandbox environments, and initiating production BI capabilities.

- **Digital Transformation Opportunity Assessment & Strategy** Initiation: Enable Digital Transformation with a formalized framework, best practices, and modeling tools to identify improvement opportunities, define business priorities, sequence migration phases, develop roadmaps, and create business cases.

The primary DataOps services to support data professionals in IT and business functions are:

- **Establishing and Monitoring of Data Flow Practices**: Develop procedures and best practices for operational roles, rules, guidelines and processes to ensure that data access, distribution and quality meets performance requirements. Share information and train data users for the appropriate use of data flow practices.

- **Development of Data Flow Standards and Guidelines**: Ensure that DataOps strategies and policies are followed by developing and maintaining related framework, methodology, tools and standards.

- **Data Process Automation Capabilities Modeling**: Support the development of self-service and automated data discovery, delivery and quality by defining operational and technology architectural models.

Challenges and Problems that DataOps Resolves

Effective use of data is indispensable to your business; you should have an advanced and mature professional data management practice.. This section explains some of the causes of potential data issues, why traditional methods don't address them effectively, and how DataOps is different.

In short, the nightmare for data professionals is the inability to keep up with the dramatic increase in data complexity, variety and scale. There are three specific categories which keep data professionals awake at night: data sprawl, data drift and data urgency.

Data sprawl is the dramatic variety of data sources and their volume. Consider systems such as mobile interactions, sensor logs and web clickstreams. The data that those systems create changes constantly as the owners adopt updates or even re-platforms those systems. Modern enterprises experience new data constantly in different formats, from various technologies and new locations.

Data drift is the unpredictable, unannounced and unending mutation of data characteristics caused by the operation, maintenance and modernization of the systems that produce the data. It is the impact of an increased rate of change across an increasingly complex data architecture. There are three forms of data drift: structural, semantic and infrastructure.

- **Structural drift** occurs when the data schema changes at the source, such as application or database fields being added, deleted, re-ordered, or the data type being changed.

- **Semantic drift** occurs when the meaning of the data changes even if the structure hasn't. Consider the evolution from IPv4 versus IPv6. This is a common occurrence for applications that are producing log data for analysis of customer behavior, personalization recommendations, and so on

- **Infrastructure drift** occurs when changes to the underlying software or systems create incompatibilities. This includes moving in-house applications to the cloud, mainframe apps to client-server systems, or moving from a traditional database to big data solutions, such as Hadoop, NoSQL, Hive and Sqoop to mention just a few of the many technology options.

Data Urgency is the compression of analytics timeframes as data is used to quickly make real-time operational decisions. Examples include Uber ride monitoring, fraud detection for financial services, next best offers in e-commerce, or real-time notification of customer issues. In addition, the Internet of Things (IOT) is creating ever increasing sources of transactions that need immediate attention: doctors are demanding input from medical sensors connected to their patients, companies are investing in automating sensors on equipment, trucks and buildings to monitor their status, potential failure events, and on and on.

You might be thinking "*The issues of data sprawl, drift and urgency aren't new and have been around for years. Why do we need DataOps now and what is wrong with traditional processes?*"

You would be correct in thinking that these aren't new issues, but their increased frequency and magnitude are new requirements! In past years, these issues were generally isolated and could be dealt with using standard exceptions methods. Let's look at how data service incidents were resolved in the past (and are still used today in many enterprises).

1. **First,** an exception event occurs. Maybe it is flagged by a computer mainline job that "abends" (ends with an error) and is noticed by a data center operator or maybe a monitoring job sends out an event pager. Alternatively, a business owner may see odd results in the monthly sales performance report or a customer calls the service desk to complain about a slow website. In any event, someone in the incident management team or help desk is notified of the exception.

2. **Second,** the help desk gathers as much information as they can and makes an assessment of the "severity level". For a low severity, they send an email to the application owner and ask them to look into it when they can. If it's a Severity 1, they take a more dramatic action and initiate the "Severity 1 Group Page" which notifies dozens of staff to organize a conference call.

3. **Third,** the staff on the conference call work to a) understand the current issue and its impact, b) analyze the problem and determine the root cause and c) figure out how to correct the situation and return to normal operations. Dozens of staff are involved because it's not clear up front what the precise problem or correction is, so anyone that "might" be able to help is required to attend. In any event, usually the team is able to return functional operations and data flows are restarted. But the incident recovery often does not result in a permanent solution and the company needs to know the root cause and how to avoid future recurrences.

4. **Fourth,** a postmortem process is initiated to fully understand the root cause and how to avoid it in the future. It could be several weeks to understand what happened, followed by a group review meeting by multiple SMEs and managers, followed by a formal report and recommendations for division leaders, internal audit, or senior management. Hopefully the defined recommendations are approved and a permanent resolution is implemented.

It is clear that this process is painful, expensive, and simply won't work in today's reality of increasing data complexity, data variety and data scale. We need a better solution built on the assumption that data sprawl, data drift and data urgency are the new normal.

It’s worth noting that with DataOps you would reduce the number of incidents dramatically not because the issue frequency slows down, but because an infrastructure that executes with DataOps can automatically handle a vast majority of the typical changes and emerging characteristics of systems. These same changes could cause devastating damage if not handled via DataOps and could lead to data corruption, loss, and SLA breaches that could create cascading failures downstream.

Practices & Skills Compared to Traditional Methods

DataOps sounds compelling, but what exactly are the procedures and techniques for applying the practice? Figure 4 below compares DataOps methods with traditional means, organized by data needs. The methods listed below could leverage DataOps products (described below), or could be developed from open-source or commonly available technologies, or simply are processes that could be applied by skilled managers or subject matter experts (SMEs).

Data Usage Needs	Traditional Data Integration & Execution Requirements	DataOps Practices and Advantages
Design Solutions	Requires exact schema-to-schema mapping specifications.	Requires minimal schema specifications to accommodate a range of change (data drift) adoption without redesign.
Infrastructure Dependency	Integration logic is tightly coupled with the underlying infrastructure, therefore infrastructure cannot be changed without redesign.	Provides a decoupled "run-anywhere" semantic that is infrastructure-agnostic and allows the same integrations to operate in diverse environments. Provides portability and freedom from technology lock-in.
Data Delivery Speed	Requires applications to work with the execution modes offered by the data integration system. Batch and streaming require different technical solutions.	Enables a "run at any speed" semantic that allows integrations to adjust to application requirements. Provides portability across batch, micro-batch, streaming, and real time operations modes.
Data Map	Relies on manual design-time documentation to produce a visual map for individual integration (not complete maps for large enterprises).	Self-documented through metadata to enable live views of data interconnectedness within the enterprise, with capabilities that can track evolution over a period of time.
Dataflow Consistency & Variation	Requires consistently matched metadata scenarios. Data variations are hard-coded for specific specifications.	Intent-oriented design that applies broadly whenever the same type of data is in motion. Supports reuse across scenarios even when lower-level semantics are different.

Data Usage Needs	**Traditional Data Integration & Execution Requirements**	**DataOps Practices and Advantages**
Continuous Integration & Deployment	Provides black box integrations that must be designed to specific requirements. Requires manual and explicit handoffs for deployments in production for dataflow integrations.	Whitebox integrations that protect data corruption, loss or failure, and enable operation in the face of changing requirements. Provide criteria based the automatic promotion of integrations to ensure agility of operations.
Continuous Data Privacy	Requires data privacy and protection logic to be designed for each pipeline. Demands complex intertwined logic between orthogonal concerns like integration and security.	Tiered data privacy rules applied to integrations automatically based on governance policies empowers information security teams without requiring awareness of lower-level details. Decouples security and integration concerns.
Entity Centric Services	Active support for entity-centric services such as live catalog and glossary management through custom-developed software.	Operates on the basis that all integrations are directly or indirectly related to the entities within the business domain and therefore works to keep the reference systems like catalog and glossary in sync automatically. Provides up to the minute visibility into how and where data infrastructure is aligned with the business processes it supports.
Zero Downtime Evolution	Requires taking integrations offline while making changes to logic or underlying systems. Treats every integration as a black box during runtime and does not enable the rapid evolution of integration in the face of changing requirements.	Enables the rapid evolution of integrations while ensuring the continuity and integrity of data flowing through them even as change occurs. The capability to make this happen is an approach to "instrument everything" towards data integration using dataflow sensors. In such scenarios, every change is captured and evaluated at different context levels (that of the immediate integration, the application being integrated, the topology that this application is a part of, etc.). Different context levels result in different insights that spotlight the emergent designs within the infrastructure, thus enabling rapid iterations without the need for expensive redesigns.
Bad Record Management	Handle bad records as exceptional cases with little or no capabilities to understand, anticipate and reprocess them. As errors increase, data flows stop.	Treat bad records as a natural consequence of operations and provide support for the direct handling and reprocessing of such data, thereby ensuring that there is minimal data corruption or loss along the way.
Data Metrics and Alerts	Provide limited design time support for profiling data and consequently visibility into how data is changing and evolving at runtime.	Provide fine-grained visibility into how the data evolves at runtime based on constant runtime data profiling, ensuring the operational integrity.

Data Usage Needs	**Traditional Data Integration & Execution Requirements**	**DataOps Practices and Advantages**
Data Confidence	Requires manual subject-matter-experts and knowledge across a group of people to ensure data integrity and correctness.	Provides direct tie-in and automation to ensure the integrity and correctness of data that flows through the infrastructure at all times. The underlying technology is based on ML (Machine-learning). Use ML to process large amounts of sensor data and thus identify the outcomes that support data confidence.
Data Marketplace	Traditional DI does not provide a systematic means to further the sharing and access to data in the enterprise. Different roles such as designers, architects or data analysts generally interact with others within the same function across rather than systematically across functions.	DataOps is based on the principle that data powers business processes at the core. Sharing and enablement provides access to data in all parts of the enterprise while being compliant with security, privacy and other requirements. Consequently, DataOps lends itself towards better sharing and secure sharing of data across the enterprise.

Figure 4. DataOps Methods and Advantages Compared to Traditional Methods

Starting the DataOps Center of Excellence

There are two main strategies for implementing DataOps: bottom-up evolution or top-down transformation. The quickest way to start is bottom-up by data professionals simply applying new methods incrementally. The quickest way to finish a mature practice that is embedded company-wide is top-down by following a transformation roadmap with senior management support.

A common need for both strategies is to leverage change agents. DataOps will change how the company ingests, propagates and uses data so it is critical to have one or more change agents who:

- Are voracious learners
- Do not wait for orders to take action on new ideas
- Express excitement freely concerning new ideas and change
- Demonstrate a sense of urgency to capitalize on innovations and opportunities
- Challenge the status quo
- Transcend silos to achieve enterprise results
- Skillfully influence peers and colleagues to promote and sell ideas
- Display personal courage by taking a stand on controversial and challenging changes

Start DataOps quickly by finding a few change agents to begin applying practices. They may be new employees or long-term established staff with a network of relationships and the ability to get things done across the enterprise. You have some change agents in your organization today, so find them; maybe you are one of them. Collaborate with them, start applying simple Agile or Lean methods such as flow of value, waste elimination and fail fast, and evolve the capability as you have successes.

The top-down transformation of DataOps leverages same of the same methods as bottom-up, but with more structure and formality. The steps are:

1. Identify an Executive Sponsor
2. Define the Vision and Charter and Inform Stakeholders
3. Develop a Roadmap to Map the Journey
4. Execute and Advertise the COE
5. Periodically Assess and Renew Plan
6. Reinforce the DataOps Culture

First, you need support from an executive sponsor since you will run into resistance from team processes, policy changes, funding needs and other roadblocks. It's easier, if not essential, to have a senior director, VP or C-level officer that you can work with to support the DataOps vision.

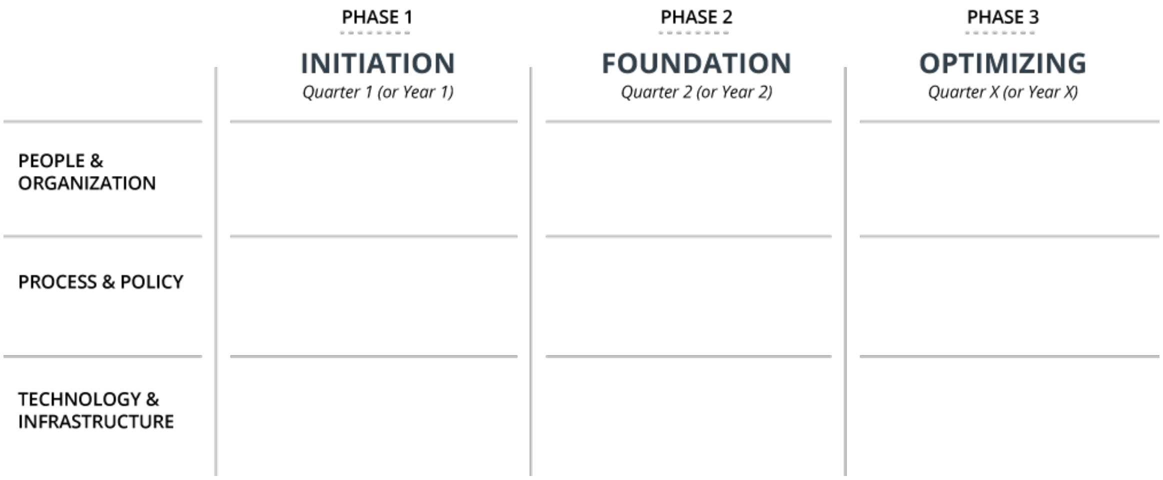
Second, formally define and document your vision and charter. One way to start is to simply ask your executive sponsor *"What keeps you up at night?"* and *"If our DataOps turns out to be successful, what would that look like from your perspective? How would you measure the results or talk about the outcomes?"* You should also review your company's annual report and incorporate priorities of the CEO or Chairman.

Developing the DataOps Roadmap and Gaining Executive Support

The third step is to launch your DataOps blueprint which consists of three elements:

- **Strategic Roadmap** is a "checklist" of milestones or outcomes arranged in multiple tracks and phases. Find specific leaders/managers to assume responsibility for the tracks and phases, but the strategic roadmap does not specify "how" the milestones are to be accomplished, only "what" the outcomes would be.
- **Program Roadmap** is to define and gain approval for specific initiates including business justification, costs, change drivers, timelines, current/future state models, risks and constraints. This map adds concrete initiates to the strategic roadmap and plays them out in phases.

- **Project Plan** details efforts for a program initiate showing a detailed breakdown of activities, resources, dependencies, costs, deliverables and other elements defined by the Project Management Body of Knowledge (PMBOK). Start your DataOps Transformation with at least the first project clearly defined.



	PHASE 1 INITIATION *Quarter 1 (or Year 1)*	PHASE 2 FOUNDATION *Quarter 2 (or Year 2)*	PHASE 3 OPTIMIZING *Quarter X (or Year X)*
PEOPLE & ORGANIZATION			
PROCESS & POLICY			
TECHNOLOGY & INFRASTRUCTURE			

Figure 5. A DataOps Strategic Roadmap Template

Figure 5 shows a Strategic Roadmap template intended to represent your entire plan in one page. The idea is to document bullet-oriented milestones or outcomes in each of the 9 boxes across the 3 phases and 3 tracks. There will be more time in later months to add additional details, but at the beginning it is important to have a holistic one-page plan that contains the critical points and to articulate the plan and **reach agreement on the roadmap across all key stakeholders**.

The main value of the Strategic Roadmap is to quickly develop the plan; a few days or one week may be sufficient. Roadmaps may be created in a 4-hour workshop. The components of the roadmap are:

- Program Owner: The person responsible for ensuring that the roadmap details are completed
- Program Sponsor: Senior staff supporting the resource needs
- Roadmap milestones ordered by tracks and phases:
    - **Tracks** defining at least three dimensions:
        - People and Organization
        - Process and Policy
        - Technology and Infrastructure
    - **Phases**
        - For a 3-year roadmap – Phases are Year 1, 2 and 3
        - For a 1-year roadmap – Phases are Quarter 1, 2, 3 and 4
        - For a 3-month roadmap – Phases are Month 1, 2 and 3

The complete blueprint may take a few months to fully define and gain agreement across key stakeholders. The Transformation strategy will take longer than the Evolution approach to get started, but it will establish a foundation to scale the DataOps capability and sustain it to become adopted company-wide.

The next steps for transformation are ongoing activities to grow and sustain your DataOps capability:

- **Execute and Advertise/Market**: The Program Owners develop detailed project plans, secure the resources, and then make it happen. Keep stakeholders and the data community up to date with progress. Make specific efforts to highlight successes and measurable outcomes.
- **Periodically Assess and Renew your Plan**: Do a periodic review with the Executive Sponsor and leadership team; depending the pace and speed of your Transformation, should do a review every month, quarter or year. At least once a year you should review and possibly reshape the plan, especially if company strategies have evolved or significant technologies or other best practices are now possible.
- **Reinforce the DataOps Culture**: This is an ongoing process and is a deep enough topic that we will expand it a subsequent paper.

In summary, start with your DataOps with minimum investment. A large investment may happen as the team or capability grows and becomes adopted across the enterprise, but when that happens, the payback will be obvious, and the investment will be justified.

Overcoming DataOps Resistance

It would be a mistake to underestimate the difficulty in leading DataOps change in a large enterprise. Some of the really difficult challenges include:

- The “not invented here” syndrome and similar behaviors of people/teams simply resisting what they have been doing for years,
- Project funding by fine-grained silos that don’t have the money for and aren’t motivated to solve the “big picture”,
- Tactical short-term investment emphasis that doesn’t appear to leave any room for strategic infrastructure investments,
- Concessions and trade-offs needed by tactical pressures that get in the way of “the right thing” in the long term,
- Autonomous operating groups in distributed geographies that will not accept guidance from a central group, and
- Fear of change and vested interests in the status quo.

The term “challenges” may be too polite when referring to the above list; these seem a lot more like immovable barriers. As insurmountable as these hurdles may appear to be, they are not unique to a DataOps implementation and have been conquered in the past. While there is no simple “silver bullet” solution, there are a number of key concepts which have been proven over and over to be effective. Here are seven of the best:

1. **Think strategically and act tactically:** Have a clear vision of the future but be prepared to get there one step at a time. It is good to keep in mind that *there is no end-state*. In other words, things are always changing. For example, if you miss a window of opportunity to establish a new architectural standard on the latest project, don’t worry. Another project will come along. If you are in it for the long run, individual projects, even big ones, are just blips on the radar screen.

2. **Credibility through delivery:** In order to be perceived as a leader by others in the enterprise, you need their trust and respect. It is not just about being open, honest and trustworthy; but do people trust that you will actually get the job done? In the final analysis it comes down to your ability to execute. To organize your work, set appropriate priorities, assign the appropriate resources to the task and maintain good communications with your customers. Above all, keep your promises.

3. **Sidestep resource issues:** In this global economy of outsourcing, offshoring and contracting, there should always be able to find the resources to get a particular job done. If you want to create a reputation as a “can do” customer service-oriented team, there should never be a time when you need to say No to a service request due to lack of resources (there may be other reasons to say No).

4. **Choose your battles:** Whenever you have the choice between a carrot and stick approach, always use the carrot. You can, and should, carry a stick in terms of having the support of senior executives for any mandated processes or standards, but you should use the power as infrequently as possible. Sometimes this might even mean deviating from enterprise standards. One way to help you choose your battles is try this exercise. Write down your DataOps principles on a piece of paper and cross out one at a time starting with the ones that you would be willing to compromise if pushed into a corner until you only have one left. That is the principle that you should use your stick for.

5. **Take out the garbage:** Accept responsibility for work that no one else wants. An interesting lesson learned is that many of the jobs that no one really wants are those that don’t serve a specific function but end up being ideal data initiatives. Sometimes these also end up being really difficult challenges, but generally they are recognized by management as such, which opens the door to asking for top-level support when needed.

6. **Leverage knowledge:** There is a well-known truism that states that “knowledge is power”. In a DataOps team you are ideally positioned to talk with just about anyone in the organization. By asking a lot of questions and being a good listener, you can gain a lot of knowledge about the organization that narrowly focused project teams or groups don’t have. This knowledge can come in very handy in terms of which projects are getting approved and where you shouldn’t spend your time, where next year’s budget will land, which groups are hiring and which aren’t, etc.

7. **Take it outside:** Another aspect of leadership is active participation in the broader community; specifically, participation in standards bodies and professional organizations. The external activities can be useful for both getting new ideas and insights, and for polishing your own ideas through discussion and debate with others. In the end, these activities can make you stronger as an individual which can help you play a leadership role inside your enterprise.

DataOps: Name It and Claim It

At some point in the roadmap execution, you will formalize the team and align it structurally with the organization. The general suggestion is to develop a name or “brand” of the DataOps team early, but keep it focused on specific needs and responsibilities. In short, don’t start with a wide scope; instead let the size and scope of responsibilities grow as the team demonstrates success and the demand for their services increase.

In addition to identifying the initial team, enterprises should also create space for individuals and change agents. They will be doing disruptive and hard work; the core team is the leading edge of the “spear” that the rest of the company can learn from and adopt as the practice executes.

In terms of the DataOps name or brand, you may be able to build on a competency center or center of excellence that already exists in the enterprise; such as a Security Competency Center, Integration COE, BI COE, Network Operation Centers (NOC), and so on. One option is to structure it as the Chief Data Office (CDO) or Data Operations Center (DOC) that serves as a center of excellence across the enterprise as it is influencing and directing the practice being developed across the enterprise.

Case Studies

Case Study 1: DataOps Capability at GLOBEX, an Online Global Marketplace

This case study is based on an online global marketplace which we renamed as GLOBEX to protect the identity of the actual company.

Executive Summary

The legacy approach to data sharing facing GLOBEX was ad hoc, slow and inefficient, leading to a tangled web of data pipelines. The goal of the effort was to increase efficiency, lower costs, and make actionable data easily accessible for sophisticated real-time analytics.

To ingest their large and varied data sources to the data lake, they initially planned to build a custom solution. They soon came to realize that a hand-coded approach was unsustainable and was preventing them from meeting the goals of the data lake. New data sources were taking weeks to onboard and the backlog of jobs was growing to the point where it was unsustainable. In addition, the reality of changing schemas, known as data drift, was leading to endless maintenance of sources already feeding the lake.

To create a sustainable solution to the data lake ingestion problem, the data team launched an exhaustive across-the-board survey of available ingestion solutions. They chose StreamSets to create a real-time self-service data exchange that ingests data from all their sources, including social media, SMTP servers, JSON, XML, unstructured and binary data, into a new unified “data lake”, available to all business users in real time.

After implementing their DataOps capabilities based on the StreamSets solution, GLOBEX was able to fulfill over a year of backlogged data ingestion requests in less than a month. They now use an automated form-based process to make new data sources and streams universally available as soon as a request arrives; the need for backlog was eliminated. In addition, data drift is automatically handled with schema changes propagated into the data lake without manual intervention. Business partners can now leverage all of the company’s data to both innovate and improve operational effectiveness.

Challenges

The business challenge for GLOBEX was to speed innovation by unifying access to data from its separate divisions and trading partners. Each entity has its own systems, schemas, data centers and staff. Each independently ingests and processes multiple real-time data streams and databases. Their starting point was a “tangled web” of logical data flows between business units which I characterize as an Integration Hairball.

Data movement between companies was historically custom-created and managed by valuable engineering talent, usually to meet a very specific need. The aggregate result of these ad hoc pipelines was a spider's web of data movement pipelines that was not only inefficient to build but also difficult to track, expensive to maintain and next to impossible to govern. Data was being moved to numerous locations and in some cases making unintentional round trips. In one example of data flowing hither and yon within the GLOBEX ecosystem, OSCORP gave data to ACME who then gave it back to OSCORP who then moved it to Nakatomi Trading who then gave it back to ACME who then gave it over to Tyrell to market the product.

The customized pipelines also could not handle the data drift or schema evolution inherent in the company's data sources due to the constantly changing nature of its businesses and its decentralized structure. Any kind of data drift in the upstream schema would cause pipelines to fail and required urgent maintenance work to get pipelines back online and maintain a functional data operation.

Also the complexity of the data sets was daunting. Across the company ecosystem, there were nine or more source types, RDBMSs like Oracle Netezza, MySQL, DB2 as well as APIs, flat files, Kafka topics and more. And every entity has more than one of those implementations. And each source had multiple schemas. For example, at ACME, their Oracle RDBMS contains one schema that has roughly 1,600 tables. It was clear that the traditional approach — hand coding — was not going to scale gracefully.

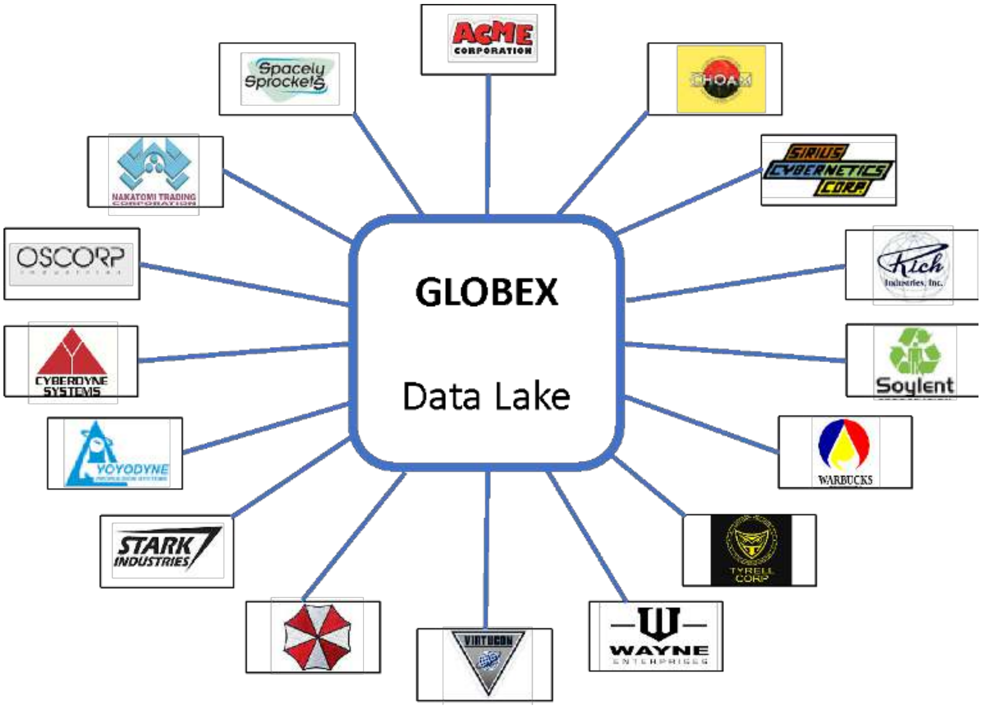


Figure 6. Exchanging Data at Scale with Independent data Silos

In short, data was not democratized, not real time, not consumption-ready and not reliable due to data drift. GLOBEX could not integrate or leverage the data power that existed in its 25 walled-off data silos.

Solution

To achieve the ambitious goal of unifying all data systems, GLOBEX decided to create a central data lake which would contain everything so any team member from any company could access and analyze timely and trustworthy data from any of its peer companies.

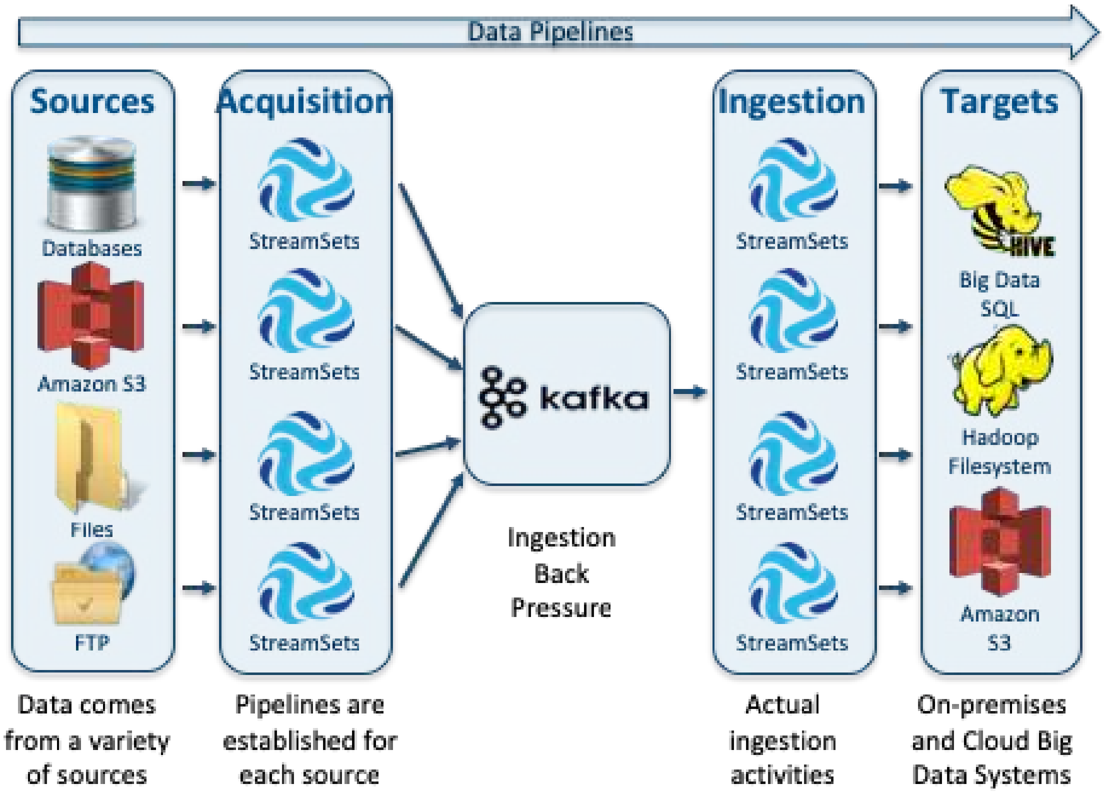


Figure 7. The GLOBEX DataOps StreamSets Architecture

The GLOBEX ingestion architecture leveraged StreamSets DataOps technology to create a heavily-automated self-service ingestion platform. A key architectural concept was the decoupling of data acquisition from data ingestion.

Pipelines were created to move data from sources such as relational databases, files, FTP servers and cloud environments into a shared-service, on-premise/cloud hybrid big data system. A load-balanced tier of pipelines fronts an Apache Kafka message bus to ensure scalability and security. It also handles both impedance mismatches from different sources and provides back pressure to the dataflows. A separate set of pipelines was used to consume data out of Kafka and into various destinations including an on-premises HDFS/Hive warehouse and Amazon S3.

Key attributes of the DataOps technology solution were:

- Acquisition pipelines are decoupled from ingestion pipelines with Kafka acting as the interceding message bus to handle.
- Acquisition pipelines are source-specific, retrieving data from a particular source, but all acquisition pipelines send data to the same destination.
- Ingestion pipelines are completely generic and can be used to route data coming from any/multiple acquisition pipelines.
- New acquisition pipelines can be brought online without any changes to the ingest pipeline tier.
- Data moving between data centers is encrypted in transit using TLS.
- Errors are handled dynamically.
- Data standards are applied to the data in motion including compression, file formats, partitioning schemes, row-level watermarks and time stamping.
- Auto-creation and ongoing management of Hive schemas during data flows.
    - New tables and partitions created automatically
    - Upstream schema changes/drift are synchronized with the downstream Hive warehouse
- Ingestion for new sources is completely automated and any field changes are dynamically and automatically reflected, eliminating pipeline breakage and the maintenance cycles that had plagued GLOBEX’s custom-coded processes.
- All data is fingerprinted with hash records. Avro schemas are created automatically on the fly, a massive productivity improvement over the legacy model of manually mapping every field in every source.

Use of StreamSets has allowed GLOBEX to massively streamline data ingestion. Rather than each ingestion job becoming a new IT project in the backlog, now a new job is automatically triggered by completing a form that specifies the source and clicking “Build.” The pipeline is automatically built in StreamSets and, leveraging its REST API, machines are deployed automatically. This all occurs within minutes of the initial request.

StreamSets handles data drift dynamically so that when a database changes its schema, for example by adding and removing columns, StreamSets deals with this seamlessly and automatically.

Results

The impact of the DataOps implementation was immediate! Before, DataOps, the data ingestion team had been processing in the range of 25 to 50 jobs on a typical month. Once the DataOps tools and methods were applied, the throughput increased to **350 jobs in a single month**; an order of magnitude or 1,000% improvement in productivity! The massive backlog of requests from the business units — some more than a year old — was cleared in a couple of weeks, freeing up the backlog with no increase in staff.

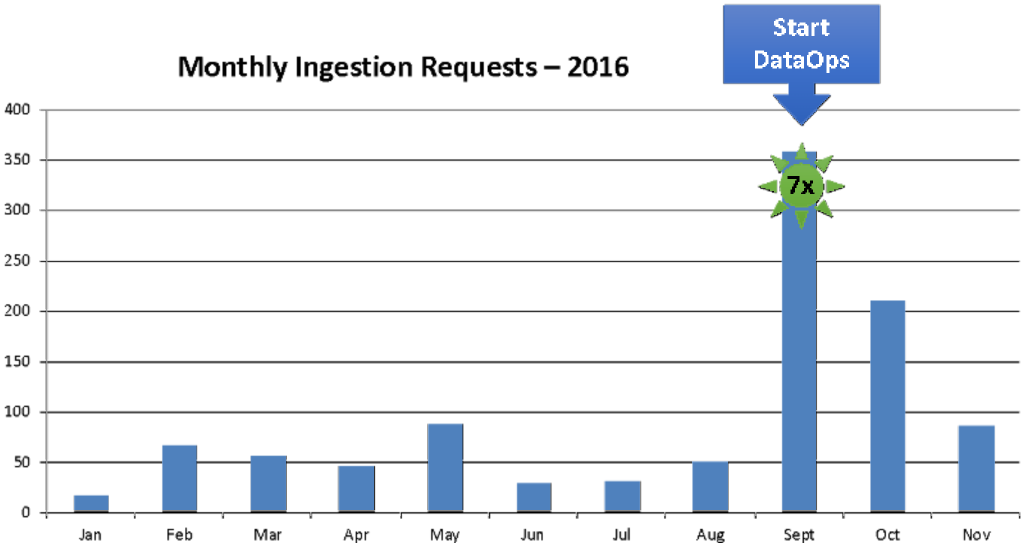


Figure 8. Data Throughput – almost an order-of-magnitude improvement via DataOps methods

Furthermore, the data availability latency, how quickly a new data source can be made available in the data lake, **fell from 21 days to 4 hours**. User demands for data can now be immediately satisfied as fast as they arrive!

Long-Term Benefits

Instead of experienced high-value engineers, interns can now manage the process. No coding required!

Access to data is centralized, federated and democratized. Data is freed from fragile, custom-coded batch processes. *Anyone* from any GLOBEX company can access the data.

Data democratization enables business innovation. Prior to StreamSets, GLOBEX business units were highly selective with their requests for new data sources. They knew that the data teams were overloaded and would not be able to service them for weeks or perhaps months. With the StreamSets COE, these limiting barriers have been torn down; any data source can be made available to the entire corporation *immediately.*

Now that GLOBEX had unified their data, they can move onto new innovations.

Summary

The problem GLOBEX faced would be easily recognizable to data professionals in many large companies – the chaos created by organizational complexity, data variety and the fragmented ownership and governance of the data.

DataOps and the StreamSets technology provided a unique path to the automated on-boarding of new data sources and reliable continual data ingestion. The key features that GLOBEX was able to take advantage of included:

- Any-to-any ingestion infrastructure to eliminate hand-coding.
- Data drift handling to reduce pipeline maintenance due to schema changes.
- REST APIs allowing deployment automation using Ansible, Puppet, Chef and other tools.

The challenge GLOBEX faced could be applied to any industry as companies endeavor to unlock the value of their data. Improving how enterprises make data universally accessible and immediately consumable should be an early priority since so much depends on it in today's data-driven world.

Case Study 2: DataOps in R&D at a Health, Pharmacy and Biotech Company

This case study is based on a global Health, Pharmaceuticals, and Biotech company which we have named as INGEN (International Genetic Technologies) to protect the identity of the actual company.

The Health, Pharma and Biotech industry is enormous, including segments such as Diagnostic Laboratories, Doctors and Health Care Practitioners, Hospitals, Medical Devices, Medical Supplies and Equipment, Outpatient Care Centers and Personal Health Care Products. R&D is an extensive process which requires recording, storing, and analyzing massive amounts of test, experimental and clinical data. This case study applies to any healthcare enterprise that has an R&D organization regardless of whether they are involved in new drugs, vaccines, medical devices, clinical trial research, health execution, or healthcare products.

Executive Summary

INGEN has 3 global businesses that research, develop and manufacture innovative pharmaceutical medicines, vaccines and consumer healthcare products. Its goal is to be one of the world's most innovative, best performing and trusted healthcare companies.

At the time of this writing, the enterprise is 3 years through the DataOps transformation and has a very mature practice. During the journey, they have made a series of policy and process changes, invested in hardware and software to build DataOps systems, and assigned and acquired dozens of staff. The DataOps team is now at the point where they are recognized at the heart of the enterprise strategy and adding tremendous business value that far exceeds the investment to-date.

At INGEN, they implemented the hybrid model with the creation of the R&D Data Center of Excellence. The team is comprised of employees focused on building the information platform, data movement, data curation, data standardization, enablement, data science, and overall program management. The DataOps COE leads the construction of the platform to provide analytics-ready data across R&D and works with research scientists regarding the use of data to answer specific questions.

Starting DataOps Center of Excellence

The journey for INGEN started with an executive memo 3 years ago to initiate the DataOps capability. The memo was extremely significant since it:

- Reinforced the support from an executive sponsor and the entire senior management team, and
- Announced the key Vision and Charter summarized as combining ALL data sources in a centralized shared infrastructure.

The value of this change had a tremendous impact because previously each independent team was creating and maintaining their own data, making it difficult to share enterprise-wide knowledge.

Memo: Data Source Load approval

As you are aware, the vision of the Data Centre of Excellence is for INGEN to be able to better utilize internal/external data to transform our business by providing better insight and improve decision making. A key step towards this vision is being able to copy all of INGEN's data from multiple source systems we have onto the single IIP (INGEN Informational Platform). From this platform, users will be able to conduct exploratory data analysis to answer specific business questions.

To ensure that data integrity, privacy, confidentiality and transparent disclosure is respected, on behalf of the IMT (Management Team) we have set up a "Data Access and Security Committee" which includes senior leaders from across INGEN including, Paul Spencer in his Chief Strategy Officer role and other IMT members. Their remit is to:

- Review and approve access to data sources across INGEN to enable data to be copied onto the IIP,
- Determine the level of access for users where confidentiality or regulatory controls may apply, and
- Provide guidance on linkage of data from various sources which may have different regulatory and security requirements.

At the first meeting in April, we discussed and reviewed the request to begin copying INGEN data onto the platform including one of the most significant types of data we possess, product R&D quality data. After some good discussion the committee approved and agreed that ALL INGEN data sources could be copied onto the IIP. INGEN DataOps COE and relevant I.T. team members will have access to this data to support the loading from source systems.

At a subsequent meeting on June 13, the committee granted permission for access to a limited number of individuals for innovation development. This memo serves as the approval, on behalf of GMT, for the loading of data from the source systems and access approval for these limited individuals (included in appendix).

Please note that discussions on the process for user access, and "who" should have access continues. The committee meets bi-monthly and will seek input from their constituencies where needed to be able to approve decisions on access. Any further decisions will be communicated as appropriate, but feel free to contact me if you have any questions.

Best wishes,
John Smith on behalf of the INGEN Data Access and Security Committee

The memo allowed functional owners to officially approve and gain access to share enterprise-wide data. The value of breaking down their functional barriers and sharing information is now estimated to create billions of dollars of profit for the enterprise.

Sharing data across independent teams may or may not be the strategic driver for your enterprise, but it is critical for an organization to explicitly articulate the strategic needs for data that will enable DataOps results. An appropriate memo by an executive in a simple example, but it this case it was a powerful milestone.

Challenge

One of the major challenges was the development of the steps required to share data across the organization. The DataOps COE created a brand-new Hadoop-based solution as the information platform. One of our main focus areas at the time was centered upon the movement of the data, and one of the major goals was to load at least 90% of the structured data within a six-month period. The only way to achieve this goal was through automation. They selected a technology which would allow them to quickly move data, have access to source data from a variety of structured sources, and move the data onto the platform.

For data acquisition, the COE selected StreamSets to move data from the source systems to the Hadoop platform. A key benefit of the StreamSets technology is the ability to augment with additional automation. Rather than manually create each of the pipelines, the team created the capability to dynamically create and execute all the pipelines directly from the inventory of data sources. Using this approach allowed them to automate the data ingestion process across thousands of structured data sources leveraging 100,000s of pipelines to load 6 petabytes with 4 petabytes refreshed every week.

DataOps COE Advice

The DataOps COE was structured in 4 major areas. First they established an information platform to be the single location for integrated data.

The second part was to enable an extended team. Some team members understand the science in discovering a new medicine, some understand R&D work, while others understand the clinical trials needed in developing a new medicine to assess efficacy and safety. With their deep understanding of the data, they act as a translator between the scientific need and the data and analytics environment. The team correlates data with business needs and attempts to consolidate it.

The third component of the team handles data science and analytics. It is very important that the R&D organization defines and derives value from the data. But in many cases, the R&D department requires assistance in understanding the tools needed. For this reason, the INGEN created a small data science and analytics team to be a catalyst in assisting the needs of the R&D scientists.

The fourth part of the team focuses on overall program management and operation. Building a well-architected, production-level data and analytics environment can be complex. INGEN established a program management office (PMO) early to assure the proper level of project management and to coordinate project plans across the sub-teams. In effect, they have a project management COE inside the DataOps COE. In addition to the PMO, the team also provides internal communication, training, finance, contract management, and leverages the common user experience program.

Finally, as part of the extended COE team, there is also a solution development sub-team which acts as a type of packaging area. It is important to understand the core team not develop software. However, as the business requires a solution to provide access to data, the solution development team leads the effort in producing the dashboards, queries, or analysis by leveraging the technologies available on the platform.

The Solution

Establishing an information platform to support the data and analytics needs for a large organization requires the integration of several technologies. The solution is not a single technology from just one company, but rather a best-in-class ecosystem that delivers a production level, large scale platform. The foundation for the INGEN platform is the Cloudera Enterprise which provides Hadoop and additional components, including security, Spark, Hive, Kafka, search, and manager.

The DataOps COE addressed their data ingestion challenge using StreamSets. For data curation, they used TAMR (www.tamr.com) which uses machine-learning to rationalize data elements and align data to industry data models. There was a lot of data sources, and a lot of similar data sources due to data fragmentation. By using machine learning, they could understand and make sound decisions by utilizing the data itself, rather than having people sitting in a room arguing over data attributes. INGEN use analytics on the data to improve the traditional Extract-Transform-Load process.

To complete the solution, the COE also used Trifacta to enable business user data wrangling, Waterline Data as the metadata repository, ZoomData to provide dashboards and data visualization, and Kinetica to enable GPU database capabilities. The power of a big data and analytics platform is in the collaboration across the ecosystem of technologies and service providers, not in a single component.

Lessons Learned

One of the more important lessons learned at INGEN is that it requires integration of several technologies to create a large scale, successful data analytics platform; and this is not easily achieved. Many of the new technologies have so much capability, they can be difficult to integrate and form a best-in-class ecosystem. It is important to understand the level of work in bringing all of this together to deliver a full production-level solution.

Another key lesson is the significance of developing use cases. Rather than selecting a single use case to use for the start of the program, a better approach is to select a portfolio of use cases from across the business to serve as a base for the program. Addressing 10+ use cases, rather than just one, drives very different decision-making related to the approach and dimensions of both the environment and processes.

Rather than "painting yourself into a corner" with a single use case, the portfolio approach drives a model that enables an easier progression to additional use cases, and addresses production-level items from the start.

WHAT'S NEXT

Stay tuned for the 2nd edition the DataOps publication which will add additional details to take DataOps to an even higher level of capability. It will include broadening the teamwork across cross-functional teams and individuals and institutionalizing DataOps as a company-wide culture. It will also expand on the roadmap introduced in this 1st edition to include more details on the DataOps technology platform, automation and continuous improvement. In addition, it will outline steps for a DataOps Scorecard and measuring productivity, maturity and business value
of the related center of excellence. You can expect to access the DataOps Playbook 2nd edition later this year.

In the meantime, get started with this paper. We would welcome your comments and suggestions by sending a note to streamsets.com/contact-us.

ABOUT STREAMSETS

StreamSets transforms how enterprises flow big and fast data from myriad sources into data centers and cloud analytics platforms. Its DataOps platform helps companies build and operate continuous data flow topologies, combining award winning open source data movement software with a cloud-native Control Hub. Enterprises use StreamSets to enable cloud analytics, data lakes, Apache Kafka, IoT, and cybersecurity.

Founded by Girish Pancha, former chief product officer of Informatica, and Arvind Prabhakar, a former engineering leader at Cloudera, StreamSets is backed by top-tier Silicon Valley venture capital firms, including Battery Ventures, New Enterprise Associates (NEA), and Accel Partners. For more information, visit streamsets.com.