WHITE PAPER

What is DataOps & Is It Worth Adoption?

PETR TRAVKIN
Solution Architect
EPAM Canada

TODD HOMA
Senior Director, Technology Solutions
EPAM US

<epam>

Contents

Introduction

Making data universally accessible within companies has been an imperative across every industry, even before the invention of the first computer. Data accessibility is still a challenge for organizations despite the evolution of business processes and data-related technologies.

Businesses today are focused on how to evolve their culture, processes, organizational structure and technologies to become truly data-driven companies. However, the focus area for addressing these issues has shifted from technology to process.

According to a Forrester Consulting survey, consisting of 900 global business leaders commissioned by Collibra, businesses that rely on data management tools to make business decisions are

58%

MORE LIKELY TO BEAT THEIR REVENUE GOALS THAN NON-DATA-DRIVEN COMPANIES.

Additionally, companies leveraging data see an 8% boost in customer trust, and a 173% advantage in efficient regulation compliance, compared to non-data-driven companies. By contrast, less data-mature organizations are 55% less likely to say internal data management strategies lead to optimal business decisions.[1]

There is no doubt that data is more important than ever before, but how do you begin to tackle the overwhelming amount of information within your organization to become truly data-driven? In this white paper, we will explore many of the challenges that companies face when addressing data accessibility, discuss data debt, introduce the concept of DataOps and how this approach can help businesses operationalize data science to glean insights and accelerate innovation.

[1] https://www.ciodive.com/news/data-driven-companies-revenue-coronavirus-covid19/578159/

Data Challenges: Old & New

Deriving business value from data is often a long and cumbersome process, regardless of how mature your company is. There are several challenges that companies typically face when addressing data accessibility:

IMMATURE DATA AND ANALYTICS PIPELINES

The process of building data and analytics pipelines continues to be a largely manual, non-repeatable process with minimal opportunities for reuse. Often, this results in a plodding, error-prone development environment that is slow to respond to change requests, making it nearly impossible to keep pace with the demands of a data-driven business. Immature development and delivery processes force business users to build their own pipelines, resulting in an ever-expanding universe of ungoverned data silos that go unnoticed until a major decision backfires.

BULKY AND UNMANAGEABLE DATA VOLUMES

While some companies are still figuring out how to leverage data, others have embraced the concept but are struggling with the magnitude of the data landscape, and the number of new systems and technologies developed to address the need for big data. Additionally, businesses often approach data platform implementation from a technical perspective rather than a business perspective, which can result in the platform becoming an ineffective dumping ground for data instead of defining use cases that are mapped to business value.

LACK OF AN IDEAL DATA UNIFICATION SYSTEM

The benefits of unifying data sources are obvious, but for enterprises operating at a large scale, there is more data than typical ETL tools can manage. Everything from accounting software to factory applications are producing data that yields valuable operational insights. The availability and value of data sources on the web compounds the scalability challenge. Moreover, enterprises are not static. Scalable data unification systems must accommodate the reality of shifting data environments.

UNQUALIFIED DATA TEAMS

Data teams are often familiar with data warehouses, which have mature architecture design patterns and are strongly integrated with other data tools. Alternatively, new concepts – such as a data lake, operational data hub or data factory – are less mature and require specific expertise that is difficult to find. In addition, since the operational tools to manage these new approaches are continuing to evolve, it is more difficult for companies to support them compared to data warehouses.

DIFFICULTY IN CREATING PERVASIVE SELF-SERVICE DATA ACCESS

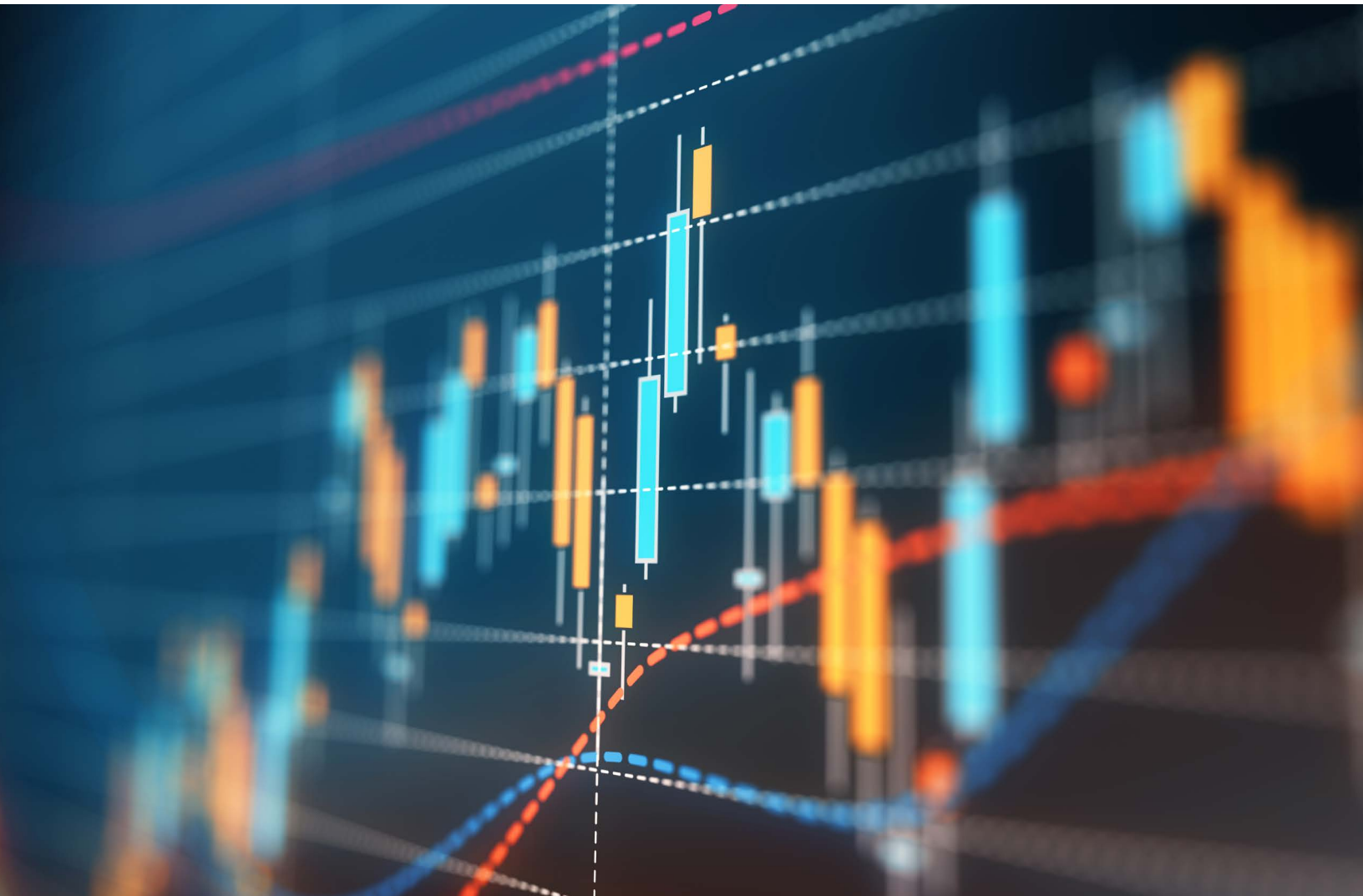
To truly democratize data, a company needs to transform data access tools and infrastructure provisioning to a self-service mode. This process requires a thoughtful, strategic and collaborative effort between the business and technology functions, while maintaining a continuous feedback loop between data consumers, analysts, scientists and data engineers.

Data Challenges: Old & New (*continued*)

IN ADDITION TO THESE CHALLENGES, COMPANIES TODAY FEEL PRESSURE FROM SEVERAL BUSINESS DRIVERS, SUCH AS:

Competition from digitally native companies	Opportunities brought about by the democratization of analytics, which is driven by new products and companies that enable the broad use of analytics tools	The need for greater agility as data that does not move at the same pace is dropped from the decision-making process	The proliferation of data sources through IoT and social media

The way companies address data challenges due to these various workarounds ultimately results in an increasing amount of data debt. The questions is: How can your organization minimize it?



Data Debt & How DataOps Can Help

Data debt stems naturally from the way companies conduct business, especially when companies are run as a loosely connected portfolio. Each line of business wants control and rapid access to their mission-critical data, so they start making "free-rider" decisions about data management and procuring their own applications, thus creating silos. Managers move talented personnel from project to project, creating turnover among data systems' owners. As a result, most large enterprises still face the reality of intensely fractured data environments and general data heterogeneity, also defined as data debt. These issues and drivers beg for a new approach to building data analytics solutions that increases agility and cycle times, while reducing data defects.

DataOps is a discipline intended to enable companies to pay down their data debt by rapidly and continuously delivering high-quality, unified data at scale from a wide variety of enterprise data sources.

THERE ARE MULTIPLE DEFINITIONS OF DATAOPS THAT ARE USED IN THE MARKET, SUCH AS:

DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization.[2]

DataOps is an automated, process-oriented methodology used by analytics and data teams to improve the quality and reduce the cycle time of data analytics.[3]

DataOps is a set of best practices that improve coordination between data science and operations.[4]

DataOps is an emerging methodology for building data analytics solutions that deliver business value. Building on modern principles of software engineering, DataOps applies rigor to developing, testing and operating code that manages data flows and creates analytics solutions.[5]

All these definitions are valid, but do not address the practicality of "how" it works. The main idea behind DataOps is that it delivers the product (production-ready data) via the process (operations), which addresses quality, timeliness, transparency and agility. Taking a cue from DevOps, DataOps looks to combine the production and delivery of data into a single, agile practice that directly supports specific business functions.

[2] Gartner, Introducing DataOps Into Your Data Management Discipline, Ted Friedman, Nick Heudecker, October 31, 2019.
[3] https://en.wikipedia.org/wiki/DataOps
[4] https://www.ibmbigdatahub.com/blog/3-reasons-why-dataops-essential-big-data-success
[5] https://www.eckerson.com/articles/dataops-explained-a-remedy-for-ailing-data-pipelines

DataOps: The Core Principles

You can read (and even sign!) the DataOps Manifesto at www.dataopsmanifesto.org.
The core principles of DataOps are summarized below:

Apply Agile & Software Engineering Best Practices

Short time-to-delivery and responsiveness to change are mandatory. Using version control, automated regression testing of everything, clear code design and factoring is mandatory too.

Integrate with your Customer & Deliver Business Value

The DataOps team benefits from customers and the engineering teams they support being in-house, readily available for daily interaction. Gather feedback as frequently as you can. Data is not an end, but a means to delivering insights that add value to the business and satisfy the customer.

Collaboration & Communication

Share knowledge, simplify communication and provide feedback at every stage of the data analytics lifecycle.

Analytics as a Code

Look at data artifacts, such as models and visualizations, like you would code and adopt software methods (version control, automated testing and continuous deployment). This also means host configuration, network configuration, automation, gathering and publishing test results, service installation and startup, error handling, etc. Everything needs to be code.

End-to-End Processes & Continuous Improvement

Avoid data silos and consider analytics an enterprise endeavor. Orchestrate data, schema, tools, code and stakeholders throughout the data landscape. Learn from mistakes, review processes continuously and adapt to changing circumstances.

Maintain Multiple Environments & Integrate Toolchains

Keep development, acceptance testing and production environments separate. Never test in production, and never run production from development. Maintain multiple environments, but within each environment, everything needs to work together. The different domains of operations require different collections of tools (or toolchains). These toolchains need to work together for the team to be efficient.

Reuse & Automate

Automate wherever possible and reuse existing artifacts to avoid unnecessary rework and repetition.

Short Cycles & Incremental Change

Avoid “big bang” releases and bloated processes. Iterate in short cycles so you can adapt quickly to new and changing needs.

Test Everything

Make quality and testing a top priority and ensure that no untested artifact reaches production. Automated testing is what allows you to make changes quickly, having confidence that problems will be found early, long before they get to production.

Full-Stack Monitoring & Data-Driven Improvement

Continuously monitor applications down to infrastructure and use those insights to enhance performance and reliability.

Keep It Simple!

Whenever an easier solution appears, it is likely also a superior one.

The Benefits of DataOps

DataOps can help organizations handle complex data landscapes and analytics solutions that require the coordination of a broad range of stakeholders and technologies. Below are just some of the benefits of DataOps:

ACCELERATES TIME TO PRODUCTION

A major driver for DataOps is speed. Streamlined, largely automated analytics pipelines help deliver new features and insights quickly and reduce manual effort. Moreover, the short feedback and testing cycles help speed up reactions to changing business requirements and increase flexibility.

INCREASES THE VALUE PROPOSITION OF DATA AND ANALYTICS

The stages and steps that must be orchestrated in a data analytics pipeline are not always serial; often multiple steps happen in parallel and once completed, a step might be repeated as part of an iterative, agile workflow to refine output until it gains user acceptance. This approach increases quality because it ensures that no untested change makes it to production. It also improves orchestration and collaboration as different stakeholders in the pipeline rely on another and work together in a fluid process.

SUPPORTS THE MANAGEMENT AND ORCHESTRATION OF HETEROGENEOUS TECHNOLOGIES

A key role of DataOps is to orchestrate and automate the flow of data and code between people and tools in an efficient manner that ensures clean handoffs and minimal errors and disruptions. With complex pipelines, this can be challenging, making orchestration and automation key requirements in any DataOps implementation.

IMPROVES COLLABORATION AND ESTABLISHES A CULTURE OF CONTINUOUS IMPROVEMENT

DataOps requires a change in culture that promotes collaboration, trust and responsibility. The goal is to blur the lines between departments and functions, encourage the exchange of knowledge, reduce conflicts and eventually increase productivity. The convergence of different roles helps align changes throughout various stages, like when a data engineer is informed about the later cleansing issues encountered by a data scientist, or the lack in performance of an ETL process in production.

ENSURES THE STABLE AND EFFICIENT OPERATION OF APPLICATIONS AND INFRASTRUCTURE

Well-defined analytics pipelines enhance both speed and robustness of data. Multiple stages of automated and manual tests prevent the deployment of flawed updates. DataOps also includes monitoring production environments to identify bottlenecks or potential issues, thereby improving the efficiency and stability of infrastructure and applications.

ENABLES SELF-SERVICE

With greater automation and machine learning algorithms that simplify development, deployment and performance of management tasks, organizations need fewer experts to build and manage data and analytics pipelines. Business users with some technical savviness can build their own pipelines or move code into production.

CONCLUSION

Over the last decade, we've seen that DataOps has transformed from an emerging trend to an applicable set of principles and technologies. DataOps adoption has skyrocketed in the past year, driven by data dissemination across hybrid and multi-cloud environments, increased data privacy regulations and the need for companies to accelerate innovation in a hyperdynamic digital landscape. Companies that are looking to transform to data-driven enterprises should rethink their approach to managing data and consider adopting DataOps processes.

ABOUT EPAM SYSTEMS

Since 1993, EPAM Systems, Inc. (NYSE: EPAM), has leveraged its core engineering expertise to become a leading global product development and digital platform engineering services company. Through its 'Engineering DNA' and innovative strategy, consulting, and design capabilities, EPAM works in collaboration with its customers to deliver innovative solutions that turn complex business challenges into real business opportunities. EPAM's global teams serve customers in over 25 countries across North America, Europe, Asia and Australia. EPAM is a recognized market leader among independent research agencies and was ranked #8 in FORBES 25 Fastest Growing Public Tech Companies, as a top information technology services company on FORTUNE'S 100 Fastest Growing Companies, and as a top UK Digital Design & Build Agency. Learn more at www.epam.com and follow us on Twitter @EPAMSYSTEMS and LinkedIn.

GLOBAL

41 University Drive,
Suite 202
Newtown, PA 18940, USA

P: +1-267-759-9000

F: +1-267-759-8989

<epam>
</epam>