Qlik and DataOps: A Vision for Business Value

DAN WOODS
CONTENTS

Introduction ................................................. 3
  Qlik and DataOps: A Vision for Business Value ............................................. 3

The Anatomy of DataOps ...................................... 5

The Fundamental Problems of Data Management and Analytics ..................... 7
  Qlik’s Early Efforts Prefigure DataOps .......................................................... 7
  Foreshadowing DataOps .............................................................................. 8
  The Productization of DataOps .................................................................. 9

Qlik’s Design for a DataOps Platform ............................................................ 11
  Qlik’s Data Integration Platform .................................................................. 11
  Qlik Data Catalyst ....................................................................................... 12
  Qlik Sense ................................................................................................. 13
  General Data Consumption ........................................................................ 13
  Productized DataOps ................................................................................ 13

Where Will DataOps Be in Five Years? ........................................................... 13
  For the Rest of Us ....................................................................................... 13
  A Future of More Guidance ..................................................................... 14
INTRODUCTION

It is an amazing time to be involved in the world of data management and analytics. The world seems to be rapidly moving toward a new synthesis of existing capabilities. This synthesis combines past victories in the realms of data lakes, data warehouses, business intelligence, and data discovery and knits them into a streamlined form that masters the complexity of the growing data landscape and allows for agile, incremental evolution.

This new synthesis is called DataOps, a set of practices supported by technology that accelerates broad access to fresh data in the same way that DevOps accelerated software development.

As an analyst of data-related technologies for the past 15 years, it is extremely exciting to see how DataOps is creating systems that are providing what we have always wanted from our data platforms.

Here's an outline of the fundamental problems that DataOps seeks to address:

- To expand the use of data we must have a larger landscape of datasets.
- Datasets don't spring from data sources ready to use; they must be crafted into the form needed to support analysis or other uses.
- To keep those datasets fresh, we need a huge number of real-time data pipelines.
- All of those data pipelines must be affordable, manageable, and governable.
- We need a raft of new, better technology to support this.
- DataOps must allow as many people as possible to participate in finding data, curating it, creating datasets and pipelines, and adapting ways to use that data to create value. This is a lot of work on an ongoing basis, so it cannot just be done by a small collection of experts.

The fundamental challenge of DataOps is to create a massive unified data supply chain supported by automation. This new data supply chain must be agile and robust to handle problems that arise and to incorporate new data in an orderly fashion. The result is more people using more data.
Anyone who cares about expanding the use of data in their organization will have to grapple with the problems DataOps addresses. The question is how? What will DataOps mean to you and your teams?

Don't worry if the goals of DataOps sound impossible or too good to be true. In the early days of DevOps, the goals seemed to be impossible as well. DevOps sprang from the need to rapidly evolve a highly complex application like Google Search, Facebook, or Amazon, but at the same time keep the system running flawlessly with little or no downtime. By bringing everyone involved together, changing assumptions, reengineering the process, and using new technology in new ways, DevOps achieves seemingly impossible goals. You can rebuild the car while it is going 100 miles an hour.

This paper takes the long view and shows how the fundamental problems being addressed by DataOps have been with us for a while and have been addressed to some extent in every generation of business intelligence and data processing technology. But now, aided by cloud computing, numerous powerful SaaS services to meet specialized needs, and growing data literacy, we are entering an accelerated phase of evolution.

By looking at the roots of DataOps, its current form, and how the technology to support it is evolving, you will be better equipped to benefit from this unstoppable trend. DataOps will teach us how to build a nervous system of data that brings awareness and power to every part of our organizations.

Technology vendors like Qlik are extremely excited about DataOps because it helps show the way to a future in which data plays a much bigger and more important role in business. My goal in this essay is to better understand the incremental path to DataOps by examining how Qlik has solved fundamental problems in the past and the moves it has made to address those problems in the present, and future. To do this, we will look at the current state of DataOps productization as it exists in the Qlik portfolio, and then look forward to the next stage in which new systems will complete the job.

I'm using Qlik as an example because I have followed the company closely for years, they have been actively seeking to accelerate DataOps, they are committed to using DataOps as a way to make sense of the current landscape, and they wanted to co-sponsor this analysis.
DataOps must bridge more silos and serve more people. While not everyone develops software, everyone at every level in an organization needs data.

But even if you never plan on using Qlik, this essay should give you a better understanding of what DataOps is, why it is important, and how it could fit into your organization's future.

THE ANATOMY OF DATAOPS

For a deep dive into the origins of DataOps, the nature of the modern data landscape, and the detailed problems that must be addressed in an implementation see “The Detailed Anatomy of DataOps.”

Here is a brief introduction to DataOps, which will suffice to bring those unfamiliar with the concept up to speed.

DataOps is an evolution and adaptation of DevOps.

DevOps bridges the silos of the development and operations worlds. DevOps focuses on improving automation and productivity in the process of building and running a software product with the ultimate goal of improving user happiness. DevOps dramatically expanded automation techniques in software builds, testing, monitoring, and communication.

A great deal of technology evolved to support these needs, largely from companies with significant engineering resources. As the environment surrounding DevOps sprung up, there was a concurrent move to the cloud and SaaS, which also increased the importance of DevOps. Containers and Kubernetes made DevOps work smoothly. It’s my opinion that the same thing will happen with DataOps.

DataOps is more ambitious than DevOps. DataOps must bridge more silos and serve more people. While not everyone develops software, everyone at every level in an organization needs data.

DataOps seeks the same type of integration and unified team as DevOps. DataOps moves beyond the silos that divide data engineers, data scientists, data stewards, and end users. It empowers companies to think about data engineering at the start to create a process that can meet the needs of all users throughout the life cycle of data usage.
In order to succeed in dramatically expanding the value that companies get from data, DataOps has a number of tasks:

- DataOps will support a vast increase in the number of data pipelines needed to support broader use of data by creating a new kind of agile data supply chain.
- DataOps will support an expanded set of canonical data presented in layers of data models to support the use of data and the creation of pipelines. This set of models will become the data product inside an organization.
- DataOps will make supporting the larger number of pipelines economical by increasing self-service through simplified abstract methods for creating pipelines and by following the lead of DevOps in automating the flow and transformation of data from end-to-end in a way that is resilient, robust, and allows for rapid change.
- DataOps will seek to break down organizational silos that have thwarted communication, collaborative design, and progress toward providing all data needed for supporting decisions in an appropriate form and a timely matter instead of in
- DataOps infrastructure will be built and run by teams that include all needed skills working together in a unified manner, the way that developers and operations staff were brought together in DevOps.
- The shape and operation of the systems supported by DataOps will seek to optimize the business value created through data, just as DevOps is focused on maximizing customer happiness.
- A new generation of tools and technologies is being created and adapted to support DataOps.

This all sounds good, but it is a tall order. There are many moving parts and lots of complexity. And at this point, DataOps is not a product or even a stable set of requirements. Many will disagree with the definition just presented. Some practitioners do only a portion of what is described above and call it DataOps.

The point is not to adhere to my or anyone else’s definition. The important thing is to expand the use of data, which requires implementing the capabilities listed above.
The need to create canonical datasets and to refresh them has been part of business computing from day one. Mainframes created lots of data supply chains using batch jobs.

THE FUNDAMENTAL PROBLEMS OF DATA MANAGEMENT AND ANALYTICS

To better understand the roots of DataOps and its relevance, it’s helpful to look back at previous efforts to support data pipelines. By understanding these previous developments, we can understand the power of DataOps today more deeply. We can also see how solutions to older problems must be updated for the modern world to achieve the goals of DataOps.

Qlik’s Early Efforts Prefigure DataOps

The need to create canonical datasets and to refresh them has been part of business computing from day one. Mainframes created lots of data supply chains using batch jobs.

The advent of data warehouses dramatically expanded the number of data pipelines. Several generations of ETL technology arrived to support data warehouses and were later employed to feed data to next-generation dashboarding and data discovery systems like Qlik. These ETL systems were created to make the work of accessing data sources, transforming it, and moving it around easier. More people could be involved and more data pipelines could be created and maintained with the same resources.

In this era, Qlik addressed the problem in the following way:

• Qlik created a technology, QlikView (and later Qlik Sense), that allowed more explorability within a dashboard based on an associative, in-memory model that supported a highly interactive user experience:
  › Qlik allows users to change the selection criteria for data and immediately see the results on a set of connected dashboards and visual elements.
  › One of Qlik’s innovations is the green, white, and gray display mechanism that shows selection criteria, matched data, and the data that doesn't match but is potentially related, respectively.

• As QlikView became popular, the demand for Qlik extract files created an early form of the challenge of DataOps: creating, managing, and operating data pipelines at scale to create many Qlik extracts.
This model worked, and many companies created data supply chains that included large numbers of data pipelines using Qlik’s technology.

Qlik decided the best way to handle this was to create its own data supply chain technology where you could land data and then use a proprietary ETL language or its data prep tools to massage that data. The outcome would be a data extract file that could be directly used in a dashboard.

This model worked, and many companies created data supply chains that included large numbers of data pipelines using Qlik’s technology.

Qlik then took it further, adding Expressor, a metadata management based solution that introduced governance over existing Qlik deployments. With Expressor, Qlik customers had visibility and governance over the contents of all Qlik data and applications across the enterprise. But it wasn’t prescriptive.

Foreshadowing DataOps

Qlik’s efforts foreshadowed DataOps in several ways. Qlik created new tools to simplify and productize functions that had been reliant on custom code and applications. The tools were built to work together in a coherent system. The data supply chain in the QlikView era allowed data engineers to create data pipelines with tools that helped increase productivity.

But the approach suffered from many of the limitations of all technology of that era:

- Data pipelines could only be created and maintained by experts.
- Coding in relatively low-level languages was the primary means of creating data pipelines.
- Operations of data pipelines and data management were not supported using products. Most of it was roll your own.
- The data supply chain was focused on one method of using the data: in this case, QlikView.
- The expansion of self-service to create and maintain data pipelines started with easier to use, visual, and guided interfaces to ETL, but the need for self-service data prep tools to address the last mile of data was not initially addressed.
Despite these challenges, which were the same across the entire business intelligence industry, powerful data supply chains were created. Companies have to be able to land data, put it into a purposeful form, and then deliver it to users.

The Productization of DataOps

In the big data era, many practitioners were already doing something like DataOps without actually using the name. Starting around 2010, innovators and early adopters used open source technologies like Hadoop and Spark to craft the kind of highly automated end-to-end data pipelines that DataOps is all about.

But even though many of these pipelines achieved high levels of automation, these early efforts were anything but agile. Companies with large engineering teams painstakingly built handcrafted data pipelines. One great part of this era was that the range of tools and databases for building data pipelines rapidly expanded. The Hadoop ecosystem was especially creative in adding projects for many specialized tasks.

The data pipelines created in this era could be large and complex and handle data volumes never before contemplated and employed powerful transformations and analytics.

These systems were hard-fought end-to-end victories and accelerated analysts' ability to assemble data, create a pipeline, and create useful analytics. If DataOps required this type of effort, it wouldn't really be feasible.

This brings us to a key point. DataOps will never come to life at scale if it requires an army of developers. In order for DataOps to succeed it must be productized, which is the only viable path to widespread adoption.

The good news is that even though we are in the early days, it is clear that the needed productization is underway and accelerating.
A number of new processes and tools have already been developed to support the achievement of these goals. These technologies include:

- Modern ETL systems
- Various systems for using object storage
- Data catalogs
- Change data capture systems
- Data prep tools, some of which offer self-service

It is not clear that this entire process will ever be fully productized. But what is already clear is that the new collection of components that have emerged can be integrated into a form that gets closer and closer to DataOps as I have described it.

The companies that get the most value from DataOps will be those that craft a long-term vision to fit the needs of their business and then provision systems to bring it to life. The challenge will be to create an orderly strategy to move closer and closer to a complete vision of DataOps.

The companies that create a vision for what DataOps means to their business will enjoy much higher levels of self-service and transparency with respect to how data is used and managed within the business. As a result:

- Data itself will become a product that can be managed as such.
- Data governance will be integrated.
- Companies can see where data is being used and prioritize efforts to improve the quality or quantity of data.
- The entire DataOps process can be optimized for continuous improvement.
- The number of people who use data to make business decisions will dramatically increase.

Several vendors are rapidly expanding their portfolio of technologies to create a productized platform for DataOps. Let's look at what Qlik is doing in this realm.
Many data lakes end up becoming places for hoarding data rather than offering a way to organize, gain insights, or derive business value from data.

**QLIK’S DESIGN FOR A DATAOPS PLATFORM**

The first thing to note is that Qlik as a company has expanded its vision. The incarnation of Qlik that created QlikView and then Qlik Sense to support data pipelines has become a new company. Qlik today is of course interested in creating data pipelines to deliver data to Qlik Sense, its product designed for a world of cloud, mobile devices, AI-powered user experience, and advanced analytics. But Qlik’s acquisitions of Attunity (now Qlik Data Integration), Podium (now Qlik Data Catalyst), and several other firms have assembled an infrastructure that can work with any BI tool to productize a significant amount of the anatomy of DataOps we have discussed above.

Here is a tour through Qlik’s DataOps portfolio.

**Qlik’s Data Integration Platform**

Qlik’s Data Integration Platform addresses the challenge of accessing data at the source, landing it, and then creating pipelines to refresh various levels of modeled datasets. It provides the ability to support:

- Data engineering
- Data profiling and usage analysis
- Model-driven and abstract ETL
- Operation of production data pipelines
- Data logistics for moving and replicating data through change data capture and other methods
- Data profiling, cleansing, tagging, prepping and cataloging capabilities

Qlik’s Data Integration Platform can create productized data pipelines to deliver landed, modeled, and purpose-built data. It could be a data warehouse or a data lake, on-premise or in the cloud.

Qlik’s Data Integration Platform is especially useful for mission-critical productization and the more stable parts of departmental productization. Advanced users will be able to use Qlik’s Data Integration Platform for experimental purposes.
Allowing more people to use data directly leads to gathering more data about usage, which in turn makes the marketplace more powerful.

Qlik's change data capture system allows companies to land data in the cloud so that data engineering can occur there. Additionally, it allows models to be created in the cloud, and companies can use ETL to find all data sources and generate desired transformations as Spark code. Thus, not only can the data models be used and generated in data warehouses, but also in a Kafka output stream.

**Qlik Data Catalyst**

As the last phase of the Qlik Data Integration Platform, Qlik Data Catalyst takes data that has been delivered and refined a step farther and provides technology to dramatically broaden the number of people who create their own pipelines.

Qlik Data Catalyst provides expanded data catalog capabilities so that data can be profiled, described, and easily found. It offers end users self-service for creating new datasets and pipelines. In this way, end users create new models and purpose-built datasets, as well as making ad hoc changes as needed.

Qlik Data Catalyst tracks data lineage, including all transformations, along with usage data. The data catalog technology in Qlik Data Catalyst offers unique governance and security features. Data creators can post data in the catalog while data stewards set policies that ensure end users see all the data they are permitted to access.
Data marketplaces encourage users to share data across the organization and thus transform data into higher and higher forms of value.

Qlik Data Catalyst provides the next stage of the DataOps infrastructure by starting with the orderly data that is created by the earlier phases of Qlik's Data Integration Platform.

Here are a few relevant use cases:

- Understand the content and quality of your data warehouse and lake for regulations like GDPR and CCPA.
- Enable data engineers and data stewards to quickly and easily combine, tag, standardize, and filter data so that data consumers can select for themselves the data they need.

**Qlik Sense**

Qlik Sense allows data to be delivered into a variety of business contexts, completing the end-to-end cycle.

From a DataOps perspective, Qlik Sense supports the experimental use of data both for creating new ad hoc datasets and supporting data pipelines.

**General Data Consumption**

One of the key properties of the combined DataOps infrastructure created by Qlik's portfolio is that data can be consumed by or delivered to a wide variety of consumers, not just QlikView or Qlik Sense. Analytics tools, databases, applications, and data integrations can all receive the data they need.

**Productized DataOps**

With all of these tools, Qlik has productized its approach to DataOps and gives companies transparency into their data to understand and monitor usage. Qlik's portfolio represents the state of the art with respect to a fully productized DataOps implementation.

**WHERE WILL DATAOPS BE IN FIVE YEARS?**

If we think about what happened with DevOps, certain companies created massive DevOps systems through their own engineering powers and they benefited from that work. There were also companies that started using open source tools to implement DevOps in their organization.
For the Rest of Us

But then, there was a much larger group of users that addressed DevOps using commercial products that arose to enable even non-engineering-rich companies to succeed.

The same thing will likely happen with DataOps. There are companies with robust engineering resources that are already creating data products to support their entire businesses. They have engineering support to make them agile because the data products are so cost beneficial to the company.

For the rest of the world, though, there will be simpler, more visual platforms, such as those offered by Qlik. There are also tools focused on specific parts of the DataOps process, in addition to comprehensive platforms like Qlik.

A Future of More Guidance

Regardless of what companies adopt, there will be more guidance on how companies create and manage these large data pipelines. The world of data prep offers a variety of tools that guide users in understanding the breadth of the data available. Companies will thus be creating systems that are much more menu-based and productized than in the past.

The same will be true of operationalizing data pipelines. AI and automation that guide discovery and reveal insights from the data will become commonplace. I predict similar tools will arrive to simplify governance over all datasets.

Looking ahead, it’s clear that systems must work both in the cloud and on-premise. Systems that only work in the cloud will ignore the large amount of data that resides on-premise. Comprehensive tools must address both so as not to ignore DataOps in on-premise environments. The speed of insights will also likely accelerate. And at some point, we will see the development, possibly at the modeling layer, of a Kubernetes-like tool that gives users what they want when they want it.