Data observability's newest frontiers: DataFinOps and DataBizOps

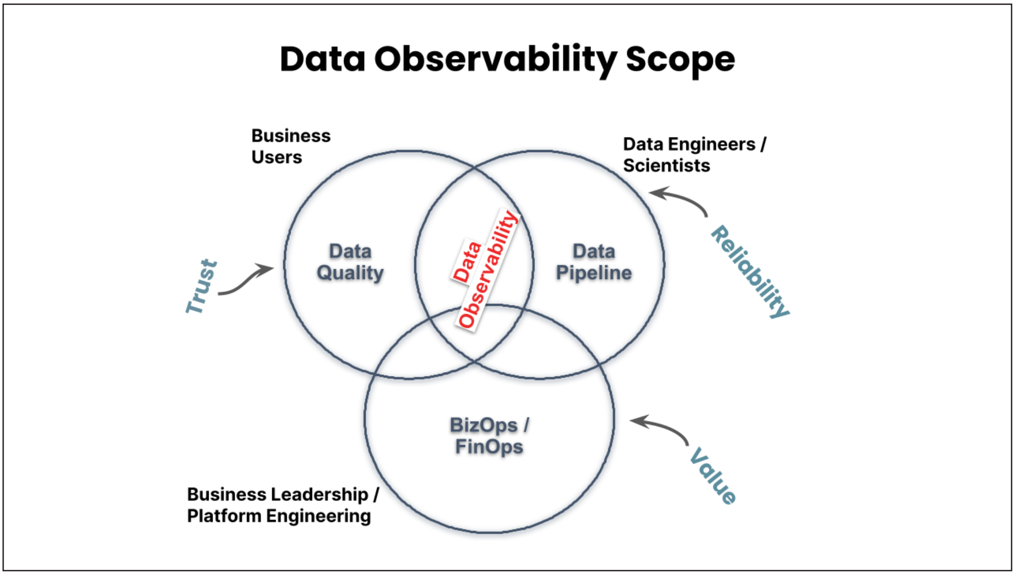
Chapter 13

Sanjeev Mohan,
Principal, SanjMo & Former Gartner Research VP, Big Data and Advanced Analytics

The sign of maturity for any new technology is when it goes from being a nice-to-have to a must-have. As we have seen in this book, data observability serves critical use cases ranging from data quality to data pipeline reliability. And, as we enter a new era defined by two major shifts – slowing economy and the rise of data products, new use cases have emerged. These pertain to helping improve business operations (DataBizOps) and control costs of our pipelines (DataFinOps).

Data observability is a key component of DataOps. It helps improve the speed and quality of data delivery, enhance collaboration between IT and business teams, and reduce the time and costs associated with data management. DataOps helps organizations make data-driven decisions faster and with more confidence. It provides them with a unified view of data across the entire organization and ensures that the data is accurate, up-to-date, and secure.

As data observability matures, we expect it will fold use cases like data security monitoring and protection into a comprehensive metadata platform. The following diagram shows the current scope of this concept.



Data observability is a multi-dimensional concept that addresses three areas: data quality, pipelines or infrastructure, and business operations including cost metrics. These three areas respectively deliver trust, reliability and value.

This chapter focuses on DataFinOps and DataBizOps.

What is DataFinOps?

The *FinOps Foundation* has defined best practices for cloud financial management (CFM) practices. When these principles are extended to data management, we call it DataFinOps. The financial metrics are tracked and monitored like data quality and performance metrics. And the data observability system recommends improving cost efficiency.

DataFinOps is a set of processes used to track costs incurred across the data stack and by different users and teams to optimize overall spending. Ideally, it provides a granular view into how costs are being incurred before they impact budgets and forecasts. Using the data observability capabilities of monitoring, alerts, and notifications, proactive actions can be taken to ensure that data pipelines run efficiently. The cost consumption is across granular or aggregate level for:

- **Resources**, like Snowflake, Databricks, Hadoop, Kafka etc.
- **Users and departments**, like business analysts, data engineers and data scientists
- **Workloads**, like data preparation pipeline, data quality, data transformations code or training models

A DataFinOps platform should provide a consolidated set of capabilities:

- **Discovery** (or observability) of the costs and spend with an ability to tag and categorize costs based on the organization's desired taxonomy. It helps to track, visualize and allocate costs. Tracking should be precise and use workload-aware context.
- **Budgeting**, **forecasting** and a **chargeback** (or optimization) of costs. It helps to eliminate inefficiencies and optimize cost. It uses AI to predict capacity rightsizing and automation to trigger preemptive corrections.
- **Recommendations** (or governance) on cost reduction strategies and opportunities. It helps to identify guardrails needed to control and prevent cost overruns. This could be across infrastructure resources, configuration, code or data. For example, datasets with low usage could be candidates for a cheaper storage tier.

When data observability is used for quality or pipeline reliability use cases, the personas are business analysts, data engineers or data scientists. For the DataFinOps use case, the primary personas are executives, such as the head of the cloud center of excellence or the CIO. These executives work with chief data officers (CDO) or the head of analytics, who manage data assets.

Interestingly, the fiscal discipline is now being pushed down to data engineers. DataFinOps practitioners inculcate a culture of financial discipline and facilitate collaboration between business, engineering, operations, and finance teams. They help improve cloud efficiency across various business teams through cost and efficiency metrics, such as resource utilization and trend analysis.

Why do we need DataFinOps?

Data is a victim of its own success. Data has gone from derided as the "exhaust" of applications to being a kingmaker. It has gone from being used by a few specialized analysts to a wide range of consumers who throw ever new use cases at it. In addition, the average number of sources that generate data has also increased significantly.

Imagine, we have gone from a few dozen data sources, like ERP and CRM, all within our firewalls to an explosion of SaaS products. *Recent studies* show that enterprises on an average used 110 SaaS products, and large companies now have close to 500.

On the one hand, we are dealing with a larger number of data sources, more data consumers, and more use cases, but on the other hand, we've also had a stampede of new tools and applications processing our data. These tools are no longer confined to our firewall-protected environments but are found on the edge, in private clouds and in public clouds. The data pipelines have grown complex and unwieldy. Organizations are now keen to understand the financial metrics of building and operating data and analytics workloads.

This scenario is the very essence for the genesis of the data observability space. Data engineers monitor the data and the pipeline to quickly detect and analyze quality and reliability issues. However, the cost of running the pipelines has never been in their scope, until now.

The economic downturn of 2023 has hit the tech sector the hardest. As infrastructure costs rise, management is keen to prioritize their spending. But how does one prioritize without clearly understanding the costs incurred at various data pipeline stages? Organizations estimated their budget and allocated capital expenditure (CAPEX) and operating expenditure (OPEX) towards their new IT initiatives. Cloud computing has upended this model by shifting most costs to OPEX. With this shift, many teams find out that the money allocated for the entire year has been consumed in the first quarter. We are now dealing with two problems - rising costs and unpredictable costs.

Studies have shown that 30% of the cloud cost is simply wasted. This is low-hanging fruit for a team looking to pare down redundant expenses. It is relatively easy to use native cloud provider cost management tools to identify unused instances and shut them down. This itself can save significant expenses. But it gets harder to pinpoint inefficiencies in complex SQL queries, ML training and data transformation workloads. Hence, DataFinOps needs a more concerted effort.

Data engineers embracing the "shift-left" approach can yield a much more proactive approach to cost containment. Sometimes rising costs are a harbinger for incorrect data or pipeline defects. So, the DataFinOps use case of data observability complements its other well-known use cases of data quality and pipeline reliability.

Challenges of DataFinOps

Cloud computing is celebrated for developing reports and dashboards infinitely faster. Within a minute, we can spin up a virtual warehouse in Snowflake and use its Zero Copy Clone to analyze the data. Similarly, we can use Terraform to spin up cloud resources on demand. But this ease results in a complex web of cloud costs.

Take the example of Accenture, which has over 1,000 teams using AWS. Its cloud bill runs into tens of millions of *lines of billing data*. The reason is so many cost elements in the cloud, such as compute instances, compute transactions, storage, storage I/O, data transfer, support, networking, monitoring, disaster recovery, and various other costs.

The second reason for cost complexity is because of the vast expansion of choices in the cloud. For example, as of February 2023, *AWS offers* 536 instance types running Linux and 427 running Windows. This leads to the challenge of calculating the cloud's total cost of ownership (TCO) and return on investment (ROI). Hence, an automated tool, such as data observability is needed.

Another challenge in data concerns the lack of ownership of data. As data journeys from the producer and goes through various transformation steps, it is unclear who owns the data. Earlier in this section, it is a cinch to spin up Snowflake compute resources, the lack of ownership leads to a lack of accountability.

Finally, cloud providers allow alerts to be set up, but sometimes by the time the administrator can act, the damage is done, and the high cost of the erroneous query is now on our books. Hence, a proactive approach is needed, which is what DataOps provides.

DataFinOps Case Studies

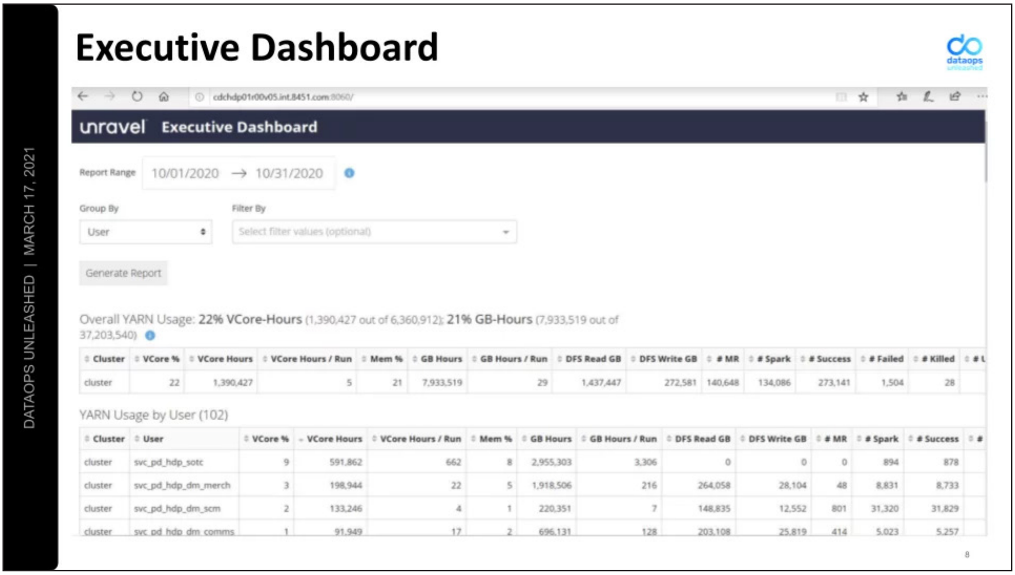
In this section, we will look at some successful customer case studies where the DataFinOps capabilities have delivered tangible value to the business. Let's first create a baseline as to what should an organization expect their DataFinOps to uncover:

- What parts of the pipeline are experiencing costs higher than the historical trends?
- What will be the impact of disrupting the job causing runaway cost?
- What are the cost trends, and when will we exceed the budget?
- Will the cost saving offset the cost of procuring and running the data observability tool?
- Will it provide a fine-grained cost breakdown into jobs and their consumption?

The first case study is for the US's largest supermarket retailer, Kroger. It has a retail data science wholly owned subsidiary called 84.51 (named after the longitude of its headquarter city, Cincinnati, Ohio). This organization provides retail analytics to its parent company, suppliers, and partners. One of its tasks was optimizing spiraling IT costs, consisting of an on-premises Hadoop ecosystem and Oracle Exadata. Using the newly gained DataFinOps capabilities, 84.51 has migrated some of its workload into the cloud. They use a combination of Google Cloud, Snowflake and Databricks for the persistence of data.

84.51 utilized Unravel Data's data observability suite to profile and identify inefficiently tuned Spark configurations. After initial success at pinpointing misconfigured clusters causing bottlenecks, the organization has now integrated Unravel into its development process. According to Jeff Lambert, VP of Data Solutions, *a key challenge* is ensuring these tools' adoption. His team had to ensure that DataFinOps capabilities were embedded into the developer ecosystem and that the notifications were handled promptly.

The figure below shows a snapshot of the executive dashboard that acts as a single pane of glass to display key metrics. According to the *case study*, it cut down aggregating workloads to just 10% of the original effort.


DATAOPS UNLEASHED | MARCH 17, 2021
dataops unleashed
cdchdp01r00v05.int.8451.com:8060/
unravel

Executive Dashboard

Executive Dashboard

Report Range 10/01/2020 → 10/31/2020

Group By: User

Filter By: Select filter values (optional)

Overall YARN Usage: 22% VCore-Hours (1,390,427 out of 6,360,912); 21% GB-Hours (7,933,519 out of 37,203,540)

Cluster	VCore %	VCore Hours	VCore Hours / Run	Mem %	GB Hours	GB Hours / Run	DFS Read GB	DFS Write GB	# MR	# Spark	# Success	# Failed	# Killed	#
cluster	22	1,390,427	5	21	7,933,519	29	1,437,447	272,581	140,648	134,086	273,141	1,504	28	

YARN Usage by User (102)

Cluster	User	VCore %	VCore Hours	VCore Hours / Run	Mem %	GB Hours	GB Hours / Run	DFS Read GB	DFS Write GB	# MR	# Spark	# Success	#
cluster	svc_pd_hdp_sotc	9	591,862	662	8	2,955,303	3,306	0	0	0	894	878	
cluster	svc_pd_hdp_dm_merch	3	198,944	22	5	1,918,506	216	264,058	28,104	48	8,831	8,733	
cluster	svc_pd_hdp_dm_scm	2	133,246	4	1	220,351	7	148,835	12,552	801	31,320	31,829	
cluster	svc_pd_hdp_dm_comms	1	91,949	17	2	696,131	128	203,108	25,819	414	5,023	5,257	

8

84.51 Data Observability dashboard powered by Unravel

Unravel Data is a data management platform that helps companies optimize their big data applications. Helped by Unravel Data, Kroger could identify performance bottlenecks in their big data applications and reduce infrastructure costs by 84%. Additionally, Kroger improved application efficiency by 51%, increasing customer satisfaction and revenue.

The second case study examines the DataFinOps features in the data observatory platform from *Acceldata*. This platform creates a spend analysis chart that displays cost consumption across services, accounts and workloads. Tags are used to categorize the data so that departmental chargebacks can be accurately calculated. The product provides a single pane of glass to see consolidated costs across all cloud and on-premises resources at individual and aggregate levels and to facilitate financial reporting to stakeholders.

Other capabilities of Acceldata are budget and recommendations. The platform provides tools to create and track budgets and generate data usage reports. Through inferred usage patterns, it also provides cost-saving recommendations.

Introducing DataBizOps

As seen in the first diagram in this document, the purpose of DataBizOps (and DataFinOps) is to demonstrate value from our data assets. It is a set of metrics that help in calculating productivity and reducing cost. For example, an organization may have built hundreds of data artifacts - reports, dashboard, views etc. However, by analyzing their usage, businesses can retire the whole set of processes leading to unused artifacts and saving costs.

The newest use case of data observability, DataBizOps, is often related to the rise in creating data products. "Data as a product" was introduced as one of the four principles of data mesh, while *Data products* turns the principle into business-outcome-driven consumable entities like a report, a dashboard, a table, a view, an ML model or a metric. DataBizOps gathers "data telemetry", like the frequency of data product releases, usage of data products and anomalies, which can lead to other indicators, such as poor data quality.

Data products take a business-first approach instead of the prevailing technology-centric bent of producing data artifacts. This approach can help mitigate challenges and frustrations of the modern data stack. For example, delivering outcomes involve a tedious and complex pipeline collection that leads to increased cost, effort and time. These hard-to-debug processes cause reliability and downtime issues, requiring data observability. Another example is the poor adoption of data catalogs, as we start by trying to collect and tag all the data available. This 'boil the ocean' approach often fails and reduces trust in the data governance initiative. If we catalog the data for the data products being built, we will have a better chance of a successful data governance outcome.

So, what is the role of DataBizOps?

DataBizOps can help establish data's ROI through metrics, like the number of data-related goals and objectives met and the level of data consumer and management buy-in. It can help support the data strategy within the organization. The best part is that the business strategy drives the data strategy. They can be in perfect alignment, maximizing the potential of corporate data assets.

An organization can benefit from DataBizOps:

- Launch a data marketplace and collect metrics to ensure that the data is being used for the intended purpose, by the intended consumers, and within the established guardrails. DataBizOps can augment the data producers' data governance efforts by extending them to data shares and exchanges.

- Strengthen the DataOps processes by providing the telemetry so intelligent decisions can be made for automation, testing and orchestration. Today, many orchestrators are rule-based and act like a complex collection of CASE statements. However, DataBizOps can feed contextual information, so the orchestrators can assess the next steps.

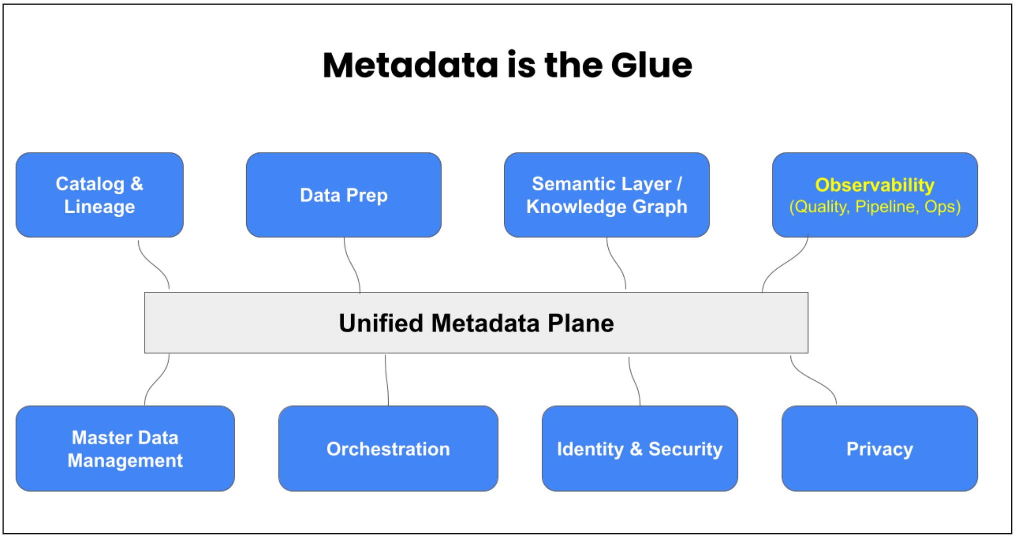
This is an overview of the potential of DataBizOps capabilities of data observability. Although some vendors are developing standalone products to provide the above-mentioned features, it should be part of the comprehensive data observability product.

Summary

DataFinOps (Financial Operations) refers to the set of practices, processes, and technologies used to assess and manage costs of the organizations' data assets. It aims to provide a streamlined and efficient approach to managing the data lifecycle, from collection and storage to analysis and decision-making. As the case studies reflect, it helps reduce costs and improve efficiency.

DataBizOps is the newest member of the data observability space. It has the potential to help establish data's ROI finally.

As we end the chapter, various use cases of metadata should be unified into a common metadata plane. These use cases are depicted in the figure below.



Metadata is the key to the success of data and analytics projects. This book has examined the observability capabilities. Organizations will benefit from a unified approach to handling metadata.

DataFinOps and DataBizOps metrics can help integrate the overall metadata management initiatives so our assets have high quality, reliability and security, and developers are more productive. A comprehensive metrics-driven metadata management plane can reform a data engineer's role from being reactive to becoming proactive. In addition, the data teams integrate into the rest of the business and deliver on the strategic imperatives. This will help with mainstream data observability adoption in an environment where data, data, and analytics workloads are constantly exploding.