WHITE PAPER

The DataOps Revolution

INFOSTRUX

DATAOPS IS AN ABSOLUTE MUST FOR ORGANIZATIONS LOOKING TO BECOME MORE DATA-DRIVEN.

What is DataOps?

One way to define DataOps is through a technical definition: a way to manage your entire data infrastructure through code. This data automation includes schemas, data, testing, and all the orchestration around them in an easily manageable, fully auditable package, including governance.

However, another way to look at DataOps is through a lens of culturally-focused transformation.

It’s about democratizing data and using agile, collaborative methods to increase data usage while making it more reliable.

This is important because much of the big data projects’ problems are due to bad data, and the problem is so widespread.

> “DataOps is a data management method that emphasizes communication, collaboration, integration, automation, and measurement of cooperation between data engineers, data scientists and other data professionals.”
>
> Andy Palmer—the person who popularized the term.

C.A.L.M.S.

There's a DevOps Framework, C.A.L.M.S., that we can use as a point of reference for what good DataOps looks like.

DEVOPS	DATAOPS
Culture	Culture
• Focus on people • Embrace change and experimentation	• Open data, reaching everyone • Embrace change and experimentation
Automation	Automation
• Continuous delivery • Infrastructure as code	• Automated certified data sets • Environment management through Infrastructure as code
Lean	Lean
• Focus on producing value for the end user • Small batch sizes	• Focus on high-value data sets and features • Small batch sizes
Measurement	Measurement
• Measure everything • Show the improvement	• Reliable data • Govern changes • Secure changes
Sharing	Sharing
• Collaboration and communication • Open information sharing	• Collaboration and self-service

As we review each of these points as they relate to DataOps, you will notice a pattern grounded in the Agile Manifesto. These principles reinforce the modern DataOps movement.

Culture

Culture is an essential aspect of the Agile Manifesto, and so it continues in DataOps. The most crucial element of DataOps is making data openly available to people. Open data does not mean open-sourcing all data. It means having readily accessible data to the data users, for example, a corporation having data readily accessible to all its employees.

This level of access to data enables change and experimentation that is so important in fostering a culture of data. You cannot achieve a culture of data without DataOps.

> **THE MOST CRUCIAL ELEMENT OF DATAOPS IS MAKING DATA OPENLY AVAILABLE TO PEOPLE.**

Automation

For the longest time, data was considered a second-class citizen in the automation world, but things have changed.

So much can be said about this topic, but let's consider the basics. DataOps requires that you manage all environments and their dependencies using automation.

Things such as Infrastructure as Code, CI/CD deployments, self-documentation via code, and automated testing for your data and transformation pipelines.

This high degree of automation grants the ability to launch net new environments with the push of a button, including data schema creation and data generation/ restore.

DATA PRODUCTS FOR AUTOMATION

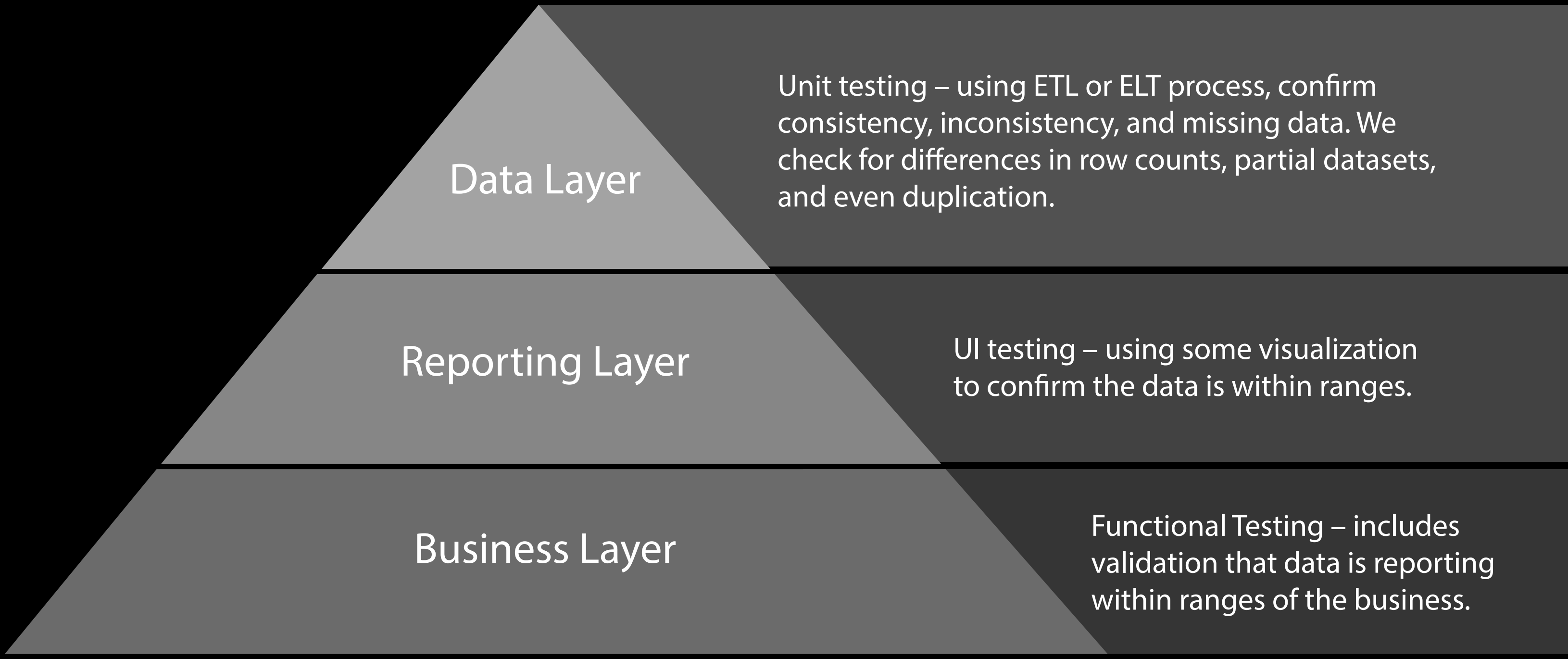
Here are some data-specific products that help you accomplish automation. Some are open source, and some are paid products depending on your budget or needs.

- Data Build Tool (dbt)
- Apache Airflow
- Airbyte
- Streamsets
- Fivetran
- Atlan
- RightData
- DataKitchen
- Prefect

TESTING PYRAMID

DataOps should follow the testing pyramid and use unit testing, functional testing, and UI testing.

From a data perspective, tests run during builds and deployments should validate all the data objects, schemas, and views.



Data Layer

Unit testing – using ETL or ELT process, confirm consistency, inconsistency, and missing data. We check for differences in row counts, partial datasets, and even duplication.

Reporting Layer

UI testing – using some visualization to confirm the data is within ranges.

Business Layer

Functional Testing – includes validation that data is reporting within ranges of the business.

Lean

Many data projects fail because the scope is too large. Smaller scopes tend to deliver results faster and with better quality.

DataOps following an Agile methodology is a must. Pick your preferable framework; Scrum, Kanban, or Scrumban. Most importantly, focus on high-value items and delivering quick and immediate business results and value.

It is pointless to spend three years delivering a project that defines all data modelling well, figures out all security and data infrastructure, but fails to deliver value for the organization.

Measurement

Measurement in DataOps touches on governance in an agile manner, which means tracking changes and keeping the data and data pipeline reliable.

We often see companies making strategic business decisions based on unreliable data.

Proper governance gives you the chance to make sure everyone in your organization trusts your data.

While the list of things you should monitor is long, you can use this list below as a guideline. Remember to implement an SLO, and it is best to use concrete metrics to measure success.

For example, you can set a target report load time of 10 seconds and measure the success of this metric.

- □ Data Transformation error rates
- □ Accuracy or ratio of data to errors is a good indicator to measure
- □ Completeness, for example, keep track of empty values
- □ Consistency, as confirming that various records from different data sets match as they should
- □ For all these above, measure the processes over time, using ML-based tools to predict anomalies in the ingest and transformation processes.
- □ In general, monitor all infrastructure (cloud or on-premise, including PaaS and Saas).
- □ Keep track of your costs so you can adapt things to save money and detect if someone’s use case changed.

Sharing

The sharing principle in DataOps means self-service and the ability to collaborate over the data while maintaining the data’s governance.

Accessible and understandable data empowers data scientists and analysts and the sharing principle in this framework provides that lift.

Another important aspect of sharing is continuous improvement. As you create a culture of transparency through open sharing, your data program will continue to improve with retrospectives - lessons learned and things to avoid or prevent in the future.