DevOps – DataOps

The ultimate guide to DataOps

Saagie

Summary

Not so long ago, data processing was a function of the IT department. Big Data was not so big and analyzing it actually meant analyzing structured data. As for data storage, data was fed to and initially resided in databases, and then data warehouses or server farms.

The Big Data landscape changed around 2010 with the invention of Data Lakes. Tools allowed us to extract data no matter its origin, and sources multiplied. Unstructured data (images, texts, audio…) could be processed in larger quantities, which is what we now call Big Data.

With the substantial amount of data being produced, new technologies and new profiles such as Data Scientists and Data Engineers, emerged and led to the creation of the first Data Labs. These changes brought freedom and independence to the teams but deployment remained a central issue because Analytics ecosystems do not share the same IT delivery criteria as software development frameworks. Converting an idea into a Proof of Concept became possible, but getting it deployed was still problematic.

Today the reality is that only half of AI projects have been deployed today. Why? Because it often takes longer than planned — between 12 and 18 months as noted by Gartner, Capgemini, or BCG. The main challenge is industrialisation. Let’s take a look at the past and what could be improved to help make data projects come to life.

Data & Analytics Project Challenges

Today, businesses planning on launching Big Data & Analytics initiatives face harsh realities:

- **Governance**: 28% of project stakeholders show inability to adequately secure or govern the data or analytical inputs and outputs from their initiatives
- **Time to market**: It takes a long time to implement — from 12 to 18 months to build and deploy AI pilots.
- **Value delivering**: only 27% of CxO consider their Big Data projects valuable; 38% of those projects show inability to demonstrate return on investment

In this context, the challenge for such organizations is to operationalize Big Data & AI projects to deliver value and prove the profitability of these initiatives. It all boils down to three distinct issues found in most companies: people, technology, and process.

People

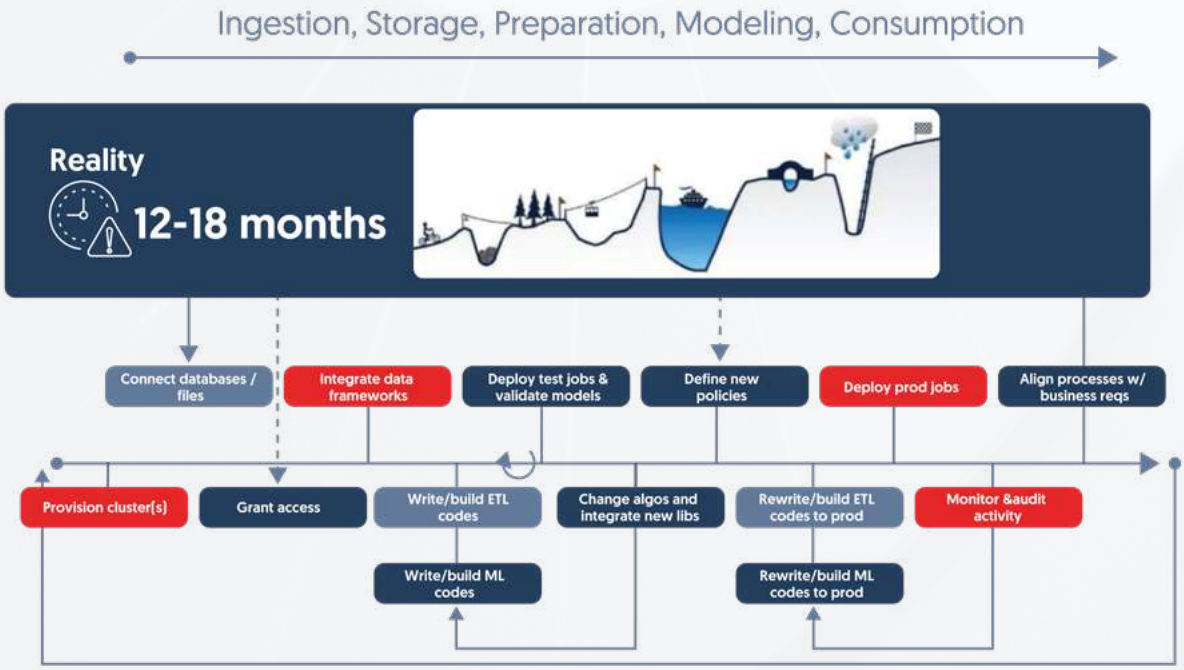
IT, Data, and Business teams are under immense stress when it comes to conducting and deploying data projects. These teams do not share the same expertise, objectives, or working methods, which prevents them from being aligned and working together productively.

Business teams want to be proactive in order to seize business opportunities quickly, data teams need to continuously upgrade their tools to better respond to business needs, and while IT teams require a robust and secure infrastructure. As these needs and aspirations clash, data project initiatives often fail to gain internal traction.

Process

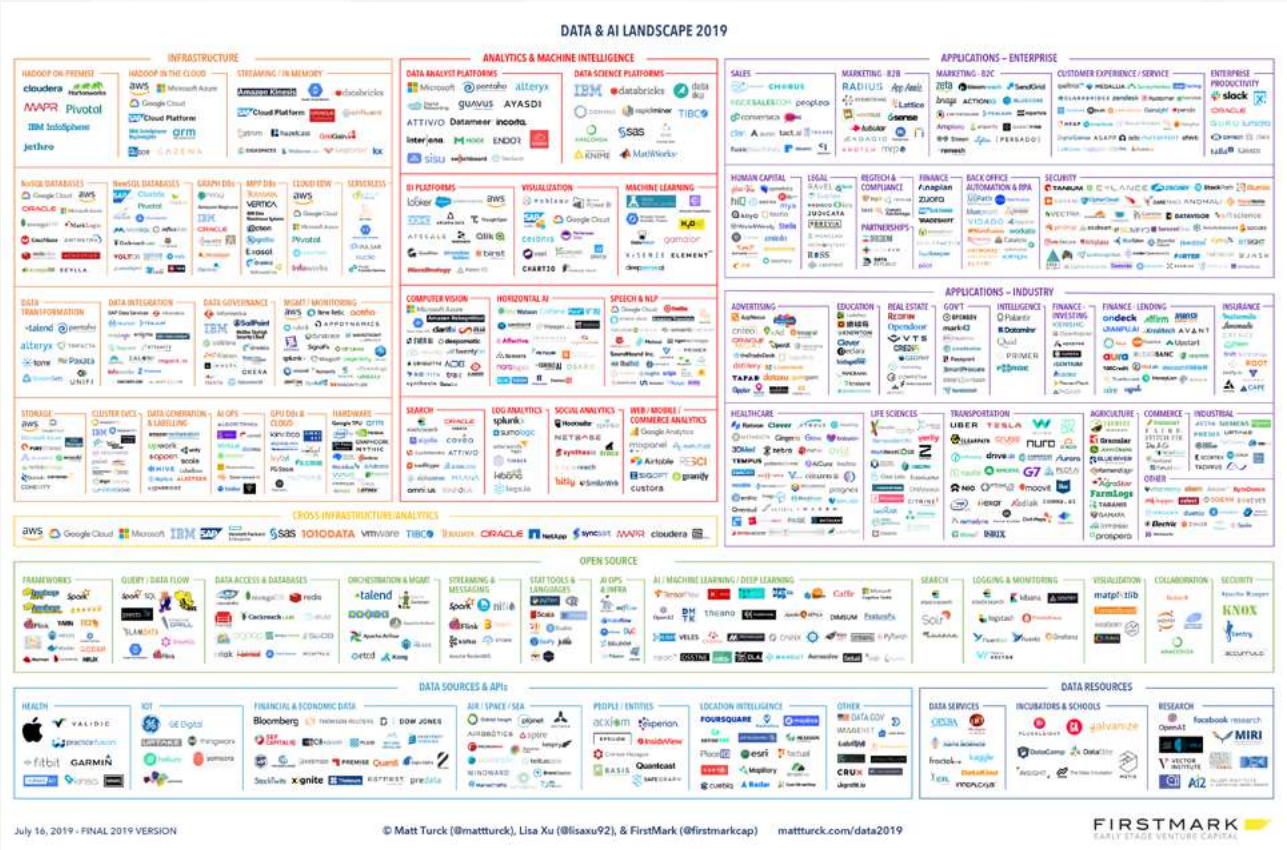
Companies tend to overlook the process requirements necessary to equip teams, write code/scripts and bring them from test to production environments.

This process covers diverse activities from provisioning servers, integrating frameworks and libraries, writing code in various languages, and collecting and integrating feedback from business users. These steps can collectively take more than a year and offer very limited reproducibility for later projects. It becomes a long, risky, and costly endeavor that creates tension between teams.



Technology

Today, the AI and Big Data ecosystem is massive (see Matt Turck's latest landscape above). Countless technologies exist, new ones appear every day, and others fall off the map before even getting started. On top of that, established frameworks are constantly updated by their developers. In the context of complete data projects, maintaining such an evolving and heterogeneous technological stack becomes very complex. As an effect, making technological choices often translates to risks — both in terms of investment and IT infrastructure.



Addressing these three challenges may involve numerous changes in terms of organization, processes and technological choices. DataOps, a practice identified by Gartner as a game changer in their last Hype Cycle report, seems to be the obvious option to make data projects come true.

What is DataOps?

This new approach is getting more and more attention as it seems to answer the numerous challenges we mentioned above.

Gartner defines DataOps as “a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers (Data Engineers, Data Architects, Data Stewards) and data consumers (Data Scientists, Business Analysts, Business teams) across an organization.”

At Saagie, we define DataOps as an organizational & technological framework derived from DevOps, which aims to bring agility, automation and control between the different data project stakeholders, including the IT team (IT Ops managers, application developers, architects), the Analytics team (data product owners, data scientists/ engineers, data stewards) and the Business team. DataOps operationalizes analytical workflows by leveraging the large, foreverchanging big-data ecosystem and the skills of all data practitioners.

Its main pillars are **flexibility** (DevOps approach, agile methods, repeatability), **governance** (monitoring, process control, security management), and **orchestration** (conditional pipelines, batch/streaming, containerization, auto-scaling/healing, advanced load balancing).

DataOps aims to improve and optimize the Data & Analytics lifecycle in terms of speed and quality.

DataOps uses technology to automate conception, deployment and management of data deliveries. It serves as a technological orchestrator for data projects.

DataOps Main Principles

Although it has only been used rather recently, it is based on two well-known approaches:

Agile

Data Agility is about setting up use cases that can quickly be deployed to reinforce confidence within teams and demonstrate value.

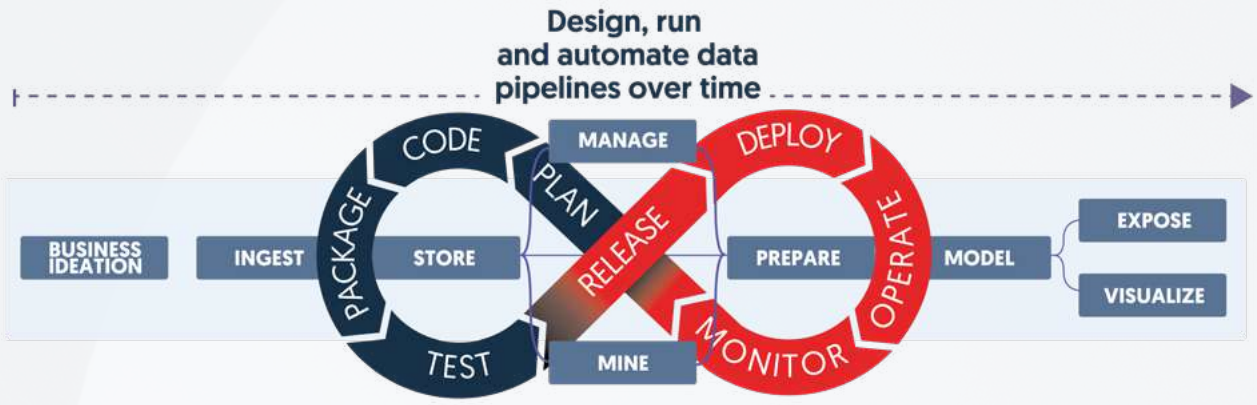
These practices favor communication and collaboration between teams, allowing quicker project deployment and reduced costs.

DevOps

DevOps is based on two main concepts: Continuous Integration (CI) and Continuous Delivery (CD):

- **Continuous Integration** consists of building, integrating and testing new code in a repeated and automated way. It allows to quickly identify and solve potential issues.
- **Continuous Deployment** automates software delivery. As soon as an app has gone through every step of qualification testing, DevOps allows it to go to production.

To put it simply, the DevOps approach ensures the alignment of the development and operations teams to automate every step of the software creation cycle, from its development and deployment to management.



A few practices are common to DevOps and DataOps:

- Automation (CI/CD)
- Unit tests
- Environment management
- Versions management
- Monitoring

DataOps is a mix of both, but harder to set up as it is applied to data as well as software development, meaning the Data and IT teams don't typically work for the same department.

What are the Main Profiles Involved in DataOps?

IT team

IT Ops Manager	
	Maintaining and monitoring a safe, sustainable, and available IT stack (hardware & software) composed of different environments

IT Architects	
	Designing IT architectures and networks, setting up databases and creating data flows

Software Developers	
	Writing structured code to build software aimed at data consumers

Data team

Data Product Owners	
	Designing data products and facilitating collaboration between data and IT teams to satisfy business needs

Data Engineers	
	Creating fast data flows / ETL workloads, preparing data samples for Data Scientists and helping them

Data Scientists	
	Transforming insights into business value by writing algorithms for preprocessing, feature engineering & Machine Learning

Business team

Data Stewards	
	Ensuring data quality across the organization by identifying data owners and documenting metadata

Business Analysts	
	Sorting and processing data to produce data visualization dashboards for business experts

Business Experts	
	Leveraging data to create business value in their respective activities: marketing, finance, logistics, engineering, HR, etc

How to Implement DataOps

A common assumption is that technology alone can solve data and analytics project issues, as if implementing effective tools could increase value creation by itself. While it is true that using the right tools helps immensely, the real solution is multifaceted and involves both cultural and process changes.

Regarding culture, DataOps involves a shift in mindset where cross-team collaboration must become natural to all stakeholders. Naturally, only an effective alignment between departments can help obtain successful long-term results. This agile culture doesn't operate on its own; it requires the implementation of new processes as well as changes in management to support data transformation and adoption. In addition, a transverse governance layer is crucial to ensure data quality and security across the entire company.

Data Processing Challenges

While DevOps offers automation and agility, it has limitations when it comes to creating applications that are meant to process data in real time. Data and Analytics projects require building and maintaining data pipelines (or data flows). **A data pipeline represents a data flow from its conception to its consumption.**

Data enters one end of the pipeline, goes through numerous preparation and processing steps, and exits as models, reports and dashboards. This pipeline is the "Ops" aspect of data analytics.

Other differences that come from Data Science projects requirements:

- Results repeatability
- Model Performances Monitoring
- Models Exposition

DataOps Orchestrator

> *We often hear 'data is the new oil' but as important as data may be, the key is what you do with it that makes it valuable. What you need is a refinery, because people don't want oil, they want gas.*

Adrien Blind, DataOps Evangelist - Saagie

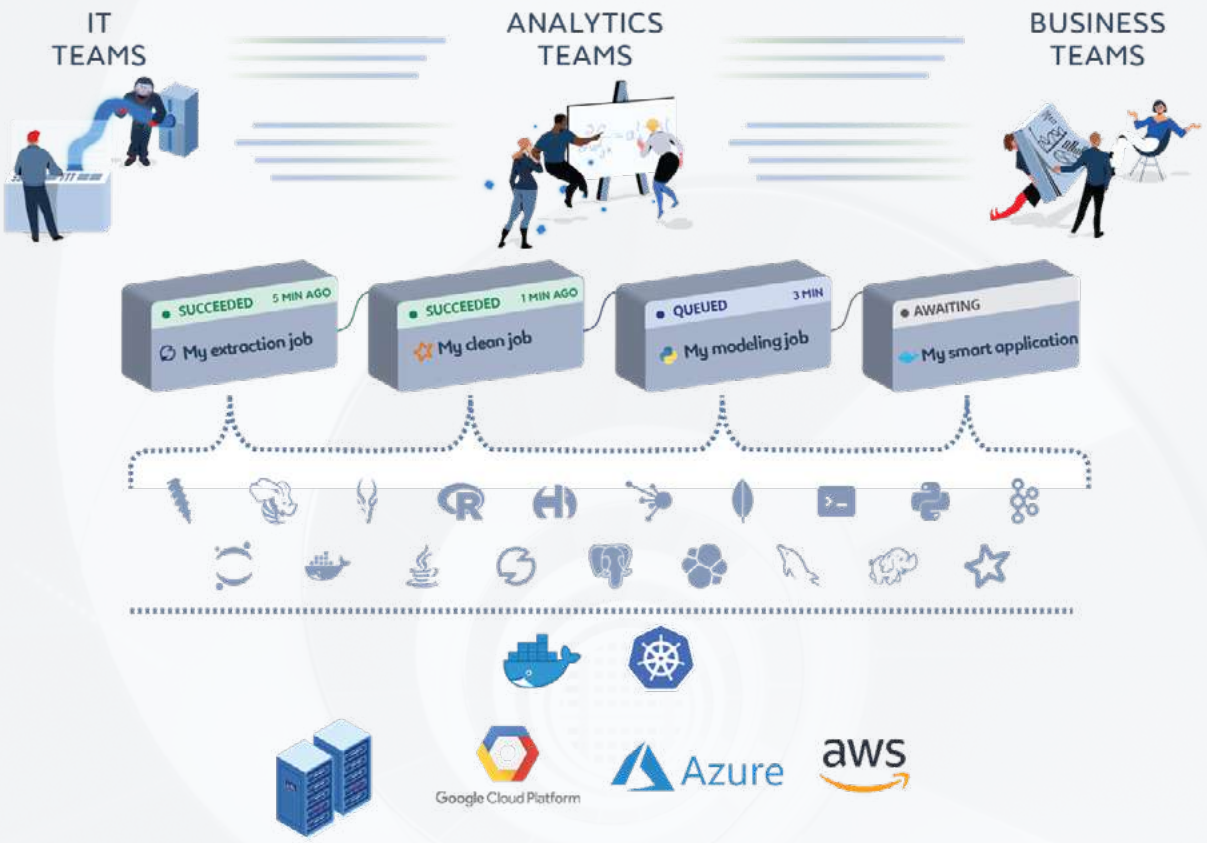
What if the "refinery" was the combination of an approach, DataOps, and a technology? A tool we call a DataOps orchestrator. It will help you:

- Manage data from extraction to consumption, including storage, preparation, processing, and visualization.
- Ease and accelerate Data & Analytics projects deployment because all the technologies you need are gathered, updated, and available in one location (Elasticsearch, PostgreSQL, Talend, Java, Scala, Jupyter, Docker, Mongo DB, and MySQL).
- Improve collaboration and communication within the company as every member of the team is involved and work on a unique centralized tool.

Orchestrator Features:

- It is a true technological toolbox that allows managing the whole data cycle (preparation, processing, valorization and visualization) while being operational (stable infrastructure)
- It doesn't only gather these technologies, it needs to organize them by tasks (preparation, processing...), so it needs to have an orchestration feature
- It has governance features so your team is allowed to manage data access in a secured way

- ergonomics and usability also matter as business teams are involved. Therefore, it needs to have a simple and intuitive interface that everyone can use.
-



Why is an orchestrator so valuable? Because having such a tool allows to manage processing data pipelines and deploy them automatically under a specific schedule from one environment to another.

In this context, Saagie provides an Orchestrator for DataOps that aims to bring agility and process automation to each step of the data value chain: ingestion, storage, preparation, modelization and sharing.

On the Data and Development fronts, Saagie orchestration offers code and infrastructure flexibility for data pipelines and the continuous improvement of Data Science & Engineering processes. Regarding Operations, Saagie allows process monitoring, version management and log centralization.

References

Gartner, CIO Survey, 2018

Capgemini & Informatica, The Big Data Payoff: Turning Big Data into Business Value, 2016

BCG, Putting Artificial Intelligence to Work, September 2017

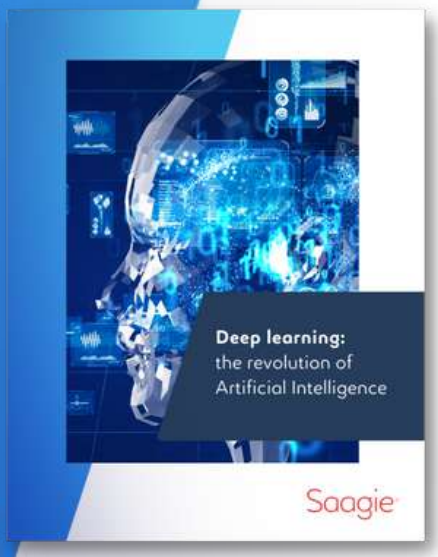
Gartner, Innovation Insight for DataOps, December 2018

Gartner, Market Share: all software markets, worldwide, 2018

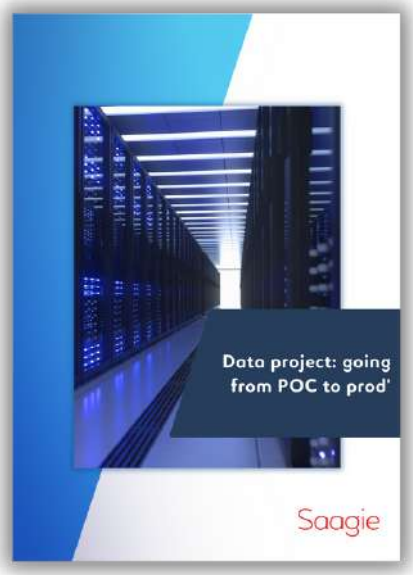
Saagie

Thank you !

To go further...



Deep Learning



Data project