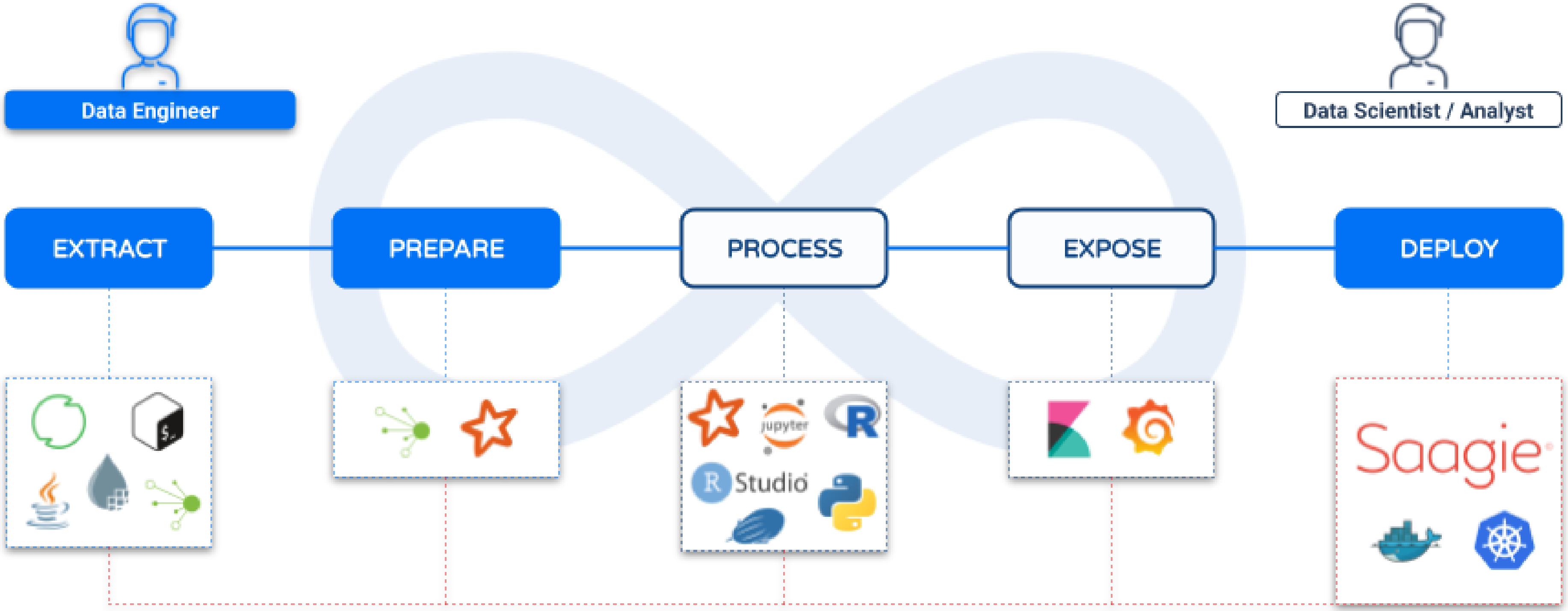
Saagie©

DataOps Platform

to easily, quickly and reliably
deliver data projects

Saagie | Positioning

Saagie is a DataOps Platform for Data Engineers



Why empower Data Engineers?

Data scientists, engineers, analysts... People might think building a data team - or data lab - consists in choosing people who have “data” in their job title, starting with the data scientist. But if you were to start your own F1 racing team, would you hire the pilot first? Of course not, you would start by hiring the man who will design the best car: that’s your data engineer. **Saagie provides data engineers with the means to easily, quickly and reliably deploy any data project into production**, the only thing left for them to do is to deliver.

Why choose our Platform?

Gartner stated that it takes between 12 and 18 months to build a data platform from scratch. Your teams have their own technological habits. Do not waste time in looking for the perfect tools in the ever-changing market of data technologies to fit both their preferences and your IT requirements. **Our DataOps platform is open to fit your architecture, fully managed and secured**. Above all, Saagie integrates and orchestrates **the best of open-source and commercial worlds** so your teams can start right away. Take a pick.

Why implement DataOps?

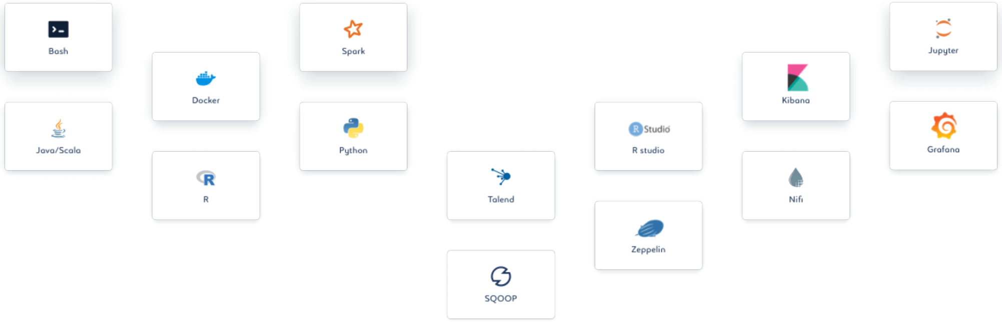
Operating a data project means lots of iterations between teams and more juggling between different technologies and their versions. That’s when you need DataOps: already announced as the DevOps successor, **DataOps is an organizational & technological approach** to deliver data projects. Take back control with our DataOps platform: **build and run automated pipelines and enjoy multiple environments to explore, test and deploy**. Saagie allows you to manage the entire data lifecycle while meeting production, security and traceability criteria.

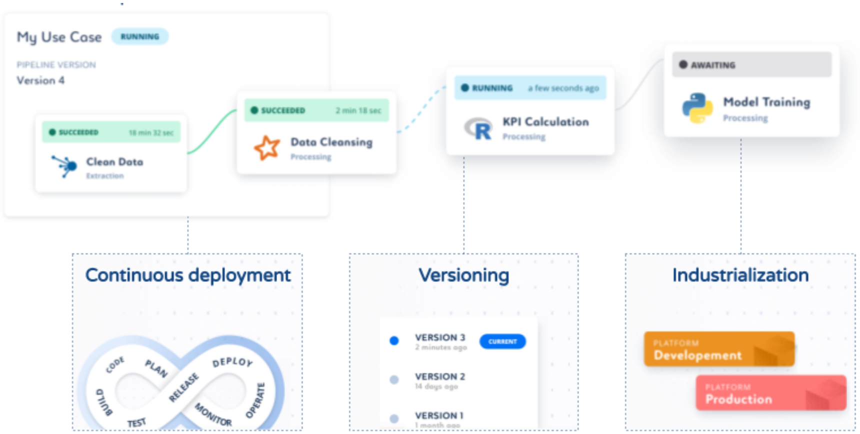
Saagie® | Value Proposition

A DataOps platform to deliver and run data projects easily, quickly and reliably

Make it easier

Building your own technological tool may sound nice: Ops can have their hands on it, Data teams can customize it as they wish. But once you have, it is very hard to manage, maitain and update. We provide a ready-to-use platform that combines and orchestrates the data market leading technologies to offer you a single point of entry to the best of the ecosystem. It is also managed so you can focus on creating business value instead of worrying on how to make it work or maintain. Easy.





Make it quicker

Saagie is production-ready by design so your project does not fall apart when you try to deploy it in production. You create a project, choose your favorite technologies, create data jobs and make it run into automated pipelines. You can replicate those jobs and pipelines and promote them from an environment to the other. To make it sound simpler, we provide you with the means to build an automated production line that you can reuse whenever you need. Quick.

Make it reliable

We provide you with the technologies to manage the entire data lifecycle and the features to monitor and trace every step of the way. You can isolate projects, manage accesses, monitor status and store and centralize logs to oversee activity. You can leverage a robust platform that simplifies configuration and maintenance to operationalize your many use cases. Once your projects are up and running, Ops can track everything to constantly improve both performance and security. Reliable.

Saagie | What is DataOps?

Over the past two years, **only 53% of POCs have made it beyond the lab into production**, taking an average of **9 months**

Gartner, 2020

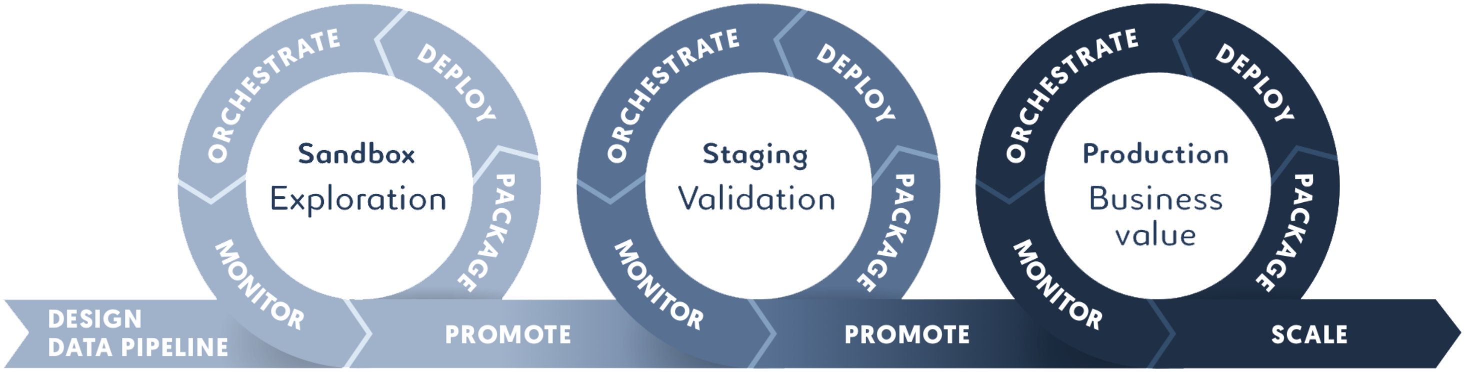
Three teams are involved in data projects and often find themselves **working in silos**. The business team needs quick results, the data team craves for agility, while the Ops need control over their architecture. **Aligning those teams and goals** can be challenging.

Project stakeholders tend to oversee processes. **1 out of 3 companies states that the lack of DevOps skills caused their project to fail.** To be able to quickly deploy, companies need to implement automation and monitoring processes.

Data teams need ready-to-use and orchestrated tools to manage the entire data lifecycle: extraction preparation, processing and visualization. **45% of companies point out their technological choice when asked what prevent their project to be deployed.**

Data projects are dynamic processes continuously evolving as data comes in and more and more people get involved using different technologies.

DataOps, often described as the **DevOps approach applied to data projects**, is a set of agile and collaborative practices aiming to bring automation and control in data projects delivery.



DevOps is based on two main concepts:

- **Continuous Integration** consists of building, integrating, and testing new code in a repeated and automated way. It allows you 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.

The DevOps approach enables teams to **automate every step of the software creation cycle**, from its development and deployment, to its management.

While the approach offers automation and agility, **DevOps has limitations when it comes to creating applications that are meant to process data**. Data and Analytics projects require building and maintaining data pipelines (or data flows).

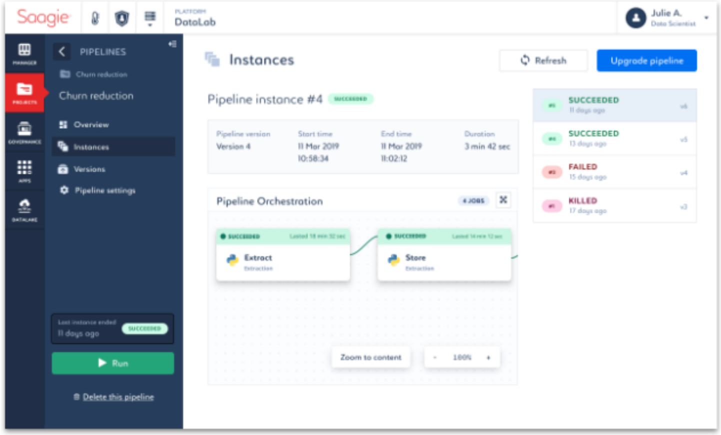
DataOps is an organizational & technological approach to deliver Data projects.

Thus, **DataOps operationalizes analytical workflows** by leveraging the large and various big data ecosystem and the skills of all data practitioners.

Its main pillars are:

- **Orchestration** through containerization, scale up / advanced load balancing, jobs and pipelines creation and scheduling, and batch / streaming modes.
- **Agility** through the Agile Methodologies and practices, jobs and pipelines replicability, versioning, rollback and portability from environments.
- **Control** through full traceability, process monitoring, centralized logs, network isolation and security management.

Saagie | DataOps Platform features



Instances

Pipeline instance #4 SUCCEEDED

Pipeline version: Version 4
Start time: 11 Mar 2019 10:58:34
End time: 11 Mar 2019 11:02:12
Duration: 3 min 42 sec

Pipeline Orchestration 4 JOBS

SUCCEEDED Extract
SUCCEEDED Store

#4 SUCCEEDED 11 days ago
#3 SUCCEEDED 13 days ago
#2 FAILED 15 days ago
#1 KILLED 17 days ago

Last instance ended 11 days ago SUCCEEDED

Agility for Data Teams

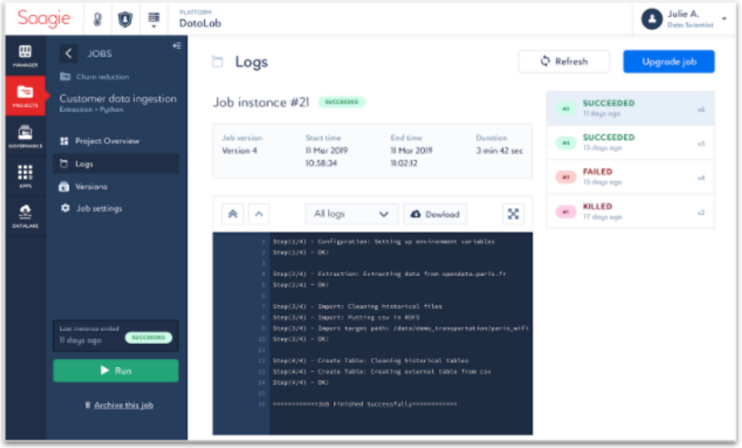
The DataOps Platform allows data teams to work in autonomy:

- Get started right away on a **pre-configured processing cluster** and focus on creating value;
- **Create projects** and assign users within your teams;
- Collaborate easily with teams thanks to **isolated and secured work environments** for each project;
- Create jobs to **manage the entire data lifecycle**, from data extraction to visualization;
- **Choose different technologies** for each job category (extraction, preparation, processing, visualization);
- Build and **automate multi-framework pipelines;**
- **Integrate Saagie into CI/CD pipelines** through plugins and APIs;
- **Make your jobs reproducible**: run processing jobs in exploration mode as well as in production;
- **Monitor** job statuses (by author or category), logs, version and instance history, as well as allocated resources.

Control for Ops

The DataOps Platform enables Ops to oversee every step:

- Get a pre-configured cluster with **automated maintenance and updates** using orchestration stantards: **Docker** & **Kubernetes;**
- Work on a **Kerberos**-compatible data lake;
- Manage rights to **control data access**:
  - User rights per group and profile;
  - Sentry rights;
  - Dataset protection (reading/writing).
- Isolate projects and **increase traceability:**
  - Centralized logs and history;
  - Possibility to route logs to your log management system (**ELK, Splunk, CloudWatch**);
  - Job and pipeline status notifications.
- Schedule and **monitor processes:**
  - Jobs scheduling inside Saagie or through an external scheduler (**$U, Control-M, Tivoli/Websphere Workload Scheduler**);
  - Status monitoring (via UI or API) activity overview.



Logs

Job instance #21 SUCCEEDED

Job version: Version 4
Start time: 11 Mar 2019 10:58:34
End time: 11 Mar 2019 11:02:12
Duration: 3 min 42 sec

```
Step(1/4) - Configuration: Setting up environment variables
Step(1/4) - OK!

Step(2/4) - Extraction: Extracting data from opendata.paris.fr
Step(2/4) - OK!

Step(3/4) - Import: Cleaning historical files
Step(3/4) - Import: Putting csv in HDFS
Step(3/4) - Import target path: /data/demo_transportation/paris_wifi
Step(3/4) - OK!

Step(4/4) - Create Table: Cleaning historical tables
Step(4/4) - Create Table: Creating external table from csv
Step(4/4) - OK!

=============Job Finished Successfully=============
```

#4 SUCCEEDED 11 days ago
#3 SUCCEEDED 13 days ago
#2 FAILED 15 days ago
#1 KILLED 17 days ago

Last instance ended 11 days ago SUCCEEDED

Saagie | Orchestrated technologies

Supported technologies

Python
2.7, 3.5, 3.6, 3.7

R
3.4.4, 3.5.3, 3.6.3

Java
7, 8, 11

Talend
8.121, 8.131

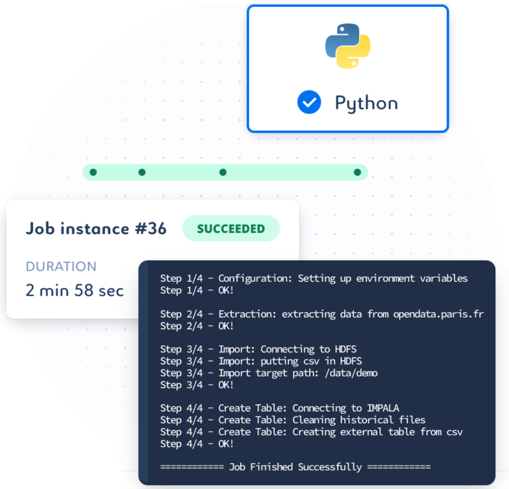
Spark
2.4 (Java: 8,11 – Python: 2.7, 3.5, 3.5, 3.7)

Bash
Debian 9, CLI AWS, CLI Azure, CLI GKE

Docker

Sqoop
1.4.6

Build and run automated data pipelines



Job instance #36 SUCCEEDED

DURATION
2 min 58 sec

```
Step 1/4 - Configuration: Setting up environment variables
Step 1/4 - OK!

Step 2/4 - Extraction: extracting data from opendata.paris.fr
Step 2/4 - OK!

Step 3/4 - Import: Connecting to HDFS
Step 3/4 - Import: putting csv in HDFS
Step 3/4 - Import target path: /data/demo
Step 3/4 - OK!

Step 4/4 - Create Table: Connecting to IMPALA
Step 4/4 - Create Table: Cleaning historical files
Step 4/4 - Create Table: Creating external table from csv
Step 4/4 - OK!

============ Job Finished Successfully ============
```

Choose your technologies

Start now by mixing ready-to-use frameworks and their various versions from the open-source and commercial worlds.

Make your jobs reproducible

Meet production criteria with our containerization capabilities and run processing jobs in exploration mode as well as in production.

Run your jobs in pipelines

Build workflows to run ETL, pre-processing and processing jobs and manage the entire data lifecycle through versioning and logs.

Apps Catalog

R Studio

Jupyter

Zeppelin

Nifi

Grafana

Kibana

External Technologies

Add your own technologies using the **Saagie Technology SDK** (software development kit). Package your technologies, apps, and their associated dependencies and libraries into Docker images to create your own execution contexts. Run jobs and apps directly on Saagie's Kubernetes cluster.