devoteam

January 2022

From data lakes to machine learning platforms:

The new challenges of cloud-based data

Creative tech for Better Change

About Devoteam

Devoteam is a leading consulting firm focused on digital strategy, tech platforms and cybersecurity.

By combining creativity, tech and data insights, we empower our customers to transform their business and unlock the future.

With 25 years' experience and more than 8,000 employees across Europe, the Middle East and Africa, Devoteam promotes responsible tech for people and works to create better change.

Creative tech for Better Change

The data market is evolving very fast.

Until recently, data engineering platforms were dedicated to processing information and routing data. Today, more than half of our customers are looking for a global platform which iintegrates data science layer with data lakes and data pipelines. The industrialisation of tools for data scientists and data engineers, to help them prepare models and put them into production, is now a strong market trend.

Data lakes are now an asset whose best practices are known and mastered. The situation is different for data science, where what might be considered state of the art is not yet set in stone. On these platforms, our work sits between R&D and industrialisation. Our aim is recognise best practices and apply them within such platforms alongside supporting our clients adopt. This is a transitional stage before further industrialisation.

In five years' time, machine learning operations (ML Ops) platforms will be deployed in an automated manner on AWS; practices will be standardised, and there will be no more "how to" unknowns.

In this ebook we bring you an overview of 2021's data challenges:

- Data lake best practices
- Emergence of ML Ops
- Feedback from our customer experience - Olympique de Marseille case study
- Anomaly detection use case
- Focus on the data scientist profession

Table of contents

1 The evolution of a data lake to Cloud services

A data lake is a complex product that is usually custom-built for each individual business need. In the customer feedback we present here, the issue was the replacement of an on-premise data lake based on the Cloudera solution. The reason for this change was recurring scalability problems and storage that had reached saturation point. In this section, we will review which Cloud services and which organisation were chosen to best meet their needs.

The nature of work within the context of a data lake project can often bring with it an additional complexity factor, due to the number of actors involved.

For this project, it was a joint effort between the traditional IT department and the team dedicated to innovation. The first need conveyed by the client was to start building the service from day one to help them achieve results quickly.

To address this challenge, we simultaneously began to launch the build phase and onboarding of IT department. The objective was to build a data lake with Dataiku, with no distributed computing capacity to keep the architecture simple, and then to switch the data lake and current uses to the AWS Cloud.

Project Context

After the implementation of the first version of the platform, the Devoteam A Cloud's mission was to take the technical lead and to accompany the data lake team (composed of internal and other external service providers), on to the next steps of the deployment.

Target Architecture

The chosen target architecture is centred around three storage zones, based on as many S3 buckets:

- Storage of raw, unmodified incoming data
- Storage of cleaned and transformed data
- Storage of refined data, for data lake users

The flows and processing between these spaces are done by AWS Lambda (python) and AWS Glue.

Use Case

The data processed by the data lake is intended both for internal use (by analysts and data scientists) and external use (clients and partners) via applications or portals. The data handled is personal, named and of a medical nature. Processing of such data raises serious security, compliance and traceability issues. To meet the GDPR and compliance constraints, some personal data has been anonymised: information has been deleted, precise information such as postcode or date of birth has been reduced, hashing implemented, etc.

The analysis of this data makes it possible to respond to two major types of use cases:

- customer churn, i.e. the preventive detection of customer loss
- the management and monitoring of absences.

The Challenge of a scalable solution

As Tony Phe, Cloud & DevOps Consultant, Devoteam A Cloud, explains:

"We have already implemented this kind of pipeline for other clients, but the infrastructure is never the same from one project to another. It's adapted to the context and the use cases, and this was the first time we were dealing with data as sensitive as France's Nominative Social Security Declaration (DSN).

Our challenge was twofold: to respond globally to the client's need to use the cloud to make economies of scale, and to very quickly offer new use cases.

To meet these expectations, we chose AWS Glue, which still brings certain constraints on use cases with high processing intensity. In this particular situation, using Glue allowed us to accelerate the implementation of the Cloud architecture. The challenge of this project was not so much about the hard skills, but more about the reflection on the implementation of the solution and its evolution over time."

AWS step functions

AWS Step Functions was chosen as the orchestrator for the data pipeline. AWS Step Functions offered the best solution to the challenge of moving the project forward very quickly despite the small size of the team at the beginning. As this choice made it possible to accelerate the project start, the team is currently studying the possibility of using an additional solution with Airflow: their objective is to adapt the infrastructure to the life cycle of the project and the company's current needs.

As Romain Pierlot, Delivery Manager, Devoteam A Cloud explains:

“It was a question of aggregating the bricks that were already operational from the platform, and to have a stabilised base to be able to migrate the existing use cases. We had to replace the on-premise data lake by migrating these use cases quickly, and develop new use cases on the cloud platform. The success of a cloud project also depends on internal adoption, so having ‘quick win’ use cases is essential.”

The second objective concerned a specific use case for an application used by field salespeople. It had to be deployed quickly with high performance constraints that led to ingenious architecture choices to meet the end-users’ business requirements. We built this use case in parallel with the data lake base.

A new nature of work with the lockdown

The challenge of building the data lake while creating a separate pipeline for a specific use case was met despite the lock down constraints.

A new process was put in place overnight:

- Migration of all day-to-day management and control tools,
- Replacement of Jira by Miro, which facilitated remote access and allowed for better task distribution and visibility.

This new organisation allowed us to meet the deadlines for delivering the platform and the first use cases. The goal now is to migrate more than 30 use cases, and to continue to strengthen and stabilise the platform, while ensuring it remains operational.

2 ML OPS

The industrialisation of data science is something that interests many companies these days. But it is not simply a technical topic: IT, business and organisational set-up must be addressed at the same time. In this chapter, we will introduce you to ML Ops, and the challenges of this approach.

A few decades ago, data science was a purely theoretical discipline, studied mainly in academia and research teams. Today, the business world invests in data science projects with the aim of creating value.

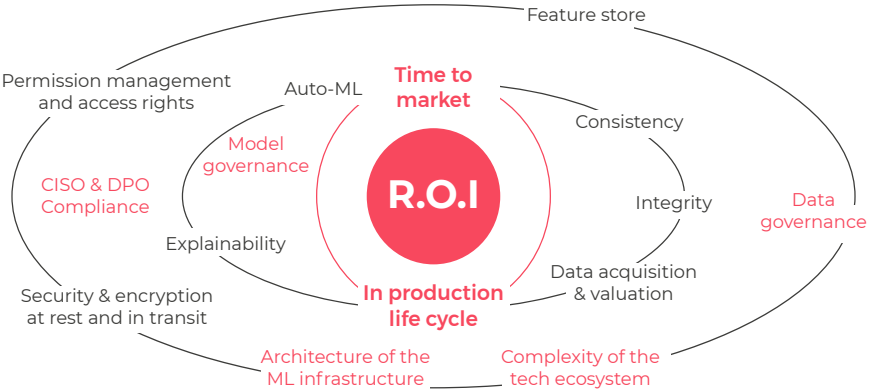
This requires a significant maturity in operational processes, which were built with traditional IT system requirements in mind, but unfortunately less suited for machine learning models.

50%

A study conducted by Algorithmia shows that half of the companies spend between 8 and 90 days to implement a single AI model into production. This considerable time is due to constraints which slow down the production process.

The ML Ops approach allows businesses to respond to these challenges in a more timely manner.

ML systems' pain points



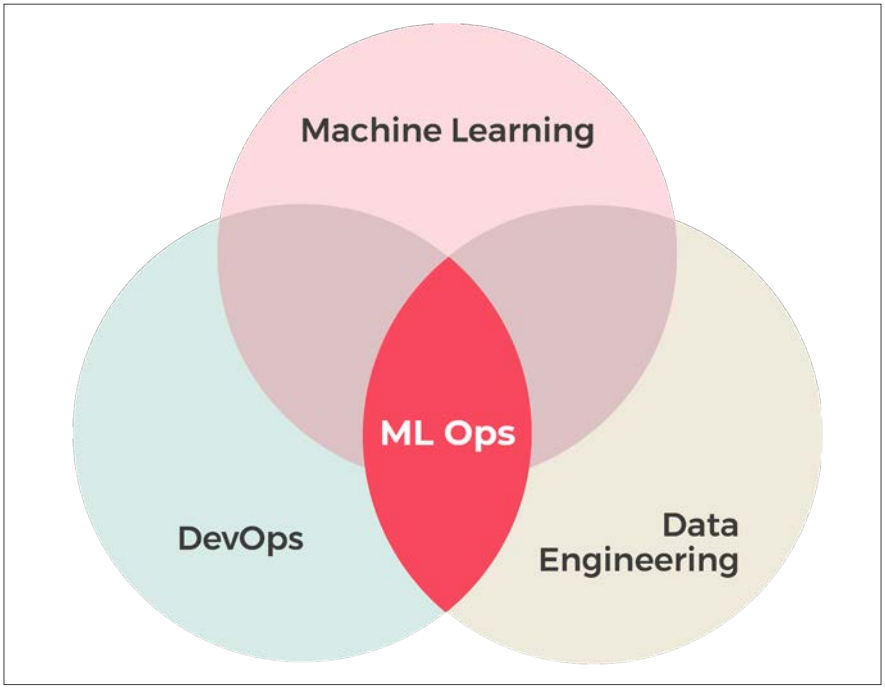
What is ML Ops ?

ML Ops is a recent discipline focused on ensuring collaboration and communication between data scientists and IT professionals (such as data engineers, infrastructure engineers and those in IT operations) to automate recurring tasks in the machine learning lifecycle.

Using a variety of tools and practices, ML Ops establishes a culture and environment in which ML technologies can generate the greatest value.

In this approach, the model is not considered the end product to be delivered to the operations team, but will be encapsulated along with other components that are essential to inference, such as data or code pipeline.

ML Ops at the intersection of several areas



Challenges of a ML approach

The challenges associated with the ML Ops approach are similar to those of DevOps. However, there are also some that are specific to ML, as shown in the table below.

Specific ML Features	Description
Data versioning, Model, Hyperparameters	Versioning the data Versioning the model Versioning hyperparameters
Trials	Feature Engineering Tracking experiments
Testing	Data quality tests Pre-processing tests
Monitoring	Continuous monitoring of health and performance Performance metrics of the model in production. Ressource consumption
Automation	Data transformation Model evaluation Model re-training Model selection/optimisation
Reproducibility	Revert to an older version of the model and it's inputs (training/test datasets, hyperparameters)
Auditability	Ensure data/model integrity Continue audit logs Ensure compliance with company policies and associated regulations
Scalability	Scaling the infrastructure that hosts the model to ensure the expected level of service


LIVE DATA
LIVE DATA
LIVE DATA

3 Client feedback: The Olympique de Marseille data lab

The French football club Olympique de Marseille (OM) launched a digital transformation project in January 2019, aiming to accelerate innovation and provide them with a new technological platform to achieve its sporting and business goals.

After taking over the commercial operation of the stadium, OM found that its IT infrastructure did not factor in either the challenges or potential usefulness of its data. The club envisioned how its new tech ecosystem could act as a catalyst for OM's business and sporting development, ensuring data was at the heart of this strategy.

The club spent 18 months redesigning the entire information system, from website, mobile application and accounting to CRM, commercial management and HR tools. At the same time, the scope for data as a development lever was promoted internally. The aim was to make quicker and better decisions, improve sales and find new opportunities.

The club built a data lake, designed as the foundation to build further projects. Furthermore, to foster innovation and identify new opportunities for their data, OM brought in external partner support and combined several existing data-related initiatives within an innovation laboratory; the Data Lab.

This lab allowed them to bring together partners, startups, engineering schools, researchers, companies such as AWS and Devoteam A Cloud to develop joint innovation projects. For example, optimising the filling of the stadium using ML models which leveraged on the stadium's historical sales data.

OM also had big expectations from Machine Learning and video detection to help them improve the way they identify young football talent. The club therefore grouped these internal initiatives within the lab to find partners to support these areas, with data as the common thread.

Rethinking the use of data

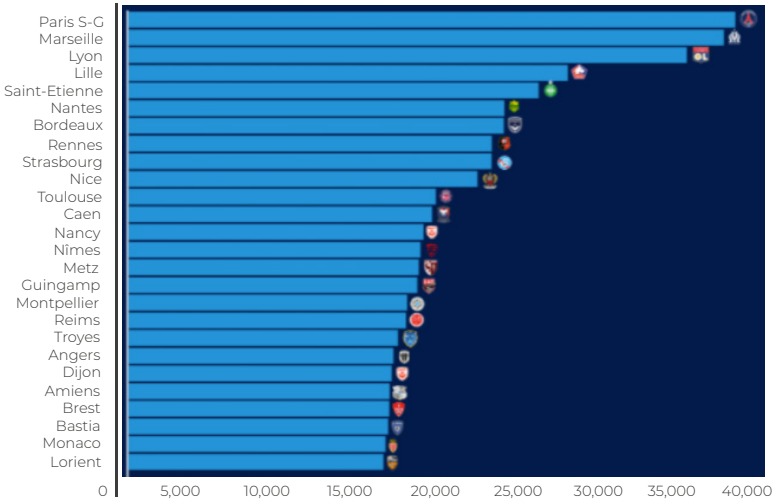
To capitalise on data's potential to help meet its marketing and sporting goals, OM developed a data-based strategy

- Build a data ecosystem to allow information to flow between the various databases
- Data storage and segmentation to forecast current and future needs
- Development of a genuine sports data strategy, around talent scouting, performance analysis, medical, etc.

Overall, the use of its data has been a success.

Average attendance per team (home and away)

1st league 2016-17 to 2019/20 excluding games played behind closed doors[1]



Attendance at a match depends on many factors, including the month, the day and time, and the clubs playing.

But what these figures do not tell us is the individual impact of each club on attendance levels, regardless of the opponent or context variables.

By controlling for additional variables such as stadium, day, time and month of the year, we can extract the individual impact of each club on stadium attendance

[1]**Data: LFP** https://infogram.com/etude-1-impact-om-data-lab-1hxj48yrl5152vg

The AWS cloud as a data support

As Frédéric Cozic, CTO of Olympique de Marseille, explains, the AWS Cloud was chosen to support this project.

“To deploy this kind of solution and ensure its sustainability is a real challenge. We quickly ruled out on-premise solutions. Not only did we not have the capacity for it but they required a lot of management and, above all, they risked limiting us technologically.

We therefore looked at the main players in the cloud market, but saw a higher level of technical maturity when it came to data, AI, and ML stacks on the AWS side, hence our choice of AWS Cloud. With Amazon SageMaker and managed services storage bricks for data lakes, AWS remains the most advanced player on these topics.

Data-related needs triggered our move to the Cloud: scalability, storage capabilities, the fact that we can store in the areas of our choice. The Cloud gives us the widest possible range to meet our needs. Beyond the data, we have a multi-year Go to Cloud project underway and also use the Cloud for our IT system.

Having the option of going to the Cloud in addition to the on-premise options allows us to move more quickly and to set up more scalable, advanced and secure IT architecture. It also helps to lower costs: in addition to run cost savings, it takes less time, so there is also a man power saving. Cloud storage is economically very advantageous, and a serverless approach is also very interesting. Our data lake uses many serverless workloads that take in data, then dispatch, classify and, clean it up, and most of these workloads are below the ‘free tier’ offer, or cost very little. Serverless data processing has many advantages compared to on-premise solutions.”

4 Use case: anomaly detection explained by football

It's Sunday morning, you're having your coffee and reading the paper as usual. As a football fan, you go to the sports page to see the latest match results and then something catches your eye (and you almost spill your coffee!) A second-division team has won the most competitive championship! How is this possible?

Spotting or discovering something unusual or strange enough to be noticed and give you a surprise is what we call anomaly detection (Cambridge Dictionary).

Anomaly detection is a technique that can be applied to different situations

(Dutta & Vallabhajosyula, 2017):

- **TELECOMMUNICATIONS:** Detection of roaming abuse, revenue fraud and service interruptions
- **BANKING:** Identifying abnormally high purchases/deposits and detecting cyber-intrusions
- **FINANCE AND INSURANCE:** Detect and prevent fraudulent spening patterns and travel expenses
- **HEALTH:** Detect fraud in claims and payments
- **INDUSTRY:** Detect abnormal machine behaviour to prevent increased production costs
- **SOCIAL NETWORKS:** Detect compromised accounts and bots that produce false reviews
- **NETWORK:** Detect network intrusions
- **SMART HOME:** Detect energy leaks
- **VIDEO SURVEILLANCE:** Detect or track objects and people

But how can you use this technique to identify atypical events in your favourite football team? For example, unexpected wins or losses, with unusual scores.

Before you start looking for data, it is important to understand another key concept for applying anomaly detection techniques. This concept is called “time series.”

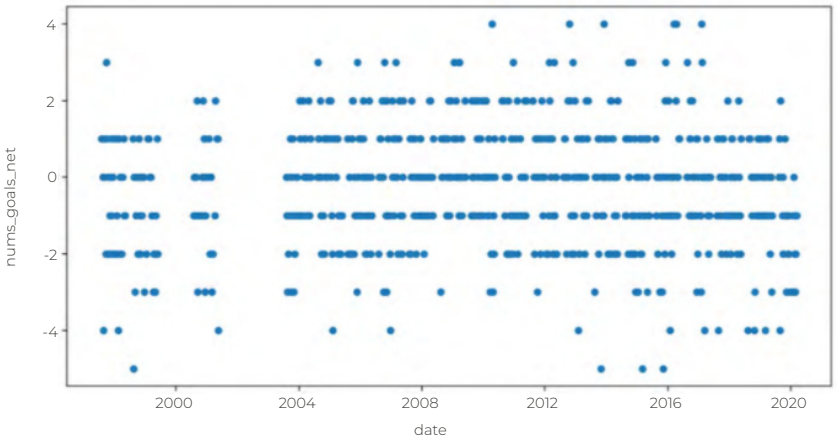
A time series “is a set of regular observations ordered in time (...) taken at successive, mostly equidistant periods/moments” (OECD.org, N.A.). In our football example, the time series data is the historical data of all matches and the number of goals scored either for or against one’s team.

By merging the anomaly detection and time series concepts, we could say that detecting time series anomalies is the identification of rare events that have distinctive characteristics compared with the majority of data processed over time (DeepAI.org, N.A.).

To return to our example, we could aim to identify – using this historical data - the matches where our team won or lost with an unusual score difference.

To simplify the analysis, we will create a new variable called “net goals” (num_goals_net) which will represent the number of goals scored minus the number of goals conceded. After an internet search for the data set, and using data wrangling (preparation of data from their raw format), we arrive at the following graphic for our team.

It illustrates the values of our new variables over time.

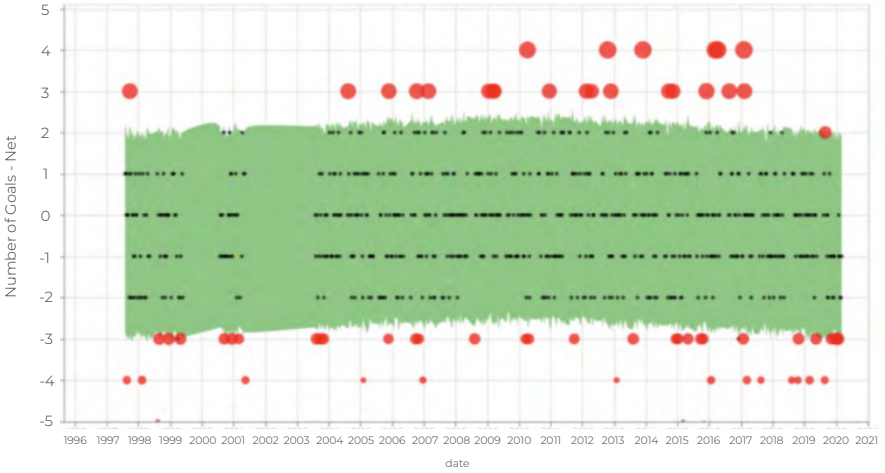


A quick glance at the graph shows that most of the results over time are between -2 and 2 in terms of net goals. The lowest value corresponds to matches lost with a difference of two goals, and the highest value corresponds to games won with a two-goal difference.

Anything outside of this range are events that are considered to be unusual / atypical / occurring less frequently.

After using the anomaly detection techniques, we could clearly see that the interpretation seems correct:

Anomaly Detection



The green area shows the majority of events (about 90% of cases)

The red dots are the anomalies detected by the algorithm.

References - checked on 22 February 2021:

Cambridge Dictionary, N.A. ‘Anomaly’ [ONLINE]. Cambridge Dictionary, N.A. ‘Detection’ [ONLINE]. DeepAI.org, N.A. ‘Anomaly Detection’ [ONLINE]. Dutta, S., Vallabhajosyula, R. R., 2017. ‘Anomaly Detection – Real World Scenarios, Approaches and Live Implementation’ [SLIDESHARE]. OECD.org,N.A. ‘Time Series’ [ONLINE].

5 The data science profession

At the crossroads of several disciplines, data science relies on methods and algorithms to obtain information and knowledge from both structured and unstructured data. Still unknown a few years ago, the job of a data scientist is evolving very quickly; on top of working on data sets, plus training and model refinement, ops and cloud-based skills are now a key part of the role.

To assess this evolution and compare our experience against that of the wider market, we interviewed Christophe Thibault, an external data scientist from outside Devoteam A Cloud with whom we have previously collaborated on a client project.

Christophe Thibault, Data Scientist

After a research career across energy-related subjects, data scientist consultant Christophe Thibault changed his career path to focus on data processing and become a data scientist. After three years in the profession, he shares his thoughts on the data science profession, its daily challenges, the variety of use cases, problems and the main tools used.

Please introduce yourself in a few words

I have been a data scientist consultant for three years. I have a background in physics, with a degree in chemistry and a PhD in physics. I spent part of my career in fundamental and experimental research on energy-related topics, such as hydrogen and batteries. Having worked in France, Europe and Asia, I was looking for a change. I thought about someone I had worked with during my PhD on data processing and data science, an emerging subject at that time. I applied for the Telecom ParisTech Master's degree in Big Data, and trained for 18 months. I am now a data

scientist, and for the last two and a half years I have been working within the digital team of a large French industrial group.

Which skills are most useful to you in your work?

I am lucky to have a background in physics and chemistry, which is a plus when working in this environment. We mainly work on physical quantities (flow rates, concentrations, conductivities, etc.); concepts I know well and which helps when working on the data.

It is very important to understand the relationships between the different quantities and concepts. The business experts are also on hand to explain to us the different types of factories and the processes implemented. They also help us to validate the results.

For the data, we work on time series and forecasts. I use what I have learned both during my master's and over time with experience; I conduct the first analyses on a Python notebook (which is handy for a first exploration and to present the findings to the business), to explore the relationships between variables, and understand the findings. The data scientist's job is to bring his knowledge of data to the business experts, and to explain the different methods and algorithms that could help them achieve their objectives. At the moment we mainly work on time series, which is different from datasets, where time is irrelevant.

How would you define the role of a data scientist?

The profession is changing very quickly. During my second Master's degree, we worked on data sets, and tried to get good results by applying different algorithms depending on the topic. As time goes by, we are realising that the data scientist must do more than that. He or she interacts with the clients to identify and fully understand their needs and must also be able to communicate with the business experts. Technically, the data scientist job is evolving fast. We are now working on the Cloud, and even on ML Ops and industrialisation. It is no longer the data scientist of five years ago who might simply have applied an algorithm to a dataset - a bit like Kaggle - which is certainly information, but a long way from what we can do for companies these days.

What kind of tools do you use?

Initially we work on Python notebook. It is a simple tool, a "draft" where we can test ideas and numerous approaches, write text or include graphics, which allows us to present the results clearly to the client.

It is essential to involve the client in the process: if he or she does not understand what is being done, it will be difficult to move forwards and discuss. It is therefore a good tool for exploring data, testing algorithms, and there is also an educational aspect to it.

For the machine learning platform, we work on AWS: it allows us to implement algorithms and pipelines, then develop and put our work into production. We use

storage services (Amazon S3, Amazon DynamoDB), automation functions (like AWS Glue), and Amazon SageMaker for the ML part. There are also notebooks, and "dockerised" algorithms. The service is ready to use and fully customisable.

Which AWS tools do you use the most?

We use:

- S3 Buckets for data storage
- AWS Glue, which is an ETL (extract, transform, load) dynamoDB tables,
- Amazon Athena, which is a SQL query service,
- AWS Lambda functions (Python code), to automate relatively simple task which are limited in terms of memory and time,
- AWS Step Functions to orchestrate Lambda functions and training/prediction jobs.

When it comes to a processing pipeline with functions, we use:

- AWS Step Functions to automate everything,
- Amazon CloudWatch to run these functions on an ad hoc basis.

AWS Step Functions also allows us to add all the algorithms implemented with Amazon SageMaker, as well as the training and prediction jobs.

- Finally, we also use Insight for data visualisation.

What tools or features are missing today?

Beyond the tools or functionalities, it would be how to manage a project better and better understand Agile methods, use of versioning, etc. to be even more efficient

What are the business use cases?

We mainly work across two projects.

The first is a project for water treatment plants, where filters are used that clog up over time. The data analysis aims to predict the best time to clean or change these filters. Given the cost of replacing these components, the economic stakes are high. Similarly, cleaning must be scheduled weeks in advance, and involves certain logistical constraints, so it is essential to be able to anticipate this need. This project is currently in progress. After an exploration phase with business experts to validate the algorithms and results, we launched a first pilot on a site, and industrialisation began in early 2021. The project is progressing well, the client is happy with the algorithms, and the results are good. The difficulty with this project is that the specifications require a common algorithm across all the sites, but they all work differently and have different processes.

In data science, it is often said that an algorithm works well with a dataset. But if you change the set the challenge is to find a algorithm that performs well on the ten or so datasets.

The second use case is related to water treatment, more specifically to a water clarifier (in the case of a water treatment plant, for example). The aim is to predict the water clarity to optimise (in terms of price) the quantity of chemicals to be introduced. We are also working on forecasting the chemicals and products needed throughout the process (logistics optimisation).

How often are production models updated?

Some models are updated daily and revised every day with new data, while others are updated every two weeks. We decide to revise the models according to the results of the algorithms: when they fall below a certain threshold, we try to readjust. In general, we monitor a lot, so are always looking to improve the models and find new ones. Some aspects of the solution work well but could still be improved, and that's why I'm always looking for new algorithms. Our production model works, but I'm not 100% happy with it, so I'm trying to find the time to study new algorithms and test them out.

You have to experiment, document your work, find the algorithm that could be the right one, and then implement and industrialise it. And, in doing so, we will change a lot of things in production.

What about models in the R&D or proof of concept (POC) phases?

It's true that we also have many models in the testing phase. Everything depends on the specifications. Some data scientists are trying out deep learning and neural networks. The training phase is long and you need a lot of patience before going into production. However, we often need a short lead time to go to market, so we try to find simpler algorithms. There is always a trade-off between the effectiveness of the result and the complexity of the algorithm.

We also have to present the models to the business experts, and need to take into account that they don't have the necessary knowledge to validate a deep learning approach.

I have already offered deep learning on time series, but for the person I was dealing with it was like a black box. The client needs to understand what's going on behind it, and must be able to explain the process. In fact, in three-quarters of the cases, we apply a simple linear regression.

So is there a need to raise awareness of the most advanced methods?

It depends on the ecosystem, but we need to work on data acculturation within the digital teams, raising awareness of the challenges and problems of data science. Machine learning is closely linked to the dataset, yet we still see many cases where datasets continue to be regularly modified, which complicates the work of a data scientist. The behaviours can then change completely and impact the quality of the results. So yes, there is a lot of educational work still to be done within data science.

What is the organisation of the team you work in?

In our data science team we have a product owner, a DevOps on AWS, a front-end team, the business experts who validate the results, and a data scientist - me - with ML Ops and DevOps skills. We're a team of six.

What are the main challenges?

The client is not always aware of the difficulty of running a data project without a detailed specification. In order to deliver functionality, we need to fully understand the requirements, which is not always clear.

If the specifications are not clear, we will code in a certain way, but we will not get the expected result and will have to work on the code again. Sometimes we must go live while the specifications are still being determined. For the data scientist, it is very complicated to work on this basis. It may seem like a small detail, but changing an aggregation or a frequency can impact the whole pipeline.

Sometimes the datasets are heterogeneous. You start working on it, then the data changes because the measurement methods change, and the algorithm can't necessarily be adapted to these changes. It is not always easy for the client to understand, so further awareness work is needed here too.

What is a typical day in the life of a data scientist?

Usually every other morning we have follow-up meetings. Everyone explains what they have done, what they are going to do, any sticking points, etc. Then I start working on my current topics. This could be industrialisation or monitoring because the subject evolves so very quickly. Continuous training and discovering new algorithms and tutorials is also part of the job.

In a large group, we also have a lot of meetings, which take up a significant part of the day - you have to be able to deal with that.

I also work with apprentices, trying to train them and work together on common issues.

We also work with the development and front-end teams. It's a fairly varied job, where no two days are ever the same. In the last fortnight I've been working alone on data mining but my next two weeks will be devoted to production, working alongside the DevOps team.

And an exceptional day?

When you have a hunch! When you are trying to solve a problem in data science, it is very difficult to switch off: you think about it constantly, wondering "what if I tried this or that?". The next day, we put it to the test, spending 7 to 8 hours coding. And if we get the expected result, it's a real feeling of satisfaction.

The job is quite demanding. You can up against a lot of sticking points. It takes up a lot of mental space.

You have to think about two things:

The code, finding the right way to code an idea is not necessarily a trivial matter.

And then the behaviour of the mathematical/intuitive aspect linked to the algorithms must be understood.

About Devoteam A Cloud

Devoteam A Cloud is AWS Premier Consulting Partner that offers excellent know-how on AWS technologies since 2012, and is AWS Premier Consulting Partner. Our team of 500+ AWS experts supports customers with scalable infrastructure, new ways of thinking and operating enabled by the AWS ecosystem to re-invent their business, and evolve into an enterprise platform. In 2020 it was awarded AWS Consulting Partner of the Year.

Key figures

500+ clients

500+ specialists

600+ certifications

4 competencies:

- DevOps
- Data & Analytics
- Security
- Migration

1000+ people trained in 2020

1 Centre of excellence in Lisbon

AWS Well-Architected

aws partner network
Training Partner

aws partner network
Premier
Consulting
Partner

devoteam
A Cloud

acloud.devoteam.com


Creative tech for Better Change