

Capgemini



x



Good morning!

We start at 9 am



How to manage an end-to-end data science case

Data BootCamp 2022

Summary

1. Formation objectives 10'
2. How does a data project work? 50'
3. Dream team for data projects 30'
4. Wrap-up 15'





**Adrien
COURATIER**
Data Scientist Consultant
Next Frontier AI



K E R I N G

STELLANTIS



SOCIETE
GENERALE

QOLMAT



FORMATION OBJECTIVES

What knowledge do you have on the conduct of a data project?



There are several types of use cases ranging from the display of indicators to artificial intelligence, it depends on the maturity of your customer



Visualization

Capitalize on quality and accessible data to share information



Monitoring

Track what's happening on a process or within a team/organization



Analytics & Prediction

Discover patterns to extrapolate the rest



Optimization

Extract insights on the elements to be anticipated or controlled as well as the optimization levers



Automation

Assist employees through self-adaptive/scalable algorithms that automate tasks



THE CONDUCT OF A DATA PROJECT

Illustration of data science project

CONTEXT

In a context of **natural disasters increase** and **meteorological disturbances**, a major French Insurer observed a saturation in its claim management centre. Therefore, we were asked to **underline and quantify** those observations and **identify root causes** and process bottle necks that could affect the claim management process.

OBJECTIVES



Objectify and measure saturation in claims management centers that have been going on for years



Identify and qualify optimization levers

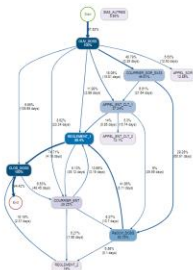


Put light on the root causes that explain this saturation: around the processes and claims themselves

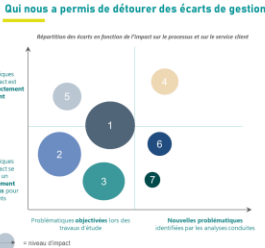


Evolve the data model and the process mining solution

RESULTS



More than **30 spot process analyses** worked with expert teams



9 root causes identified



5 Optimization LEVERs listed to support the disaster throughout its value chain

DATA

8,4M

Customers in 2019

12,8M

Insurance contracts

4,1B€

Turnover

SCOPE



1,2 millions of claims to analyse



7 weeks project delivered in 3 sprints



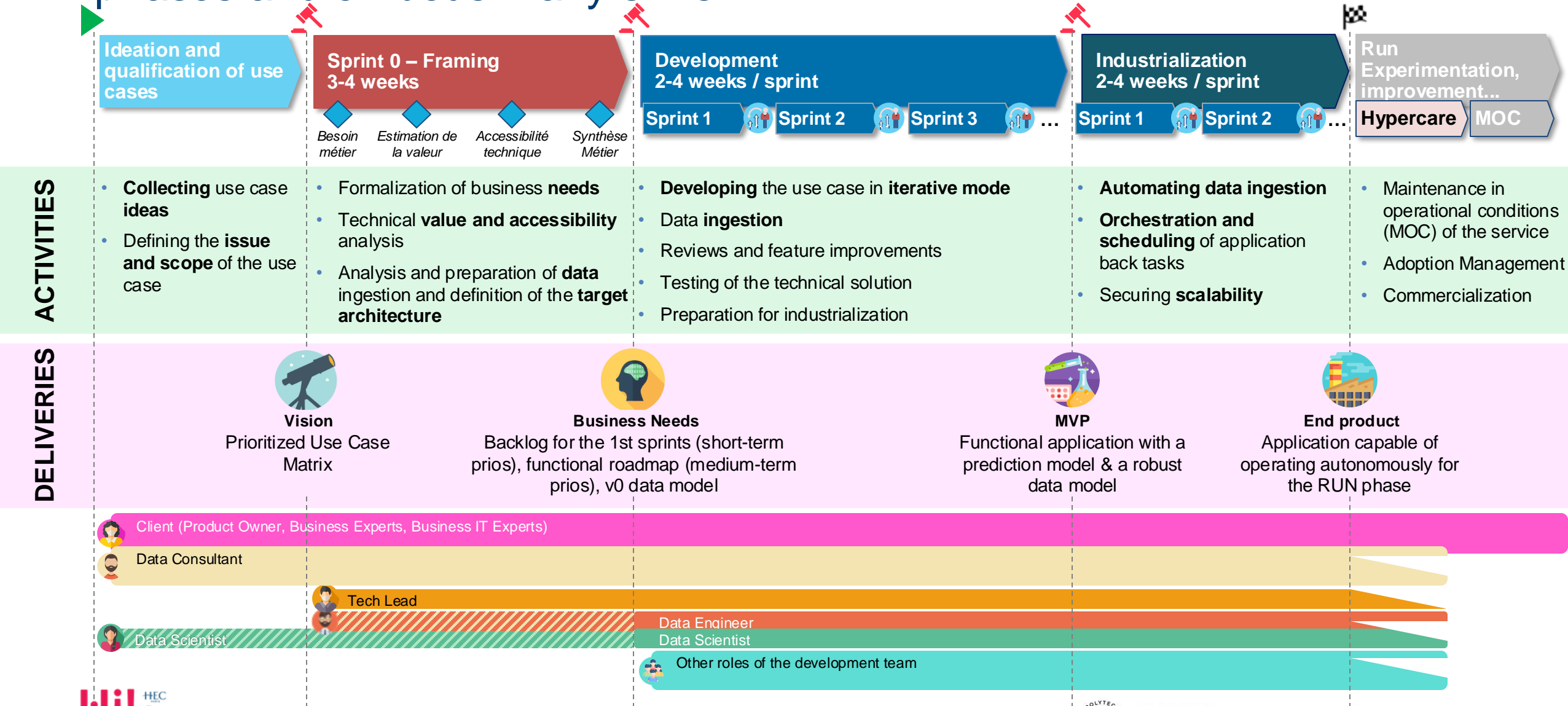
More than 20 type of insurance contract

TOOLS



In-house process mining solutions

The standard agile development life cycle of a project consists of 4 phases and embeds many skills



- Collection of use case ideas (ideation) then definition of the issue and scope (qualification)

Roles and activities



Data Engineer

Can be in support of the prequalification of data (model, techno, quality ...)



Business Analyst

- Supports the profession in the definition of the **product vision** (brainstorm, design thinking, rexs...)
- Identifies the **challenges and objectives of the data product** in the form of macro-functionalities
- Frames the product **scope** (data, users, MVPs...)
- Defines the **issue and accessibility** of use cases



Data Scientist

- Supports ideation by proposing ideas for data science applications
- Assesses the feasibility/complexity of the approach

Key actions



Precisely characterize the purpose of the data product (AI, dataviz...)



Insert the use case into a reflection around a business process



Clarify business usage and target users of the use case



Prioritize use case ideas with regard to their business challenge and technical accessibility



Define the MVP macro perimeter

Developing a Use Case

UDP Offers the possibility to businesses to develop their own use cases. Each building block of the platform was built to enable all ingredients of a Use Case



« As a truck owner, I would like to develop a Dashboard to predict engine failure of my truck fleet »

1

I collect all the data I need to build on the different KPIs of my Dashboard

The data I need can come from different sources (external, internal, sensors ...) and might even be already available on UDP !

If not, I use the ingestion framework on Data Foundation to acquire and ingest my data (or I request UDP team to add it into their ingestion roadmap)

- Engine sensors data flow
- Trucks type & Segment
- ...

2

To modelize the KPIs of my dashboard, I need to transform the data into tangible metrics

Once the data is transformed through the ingestion framework, it is accessible through all BI & AA tools available on UDP. Therefore, I use the tools create the models I need.

- I can use already setup KPI in the KPI catalogue
- Try some models in MLOps patterns
- ...

3

I'm almost there, KPIs and models are ready, I need to visualize them

UDP offers a wide portfolio of analytics and business intelligence execution tools. Being accessible and modelize, I can use some tools in my environment to build on my dashboard

- Power BI/Qlik for dynamic dashboard
- Customised AI solutions I can develop
- ...

DATA FOUNDATION

AI, ANALYTICS & BI FOUNDATION

AI, ANALYTICS & BI EXECUTION

DATA TRUST FOUNDATION

PLATFORM FOUNDATION

Phase 1 – Ideation and qualification



- Illustration of a data use case qualification sheet with "issues-accessibility" matrix

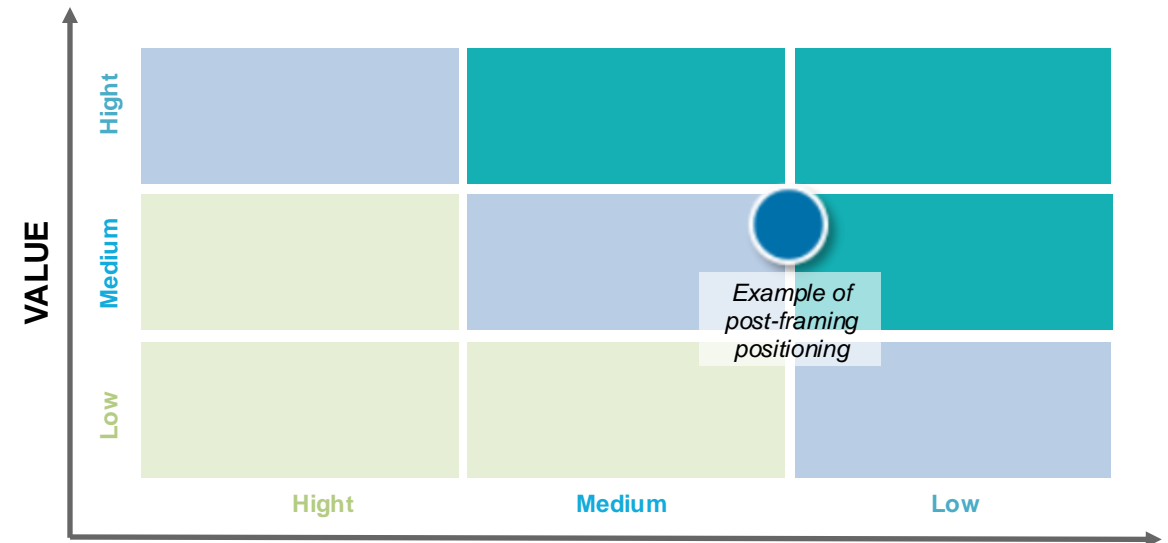
Anticipated complexity

DATA	Quality effort	Low
	Accessibility effort	Low
	Novelty effort	Hight
RSE & ETHICS	RSE contribution/ Green / Energy efficiency	Low
	Ethical risk (data privacy subject, algorithmic bias, etc.)	Hight
	Anticipated complexity/explainability of algorithms	Medium
PROFESSION	Internal complexity: actors to be aligned, visibility of the project etc.	Low
	Business complexity	Low
	Interdependencies with other projects	Medium
	Sponsorship Level	Medium
	Business Maturity	Low
TECHNICAL COMPLEXITY	Complexity data science	Hight
	Complexity DevOps / ML Ops	Hight
	Security issues	Hight
	SLA	Low

Anticipated use case value

Alignment with the organization's strategic priorities	Hight
Anticipated financial value	Hight
Contribution to the advancement of SAFE projects already launched	Low
Technical contribution to the Data Factory: Ingestion of key data, reuse of algorithms...	Medium

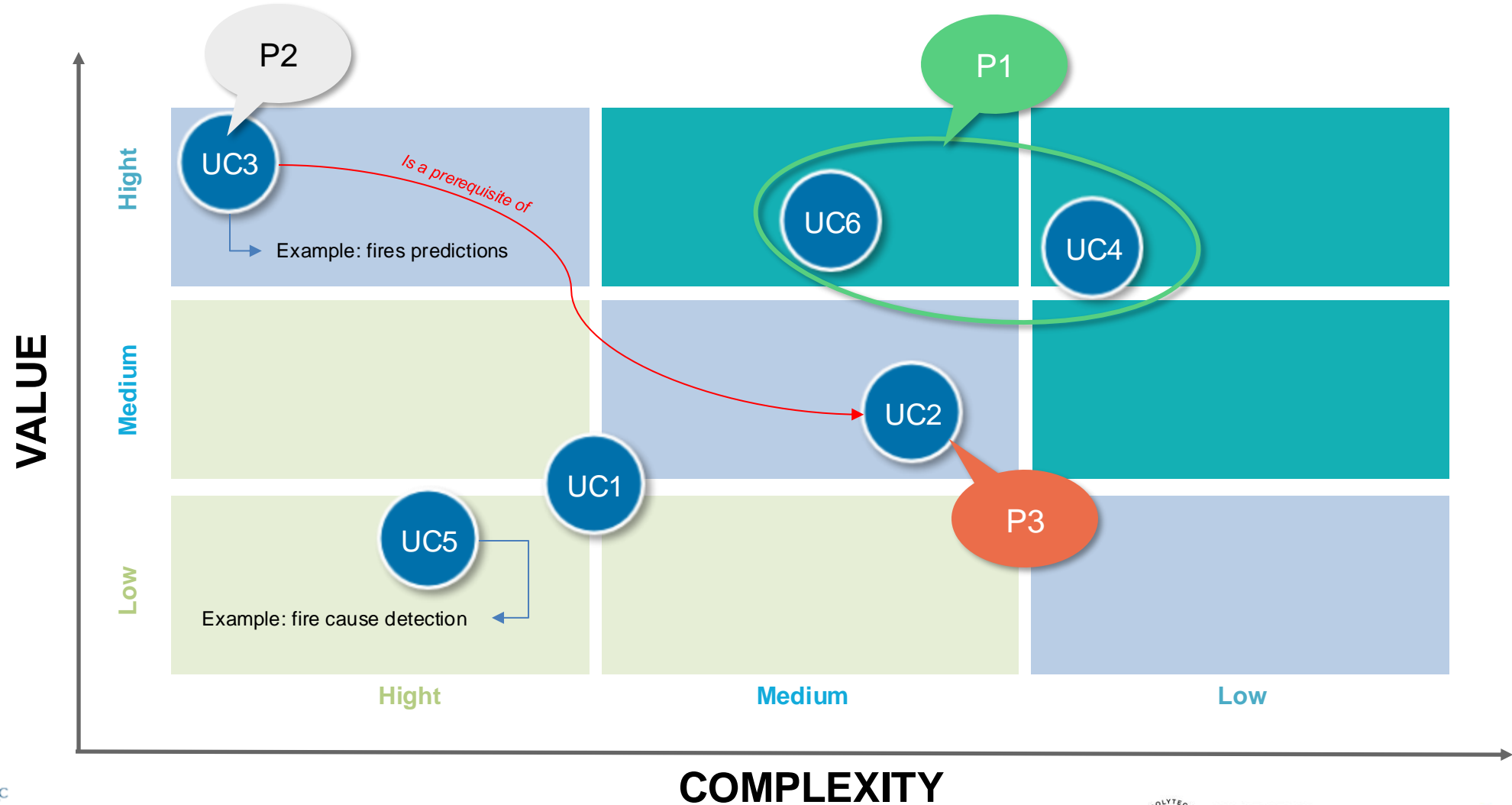
Value / Complexity Position



Phase 1 – Ideation and qualification



- Illustration of a business use case prioritization matrix
-



- Formalization of the need, analysis of the value, preparation of the ingestion and definition of the architecture

Roles and activities

Key actions



Data Engineer

- Performs **data quality analysis**
- Pre-qualifies the **data model**
- Sets up its dev environment

If Tech Lead:

- Helps define the **macro architecture** of the use case
- Ensures the **technical feasibility** of the ambition



Business Analyst

- Details with the business the list of **macro-features**
- Turns the product vision into a **mock-up**
- Identifies the **data needed** for the use case
- Defines the **business value** of **macro-features** and prioritizes them
- Secures **technical** and **legal prerequisites** (GDPR)



Data Scientist

- Advises on the **data approach** to be preferred if necessary – study of the state of the art
- Ensures the **technical feasibility** of the ambition, in terms of model (depending on the quality and quantity of the data)



Go as far as possible in defining **priority macro-features** before starting devs



Dissociate the functional need from the **technical response** in the backlog structure



Use **mockups** to help define business needs and frame user stories



Anticipate the **industrialization of the use case** (e.g. management of user profiles)



Co-construct the use case architecture with the client's technical teams (including data architects)



Adjust ambition with regard to available data and their quality

Product backlog

Ideation



Sprint 0

Development

Indus.



All Epics offer the complete view of the product's features

Epic	Feature	User Story	Release	Business value	Sizing	Data	Back End	Front End	Mock-ups validated
1. Launch an extraction	Tag prospects clients	As an analyst, I want to be able tag it as "Joker" so that I can upload the excel loader previously created in CRF to retrieve its information	S6	3.5/5	1	0	1	0	To be updated
2. Extract data	Extract text data	As an analyst, I want to make sure that the algorithm recognizes requests that have already been done, and show me the existing one	S6	TBC	2	0	2	0	Not needed
		As an analyst, when some items are not found in tables (about ten, defined in workshops) I want the application to search them in text and scrap the relevant paragraph(s) to display them	S6	3/5	11	5	3	3	Done
	Manage different languages	As an analyst, I want the application to be able to handle every language in scope in latin alphabet	S6	3.5/5	3	2	1	0	Not needed
3. Check and Complete	Error notification	As an analyst, I want to be able to know when the data extraction has been unsuccessful, so that I can relaunch it	S6	TBC	1	0	1	0	Not needed
	Visualize textual notes	As an analyst, I want to visualize relevant textual notes for defined items that have not been found in tables so that I can quickly fill them manually	S6	TBC	3	0	0	3	To be updated
Manage access	Read only	 <p>The character always appears in the user stories because it identifies the end user of the feature</p>							
	Read and modify								
	Admin access								
Dashboard and KPIs	Track the accuracy	 <p>The MVP is a subset of the epics to be developed. Epics that are not part of the MVP are not to be detailed through user stories</p>							
	Visualize results								

Architectural diagram

Ideation

Sprint 0

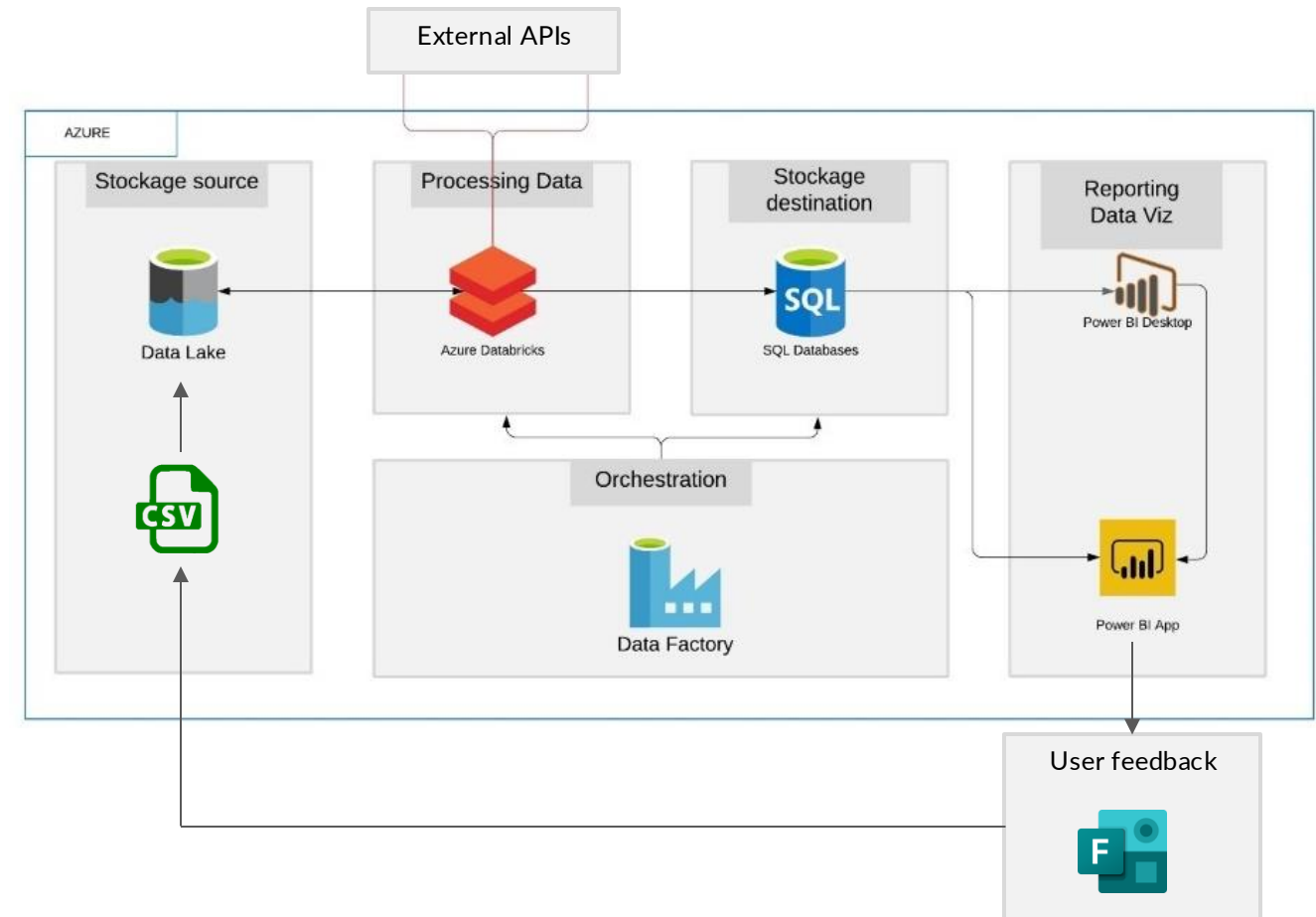
Development

Indus.

Illustration

Technical Questioning

- Understanding the services to be used for data **storage**
- Use of appropriate tools for data **processing**
- Implementation of an ingestion logic
 - Customer data sources
 - External APIs from Data providers
 - Integration of a **feedback loop** to improve / correct association algorithms
- Automated architecture deployment**
 - CI/CD from DEV to PROD in close collaboration with DevOps teams



Once the sprint 0 has passed and the need has been identified, the development pipe is set up



Business need



Data ingestion

Model and training

Preparation of data and its transfer

**Restitution
(cartography, graphs, KPIs...)**



DE



DS



DE



Dev Front



Tech lead

Construction of the application itself, in iterative mode (sprints)

Roles and activities



Data Engineer

- Builds data **pipelines**
- Cleans and structures source datasets
- Builds the target **data model**:
 - Creates aggregates (data cross-referencing)
 - Calculates metrics



Business analyst

- Framework **business needs** to feed and prioritize the backlog
- Collects **business rules**
- Identifies **missing data** and its source
- Initiates and drives ingestion requests
- **Tests** developed US
- Collects **user feedback**
-



Data Scientist

- Builds **features**
- Expands the **model**
- Builds explainability
- Defines with the business the **metrics for evaluating** the model
- Builds complex visualizations (e.g. distance graphs between business objects)

Key actions



Explain the data source as soon as the features are written



Involve the Tech Lead as soon as the features are written



Don't always wait for industrialized data



Write the US with the DE and the DS



Spend time on the data model



Estimate the complexity of the US with the entire data team



Have the US pre-tested by the consultant

Illustration

Context

- Anomalies and delays in the supply chain of an aircraft manufacturer are discovered too late
- The objective of the project is to anticipate the risks of a decline in performance in a horizon of 6 months to two years.

	As is	Random Forest	Linear Regression	Decision Tree	Gradient Boosting
True Negatives	585	611	524	611	607
False Negatives	7	11	6	12	6
True Positives	5	1	6	0	6
False Positives	26	0	87	0	4
Precision	0.16	1	0.06	-	0.60
Recall	0.42	0.08	0.5	0	0.5
F1-Score	0.23	0.15	0.11	0	0.55

True Negatives (TN): Number of alerts not raised at a rate

False Negatives (FN): Number of alerts wrongly not raised

True Positive (TP): Number of alerts raised at a rate

False Positives (FP): Number of wrongly raised alerts

Precision: Proportion of true alerts among the total number of alerts raised

$$Precision = \frac{TP}{TP + FP}$$

Recall: Proportion of actual alerts raised among the total number of alerts that should have been lifted

$$Recall = \frac{TP}{TP + FN}$$

Illustration of the performance of different models and selection of the appropriate model for the use case

Automate and orchestrate pipelines, secure operation in perennial mode

Roles and activities



Data Engineer

- Revises and **freezes the data model**
- **Parameter for updating datasets**: frequency, response if error, aggregate calculations
- **Freezes the mapping of ingestion flows**
- **Implements quality control rules**: freshness, completeness, consistency, mandatory metadata...



Consultant

- **Drives industrialization**: monitors the consistency of results and the proper ingestion of data
- Harvests and **qualifies user feedback** in the backlog
- Participates in the **acculturation of professions** to new uses
- **Ensures the transparency of the deliverable** to the customer: sources, quality, freshness, explainability, limits of the model



Data Scientist

- **Optimizes its model**: resources consumed, response time, bias, extreme values
- **Plans training and model inferences**

Key actions



Knowing how to identify industrialization: it is sometimes done over the water



Don't wait for the end of the MVP to think about industrialization



Encourage the team to mobilize the ecosystem Focus on the problems encountered



Automate, automate, automate



Keep a sprint to "refine" the app: a hypercare sprint

- Define with the business the refresh rate of information

Training

- **Model training** consists of generating a new model based on a fixed dataset.
- **Plan the training of the model:**
 - How often should the model be trained to account for new data?
- **Optimize the model:** resources consumed, response time, bias, extreme values

Inference

- **Model inference** involves applying the model to a sample to obtain a prediction.
- **Plan model inference:**
 - How often do predictions need to be refreshed?
- **Monitor performance:** relevance of the results obtained, comparison to reality and business indicators

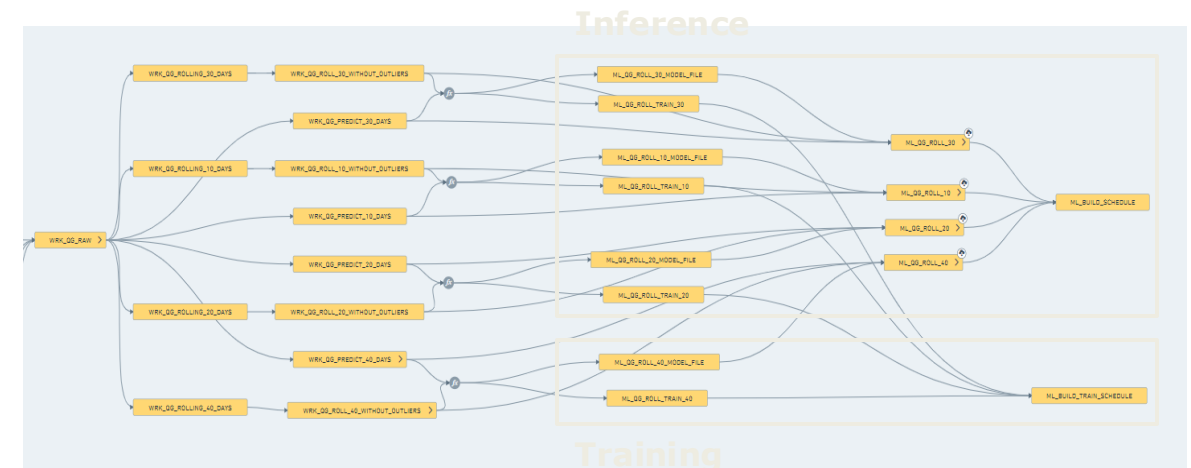
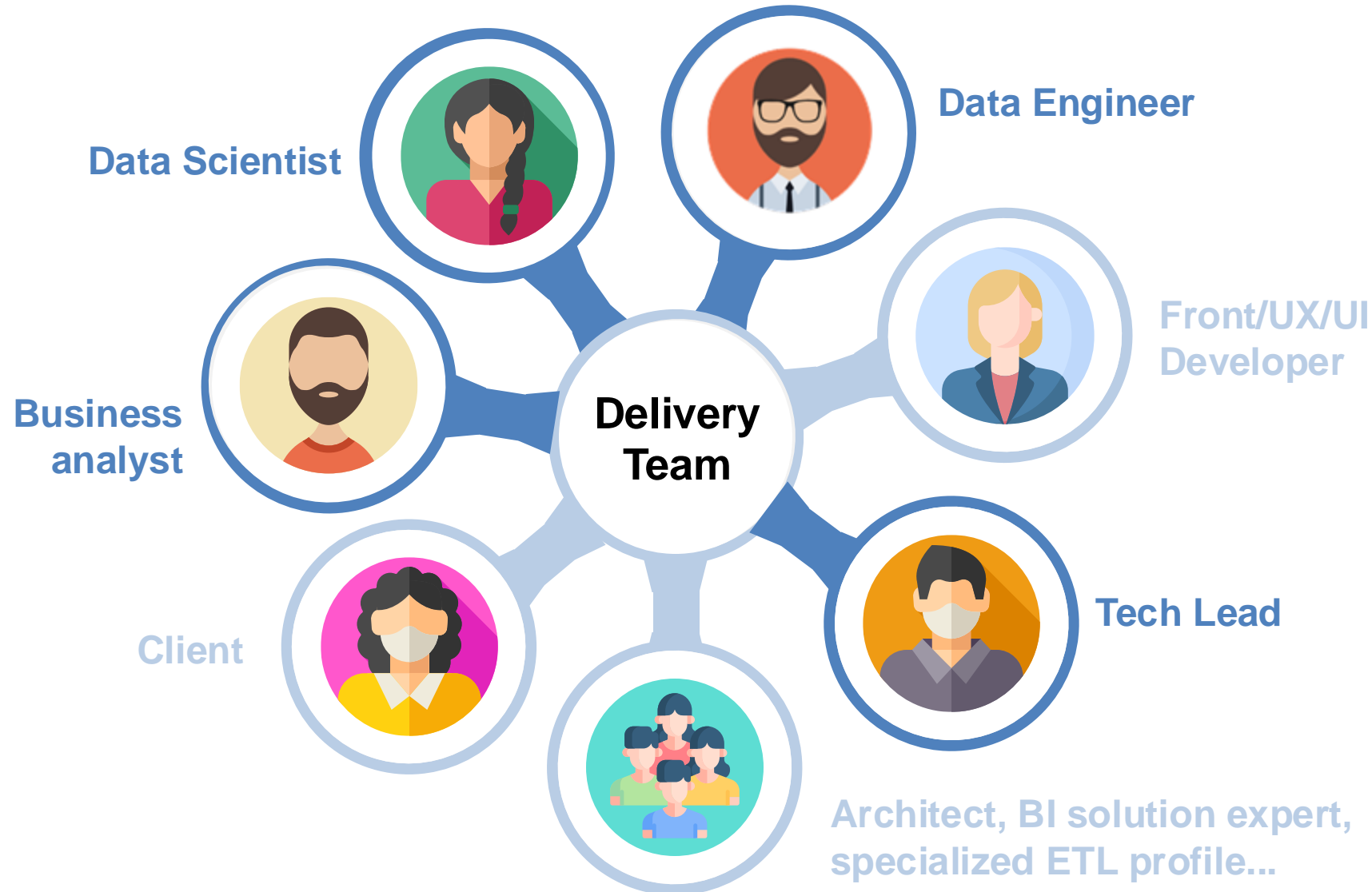


Illustration of a data pipeline and training and inference phases

The background of the slide is a light gray with a pattern of blue and red geometric shapes, including rectangles, circles, and exclamation marks. A large, solid blue rectangle is centered on the slide, containing the title text.

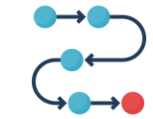
THE ROLES IN A DATA PROJECT

A delivery team requires specific profiles according to needs, 4 of them interest us today





Data Scientist



Modeller

- Turns a business problem into a mathematical problem
- Uses machine learning models, statistics and algorithms to meet business needs
- Iterates with the business in order to understand the data and know how to use it



Developer

- Carries out exploratory analyses on the available data
- Prepares the data (in collaboration with the DE) so that it can be used by its models
- Packages his model so that it can be industrialized
- Monitors the performance and durability of its model over time



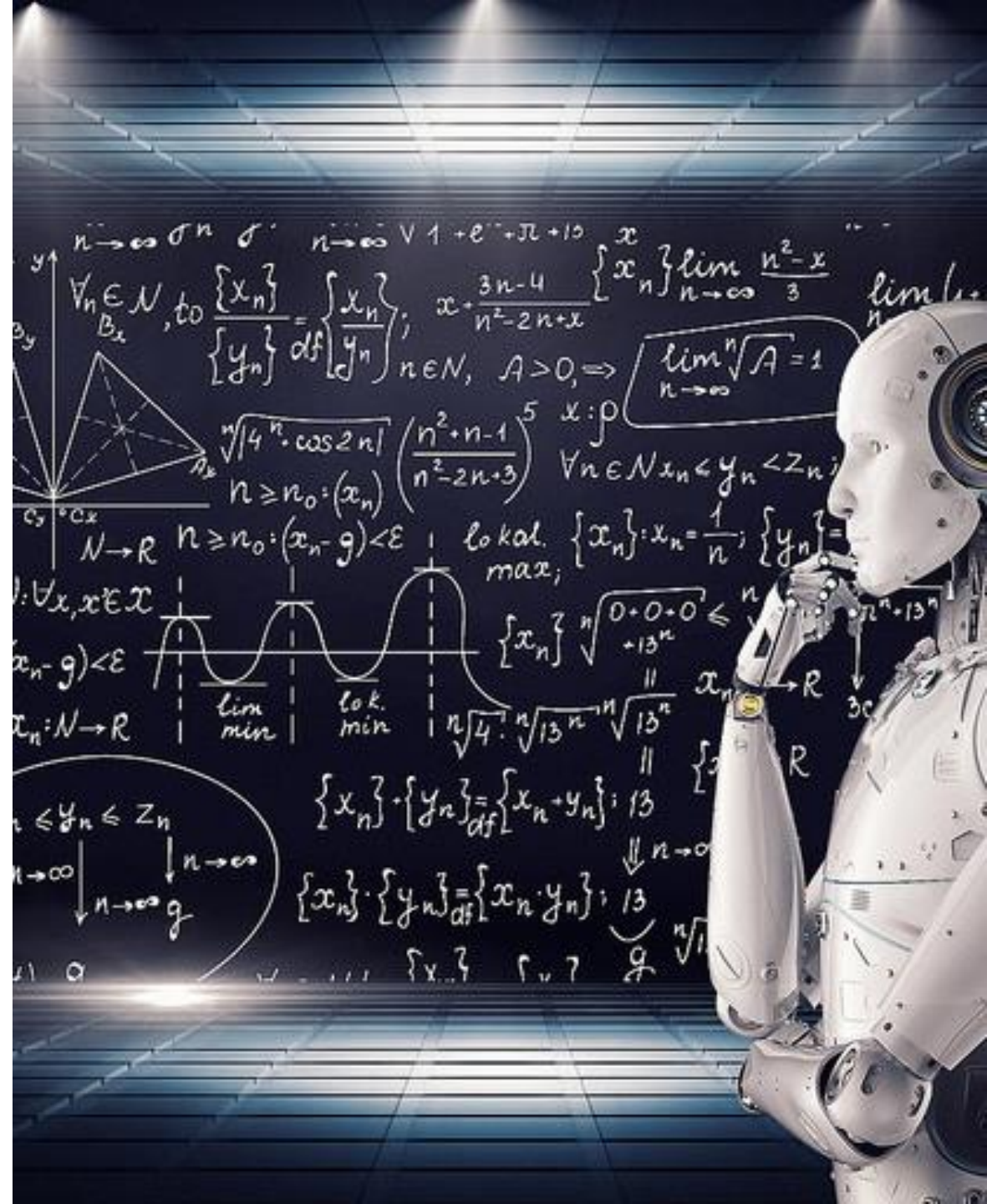
Technical referent

- Advises on technical choices
- Proposes architecture concepts for data science solutions
- Frames and prioritizes datascience use cases
- Puts his technical skills at the service of the client (e.g. production of statistical analysis)



Popularizer

- Presents its results in a format understandable by all
- Organizes acculturation sessions in data science
- Organizes hackathons



To fulfill his missions, the Data Scientist can use several technologies and learning models

Supervised methods

Predict values (continuous or discrete) based on a data history where the result is known

Ex: classify a new incident in a known taxonomy



Unsupervised methods

Group data according to their similarities (without prior information on which families to predict)

Ex: check the robustness of a taxonomy

Knowledge graphs

Connecting entities, their properties and relationships to generate new knowledge

Ex: detect fraud in financial systems by highlighting unusual behavior



Computer vision & image recognition

Acquire, process, analyze and understand images, text (OCR) or sequences of images (videos)

E.g. contract/attachment scan analysis for anomaly detection/fraudulent scheme

Natural Language Processing / Understanding

Processing and understanding human language from textual corpora

Ex: extract representative topics (build a taxonomy), analyze sentiment (customer satisfaction), ...



Natural Language Generation

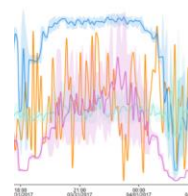
Generate textual human languages (Chatbot, weather reports, ...)

Ex: support for new recruits / increased customer relations

Time series

Analyze time-ordered data to detect unusual behavior

Ex: detect deviations in the usual operation of aircraft sensors





Data Engineer



Developer

- Develops, builds, tests and maintains an architecture



Responsible
of the data

- Processes the data:
 - Its ingestion (Datalake, API, ...)
 - Through its treatment (Spark, Scala, ...)
 - Until its final return (app., storage)



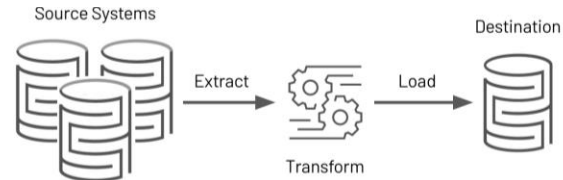
Architect

- Builds the data model to be used/rendered in an application



Orchestrator

- Guarantees the data flow (dataflow), the orchestration of programs /scripts



- Helps the Data Scientist in:
 - data preparation (pre-processing)
 - deploying models



Crew



The specializations of a DE

A DE differs in the areas it masters



Ingestion

- Ingestion of different file formats: datalake (parquet, delta), csv dumps, ...
- API requests
- Scrapping Web



Modeling

- Creating the Data Model from the need
- Creation of the Data Flow to feed the Data Model



Dev. Methodology

- Using Git (code versioning)
- Setting up an architecture on a platform



ETL

- Language used: python, spark, scala, ...
- Script orchestration
- Big data



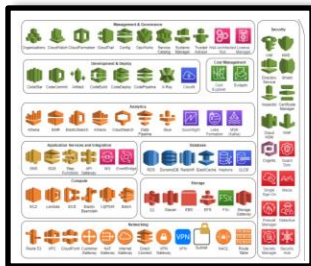
IA

- Feature engineering
- Industrialization of models

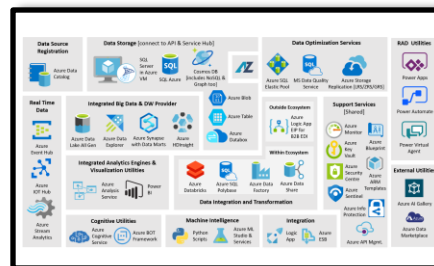
An DE can specialize on a cloud platform



Amazon Web Services



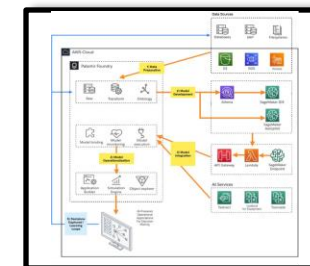
Microsoft Azure



Google Cloud Platform



Palantir Foundry



and others ...



Tech Lead



Customer Interactions

- Responsible for the technical deliverables to the customer (technical and documentation).
- The interface between the customer, the development teams and the run teams.
- Guarantees the consistency of the solution developed, and the right choices of dependencies.



Technical referent

- Arbitrates the implementation choices made by his team.
- Helps in the resolution of the blocking topics of the project.
- Is at the heart of the development validation process, and brings his technical expertise to it.



Host

- Ensures the implementation and proper application of collaboration processes (Git, CI, Code Review).
- Streamlines communication between the technical teams and the client, with a global view of the project. It identifies the right proxies to foster seamless collaboration.

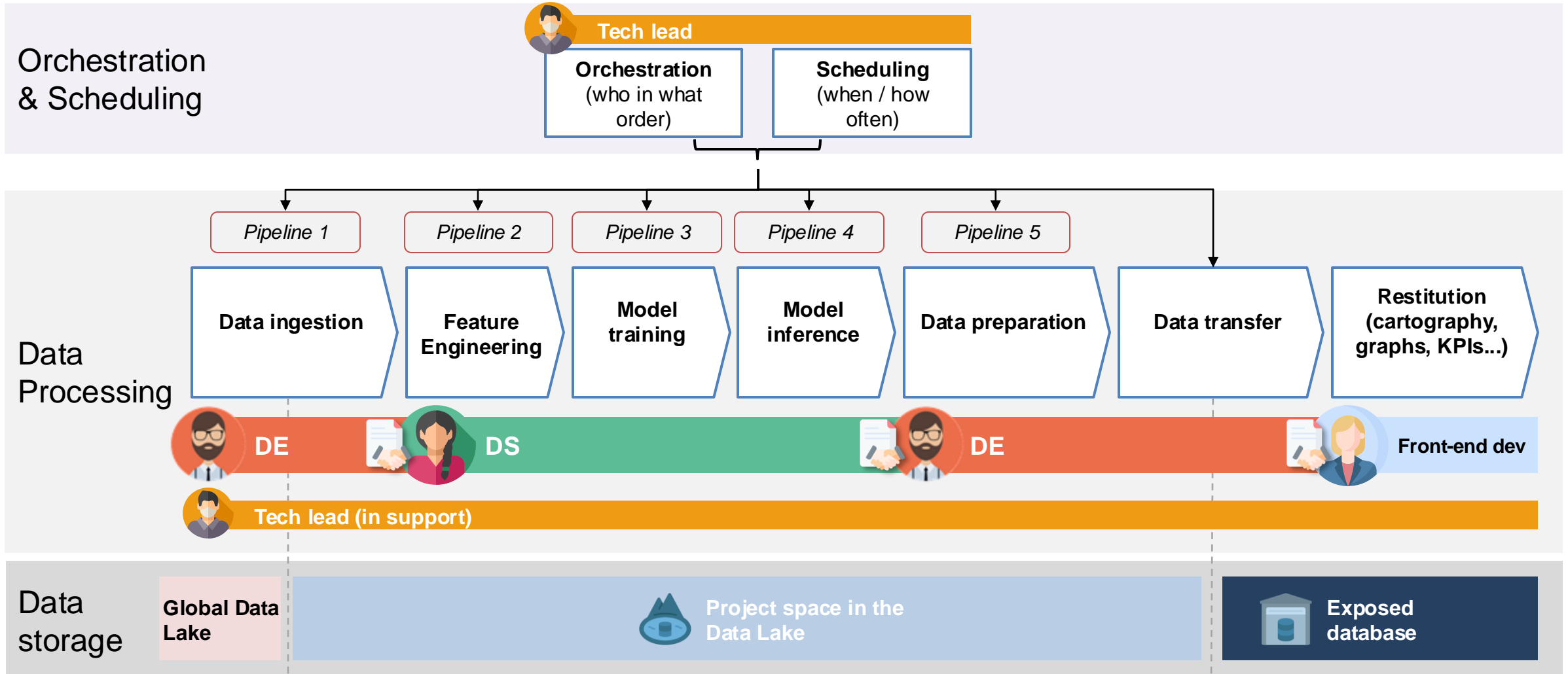


Innovative Trainer

- Carries out a technological watch on its subjects of expertise.
- Ensures the development of technical skills of its teams.
- Makes his teams aware of the state of the art (participation in Tech events, intervention of experts on project).



Illustration of role intervention steps during the construction of a data product





WRAP-UP



DATA LAKE

Data scientist



Data engineer

Data Consultant





Key success factors of a data project



Generate business value



Define model evaluation metrics with the business



Identify accessible & quality data



Engage users and capture business rules

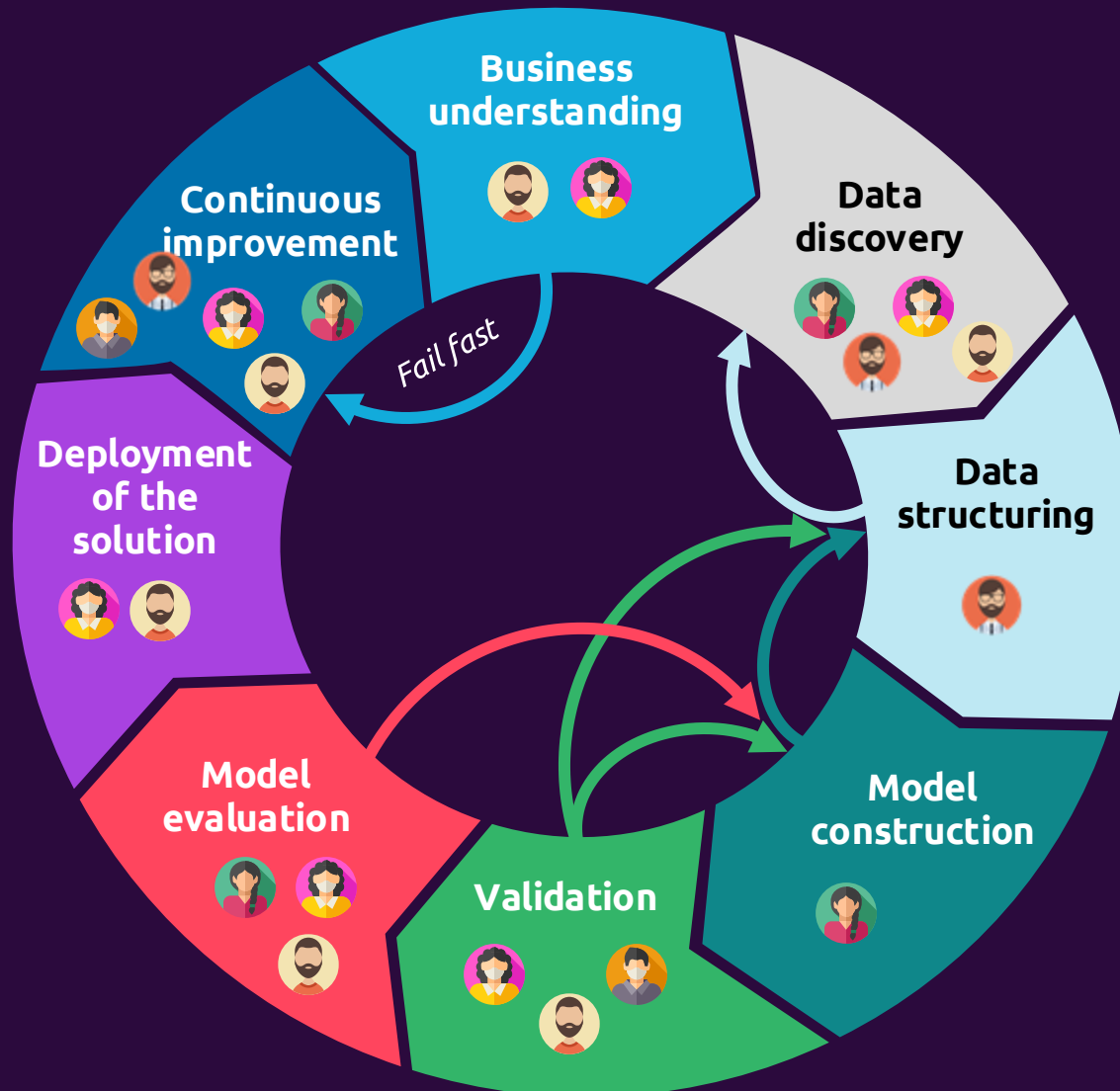


Acculturate the profession to the AI model



Anticipating industrialization

In short, the wheel of a Data project



Roles



Data Consultant



Data Scientist



Data Engineer



Client / PO



Tech Lead