01. Data platform Foundation

Atos
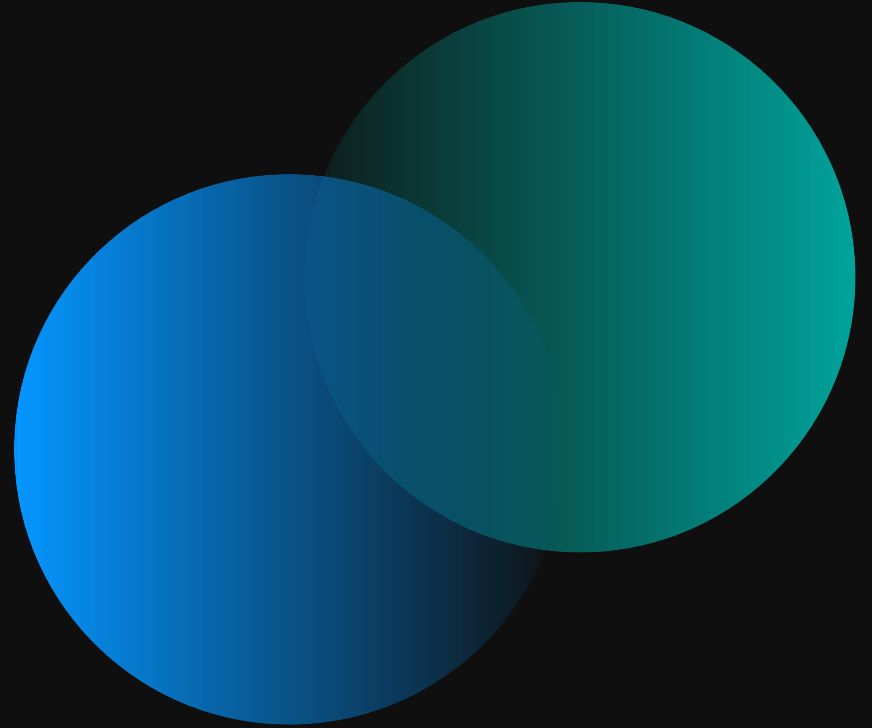
Dissection of an application
Many layers/components needed

- Applications
- Data
- Runtime
- Middleware
- OS
- Virtualization
- Servers
- Storage
- Networking

Atos
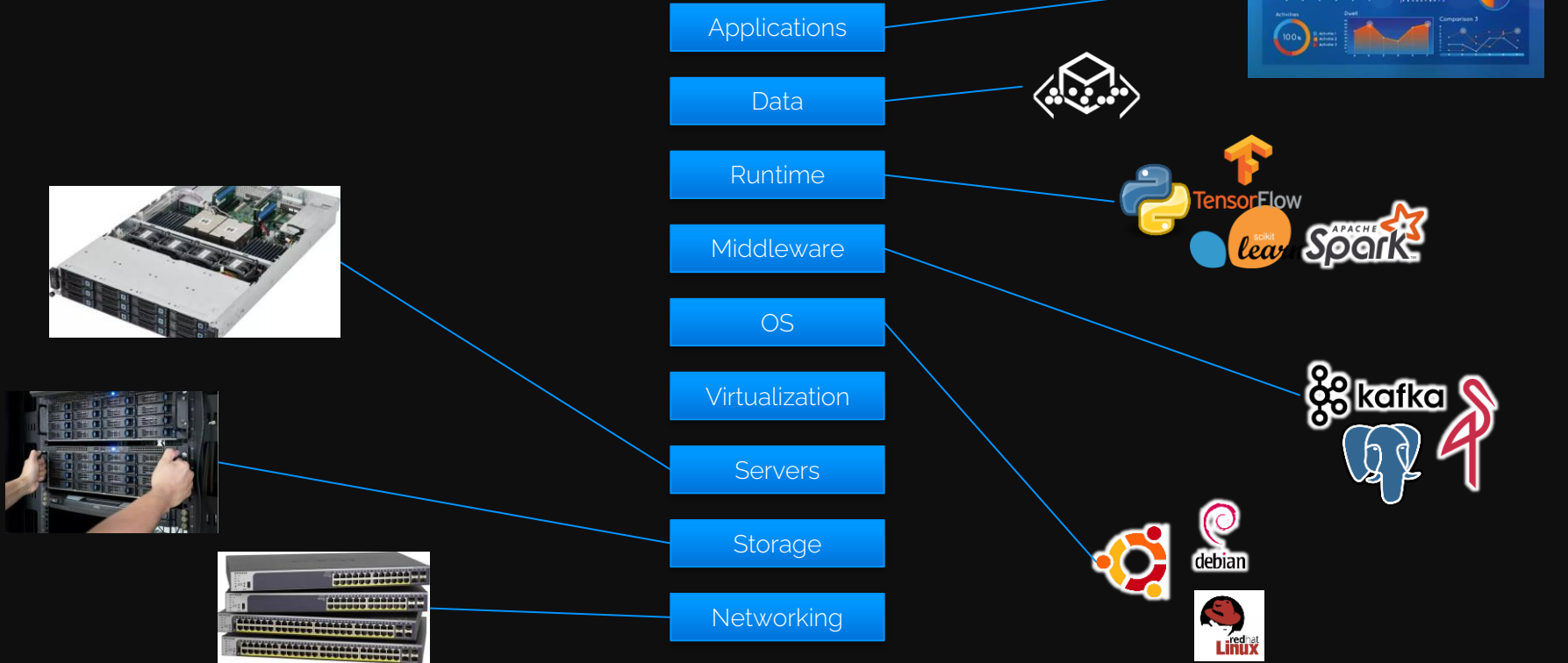
Many ways to manage these layers
Different service offers

Private cloud

Public cloud

On premise	BMaaS	IaaS	CaaS	PaaS	FaaS	SaaS
Applications	Applications	Applications	Applications	Applications	Applications	Applications
Data	Data	Data	Data	Data	Data	Data
Runtime	Runtime	Runtime	Runtime	Runtime	Runtime	Runtime
Middleware	Middleware	Middleware	Middleware	Middleware	Middleware	Middleware
OS	OS	OS	OS	OS	OS	OS
Virtualization	Virtualization	Virtualization	Virtualization	Virtualization	Virtualization	Virtualization
Servers	Servers	Servers	Servers	Servers	Servers	Servers
Storage	Storage	Storage	Storage	Storage	Storage	Storage
Networking	Networking	Networking	Networking	Networking	Networking	Networking

3

Atos

Many ways to manage these layers
Some examples

Private cloud

Public cloud

On premise	BMaaS	IaaS	CaaS	PaaS	FaaS	SaaS
- Datacenter	- X86 servers - NAS storage	- AWS EC2 - Azure VM - Google GCE	- AWS EKS - Azure AKS - Google GKE	- AWS Appsync - Azure App Service - Google App Engine - Heroku	- AWS Lambda - Azure Function - Google Cloud Function	- Salesforce - Office360 - AWS Redshift - Azure Synapse - Google BigQuery

Few or no vendor locking

Migration can be very complex depending on the data service used.
Strong vendor locking, but Dev and Ops teams are strongly reduced.

« Cloud provider agnostic »

« NoOps experience »

Atos

Cloud or not cloud ?

No « one size fit all » solution, many deployment modes available

Full cloud	Multi cloud	Hybride	On premise
Agile, no ops, secured	Best of breed	Trade off between both worlds Offloading	Usually well mastered by companies
Could be expensive	Complex management Lots of external network	Complex management	Not scalable, lake of agility and TTM

Atos

Vision
Atos is investing a lot in Sovereignty & Data securization

Atos is a founding member of Gaia-X, launched in may 2020. Gaia-x is a European initiative for improving the interoperability and the sovereignty of the cloud. Atos is working closely with Gaia-x organization to define the future standards for sharing data (data space).

In July 2020, European court of justice has denounced the privacy shield between USA/Europe especially for GDPR compliancy issues & cloud act extraterritoriality application. This decision has accelerated the needs and investments on sovereign approach especially for public sector.

Atos has been selected by French Ministry of Defense for developing its sovereign BigData platform. This success has given birth in 2021 to a joint venture between Thales and Atos called Athea for handling the next steps of this strategic program. Atos Codex Data Platform is a civil fork of this sovereign platform based on open sources.


Reccently Atos has launched Atos OneCloud Sovereign Shield, which is a comprehensive edge to cloud platform ecosystem and highly secure service that improves the level of control clients have over the data they produce and exchange, helping them regain control and effectively deal with legal dependencies.

AtoS

Cloud native architectures and applications
The core concepts
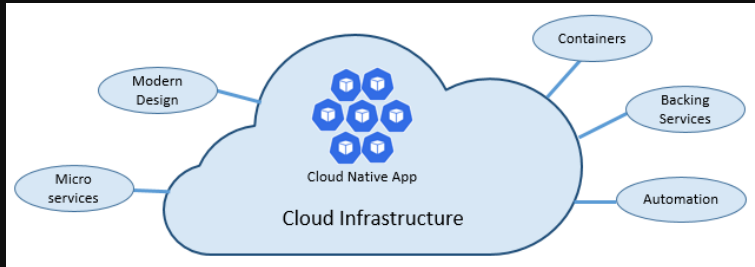
- Modern design => 12factor
- Microservices => cut down the monolith
- Containers => isolation
- Backing Services => Don't do all by yourself
- Automation => code once, run many

The Twelve-Factor App

I. Codebase
One codebase tracked in revision control, many deploys
II. Dependencies
Explicitly declare and isolate dependencies
III. Config
Store config in the environment
IV. Backing services
Treat backing services as attached resources
V. Build, release, run
Strictly separate build and run stages
VI. Processes
Execute the app as one or more stateless processes
VII. Port binding
Export services via port binding
VIII. Concurrency
Scale out via the process model
IX. Disposability
Maximize robustness with fast startup and graceful shutdown
X. Dev/prod parity
Keep development, staging, and production as similar as possible
XI. Logs
Treat logs as event streams
XII. Admin processes
Run admin/management tasks as one-off processes

XIII. API First
XIV. Telemetry
XV. Authenticaioin/Authorization

AtoS

ML in production
The real challenges

Performance

How to infer at scale to handle prediction spikes?

When do we need to start re-training the model and which datasets should we use?

Are resources (CPU, GPU) being used efficiently ?

Is the model over or under scaled ?
Are the inference response times acceptable ?

Exploit

How can we ensure that the right versions of the models are deployed and that they use the right data for their prediction?

How to recover the training datasets in order to analyse the deviant behaviours of the model afterwards?

Build

Do you know the answers of all these questions ?

How to containerize a model ?

Is there prediction protocol standards (http, grpc) ?

How to handle different framework and model format (Ttensorflow, Pytorch, Onnx,...) ?

Business impact

How can you calculate accuracy if you don't know/get the truth immediately (feedback loop, ground truth)?

Is it necessary to run in production the best model or a combination of several models?

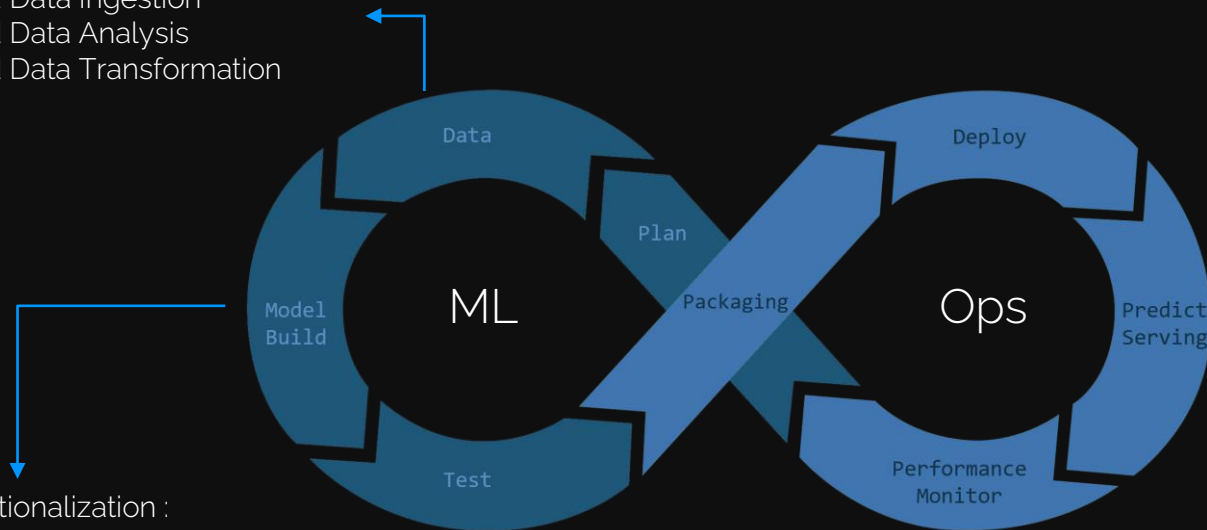Which metrics to follow on a model (precision, recall, input data distribution, ...)?

AtoS

MLOps
A methodology to industrialize ML

Data WorkFlow :
- Automated Data Ingestion
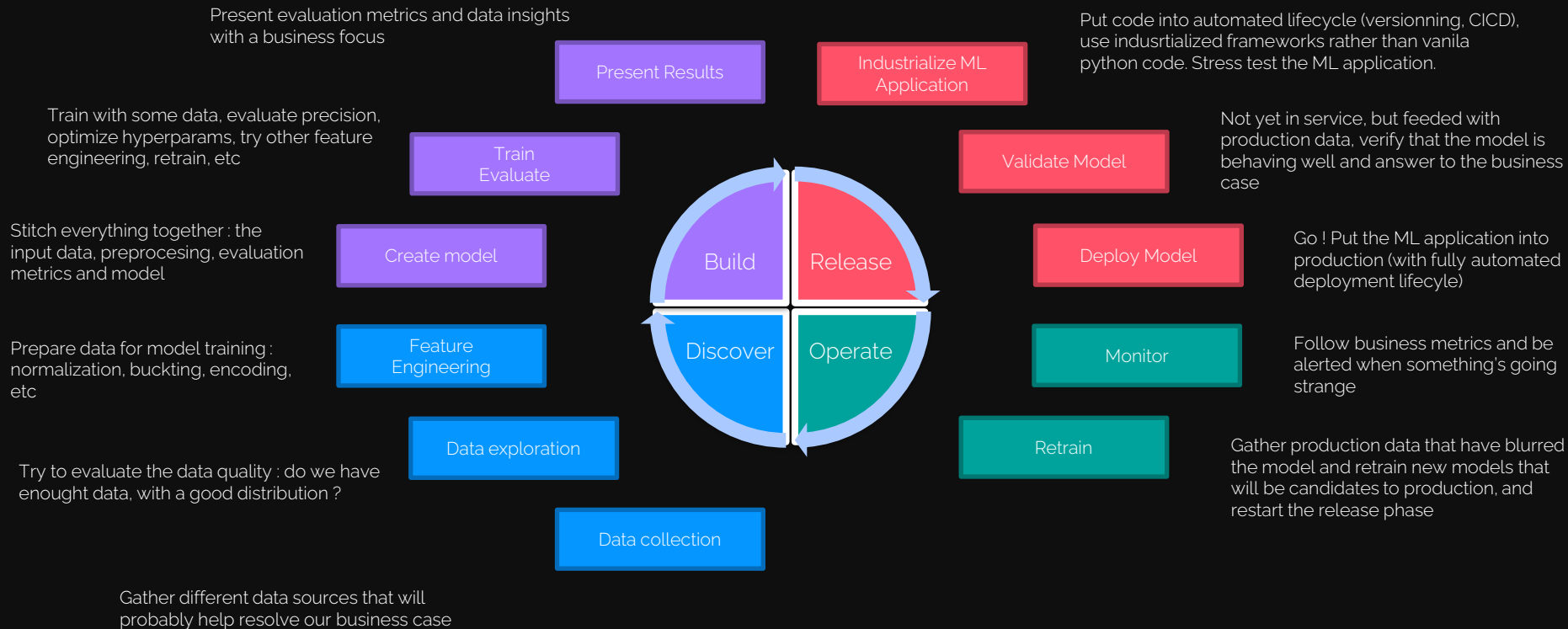- Automated Data Analysis
- Automated Data Transformation

Deploy & Run :
- Serving at scale
- Explaining
- Monitoring & logging

Model Operationalization :
- Training at scale
- Tuning
- Governance

Data

Plan

Packaging

Deploy

ML

Ops

Model
Build

Predict
Serving

Test

Performance
Monitor

Atos

ML lifecycle

Present evaluation metrics and data insights with a business focus

Put code into automated lifecycle (versionning, CICD), use indusrtialized frameworks rather than vanila python code. Stress test the ML application.

Present Results

Industrialize ML Application

Train with some data, evaluate precision, optimize hyperparams, try other feature engineering, retrain, etc

Train Evaluate

Validate Model

Not yet in service, but feeded with production data, verify that the model is behaving well and answer to the business case

Stitch everything together : the input data, preprocesing, evaluation metrics and model

Create model

Build

Release

Deploy Model

Go ! Put the ML application into production (with fully automated deployment lifecyle)

Prepare data for model training : normalization, buckting, encoding, etc

Feature Engineering

Discover

Operate

Monitor

Follow business metrics and be alerted when something's going strange

Try to evaluate the data quality : do we have enought data, with a good distribution ?

Data exploration

Retrain

Gather production data that have blurred the model and retrain new models that will be candidates to production, and restart the release phase

Data collection

Gather different data sources that will probably help resolve our business case

Atos

ML lifecycle and artifacts

Discover

| Data Collection | Data Exploration | Feature Engineering |

Build

| Model Building | Training and Evaluation | Present Results |

Release

| Operationalize Model | Validate Model | Deploy Model |

Operate

| Monitor Model | Retrain |

Data

Raw data

Prepared data data

Training data

Test data

Production data

Production data

Retrain data

Insights

Data Quality

Models metrics

Models Quality

ML Application Quality

Code

Training code

Test code

Production code

Training code

Model

Candidate models

Choosen model

Productionized model

Production model

Retrained model

12

Atos

DataOps
What the heck is behind this buzz word ?

- Comes from Agile + Devops + Lean Development
- Same ambitions than now wellknown devops
  - Federate different teams around the product/value
  - Industrialize human and technical process
    - Automate most of dev/build/push actions into target environment
  - Help to handle the technical complexity of the ecosystem
- Key concepts
  - Value first
  - Collaboration
  - Automate
  - Orchestration, test, monitor
  - Security
- Objectives
  - Improve data and analytics quality
  - Reduce TTM

DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization.

Source : Gartner

AtoS

Data Platform Functional Architecture
Lets create it
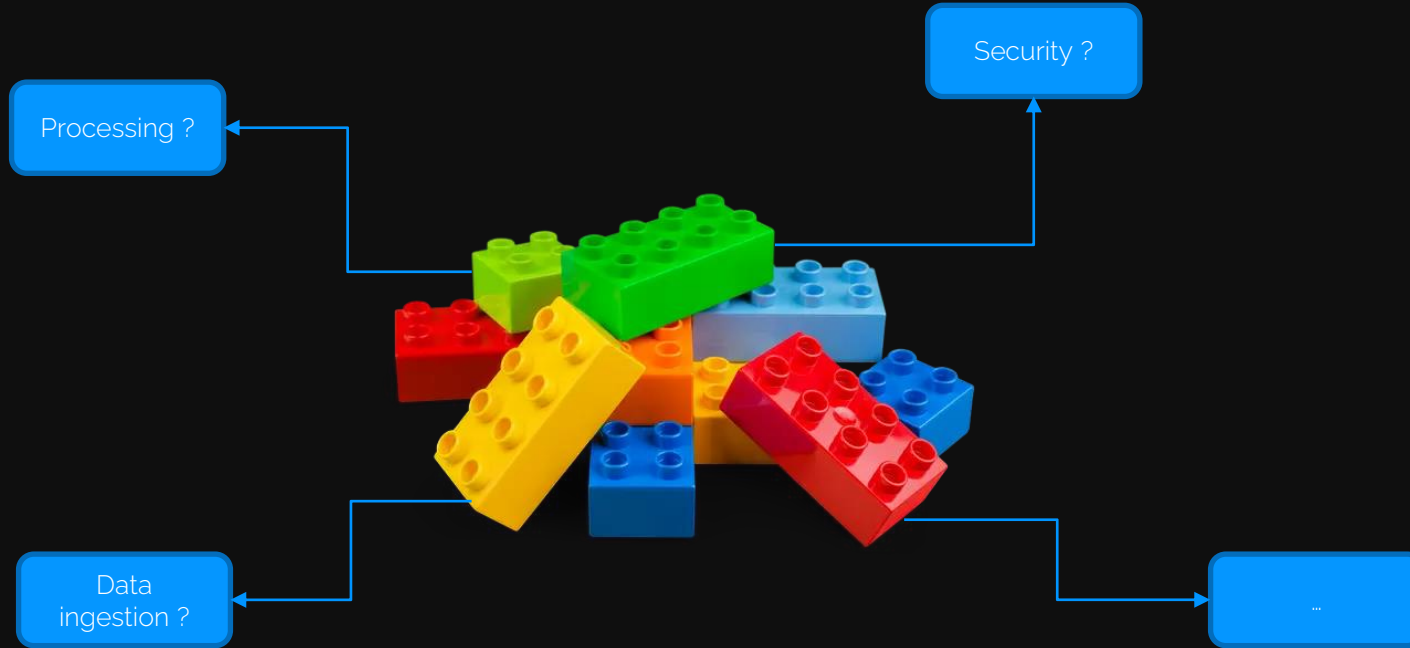
Processing ?

Security ?

Data ingestion ?

...

Atos
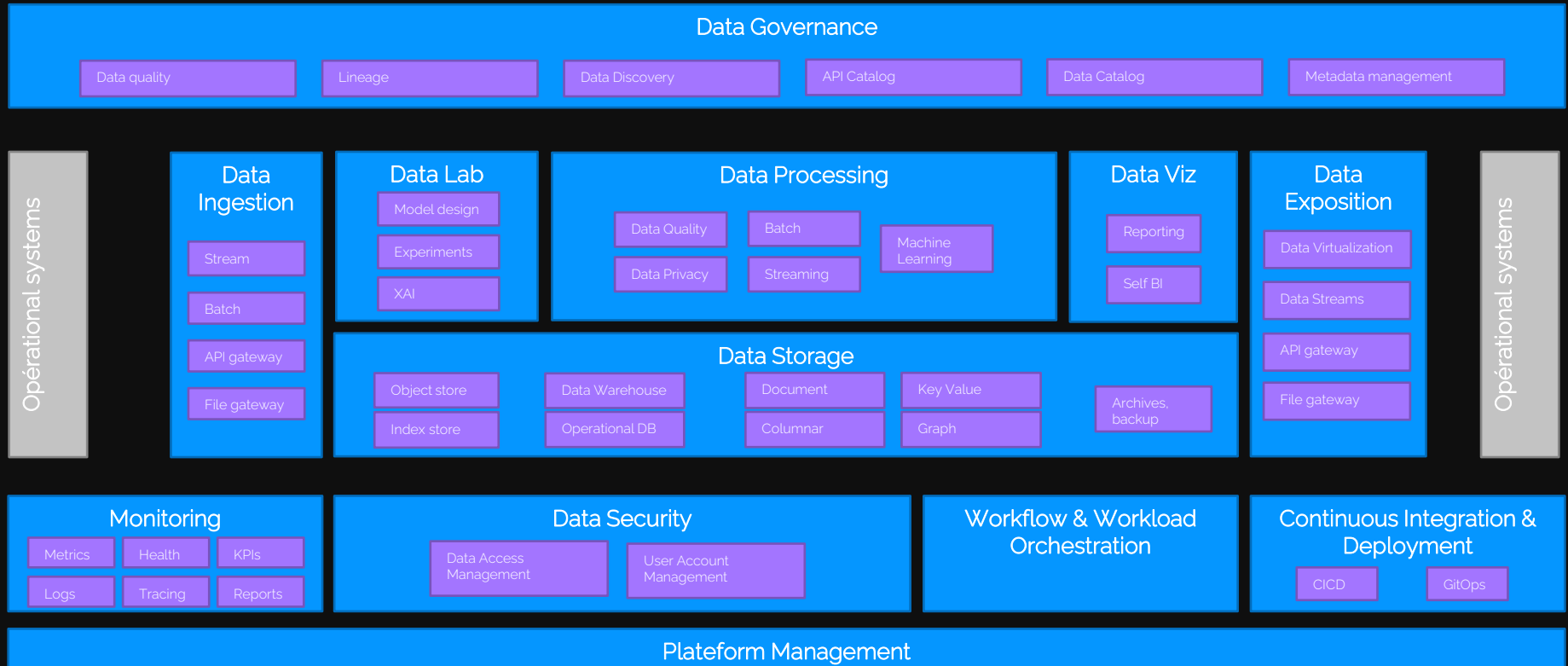
Data Platform Functional Architecture
MLOps Focus

Data Lab

- Studio
- Tracking Experiments
- AutoML
- XAI
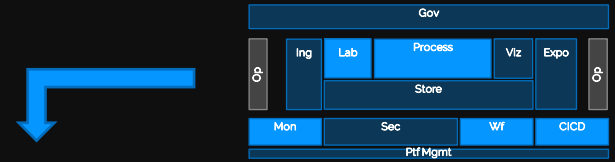
Data Processing - ML pipelines platform

Serving endpoints → ML preprocessing → Data testing → Inference → Monitoring → XAI

Monitoring

- Business impact
- Model performance
- Retrain Automation
- Alerting

Flow Management

- Orchestration

Forge

Devops tools
- Code Repository
- CI/CD - Deploy Automation

MLOPS specific tools
- Model registry
- Feature store

Atos
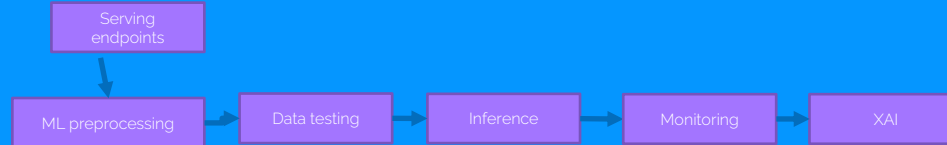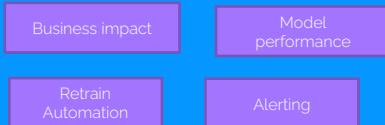
From datalake to mesh
The beginning

Late 1980s — Data Warehouse

Late 2000s — Data Lake

Mid 2010s — Cloud Data Platform

Early 2020s — Data Lakehouse
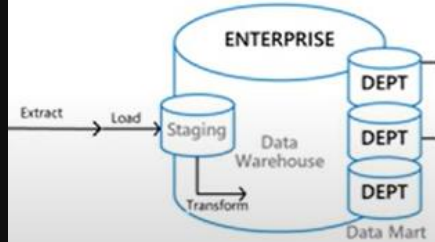
Data Warehouse : structured storage that concentrate all the data of the enterprise, used for analytical purposes (reports, dashboards)
Issue : Hard to scaleup

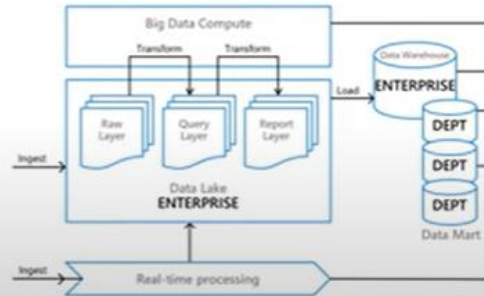**Data Mart** : structured storage oriented for a specific use (reformatting, filtering, renormalization, etc)
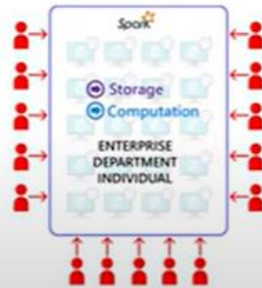
Data Lake : Huge amount of unstructured (and structured) storage with a scalable compute power and a centralized point for data analysis.
Issue : slow (batch oriented technologies), strong coupling between storage and compute
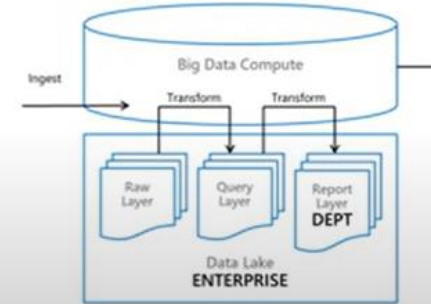
Data Platform : Cloud offering (easy access, agile, scalable) with a complete set of data services : from the enterprise data lake to different complementary products (streaming layer, data mart, etc) with a best of breed approach
Issue : requires strong technical skills to use or operate (especially onprem)
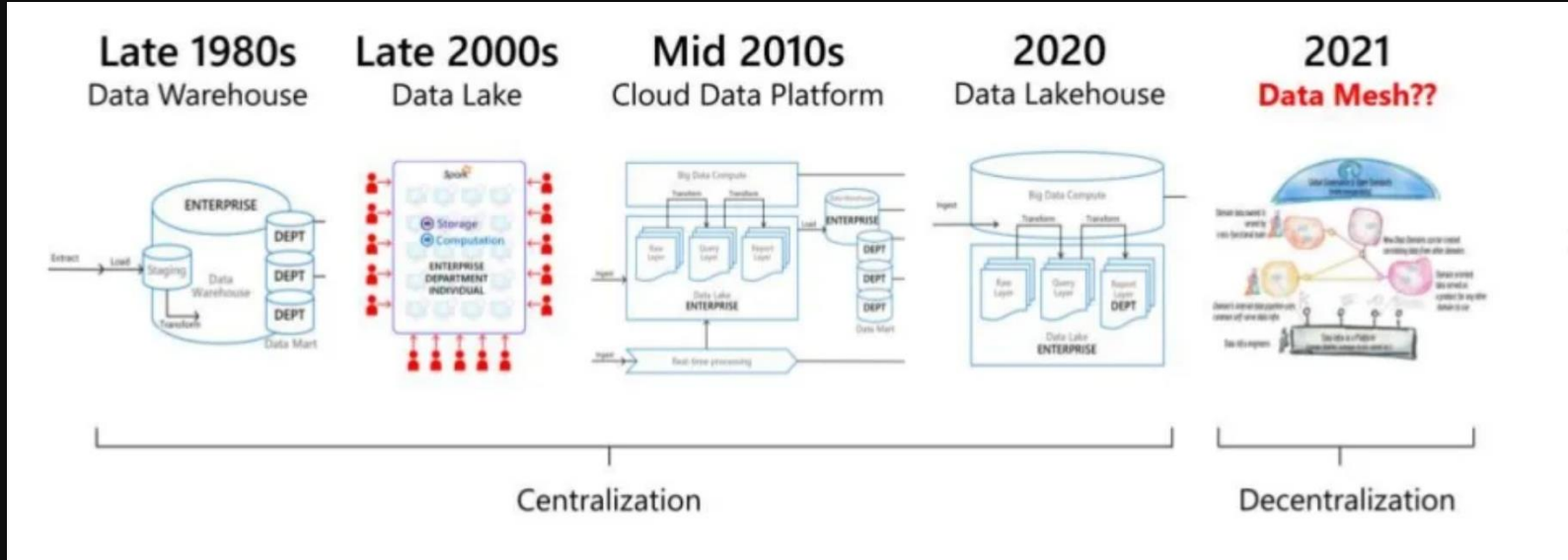
Data Lakehouse : scalable structured processing on top of heterogenous data in a lake
Issue : still a centralized approach with potential bottleneck on central data engineering team

AtoS

From datalake to mesh
Mesh, the new paradigm

Data Mesh : instantiation of several data platform for each business domain of the enterprise that all are connected through the mesh (catalogue, norms, APIs, etc)
-> Full details in data governance course later

Quizz
What we've learn

Question				
In IaaS mode, should we manage the storage layer	Y	N		
In BMaaS mode, can we use our own Middelware	Y	N		
I need to focus on data application development, which service do I need ?	IaaS	CaaS	PaaS	SaaS
Hybride mode is when a client use mutliple cloud providers	Y	N		
A cloud native application is mainly composed of stateless processes	Y	N		
Data exploration is part of MLOps lifecycle	Y	N		
What is NOT an output of the "Presenting Results" phase during Build	Metrics on models quality	Evaluation data	A choosen model	Business vizualisation
DataOps is a technology to industrialize data	Y	N		
Are data warehouses an extinguished specie since data lakes ?	Y	N		
Is there API gateway in an MLOps Architecture ?	Y	N		

AtoS

In Practice
Lab Content

- Discover
  - Notebook on KubeFlow
    - Exo1: explo/viz
      - Getting open data from public api and push it to s3
      - Quick analysis with python
    - Exo2: dwh/viz
      - Push data to CH table + postgres table
      - Visualization with superset
    - Exo3: stream
      - Push un event to a kafka topic
      - Event visualization with akhq
      - Read it form a consumer
      - Bonus : use a kafka engine in CH and see event within superset

AtoS