M2 SIAD Data Science
2021 - 2022

La Data Science Orientée Production

1. Cloud Computing

Corentin Vasseur

SOMMAIRE

1 **Introduction**
- Qu'est ce que le Cloud Computing ?
- 3 grands acteurs
- Les fondamentaux
- Comment l'utiliser ?

2 **Le Stockage**
- Différents types de stockage
- Accès CLI
- Accès SDK

3 **Le Calcul**
- Machines Virtuelles
- Les coûts

Machine Learning 4
- Machine learning system
- ML Pipeline
- Tools chain
- API ML

Platform AI 5
- SageMaker (AWS)
- AI Platform (GCP)
- Azure ML (Microsoft Azure)

Workflow ML 6
- Full Data Engineering Pipeline
- Ex. GCP
- Ex. AWS

1

Introduction : les fondamentaux

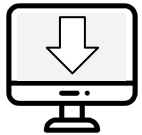
RETOUR SOMMAIRE

Qu'est-ce que le cloud computing ?

On-demand self-service

Broad network access

Ressource pooling



Rapid elasticity

Pay for what you use

3 grands acteurs

AWS

+ 165 services

2006 >

GCP

+ 90 services

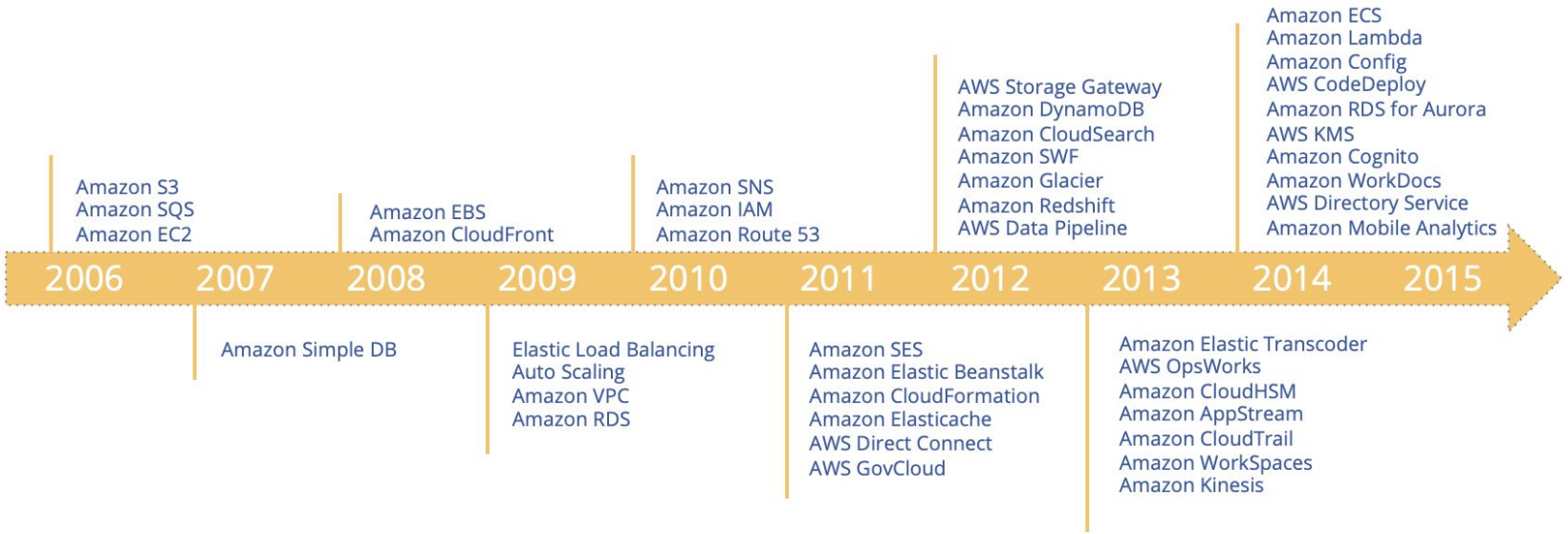
2008 >

Microsoft Azure

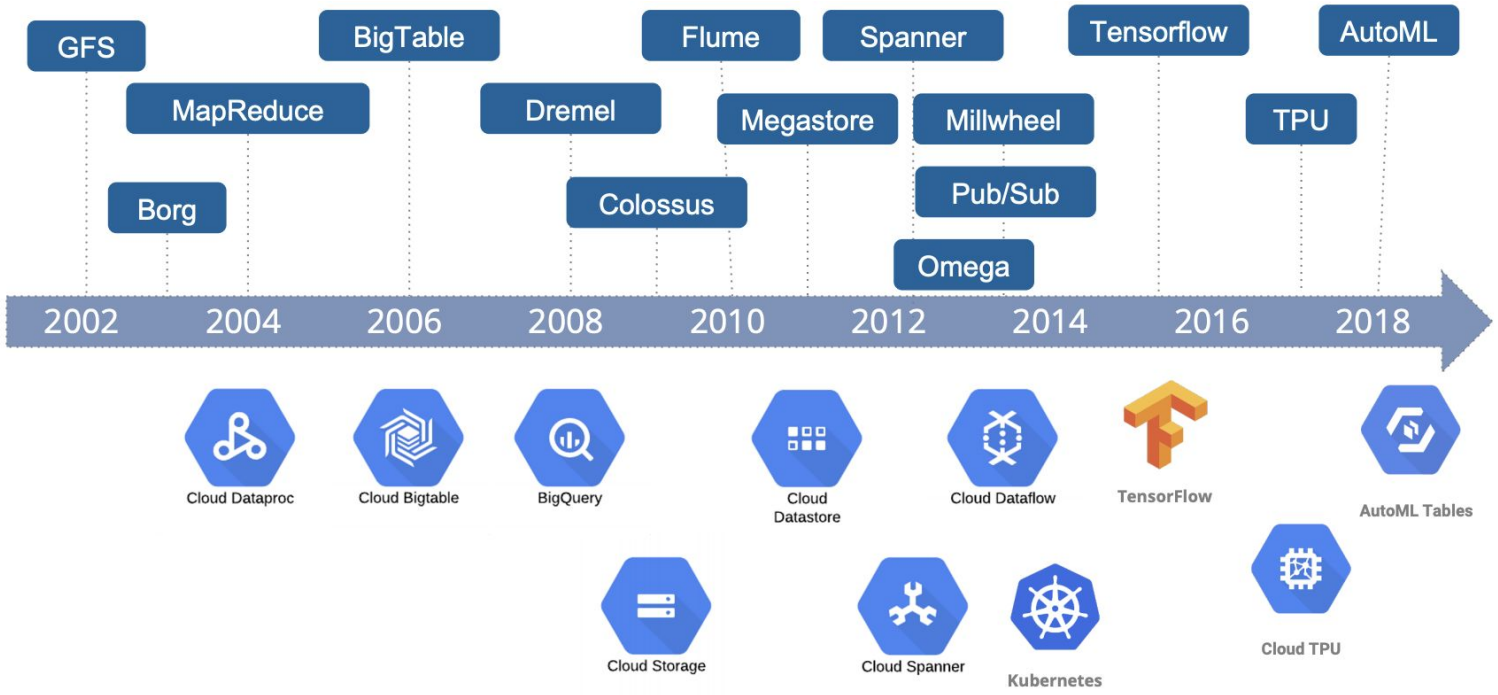
+ 600 services

2010 >

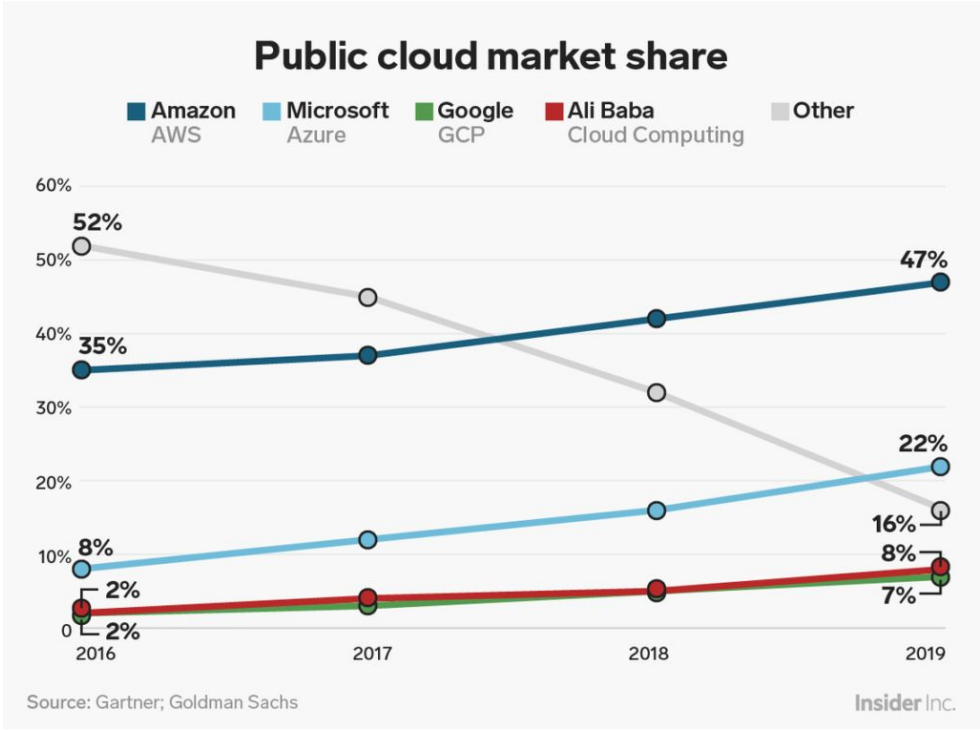
AWS


2006
Amazon S3
Amazon SQS
Amazon EC2
2007
Amazon Simple DB
2008
Amazon EBS
Amazon CloudFront
2009
Elastic Load Balancing
Auto Scaling
Amazon VPC
Amazon RDS
2010
Amazon SNS
Amazon IAM
Amazon Route 53
2011
Amazon SES
Amazon Elastic Beanstalk
Amazon CloudFormation
Amazon Elasticache
AWS Direct Connect
AWS GovCloud
2012
AWS Storage Gateway
Amazon DynamoDB
Amazon CloudSearch
Amazon SWF
Amazon Glacier
Amazon Redshift
AWS Data Pipeline
2013
Amazon Elastic Transcoder
AWS OpsWorks
Amazon CloudHSM
Amazon AppStream
Amazon CloudTrail
Amazon WorkSpaces
Amazon Kinesis
2014
Amazon ECS
Amazon Lambda
Amazon Config
AWS CodeDeploy
Amazon RDS for Aurora
AWS KMS
Amazon Cognito
Amazon WorkDocs
AWS Directory Service
Amazon Mobile Analytics
2015

GCP


GFS
MapReduce
BigTable
Dremel
Flume
Megastore
Spanner
Millwheel
Tensorflow
TPU
AutoML
Borg
Colossus
Pub/Sub
Omega
2002
2004
2006
2008
2010
2012
2014
2016
2018
Cloud Dataproc
Cloud Bigtable
BigQuery
Cloud Datastore
Cloud Dataflow
TensorFlow
AutoML Tables
Cloud Storage
Cloud Spanner
Kubernetes
Cloud TPU

Comment ces acteurs se partagent-ils le marché ?



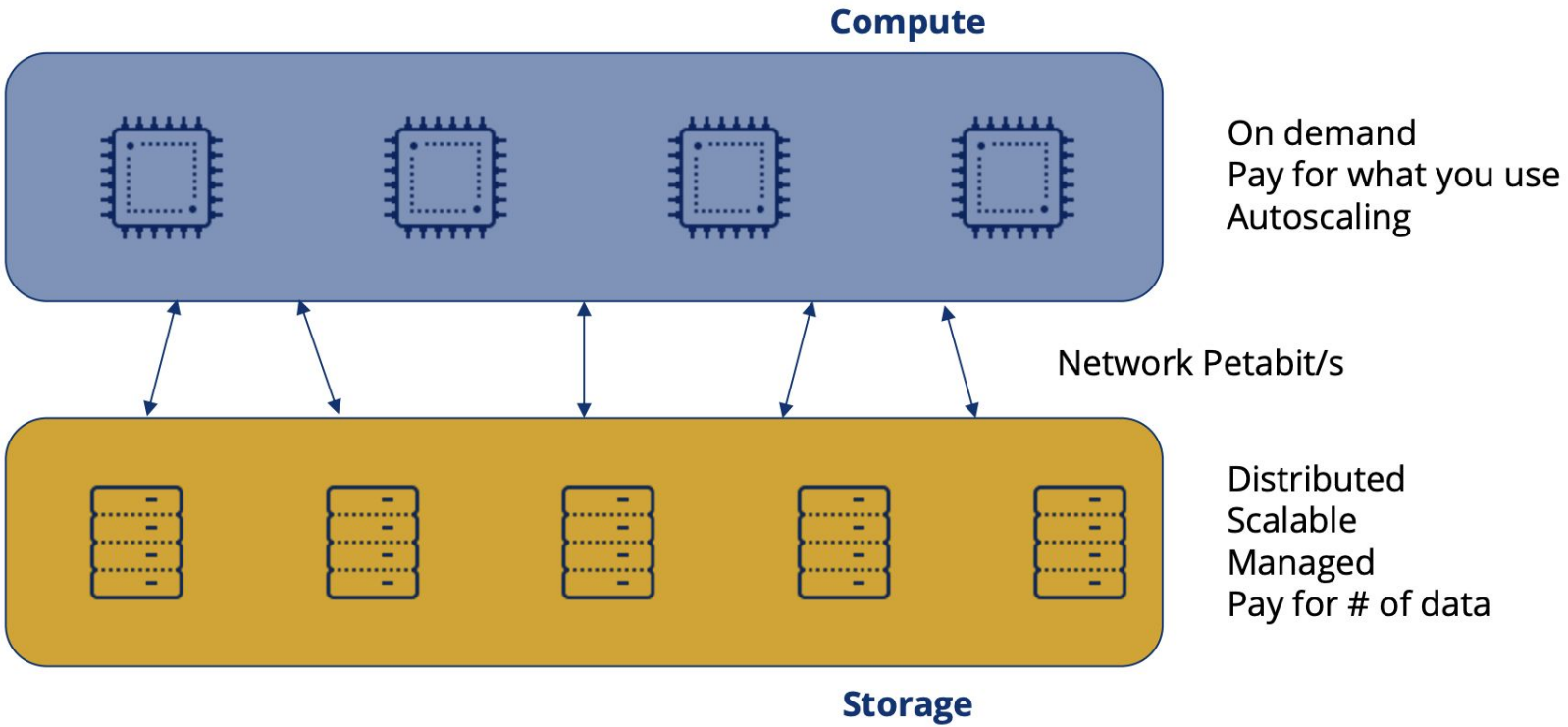
Les solutions

Physical

Virtualized

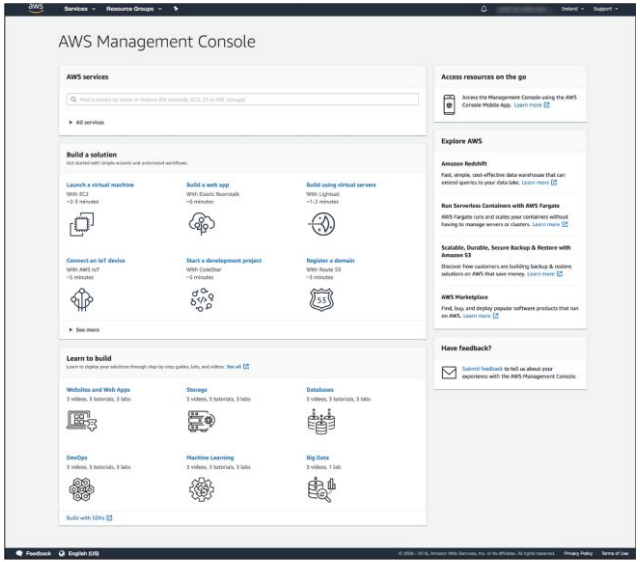
Serverless

Le stockage et le calcul

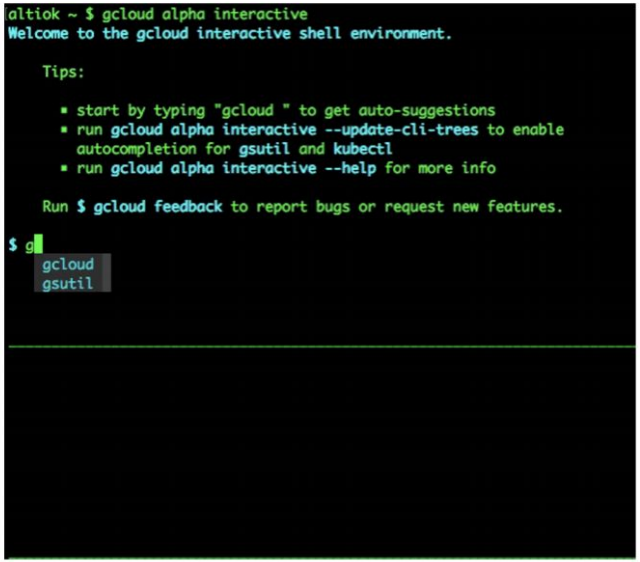


Comment utiliser le cloud ?

UI



CLI



API

In Production - Infrastructure as Code : Terraform, Cloudformation, Deployment Manager, ARM

Le SDK (Software Development Kit)

AWS

Boto 3

```
pip install boto3
```

https://boto3.amazonaws.com/v1/documentation/api/latest/index.html

GCP

```
pip install --upgrade google-cloud-XXX
```

https://github.com/googleapis/google-cloud-python#google-cloud-python-client

Azure :

```
pip install azure-XXX
```

https://pypi.org/project/azure/

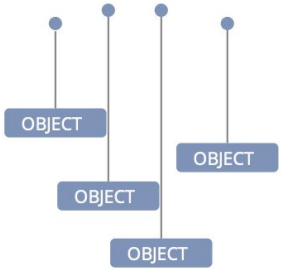
2

Le Stockage

RETOUR SOMMAIRE

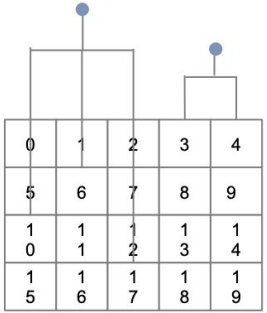
Object Storage vs File Storage vs Block Storage

Object Storage



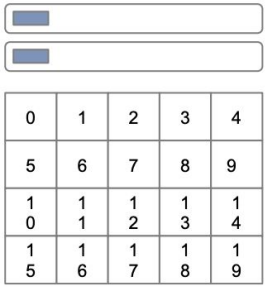
Client app (API)
Internet
Images, PDFs, Video
Custom Metadata

File Storage



User, Client via OS
LAN / 10Gb
user data, web content
Fixed FS Attributes

Block Storage



Operating System
Dedicated Fibre / 100Gb
OS, Database, Transactional
Fixed Sys Attributes

Cloud storage

- Stockage d'objets sécurisé, durable, hautement évolutif et rentable
- Les objets peuvent être structurés ou non structuré
- Facile à utiliser avec une interface de services Web qui rend facile le stockage et le téléchargement des objets
- Il est possible de stocker de grande quantité de données de n'importe où sur le Web
- On parle souvent de 'Data Lake'

Le stockage à plusieurs niveaux

- Stockage d'objets distribué (**Distributed object storage**) ou magasins de valeurs-clés redondants dans lesquels vous pouvez stocker des données objets.
- “**Cool Storage**” ou services de stockage conçus pour stocker des sauvegardes de données. Accès peu fréquent
- Stockage d'archivage (**Cold (archival) storage**) ou services de stockage conçus pour stocker des données d'archivage à des fins de conformité ou à des fins d'analyse.

AWS S3
Standard
IA
One Zone IA
Intelligent Tiering
Glacier
Glacier Deep Archive

Google Cloud Storage
Standard
Nearline
Coldline
Archive

Azure Blob Storage
with GRS or RA-GRS
with ZRS
with LRS
Cool tier
Archive Tier

Accès avec CLI

https://docs.aws.amazon.com/cli/latest/reference/s3/index.html

https://aws.amazon.com/blogs/developer/leveraging-the-s3-and-s3api-commands/

```
aws s3 mb s3://mybucket
```

```
aws s3 cp test.txt s3://mybucket/test2.txt
```

```
aws s3 cp myDir s3://mybucket/ --recursive --exclude "*.jpg"
```

https://cloud.google.com/storage/docs/gsutil

```
gsutil mb -l asia gs://some-bucket
```

```
gsutil cp *.txt gs://my-bucket
```

```
gsutil cp -r dir gs://my-bucket
```

Accès avec le SDK

AWS

Boto 3

```
pip install boto3
```

https://boto3.amazonaws.com/v1/documentation/api/latest/index.html

GCP

google-cloud-storage

```
pip install --upgrade google-cloud-storage
```

https://pypi.org/project/google-cloud-storage/

Azure :

```
pip install azure-storage-blob
```

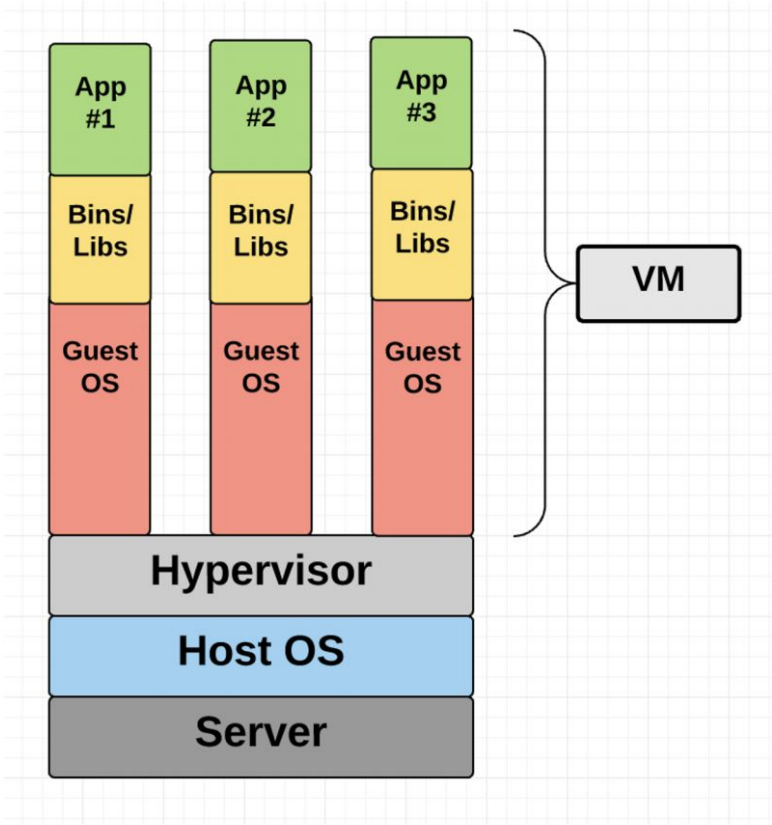
https://github.com/Azure/azure-sdk-for-python

3

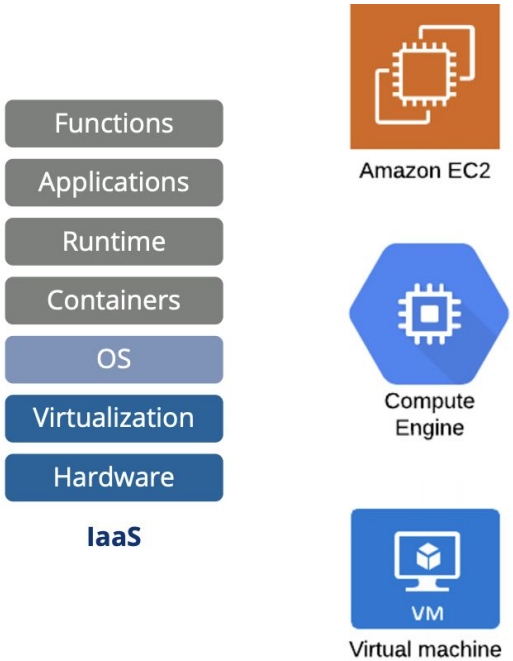
Le calcul

RETOUR SOMMAIRE

Les machines virtuelles



Les machines virtuelles



Secure, resizable compute capacity in the cloud

- Scalable and high-performance virtual machines
- Create Linux and Windows virtual machines (VMs) in seconds.

AWS : EC2 for ML

EMR supports GPU instances types

Some EC2 instances types targeted at ML tasks
- compute optimized
- accelerated computing (GPUs)

ml.* instances only for SageMaker

Appropriates instance for deep learning :
- P3: 8 Tesla V100 GPU's
- P2: 16 K80 GPU's
- G3: 4 M60 GPU's (all Nvidia chips)

Conda-based Deep Learning AMI's
AWS Deep Learning AMIs
Amazon Elastic Inference

Librairies
- Tensorflow
- Keras
- MXNet
- Gluon
- Pytorch
- ...

GPU acceleration:
- CUDA
- cuDNN
- NCCL
- NVidia Driver

GCP : Compute Engine for ML

GPUs are independent resources

- can be added to some GCE instances (depends on regions/zones too)

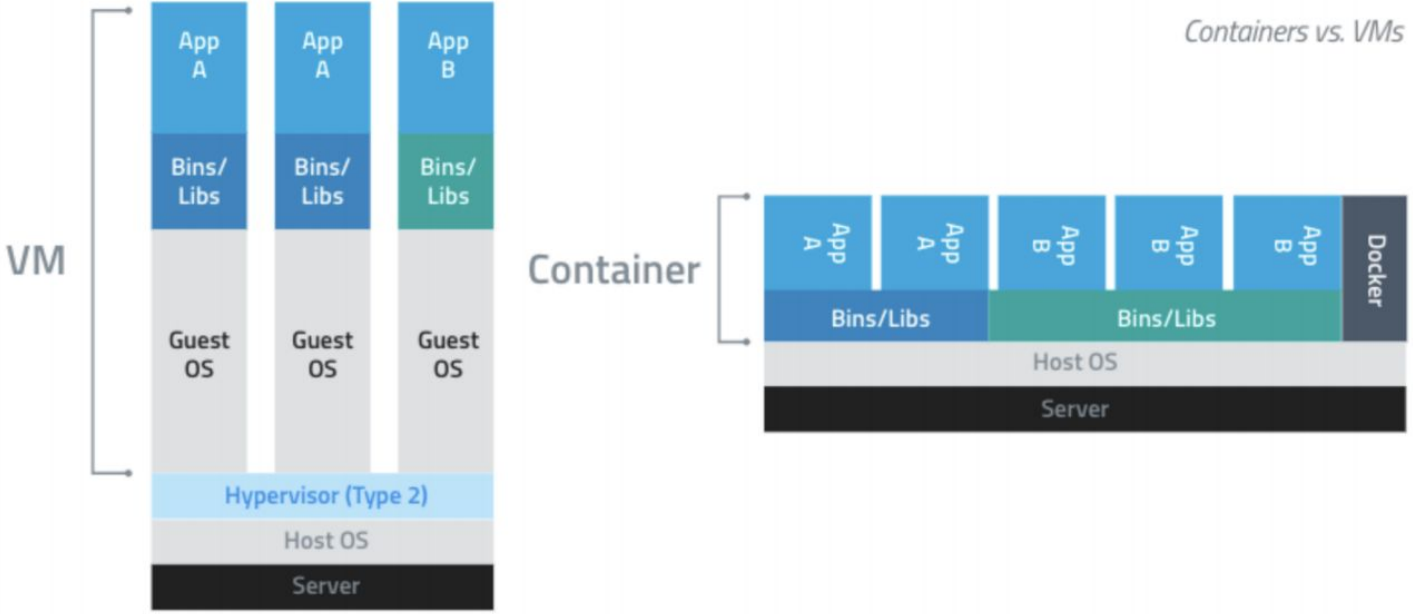
Deep Learning VMs : Deep Learning VM

Cloud TPU : NEXT TPU

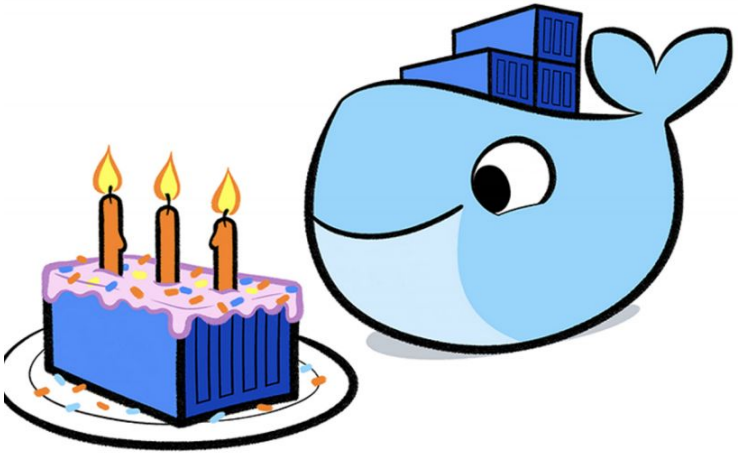
Les coûts

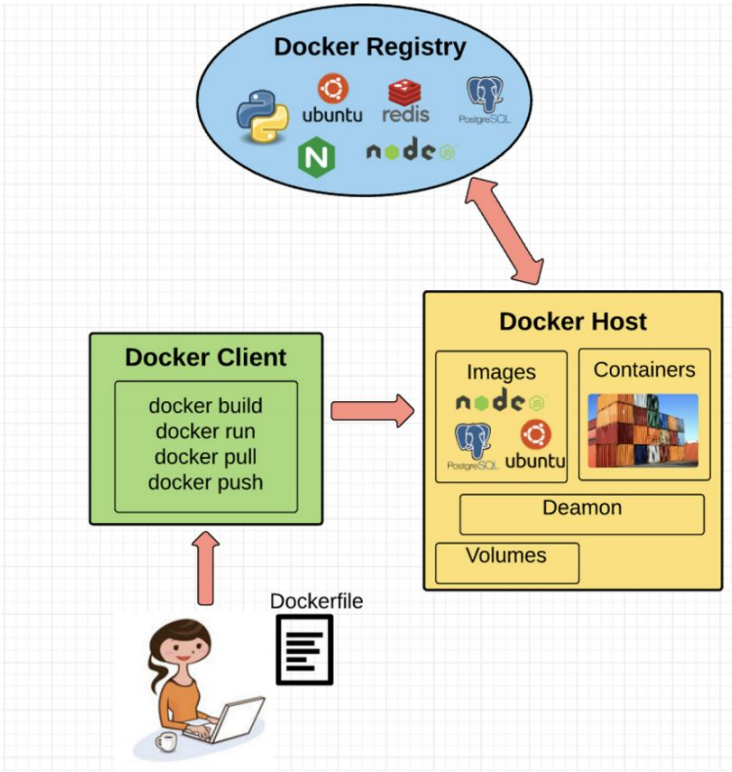
On demand	Reserved	Interruptible	Sole Tenant
On-demand Pay as you go	Reserved Sustained use discounts Committed use discounts Sole-tenant nodes	Spot Preemptible instances	Dedicated Sole-tenant nodes
$ $ $	$ $	$	$ $ $ $

Les containers

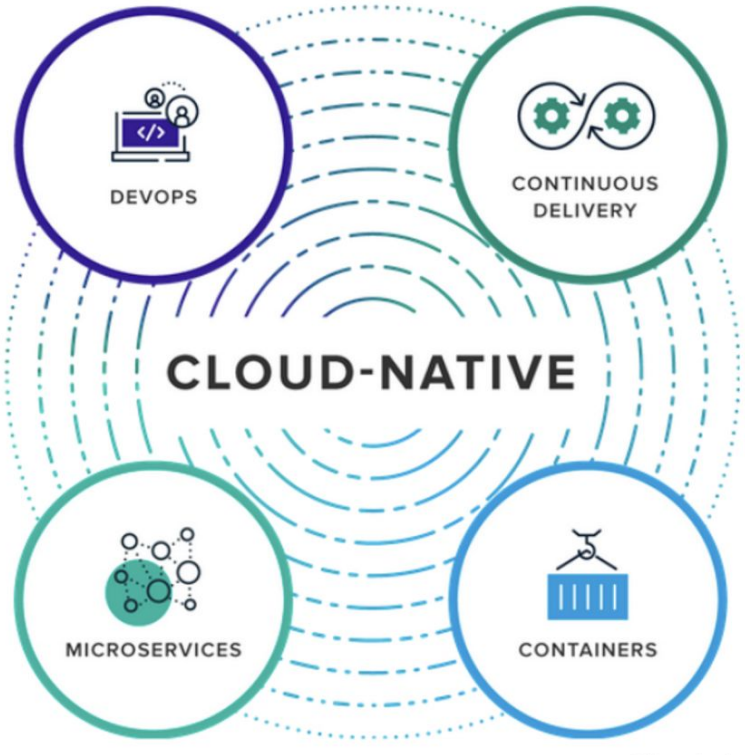


Les containers : docker





Solutions Cloud Native



Docker python web app

app.py

```
from flask import Flask
app = Flask(__name__)

@app.route("/")
def hello():
  return "Hello World!\n"

@app.route("/version")
def version():
  return "Helloworld 1.0\n"

if __name__ == "__main__":
  app.run(host='0.0.0.0')
```

requirements.txt

```
Flask==0.12
uwsgi==2.0.15
```

Dockerfile

```
FROM ubuntu:18.10
RUN apt-get update -y && \
    apt-get install -y python3-pip python3-dev
COPY requirements.txt /app/requirements.txt
WORKDIR /app
RUN pip3 install -r requirements.txt
COPY . /app
ENTRYPOINT ["python3", "app.py"]
```

4

Machine Learning Engineering

RETOUR SOMMAIRE

Machine Learning System

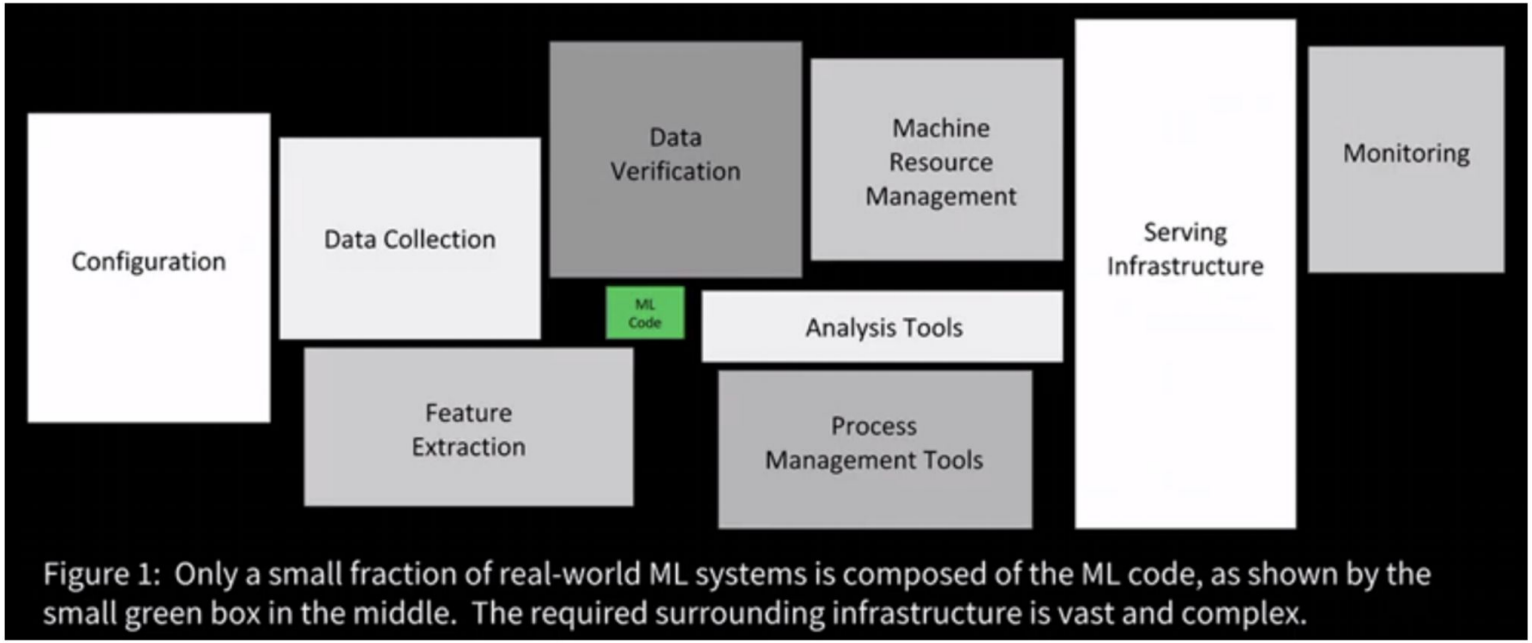
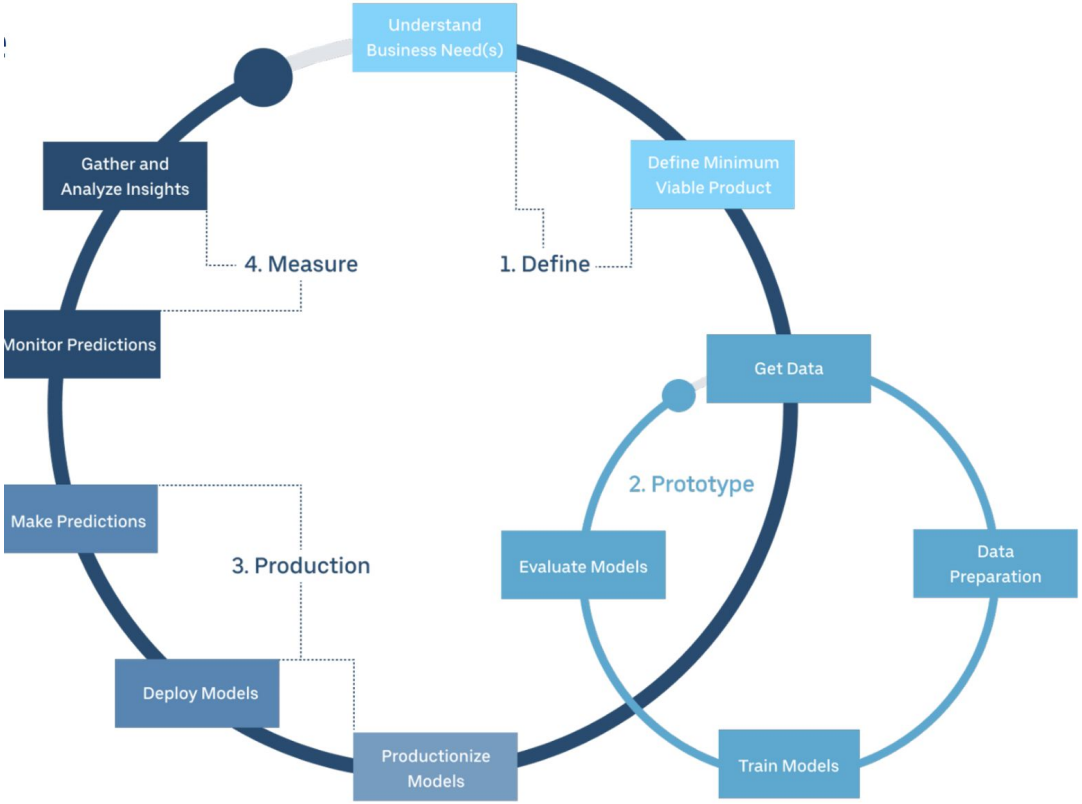


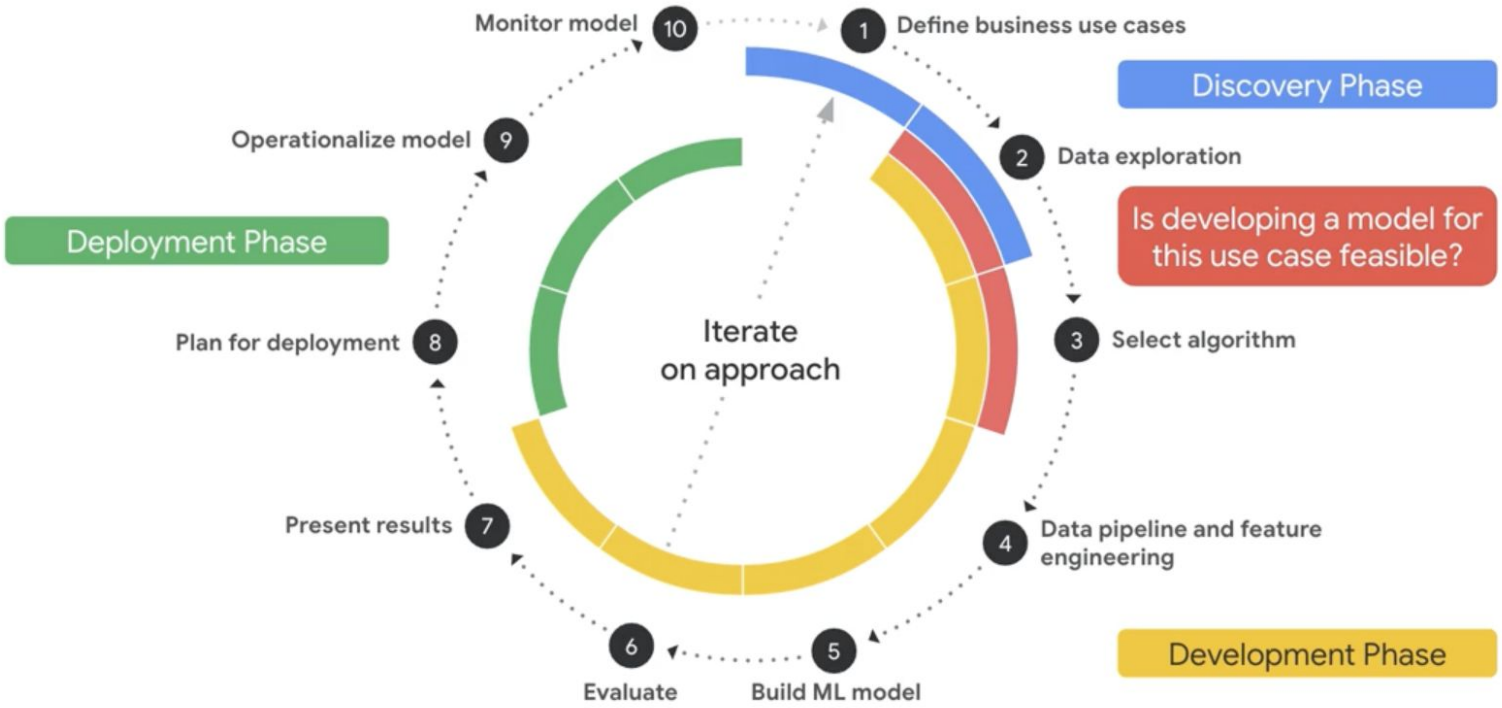
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

Source : Hidden Technical Debt in Machine Learning Systems, Google NIPS 2015

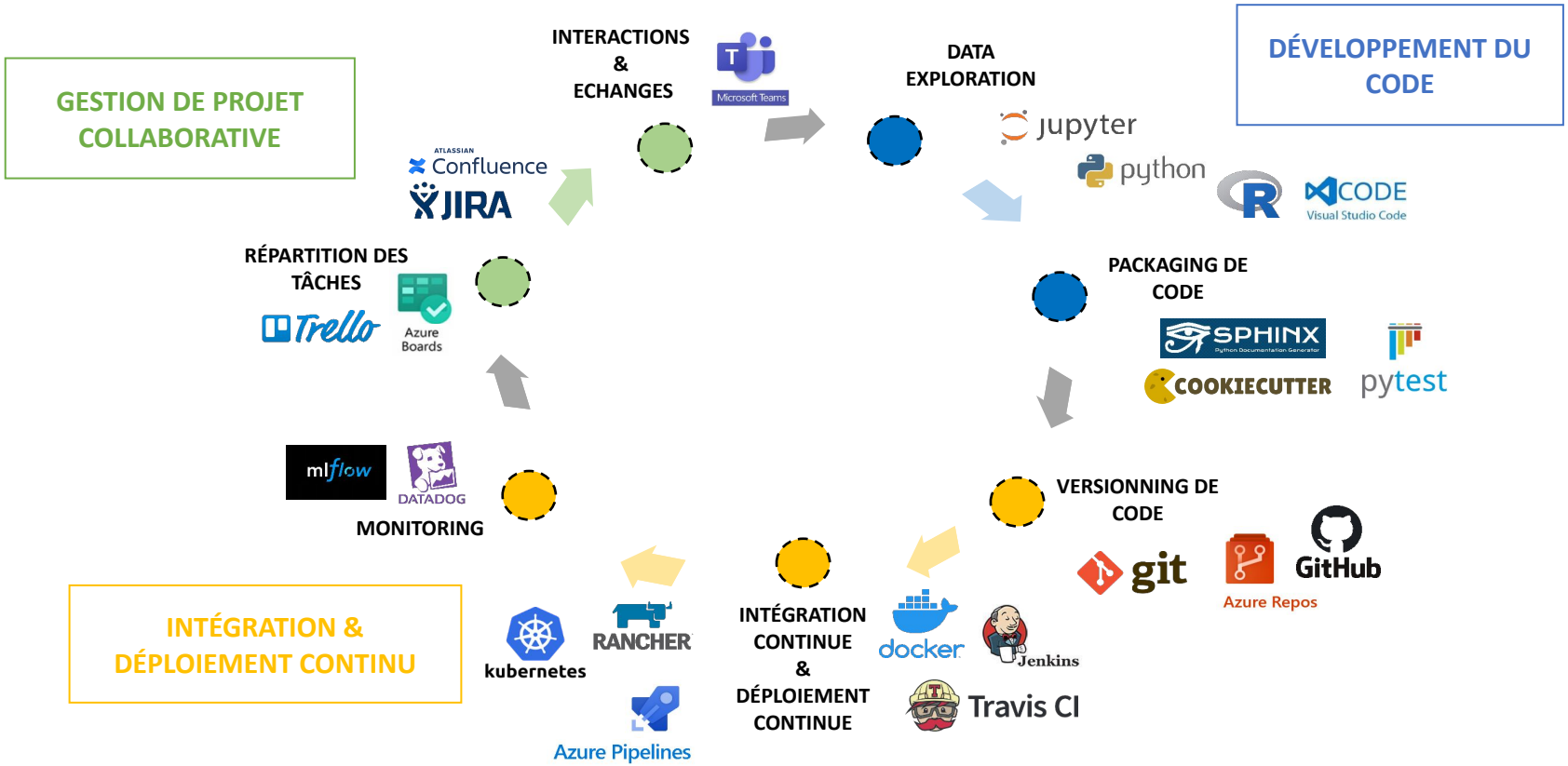
ML Pipeline


Understand Business Need(s)
Define Minimum Viable Product
1. Define
Get Data
2. Prototype
Data Preparation
Train Models
Evaluate Models
Productionize Models
3. Production
Deploy Models
Make Predictions
Monitor Predictions
4. Measure
Gather and Analyze Insights

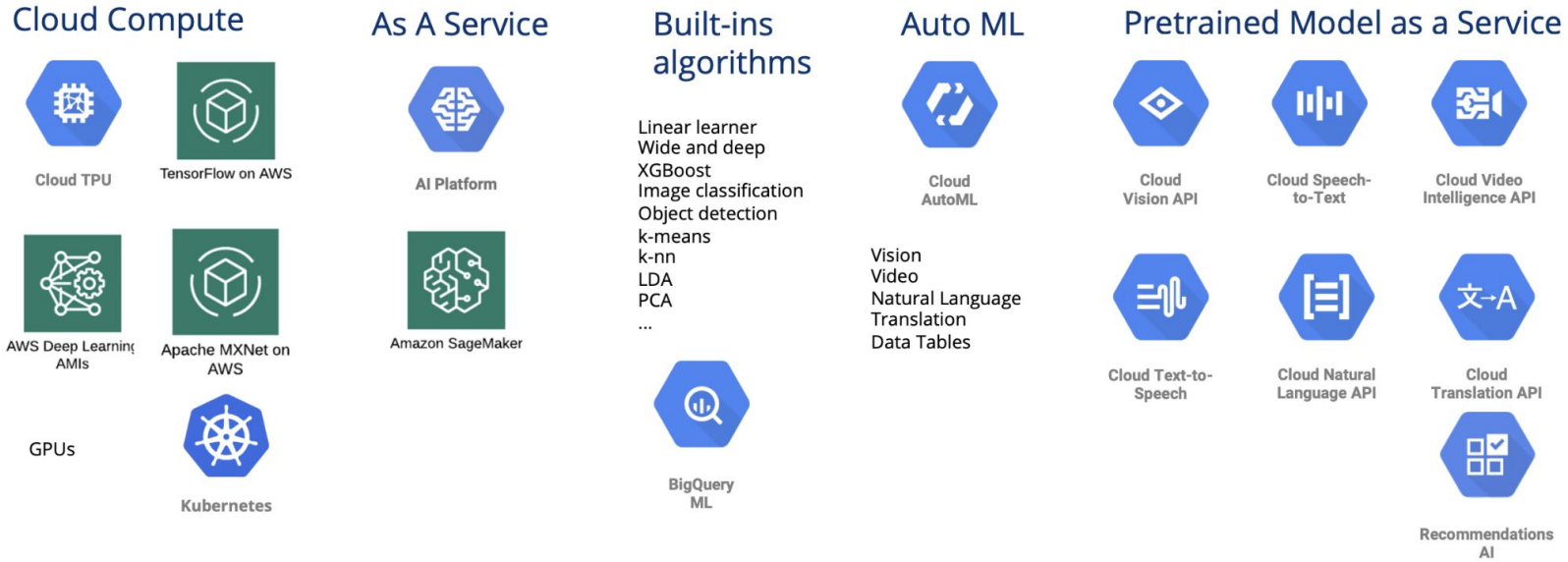
ML Pipeline


Monitor model
10
1
Define business use cases
Discovery Phase
Operationalize model
9
2
Data exploration
Is developing a model for this use case feasible?
Deployment Phase
Plan for deployment
8
Iterate on approach
3
Select algorithm
Present results
7
4
Data pipeline and feature engineering
6
5
Development Phase
Evaluate
Build ML model

Tools chain


GESTION DE PROJET COLLABORATIVE
INTERACTIONS & ECHANGES
Microsoft Teams
DATA EXPLORATION
DÉVELOPPEMENT DU CODE
jupyter
python
CODE
Visual Studio Code
ATLASSIAN
Confluence
JIRA
RÉPARTITION DES TÂCHES
Trello
Azure Boards
PACKAGING DE CODE
SPHINX
COOKIECUTTER
pytest
mlflow
DATADOG
MONITORING
VERSIONNING DE CODE
git
GitHub
Azure Repos
INTÉGRATION & DÉPLOIEMENT CONTINU
kubernetes
RANCHER
Azure Pipelines
INTÉGRATION CONTINUE & DÉPLOIEMENT CONTINUE
docker
Jenkins
Travis CI

ML options on the cloud

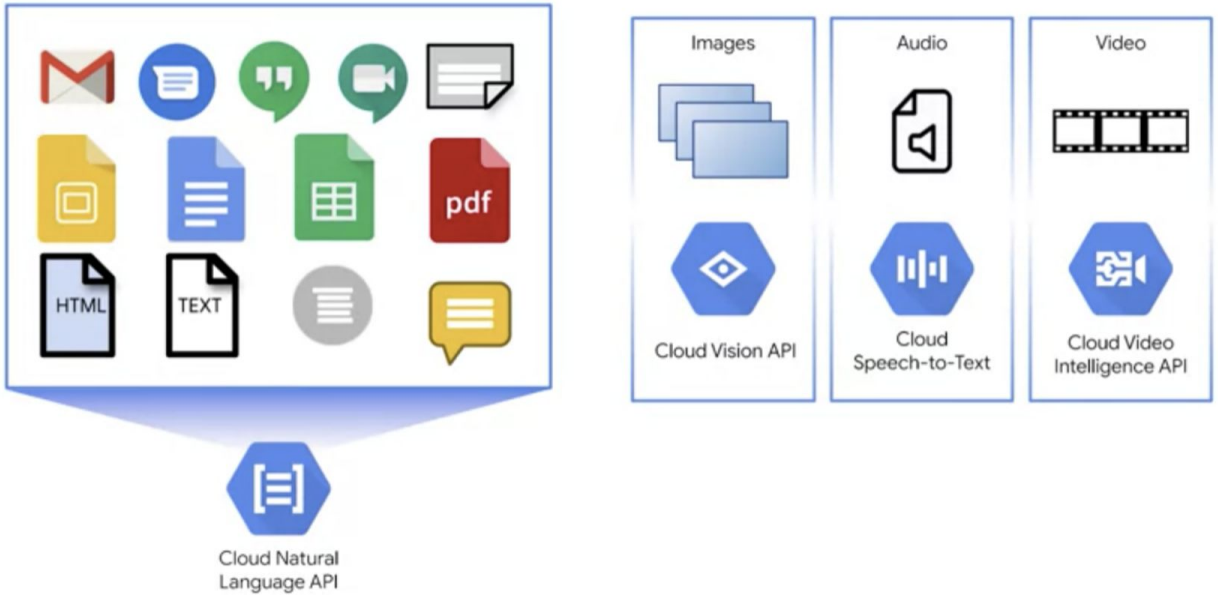


More control for advanced users

Bring your own data

Ready to go

API ML (GCP)


pdf
HTML
TEXT
Cloud Natural Language API
Images
Cloud Vision API
Audio
Cloud Speech-to-Text
Video
Cloud Video Intelligence API

API ML (AWS)

ımazon Comprehend

Amazon Forecast

Amazon Personalize

Amazon Textract

Amazon Translate

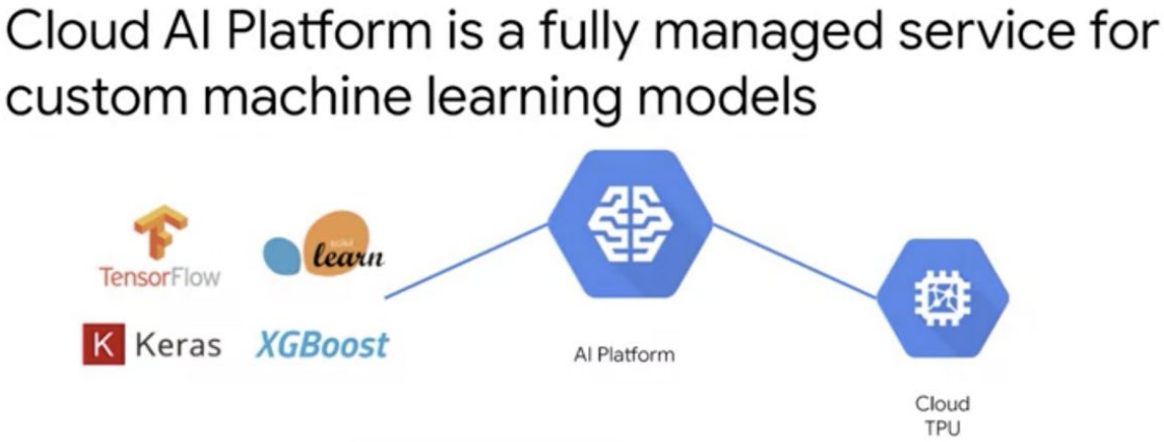
Amazon Lex

Amazon Polly

Amazon Rekognition

Amazon Transcribe

ML Platform : AI Platform (GCP)



ML Platform : SageMaker (AWS)

Amazon SageMaker

Build

- Preprocessing
- Ground Truth
- Notebooks

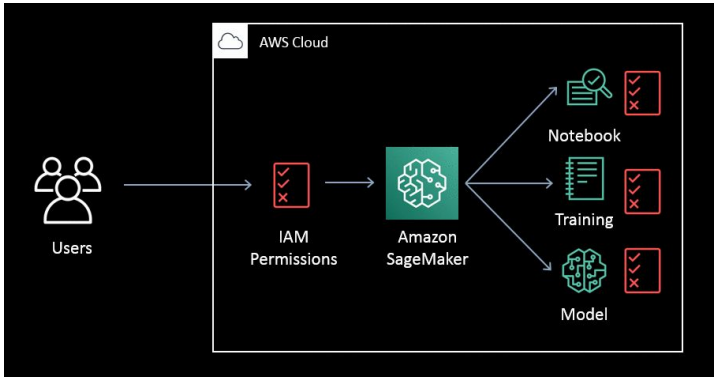
Train

- Built-in algorithms
- Hyperparameter tuning
- Notebooks
- Infrastructure

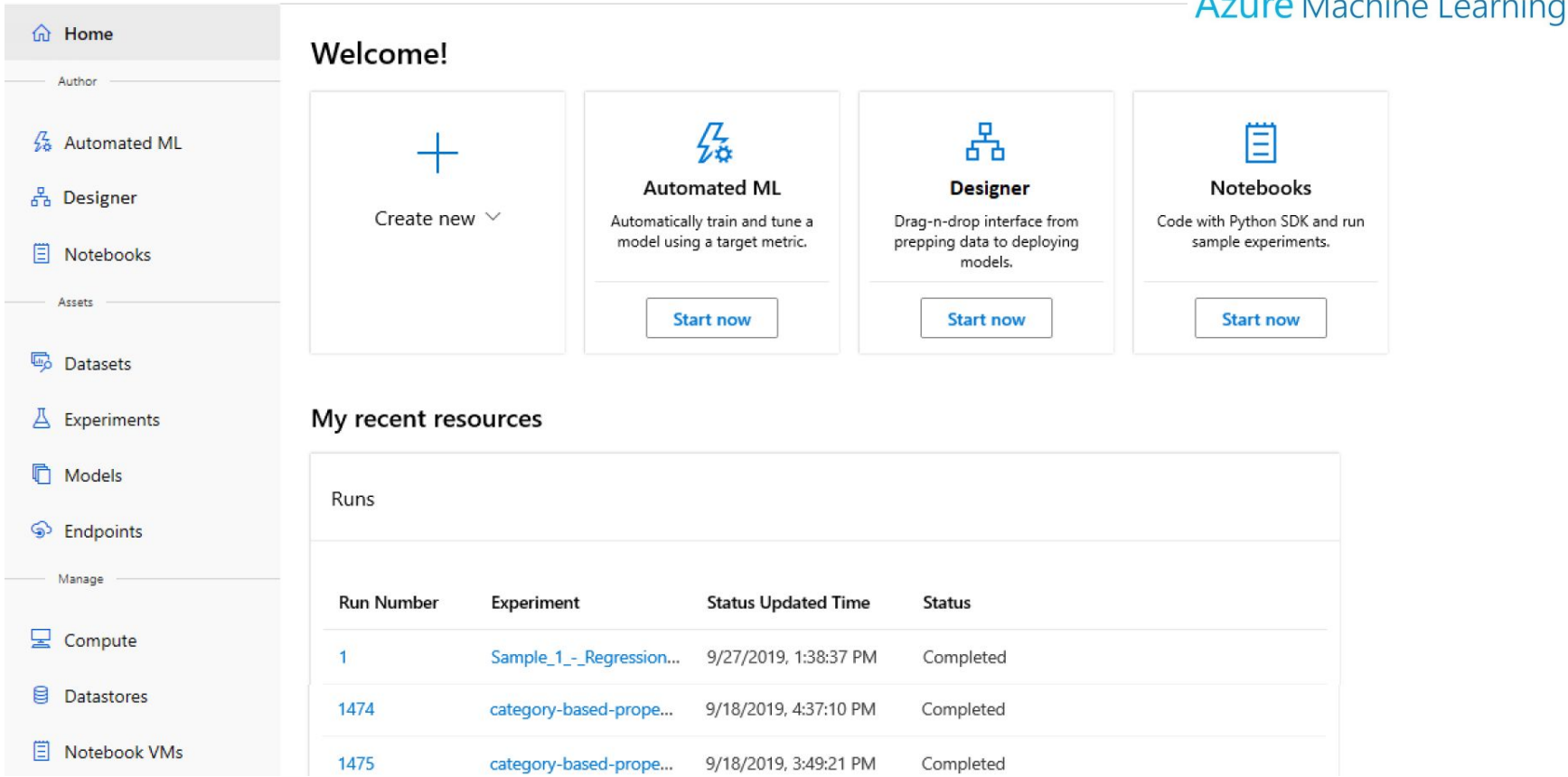
Deploy

- Realtime
- Batch
- Infrastructure
- Notebooks
- Neo (Edge)

Control :
Console (UI), SDK, Notebooks



ML Platform : Azure ML

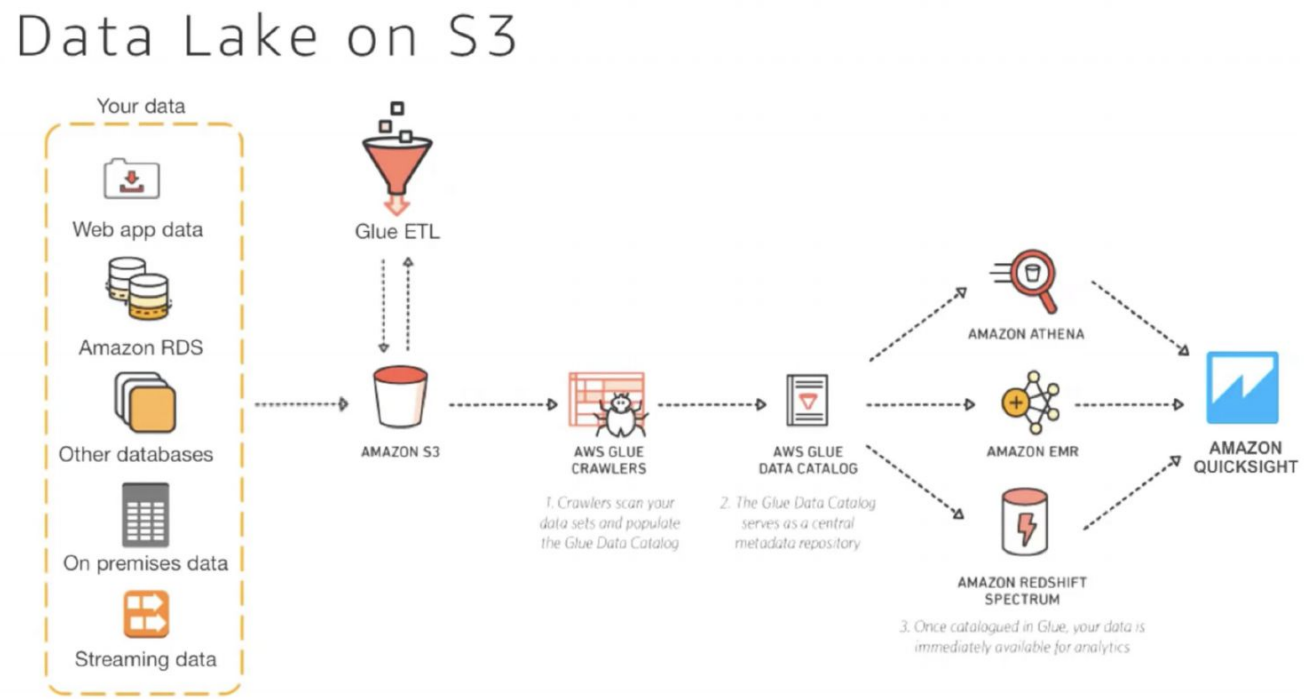


5

Ex. Workflow

RETOUR SOMMAIRE

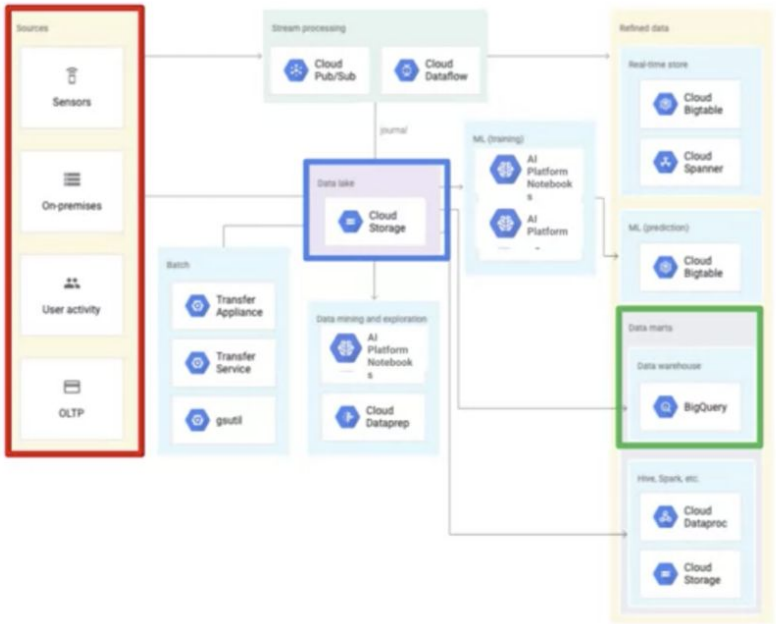
Full Data Engineering Pipeline : AWS



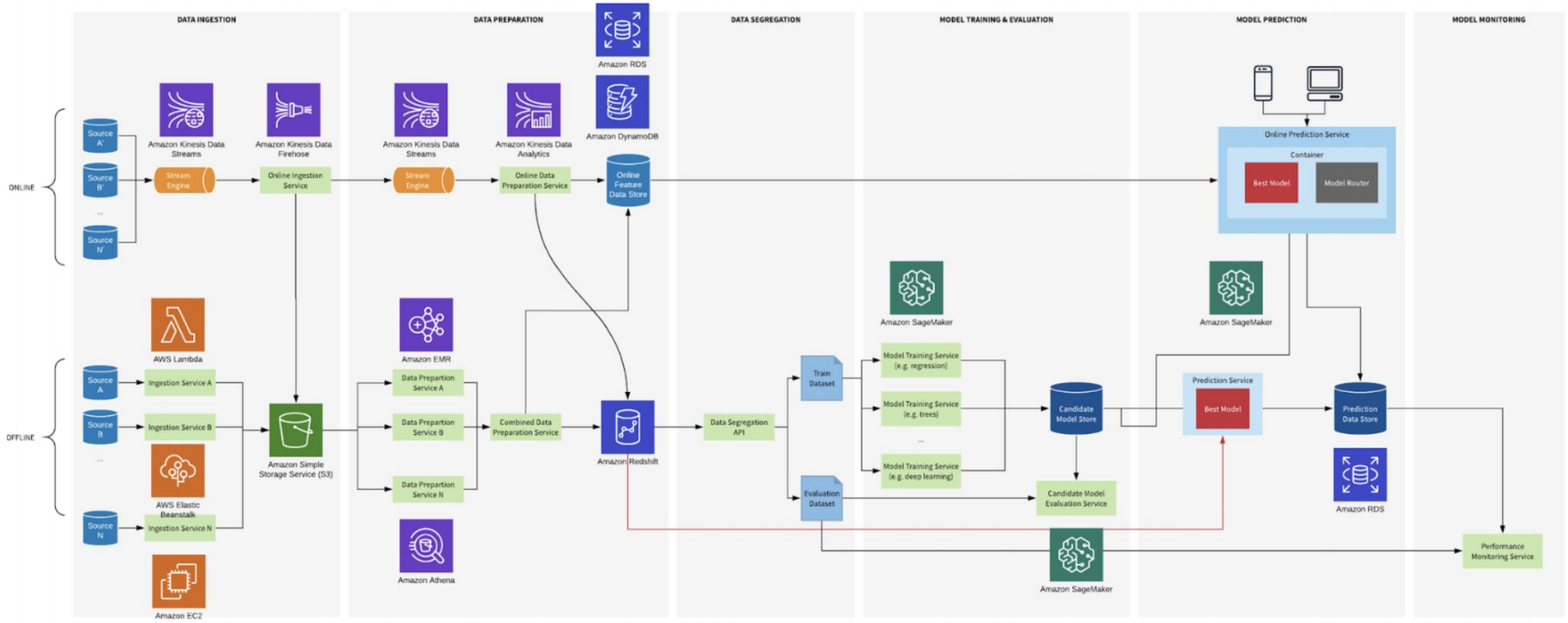
Full Data Engineering Pipeline : GCP

Example Architecture

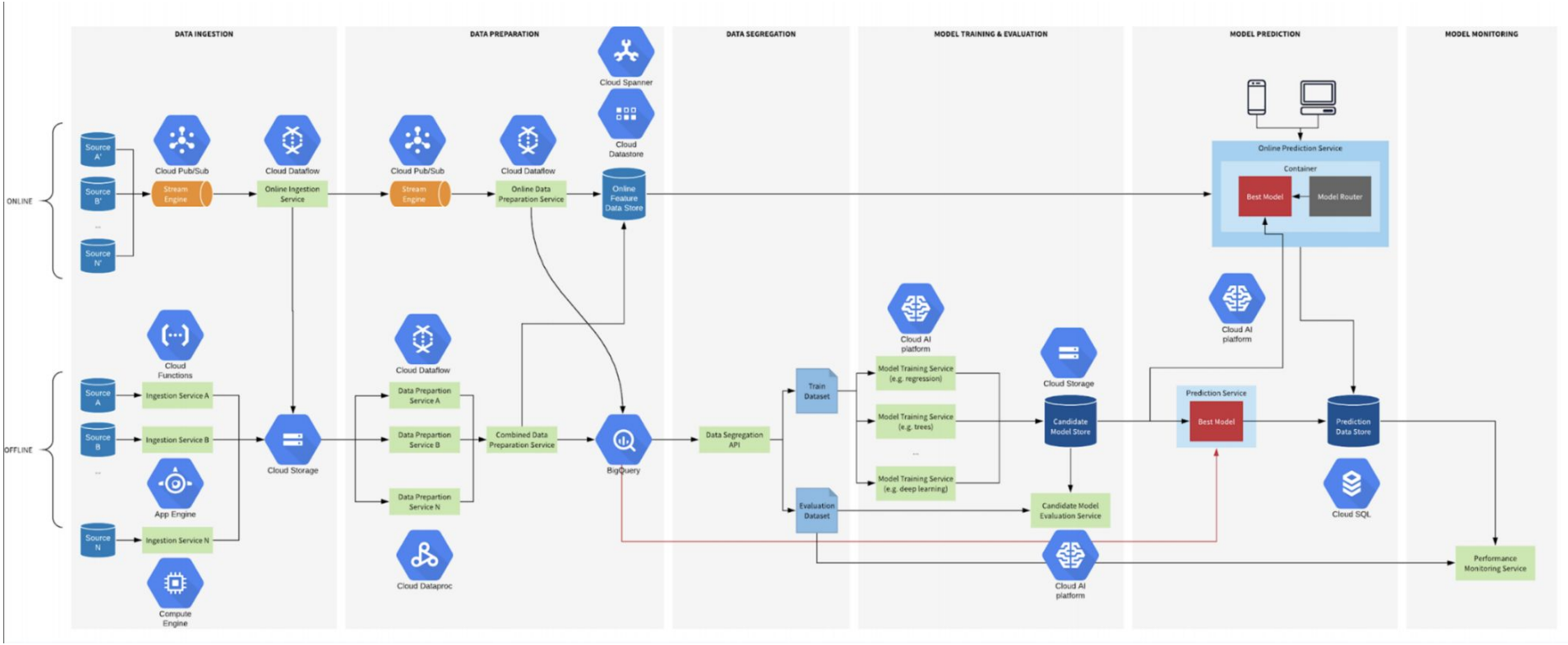
1. Data sources
2. Data Lake
3. **Data Pipelines**
4. Data Warehouse
5. Used for ML and analytics workloads



Ex. full ML Pipeline (AWS)



Ex. full ML Pipeline (GCP)



UNE QUESTION ?

corentin.vasseur@decathlon.com

Machine Learning Engineer - MLOps