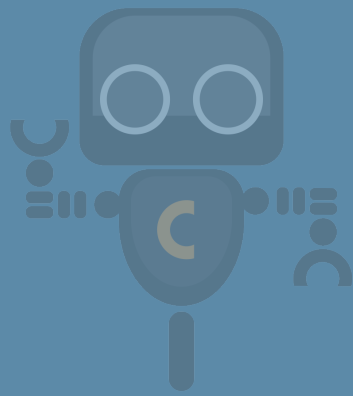
WGESAD 2020

Aula 1



Francisco Nauber Bernardo Gois

Canal da Ciência

27 de fevereiro de 2

Instrutor

F. Nauber Bernardo Gois

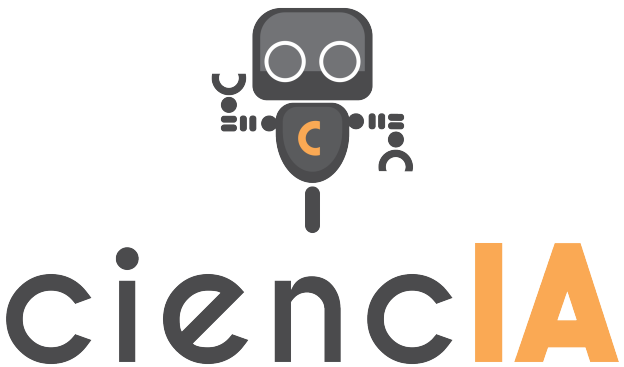
Dsc. Informática Aplicada
Lider de Aprendizado de Máquina na Secretaria de Saúde do Ceará
Engenheiro de Aprendizado de Máquina
Professor da UFC (2018-2019)
Analista de Desenvolvimento Serpro (2004-2018)

Instrutor

https://www.linkedin.com/in/n

https://www.linkedin.com
/in/naubergois/
Envie recomendações, depoimentos e competÊncias

Instrutor



https://www.youtube.com/can

https://www.youtube.com/
canaldaciencia
Instagram: @canaldaciencia

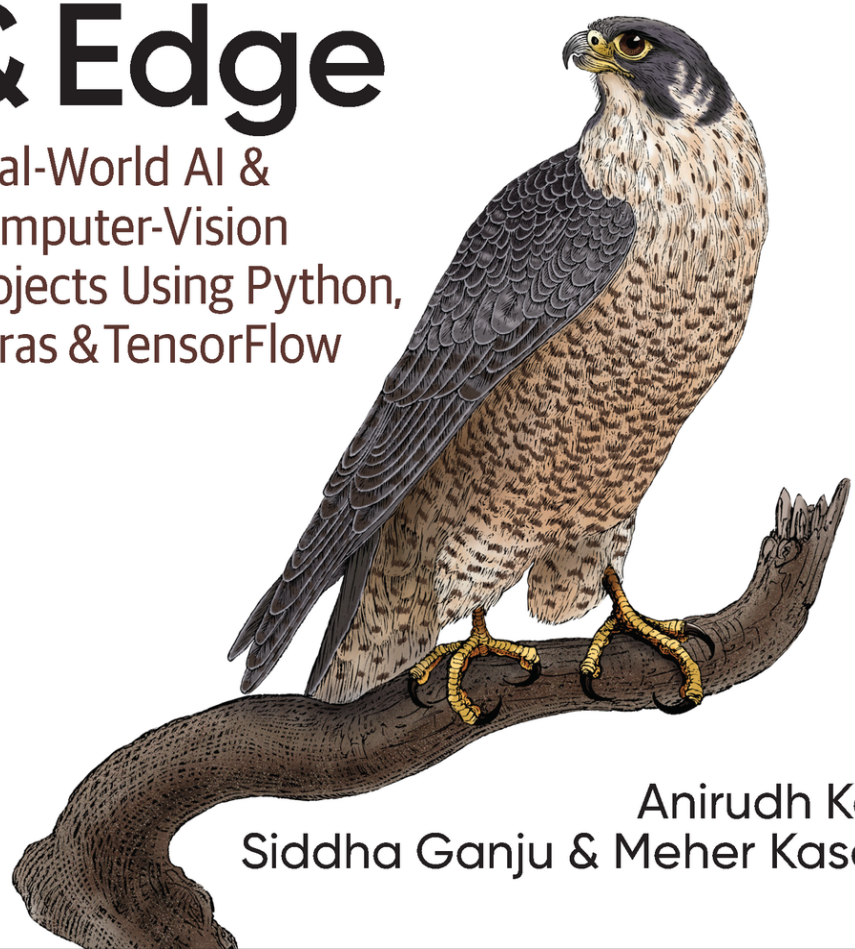
Índice

Índice

O'REILLY®

Practical Deep Learning for Cloud, Mobile & Edge

Real-World AI & Computer-Vision Projects Using Python, Keras & TensorFlow



Anirudh Koul,
Siddha Ganju & Meher Kasam

O'REILLY®

The AI Organization

Learn from Real Companies and Microsoft's Journey How to Redefine Your Organization with AI

David Carmona

O'REILLY®

Building Machine Learning Pipelines

Automating Model Life Cycles with TensorFlow

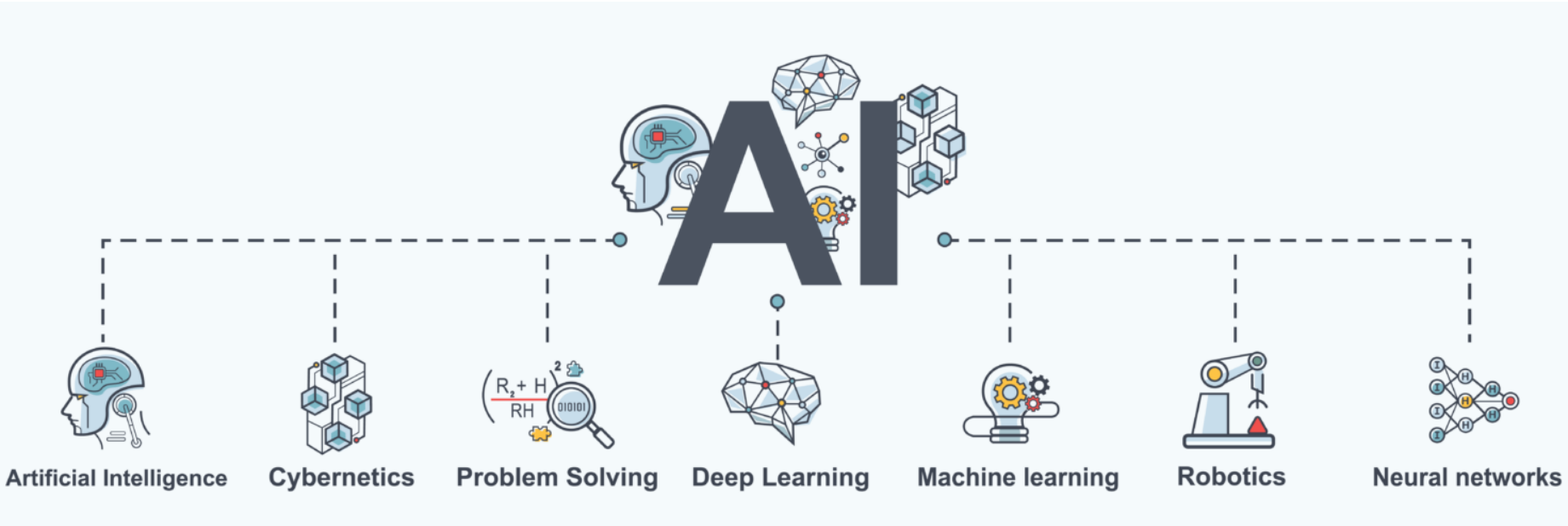
Early Release

RAW & UNEDITED

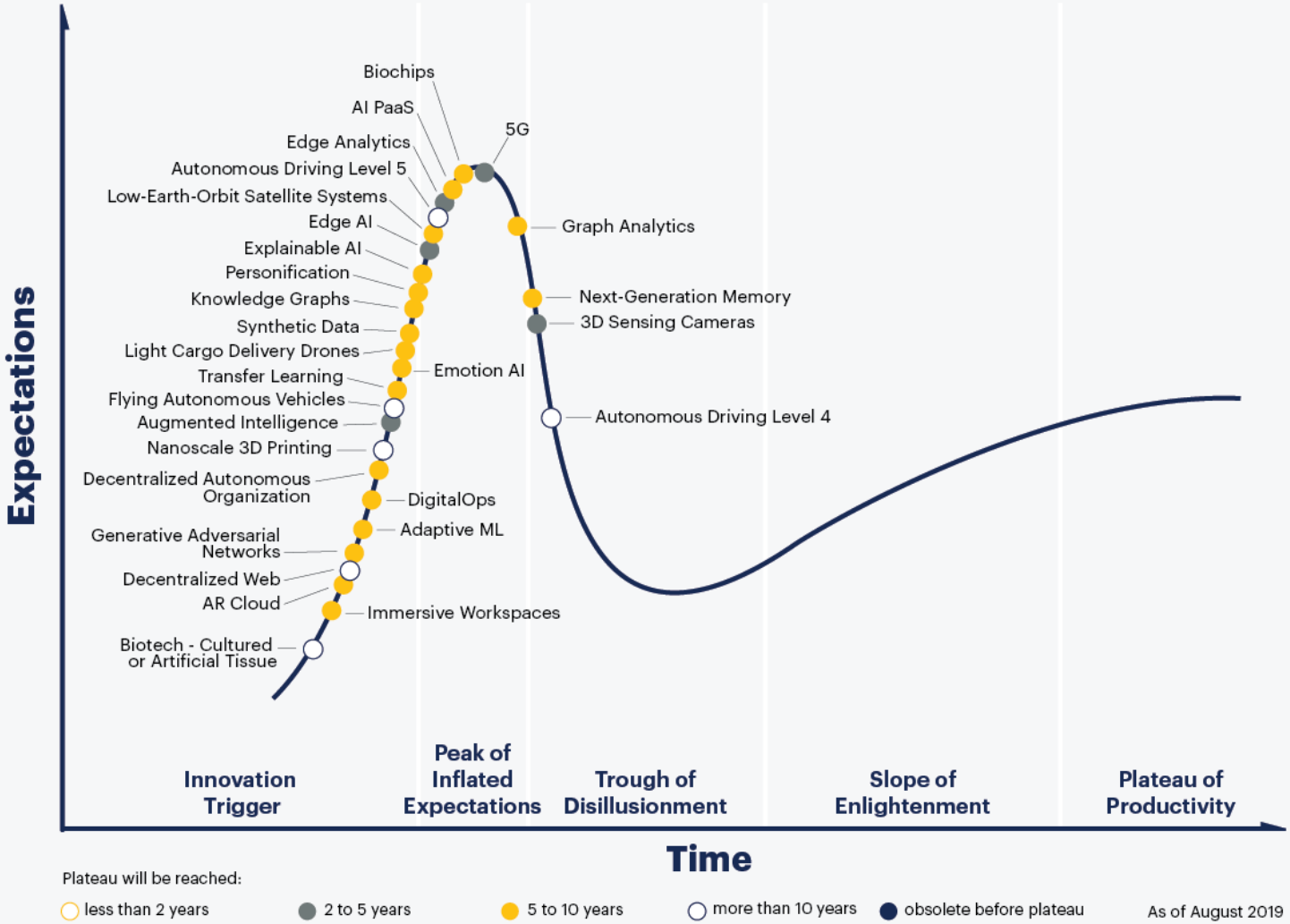
Hannes Hapke &
Catherine Nelson

Índice

Índice


AI
Artificial Intelligence
Cybernetics
Problem Solving
Deep Learning
Machine learning
Robotics
Neural networks

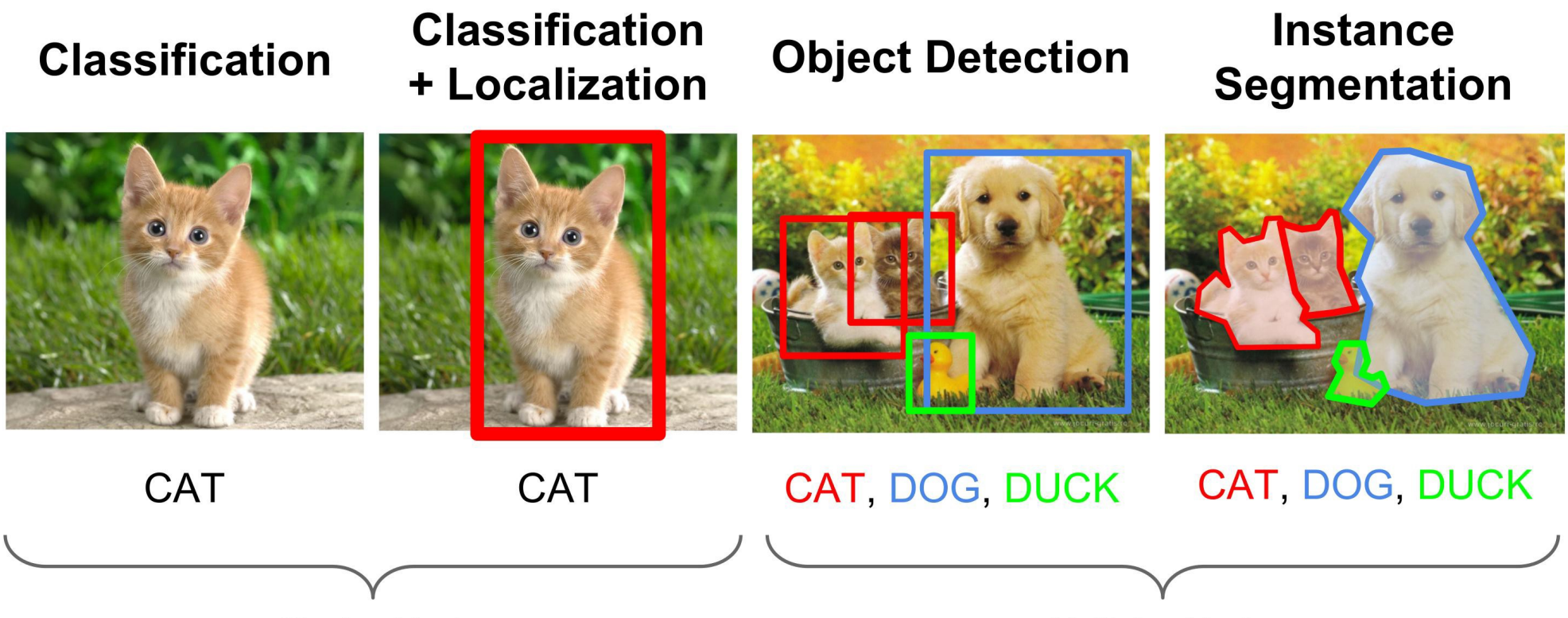
Gartner Hype Cycle for Emerging Technologies, 2019

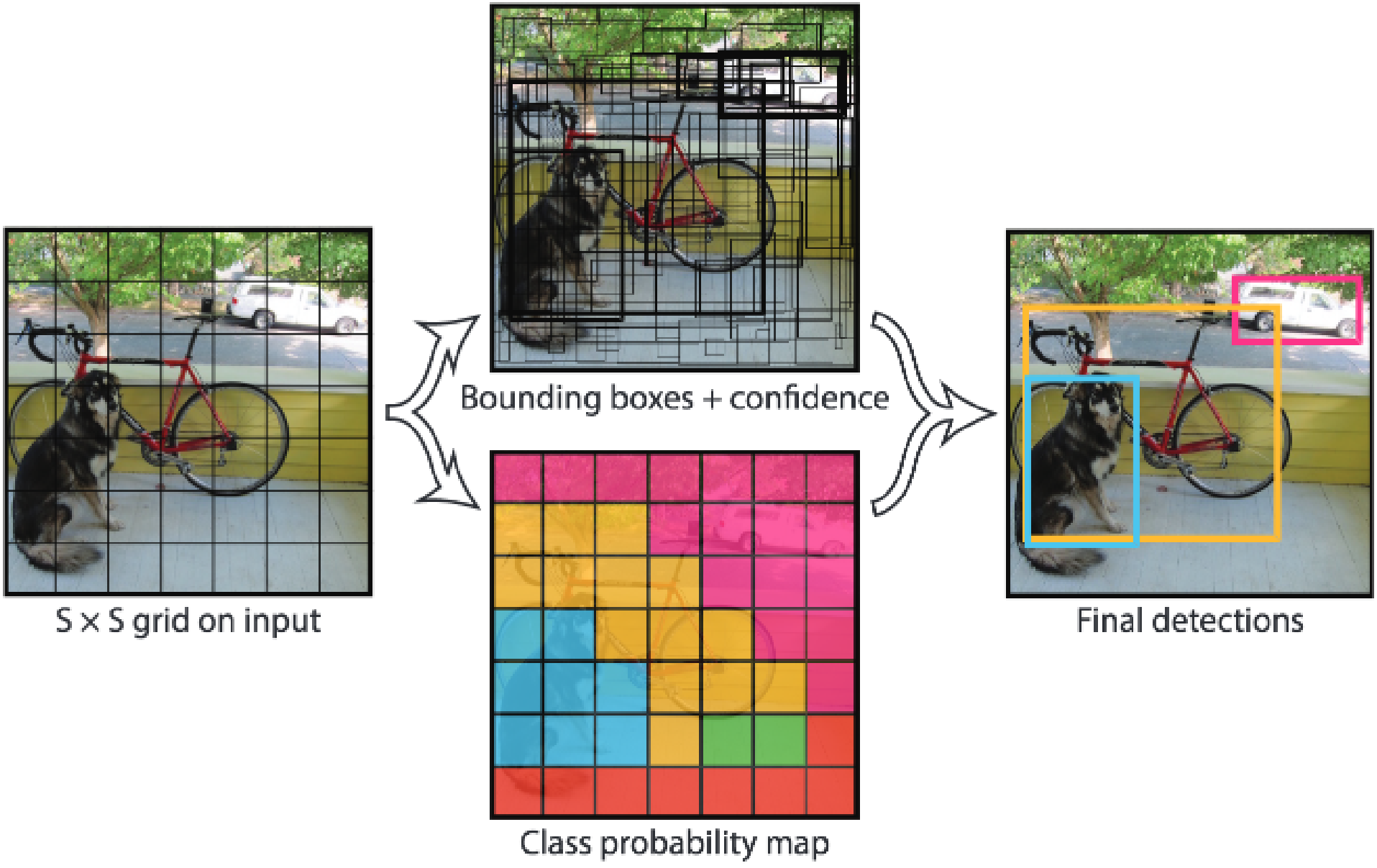


gartner.com/SmarterWithGartner

Source: Gartner

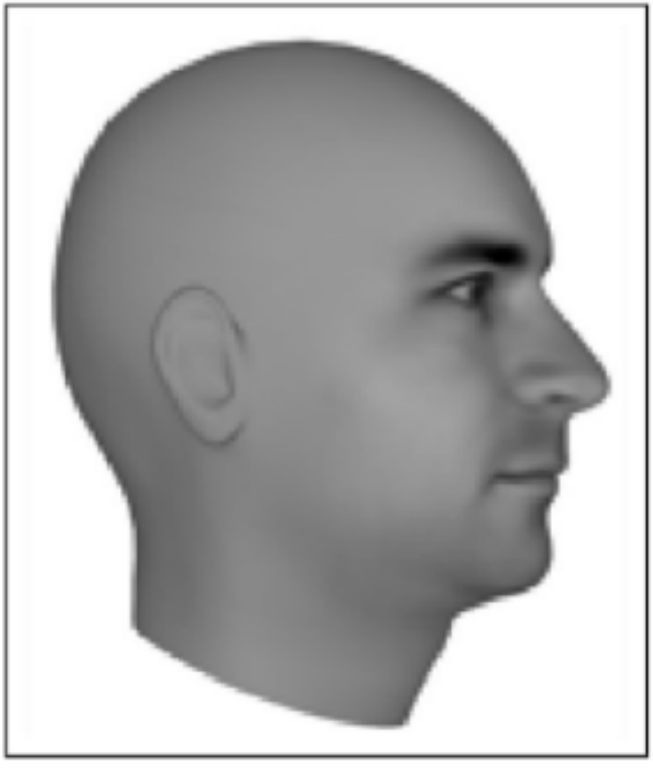
Gartner


Classification
Classification + Localization
Object Detection
Instance Segmentation
CAT
CAT
CAT, DOG, DUCK
CAT, DOG, DUCK
Single object
Multiple objects

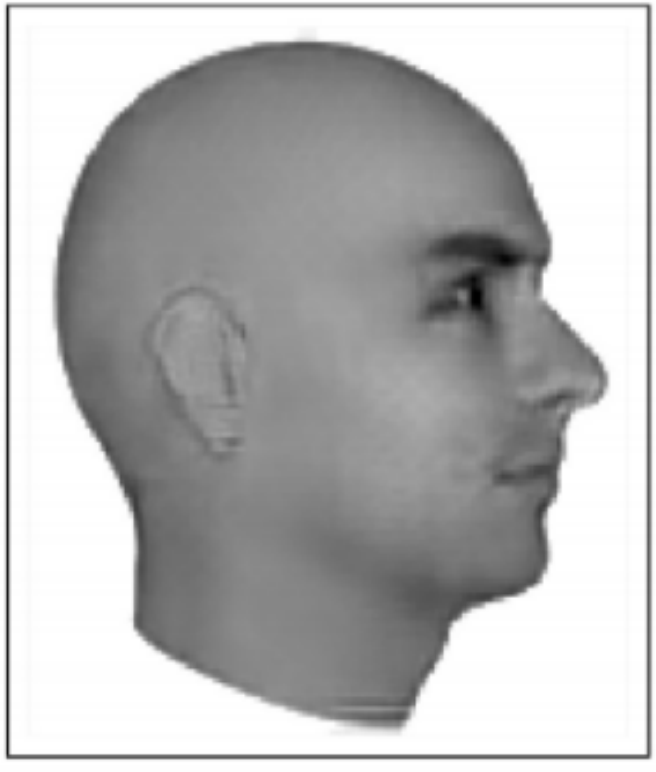

Bounding boxes + confidence
S × S grid on input
Class probability map
Final detections

Magic of GANs...

Ground Truth

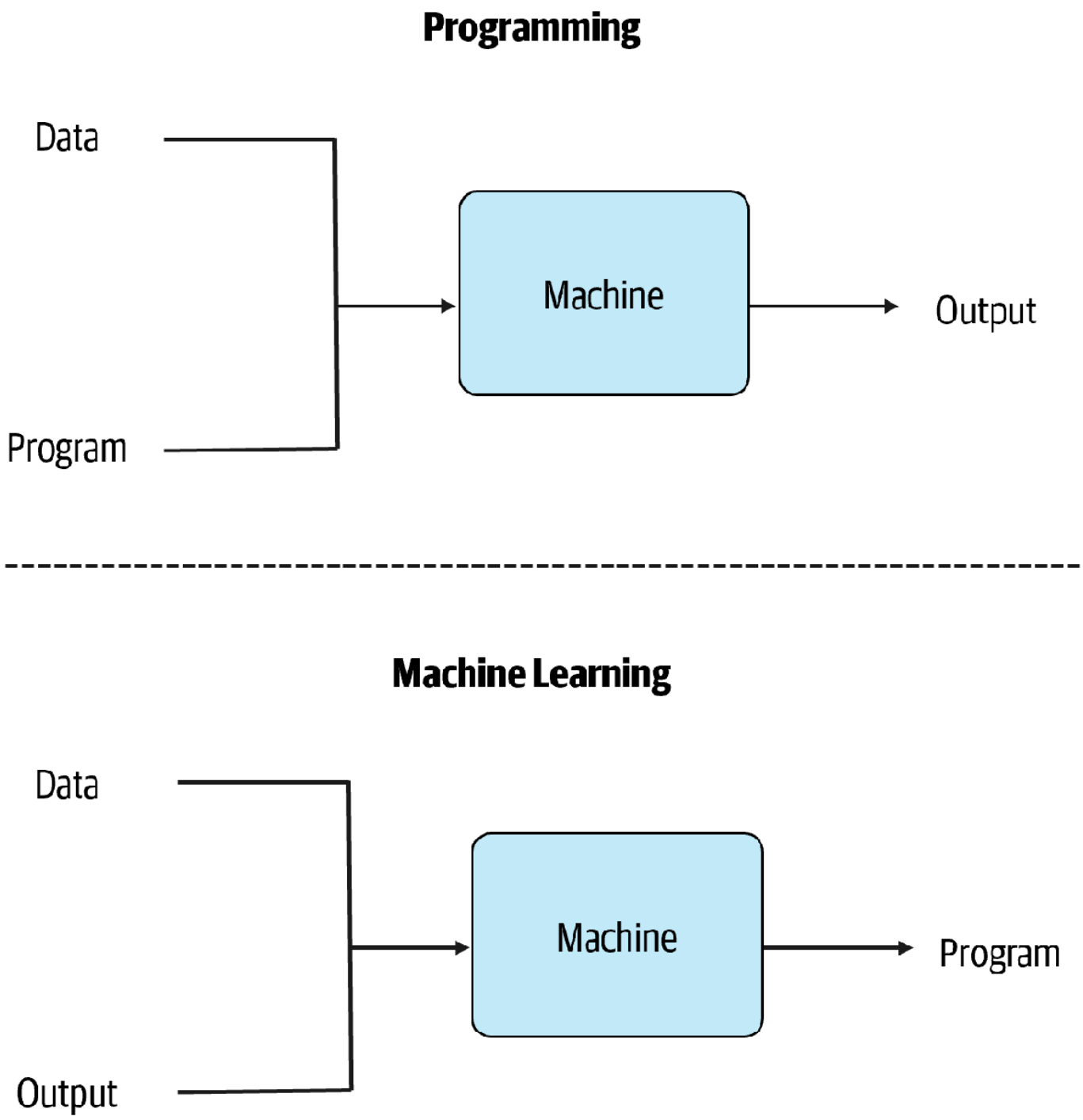


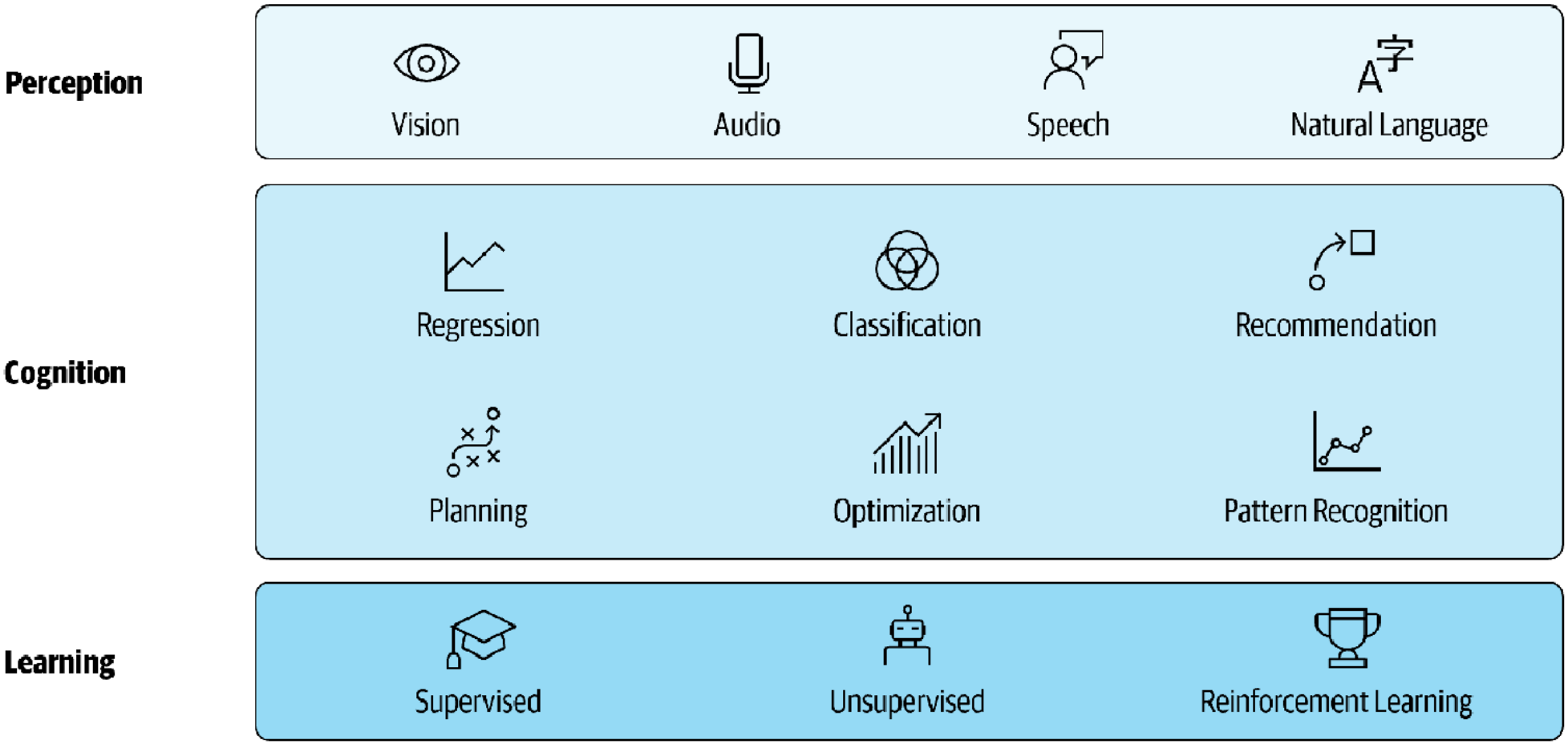
Adversarial




Imagined by a GAN (generative adversari

(a) Photo (b) Hand-drawn (c) Gatys et al. [2015] (d) Zhu et al. [2017a] (e) Ours (with ref) (f) Ours (with noise)


Programming
Data
Program
Machine
Output
Machine Learning
Data
Output
Machine
Program


Perception
Vision
Audio
Speech
Natural Language
Cognition
Regression
Classification
Recommendation
Planning
Optimization
Pattern Recognition
Learning
Supervised
Unsupervised
Reinforcement Learning

Índice

Índice

Como Iniciar o uso de Inteligência Artifical na sua Organização

1. Entender o portfolio da sua aplicação
2. Infusing AI into Your Applications
3. Creating More Engaging Applications
4. Creating New Applications with Conversational AI Interfaces

From Technology to Bussiness

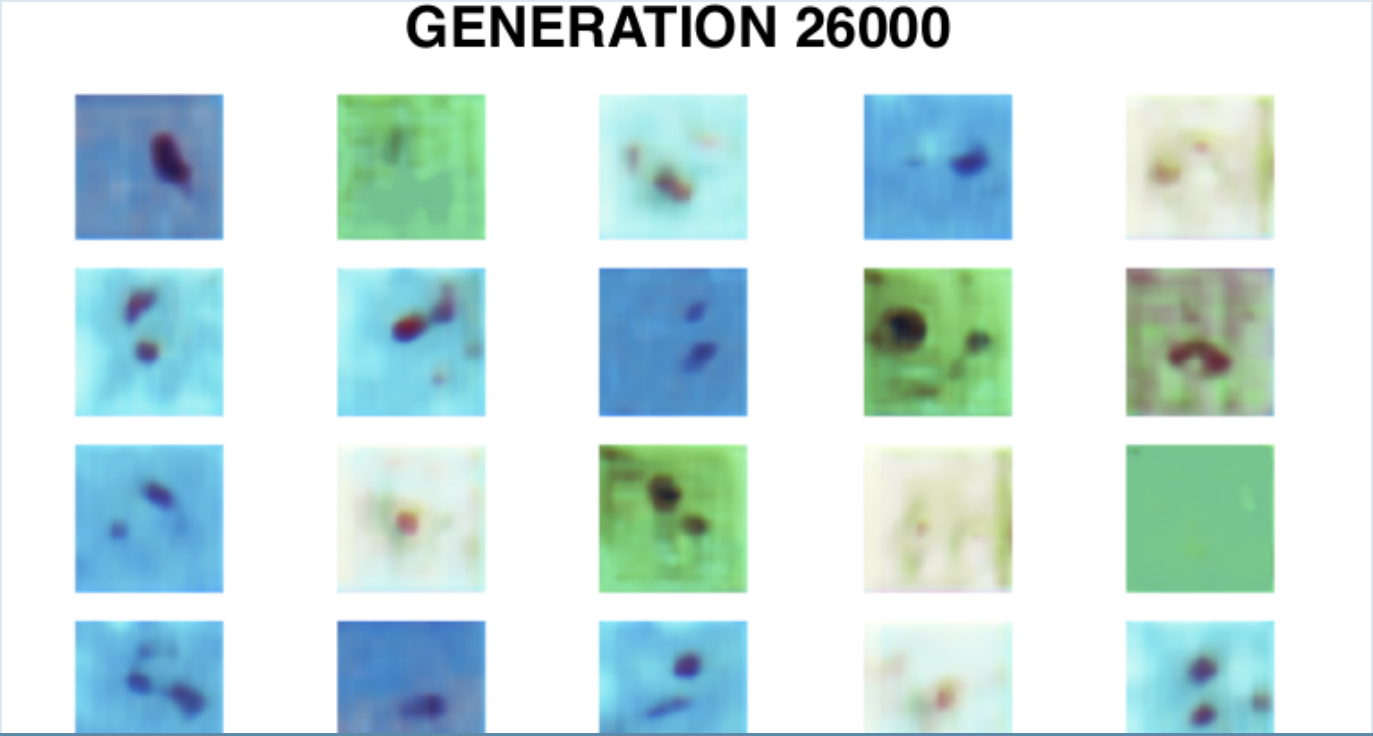
Tech to Bussiness

Start with what AI can do. Basic training for the business users on the three primary types of AI capabilities (learning, perception, and cognition)

From business to technology

Bussiness to Tech

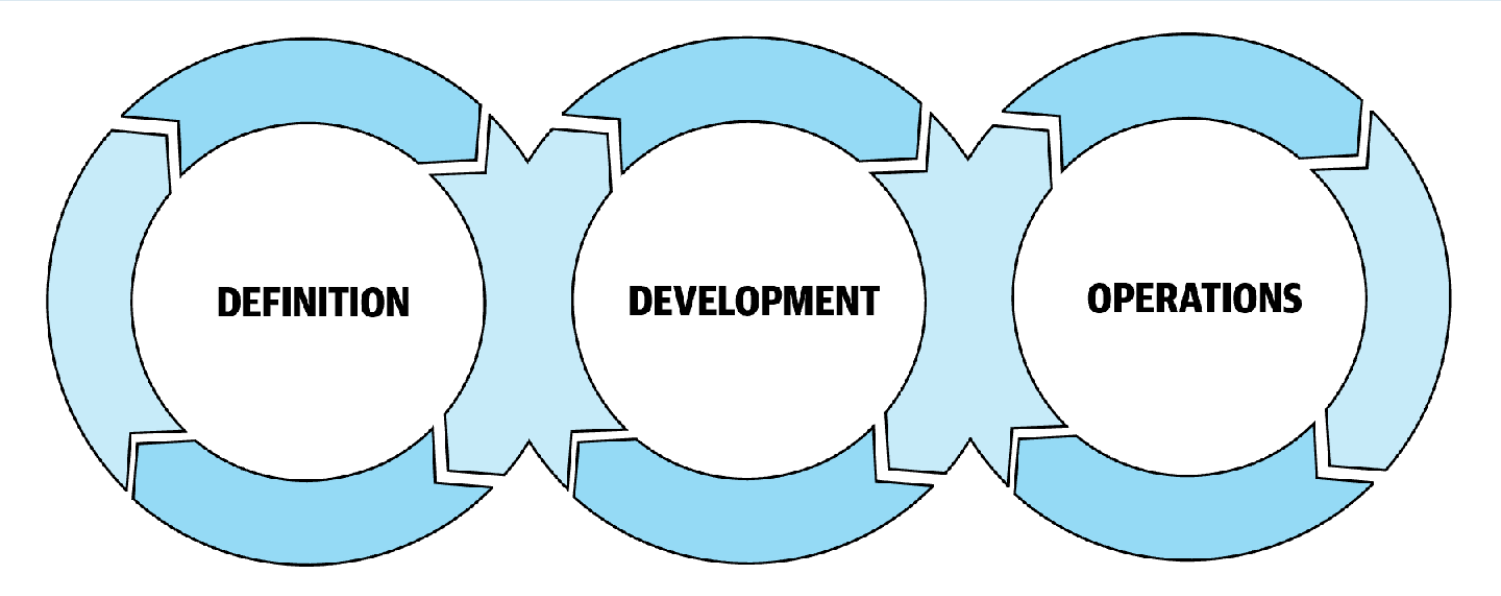
Start with business scenarios, and jointly explore how they can be improved or redefined with AI.



MLOPs

Conceito

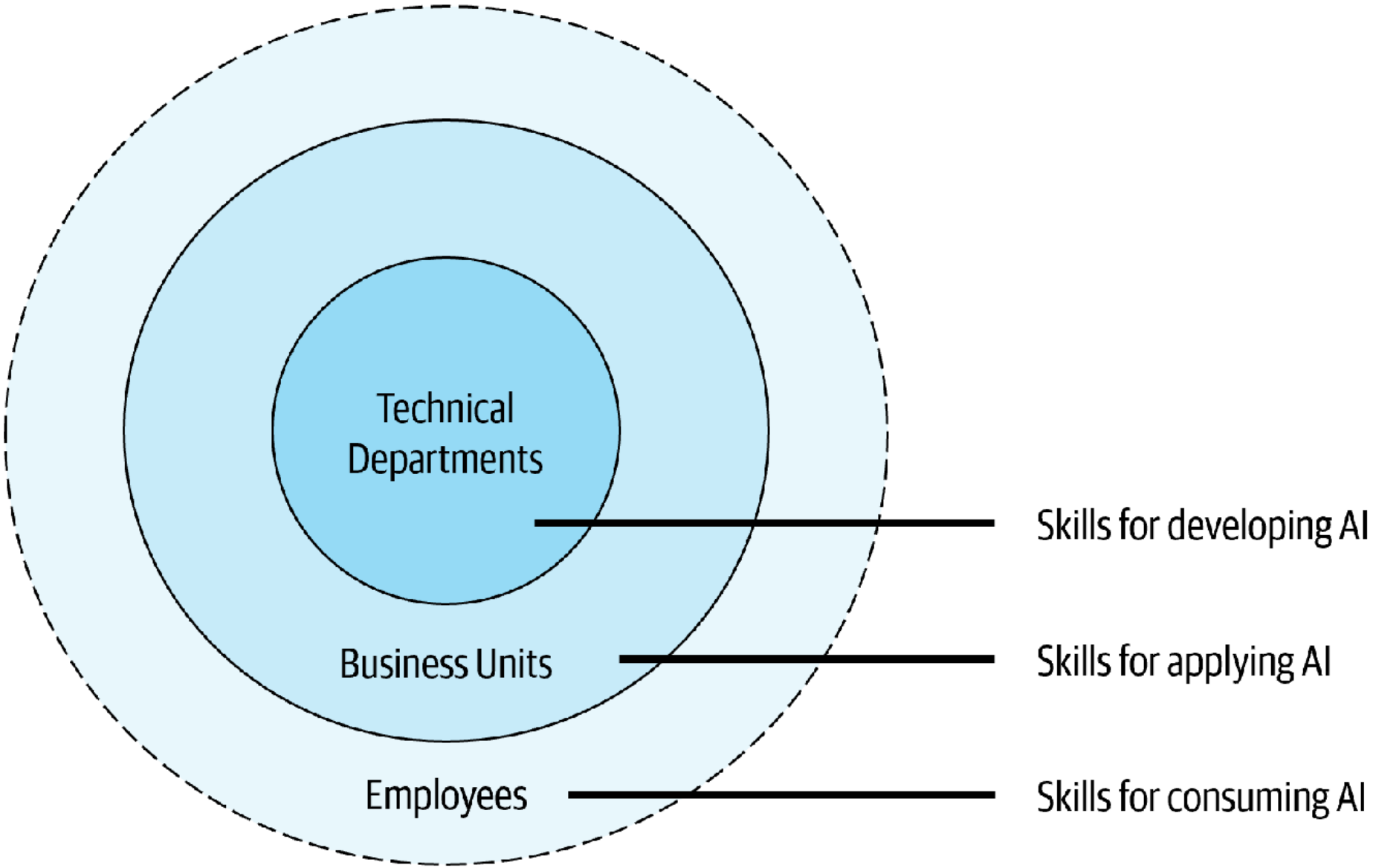
MLOps é a comunicação entre Cientistas de Dados e a equipe de operações ou produção.



Papeis

Papeis

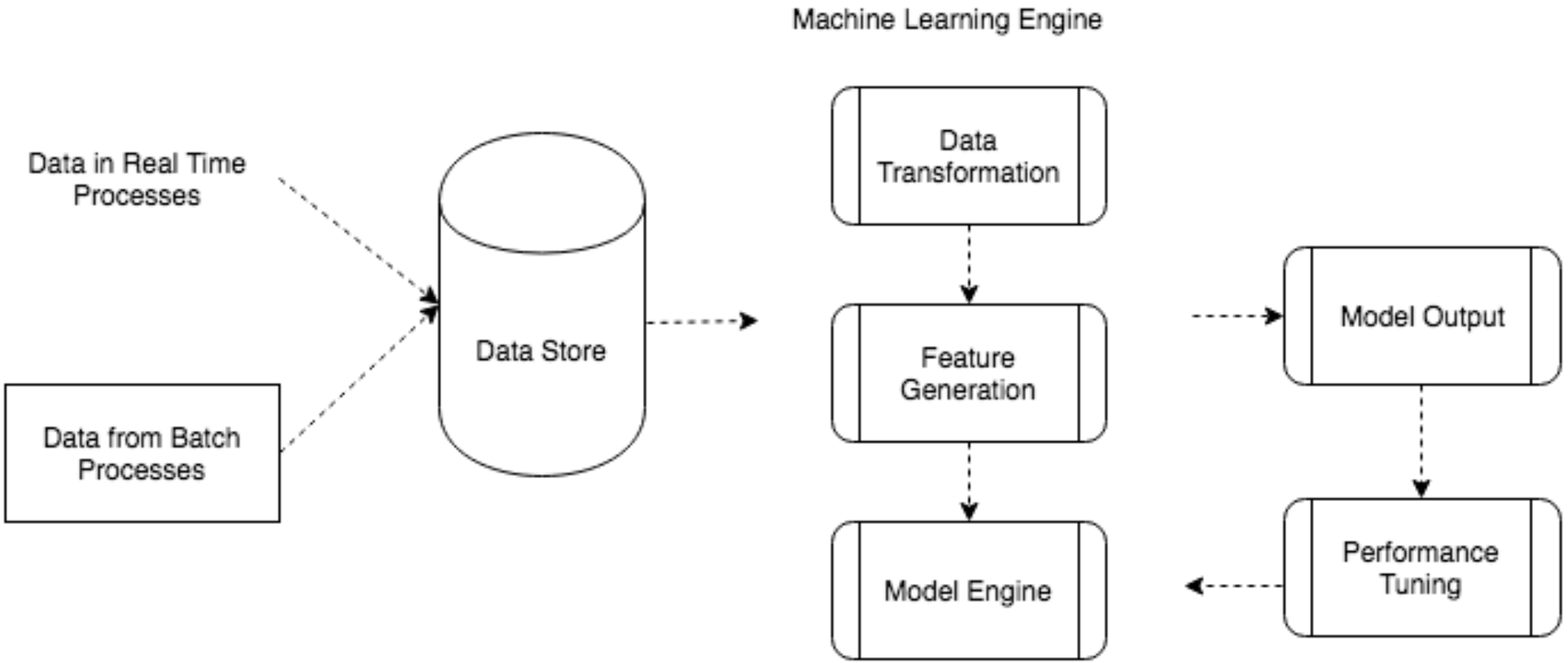
1. Pesquisador em IA
2. Arquiteto de IA
3. Desenvolvedor em IA
4. MOPs
5. Setor de Treinamento em IA


Technical Departments
Business Units
Employees
Skills for developing AI
Skills for applying AI
Skills for consuming AI

Arquitetura de Soluções de IA

Conceito

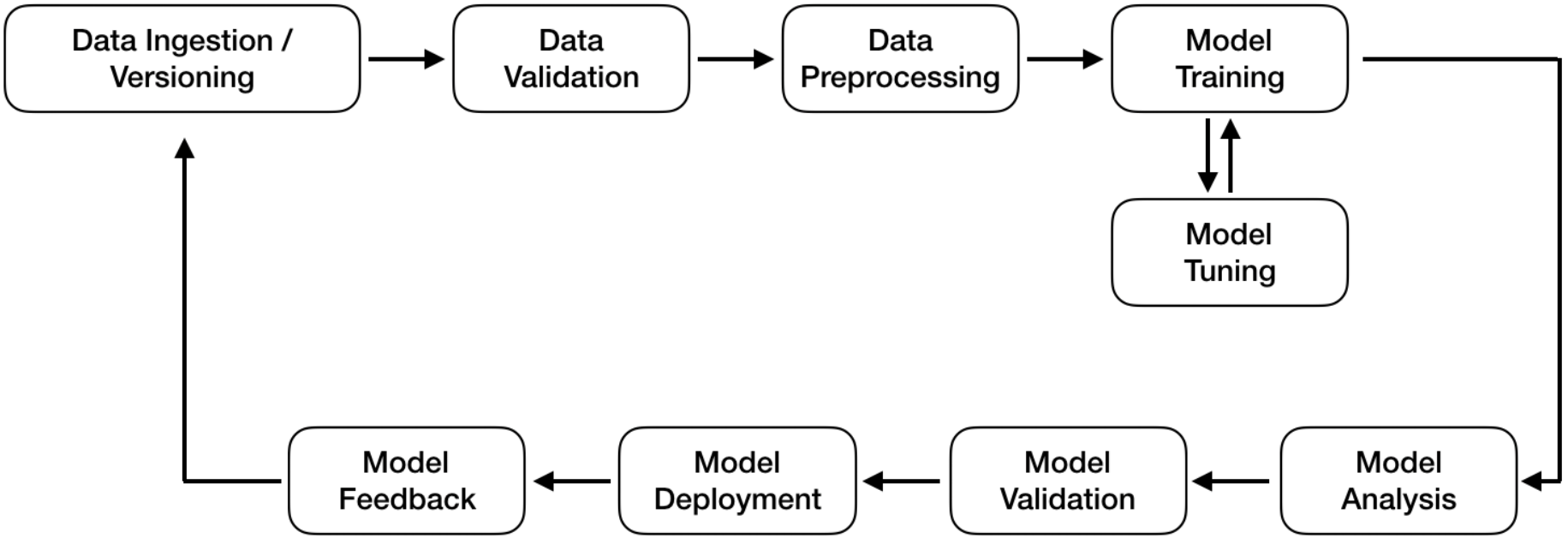
Um sistema típico de aprendizado de máquina compreende um pipeline de processos que ocorre em sequência para qualquer tipo de sistema de aprendizado de máquina, independentemente do setor.


Machine Learning Engine
Data in Real Time Processes
Data from Batch Processes
Data Store
Data Transformation
Feature Generation
Model Engine
Model Output
Performance Tuning

Pipeline

Conceito

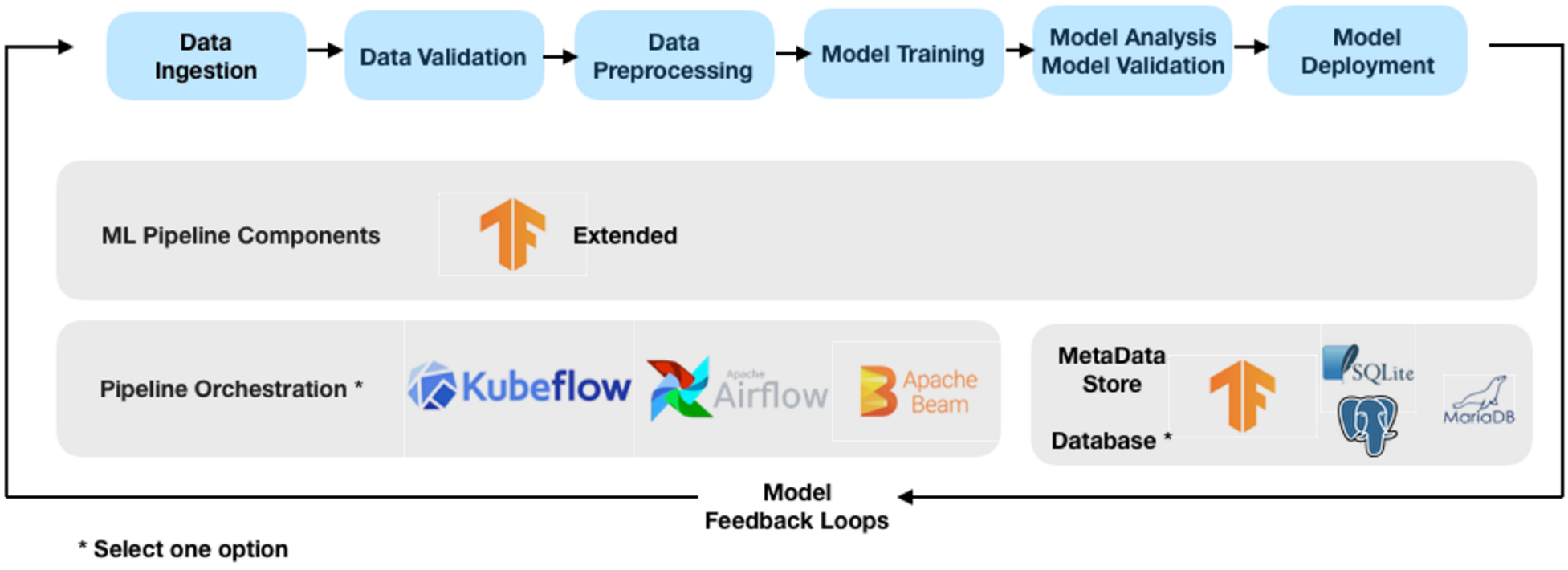
Os pipelines de aprendizado de máquina são processos para acelerar, reutilizar, gerenciar e implantar modelos de aprendizado de máquina. A engenharia de software passou pelas mesmas mudanças há uma década, com a introdução da Integração Contínua.


Data Ingestion / Versioning
Data Validation
Data Preprocessing
Model Training
Model Tuning
Model Analysis
Model Validation
Model Deployment
Model Feedback

Pipeline

Conceito

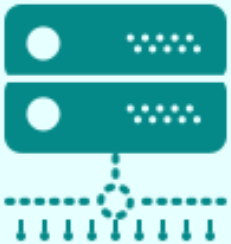
Em 2014, um grupo de engenheiros de aprendizado de máquina do Google concluiu que uma das razões pelas quais os projetos de aprendizado de máquina falham é que a maioria dos projetos vem com código personalizado para preencher a lacuna entre as etapas do pipeline de aprendizado de máquina


Data Ingestion
Data Validation
Data Preprocessing
Model Training
Model Analysis Model Validation
Model Deployment
ML Pipeline Components
Extended
Pipeline Orchestration *
Kubeflow
Apache Airflow
Apache Beam
MetaData Store
Database *
SQLite
MariaDB
Model Feedback Loops
* Select one option

DATA WAREHOUSE

VS

DATA LAKE



1110001101110

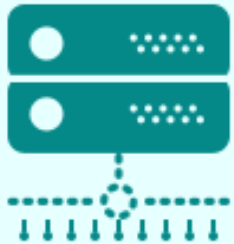
011011000110

11111000110

Data is processed and organized into a single schema before being put into the warehouse

The analysis is done on the cleansed data in the warehouse

Raw and unstructured data goes into a data lake



1110001101110

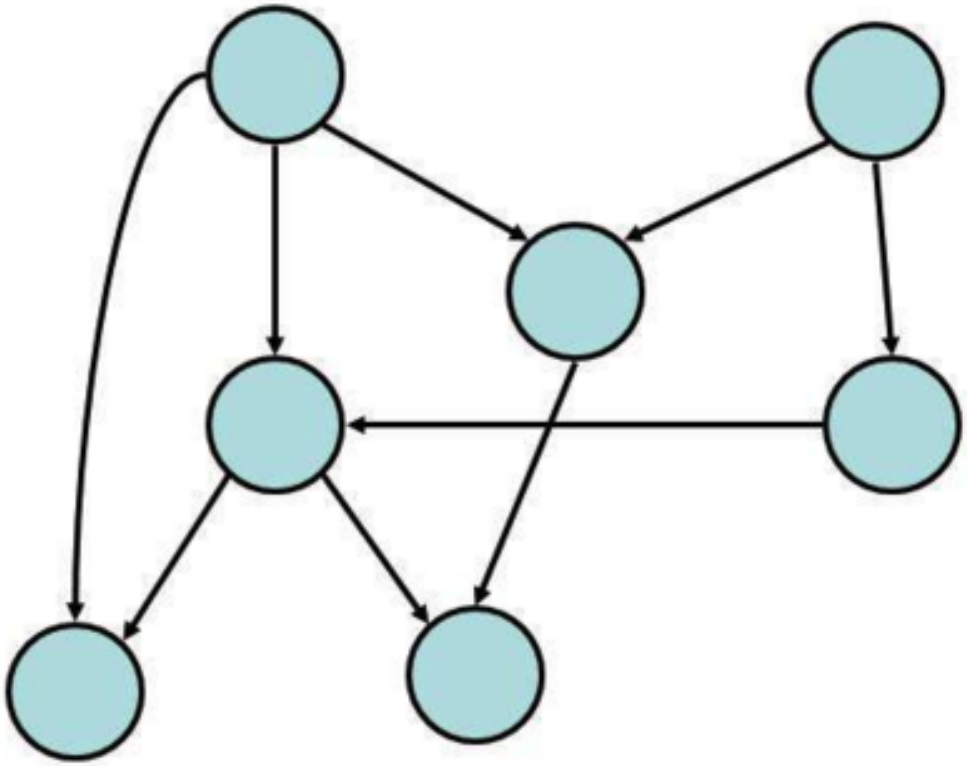
011011000110

11111000110

Data is selected and organized as and when needed

Directed Acyclic Graph

- DAG – directed graph with no directed cycles



Apache Airflow

Airflow is a platform created by community to programmatically author, schedule and monitor workflows.

Install

Airflow | DAGs | Data Profiling | Browse | Admin | Docs | About | 2019-07-28 23:43:44 UTC

DAGs

Search:

		DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
	On	tfx_caravel	None	Airflow	1 1	2019-03-04 22:52	3	
	Off	tfx_example	None	Airflow	5	2019-02-26 21:02	2	
	Off	tfx_example_solution	None	Airflow	5 1 8 6	2019-03-02 18:50	1 1	

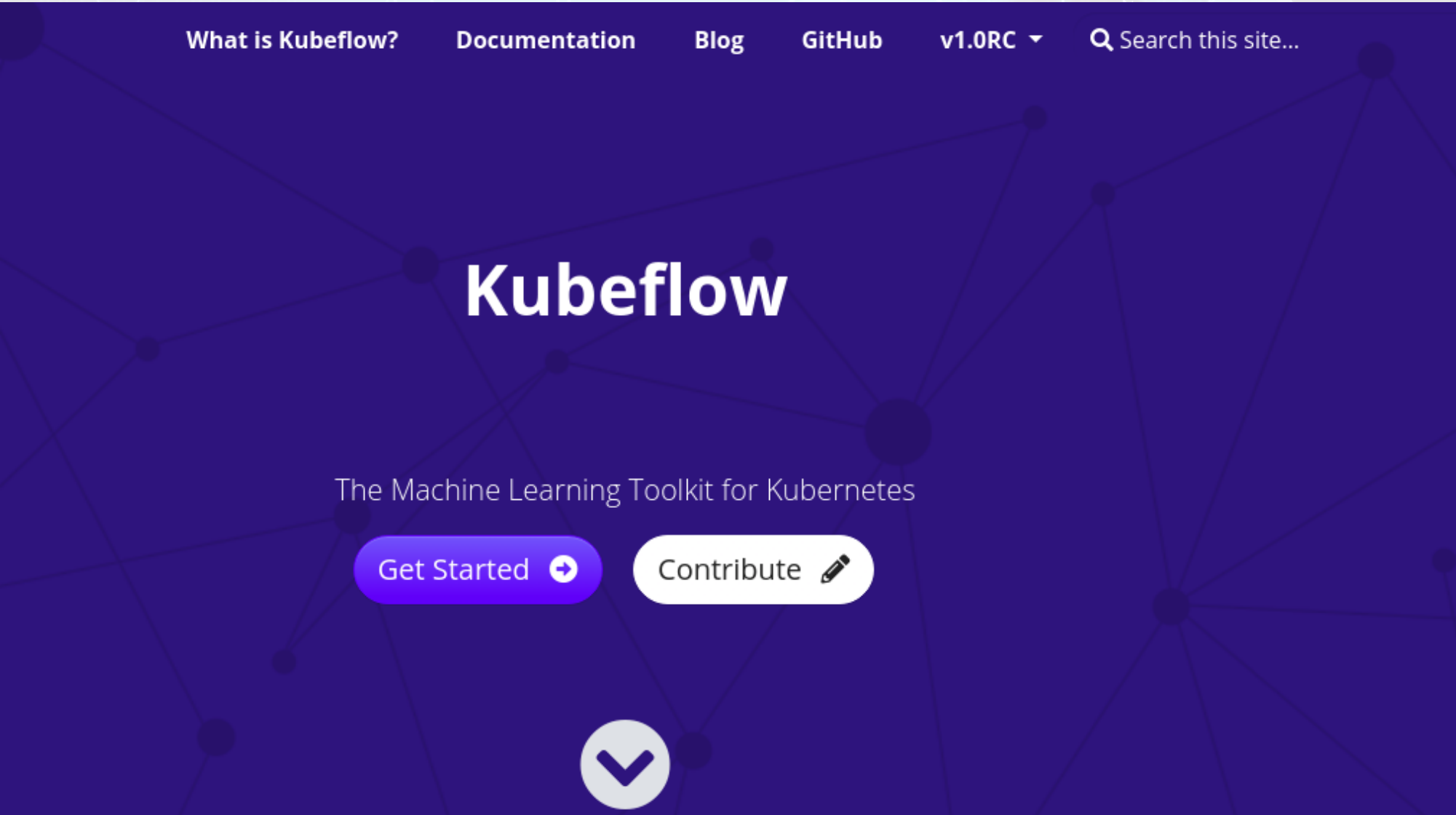
Showing 1 to 3 of 3 entries

« < 1 > »

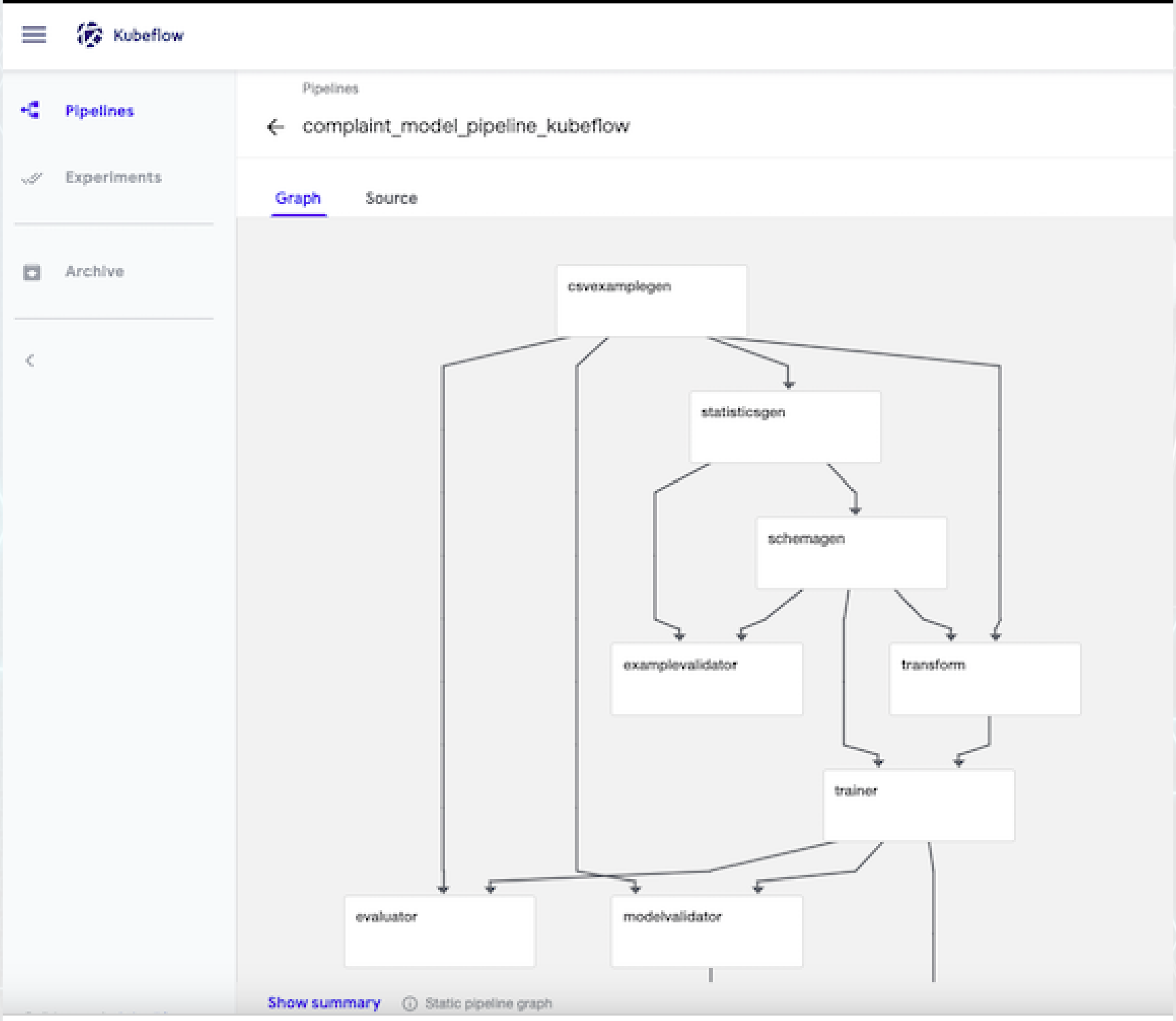
Criando DAGs no Airflow

```
dag = DAG(
    'tutorial',
    default_args=default_args,
    description='A simple tutorial DAG',
    schedule_interval=timedelta(days=1),
)

# t1, t2 and t3 are examples of tasks created by instant
t1 = BashOperator(
    task_id='print_date',
    bash_command='date',
    dag=dag,
)
```

https://airflow.apache.org/docs/stable/tutorial.html


What is Kubeflow?
Documentation
Blog
GitHub
v1.0RC
Search this site...
Kubeflow
The Machine Learning Toolkit for Kubernetes
Get Started
Contribute

```python
@dsl.pipeline(
    name='XGBoost Trainer',
    description='A trainer that does end-to-end distributed training for XGE
)
def xgb_train_pipeline(
    output='gs://your-gcs-bucket',
    project='your-gcp-project',
    cluster_name='xgb-%s' % dsl.RUN_ID_PLACEHOLDER,
    region='us-central1',
    train_data='gs://ml-pipeline-playground/sfpd/train.csv',
    eval_data='gs://ml-pipeline-playground/sfpd/eval.csv',
    schema='gs://ml-pipeline-playground/sfpd/schema.json',
    target='resolution',
    rounds=200,
    workers=2,
    true_label='ACTION',
):
    output_template = str(output) + '/' + dsl.RUN_ID_PLACEHOLDER + '/data'

    # Current GCP pyspark/spark op do not provide outputs as return values,
    # we need to use strings to pass the uri around.
    analyze_output = output_template
    transform_output_train = os.path.join(output_template, 'train', 'part-*'
```


Kubeflow
Pipelines
Experiments
Archive
Pipelines
complaint_model_pipeline_kubeflow
Graph
Source
csvexamplegen
statisticsgen
schemagen
examplevalidator
transform
trainer
evaluator
modelvalidator
Show summary
Static pipeline graph

Apache Beam: An advanced unified programming model

Implement batch and streaming data processing jobs that run on any execution engine.

LEARN MORE | TRY BEAM | DOWNLOAD BEAM SDK 2.19.0

JAVA QUICKSTART | PYTHON QUICKSTART | GO QUICKSTART

The latest from the blog

Apache Beam 2.19.0

FEB 4, 2020

Apache Beam 2.18.0

JAN 23, 2020

Apache Beam 2.17.0

JAN 6, 2020

Pipeline criado no ApacheBeam

```
PCollection<KV<String, String>> userAddress =
pipeline.apply(JdbcIO.<KV<String, String>>read()...);
PCollection<KV<String, String>> userOrder =
pipeline.apply(KafkaIO.<String, String>read()...);
final TupleTag<String> addressTag =
new TupleTag<String>();
final TupleTag<String> orderTag =
new TupleTag<String>();
// Merge collection values into a CoGbkResult collection
PCollection<KV<String, CoGbkResult>> joinedCollection
....
joinedCollection.apply(...);
```

https://beam.apache.org/documentation/pipelines/design-your-pipeline/

Validação dos dados

Conceito

No aprendizado de máquina, estamos tentando aprender com os padrões nos conjuntos de dados e generalizar esses aprendizados. Isso coloca nossos conjuntos de dados na frente e no centro de nossos fluxos de trabalho de aprendizado de máquina, e a qualidade dos dados se torna fundamental para o sucesso de nossos projetos de aprendizado de máquina.

In [5]:
```
tfdv.visualize_statistics(train_stats)
```

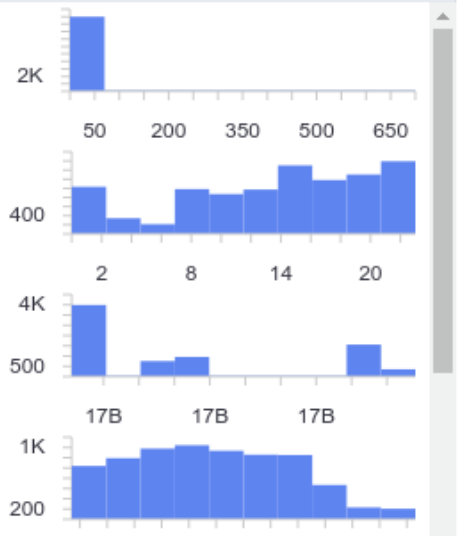
Sort by: Feature order ☐ Reverse order Feature search

Features: ☑ int(5) ☑ float(10) ☑ string(2)

Numeric Features (15)

Chart to show: Standard ☐ log ☐ expand

	count	missing	mean	std dev	zeros	min	median	max
fare	10,000	0%	11.74	12.13	0.17%	0	0	700.07
trip_start_hour	10,000	0%	13.63	6.61	4.14%	0	0	23
dropoff_census_tract	7,148	27.12%	17.0B	331k	0%	17.0B	0	17.0B
trip_start_timestamp	10,000	0%	1.41B	29.2M	0%	1.36B	0	1.48B



Gerando estatísticas dos Dados

```
train_stats = tfdv.generate_statistics_from_tfrecord(
data_location=train_tfrecords_filename)

val_stats = tfdv.generate_statistics_from_tfrecord(
data_location=val_tfrecords_filename)
tfdv.visualize_statistics(lhs_statistics=val_stats,
rhs_statistics=train_stats,lhs_name='VAL_DATASET',
rhs_name='TRAIN_DATASET')
```

https://www.tensorflow.org/tfx/data_validation/get_started

Sort by

Feature order

Reverse order

Feature search (regex enabled)

Features: int(2) float(1) string(11)

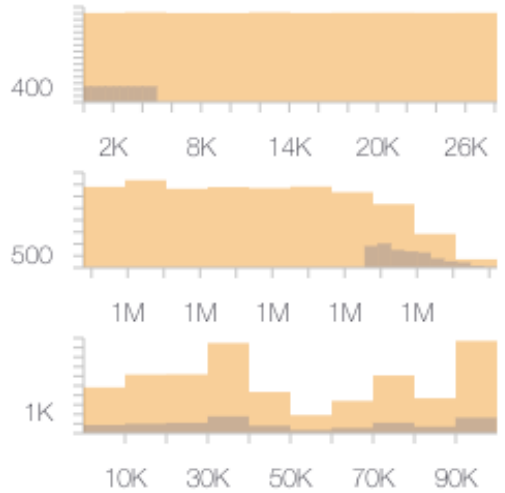
VAL_DATASET TRAIN_DATASET

Numeric Features (3)

Chart to show: Standard

log expand percentages

	count	missing	mean	std dev	zeros	min	median	max
col_0								
	4,998	0%	2,498.5	1,442.8	0.02%	0	2,498	4,997
	28.2k	0%	14.1k	8,127.94	0%	0	14.1k	28.2k
complaint_id								
	4,998	0%	1.27M	7,950.37	0%	1.26M	1.27M	1.29M
	28.2k	0%	1.23M	28.4k	0%	1.18M	1.22M	1.29M
zip_code								
	4,960	0.76%	49.8k	30.4k	0%	20	44.1k	99.9k
	27.9k	0.99%	49.9k	30.6k	0%	9	44.1k	100k



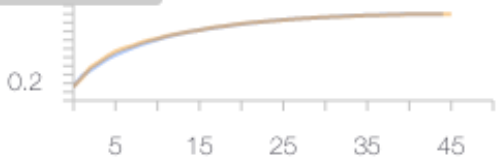
Categorical Features (11)

Chart to show: Standard

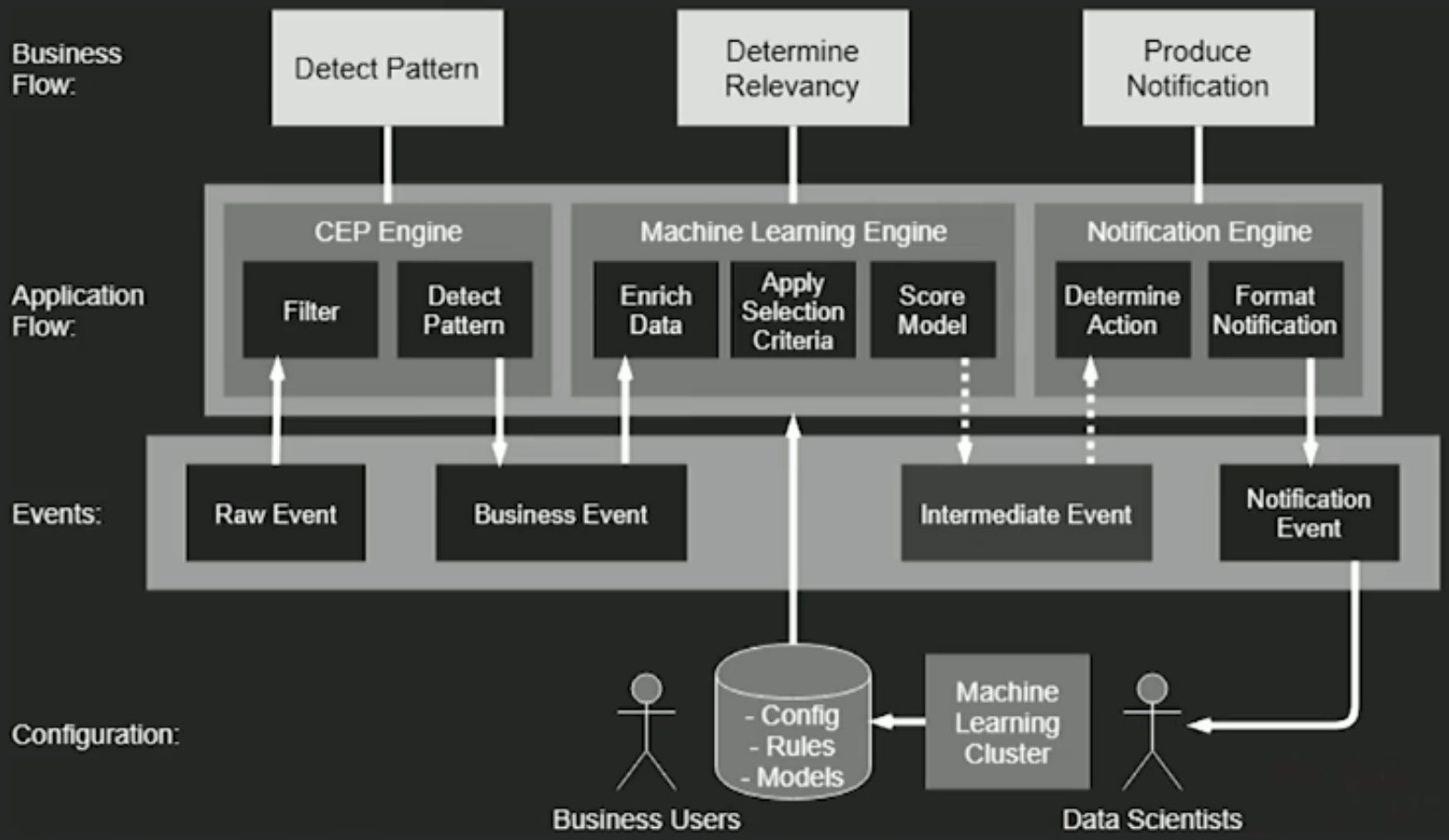
log expand percentages

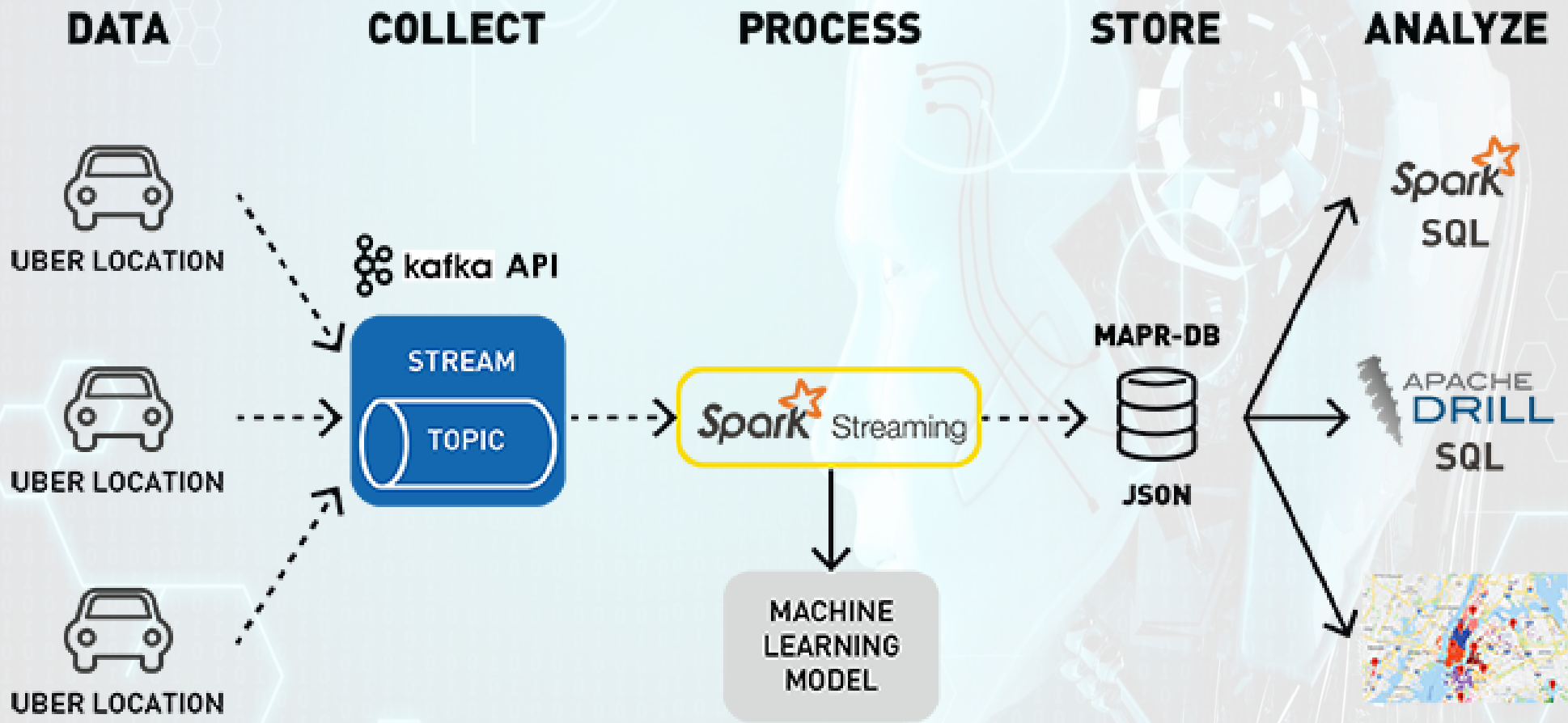
	count	missing	unique	top	freq top	avg str len
sub_issue						
	2,524	49.5%	45	Debt is n…	410	25.05
	13.2k	53.08%	46	Debt is n…	2,125	24.42

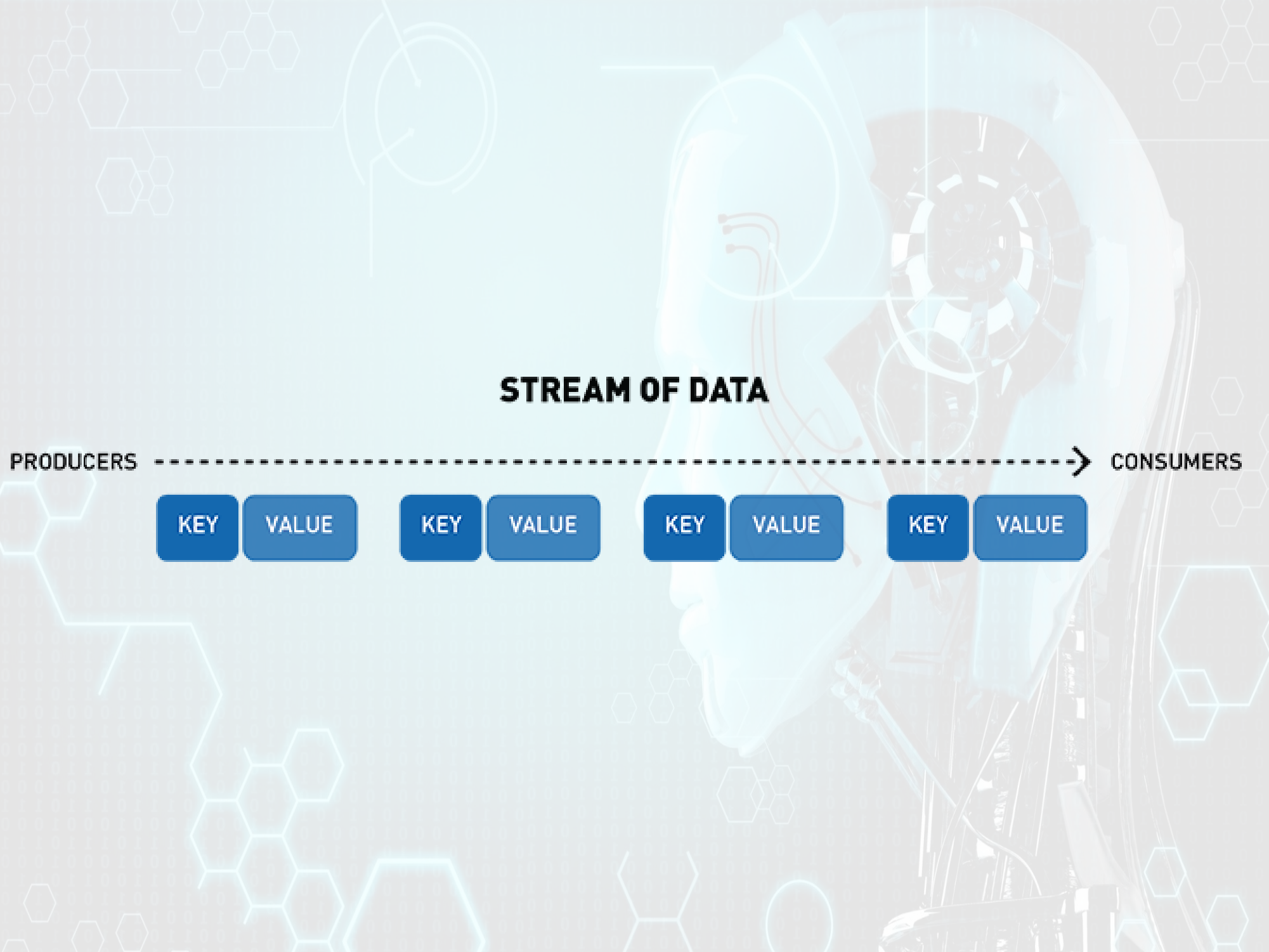
SHOW RAW DATA

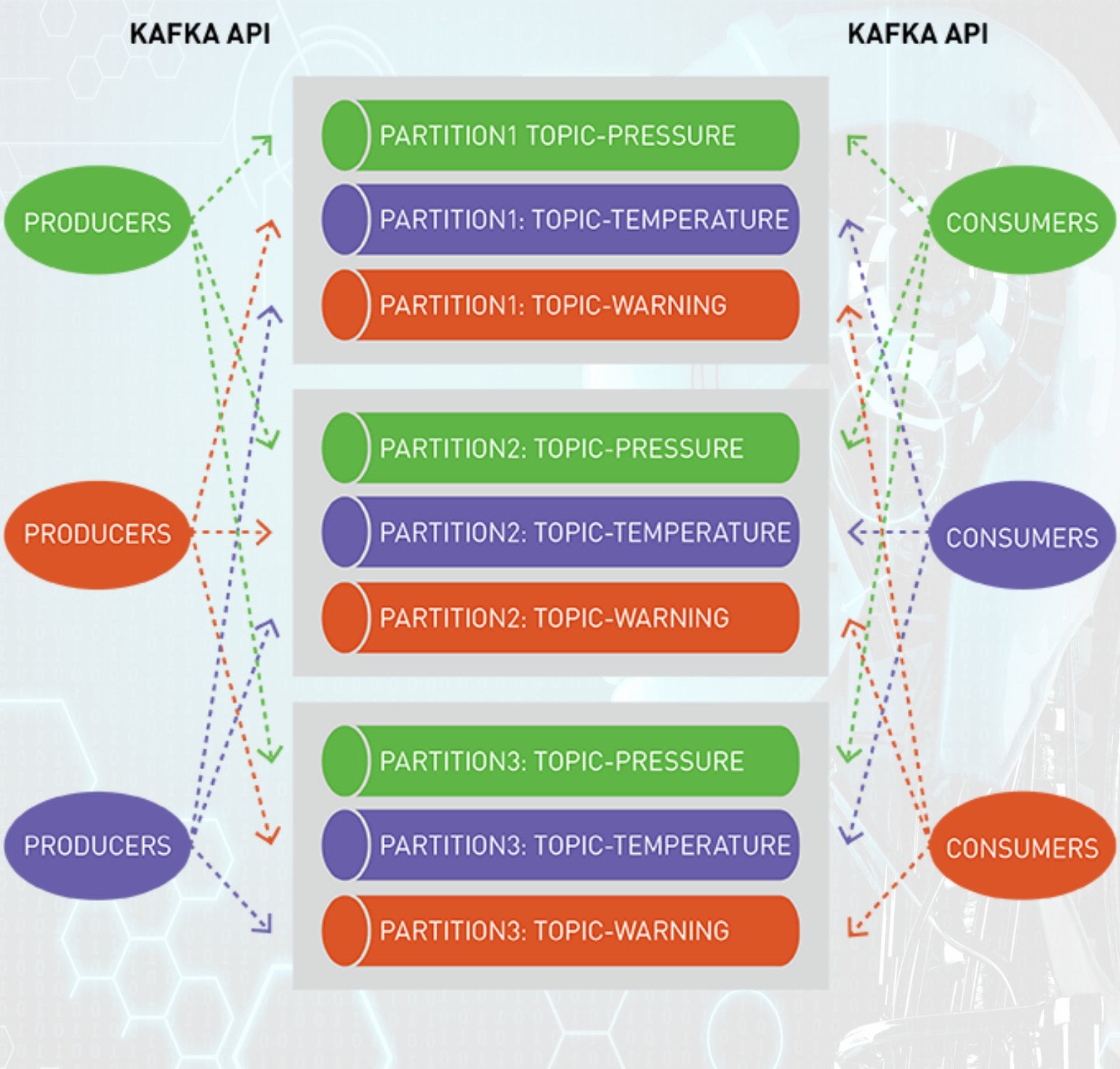


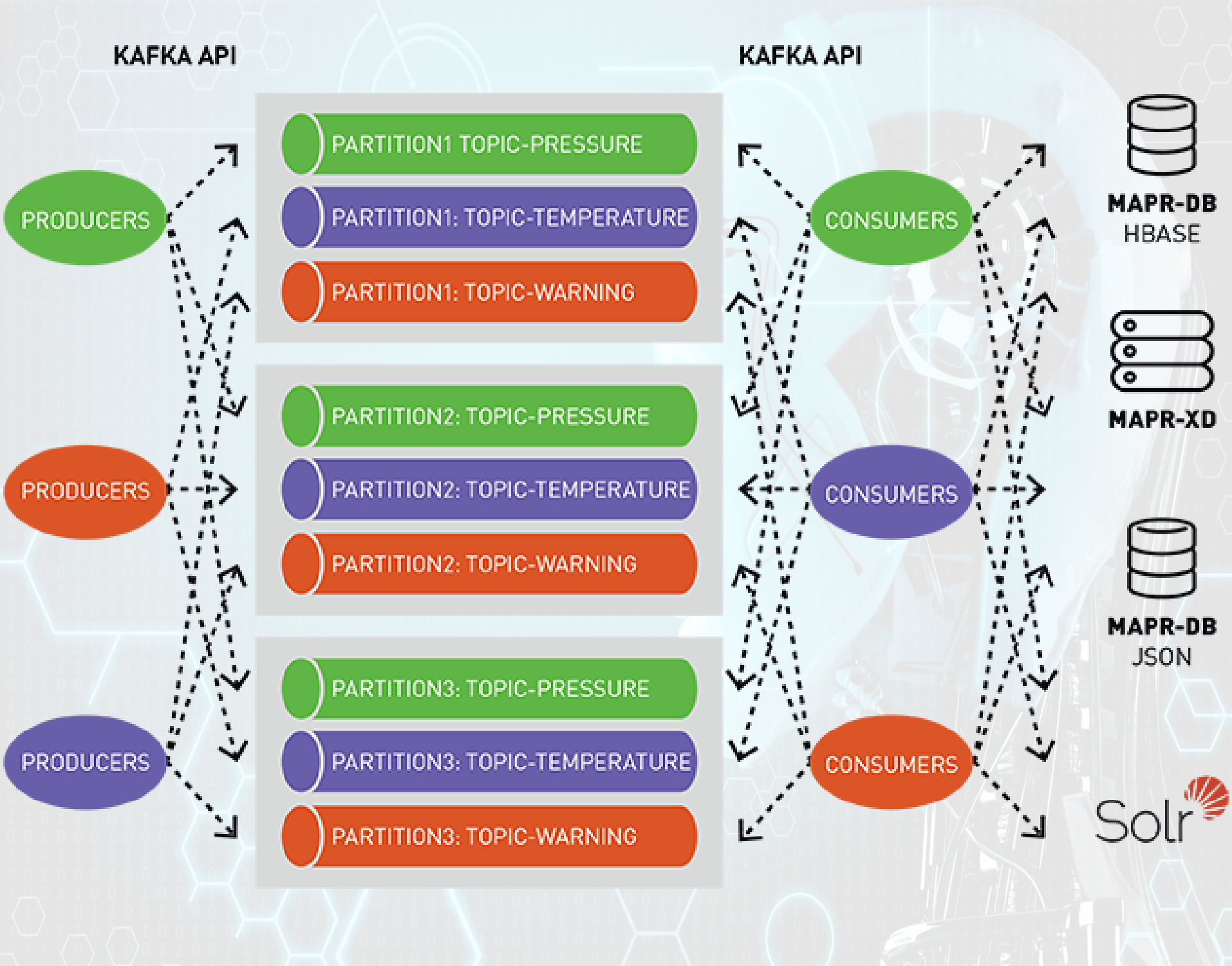
ARCHITECTURE

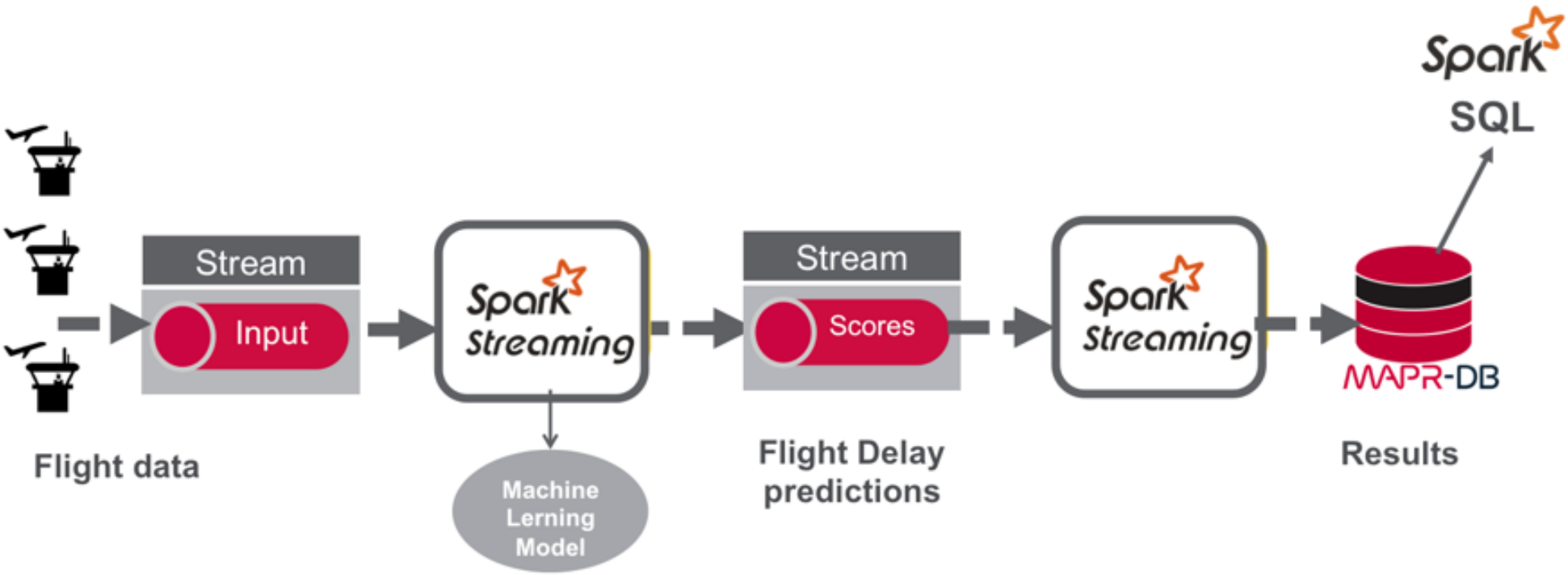



DATA
COLLECT
PROCESS
STORE
ANALYZE
UBER LOCATION
UBER LOCATION
UBER LOCATION
kafka API
STREAM
TOPIC
Spark Streaming
MACHINE LEARNING MODEL
MAPR-DB
JSON
Spark SQL
APACHE DRILL SQL


STREAM OF DATA
PRODUCERS
CONSUMERS
KEY
VALUE
KEY
VALUE
KEY
VALUE
KEY
VALUE


KAFKA API
KAFKA API
PRODUCERS
PRODUCERS
PRODUCERS
PARTITION1 TOPIC-PRESSURE
PARTITION1: TOPIC-TEMPERATURE
PARTITION1: TOPIC-WARNING
PARTITION2: TOPIC-PRESSURE
PARTITION2: TOPIC-TEMPERATURE
PARTITION2: TOPIC-WARNING
PARTITION3: TOPIC-PRESSURE
PARTITION3: TOPIC-TEMPERATURE
PARTITION3: TOPIC-WARNING
CONSUMERS
CONSUMERS
CONSUMERS


KAFKA API
KAFKA API
PRODUCERS
PRODUCERS
PRODUCERS
PARTITION1 TOPIC-PRESSURE
PARTITION1: TOPIC-TEMPERATURE
PARTITION1: TOPIC-WARNING
PARTITION2: TOPIC-PRESSURE
PARTITION2: TOPIC-TEMPERATURE
PARTITION2: TOPIC-WARNING
PARTITION3: TOPIC-PRESSURE
PARTITION3: TOPIC-TEMPERATURE
PARTITION3: TOPIC-WARNING
CONSUMERS
CONSUMERS
CONSUMERS
MAPR-DB
HBASE
MAPR-XD
MAPR-DB
JSON
Solr


Spark
SQL
Stream
Input
Spark
Streaming
Stream
Scores
Spark
Streaming
MAPR-DB
Flight data
Machine
Lerning
Model
Flight Delay
predictions
Results

Flight input data

Enrich with Prediction

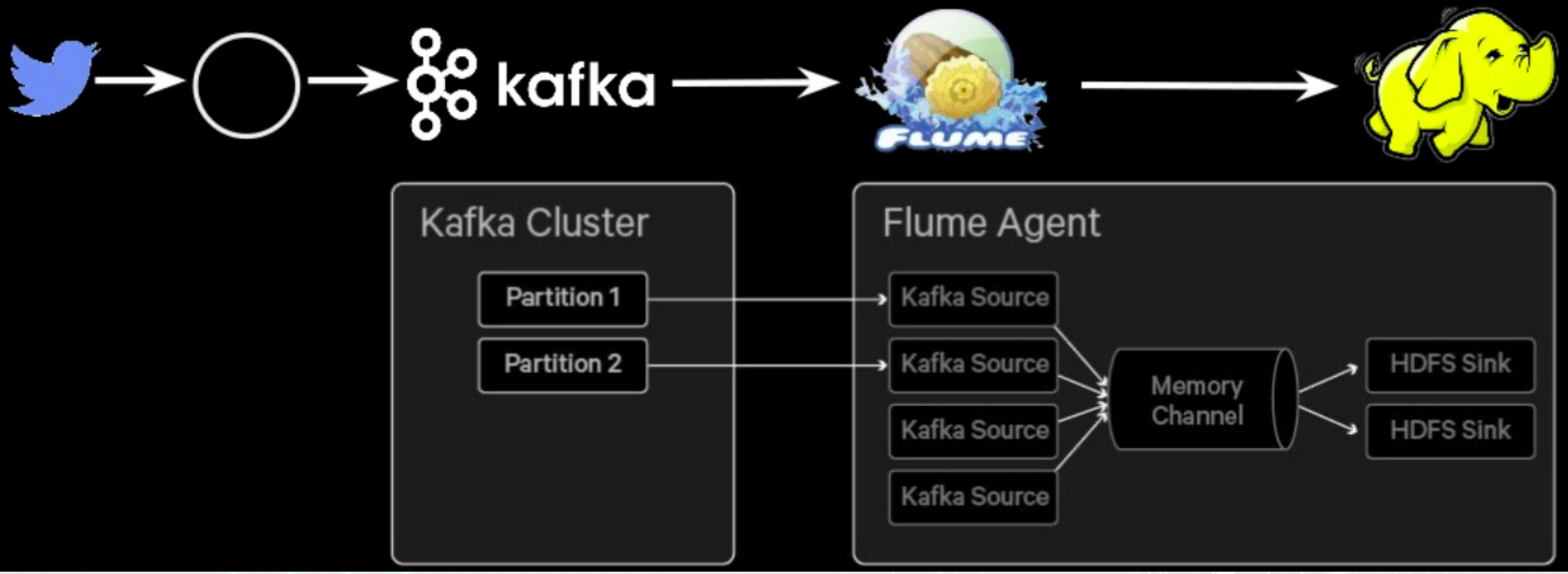
Flight delay predictions

{"_id":"AA_2017-02-16_EWR_ORD_1124",
"dofW":4, "carrier":"AA", "origin":"EWR",
"dest":"ORD", "crsdeptime":705,
"crsarrtime":851, "crselapsedtime":166.0,
"dist":719.0}

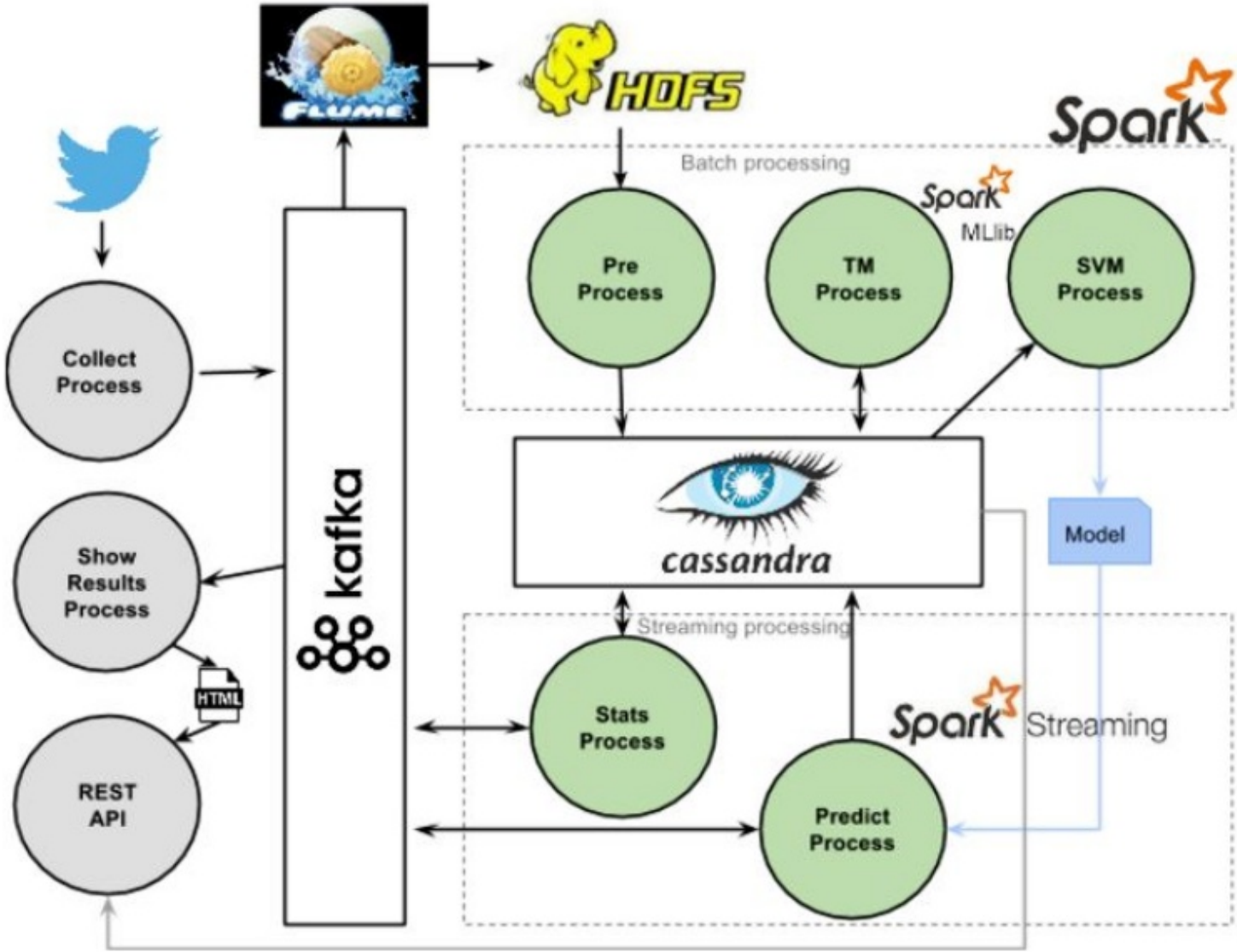
{"_id":"AA_2017-02-16_EWR_ORD_1124",
"dofW":4, "carrier":"AA", "origin":"EWR",
"dest":"ORD", "crsdeptime":705,
"crsarrtime":851, "crselapsedtime":166.0,
"dist":719.0, "pred_dtree":0}

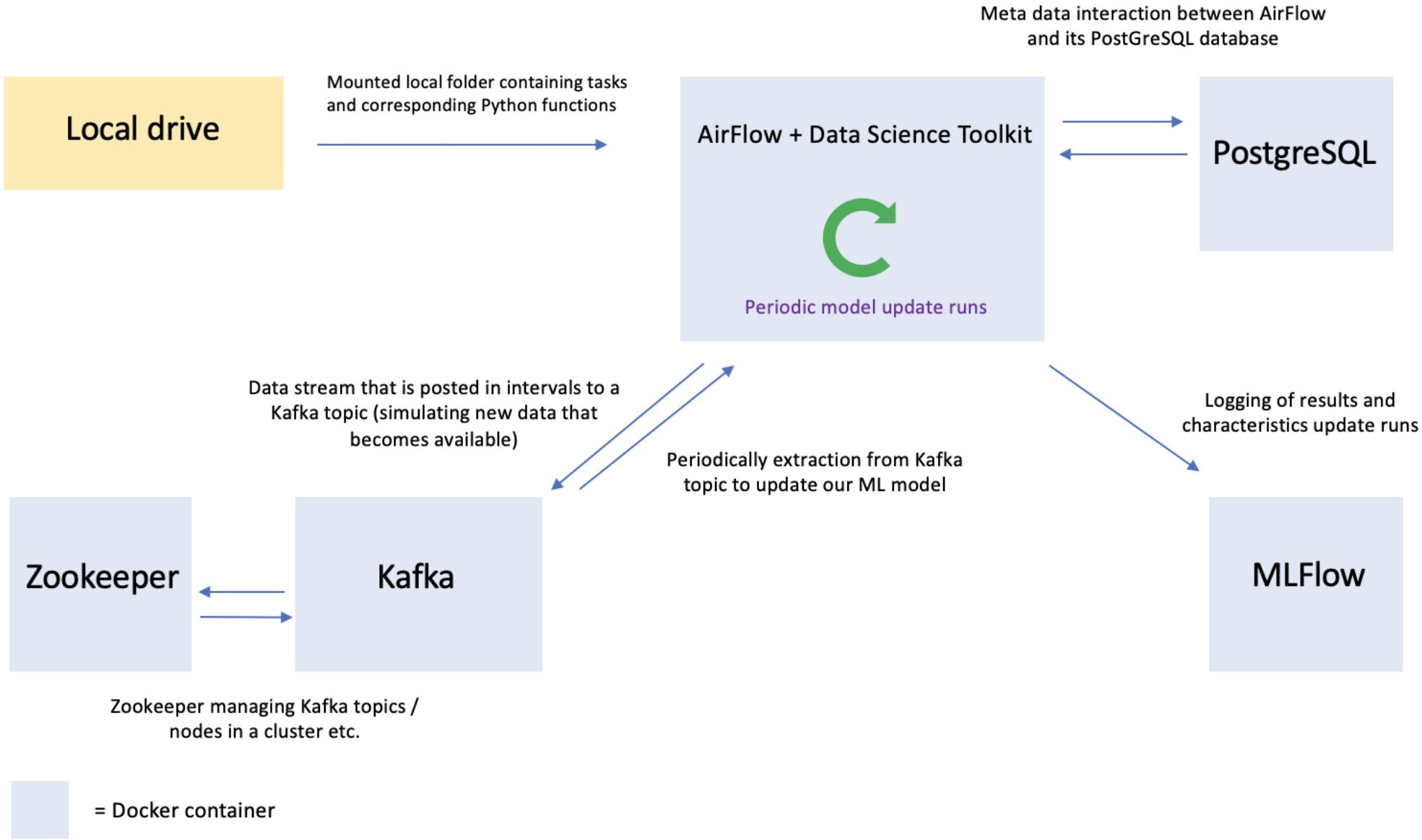
Clip slide

Data Collection: Apache Flume



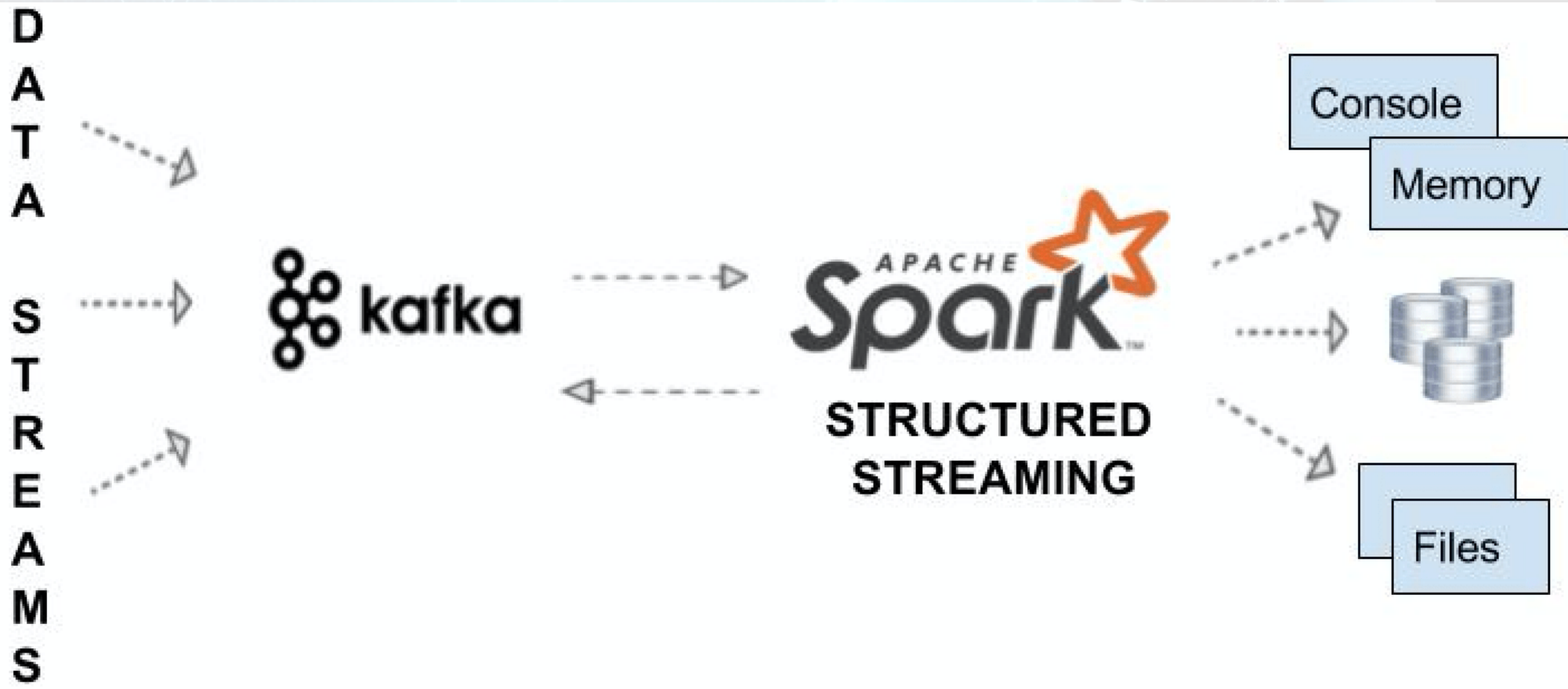
Processing Analytics Layer




Meta data interaction between AirFlow and its PostGreSQL database
Local drive
Mounted local folder containing tasks and corresponding Python functions
AirFlow + Data Science Toolkit
PostgreSQL
Periodic model update runs
Data stream that is posted in intervals to a Kafka topic (simulating new data that becomes available)
Periodically extraction from Kafka topic to update our ML model
Logging of results and characteristics update runs
Zookeeper
Kafka
MLFlow
Zookeeper managing Kafka topics / nodes in a cluster etc.
= Docker container

THE KFC STACK

- Data stream storage: **Kafka**
- Stream processing: **Flink**
- Persisting rules, models, and config: **Cassandra**
- Model scoring: **PMML** and **Openscoring.io**


DATA STREAMS
kafka
APACHE
Spark
STRUCTURED
STREAMING
Console
Memory
Files

MLflow Tracking

Exemplo de uso do MLFlow

```
mlflow.log_param("alpha", alpha)
mlflow.log_param("l1_ratio", l1_ratio)
mlflow.log_metric("rmse", rmse)
mlflow.log_metric("r2", r2)
mlflow.log_metric("mae", mae)
```


APACHE
Spark
TM

Spark

Conceito

O Apache Spark é um mecanismo de computação unificado e um conjunto de bibliotecas para processamento paralelo de dados em clusters de computadores [CZ18]

Dataframe do Spark

Lendo dados de um csv

```
flightData2015 = spark\
.read\
.option("inferSchema", "true")\
.option("header", "true")\
.csv("/data/flight-data/csv/2015-summary.csv")
```

https://community.cloud.databricks.com/

Integrando Spark e Kafka

Subscrevendo dados no Kafka do Spark

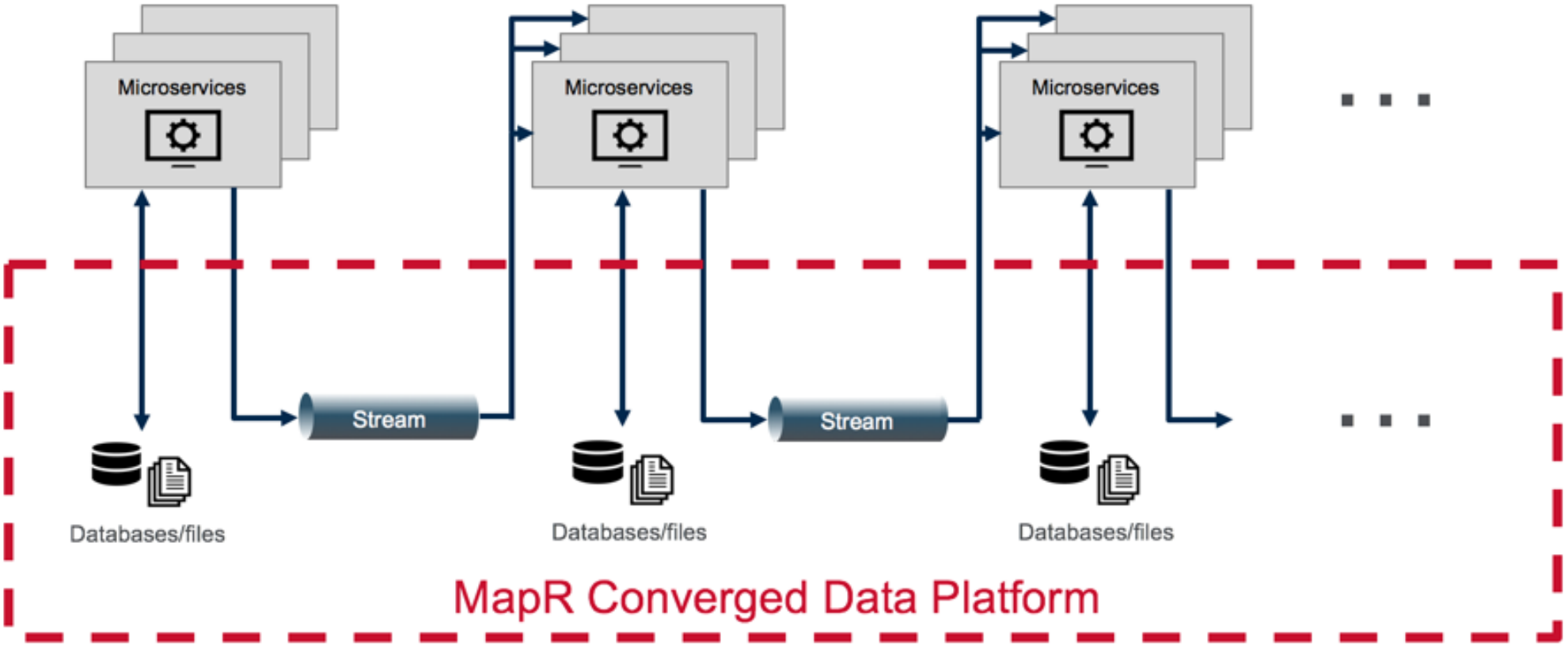
```
val ds1 = spark
  .readStream
  .format("kafka")
  .option("kafka.bootstrap.servers",\
   "host1:port1,host2:port2")
  .option("subscribe", "topic1")
  .load()
ds1.selectExpr("CAST(key AS STRING)", \
  "CAST(value AS STRING)")
  .as[(String, String)]
```

https://spark.apache.org/docs/2.1.0/structured-streaming-kafka-integration.html

Índice

Índice

Immediate access to operational (current) and analytical (historical) data in MapR

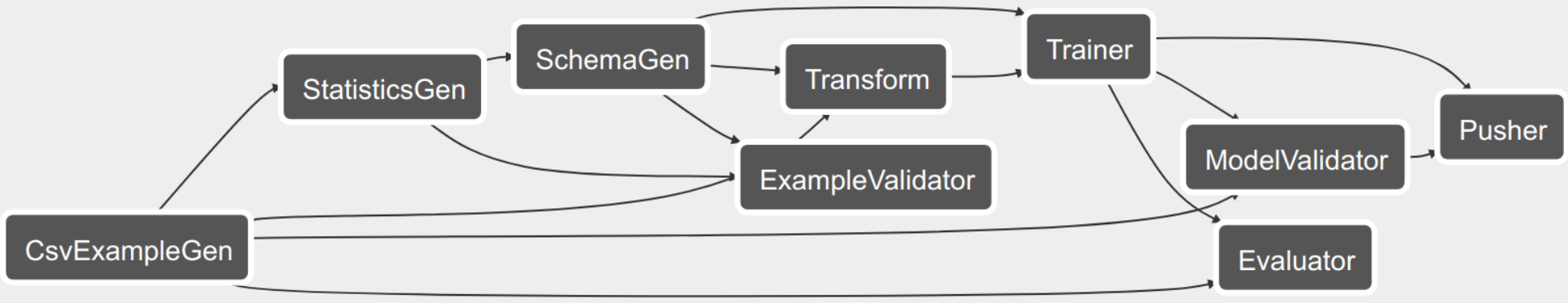


TensorFlow Extendend

Conceito

O TensorFlow Extended (TFX) é uma plataforma de ponta a ponta para implantar pipelines de ML em produção

`https://www.tensorflow.org/tfx`


CsvExampleGen
StatisticsGen
SchemaGen
Transform
Trainer
ExampleValidator
ModelValidator
Pusher
Evaluator

TensorFlow Extended

TensorFlow Extended usando Airflow

```
from tfx.orchestration import pipelinefrom

tfx.orchestration.airflow.airflow_runner import AirflowD

_airflow_config = { 'pipeline_name': 'your_ml_pipeline',
'schedule_interval': None,
'start_date': datetime.datetime(2019, 10, 28),
'pipeline_root': pipeline_root,
'enable_cache': True}

_pipeline = pipeline.Pipeline(
    pipeline_name='your_ml_pipeline',
    pipeline_root=pipeline_root,
    components=components, )

    AirflowDAGRunner(_airflow_config).run(_pipeline)
```

PMML

Conceito

A PMML (Predictive Model Markup Language) é uma linguagem baseada em XML que visa fornecer uma maneira de trocar diferentes modelos preditivos, para fins de classificação ou regressão, gerados usando uma técnica de mineração de dados ou aprendizado de máquina. O PMML foi originalmente desenvolvido pelo Data Mining Group (http://www.dmg.org/) em 1997 e sua versão mais recente (4.2.1) data de maio de 2014.

PMML book available on Amazon.com

- PMML is an **XML-based language** used to define statistical and data mining models and to share these between compliant applications.

- Mature **standard** developed by the DMG (Data Mining Group) to avoid proprietary issues and incompatibilities and to deploy models.

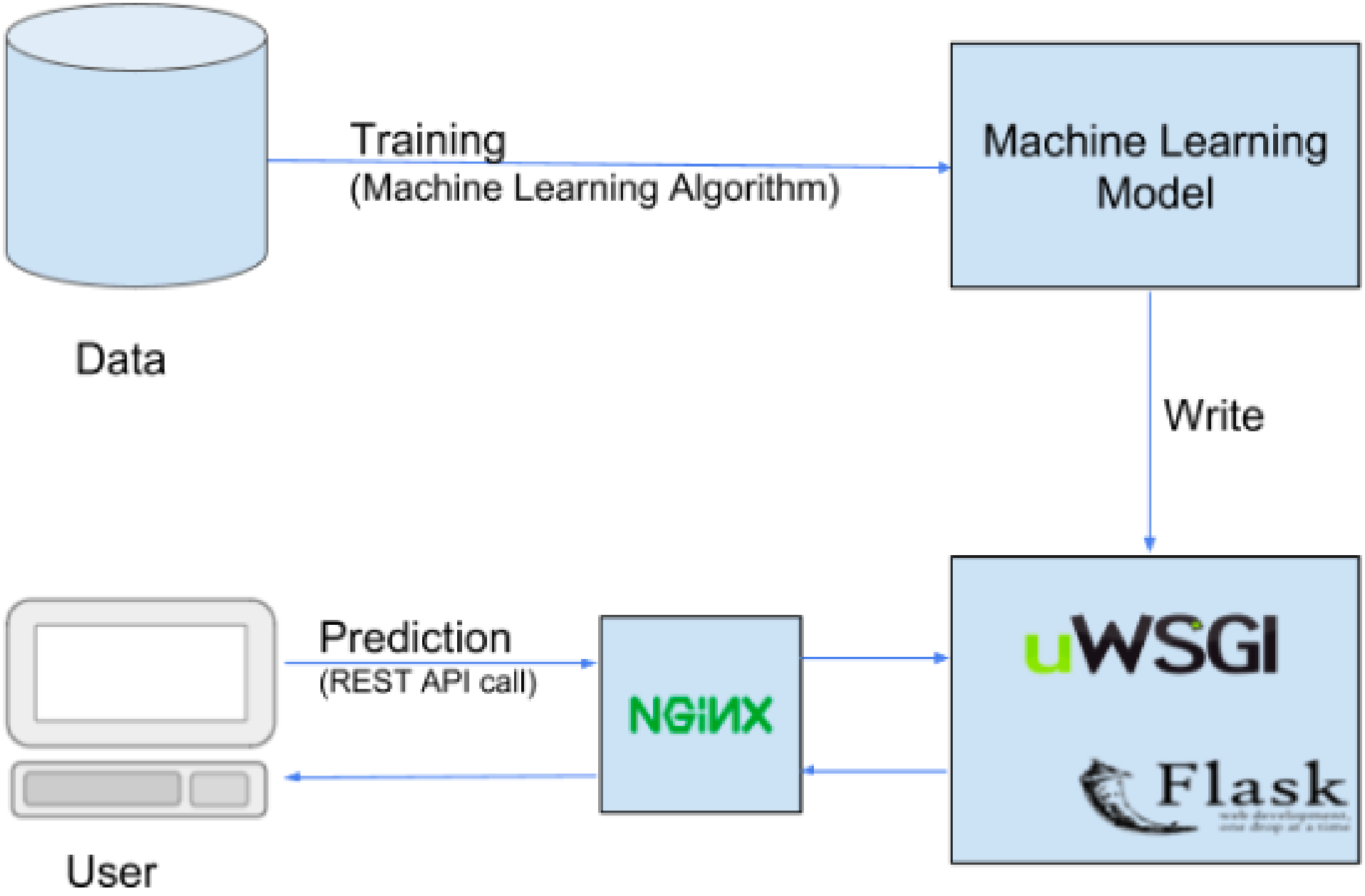
- Supported by all leading data mining tools, commercial and open-source.

- Allows for the **clear separation of tasks**: Model development vs. model deployment.

- **Eliminates the need for custom code** and proprietary model deployment solutions.

```
import pandas

iris_df = pandas.read_csv("Iris.csv")

from sklearn.tree import DecisionTreeClassifier
from sklearn2pmml.pipeline import PMMLPipeline

pipeline = PMMLPipeline([
        ("classifier", DecisionTreeClassifier())
])
pipeline.fit(iris_df[iris_df.columns.difference(["Species"])], iris_df["Species"])

from sklearn2pmml import sklearn2pmml

sklearn2pmml(pipeline, "DecisionTreeIris.pmml", with_repr = True)
```

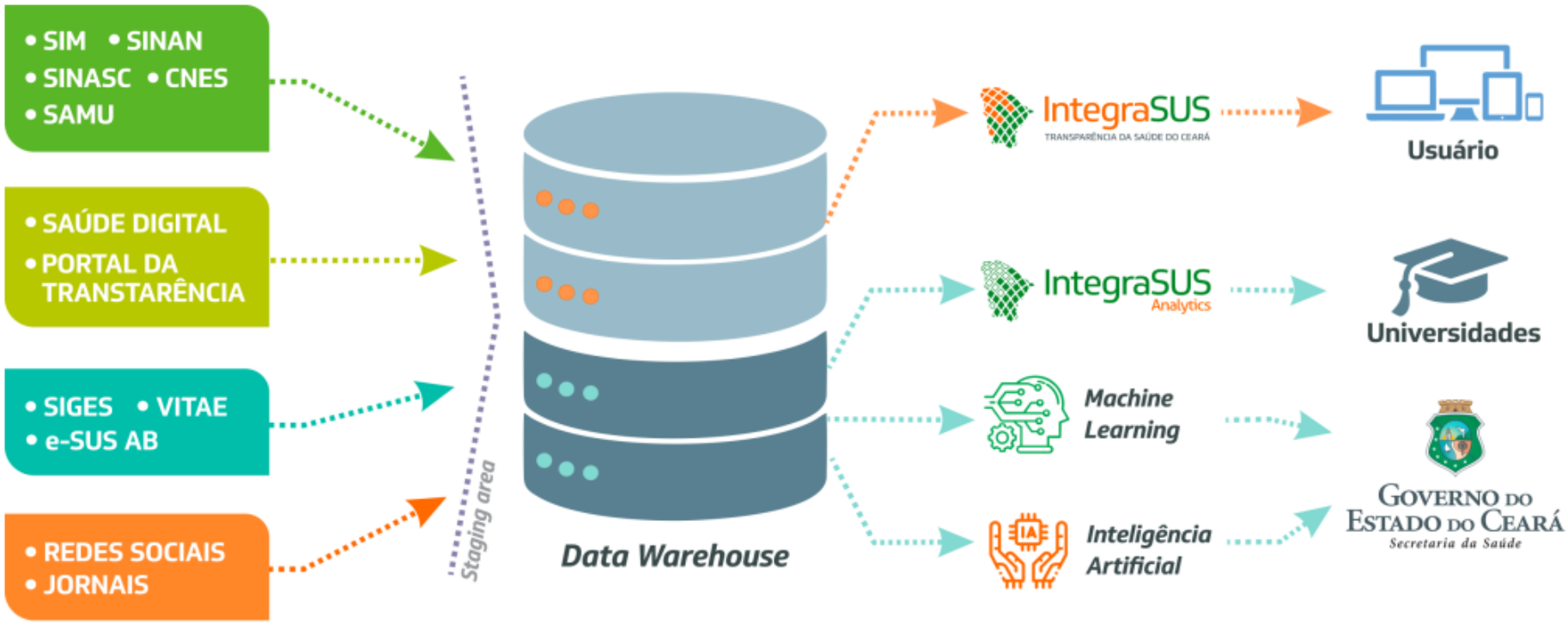

Flask
web development,
one drop at a time


Training
(Machine Learning Algorithm)
Machine Learning Model
Data
Write
Prediction
(REST API call)
NGINX
uWSGI
Flask
User

Exemplo de uso de modelo com Flask

```
from flask import Flask, request,\
      redirect, url_for, flash, jsonify
import numpy as np
import pickle as p
import json

app = Flask(__name__)

@app.route('/api/', methods=['POST'])
def makecalc():
    data = request.get_json()
    prediction = np.array2string(model.predict(data))

    return jsonify(prediction)
```

ORGANIZAÇÃO E FLUXO DA INFORMAÇÃO



Conclusão

Você esta apto a:

Compreender o que é Aprendizado de Máquina e DataScience
Entender o que é classificação, regressão e clustering

Bibliografia I

Bill Chambers and Matei Zaharia, Spark: The definitive guide: Big data processing made simple, ”O’Reilly Media, Inc.”, 2018.