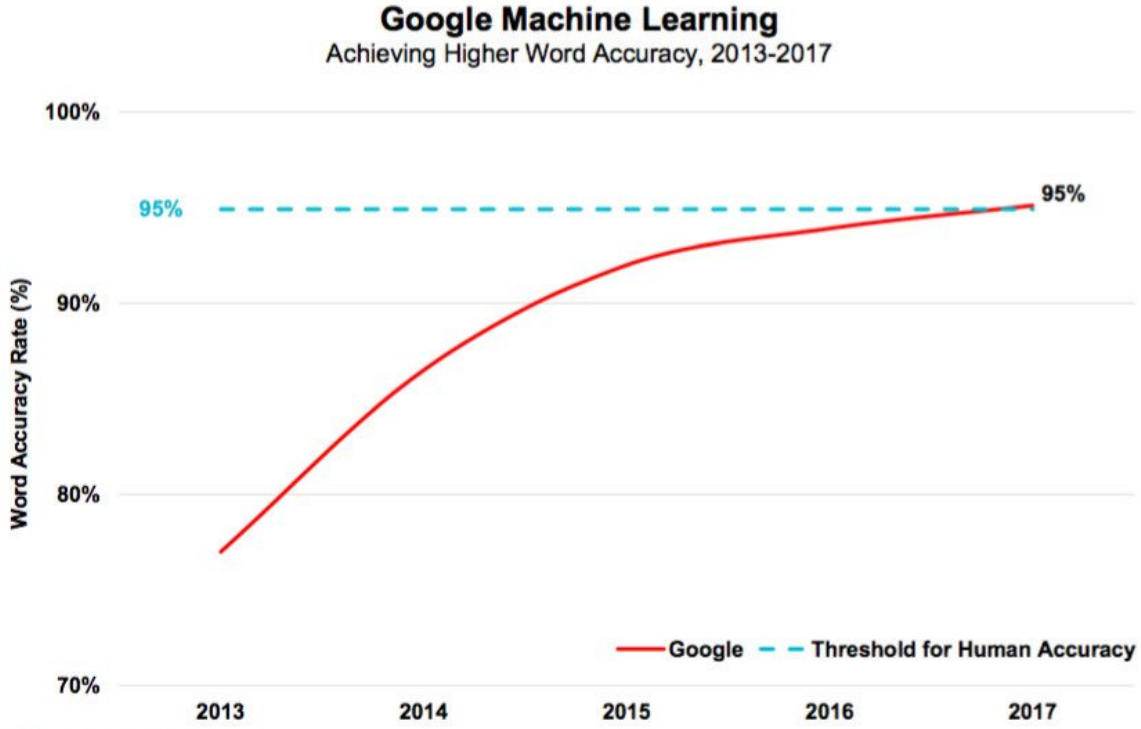
Engenharia de Dados e Engenharia de Machine Learning

como fazer projetos de Data Science usando um processo sustentável

Fabiane Bizinella Nardon
Chief Data Scientist
@fabianenardon

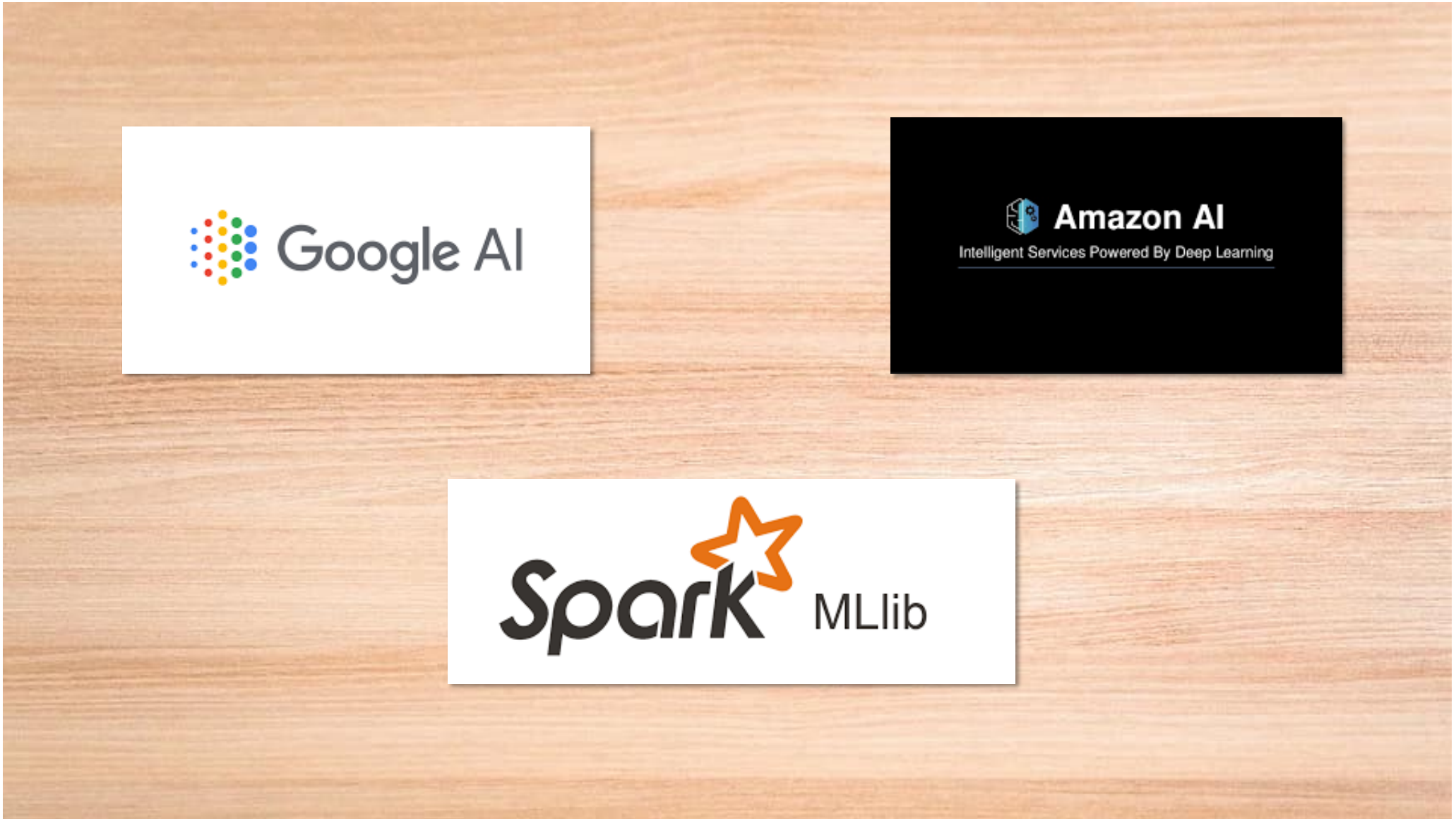

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
Harvard Business Review
DATA
Data Scientist: The Sexiest Job of the 21st Century
by Thomas H. Davenport and D.J. Patil
FROM THE OCTOBER 2012 ISSUE
SUMMARY SAVE SHARE COMMENT TEXT SIZE PRINT $8.95 BUY COPIES
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join.
THE NEW ARTIFICIAL INTELLIGENCE
THE FUTURE OF NANOBOTS
DETECTING CANCER
KILLER ROBOTS?
Forbes
OS DOIS RITMOS DA LAVA-JATO
Curitiba: 107 condenados
Brasília: nenhum
veja
DE MÃOS DADAS COM A INTELIGÊNCIA ARTIFICIAL
Longe dos cenários futurísticos da ficção científica, ela já faz parte do presente. Mas em que medida pode servir ao ser humano e, ao mesmo tempo, ameaçá-lo?
EXCLUSIVO Artigo de Yuval Noah Harari, autor de Homo Deus: uma Breve História do Amanhã
Science
SPECIAL ISSUE
ARTIFICIAL INTELLIGENCE
INSIDE: A 14-PAGE SPECIAL REPORT ON FINANCIAL TECHNOLOGY
The Economist
How to fix America's inner cities
The self-service economy
Time to open up Indonesia
Inside the anti-bribery business
Why humans cause heatwaves
Artificial Intelligence
The promise and the peril
MAY 9TH - 15TH 2015
Worldwide ex UK

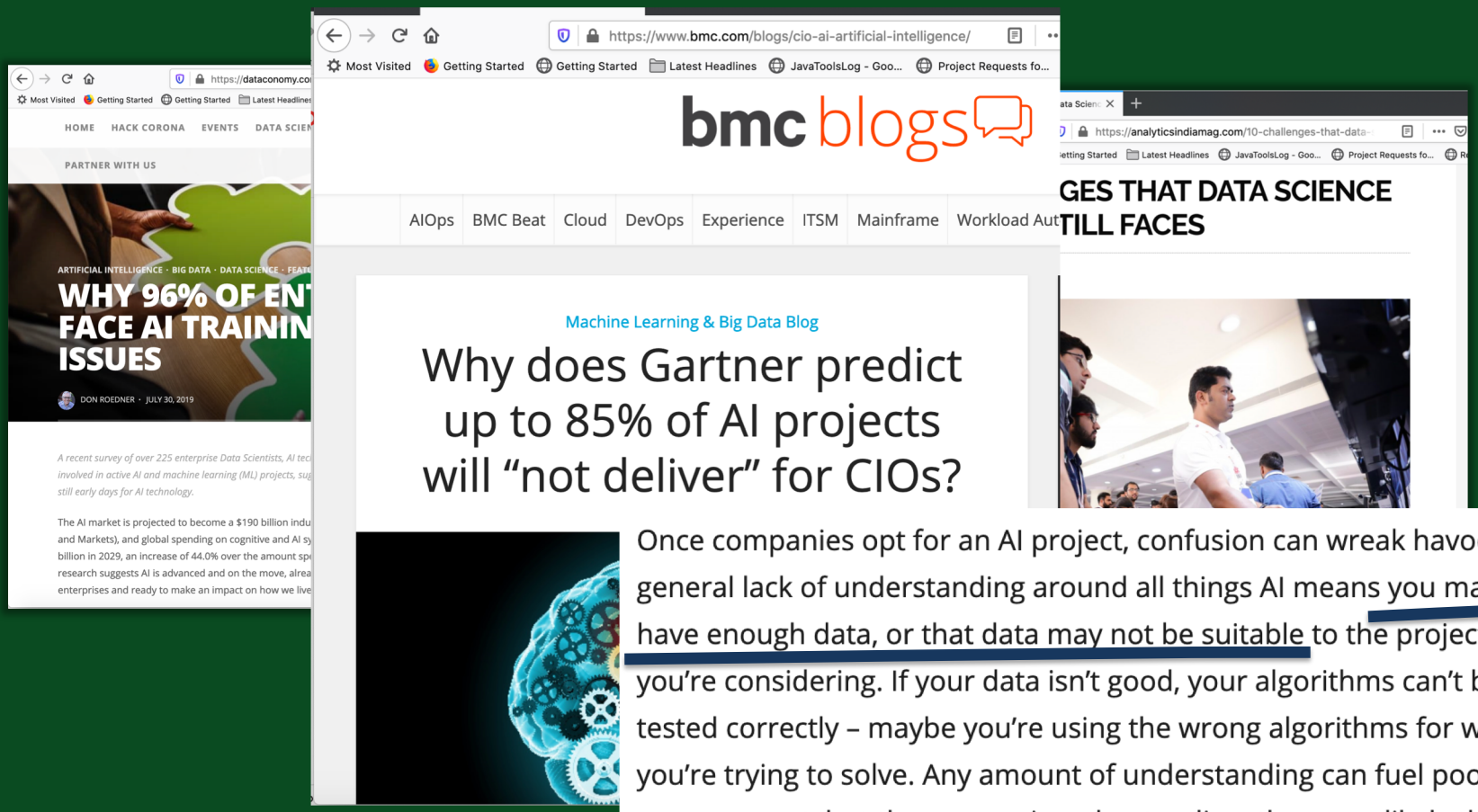
…Voice-Based Platform *Back-Ends* = Voice Recognition Accuracy Continues to Improve



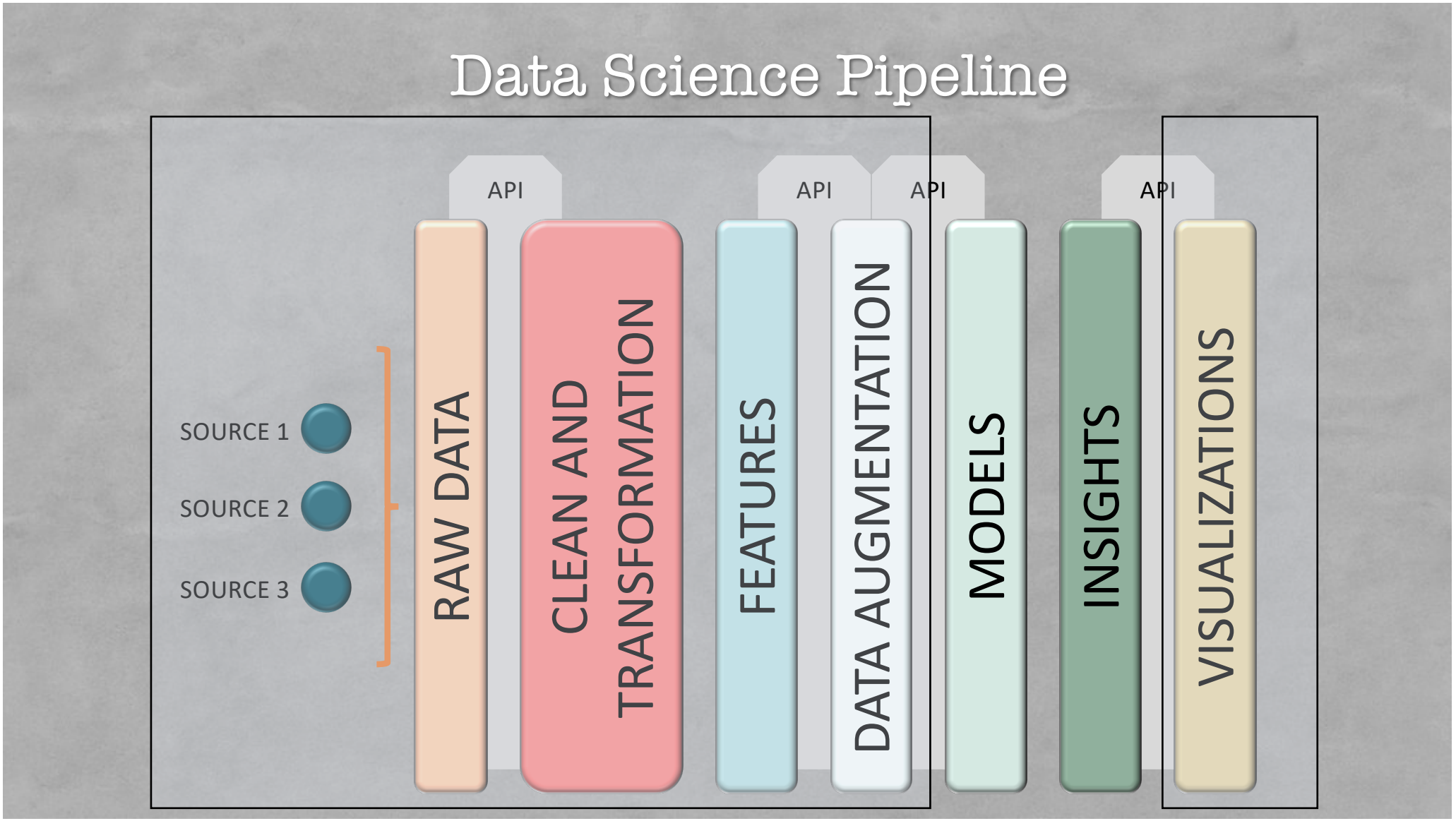
KLEINER PERKINS

Source: Google (5/17)
Note: Data as of 5/17/17 and refers to recognition accuracy for English language. Word error rate is evaluated using real world search data which is extremely diverse and more error prone than typical human dialogue.


Google AI
Amazon AI
Intelligent Services Powered By Deep Learning
Spark
MLlib

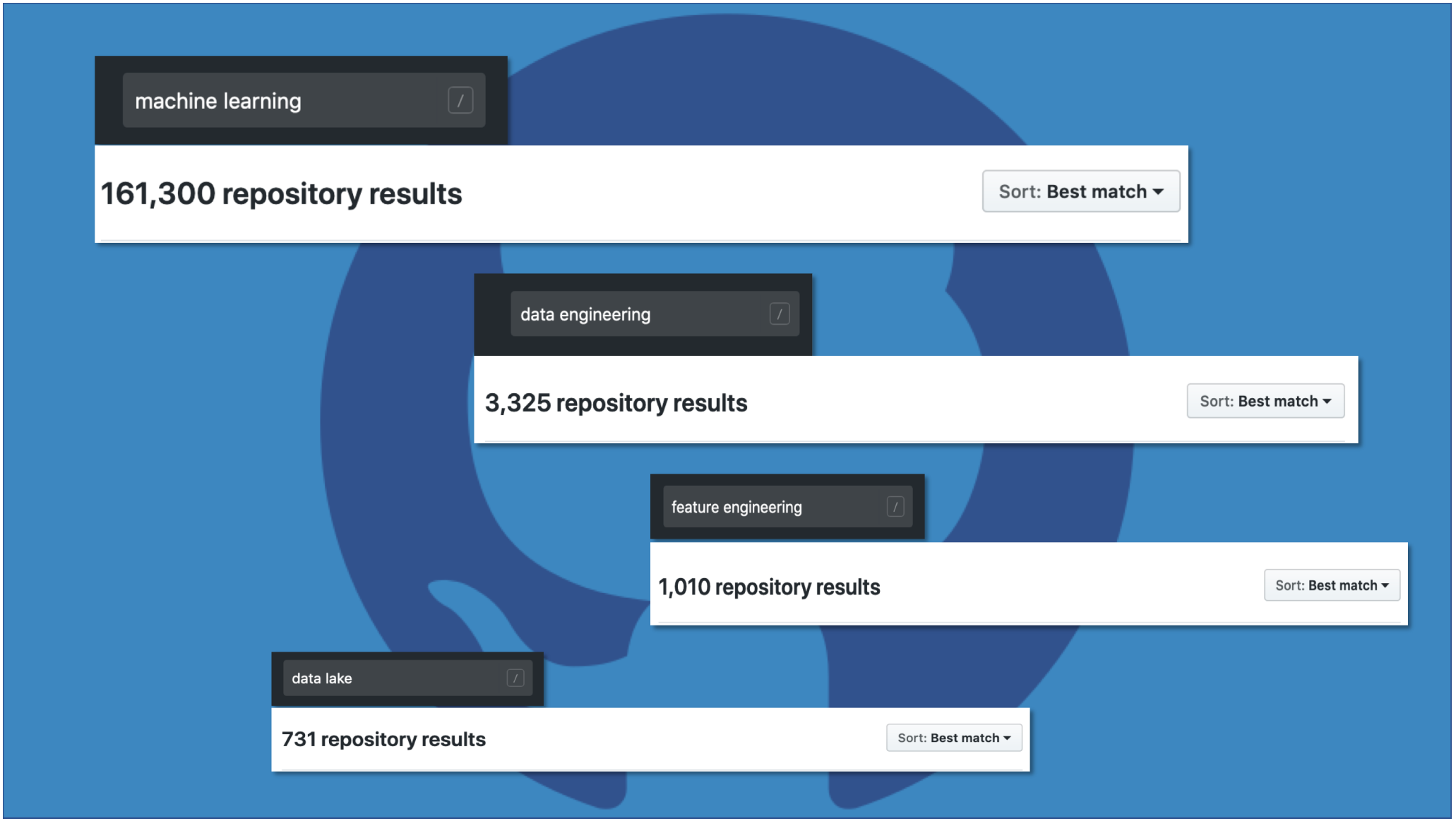


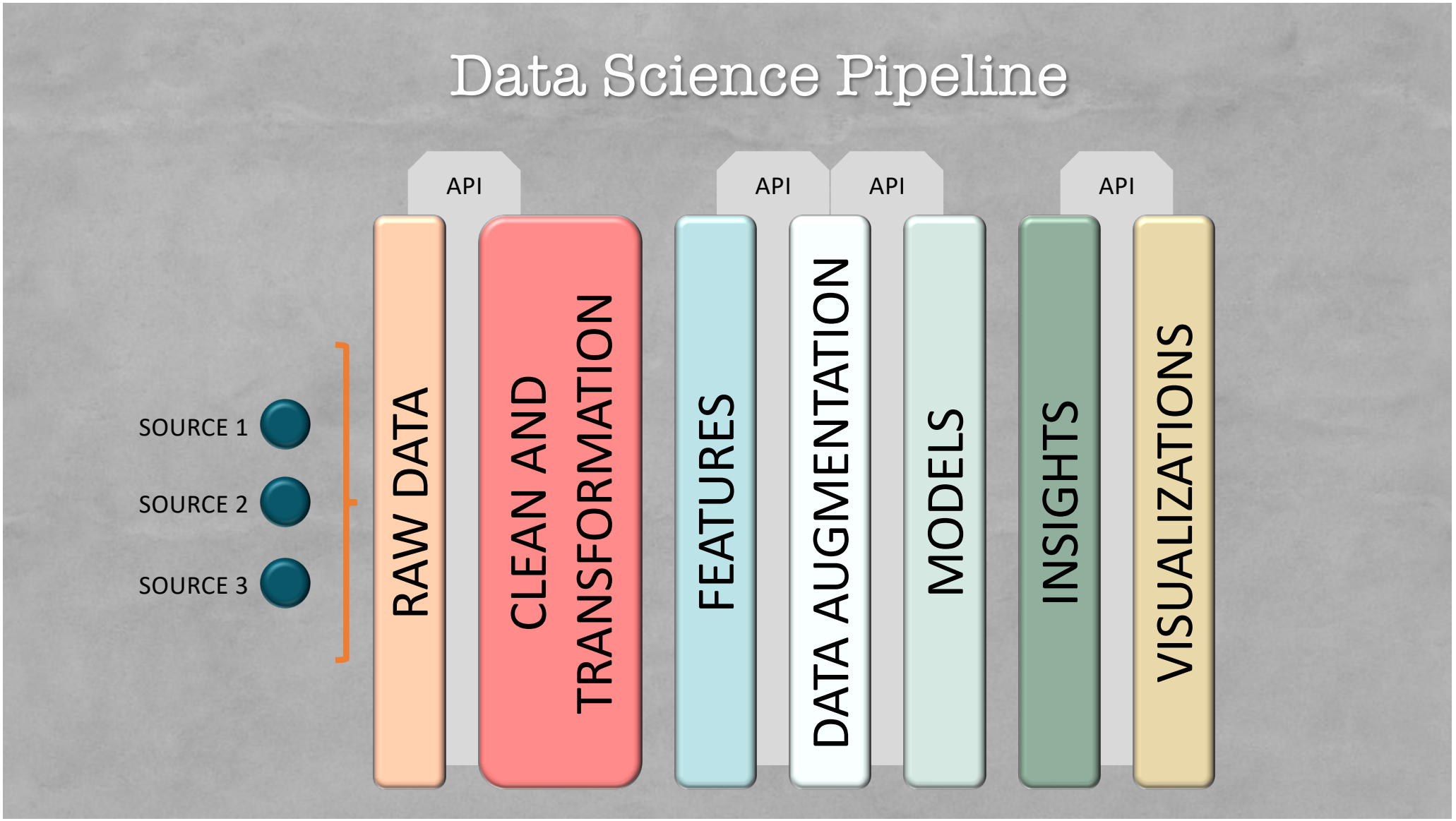
Once companies opt for an AI project, confusion can wreak havoc. A general lack of understanding around all things AI means you may not have enough data, or that data may not be suitable to the project you’re considering. If your data isn’t good, your algorithms can’t be tested correctly – maybe you’re using the wrong algorithms for what you’re trying to solve. Any amount of understanding can fuel poor team management, but the more misunderstanding, the more likely the team is just wasting time.


Data Science Pipeline
API
API
API
API
SOURCE 1
SOURCE 2
SOURCE 3
RAW DATA
CLEAN AND TRANSFORMATION
FEATURES
DATA AUGMENTATION
MODELS
INSIGHTS
VISUALIZATIONS

90% do tempo de um projeto de Data Science é gasto em Engenharia de Dados

https://www.kdnuggets.com/2019/03/most-impactful-ai-trends-2018-rise-ml-engineering.html


machine learning
/
161,300 repository results
Sort: Best match
data engineering
/
3,325 repository results
Sort: Best match
feature engineering
/
1,010 repository results
Sort: Best match
data lake
/
731 repository results
Sort: Best match


Data Science Pipeline
API
API
API
API
SOURCE 1
SOURCE 2
SOURCE 3
RAW DATA
CLEAN AND TRANSFORMATION
FEATURES
DATA AUGMENTATION
MODELS
INSIGHTS
VISUALIZATIONS


ESCALA

3.5 bilhões de novos registros
4.680 pipelines executados

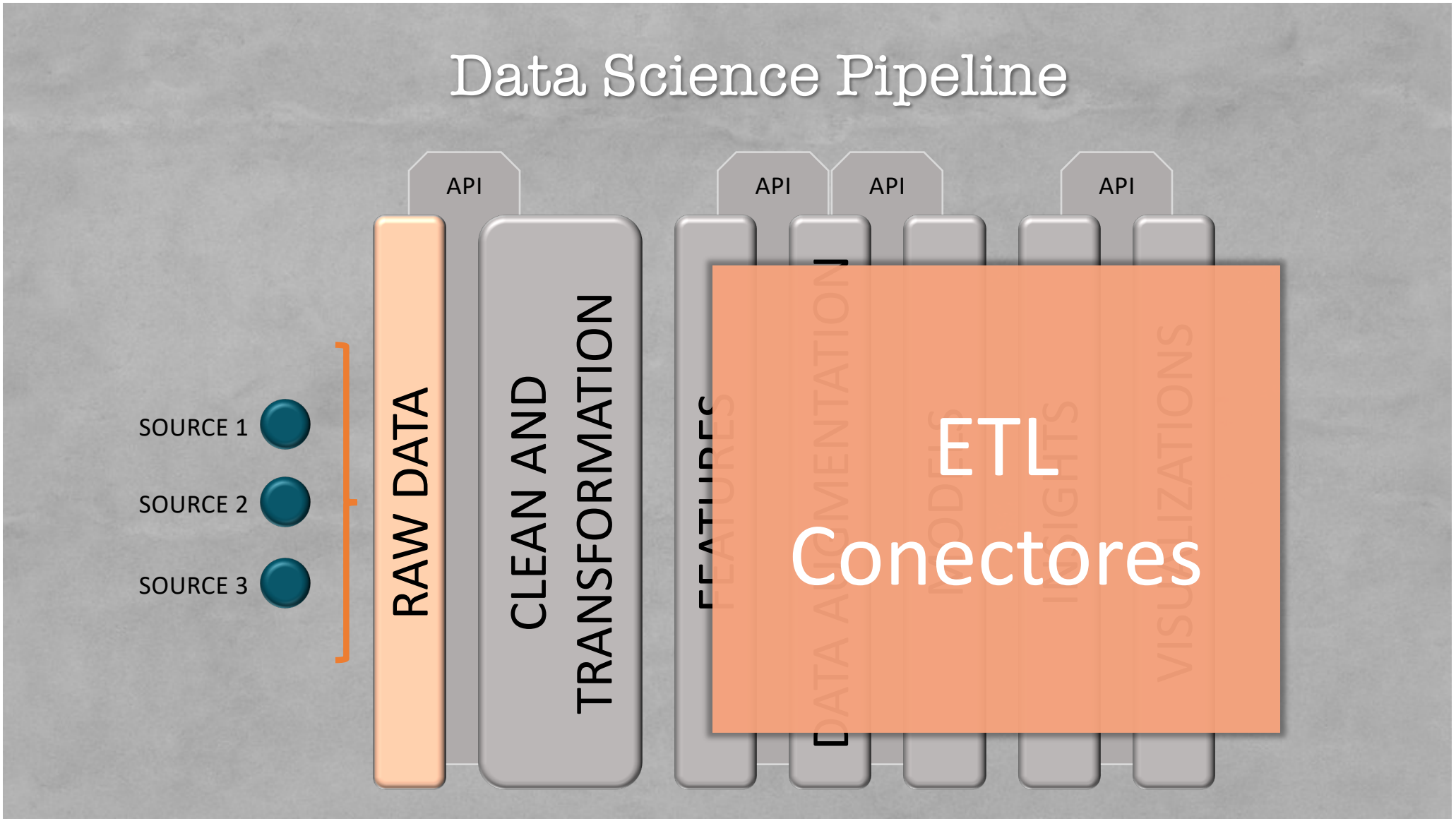
em um dia

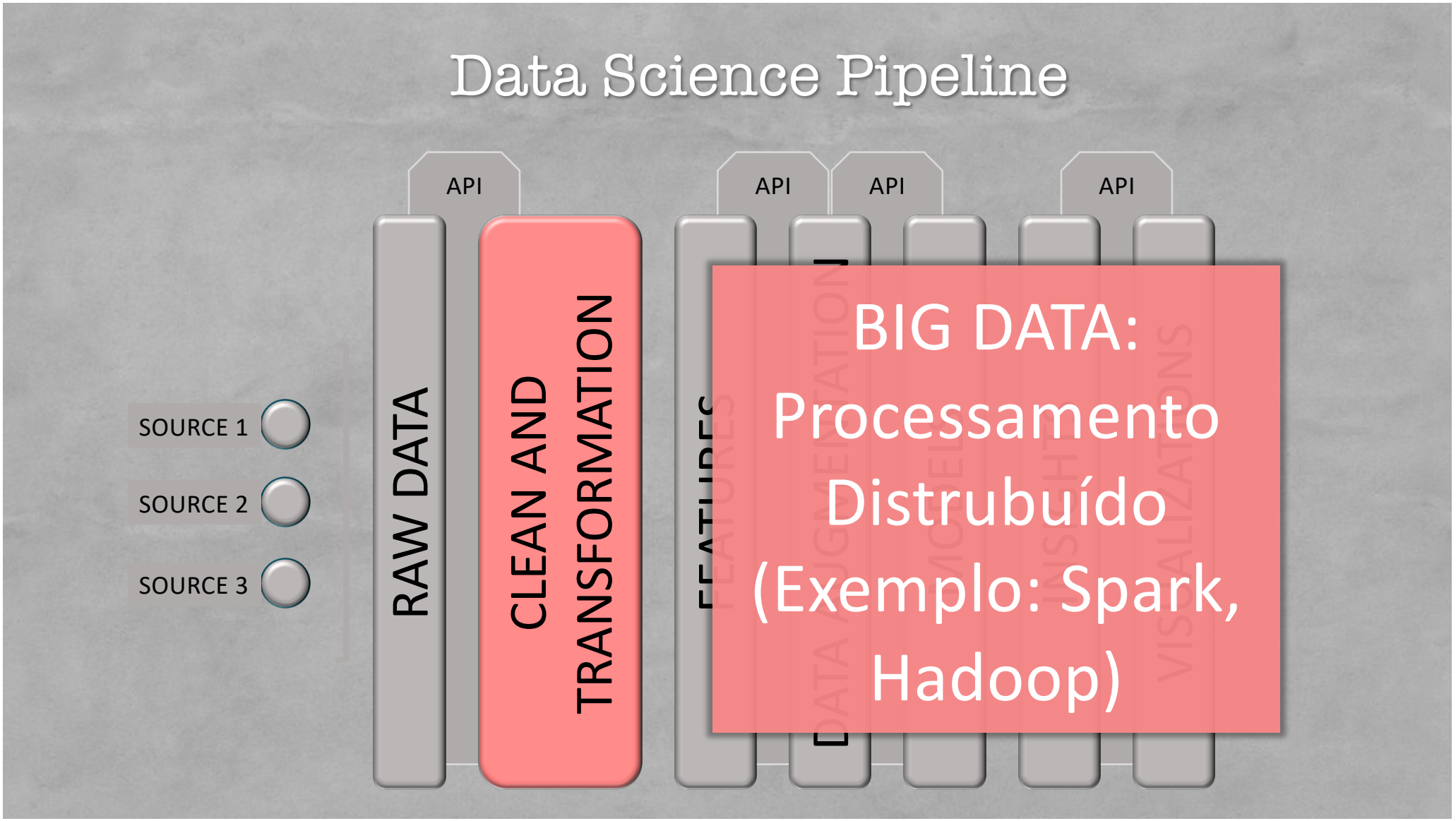
Teoria do Mundo Perfeito

1. Não existe Big Data nem código legado

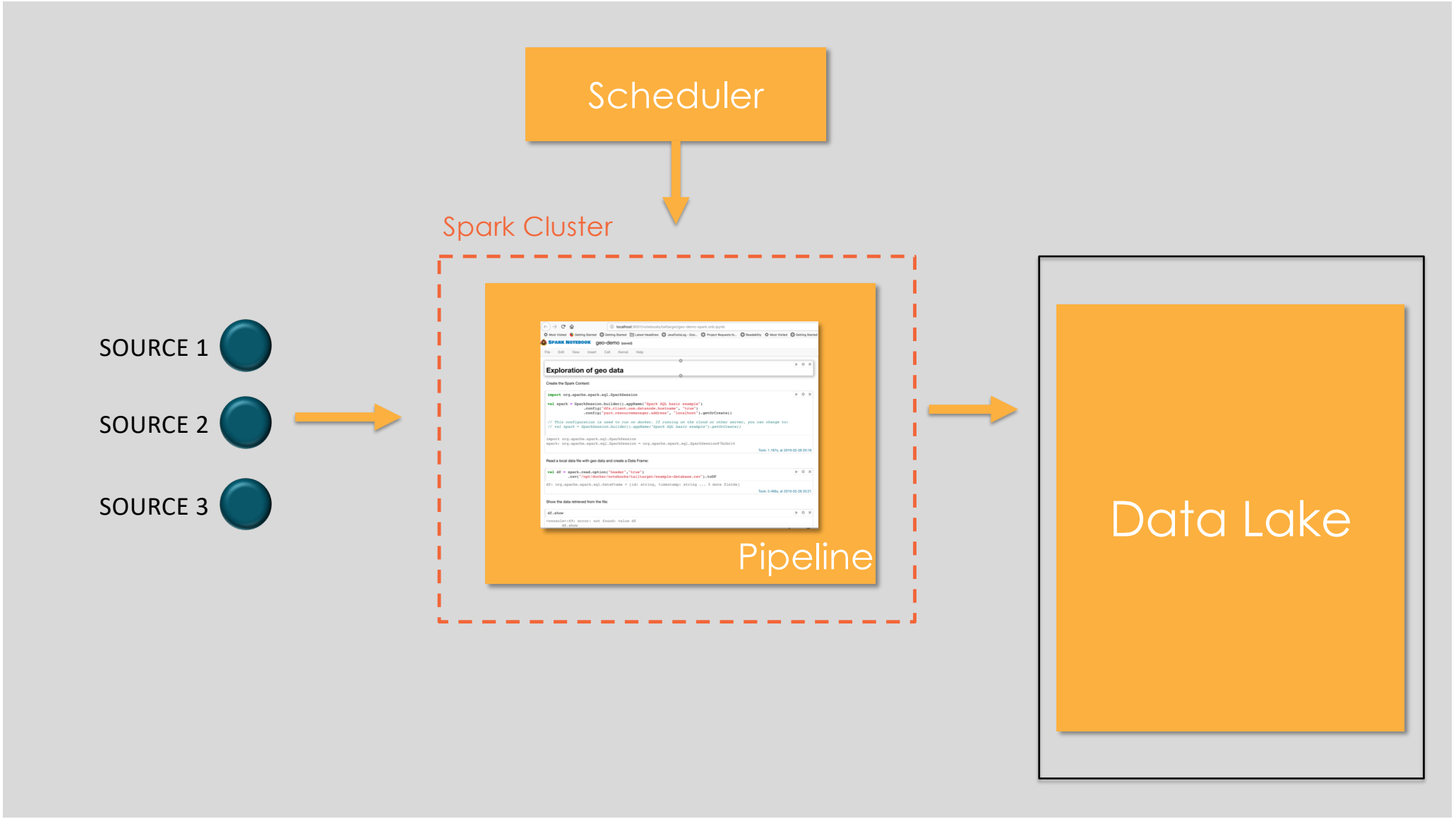
2. Depois que você faz a experimentação, o trabalho acabou

3. Existe um Data Lake com todos os dados que você precisa


Data Science Pipeline
API
API
API
API
SOURCE 1
SOURCE 2
SOURCE 3
RAW DATA
CLEAN AND TRANSFORMATION
ETL
Conectores


Data Science Pipeline
API
API
API
API
SOURCE 1
SOURCE 2
SOURCE 3
RAW DATA
CLEAN AND TRANSFORMATION
BIG DATA: Processamento Distrubuído (Exemplo: Spark, Hadoop)


> Spark SQL
val records = spark.read.option("header","true")
.csv("example-database.csv").toDF
records.createOrReplaceTempView(”records")
val transformed = spark.sql(
"SELECT ucase(name) as name from records"
)
transformed.write
.format("com.databricks.spark.csv")
.option("header", "true")
.save("newrecords.csv");


Scheduler
Spark Cluster
SOURCE 1
SOURCE 2
SOURCE 3
Exploration of geo data
Pipeline
Data Lake


>code
>lake

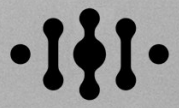

> Spark SQL
val records =
spark.read.option("header","true")
.csv("example-database.csv").toDF
val filtered =
records.filter("type = 'android'")
filtered.write
.format("com.databricks.spark.csv")
.option("header", "true")
.save("newrecords.csv");


> Spark SQL + Meu Código
val geoHash = new my.udf.SparkGeoHash();
spark.udf.register("geoHash", geoHash,
DataTypes.StringType);
val recordsWithFunction =
filtered.withColumn("generatedGeoHash",
callUDF("geoHash", col("latitude"),
col("longitude"), lit(12)))
.select(col("id"), col("latitude"),
col("longitude"),
col("geo_hash"),
col("generatedGeoHash"))


> Plugins
> Tipos de dados com semântica
> Transformações
> Agregações

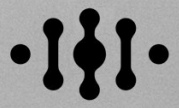
> Tipos de Dados com Semântica

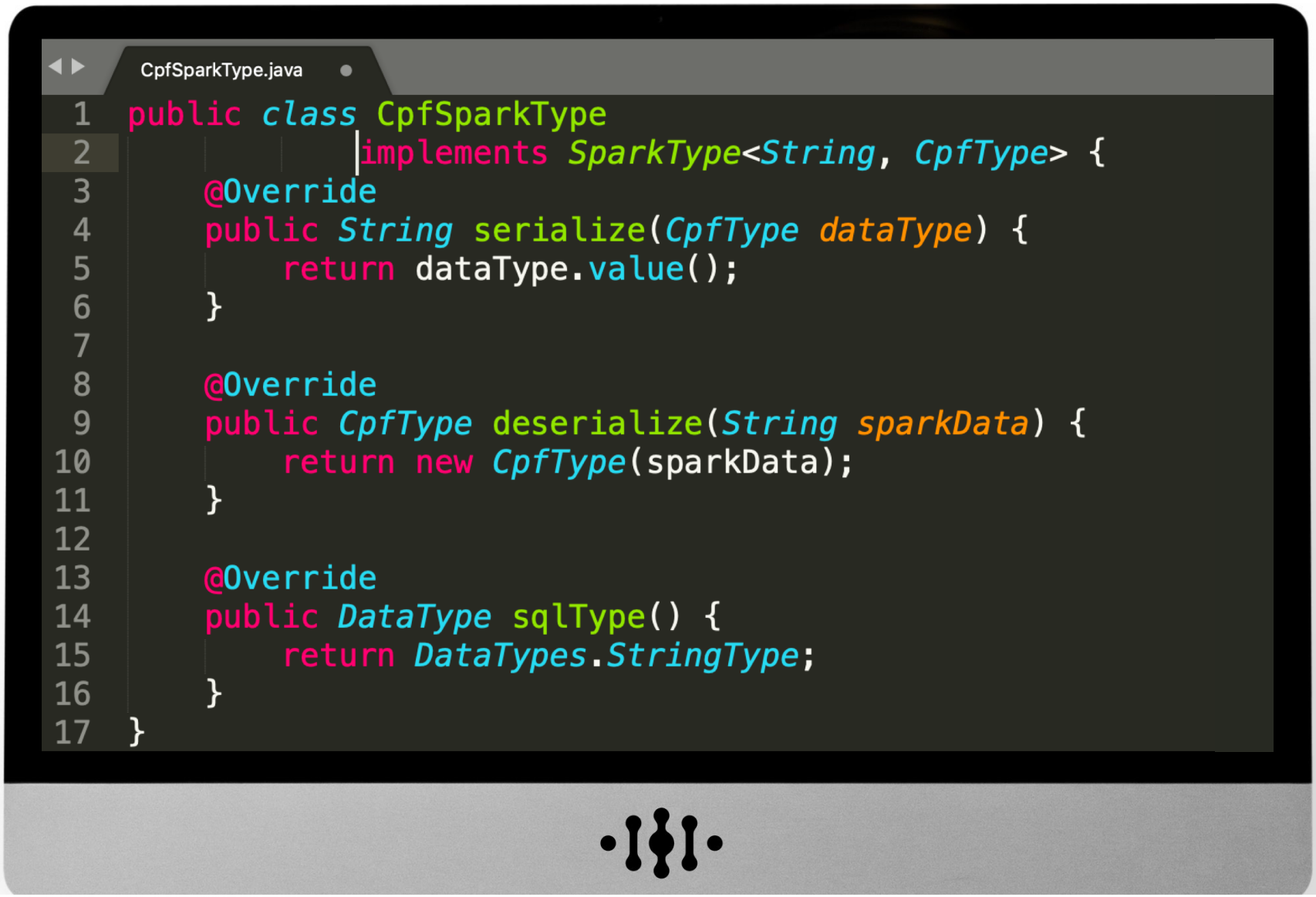
```
val records =  spark.read.option("header","true")
          .csv("example-database.csv").toDF

records.printSchema

root
 |-- id: string (nullable = true)
 |-- timestamp: string (nullable = true)
 |-- type: string (nullable = true)
 |-- latitude: string (nullable = true)
 |-- longitude: string (nullable = true)
 |-- horizontal_accuracy: string (nullable = true)
 |-- geo_hash: string (nullable = true)
```



> Tipos de Dados com Semântica

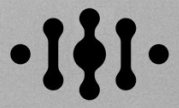
```
val schema = StructType(Array(
    StructField("ad_id",StringType,true),
    StructField("utc_timestamp",LongType,true),
    StructField("id_type",StringType,true),
    StructField("latitude",DoubleType,true),
    StructField("longitude",DoubleType,true),
    StructField("horizontal_accuracy",DoubleType,true),
    StructField("geo_hash",StringType,true)
    ))

val records =  spark.read.schema(schema)
          .option("header","true")
          .csv("example-database.csv").toDF
```




CpfSparkType.java
public class CpfSparkType
implements SparkType<String, CpfType> {
@Override
public String serialize(CpfType dataType) {
return dataType.value();
}
@Override
public CpfType deserialize(String sparkData) {
return new CpfType(sparkData);
}
@Override
public DataType sqlType() {
return DataTypes.StringType;
}
}

> Tipos de Dados com Semântica

```
val schema = StructType(Array(
    StructField("id",StringType,true),
    StructField("cpf",
       new CpfSparkType().sqlType(),true)
    )
)

val records = spark.read.schema(schema)
                        .option("header","true")
                        .csv("cpfs.csv")
```




Tipo de dados + semântica = muda o jogo
Detectores de tipo = +inteligência
Privacidade + LGPD
Validação
Feature engineering automatizada

Teoria do Mundo Perfeito

1. Não existe Big Data nem código legado

2. Depois que você faz a experimentação, o trabalho acabou

3. Existe um Data Lake com todos os dados que você precisa

EXPERIMENTAÇÃO

PRODUÇÃO

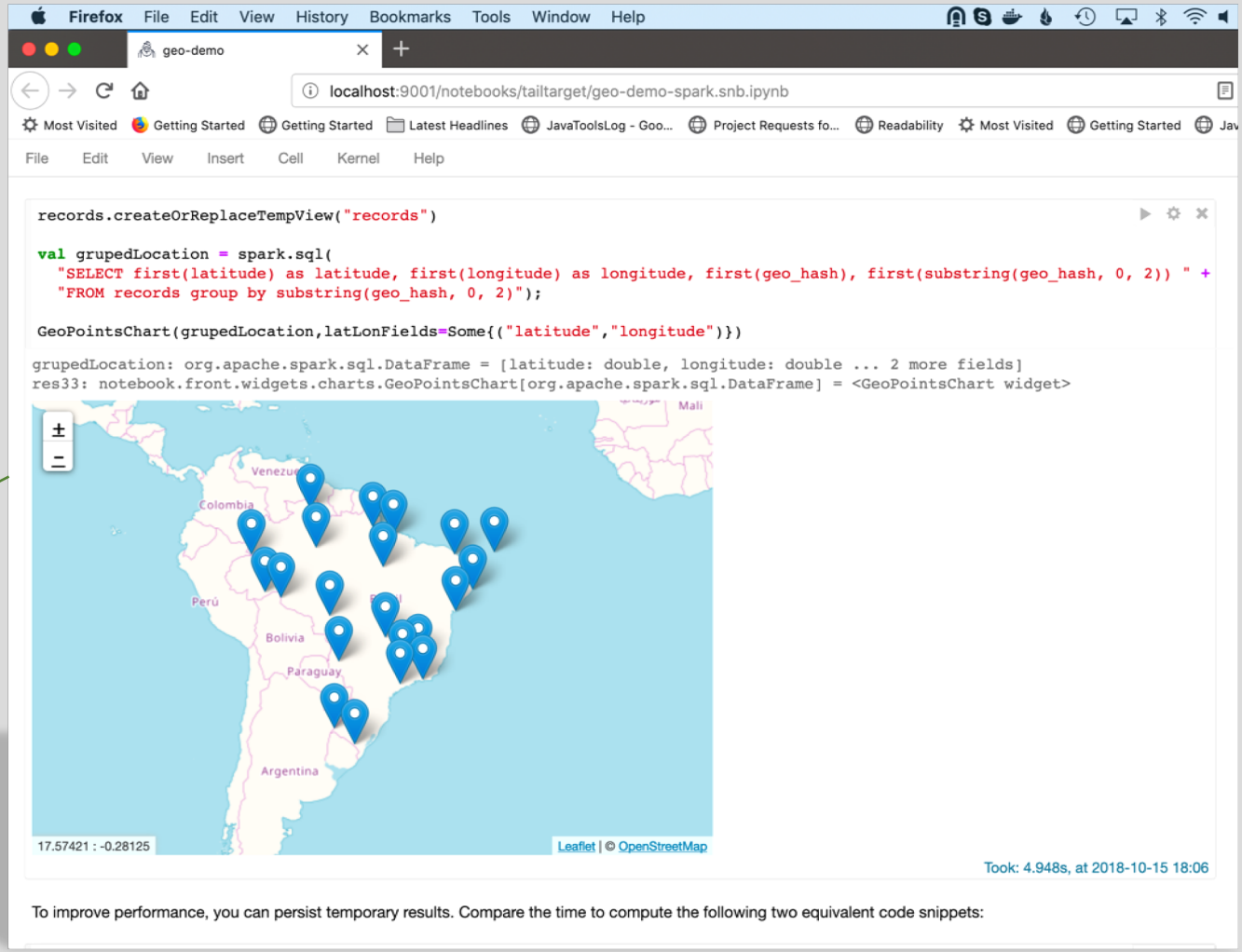
Experimentação:

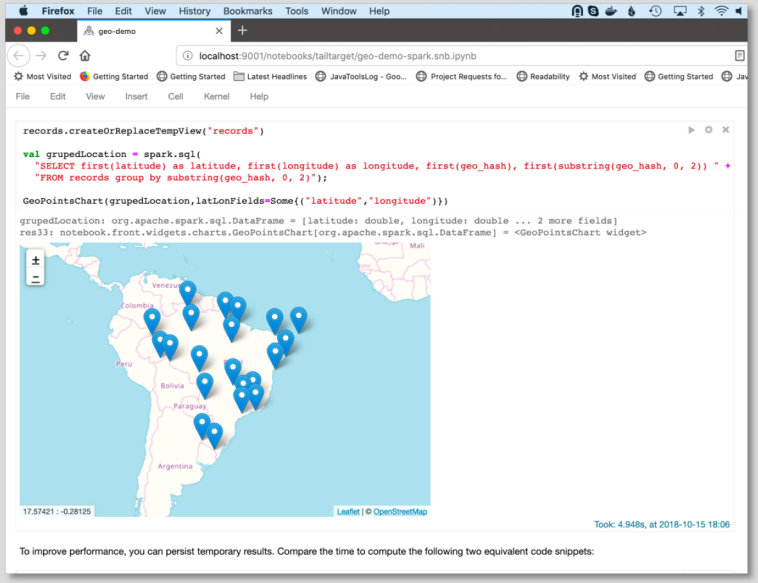
- Amostras
- Notebooks
- Execução interativa

Produção:

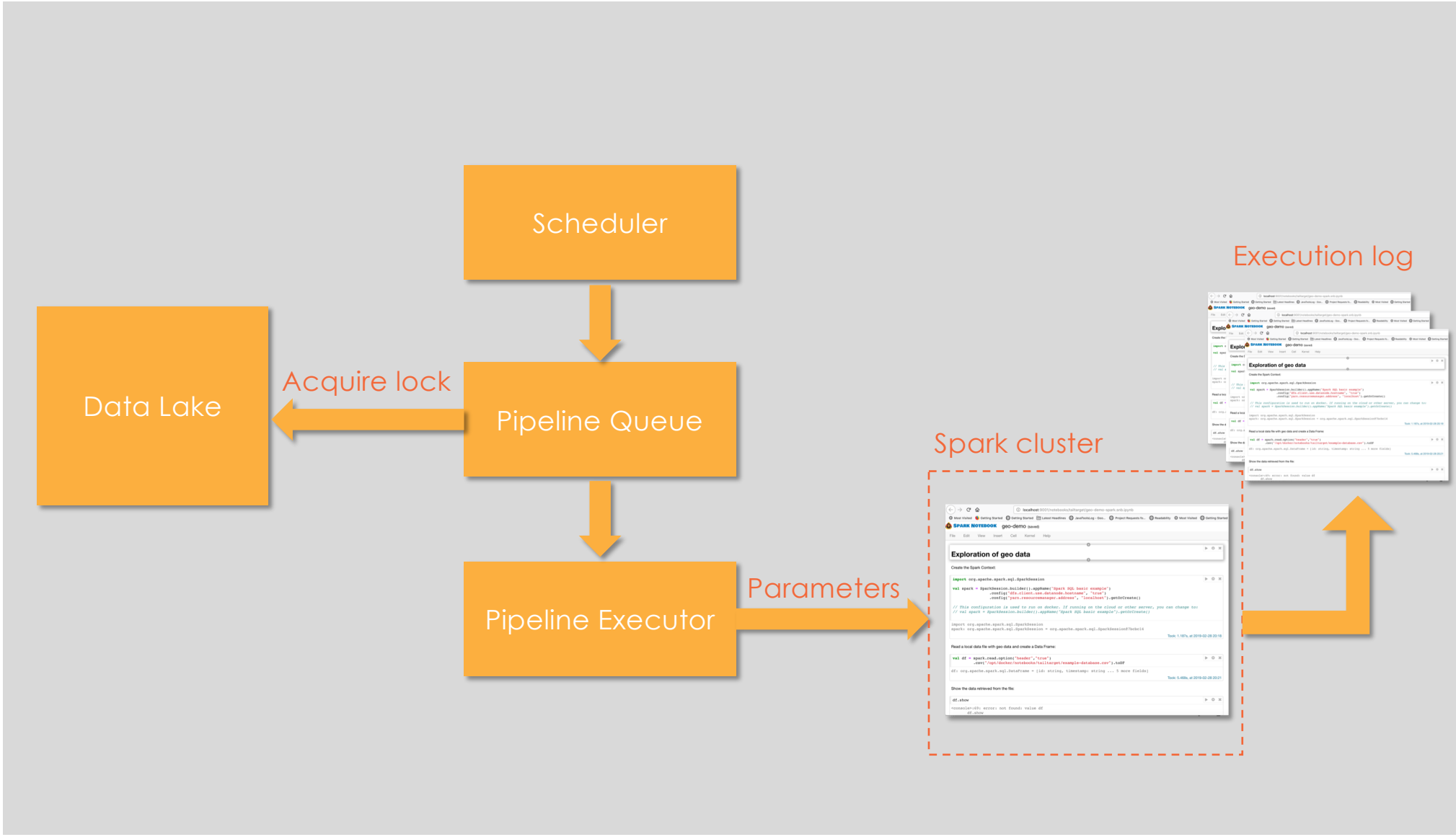
- Big Data
- Schedule
- Execução em lote
- Log

code lake
+
data lake
(amostras)





+ Parâmetros + Scheduler


Scheduler
Data Lake
Acquire lock
Pipeline Queue
Pipeline Executor
Parameters
Spark cluster
Execution log

Teoria do Mundo Perfeito

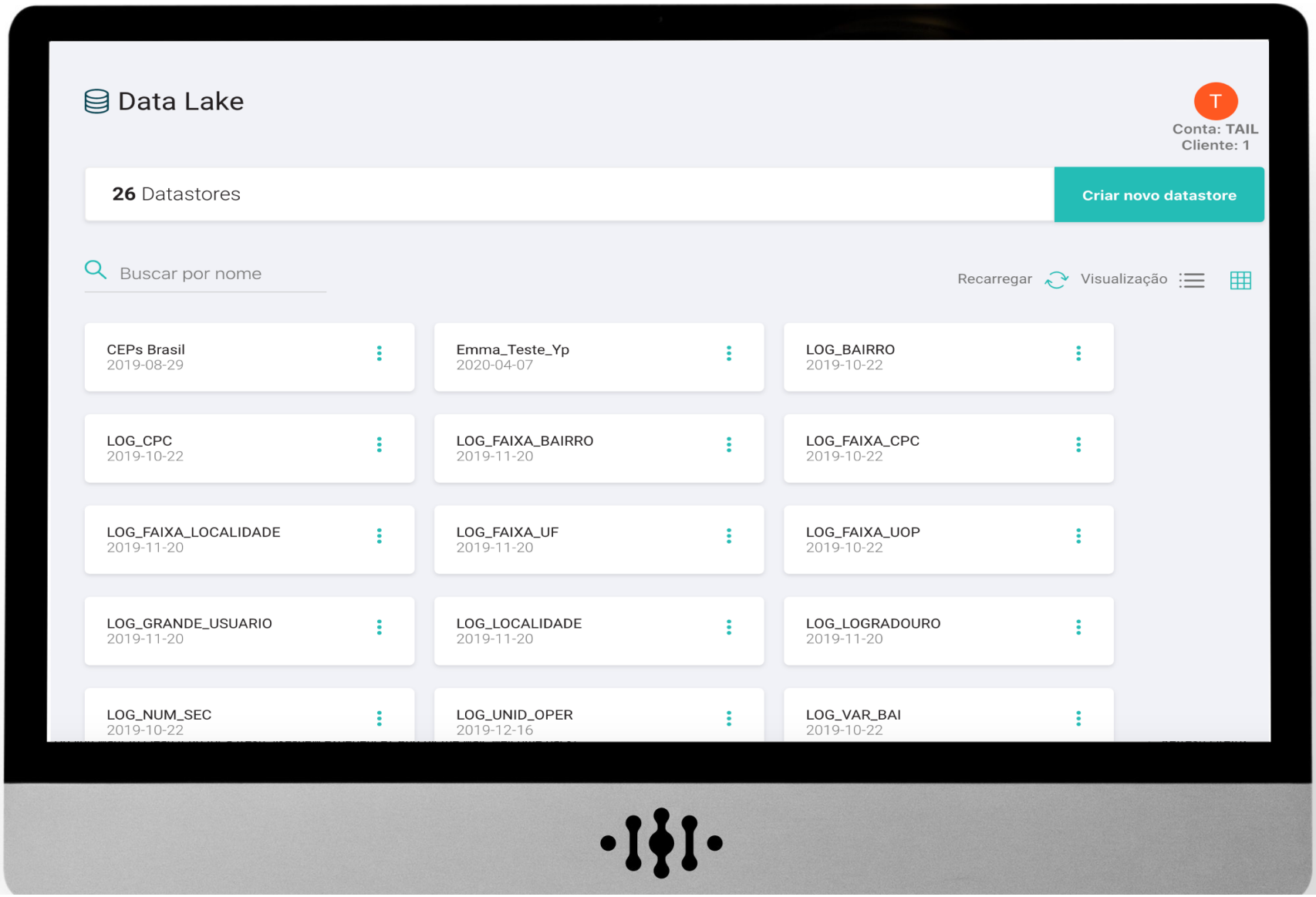
1. Não existe Big Data nem código legado

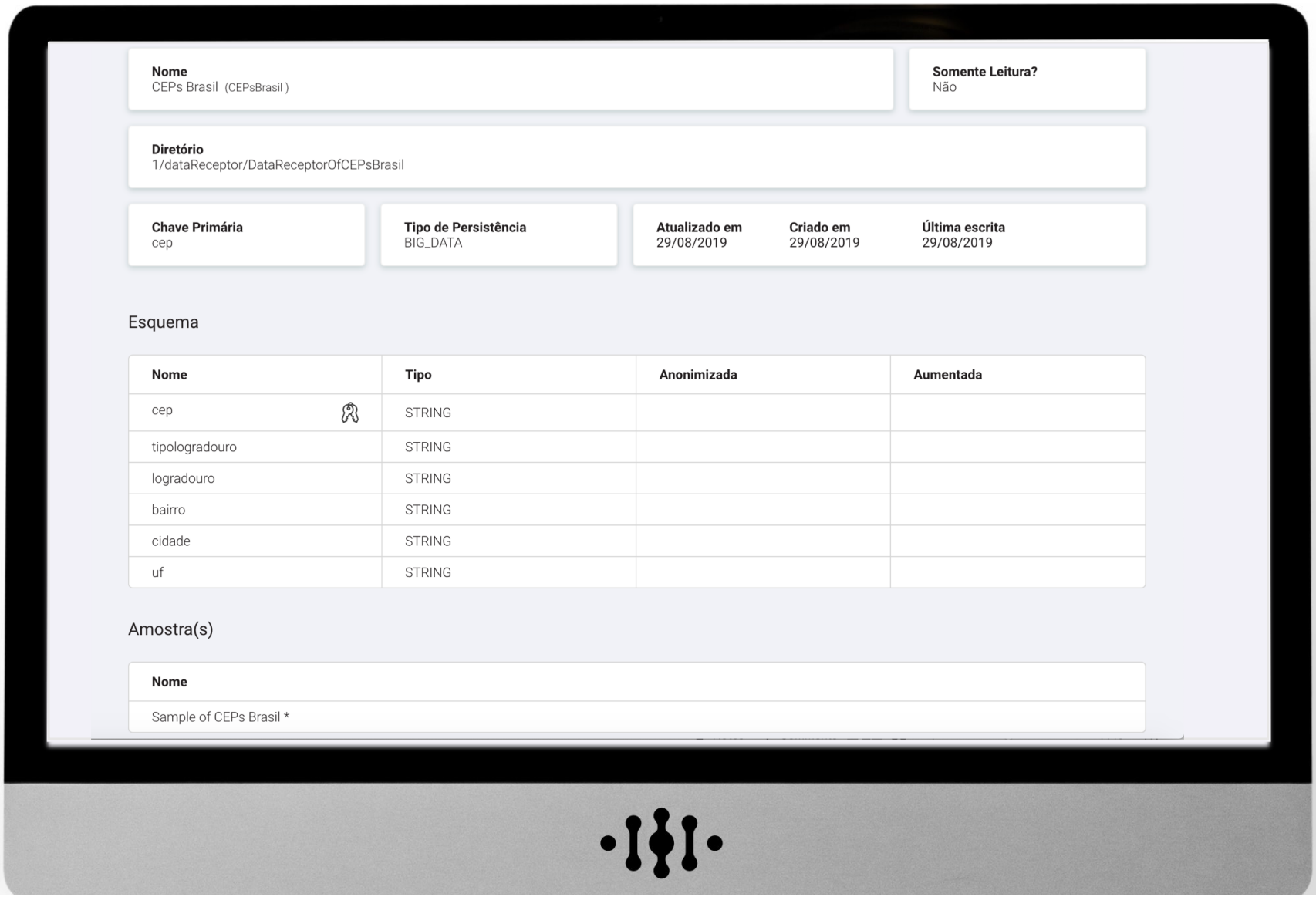
2. Depois que você faz a experimentação, o trabalho acabou

3. Existe um Data Lake com todos os dados que você precisa

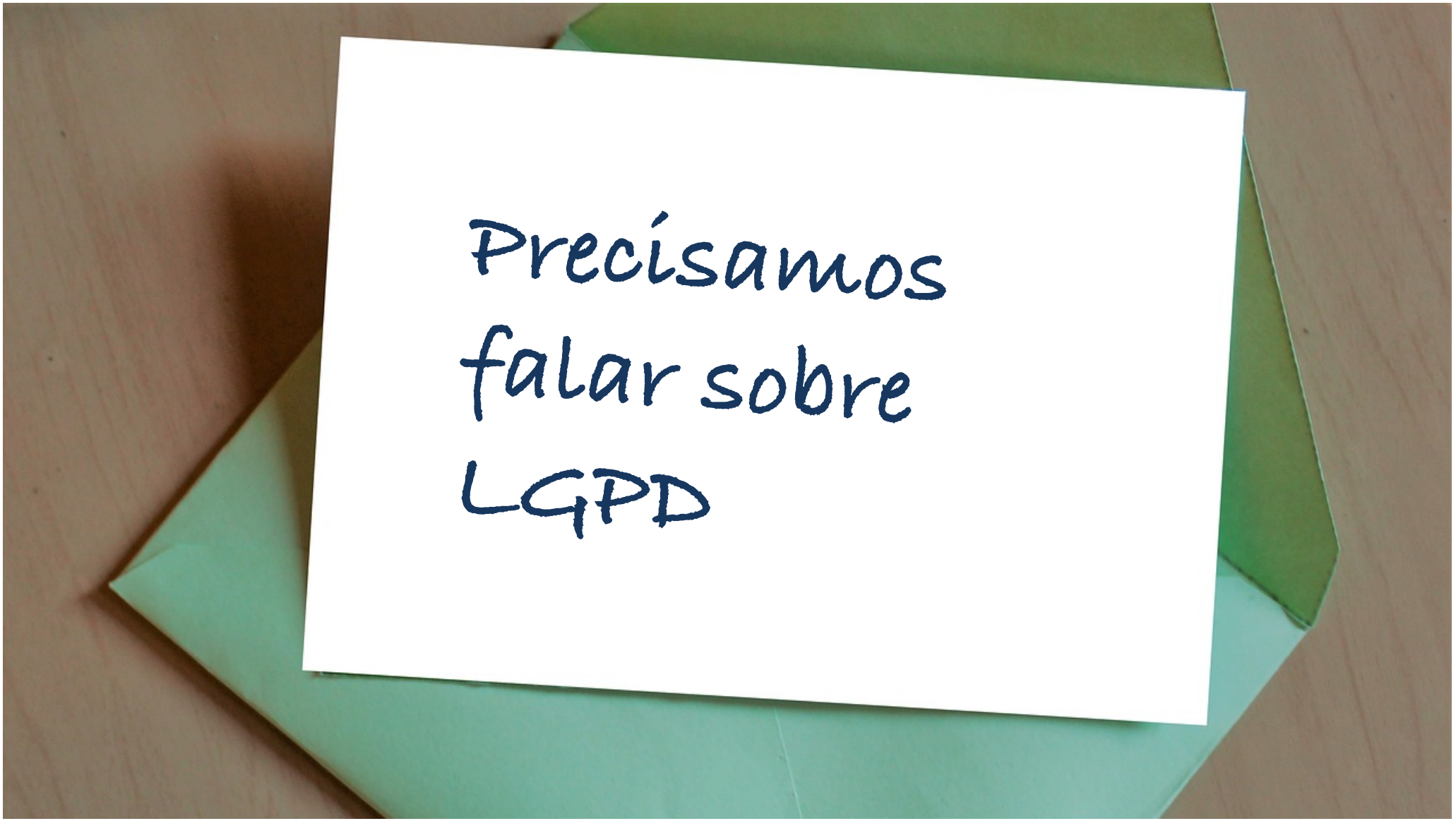

PROCURA-SE
CADASTRO CLIENTES
HISTÓRICO DE VENDAS
DADOS DE CLIMA
CLICKS EM CAMPANHAS
SCORE DE CRÉDITO
WEBSITE ANALYTICS
DADOS DE ESTOQUE
CEP


>catálogo
>de dados


Data Lake
T
Conta: TAIL
Cliente: 1
26 Datastores
Criar novo datastore
Buscar por nome
Recarregar
Visualização
CEPs Brasil
2019-08-29
Emma_Teste_Yp
2020-04-07
LOG_BAIRRO
2019-10-22
LOG_CPC
2019-10-22
LOG_FAIXA_BAIRRO
2019-11-20
LOG_FAIXA_CPC
2019-10-22
LOG_FAIXA_LOCALIDADE
2019-11-20
LOG_FAIXA_UF
2019-11-20
LOG_FAIXA_UOP
2019-10-22
LOG_GRANDE_USUARIO
2019-11-20
LOG_LOCALIDADE
2019-11-20
LOG_LOGRADOURO
2019-11-20
LOG_NUM_SEC
2019-10-22
LOG_UNID_OPER
2019-12-16
LOG_VAR_BAI
2019-10-22


Nome
CEPs Brasil (CEPsBrasil)
Somente Leitura?
Não
Diretório
1/dataReceptor/DataReceptorOfCEPsBrasil
Chave Primária
cep
Tipo de Persistência
BIG_DATA
Atualizado em
29/08/2019
Criado em
29/08/2019
Última escrita
29/08/2019
Esquema
Nome
Tipo
Anonimizada
Aumentada
cep
STRING
tipologradouro
STRING
logradouro
STRING
bairro
STRING
cidade
STRING
uf
STRING
Amostra(s)
Nome
Sample of CEPs Brasil *


> Amostras
val records =
spark.read.option("header","true")
.csv("example-database.csv").toDF
records.createOrReplaceTempView(”records")
val sample = spark.sql(
"SELECT * FROM records ”+
"TABLESAMPLE (25 PERCENT)"
)


Precisamos
falar sobre
LGPD


DADOS PII
BASES LEGAIS
DATA LINEAGE


(555) 555-1234
PII
Detector de Tipos de Dados
da2066c24702746781ff4

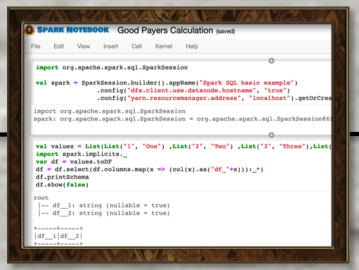

(555) 555-1234
PII
Detector de Tipos de Dados
da2066c24702746781ff4
Base Legal
(X) Consentimento
( ) Proteção ao crédito

DATA LINEAGE

CLIENTES

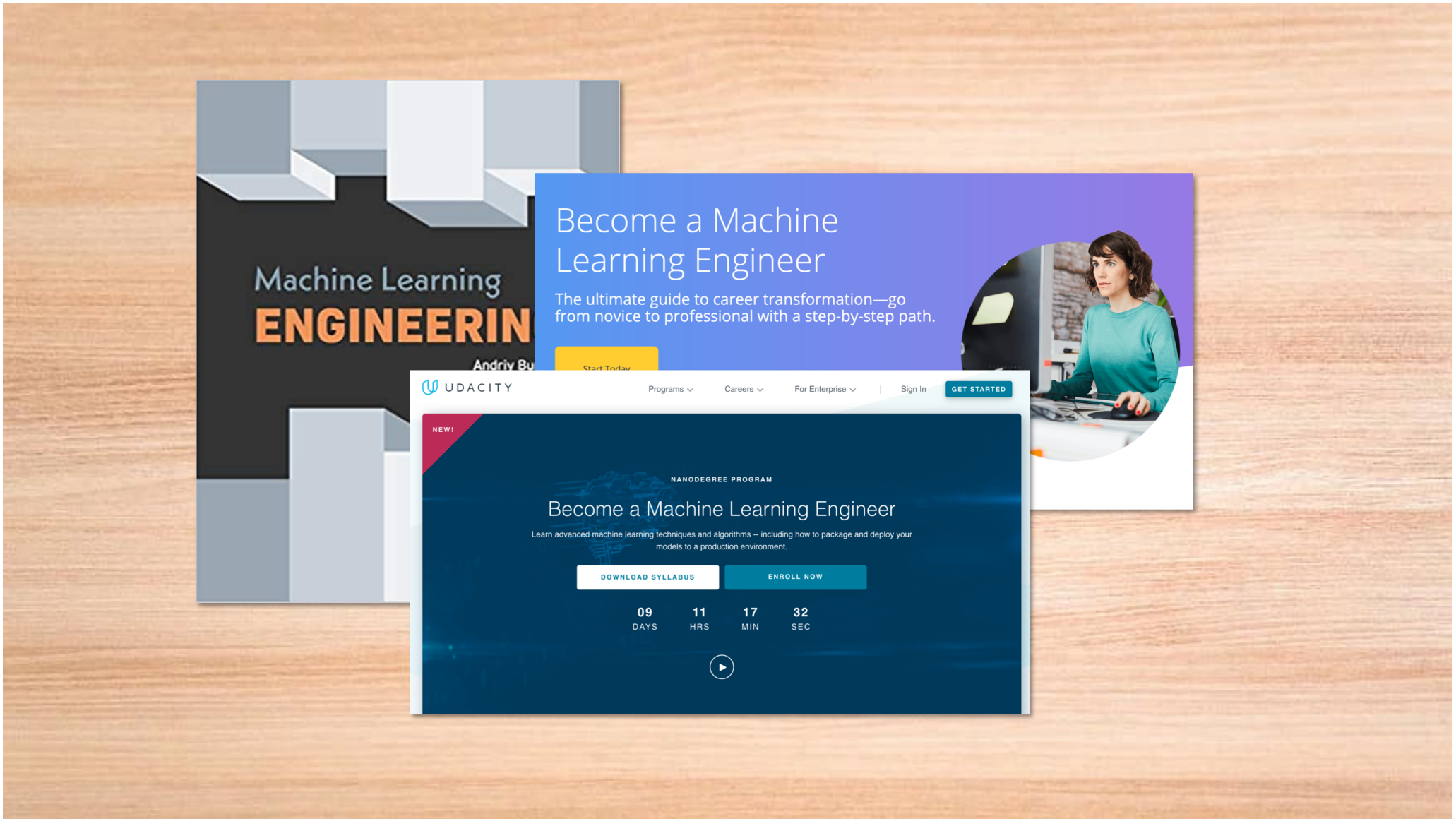
May, 5th, 2019

Scores De Crédito



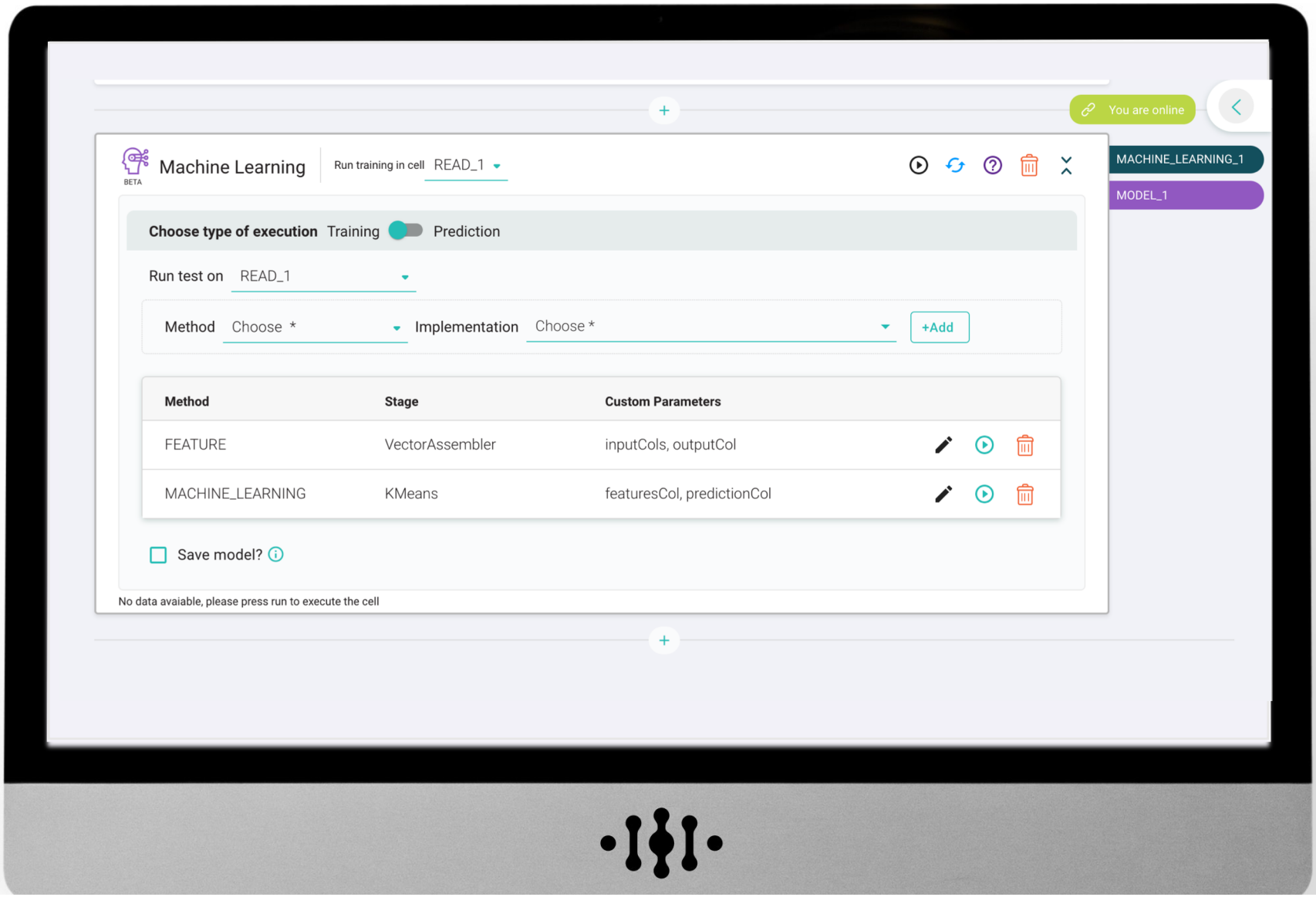
May, 6th, 2019

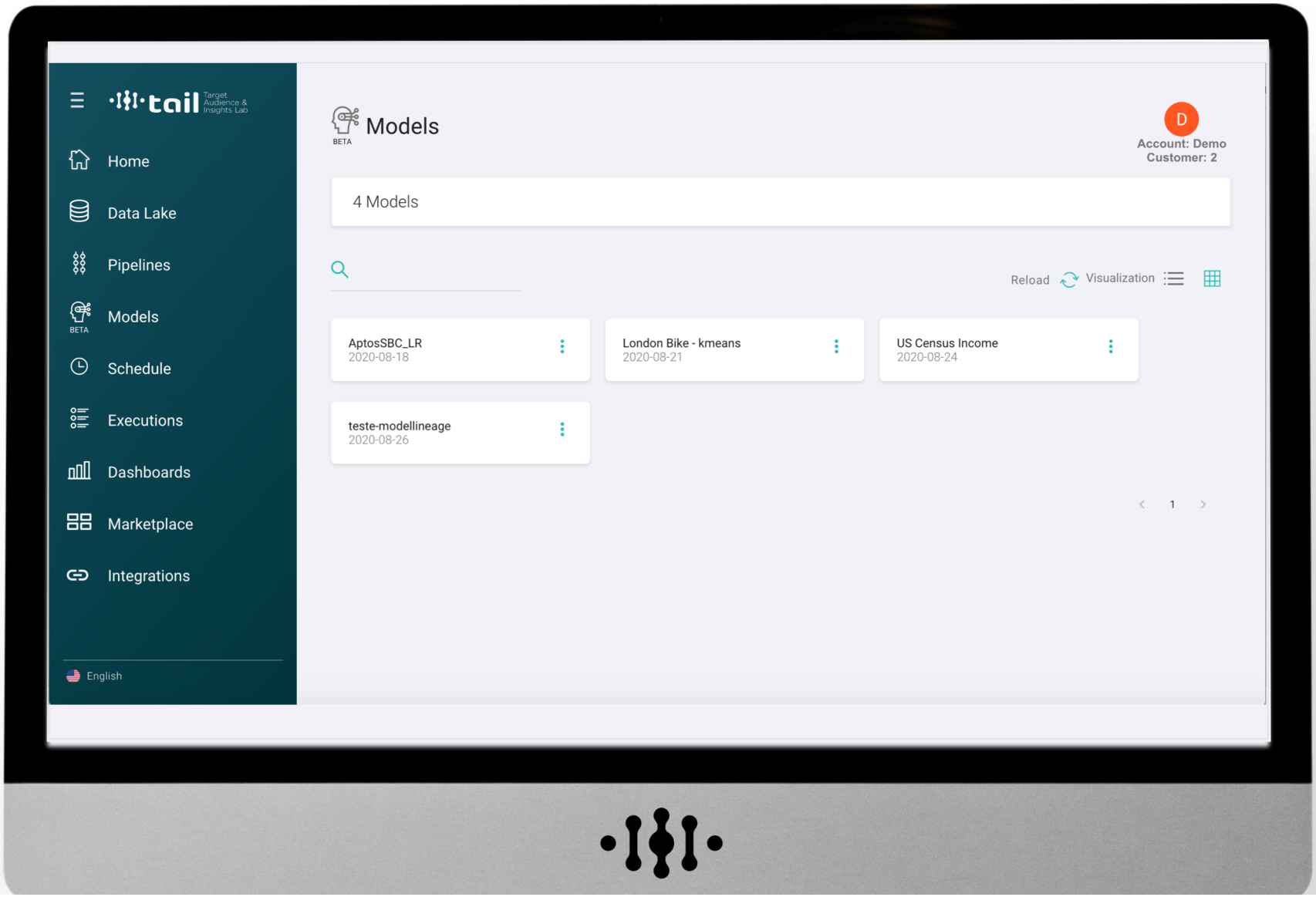
Bons Pagadores

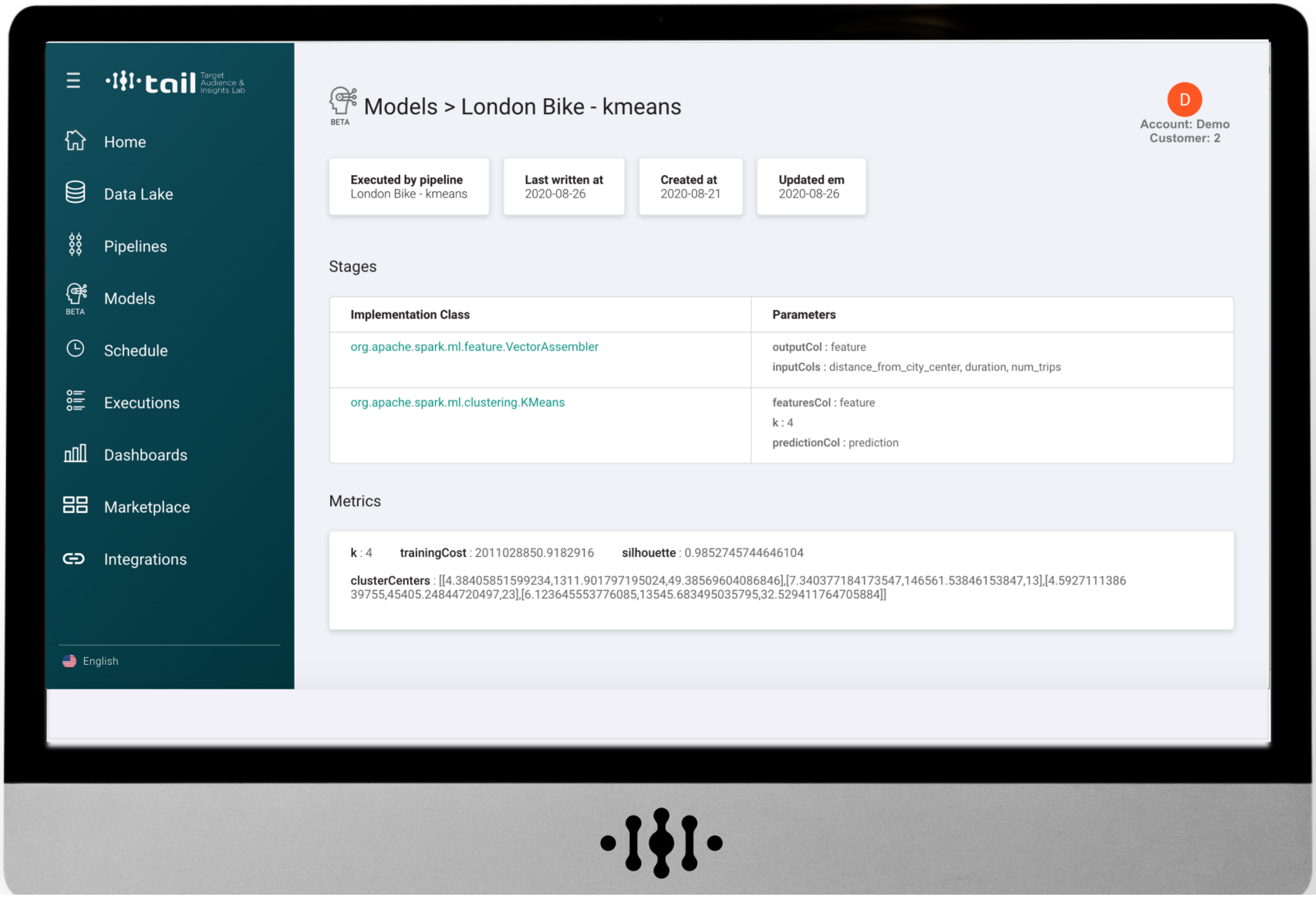

Machine Learning
ENGINEERIN
Andriy Bu
Become a Machine Learning Engineer
The ultimate guide to career transformation—go from novice to professional with a step-by-step path.
Start Today
UDACITY
Programs
Careers
For Enterprise
Sign In
GET STARTED
NEW!
NANODEGREE PROGRAM
Become a Machine Learning Engineer
Learn advanced machine learning techniques and algorithms -- including how to package and deploy your models to a production environment.
DOWNLOAD SYLLABUS
ENROLL NOW
09 DAYS
11 HRS
17 MIN
32 SEC

Engenharia de Machine Learning

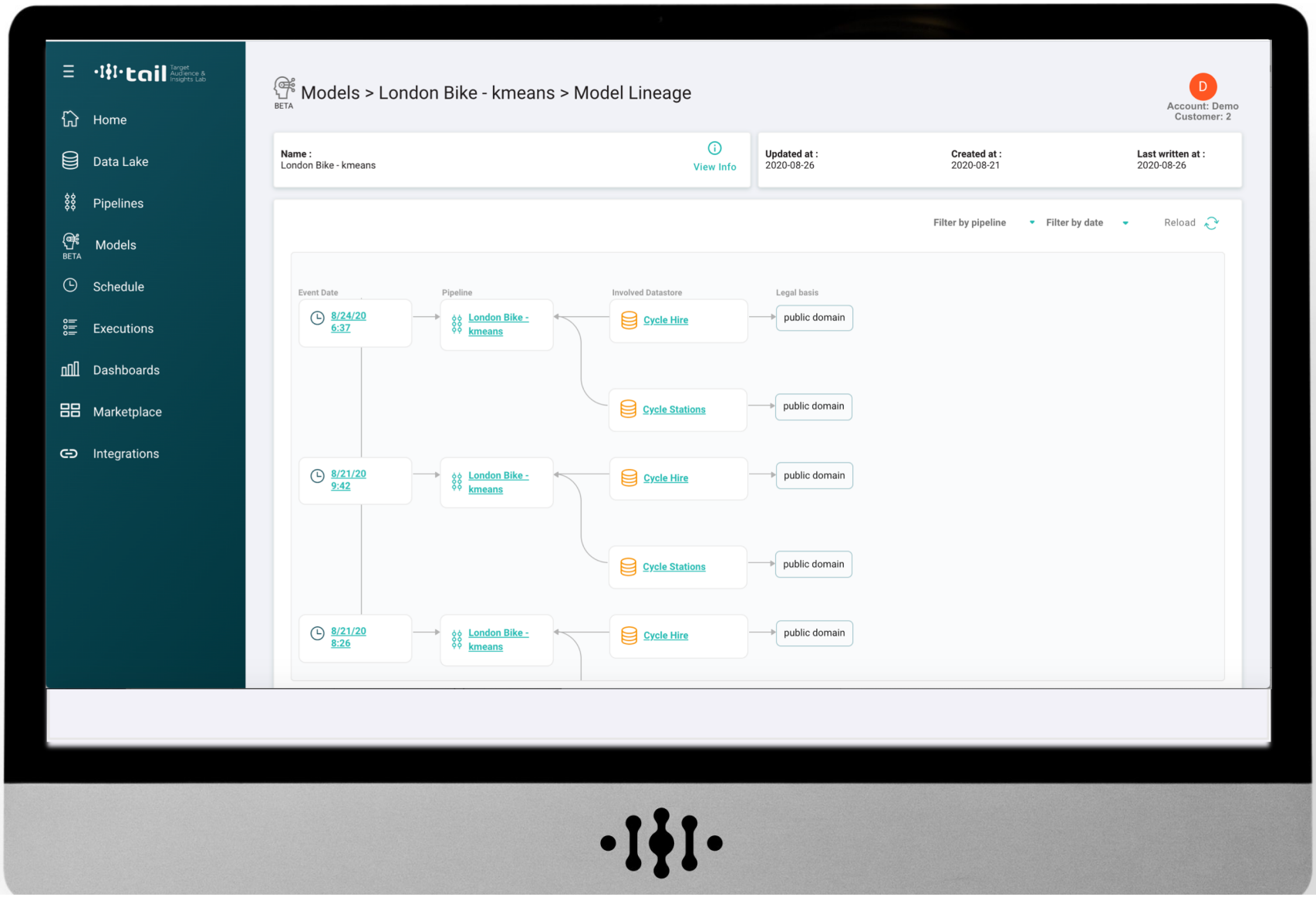
1. Implantar modelos em produção
2. Catálogo de modelos
3. Versionamento e rastreamento de modelos


You are online
Machine Learning
BETA
Run training in cell READ_1
MACHINE_LEARNING_1
MODEL_1
Choose type of execution Training Prediction
Run test on READ_1
Method Choose *
Implementation Choose *
+Add
Method
Stage
Custom Parameters
FEATURE
VectorAssembler
inputCols, outputCol
MACHINE_LEARNING
KMeans
featuresCol, predictionCol
Save model?
No data avaiable, please press run to execute the cell


tail
Target Audience & Insights Lab
Home
Data Lake
Pipelines
Models
BETA
Schedule
Executions
Dashboards
Marketplace
Integrations
English
Models
BETA
D
Account: Demo
Customer: 2
4 Models
Reload
Visualization
AptosSBC_LR
2020-08-18
London Bike - kmeans
2020-08-21
US Census Income
2020-08-24
teste-modellineage
2020-08-26
1


tail
Target Audience & Insights Lab
Home
Data Lake
Pipelines
Models
BETA
Schedule
Executions
Dashboards
Marketplace
Integrations
English
Models > London Bike - kmeans
BETA
D
Account: Demo
Customer: 2
Executed by pipeline
London Bike - kmeans
Last written at
2020-08-26
Created at
2020-08-21
Updated em
2020-08-26
Stages
Implementation Class
Parameters
org.apache.spark.ml.feature.VectorAssembler
outputCol : feature
inputCols : distance_from_city_center, duration, num_trips
org.apache.spark.ml.clustering.KMeans
featuresCol : feature
k : 4
predictionCol : prediction
Metrics
k : 4
trainingCost : 2011028850.9182916
silhouette : 0.9852745744646104
clusterCenters : [[4.38405851599234,1311.901797195024,49.38569604086846],[7.3403771841735 47,146561.53846153847,13],[4.592711138639755,45405.24844720497,23],[6.123645553776085,13545.683495035795,32.529411764705884]]


Precisamos
falar sobre
Model
Lineage


tail
Target Audience & Insights Lab
Home
Data Lake
Pipelines
Models
BETA
Schedule
Executions
Dashboards
Marketplace
Integrations
BETA
Models > London Bike - kmeans > Model Lineage
D
Account: Demo
Customer: 2
Name :
London Bike - kmeans
View Info
Updated at :
2020-08-26
Created at :
2020-08-21
Last written at :
2020-08-26
Filter by pipeline
Filter by date
Reload
Event Date
Pipeline
Involved Datastore
Legal basis
8/24/20 6:37
London Bike - kmeans
Cycle Hire
public domain
Cycle Stations
public domain
8/21/20 9:42
London Bike - kmeans
Cycle Hire
public domain
Cycle Stations
public domain
8/21/20 8:26
London Bike - kmeans
Cycle Hire
public domain

Engenharia de Dados e Engenharia de Machine Learning

como fazer projetos de Data Science usando um processo sustentável

Fabiane Bizinella Nardon
Chief Data Scientist
@fabianenardon

Recursos:
Frame photo created by rawpixel.com - www.freepik.com
Photo by Shirly Niv Marton on Unsplash
Photo by Tim Gouw on Unsplash
Image by https://pixabay.com/users/www_darkworkx_de-2044000
Image by TeroVesalainen from Pixabay
Image by MR1313 from Pixabay