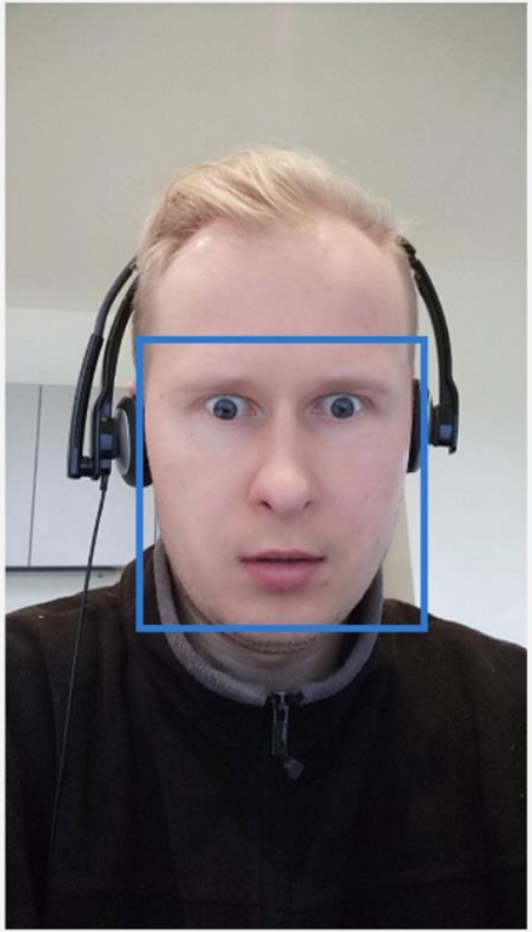
Lessons Learned on Building a Data Science Environment on a Public Cloud

valdas@maksimavicius.eu

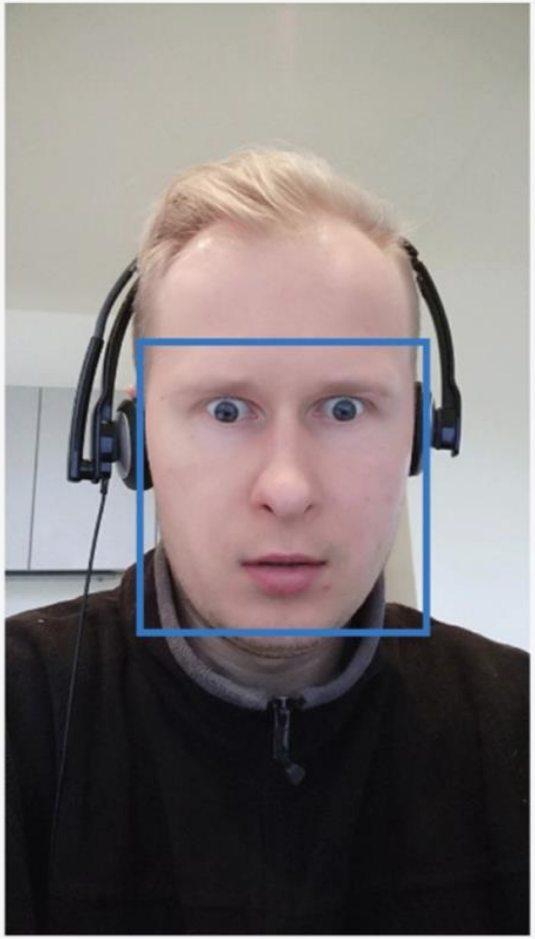
Speaker – Valdas Maksimavičius

- Big Data Architect at Cognizant
- Insurance, Pensions & Manufacturing industries
- Vilnius Microsoft Data Platform Meetup / Hack4Vilnius Hackathon
- Valdas.blog



Detection result:
JSON:


```
[
  {
    "faceId": "f783a705-c1c9-4cf1-bb24-064f951f4e52",
    "faceRectangle": {
      "top": 415,
      "left": 163,
      "width": 366,
      "height": 366
    },
    "faceAttributes": {
      "hair": {
        "bald": 0.13,
        "invisible": false,
        "hairColor": [
          {
            "color": "brown",
            "confidence": 0.91
          },
          {
            "color": "red",
            "confidence": 0.9
          },
          {
            "color": "blond",
            "confidence": 0.58
```

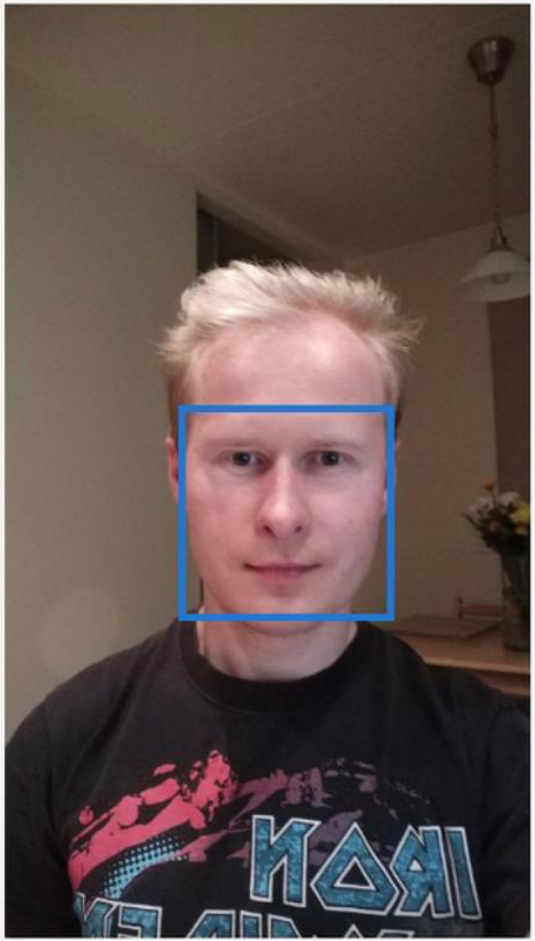


Detection result:

JSON:


```
[
  {
    "faceId": "f783a705-c1c9-4cf1-bb24-064f951f4e52",
    "faceRectangle": {
      "top": 415,
      "left": 163,
      "width": 366,
      "height": 366
    },
    "faceAttributes": {
      "hair": {
        "bald": 0.13,
        "invisible": false,
        "hairColor": [
          {
            "color": "brown",
            "confidence": 0.91
          },
          {
            "color": "red",
            "confidence": 0.9
          },
          {
            "color": "blond",
            "confidence": 0.58
```


"bald": 0.13



Detection result:

JSON:


```
[
 {
   "faceId": "588778e2-bf43-48a2-9dbd-181fb26aaf41",
   "faceRectangle": {
     "top": 1644,
     "left": 704,
     "width": 883,
     "height": 883
   },
   "faceAttributes": {
     "hair": {
       "bald": 0.17,
       "invisible": false,
       "hairColor": [
         {
           "color": "blond",
           "confidence": 0.98
         },
         {
           "color": "brown",
           "confidence": 0.73
         },
         {
           "color": "gray",
           "confidence": 0.52
```


"bald": 0.17

Let’s start with industry trends (Insurance)

INDUSTRY THEMES

- Usage-based and Pay as you go insurance
- Individual risk profiles and accurate pricing
- No paper work and manual processes
- Personalized loss exposure interventions
- Discounts to customers who link smart-devices
- Digital Claims and no touch resolutions

MEANS TO THE END

- Bots & digital advisors
- Hyper personalized Products & Comm.
- Advanced Analytics & Proactiveness
- InsurTech
- Connected Products & devices
- AI & Automation

REQUIRED CAPABILITIES

- Ingest all sources easily
- Explore data freely
- Act on customer event in real time
- Use complex algorithms in channels at scale

Select relevant end-to-end use cases

Retail

CONSUMER ENGAGEMENT

Real-time Pricing Optimization

Financial

RISK AND REVENUE MANAGEMENT

Risk and Fraud, Threat Detection

Oil/Gas & Energy

GRID OPS, ASSET OPTIMIZATION

Industrial IoT

Security

ACTIONABLE THREAT INTELLIGENCE

Security Intelligence

Healthcare

SENSOR DATA

IoT DEVICE ANALYTICS

Advertising

RECOMMENDATION ENGINE

Next Best and Personalized Offers

Media Entertainment

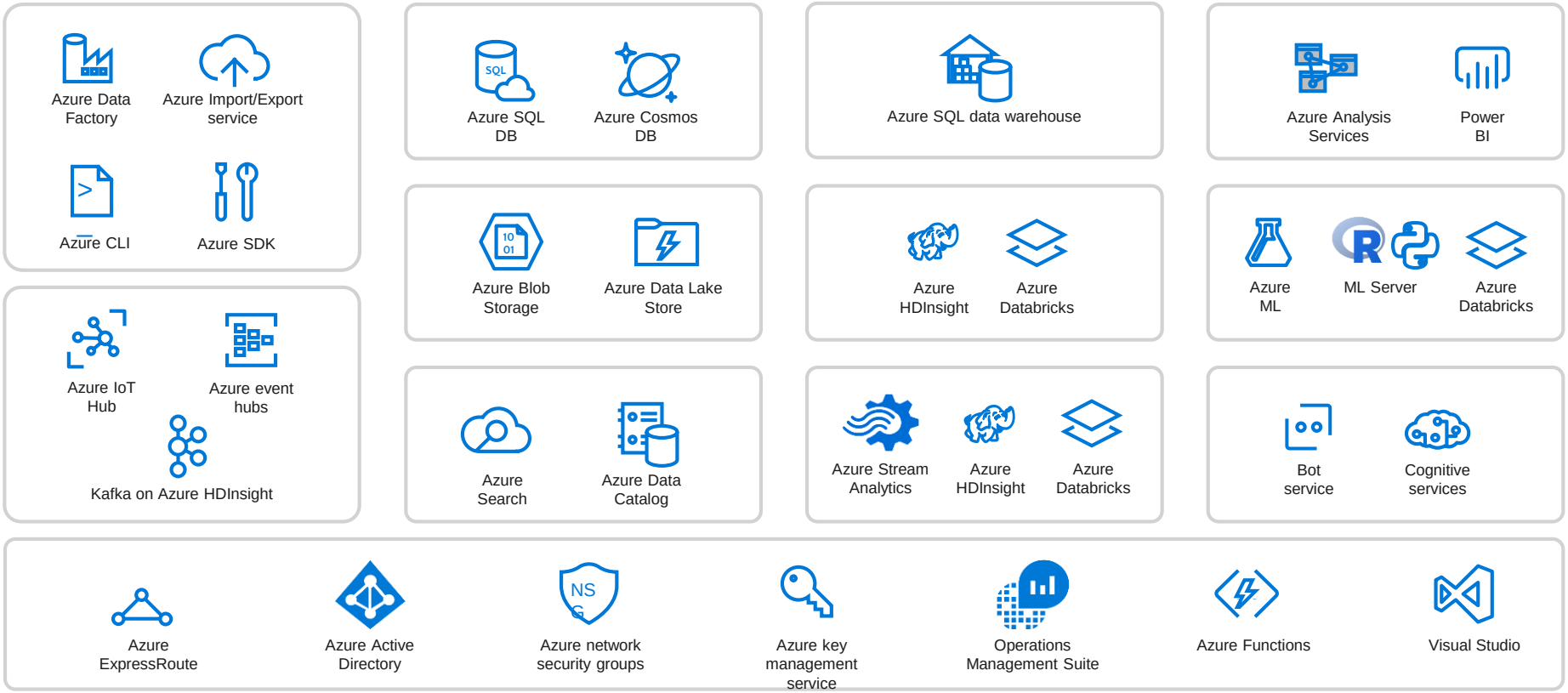
CONSUMER ENGAGEMENT ANALYSIS

Sentiment Analysis

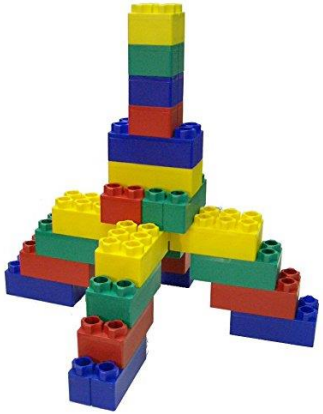
Perform gap analysis

Capabilities / Business use cases	Data pipelines	Real-time data pipelines	Data exploration (adhoc)	Data processing (production)	Machine learning	Deep learning	Serving	Master Data Management	Data governance	CI/CD	Data quality	Security
	Centralised datastore, Feature marts, batch ingestion, storage, schema discovery	Streaming, event handling	Data blending, analyst self-service, ad-hoc queries	ETL, enrichment, normalisation, in-memory processing	Choose & pick library, retraining, performance, A/B testing	NLP, image recognition, voice to text	Model serving, analytical score serving	Data integration, unification, validation	Metadata, GDPR, Data catalogs, data lineage, data collaboration	Lifecycle and devops, code and configuration repositories	Metrics, monitoring, alerting	Authentication & authorization, firewalls, VNets, etc.
Use case 1												
Use case 2												
Use case 3												
Use case 4												

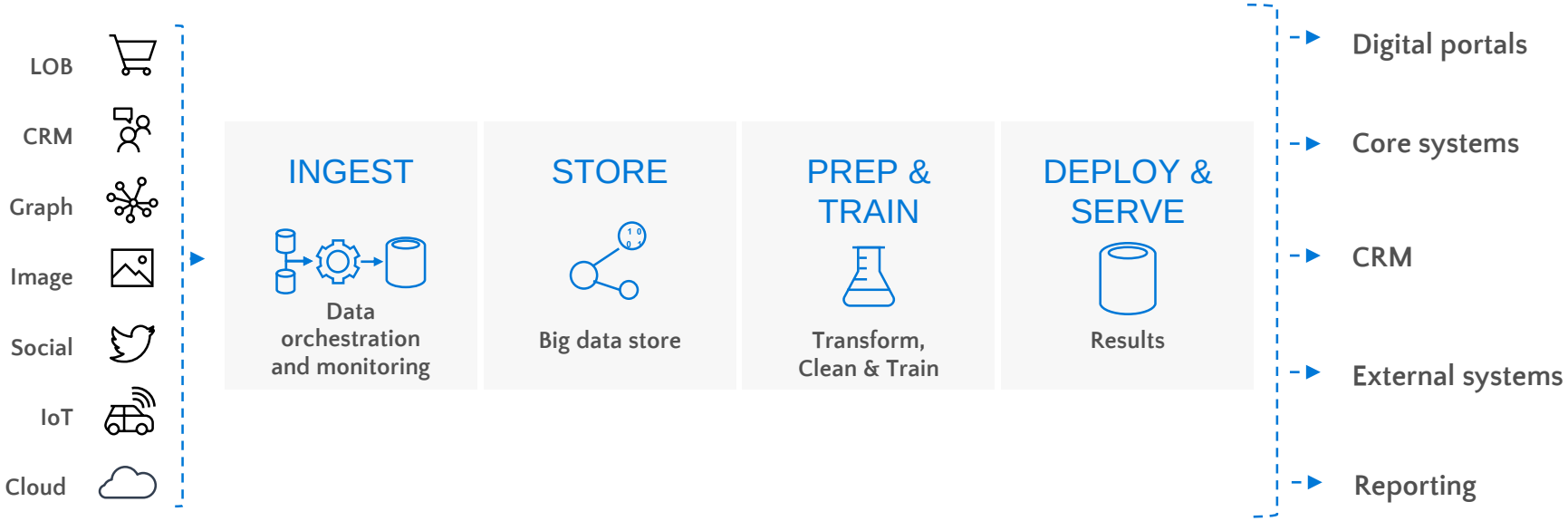
The Azure data landscape



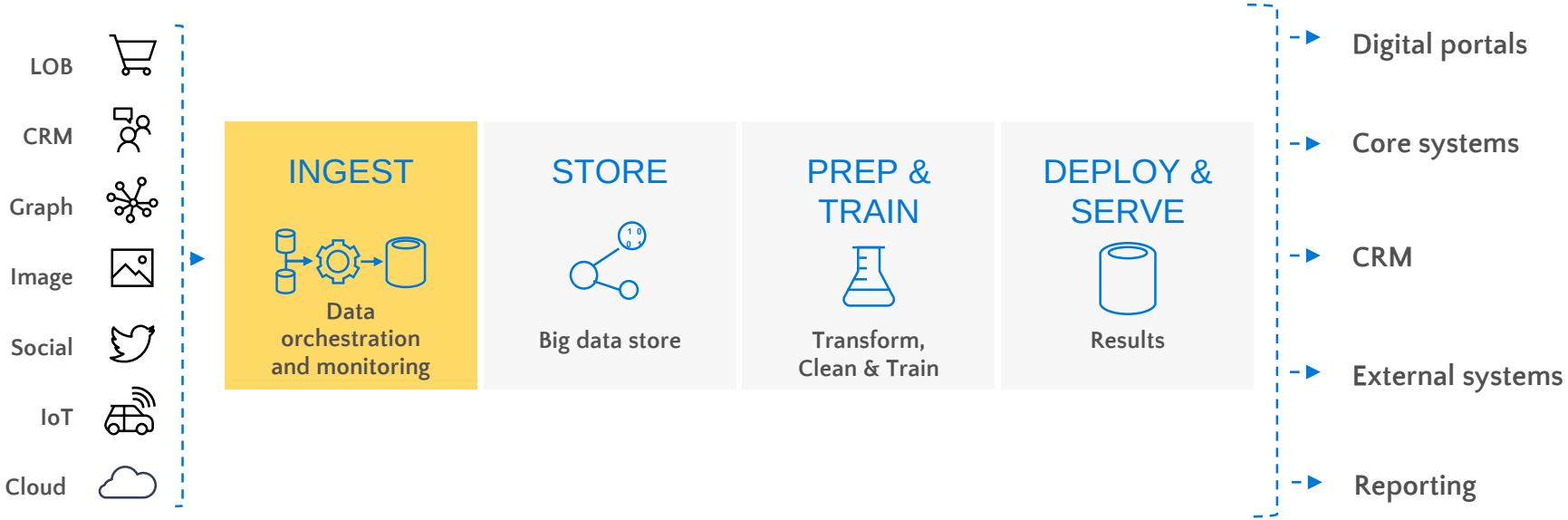
Our job, pretty much



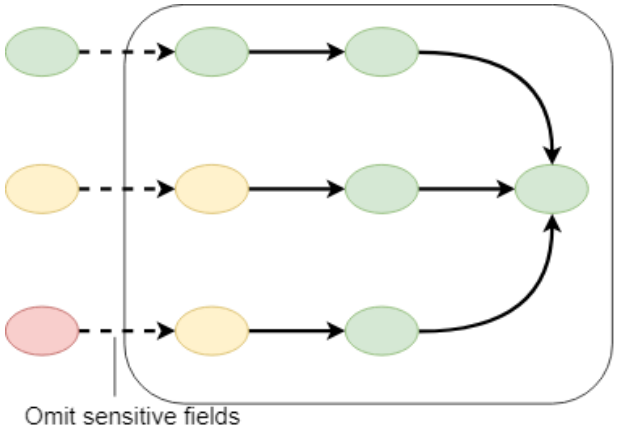
Getting things done

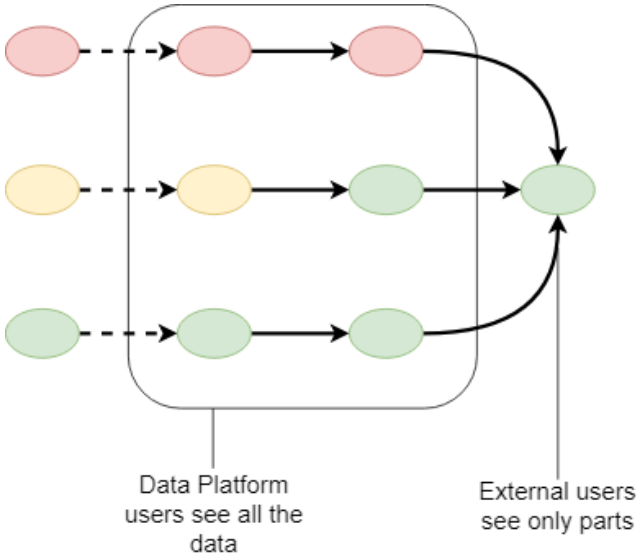


Ingest data from source systems to a data lake

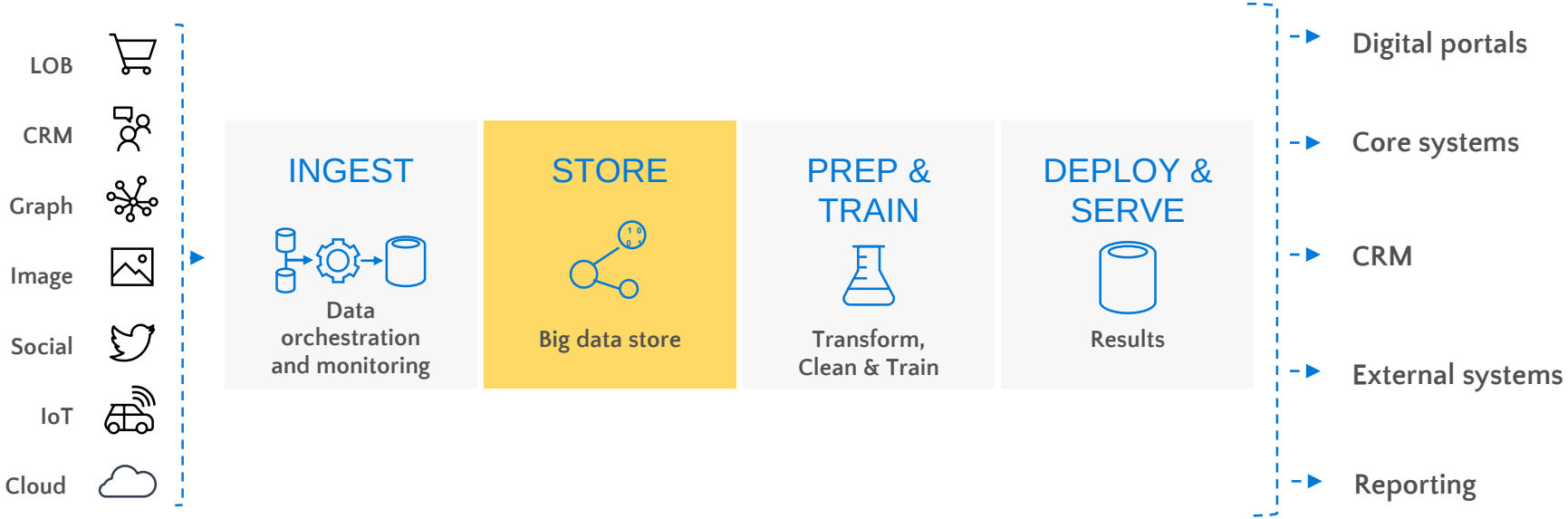


Lesson 1: Build privacy protection patterns





Store data in its native format



Lesson 2: Use cloud storage offerings instead of Hadoop

Machine Learning & Big Data Blog

Is Hadoop Dead? How Kubernetes and Cloud-Native Could Displace Hadoop

The Death of Hadoop?

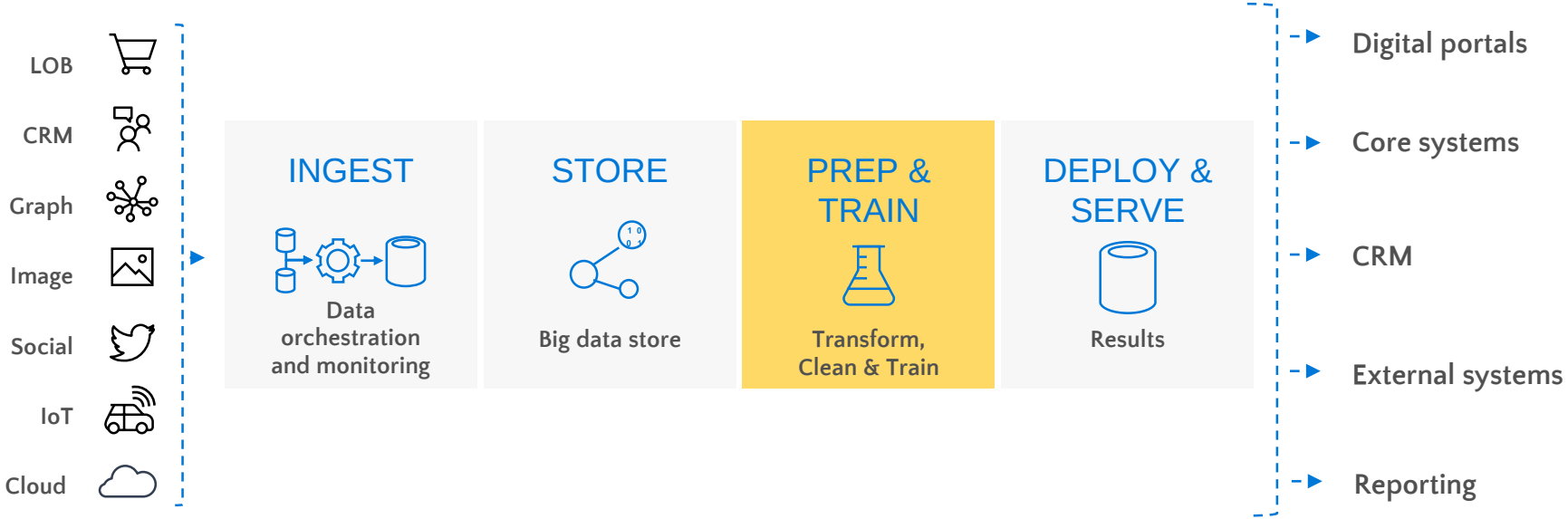
Is Hadoop dead? Not so fast. Plan on supporting multiple environments for some time to come.

By Barry Devlin

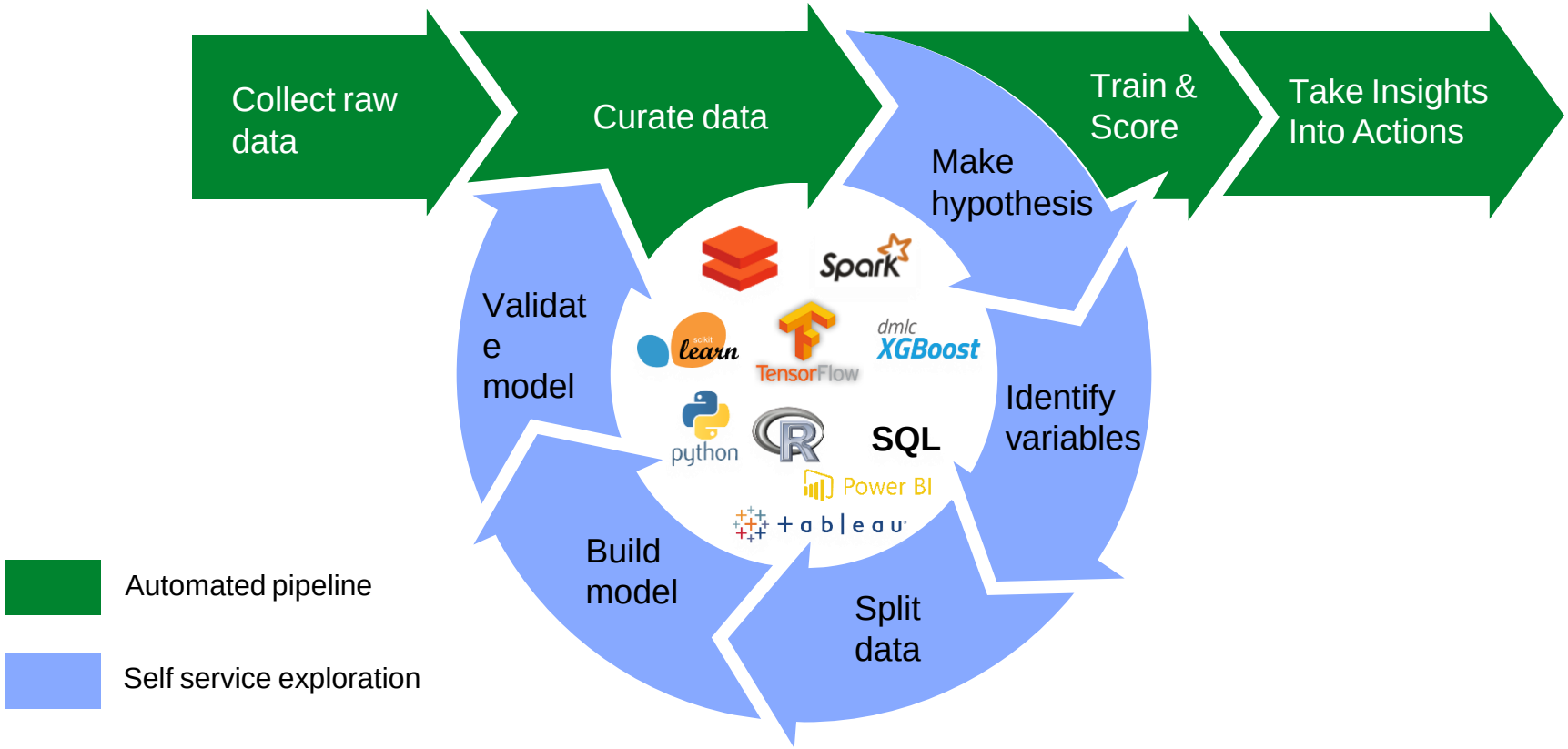
February 26, 2019

The recent "merger of equals" between Cloudera and Hortonworks has triggered speculation about the possible imminent demise of Hadoop. Market observers question if the merger indicates a shrinking Hadoop ecosystem market that can no longer support its two largest competing beasts.

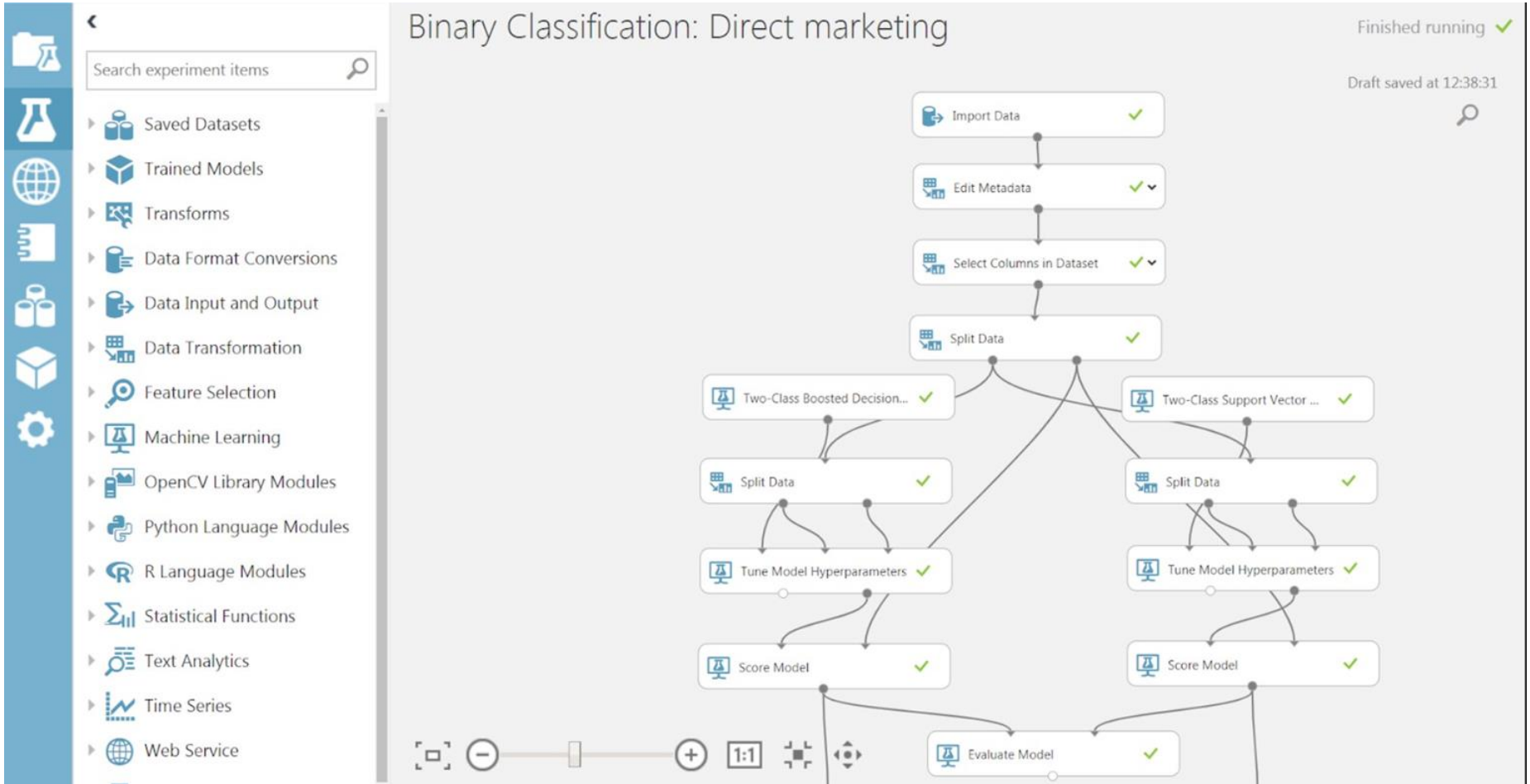
Provide tools and resources to prep & train



Lesson 3: Offer self-service tools



Azure Machine Learning Studio



Azure Databricks

Azure Databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

CarrierDelay (Python)

Attached: thecluster | File | View: Code | Permissions | Stop Execution | Clear | Schedule | Comments | Runs | Revision history

```python
from pyspark.ml.feature import OneHotEncoderEstimator, StringIndexer
from pyspark.ml.feature import VectorAssembler
from pyspark.ml import Pipeline
from pyspark.ml.regression import LinearRegression
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.tuning import CrossValidator, ParamGridBuilder

# Delete (if exists) parquet folder
dbutils.fs.rm("/mnt/parquet/carrierdelay", True)

carrierDF = (spark.read.format('csv')
            .options(header='true', inferschema='true')
            .load("/mnt/flightdelay3/*.csv"))

carrierDF = (carrierDF.withColumn("QUARTER", carrierDF["QUARTER"].cast("double"))
            .withColumn("MONTH", carrierDF["MONTH"].cast("double"))
            .withColumn("DAY_OF_MONTH", carrierDF["DAY_OF_MONTH"].cast("double"))
            .withColumn("DAY_OF_WEEK", carrierDF["DAY_OF_WEEK"].cast("double"))
            .withColumn("DISTANCE_GROUP", carrierDF["DISTANCE_GROUP"].cast("double")))

carrierDF = carrierDF.drop('DEST_AIRPORT_ID', '_c17', 'YEAR', 'DEP_DELAY').na.drop()

# carrierDF.printSchema()

carrierDF.write.parquet("/mnt/parquet/carrierdelay")

# carrierDF = spark.read.format('parquet').options(header='true', inferschema='true').load("/mnt/parquet/carrierdelay")
print("Number of carrier flights in the database: ", carrierDF.count())
```

Lesson 4: Use on-demand resources

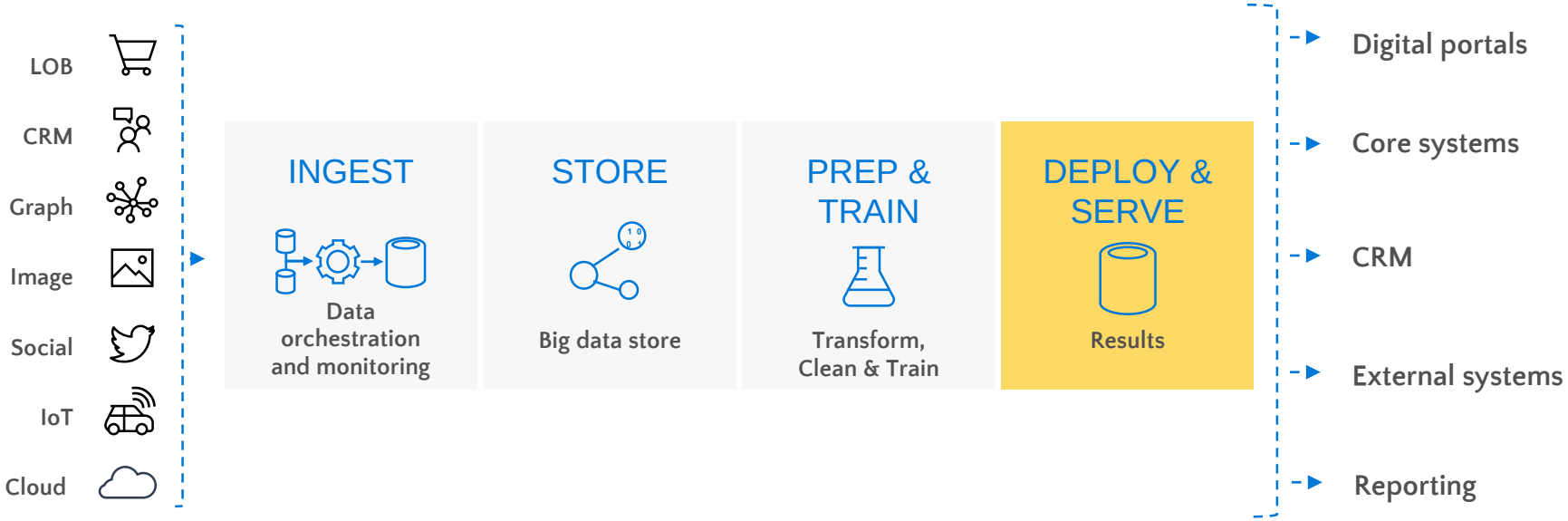
Storage

10 TB
+ 24/7
= ~300 Eur/month

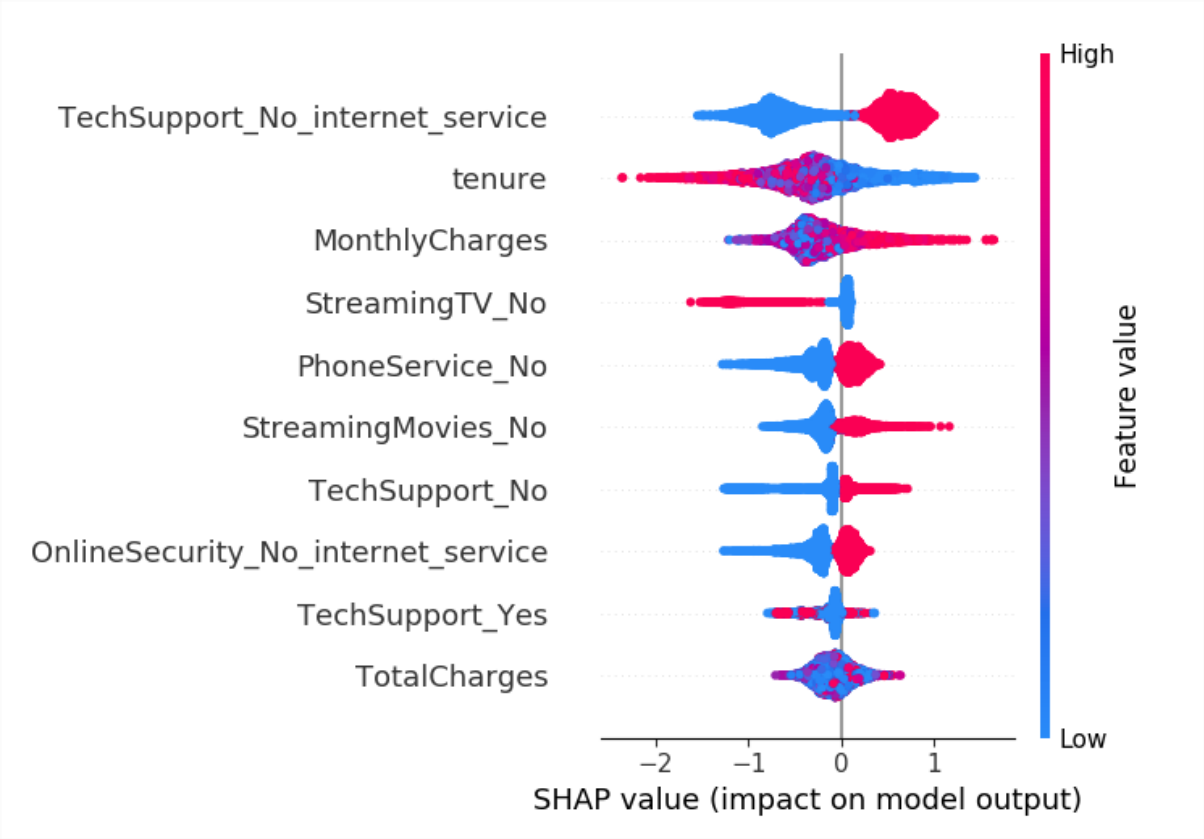
databricks

32 cores / 128 GB RAM
+ 160 hours
= ~700 Eur/month

Serve results to end consumers



Lesson 5: Explain your models to business users (e.g. SHAP)



Key Takeaways

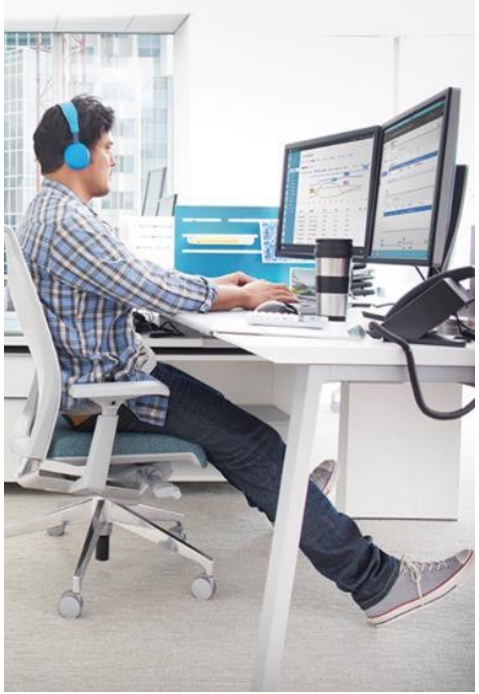
Lesson 1 Build privacy protection patterns

Lesson 2 Use cloud storage instead of Hadoop

Lesson 3 Offer self-service tools

Lesson 4 Use on-demand resources

Lesson 5 Explain ML models to business



Questions?

Valdas Maksimavičius

valdas@maksimavicius.eu