

Lecture:
Herminio Vazquez
Prof. PhD. Riccardo Tommasini

Master Class:
Data Management

Data Journey

Data Journey Master Class
Tartu, Estonia / November 19, 2020



UNIVERSITY OF TARTU



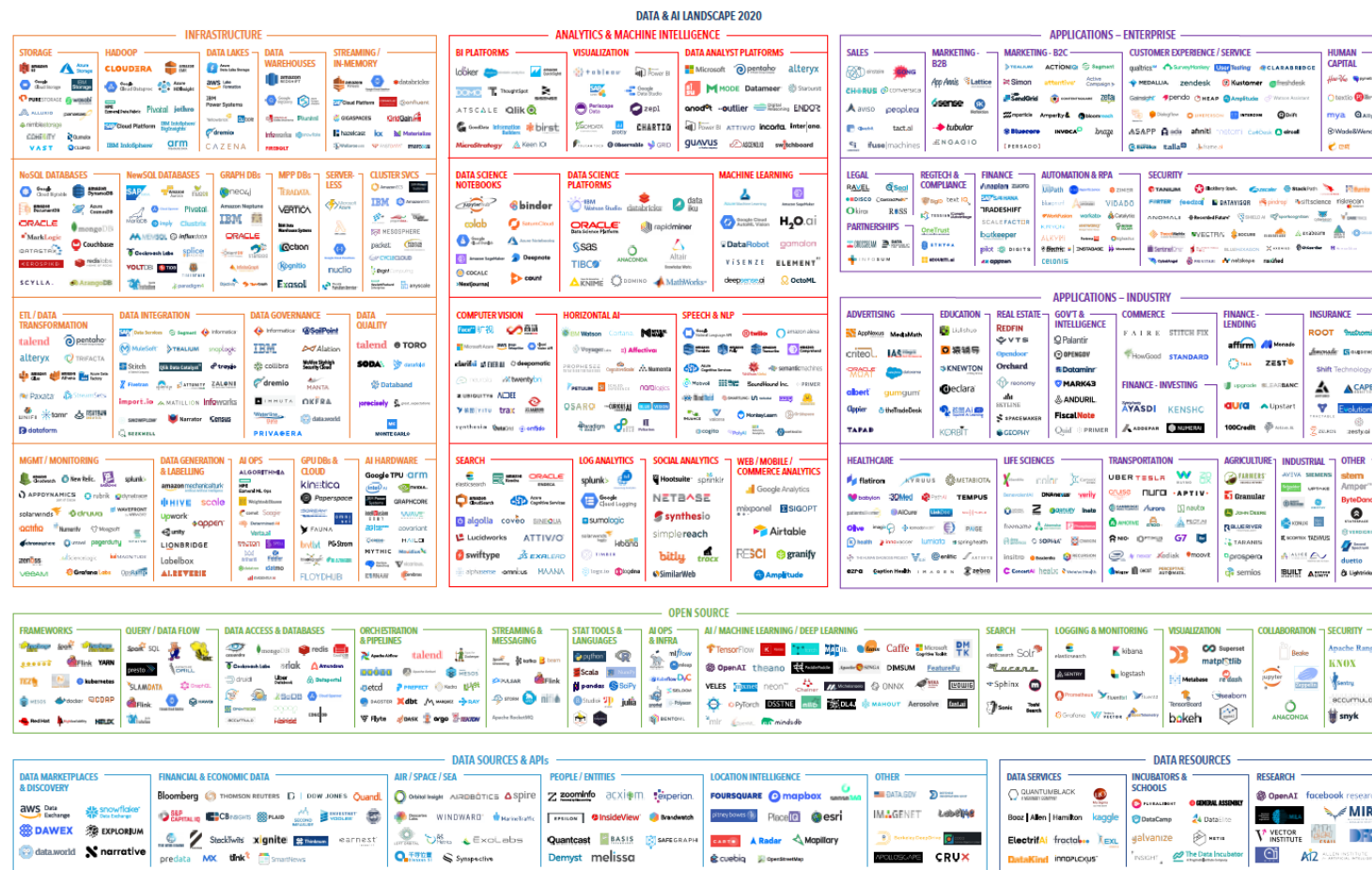
Agenda

1. The Data Journey
2. Data Ecosystem and Profession
3. Building Data Pipelines
4. Decision Engineering

Data Journey



Ecosystem



Version 1.0 - September 2020

© Matt Turck (@mattturck) & FirstMark (@firstmarkcap)

mattturck.com/data2020

<http://shorturl.at/eyBP1>

FIRSTMARK
EARLY STAGE VENTURE CAPITAL



Profession

Senior Data Engineer Big Data Architect Researcher

Data Analyst Data Engineer Data Modeller

Data Scientist Data Consultant Data Architect

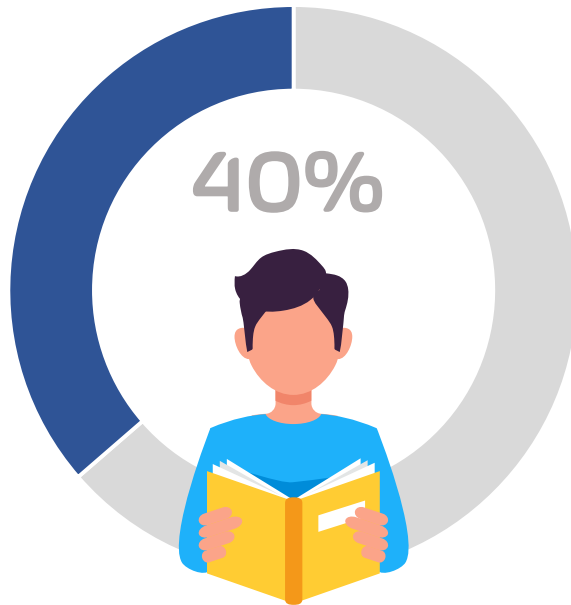
Data Steward Chief Data Officer Database Administrator

Data Product Manager ML-Ops Engineer Data Wrangler

Machine Learning Expert Data and Analytics Manager Quantum Developer

Data Designer VP of Data Governance and Jupyter Lab

Profession



Data Scientist
F1-Driver

Scientific method on data

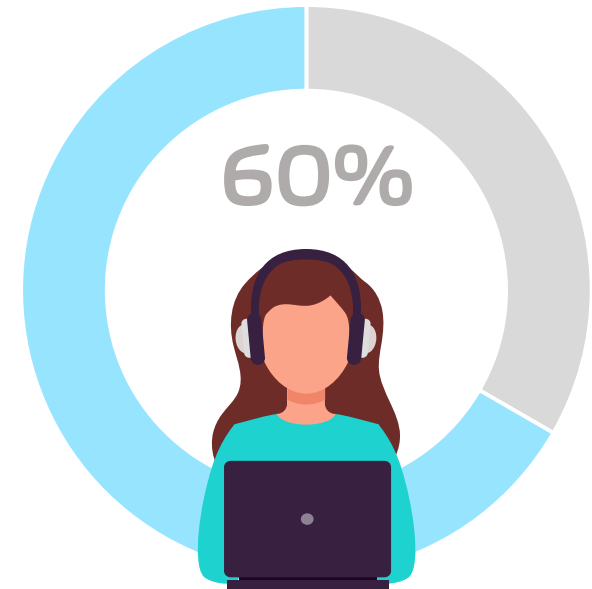


Build new knowledge



**Full Stack
Data Scientist**
F1-Team Captain

Data Engineer
F1-Mechanic



Computing methods on data



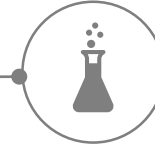
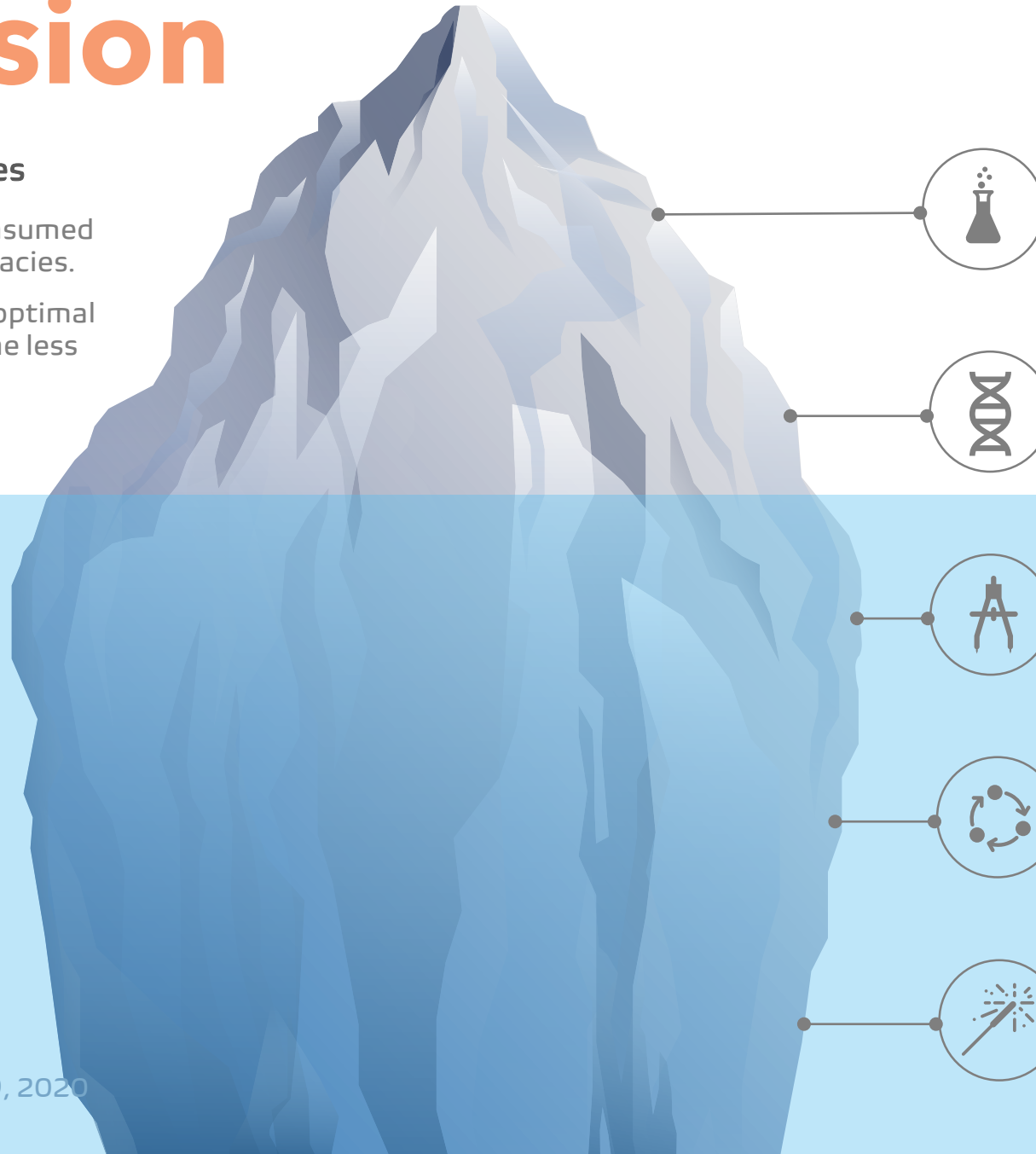
Build data pipelines

Profession

Beyond Models and Pipes

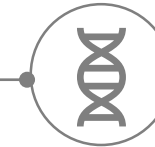
Job is done, until data is consumed and interpreted without fallacies.

Reliably, consistently, with optimal code, lower energy and in the less possible time.



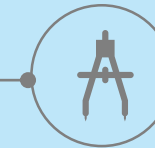
Data Science

- Mathematical representation
- Build models
- Hyper-parameter tuning
- Decision Engineering



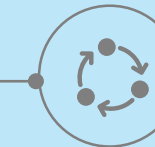
Data Engineering

- Efficient storage
- Query planning optimization
- Compute paradigms
- Idempotent jobs



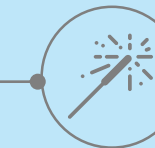
Data Architecture

- Locality
- Scalability
- Hybrid workloads
- Cost effectiveness



Data Operations

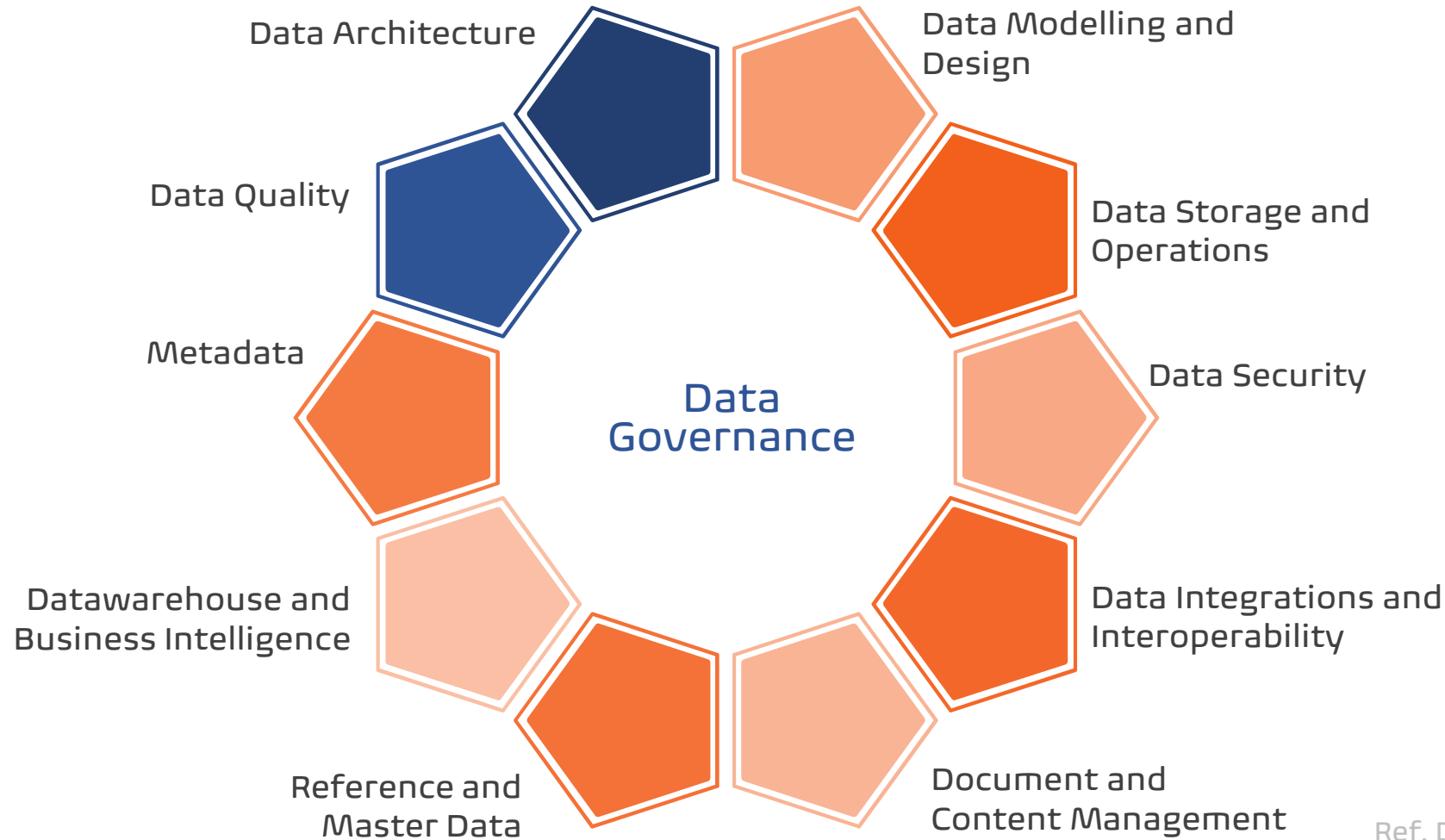
- Observability
- Monitoring and Metrics
- Resilience
- Governance



Data Visualization

- Encoding data in shape, position, color and size
- Appealing to the eye
- Cognitive automation

Functions



Ref. DATA-DMBOK 2nd Edition



Building Pipelines

Strategy

Data Format

Characters Encodings

Ingestion Ratio

Batch or NRT

Computing Framework

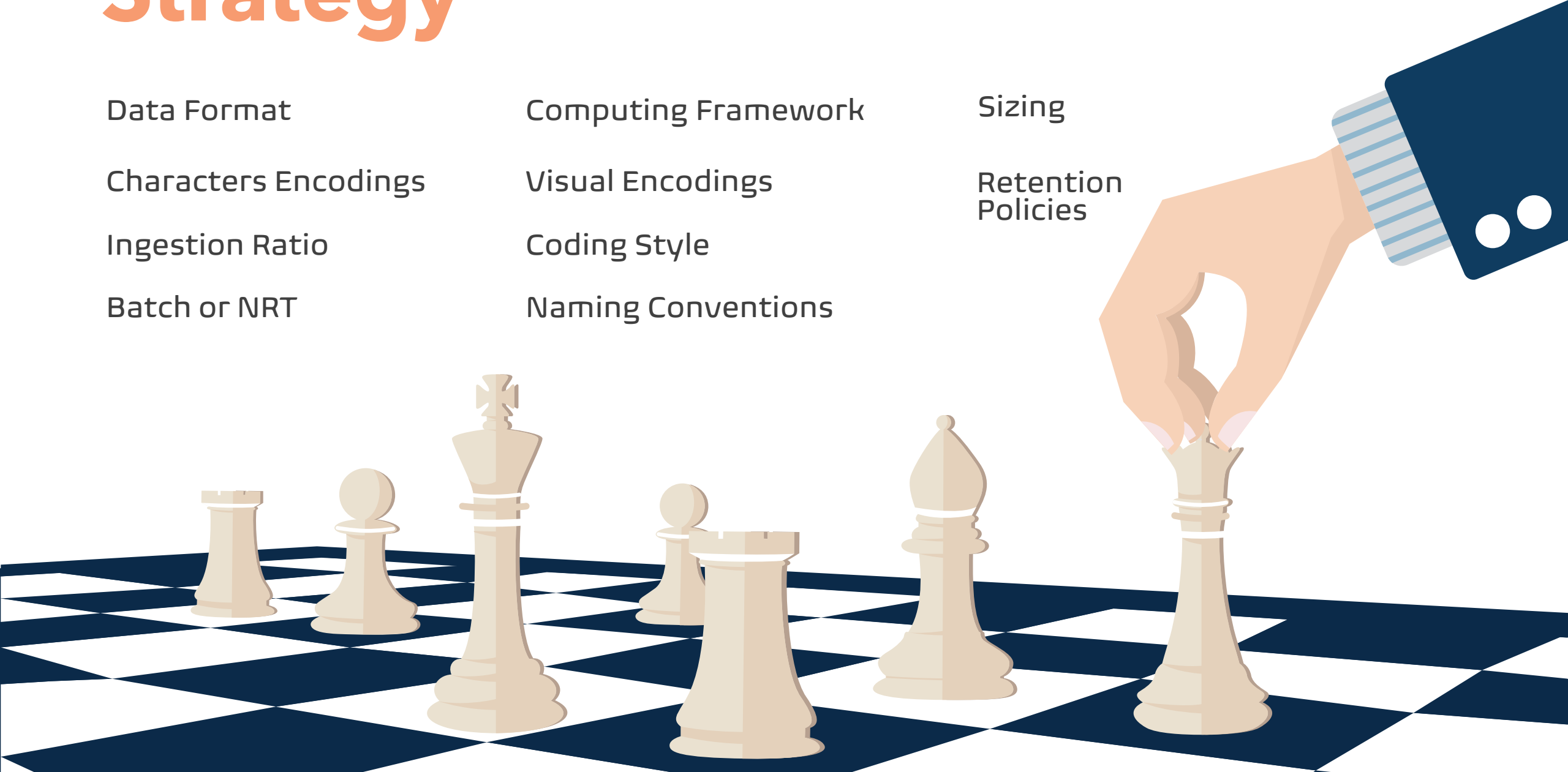
Visual Encodings

Coding Style

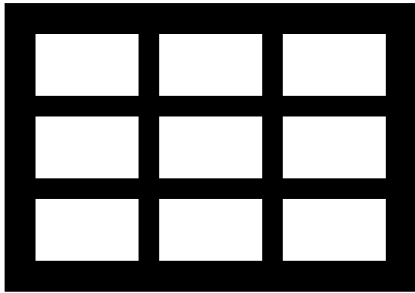
Naming Conventions

Sizing

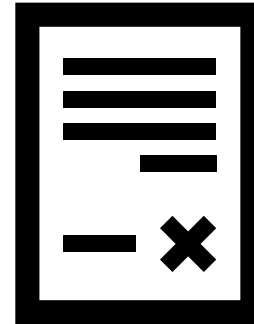
Retention Policies



Unit of Data



Query
For Structured Data



Path
For Non-Structured Data

Is Fun

Especially when you wait for more
than 5 minutes for a job

But the feeling of that 200-column
killer job ending in 1.2 seconds....

Or to find out that your functions
are non-deterministic and you
have random results

Or when Batch and NRT reconcile



Awareness

```
pyspark.sql.functions.collect_set(col)
```

Aggregate function: returns a set of objects with duplicate elements eliminated.

Note: The function is non-deterministic because the order of collected results depends on the order of the rows which may be non-deterministic after a shuffle.

```
110 # UDF: Document status over time
111 @F.udf(T.ArrayType(T.LongType()))
112 def state(m):
113     d = {}
114     for k,v in m.items():
115         d[k] = v
116     return list(d.values())
```

Order in windows matter...

```
@pytest.fixture
def spark(scope='module'):
    spark = SparkSession.builder.appName("datastore-milestones-test")\
        .config("spark.sql.mapKeyDedupPolicy", "LAST_WIN")\
        .config('spark.sql.adaptive.enabled', 'true')\
        .master('local[*]').getOrCreate()
    yield spark
```

Stay fresh...

Ref. PySpark Documentation

Advise

Make it Work

Make it Well

Make it Fast

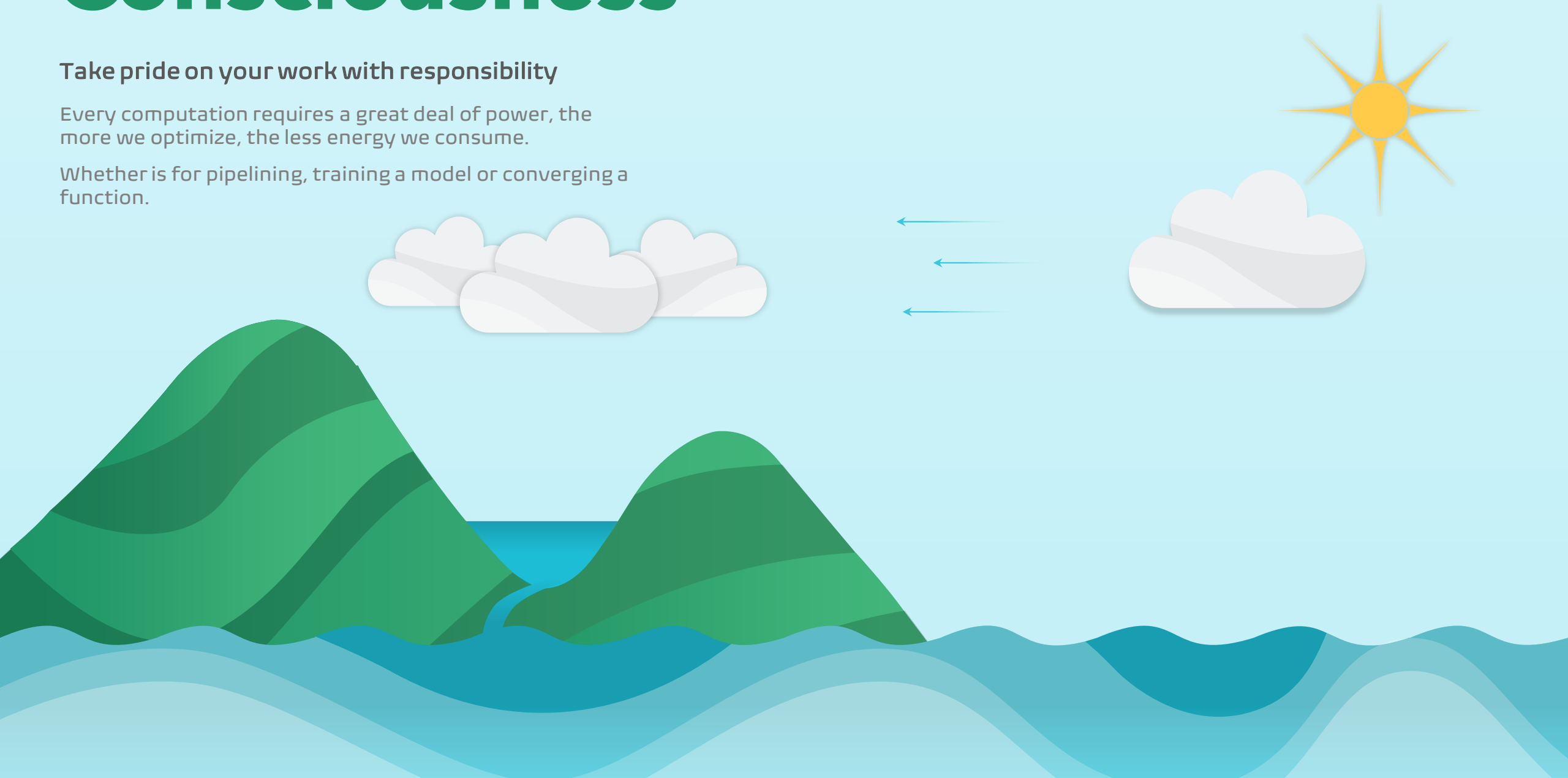


Consciousness

Take pride on your work with responsibility

Every computation requires a great deal of power, the more we optimize, the less energy we consume.

Whether is for pipelining, training a model or converging a function.



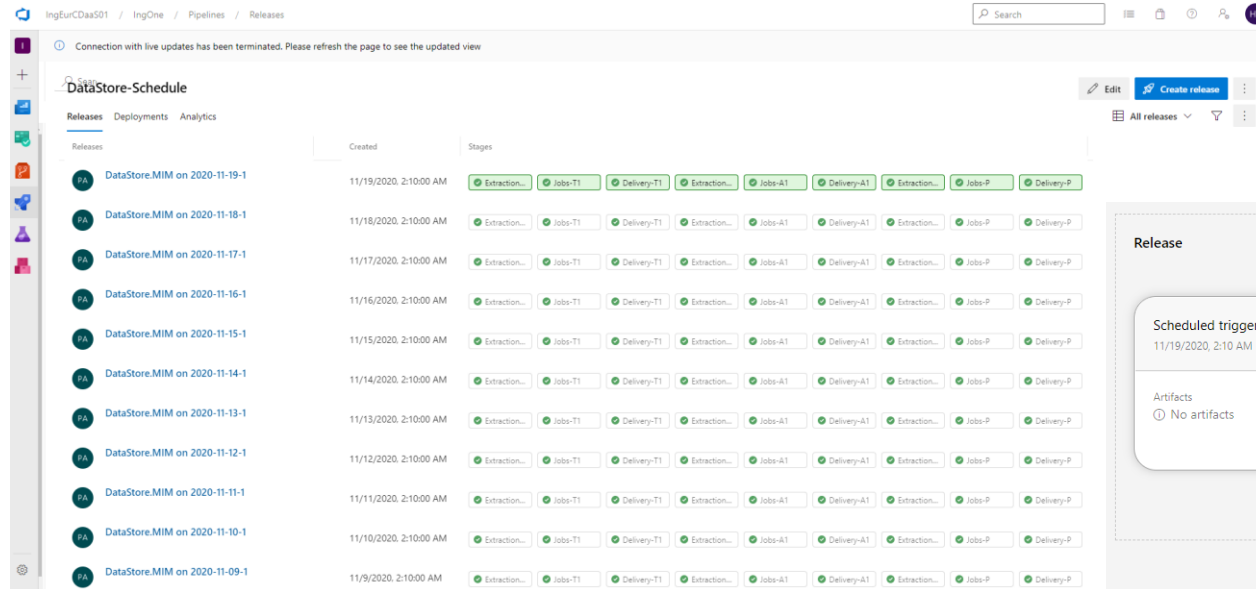
Software Energy Level?



Lines of Code
Vectorized
Data Movement
 $O'(n)$

Energy		Washing machine
Manufacturer Model		
More efficient		
A		
B		
C		
D		
E		
F		
G		
Less efficient		
Energy consumption kWh/cycle (based on standard test results for 60°C cotton cycle) <small>Actual energy consumption will depend on how the appliance is used</small>	0.95	
Washing performance <small>A: higher G: lower</small>	A B C D E F G	
Spin drying performance <small>A: higher G: lower</small> Spin speed (rpm)	A B C D E F G 1400	
Capacity (cotton) kg	5.0	
Water consumption /	55	
Noise (dB(A) re 1 pW)	Washing 5.2 Spinning 7.0	
<small>Further information is continued in product brochures</small>		

Operationalize



The screenshot shows the 'DataStore-Schedule' interface with a table of releases. The table has columns for 'Created' and 'Stages'. The 'Stages' column contains a sequence of job names (Extraction, Jobs, Delivery) for each release, each with a green checkmark indicating success. The releases are listed on the left side of the table.

Created	Stages
11/19/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/18/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/17/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/16/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/15/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/14/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/13/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/12/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/11/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/10/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P
11/9/2020, 2:10:00 AM	Extraction... Jobs-T1 Delivery-T1 Extraction... Jobs-A1 Delivery-A1 Extraction... Jobs-P Delivery-P

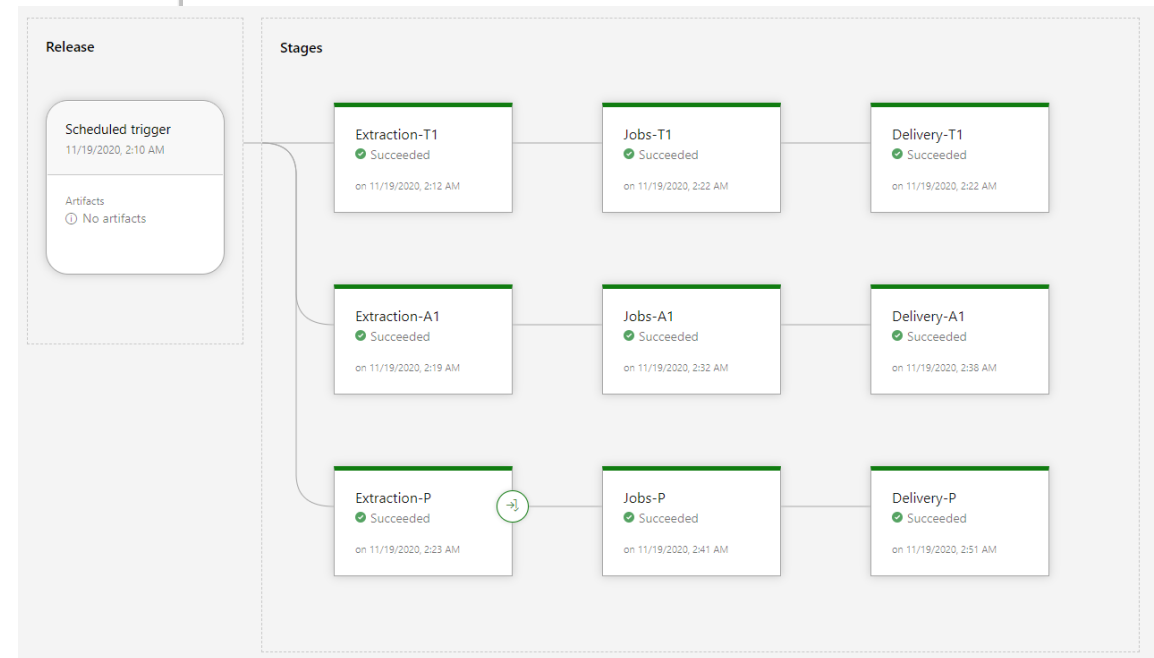
Data as releases

Audit and Traceability

History

No human intervention

Multi-schedule



Decision Engineering

Prevent human bias

Mediate qualitative with
quantitative inputs

Harness your judgement with
statistical significance

Fact-based knowledge



In God we trust; all others bring data

W. Edwards Deming

Data Provenance

Data Lineage

Canonical Model

Feature Engineering
Approach

Explainable Models | Open ML
Deep Learning Models | NN

Scores and Confidence Levels

Communicate in Domain Terms

Arrive with Data Points

Gap

You can help to close the gap



Exploring problems from a different angle

Understanding the real customer needs

Align data journey to company vision

Industry

Academia



The data industry is fascinating...

Welcome onboard!



Herminio Vazquez

