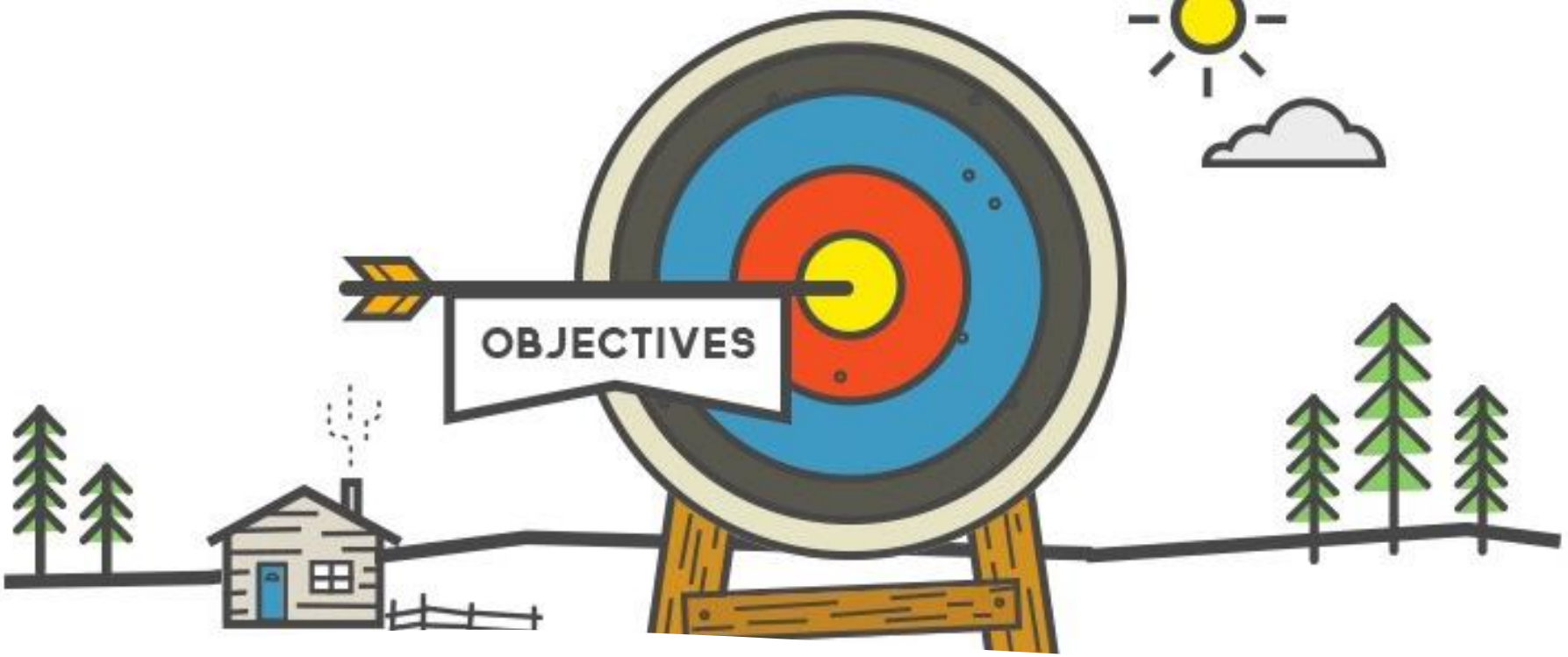
Case Studies in Sustainable Big Data

Prof. Carlos J. Costa, PhD

Saeed Angorani, DBA



Learning Goals

- Describe how big data creates measurable sustainability impact
- Identify data sources, pipelines, and KPIs in real deployments
- Select batch vs real-time processing for a given sustainability problem
- Critique dashboards/models for clarity, bias, and operational usefulness
- Draft a mini project brief for a sustainability analytics use case

How we analyze each case

For each case, capture:

- Business problem + decision owner
- Data sources (volume, velocity, variety)
- Architecture (ingest → storage → process → serve)
- Processing style: batch / real-time / hybrid (why)
- Analytics: descriptive → predictive → prescriptive
- KPIs: business + sustainability
- Risks: data quality, privacy, drift, adoption
- Outcome: impact, savings, operational change

Databricks

Upload Data

- data engineering/Data Ingestion
- Add data/Create or modify table
- Create or modify table from file upload
  - `electricity`

Databricks

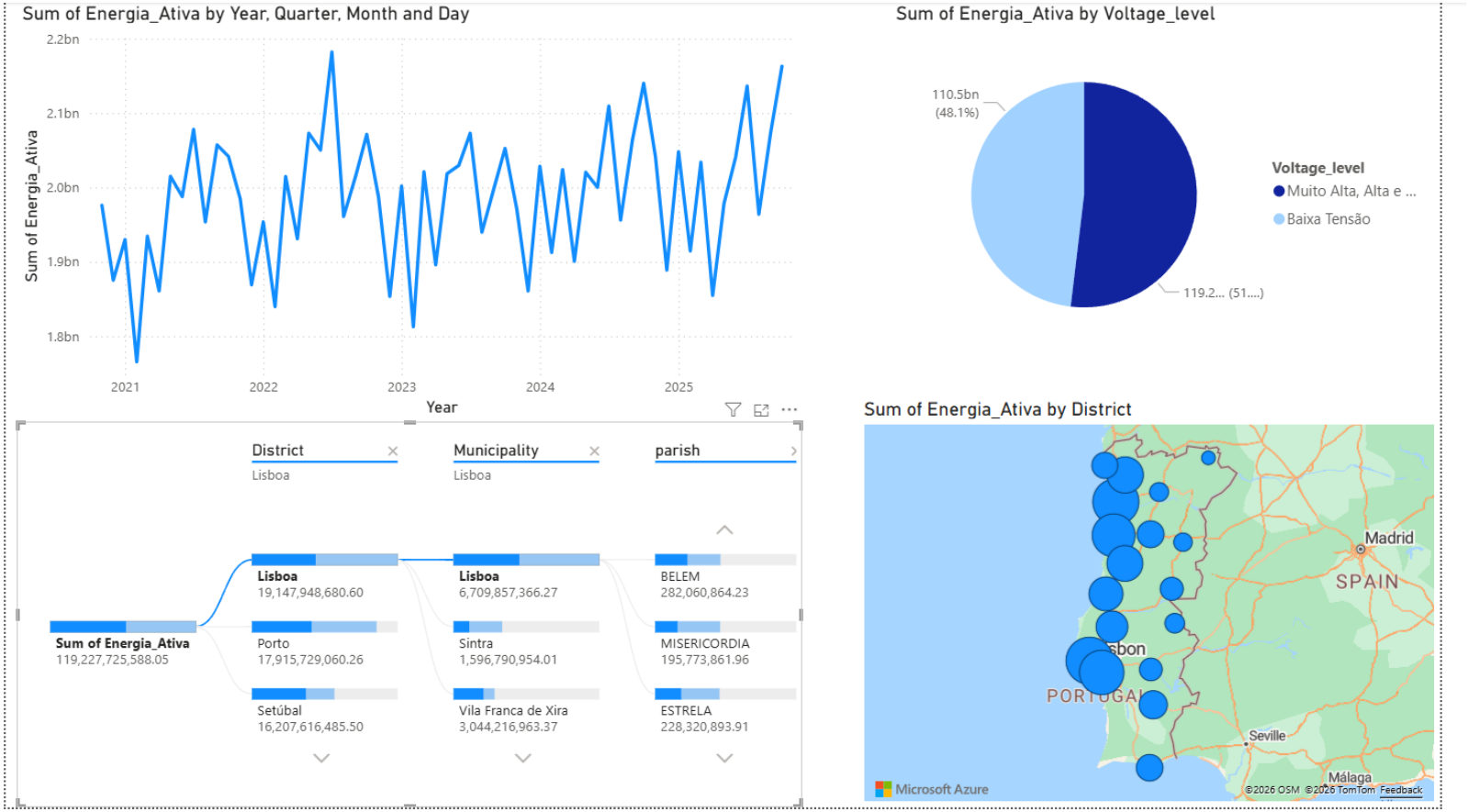
- Workspace
- Create/Notebook

```
# Read using spark.read.table
df = spark.read.table("workspace.default.electricity")
# Show first 20 rows
display(df)
```

it is possible to do: ;-)

```
# Convert Spark DataFrame to Pandas DataFrame
pdf = df.toPandas()
```

Link with PowerBI



How We Analyze Each Case (The Template)

Element	Description
Business problem	Decision owner + pain point
Data sources	Volume, velocity, variety
Architecture	Ingest → storage → process → serve
Processing style	Batch / real-time / hybrid (why)
Analytics	Descriptive → predictive → prescriptive
KPIs	Business + sustainability metrics
Risks	Data quality, privacy, drift, adoption
Outcome	Impact, savings, operational change

Case Study A: Energy Efficiency
E-REDES Municipal Consumption Data

- **Context: Private electricity consumption aggregated by municipality**

Aspect	Details
Data provider	E-REDES (Portuguese distribution network operator)
Scope	Private consumption (residential + commercial), aggregated at municipality level
Granularity	Monthly/annual totals per concelho
Access	Published datasets via E-REDES portal and dados.gov.pt
Privacy model	Pre-aggregated (no individual meter data exposed)

Case Study A: Data Characteristics
E-REDES Dataset Profile

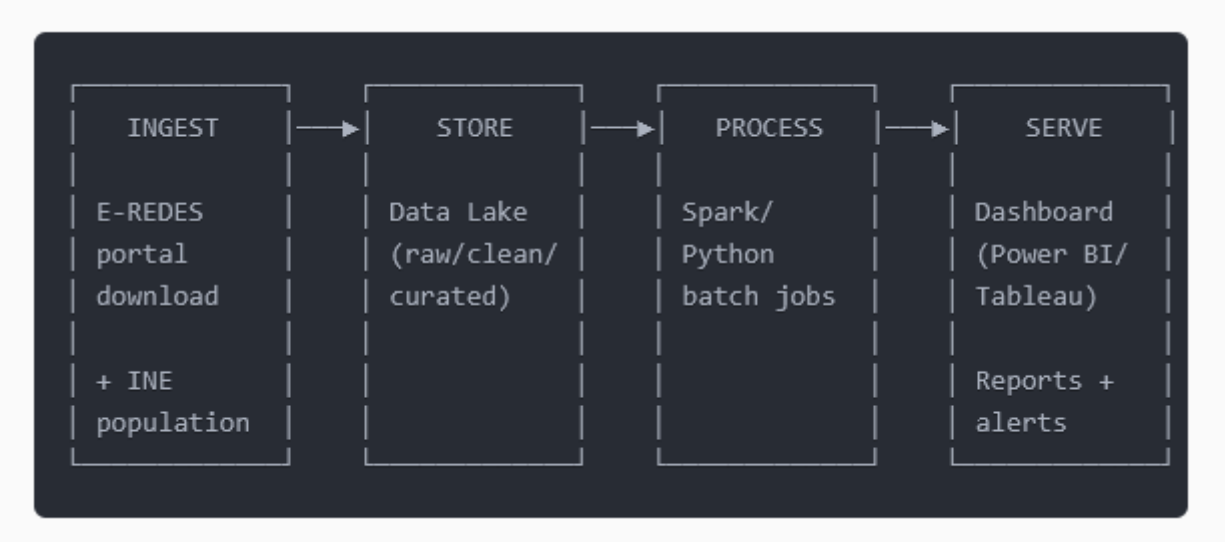
Dimension	Characteristics
Volume	~308 municipalities × 12 months × multiple years = moderate (tens of thousands of records)
Velocity	Batch updates (monthly or quarterly releases)
Variety	Structured tabular data (CSV/Excel); single source
Veracity	High (official metering data, aggregated by DSO)

Key fields typically available:

- Municipality (concelho)
- Time period (month/year)
- Consumption (kWh)
- Number of supply points (optional)
- Tariff type breakdown (optional)

Case Study A: Architecture

- **Batch Analytics Pipeline for Municipal Energy Analysis**



- **Processing style: Pure batch**
- Data arrives monthly; no real-time decisions needed
- Analysis focuses on trends, benchmarks, and planning

Case Study A: Analytics Layers

- **From Descriptive to Prescriptive**

Layer	Example Analysis	Output
Descriptive	Which municipalities consume most per capita?	Rankings, maps, time series
Diagnostic	Why did consumption spike in municipality X?	Correlation with weather, population, economic activity
Predictive	What will consumption be next year?	Forecasts per municipality (ARIMA, Prophet, LSTM, …)
Prescriptive	Where should efficiency programs be prioritized?	Ranked intervention list based on savings potential

- **Enrichment sources:**
- INE: population, housing stock, economic indicators
- IPMA: heating/cooling degree days
- DGEG: energy policy timelines

Case Study A: KPIs
Business and Sustainability Metrics

Category	KPI	Unit
Business	Consumption per capita	kWh/person/year
Business	Year-over-year change	%
Sustainability	CO_2e emissions (using grid factor)	tCO_2e/municipality
Sustainability	Progress toward municipal climate targets	% of goal achieved
Operational	Data freshness	Days since last update

- **Grid emission factor (2023–2024):**
- Portugal average: ~150–200 gCO_2/kWh (varies with renewable share)
- Source: DGEG or EEA

Case Study A: Decisions Enabled

- **What Municipal Planners Can Do With This Data**

Decision	Data-Driven Input
Target setting	Baseline consumption establishes realistic reduction goals
Program evaluation	Before/after comparison for efficiency campaigns
Benchmarking	Compare similar municipalities (size, climate zone)
Resource allocation	Prioritize interventions where impact is highest
Public reporting	Transparent progress dashboards for citizens

- **Limitation to acknowledge:**
- Aggregated data cannot identify specific buildings or behaviors
- Complements (does not replace) building-level audits

Case Study A: Risks and Mitigations

Risk	Impact	Mitigation
Aggregation masks variation	Cannot see outliers within municipality	Combine with sample audits or surveys
Delayed data	Decisions based on stale information	Build in lag assumptions; use leading indicators
Definition changes	Municipality boundaries or tariff categories shift	Document metadata; version datasets
Attribution error	Confusing correlation with causation	Use control groups or difference-in-differences
Low engagement	Dashboards built but not used	Co-design with decision-makers; embed in workflows

Quick Exercise (10 min)

Analyzing E-REDES Data

Scenario: You have 5 years of monthly municipal consumption data for all 308 concelhos.

- **Questions:**
  1. **What enrichment data would you join?** Name 2 sources and why.
  2. **Design a "consumption efficiency score"** that accounts for population and climate. What's your formula?
  3. **Pick 3 KPIs** for a municipal sustainability dashboard (at least 1 must be sustainability-focused).

Expected directions:

- Enrichment: INE population, IPMA degree days, DGEG renewable capacity
- Score example: kWh / (population × HDD+CDD)
- KPIs: per-capita consumption, YoY change, estimated CO_2e

References

- Team, I. C. E. (2021, maio 27). *Hadoop vs. Spark: What's the Difference? | IBM*. https://www.ibm.com/think/insights/hadoop-vs-spark