Python for Data Engineering

BY: Myk Ogbinar

Data & Analytics Head @ San Miguel Corp

Founder @ Data Engineering Pilipinas

DurianPy

PyCon DAVAO 2025 PANAGHIUSA

Workshop Expectations

✅ You can write basic Python apps

❌ You don’t need to know what a Data Engineer is — yet

Today we’ll:

1) Demystify what Data Engineers actually do
2) Build your first working data pipeline
3) Show how Python scripts become reliable data systems

Py-Data Engineering Workshop

From CSV to Dashboard — Building a Mini Data Pipeline in Pure Python

pandas • pyarrow • uv • streamlit

DurianPy PyCon DAVAO 2025 PANAGHIUSA

Agenda

Segment	Duration
Intro + Setup	30 min
Extract	10 min
Load (Clean + DQ)	20 min
Transform	10. min
Build	15 min
Streamlit Dashboard	15 min
Wrap-up + Q&A	20 min

DurianPy

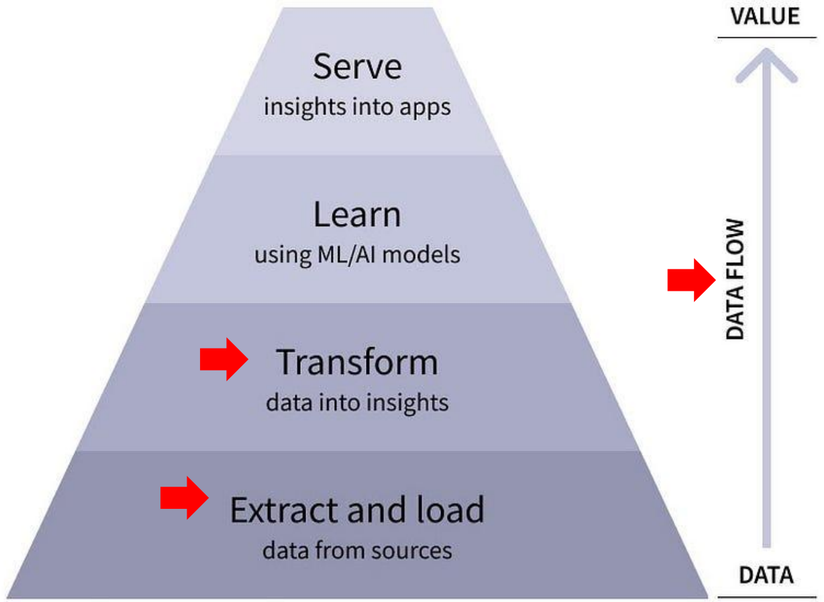
PyCON DAVAO 2025 PANAGHIUSA

Introduction to Data Engineering

DurianPy PyCON DAVAO 2025 PANAGHIUSA

What is Data Engineering?

Hierarchy of needs for data driven companies



4. Predict and improve outcomes

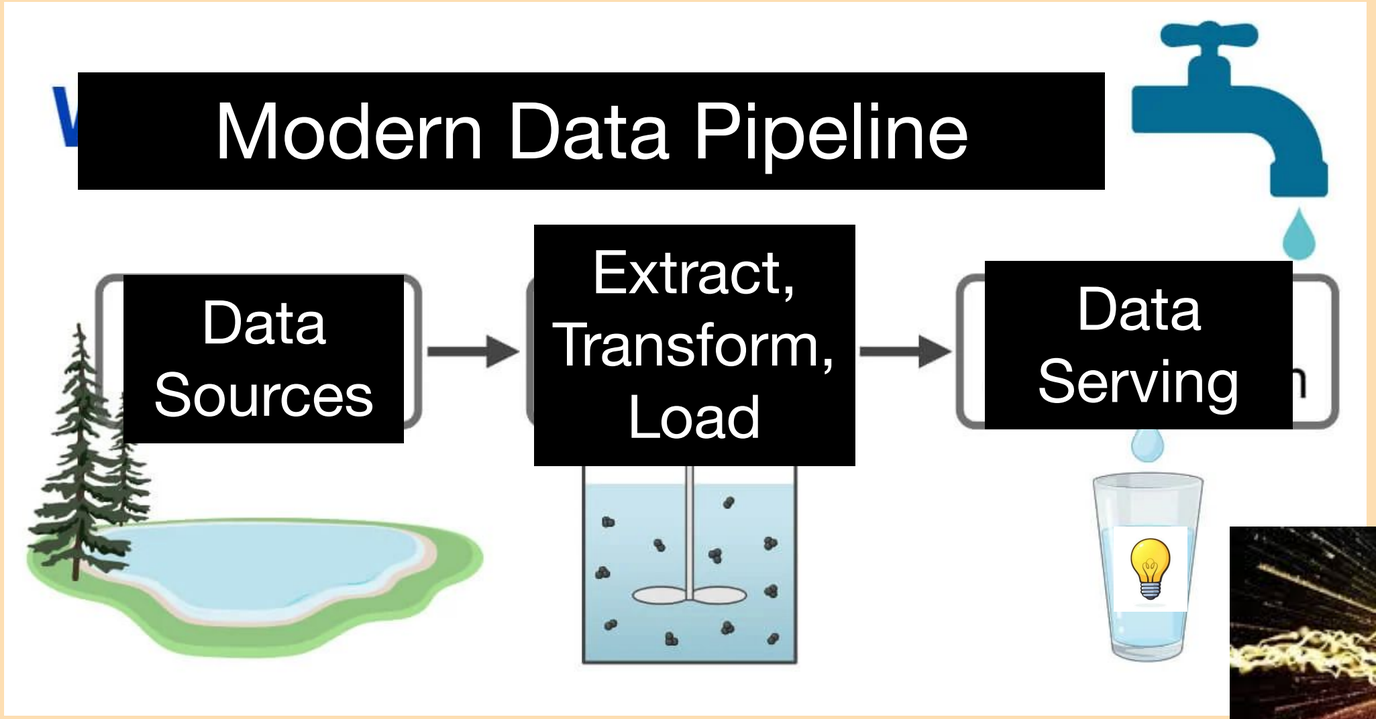
3. Analyze/categorize using ML

2. Compute & visualize KPIs/metrics

1. Unify and clean raw datasets

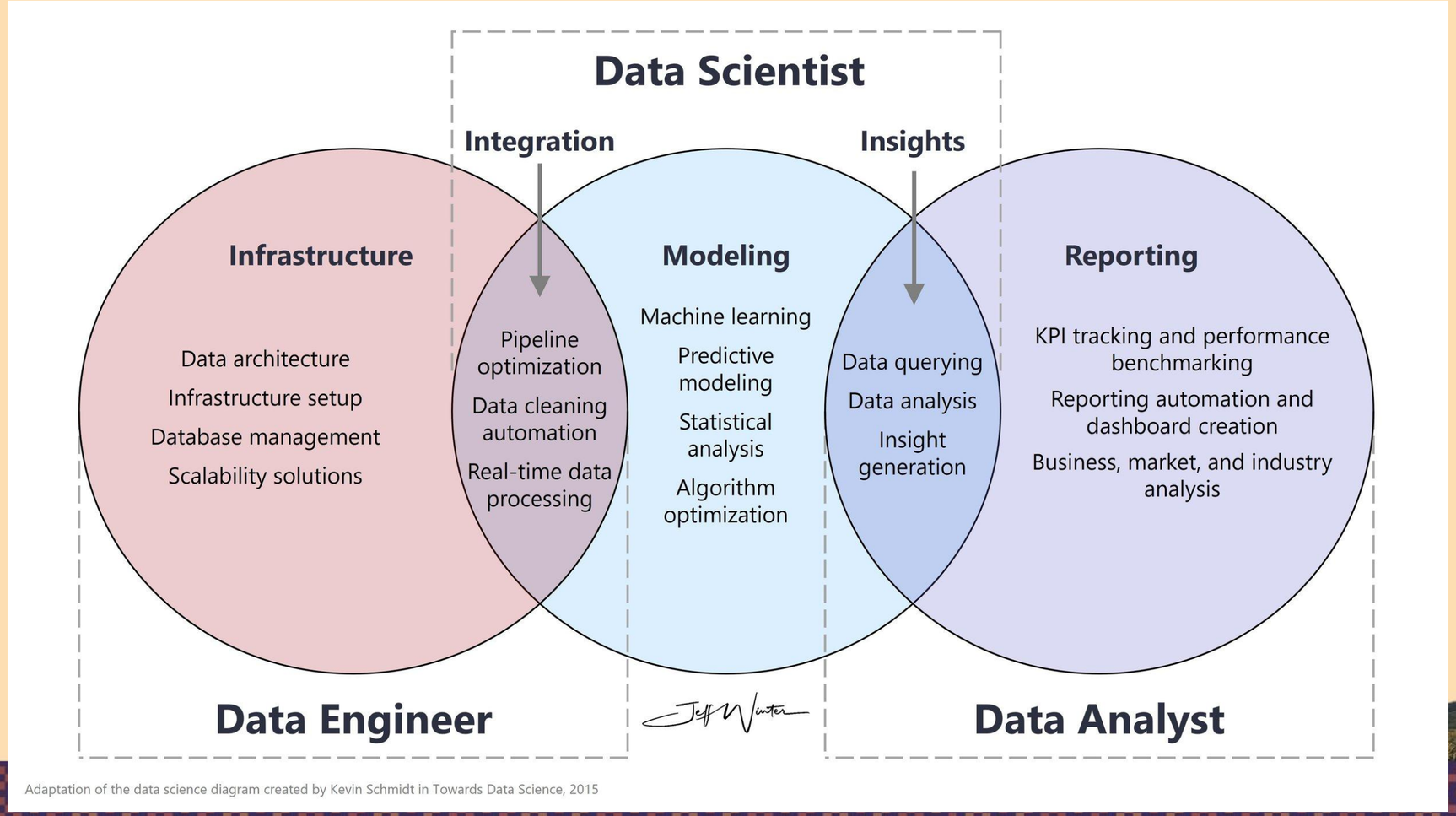
Data Engineering is tasked with building the Data Stack to support these needs

What is the DE Workflow?



Why is Data Engineering Important?

Data Engineer vs Data Scientist vs Data Analyst



Adaptation of the data science diagram created by Kevin Schmidt in Towards Data Science, 2015

We're currently looking for a Data Migration Engineer based in the Philippines to join our growing team.

Key Requiremer

- Lead data onl
- Expand and ir
- Support creat
- Reconcile inv
- Analyze data

Compensation: $1500 to $1800 per month

Hiring!

Data Engineer / Business Intelligence Developer (Power BI)
Location:Quezon City
Salary:PHP80000 - PHP90000 per month + To be discussed

Head of Data Engineering (Remote work)

Sysgen RPO, Inc. ★ 3.5 8 reviews · View all jobs

Metro Manila (Remote)

Architects (Information & Communication Technology)

Full time

₱300,000 – ₱390,000 per month

Posted 4d ago • Medium application volume

🚀 Hiring: Data Engineers (Multiple positions available) for a global technology firm in Australia 🚀
📍 Fully-Remote | Philippines
💰 Base Salary: PHP 150-170k/month + sign-on bonus + statutory benefits

Data Modeling Techniques and Methodologies in Mandaluyong

Salary: 80,000-110,000/month
Job Level: Experienced Hire

Senior Data Engineer in Taguig
Salary: 85,000-108,000/month
Job Level: Experienced Hire

Data Analyst / Data Engineer (cloud) (wfh) (day shift) 150-200K

Jobstreet

Data Analyst / Data Engineer (cloud) (wfh) (day shift) Job in Makat...

We're looking for both Data Engineer and Data Analyst that has years of experience and has experience in Cloud

🚀 **Hiring: Lead Data Engineer** 🚀

We're looking for a **Lead Data Engineer** to lead the design, development, and implementation of scalable data pipelines and infrastructure! If you're passionate about **big data, cloud platforms, and leadership**, this role is for you! 🌟

📍 **Location:** Ortigas, Pasig (Hybrid Work Setup 2x

💸 **Salary Range:** Php 200,000 - 220,000 monthly

🌟 WE'RE HIRING: SENIOR DATA ENGINEER

📍 **Location:** QC Hybrid/ GY Hours Mon-Fri
🕒 **Start Date:** ASAP
💼 **Engagement Type:** Long-Term Contract
💰 **Budget:** Up to PHP 100K/month

🚀 We're Hiring! Data Engineer (Junior to Senior Level) 🚀

📍 **Location:** UP Ayala, Quezon City (Hybrid)
🕒 **Schedule:** Day Shift
💰 **Salary:** PHP 60,000 - 115,000

Hi Engineers! We are hiring for experienced Data Engineer.

Position: Director Data Engineering

WFH set-up

Salary: Php 300, 000

2025

What is Data?

What is Data?

- ✖ Originally, data is plural for “datum”, a Latin word
- ✖ Datums are most often called “data points”
- ✖ Data represents a collection of data points (also called datasets)
- ✖ Can be used as both singular and plural form

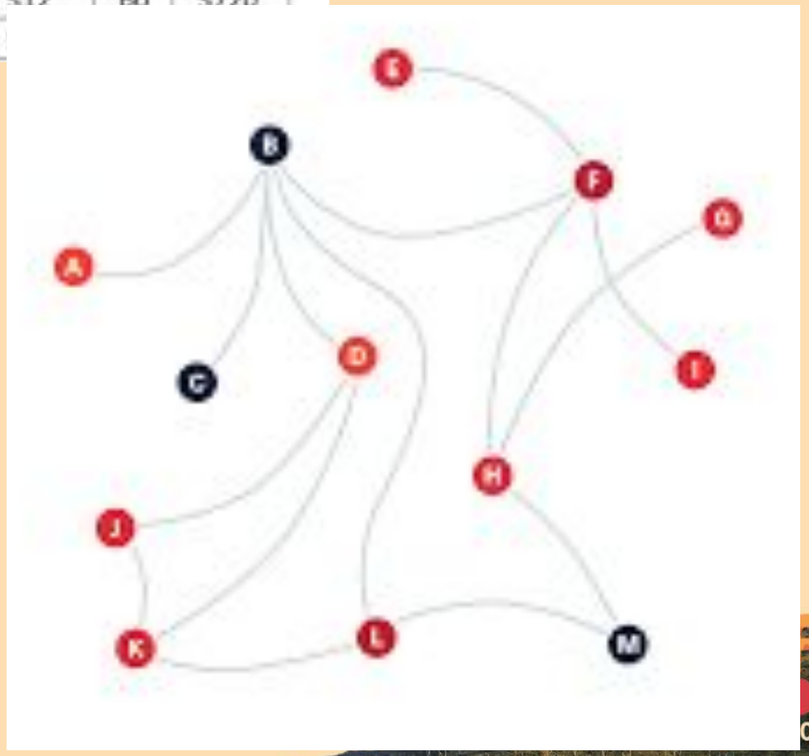
Data are characteristics or information, usually numerical, that are collected through observation.

- Wikipedia

Shapes of Data

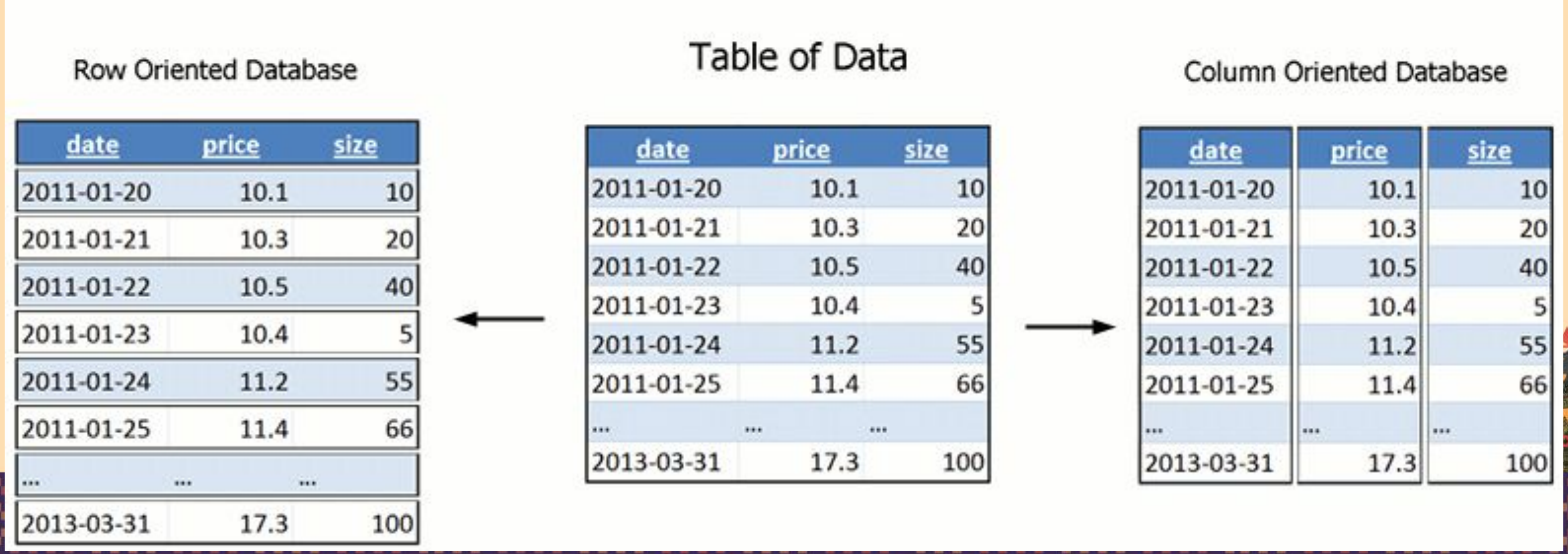
Structure	Example	Common Format
Tabular	Sales table	CSV / SQL
Hierarchical	Product catalog	JSON / XML
Graph	Social network	Neo4j, RDF
Time Series	Sensor readings	Parquet / Influx
Geospatial	Map coordinates	GeoJSON, Shapefile

D	F	G	H	I
Sales Representative	Item	Price Per Unit ($)	Units Sold	Total Sales ($)
celia Blake	HJK_10	$25	50	$1,250
nie Freeman	HJK_12	$20	100	$2,000
evieve Evans	HJK_33	$10	15	$150
don Cohen	HJK_10	$25	25	$625
an Davis	HJK_33	$10	110	$1,100
an Grant	HJK_27	$12	55	$660
ry Herrera	HJK_12	$20	120	$2,400
gan Hicks	HJK_64	$30	115	$3,450
er Fleming	HJK_27	$12	60	$720
ra Rogers	HJK_64			



Column vs Row Storage

Row-Oriented DB	Column-Oriented DB
> Stores/retrieves data by row → may read unnecessary columns	> Stores/retrieves data by column → reads only needed columns
> Fast for inserts/updates (OLTP)	> Slower writes, optimized for reads (OLAP)
> Aggregations are expensive	> Aggregations are efficient
> Lower compression efficiency	> High compression due to similar column values



date	price	size
2011-01-20	10.1	10
2011-01-21	10.3	20
2011-01-22	10.5	40
2011-01-23	10.4	5
2011-01-24	11.2	55
2011-01-25	11.4	66
...	...	...
2013-03-31	17.3	100

Special Values in Data & Data Quality

Value	Meaning	Example Sources	Data Cleansing Impact
NULL	Missing / unknown	Unfilled field	Standardize & decide: impute, ignore, or drop
NaN	"Not a Number" (invalid numeric result)	`0/0`, bad calculations	Convert to NULL or fix upstream logic
Infinity / -Infinity	Overflow or division by zero	`1/0.0`	Usually errors; filter, cap, or convert
Empty String (' ')	String with length 0 (≠ NULL)	User form input	Standardize to NULL if it means missing
Sentinel Values	Fake placeholders	`-999`, `'NA'`	Detect & replace with NULL

- Identify & standardize special values.
- Make missingness explicit (prefer NULL).
- Decide: Impute, Exclude, or Flag depending on context.

Data Modeling

Database normalization

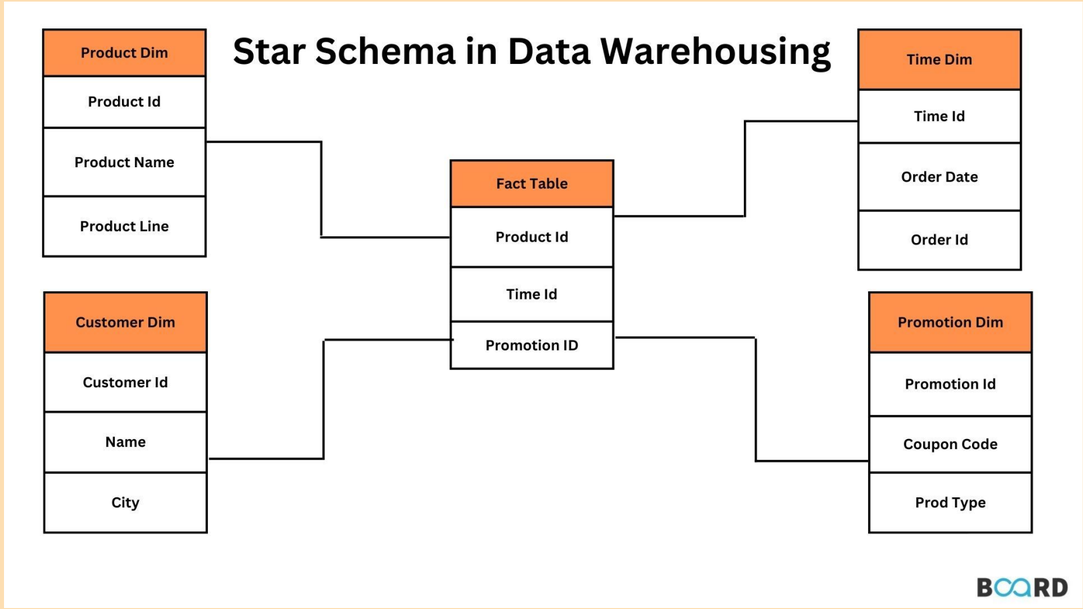
order_id	customer_name	product_name	price
1001	Mike Smith	Laptop	800
1002	Sara Jones	Mouse	20
1003	Mike Smith	Laptop	800
1004	Sara Jones	Laptop	800

Original unnormalized table

customer_id	customer_name	email
151	Mike Smith	mike@example.com
152	Sara Jones	sara@example.com

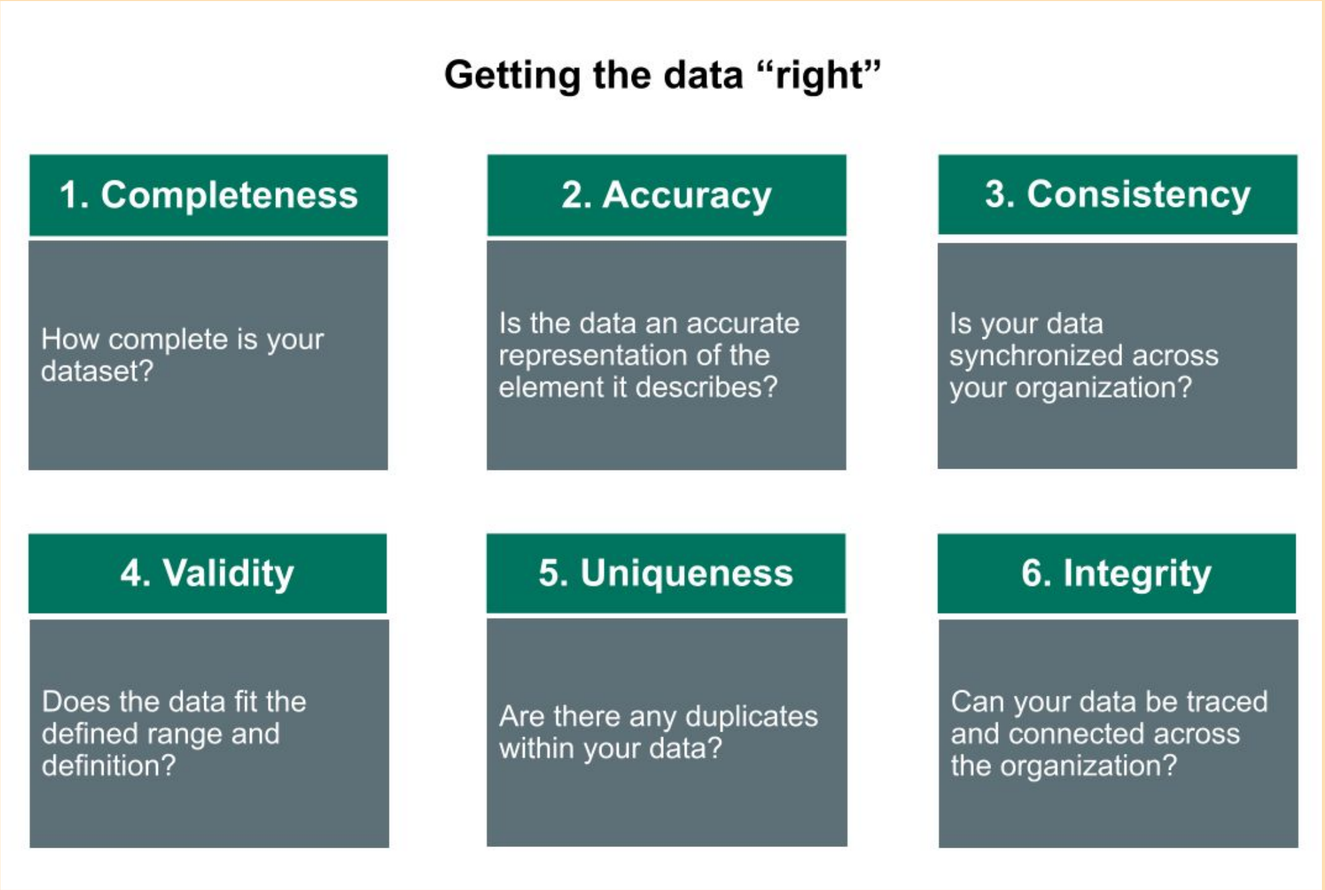
order_id	customer_id	order_date
1001	151	2024-08-16
1002	152	2024-08-19
1003	151	2024-08-22
1004	152	2024-08-25

product_id	product_name	price
501	Laptop	800
502	Mouse	20



Why Data Cleaning & Data Quality Matters

- Raw data = messy, inconsistent, incomplete
- Poor quality data → bad decisions
- Cleaning ensures accuracy, reliability, trust
- DE Mindset: Cleaning & Standardization happens in stages.



https://miosoft.com/resources/articles/data-quality-dimensions-untangled.html

Python for Data Engineering

Library	Role	Concept
`pandas`	data manipulation	SQL + Excel hybrid
`pyarrow`	Parquet I/O	columnar efficiency
`pathlib`	file system abstraction	portability
`streamlit`	lightweight visualization	serving layer
`uv`	dependency management	reproducibility

DurianPy

PyCON DAVAO 2025 PANAGHIUSA

Let’s get started!

Challenges of Data Engineering

- Inconsistent data formats
- Missing or invalid values
- Schema drift
- Scaling and performance
- Governance issues

How to Expand the Project

- Add more tables (Employees, Suppliers, Categories)
- Add new DQ rules
- Replace Parquet with DuckDB
- Schedule with Prefect
- Add new charts to Streamlit

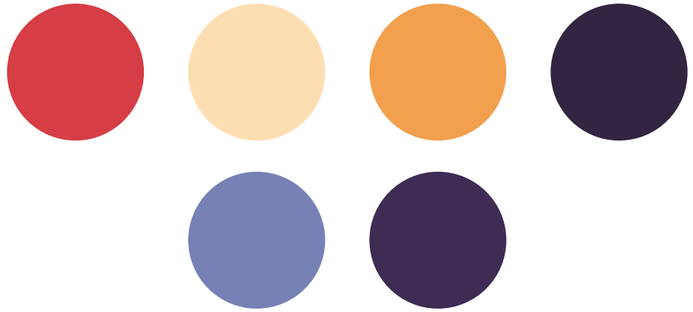
Key Takeaways

- Data Engineering = making data usable
- Python can handle full workflow
- Focus on structure, quality, reproducibility

Closing Notes

- You didn’t just learn Python — you learned how to engineer data.
- Visit: dataengineering.ph
- Add: linkedin.com/in/ogbinar

BRANDING



PIPANGANAN H1

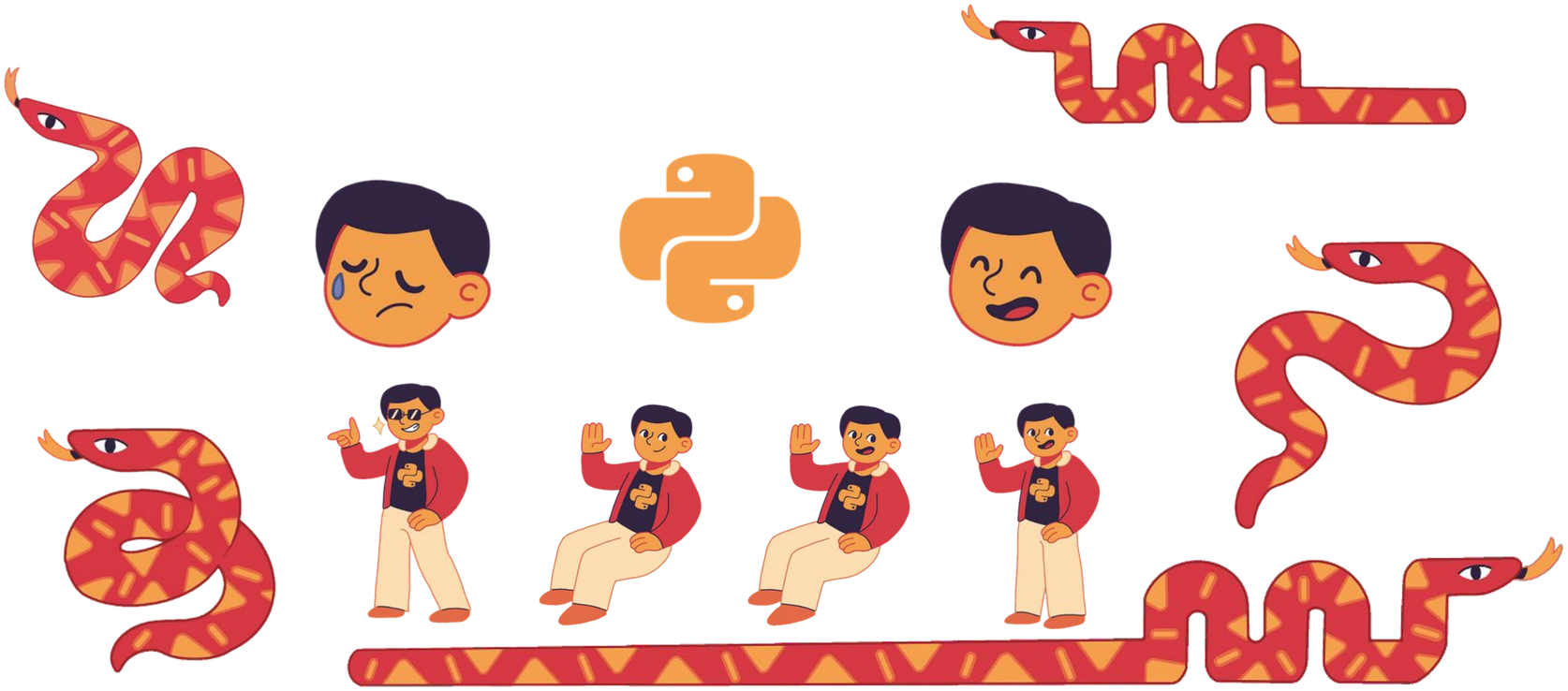
Nunito Sans (Black) H1

Nunito Sans (Bold) H2

Inter (Medium) Body

Nunito Sans (Semi-Bold) Body

ASSETS



TITLE HERE

Insert Text