

# Big Data Analytics

Piotr Menclewicz

# Why Big Data?

---

- **200 CVs for DA role**
- **3k CVs for ~10s of job offers**
- **“In any other case I’d like to be younger...  
... but not in case of searching IT job”**

# Why Big Data?






















---



**Klarna.**



# It still pays off to enter the field!

OFERTY PRACY				 Sortowanie	 Subskrybuj	 Oceń wyniki
	<b>Data Scientist – Conversational AI and Agentic Workflows – Remote</b> <span>NOWA</span>	AI Python LLM scikit-learn	16 800 – 20 160 PLN			
Link Group						
	<b>Data Scientist</b> <span>NOWA</span>	Data Python NLP AWS	24 000 – 30 000 PLN			+17
Sii Polska						
	<b>Data Scientist – Remote</b> <span>NOWA</span>	AI Python Data analysis Data mining	21 840 – 25 200 PLN			+1
Scalo						
	<b>Data Scientist</b>	Data AWS AI Databricks	23 500 – 26 800 PLN			
AVENGA						
	<b>Data Scientist (NLP &amp; LLMs)</b> <span>NOWA</span>	AI Python GenAI Azure	17 800 – 21 000 PLN			+5
Onwelo Sp. z o.o.						
	<b>Senior Data Scientist</b> <span>NOWA</span>	AI Spark Machine learning AI	20 000 – 32 000 PLN			
Neontri						

# **Main goal of this course**

---


**Understand which skills will get you there, and  
learn (some) of them!**

# About me

---



Piotr Menclewicz,  
Staff Data Analyst  
**@GetInData**

 *10+ years in data analytics  
(ex-Allegro, ex-McKinsey)*

 *not-academian, but teaching  
others is my passion*

 *family life can be an adventure*

**e-mail:** [p.menclewicz@uw.edu.pl](mailto:p.menclewicz@uw.edu.pl)

# About GetInData

---

**We work with organizations  
to grow their business  
by using data, analytics &  
technology**



**Founded in 2014 by  
ex-Spotify engineers**



**Focus on  
Big Data & Cloud  
(from day 1)**



**130+ engineers and data  
scientists**



**Community Builders  
(conference, meetups, blog  
posts, open-source)**

# Your Experience

---

1. Have you already used **Hadoop** or other **Big Data** tools?
2. Have you worked with any **data** related **cloud solutions**?
  - a. Google BigQuery
  - b. Amazon Redshift, Athena
  - c. Snowflake
  - d. Databricks
3. What is your **experience** with popular **programming languages**?
  - a. Python, SQL, R, Scala
4. What is your **experience** with **ML**?



# Goals

---

- **Introduction to Big Data**
  - Trends in data
  - Fundamental concepts and most popular technologies
- **Hands-on experience with most popular technologies**
  - Inspired by a real-world companies and their use-cases
- **Knowledge** required to **start your journey** with **Big Data**
- How to **transfer** your **current knowledge** about **data analysis** and **machine learning** to **Big Data**

## What we won't do:

- Learn **new data analysis** or **ML** concepts

# Assessment

---

- **Active attendance (hands-on quests: unique, high-quality & on-time responses)**
- **Project**

or

- **Test**

**Why such approach?** 🤔

# Logistics

---

- usually ~45 intro & discussion
- 10 minute break
- ~35 minute hands-on exercise

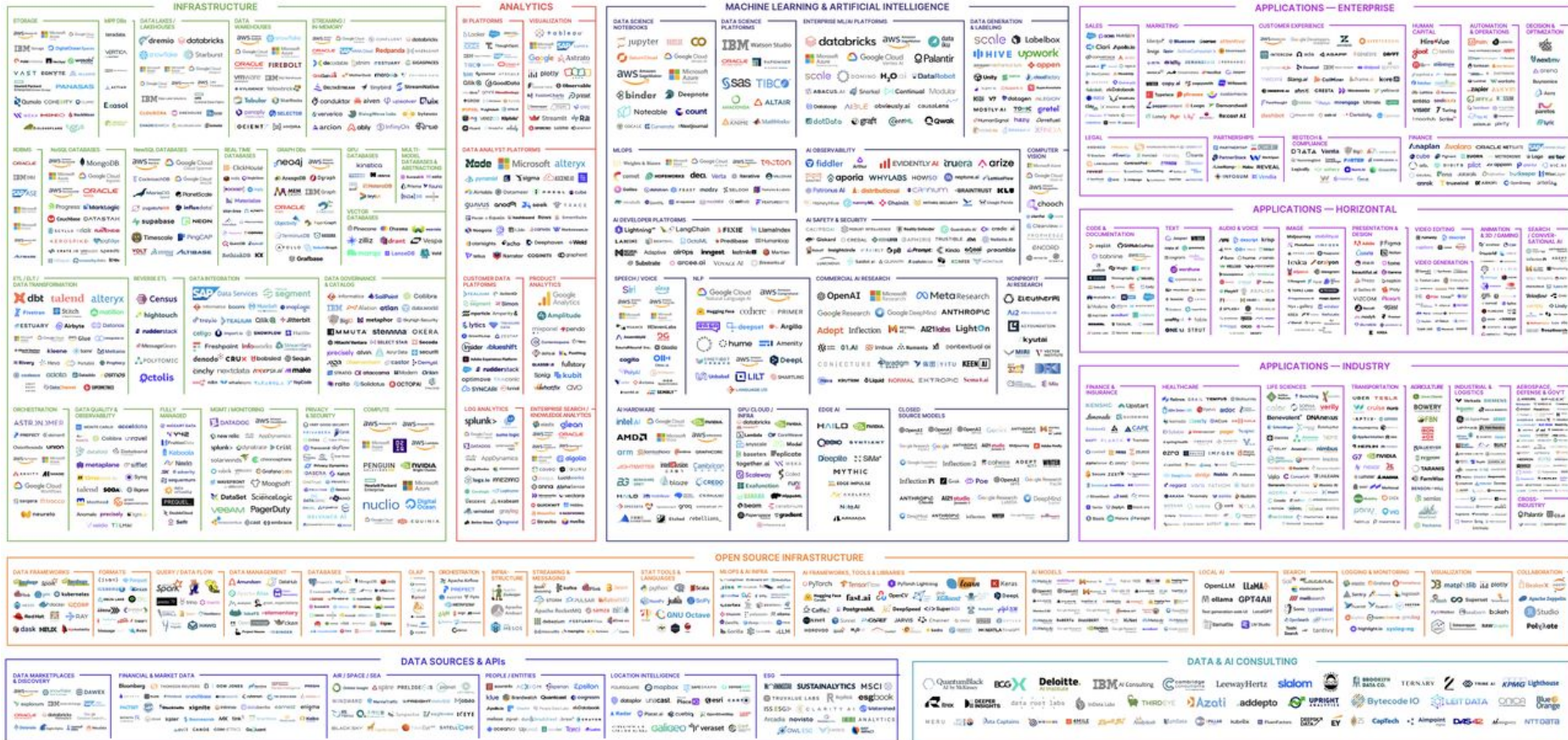
# Today you'll learn...

---

- **Core skills for Data Analysis**
- **Key trends**

# Core skills for Data Analysis

THE 2024 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



Version 1.0 - March 2024

© Matt Turck (@mattturck), Aman Kabeer (@AmanKabeer1) & FirstMark (@firstmarkcap)

Blog post: [mattturck.com/MAD2024](https://mattturck.com/MAD2024)

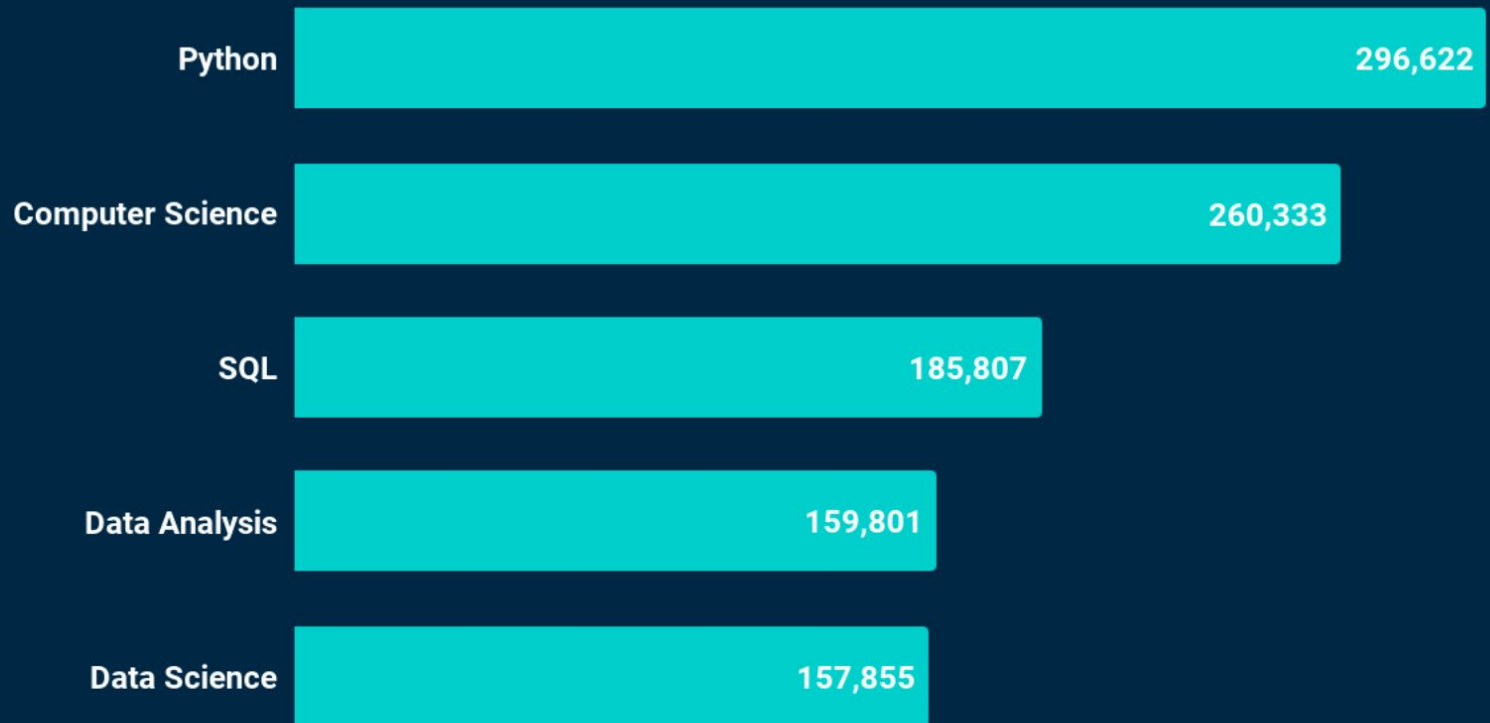
Interactive version: [MAD.firstmarkcap.com](https://MAD.firstmarkcap.com)

Comments? Email [MAD2024@firstmarkcap.com](mailto:MAD2024@firstmarkcap.com)

FIRSTMARK  
EARLY STAGE VENTURE CAPITAL

# How to choose the right ones?

## Top skills in AI job offers



source: Stanford AI index report, 2022

# How to choose the right ones?

---

$$\text{ROI} = \frac{\text{value generated}}{\text{investment}}$$

# Core skills for Big Data Analysis

---

**Data Analyst vs Big Data Analyst?**





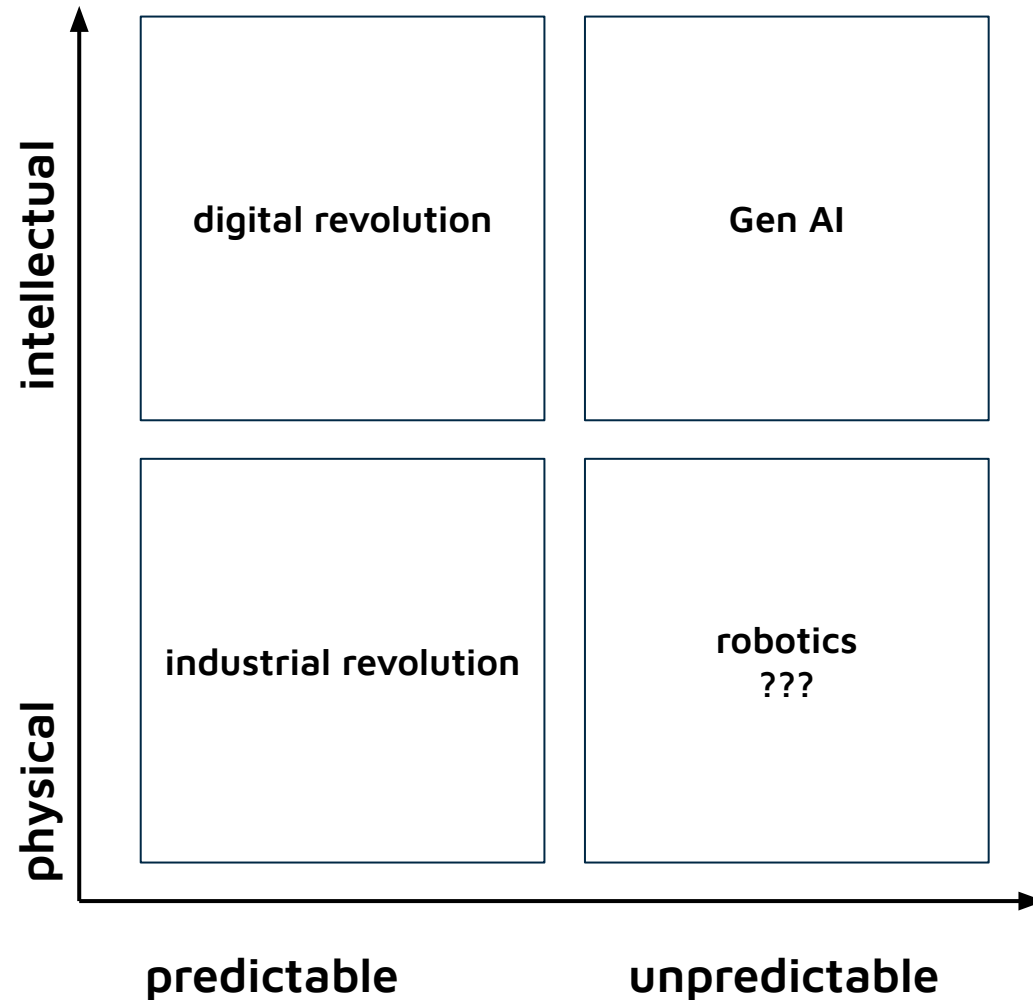
# Core skills for Big Data Analysis

---

- **SQL**
  - **BI**
  - **Python**
- } **often listed explicitly**

# Core skills for Big Data Analysis

---



# Core skills for Big Data Analysis

---

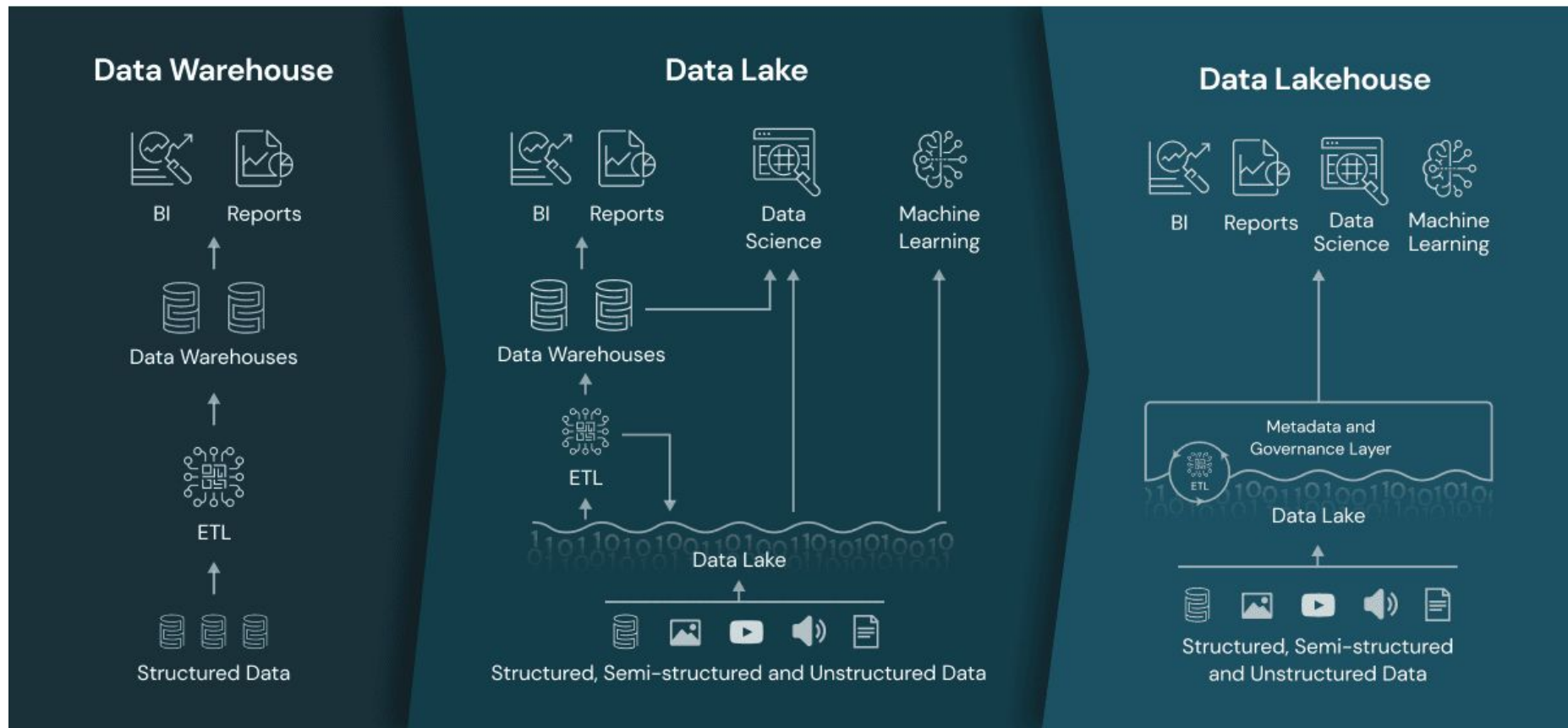
- **SQL**
  - **BI**
  - **Python**
- } **often listed explicitly**
- 
- **Business Understanding**
  - **Math/Statistics**
- } **make you perform great**

# Key trends in Data

---

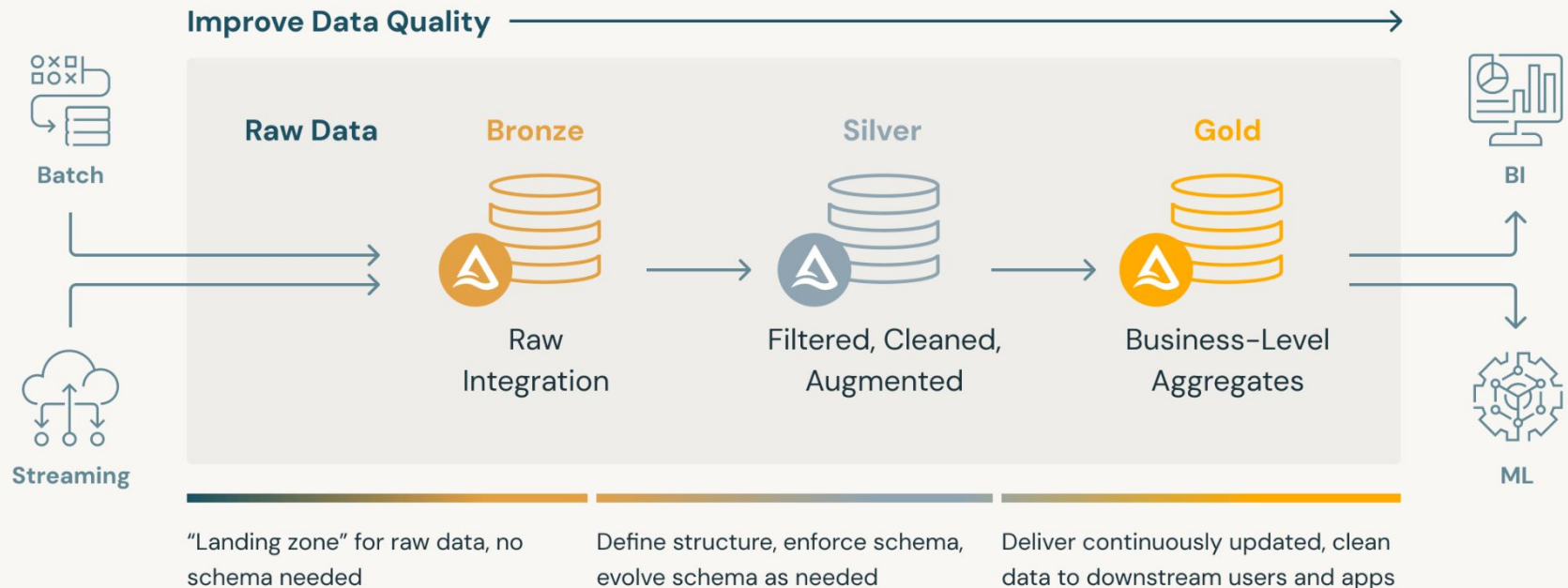
**HISTORY LIKES TO  
REPEAT ITSELF**

# Key trends in Data



# Key trends in Data

Building reliable, performant data pipelines with  **DELTA LAKE**



source: Databricks

# Key trends in Data

---

## **SELF-SERVICE**

# Key trends in Data

---



**Data  
Analyst??**



**Analytics  
Engineer**



# Key trends in Data

---

**GEN AI**

# Key trends in Data

---

## Generative AI for Natural Language to SQL

Embed Text-to-SQL right into your product with our versatile API. Get started free

Schedule Demo

 GitHub

 Discord

March 12th, 2024 | Written by Scott Wu

## Introducing Devin, the first AI software engineer

*And setting a new state of the art on the SWE-bench coding benchmark*

Meet Devin, the world's first fully autonomous AI software engineer.

Devin is a tireless, skilled teammate, equally ready to build alongside you or independently complete tasks for you to review.

With Devin, engineers can focus on more interesting problems and engineering teams can strive for more ambitious goals.

# Key trends in Data

---

**What do you think about using Gen AI tools in this course? What are pros and cons?**

**(companies face the same issue)**

# Assignment

---

**<https://forms.gle/36LsS7ttiLvB7mGF9>**