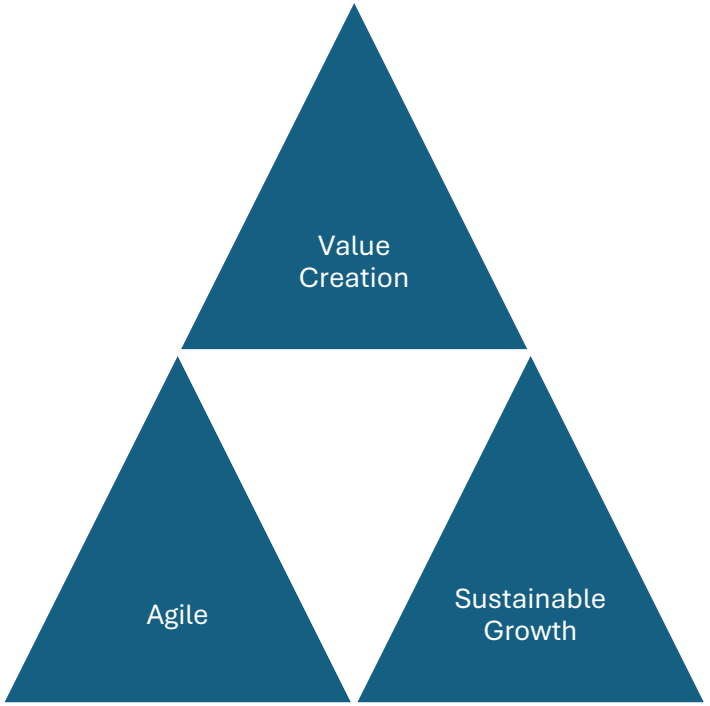
# Productionizing Data Pipeline

Ke Zhu

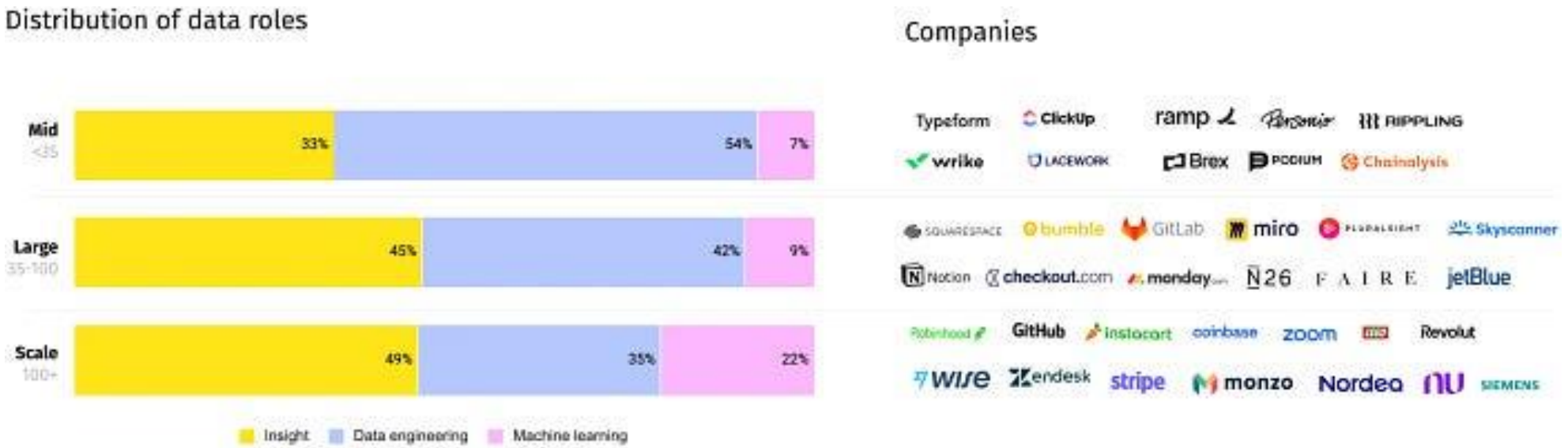
# About myself

- Lead data engineer of Beforepay – an ASX consumer lending business
- Engineer, architect and consultant on all data stuffs
- Doctor in Computer Science
- A software engineer

# What does a modern data team look like?




- A data product manager – or Head of Data / Head of Data Scientist
- A data/ML architect or principal lead
- Data analysts / data analytical engineers
- Data scientists / data analysts
- Data engineers



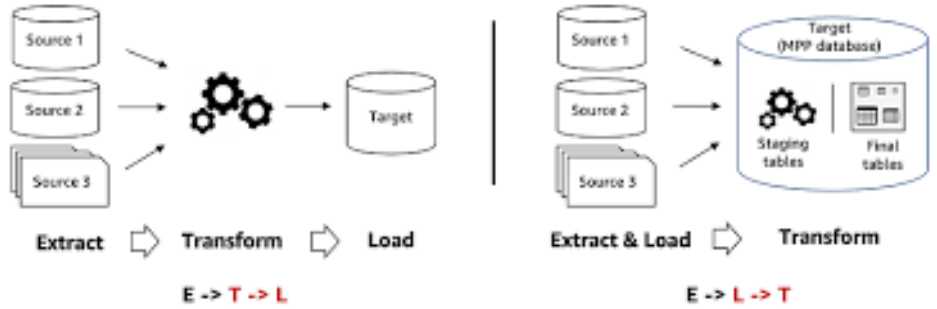

[How top data teams are structured | by Mikkel Dengsøe | Medium]

# Challenges in Productionizing a Data Pipeline

- Ensuring reliability and scalability of the pipeline
- Managing and monitoring data quality and consistency
- Integrating with multiple data sources and destinations
- Securing sensitive data and complying with regulations

# ETL vs ELT: A Comparison

- ETL stands for Extract, Transform, Load, while ELT stands for Extract, Load, Transform.
- ETL is ideal for traditional data warehousing, while ELT is better suited for big data processing.
- ETL involves transforming data before loading it into a target system, while ELT loads data into a target system before transforming it.



# A Brief Overview of Data Build Tool (dbt)

- dbt is a data transformation tool that enables teams to build, test, and deploy data models.
- dbt uses SQL and YAML to define data transformations and models.
- dbt is open source, extensible, and supports version control and collaboration.

GitHub - josephmachado/simple_dbt_project: Code for dbt tutorial