Data Deduplication in People Analytics

Lim Zi Xiang

Problem Statement

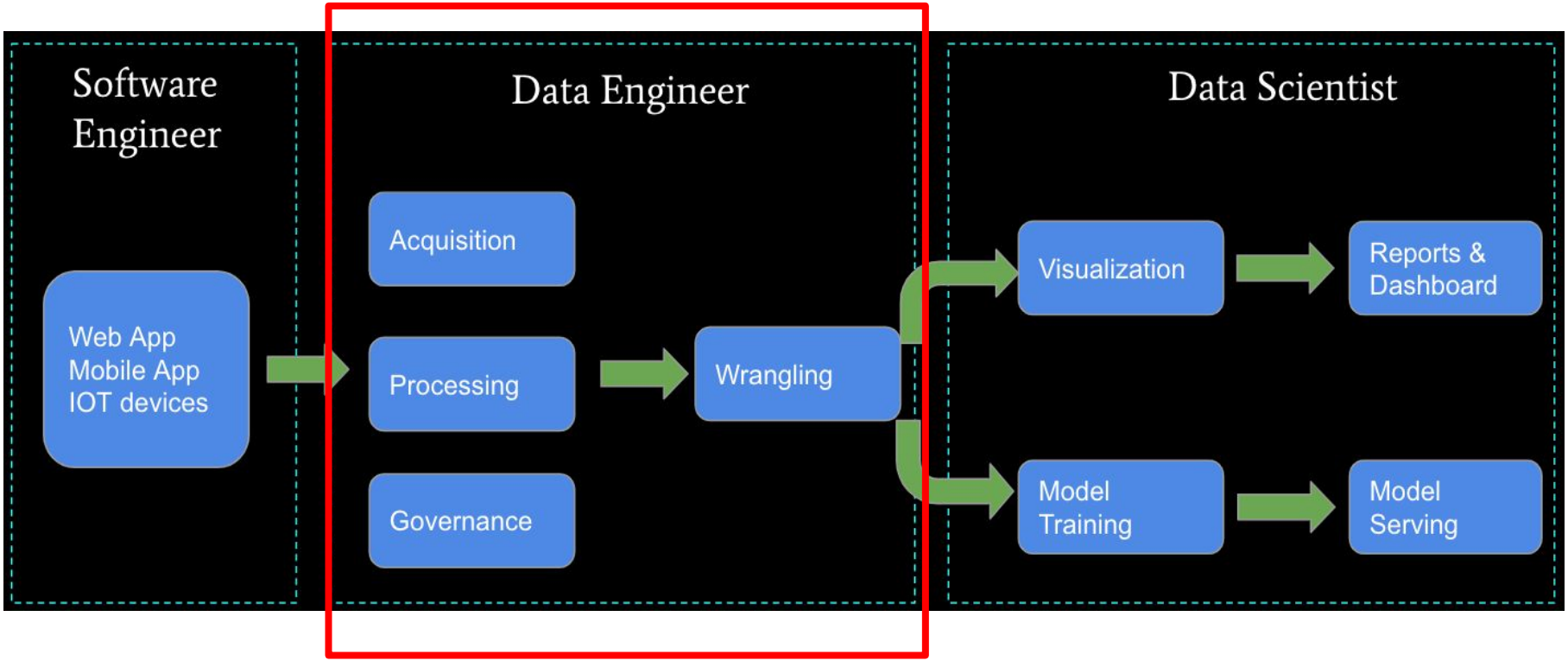
- The organization faces significant challenges in maintaining accurate and reliable people analytics data due to rampant data duplication across various systems and sources.

- Data duplication hampers data integrity, leading to erroneous insights, suboptimal decision-making, and increased operational inefficiencies.

Impact on stakeholders

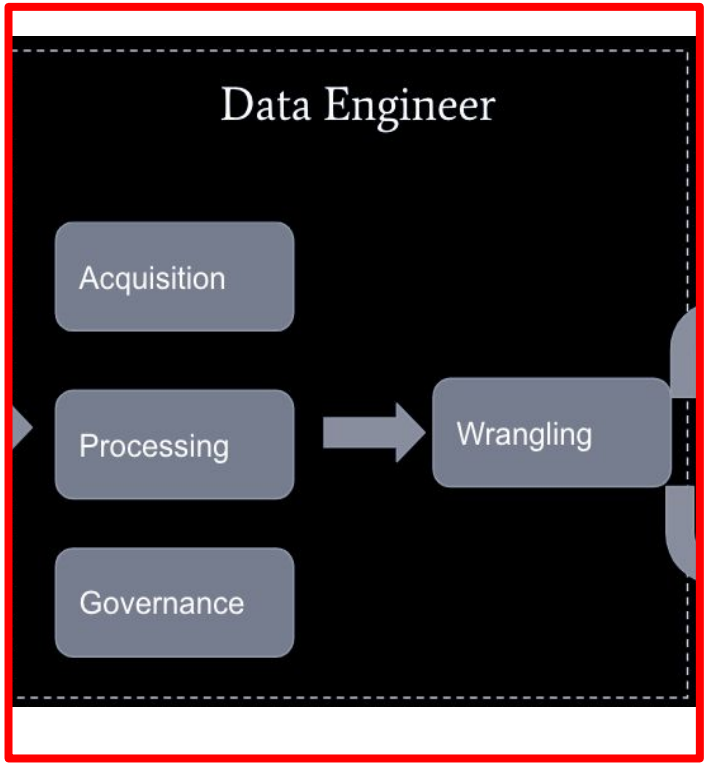
	HR	Finance	IT	Executives
Use Case	Make informed decisions about hiring, employee performance, and talent management.	To assess compensation, budget for hiring, and forecast workforce trends.	Manages data infrastructure and face resource constraints due to excessive data duplication.	Rely on data-driven insights for strategic decision-making and organizational planning.
Impact	Duplicated employee records lead to inaccurate headcounts, duplicate training efforts, and misguided performance evaluations.	Inaccurate headcounts and salary data result in flawed budgeting, leading to financial misallocation and budget overruns.	Non-optimized data storage leads to increased costs, slower data processing, and challenges in data maintenance.	Inaccurate analytics jeopardize the credibility of leadership decisions, potentially leading to missed opportunities and competitive disadvantages.

How can I help?



Chepstov, A. 2019. Data Engineering and Developer Tools for Big Data [Image] https://blog.jetbrains.com/scala/2019/12/13/data-engineering-and-developer-tools-for-big-data/. Retrieved 06 Aug 2023.

How can I help?



- Data Ingestion: Extracting data from various sources and systems.
- Data Transformation: Cleaning, transforming, and enriching data for analysis.
- Data Storage: Efficiently storing data in databases or data lakes.
- Data Integration: Combining data from multiple sources into a unified view.
- Data Quality Management: Ensuring data accuracy, consistency, and completeness.

Possible Solutions and Checks

Solutions	
Azure Data Factory (ADF) and Azure Synapse Pool Stored Procedures	Azure Functions
- An automated scheduled job to run periodically conducted by ADF with Azure Synapse Pool Stored Procedures. - Before load into data warehouse	- Write custom deduplication logic within Azure functions to dedup after fetching data and before loading into blob storage

Checks
Data Quality Tests
Implementation example: - An ADF trigger dependent on the completion of ingestion framework - Thereafter, run a stored procedure to find the number of duplicates in that pipeline run

Impact on stakeholders

1. Enhanced Data Quality
   - more accurate, reliable, and consistent people analytics data

2. Improved Decision-Making
   - confidently make informed decisions

3. Increased Operational Efficiency:
   - reduce IT resource strain and improve data processing speed

4. Cost Savings
   - optimized data storage and maintenance in data infrastructure management

End