Data Engineering Project

Module 1
Fundamentals of Data Engineering

Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech

© Nektaria Tryfona

VIRGINIA TECH

Objectives

- What is Data Engineering?
  - Definition
  - Characteristics
  - Lifecycle of a Data Engineering Project
- The Cross-disciplinary Nature of Data Engineering
  - Data Science
  - Privacy and Security
- Custom Data Engineering Pipeline

© Nektaria Tryfona

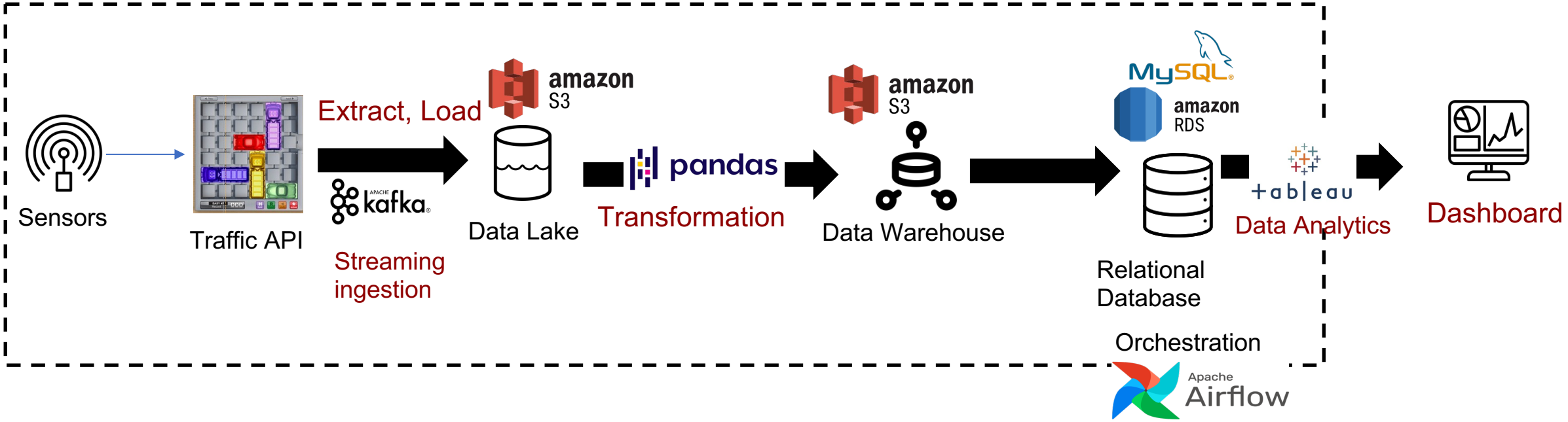
Data Engineering

Definition

© Nektaria Tryfona

What is Data Engineering? - Definition

Data engineering: designing, building, and maintaining the infrastructure and systems that support the collection, storage, and analysis of data



© Nektaria Tryfona

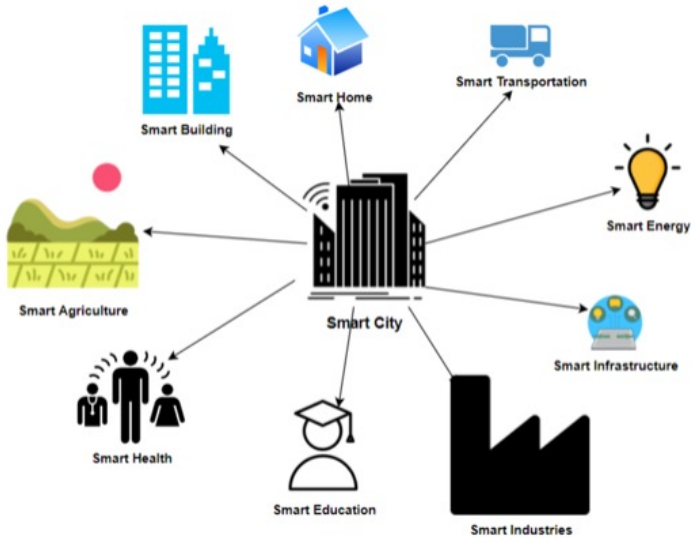
Characteristics of Data Engineering

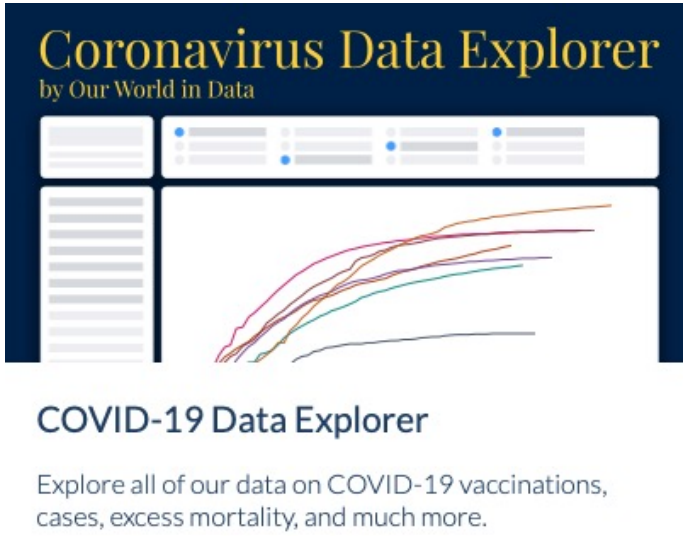
- Data can be messy, coming from different sources, in various ways
  - e.g., streaming vs. batch, in different forms, with “noise”
- Pipelines of tasks from data collection to data cleaning to data-driven solutions
  - orchestration of tasks is crucial
- Not only about data
  - *commercial and open-source* platforms and tools are part of the pipeline
- Different teams may work on the tasks
- Different teams will consume the results

© Nektaria Tryfona

Examples of Data Engineering Projects

- Build a pipeline to monitor Smart IoT infrastructure
- Detect privacy and security breaches in data platforms
- Collect data and analyze passengers' behavioral patterns to target promotions
- Build and deploy a pipeline to store, analyze and visualize Covid-19 data trends in real-time using dashboards



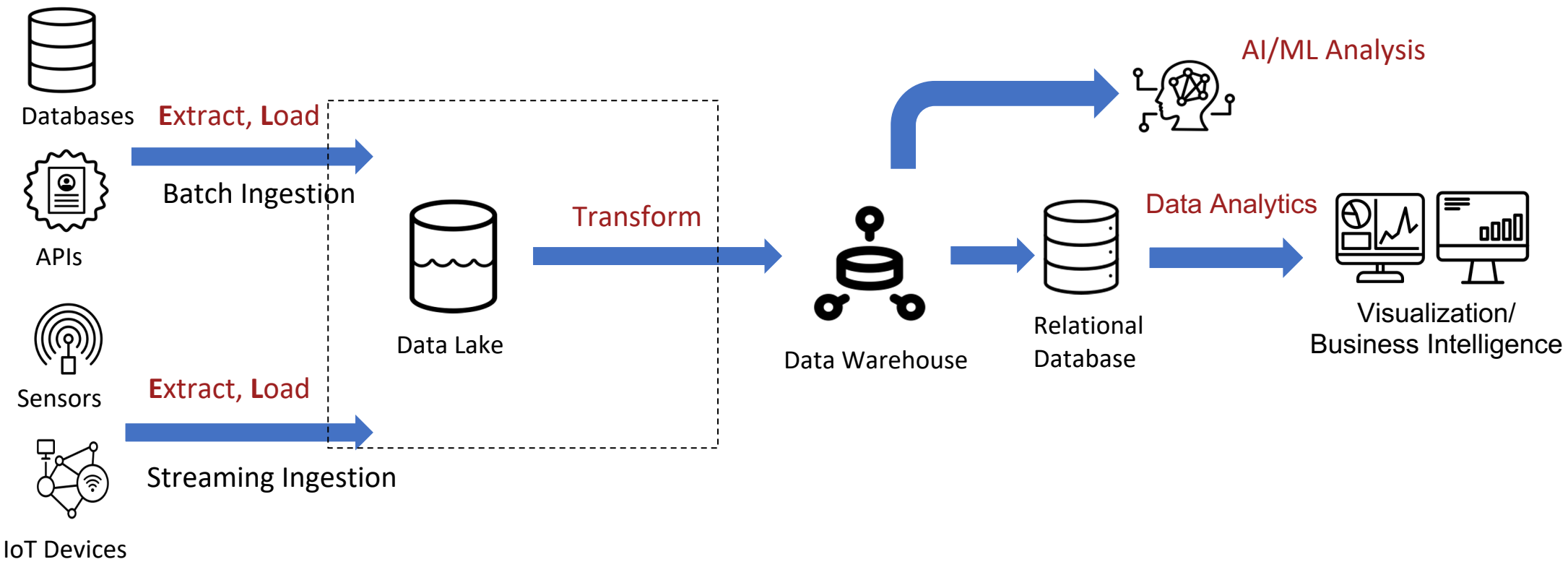


(some) Data Engineering Terms

- data collection
- data preprocessing
- data preparation
- data integration
- data storage
- data management
- data quality
- data governance
- data security

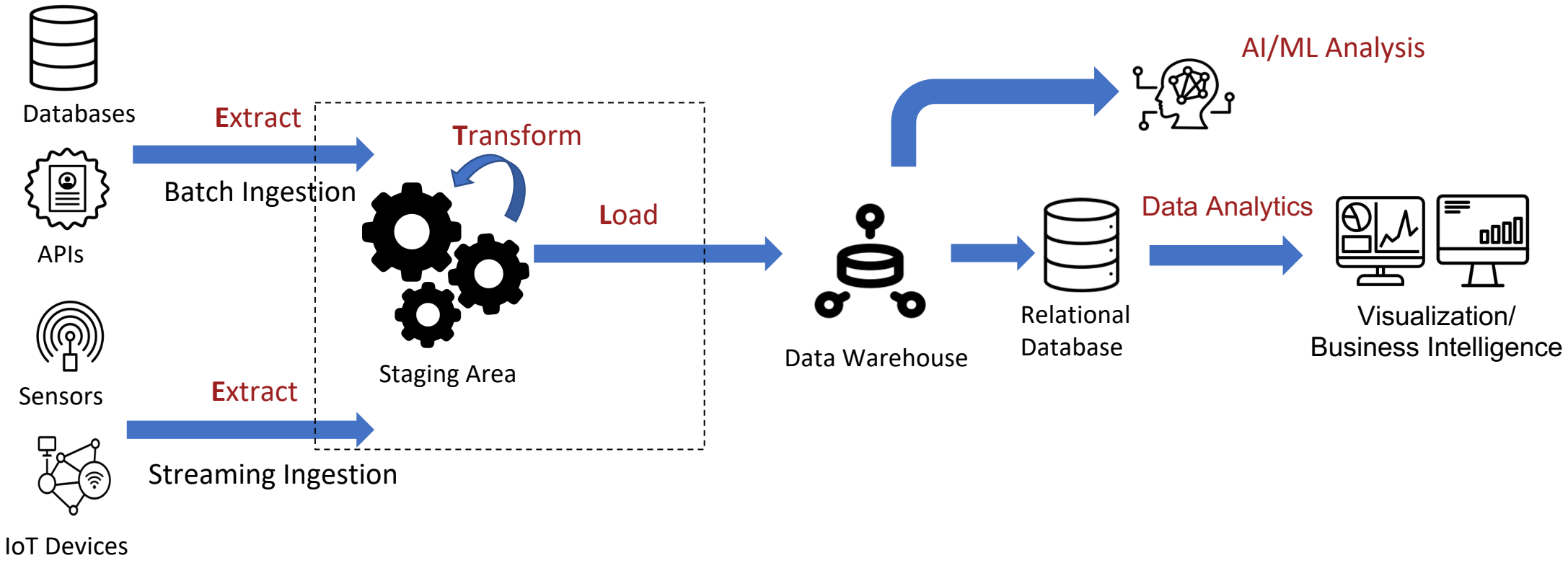
© Nektaria Tryfona

Data Engineering Lifecycle – the ELT Data Pipeline



© Nektaria Tryfona

Data Engineering Lifecycle – the ETL Data Pipeline



© Nektaria Tryfona

Data Transformation Phases

1

Data Pre-processing

- Quality assessment
- Exploratory Data Analysis (EDA)

© Nektaria Tryfona

Data Transformation Phases

1

Data Pre-processing

- Quality assessment
- Exploratory Data Analysis (EDA)

2

Data Preparation

- Data cleaning
- Data editing
- Finding outliers
- Feature extraction
- Data formatting (schema)

© Nektaria Tryfona

Data Transformation Phases

1

Data Pre-processing

- Quality assessment
- Exploratory Data Analysis (EDA)

2

Data Preparation

- Data cleaning
- Data editing
- Finding outliers
- Feature extraction
- Data formatting (schema)

3

Data Integration

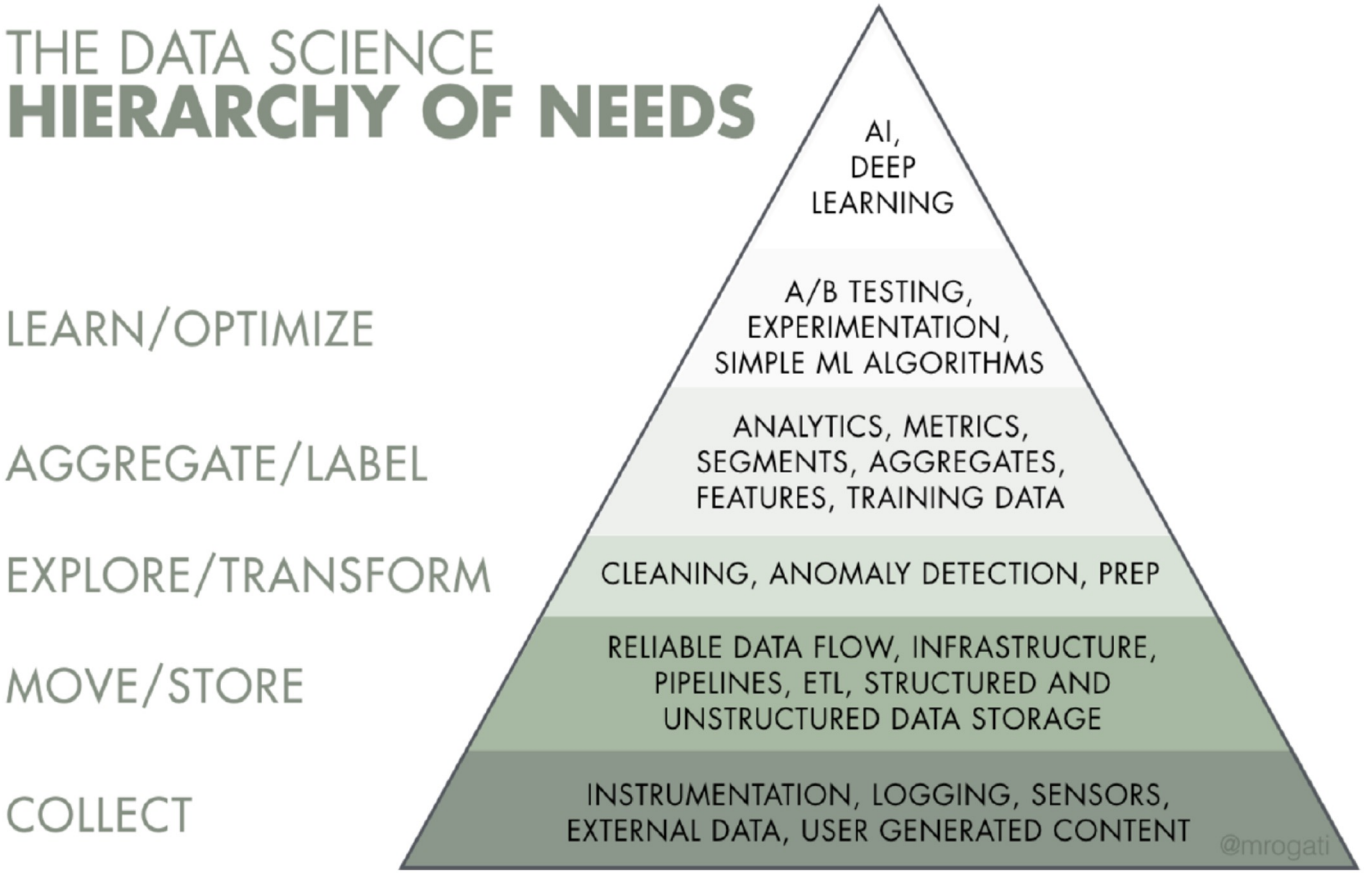
- Bringing together data from multiple sources across an organization to provide a complete, accurate, and up-to-date dataset

© Nektaria Tryfona

The Cross-disciplinary Nature of Data Engineering

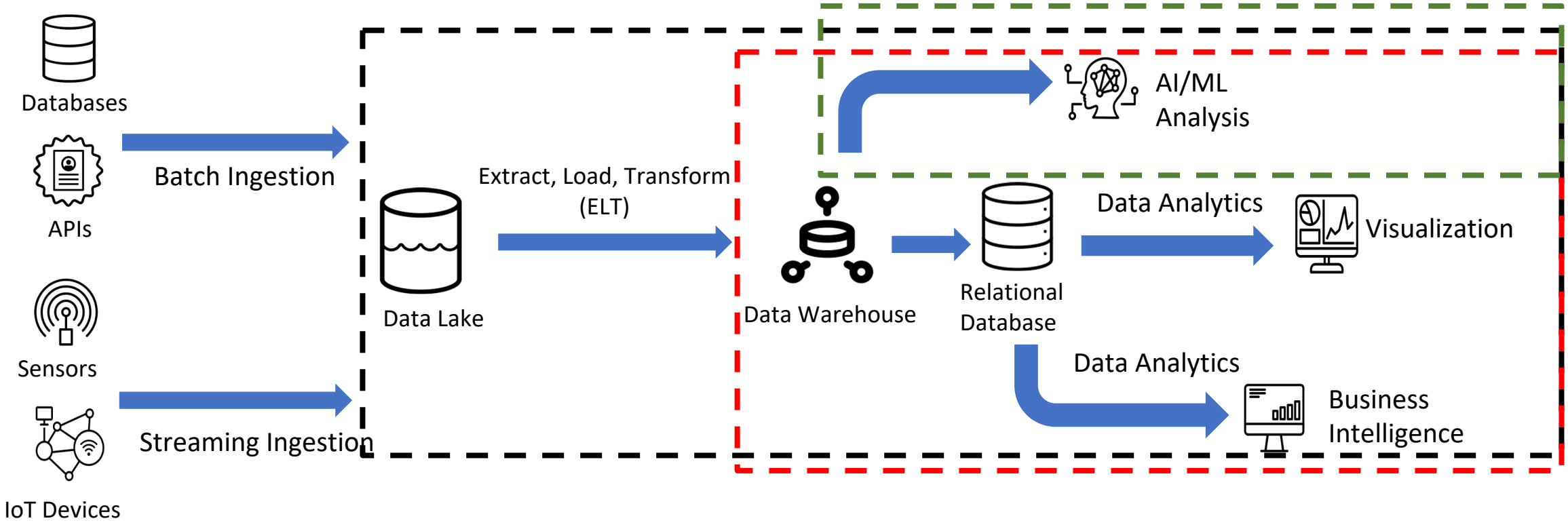
© Nektaria Tryfona

Data Science



https://medium.com/hackernoon/the-ai-hierarchy-of-needs-18f111fcc007
The AI Hierarchy of Needs, Monica Rogati, August 1017

Data Engineering and Data Science



© Nektaria Tryfona

Examples of Data Engineering and Data Science Projects

- Building and maintaining a data lake for storing large amounts of data
- Developing and implementing a data pipeline to automate the process of collecting, cleaning, and loading data from various sources into a data warehouse
- Building and deploying a real-time streaming system to process and analyze data
- Analyzing large datasets to uncover insights and trends that inform business decisions
- Building and deploying a recommendation system to suggest products or content to users
- Using natural language processing to extract insights from unstructured text data

© Nektaria Tryfona

The Role of Data Engineering in Data Science

A core component of today's data infrastructure

In data-intensive projects

- data engineers are responsible for preparing and maintaining the data infrastructure
- data scientists use that (part of the) infrastructure

© Nektaria Tryfona

The Role of Data Engineering in Privacy and Security

Data Engineering is an essential component of privacy and security

- Identify unsafe data access or practices in pipelines
- Monitor, log, and track access to the data pipeline (data repositories, containers, and code)
- Build case studies to reveal data security and privacy blind spots

© Nektaria Tryfona

The Evolution of Data Engineering

90's

- building and maintaining relational databases, data warehouses, and ETL (Extract, Transform, Load) processes to move data between systems

The big data and cloud computing era

- distributed systems (e.g., Spark) to handle large volumes of data across multiple machines
- new technologies (e.g., Apache Kafka) for real-time data processing and stream processing

Nowadays

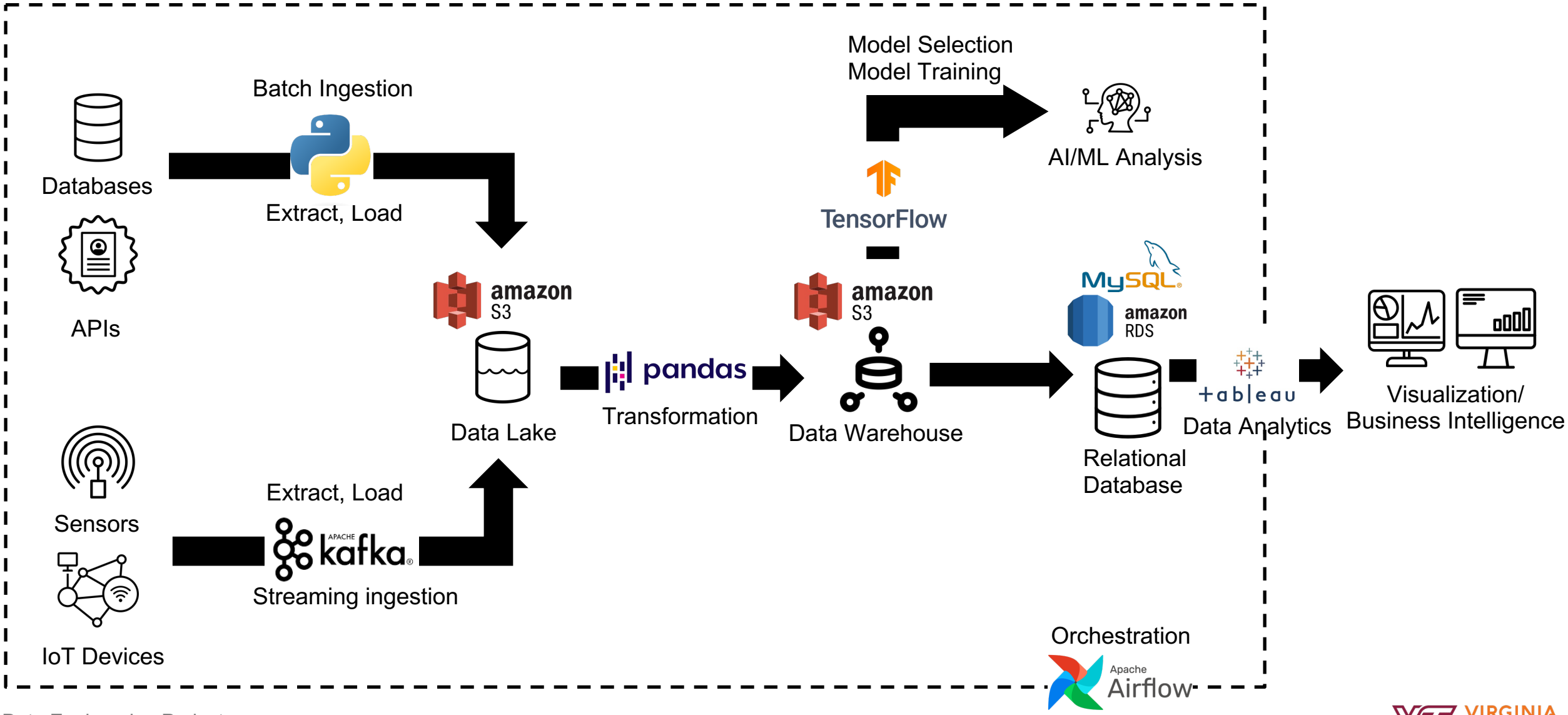
- NoSQL databases and data lake architecture to store and process unstructured and semi-structured data
- more automated and self-service, with the use of data orchestration platforms (e.g., Apache Airflow) to manage data pipelines and data cataloging

© Nektaria Tryfona

Custom Data Engineering Pipeline

© Nektaria Tryfona

Custom Data Engineering Pipeline



© Nektaria Tryfona

Tools

Task/Phase	Tools for Pipeline	Alternatives
Batch Ingestion	Python APIs	talend, Apache Spark
Streaming Ingestion	Apache Kafka	amazon Kinesis
Data Lake	AWS S3 server	Google Cloud, snowflake
Transformation	Python Library: pandas	talend, Apache Spark
Data Warehouse	AWS S3 server	Google Cloud, snowflake
AI/ML analysis	Python Library: Tensorflow	mlflow, Spark MLlib
Orchestration	Apache Airflow	Amazon Glue
Database Management	AWS RDS (MySQL)	Azure SQL
Visualization/BI	Tableau	Grafana

Data Engineering Project – How?

- Focus on data, tools/platforms, and tasks
  - Use and build customized pipelines
  - Control the orchestration, automation, and schedule of the pipeline tasks
  - Control the flow of data

© Nektaria Tryfona

Summary

- What is Data Engineering?
  - Definition
  - Characteristics
  - Lifecycle of a Data Engineering Project
- The Cross-disciplinary Nature of Data Engineering
  - Data Science
  - Privacy and Security
- Custom Data Engineering Pipeline

© Nektaria Tryfona

Data Engineering Project

Module 1
Fundamentals of Data Engineering

Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech

© Nektaria Tryfona

VIRGINIA TECH