Data Engineering Project

Module 5
Data Transformations and Data Provenance

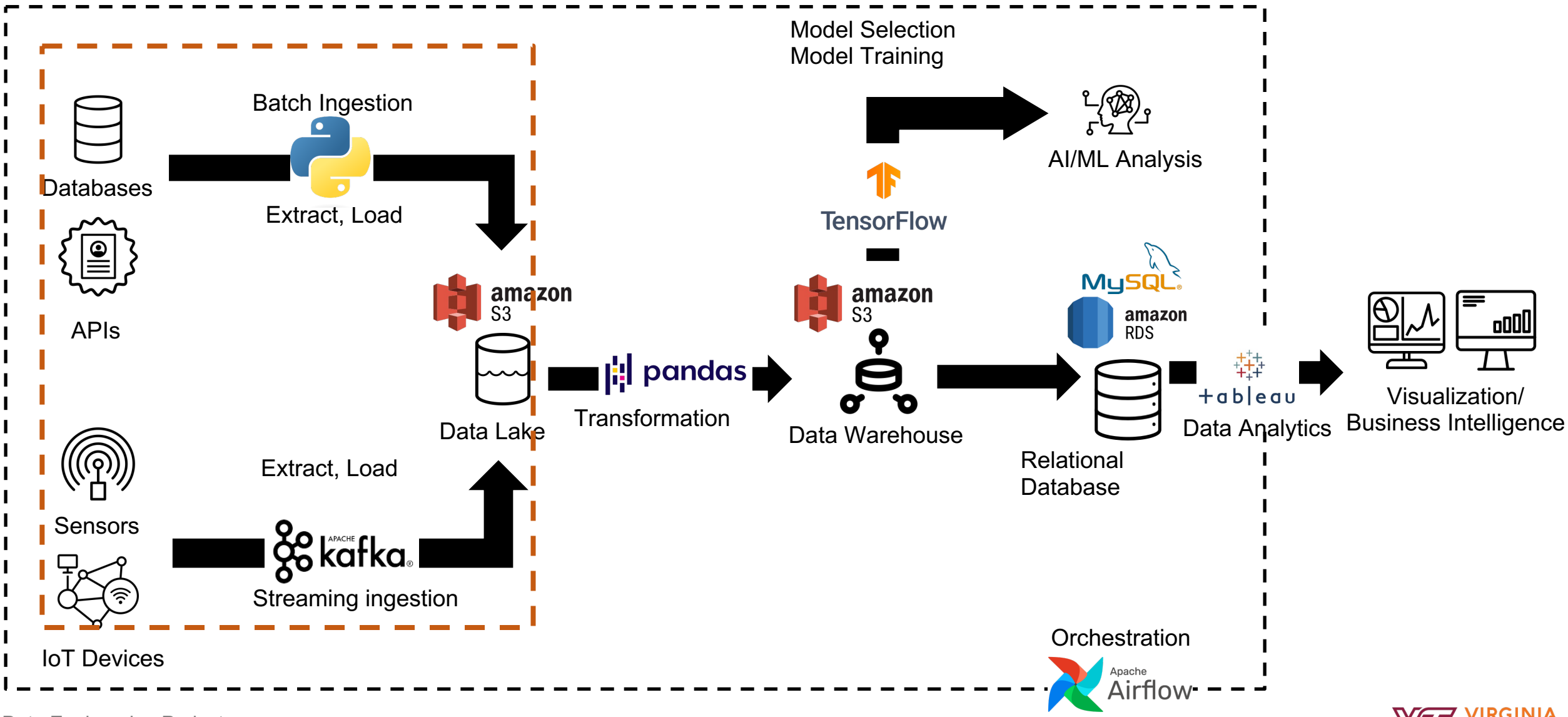
Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech

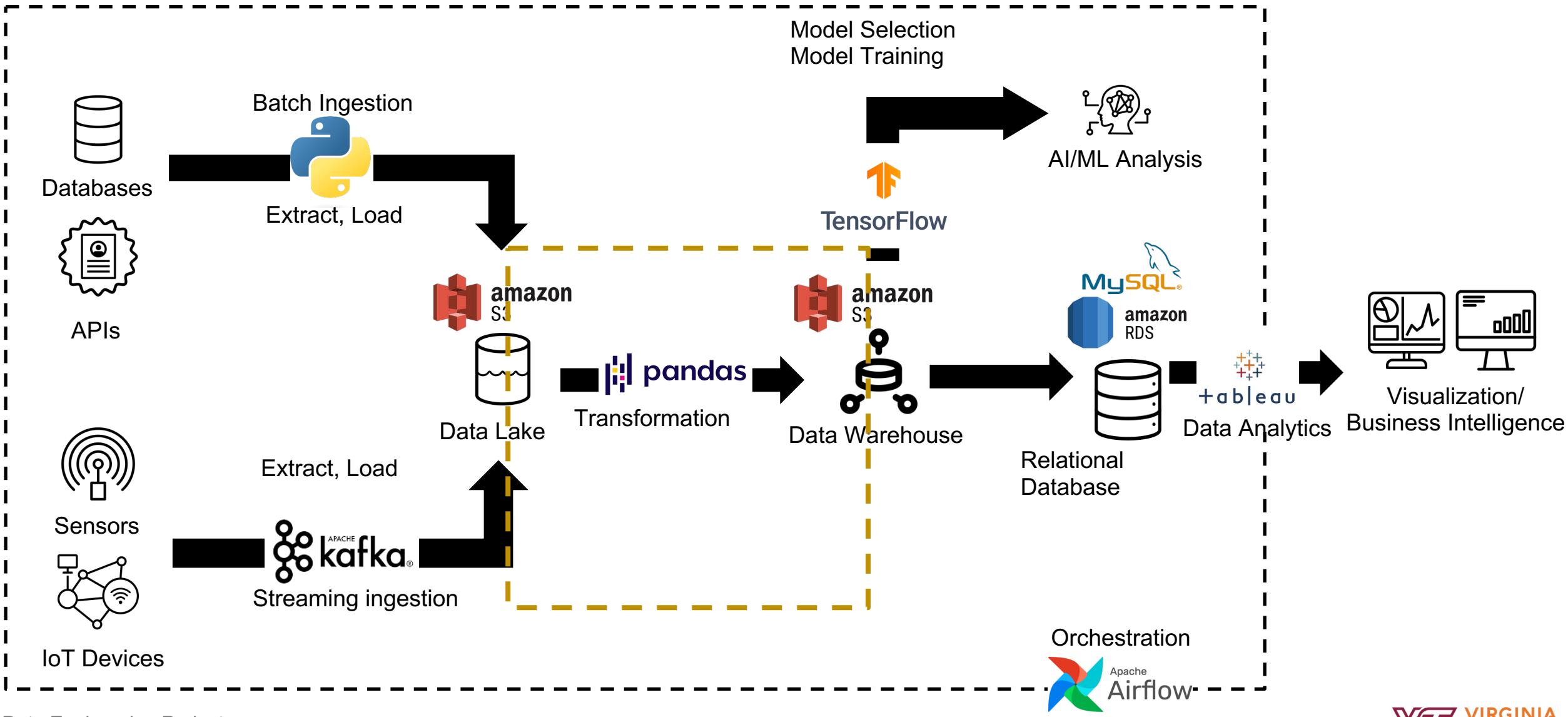
Objectives

- Data Transformations and Data Storages
- Data Provenance and Data Lineage
- The Role of Data Provenance in Data Governance
- Provenance in Cloud Services and Pipelines

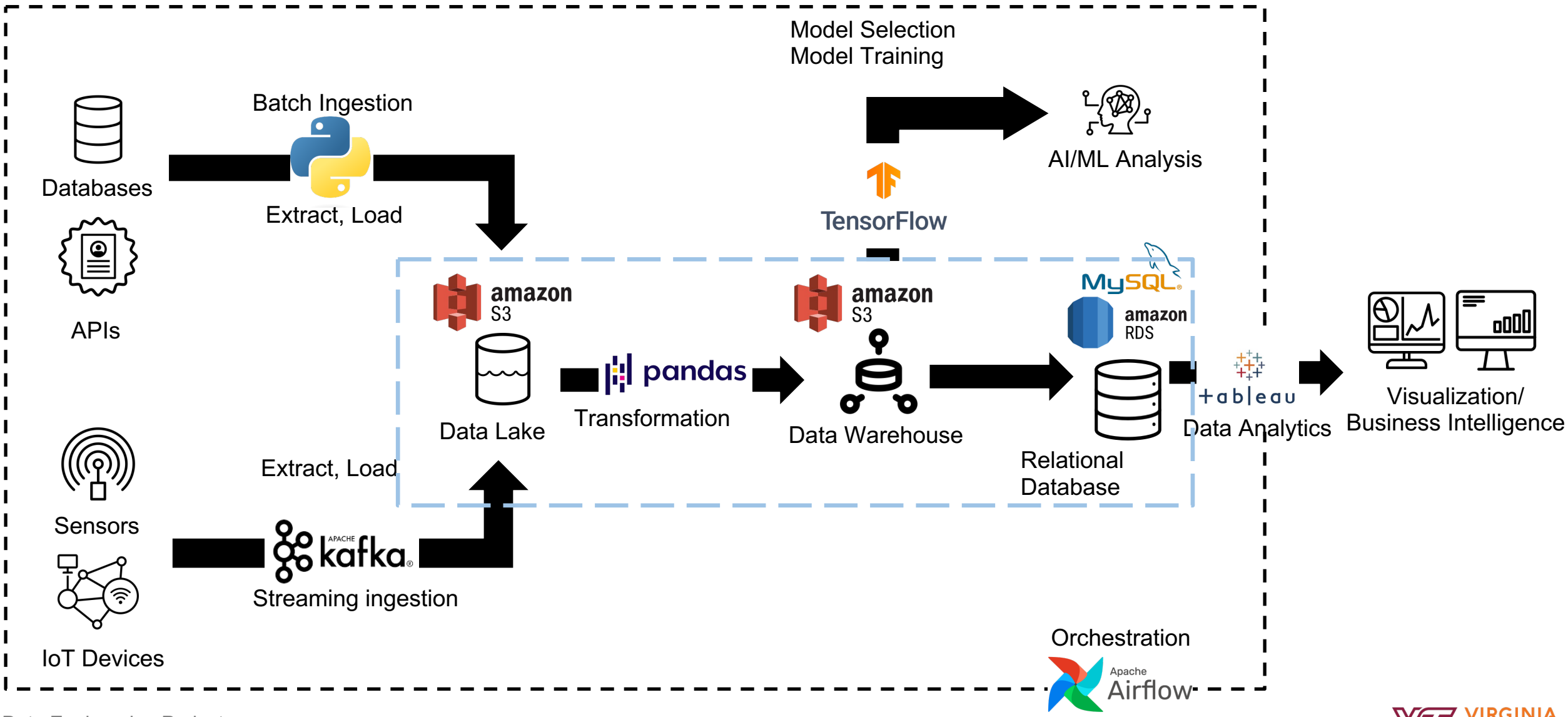
Custom Data Engineering Pipeline



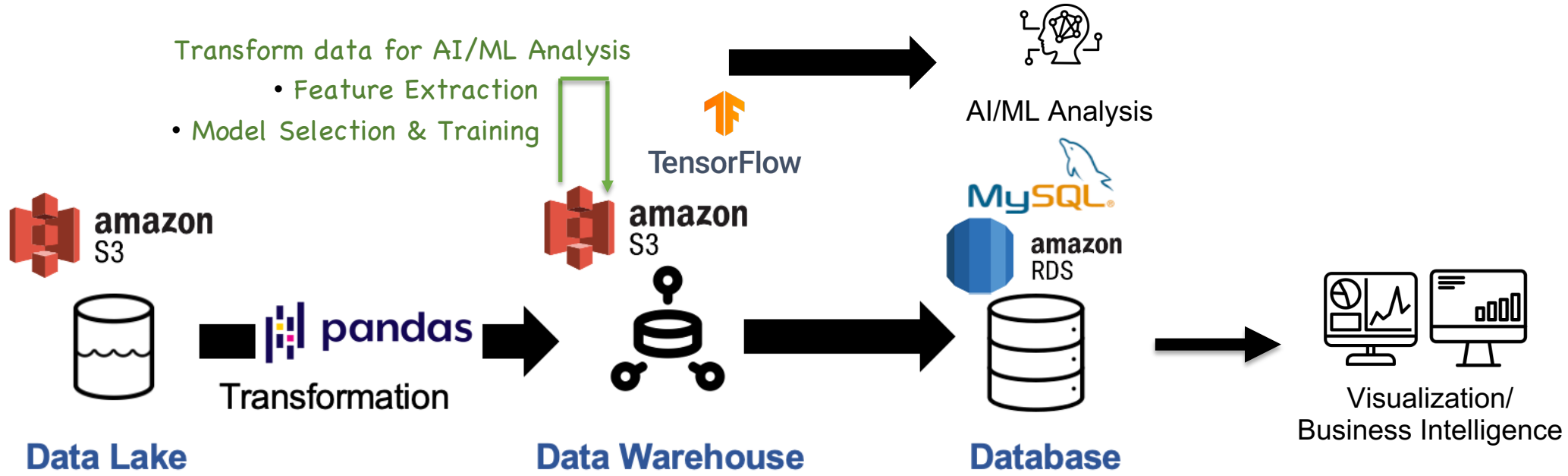
Custom Data Engineering Pipeline



Custom Data Engineering Pipeline



Transformations can happen in more that one task



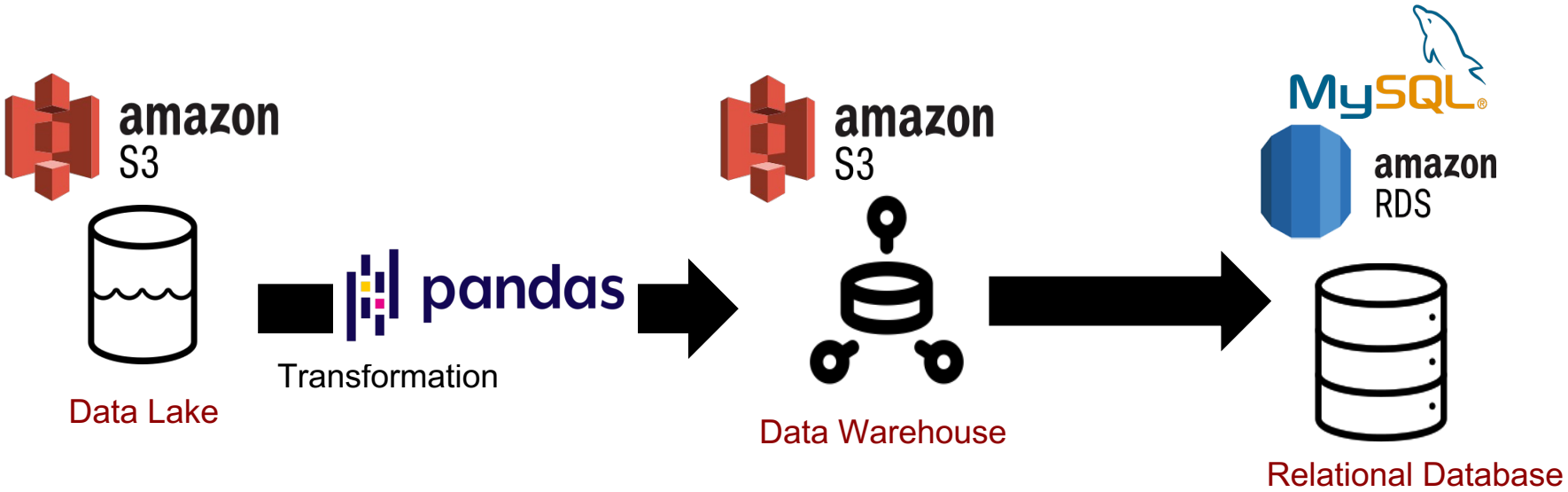
Land data in the data lake

Explore, transform data
- Remove duplicates
- Filter unwanted outliers
- Handle missing data

Select data & move it in the data warehouse

Transform data to answer queries
- Database Schema
- Data Visualization

Data Storages



Data Lake

A repository of data coming from various sources and stored in its original, raw format

About data:

- structured, semi-structured, unstructured
- no need to be transformed in order to be added to the data lake
- data can be added (or "ingested") incredibly efficiently without upfront planning

Usage:

- data analysis to gain insights
- sometimes data lakes are used as 1st staging

What is so special about data lakes?

- their ability to store data in a variety of formats including JSON, BSON, CSV, TSV, and more

Data Warehouse

A system that stores highly structured information from different sources

About data:

- usually current and historical data from one or more systems

Usage:

- combine data from different sources in order to analyze the data
- look for insights
- prepare data for Machine Learning services and Business Intelligence (BI)

What is so special about data warehouses?

A data warehouse is a giant database that is optimized for analytics

DataBase Management System (DBMS) - (Database)

An organized collection of data or information

About data:

Data is organized in rows and columns

Usage:

to store, search and report on structured data usually from a single source

What is so special about Databases:

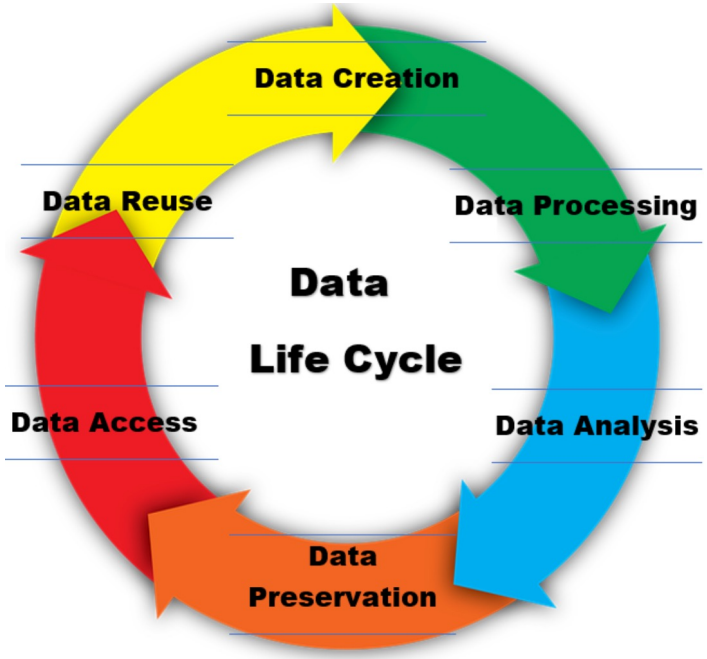
- are incredibly versatile; can be accessed on computers, tablets, and even mobile devices
- allow categorization and structuring of available data
- typically accessed electronically and are used to support Online Transaction Processing (OLTP)

Summary of Data Storage Characteristics

	Data Lake	Data Warehouse	Database
Data Source	Collected from many sources	Collected and normalized from many sources	Data captured as-is from a single source
Data Schema	Written at the time of analysis	Designed prior to the data warehouse implementation Can be written at the time of analysis Denormalized schemas, e.g., Star schema or Snowflake schema Tabular format	Highly normalized, static schemas Tabular format
Data Quality	Any data that may or may not be curated	Highly curated data	Data of high quality

Data Provenance and Data Lineage

What is Data Provenance? - Definition



Data Provenance: the process to track

- the origin of our data
- along with what type of transformations performed on them
- and the users who altered them

Purpose of data provenance:

- establish the trustworthiness, authenticity, and reliability of the data
- ensure that it is fit for its intended purpose
- track changes, if needed

Data Provenance

- Where did the data originate from?
- How was this data generated and processed?
- **How was data changed over time and by which process?**
- **Who was responsible for data modifications?**
- When was the change made?
- Why was a change made and what is the context behind it?
- Is this data trustworthy?
- Is this data authentic?
- What other data were used to calibrate, validate, and process these data?

Example: Building a Healthcare Custom Pipeline

- Collect patient data to improve health outcomes
- Analyze clinical data to improve medical research
- Build, train, and deploy an ML model towards prediction of patient outcomes, readmission rate, or disease progression

For integrity, privacy and security of PHI, need to document:

- the source of the data
- the handling and processing/transforming of the data
- the accessing, sharing and handling of the data and more

https://aws.amazon.com/blogs/architecture/building-a-healthcare-data-pipeline-on-aws-with-ibm-cloud-pak-for-data/

Data Lineage - Definition

Tracks the history of the movement of data from its source to its destination, including any transformations and processing

Traces the path of data through the various stages of its life cycle:

- Where/how it originates?
- What changes it undergoes?
- Where did it move over time?

Why is Data Lineage Important?

It contributes to:

- identifying any errors, inconsistencies, or biases
- establishing the relationships between different data sources, transformations, and outputs

Example: Tracking the movement of patient data from its original source, such as a transactional system, to a data warehouse

The lineage would include information about any transformations or processing that occurred during the data's movement, such as

- data cleaning
- aggregation
- enrichment
- any queries or reports that were generated using the data

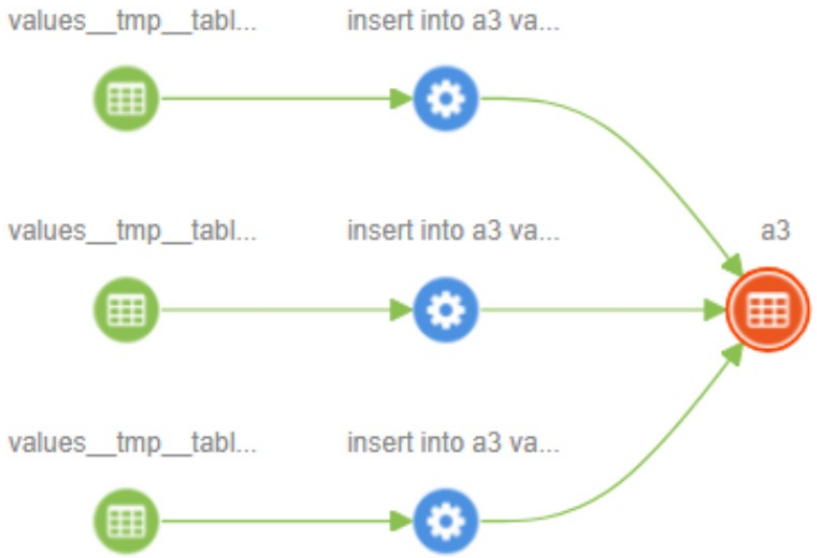
How is it different from Data Provenance?

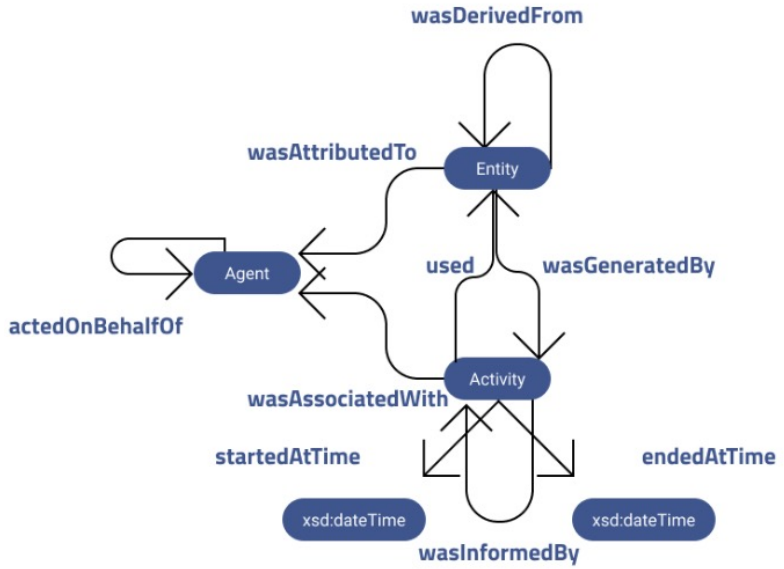
- Data provenance: high-level view and documentation of the project (or system), focusing on objects, entities, and processes
- Data lineage: details of data preparation, cleaning, and transformation
- Both relate to data quality

Data Provenance Methods and Lineage Tools

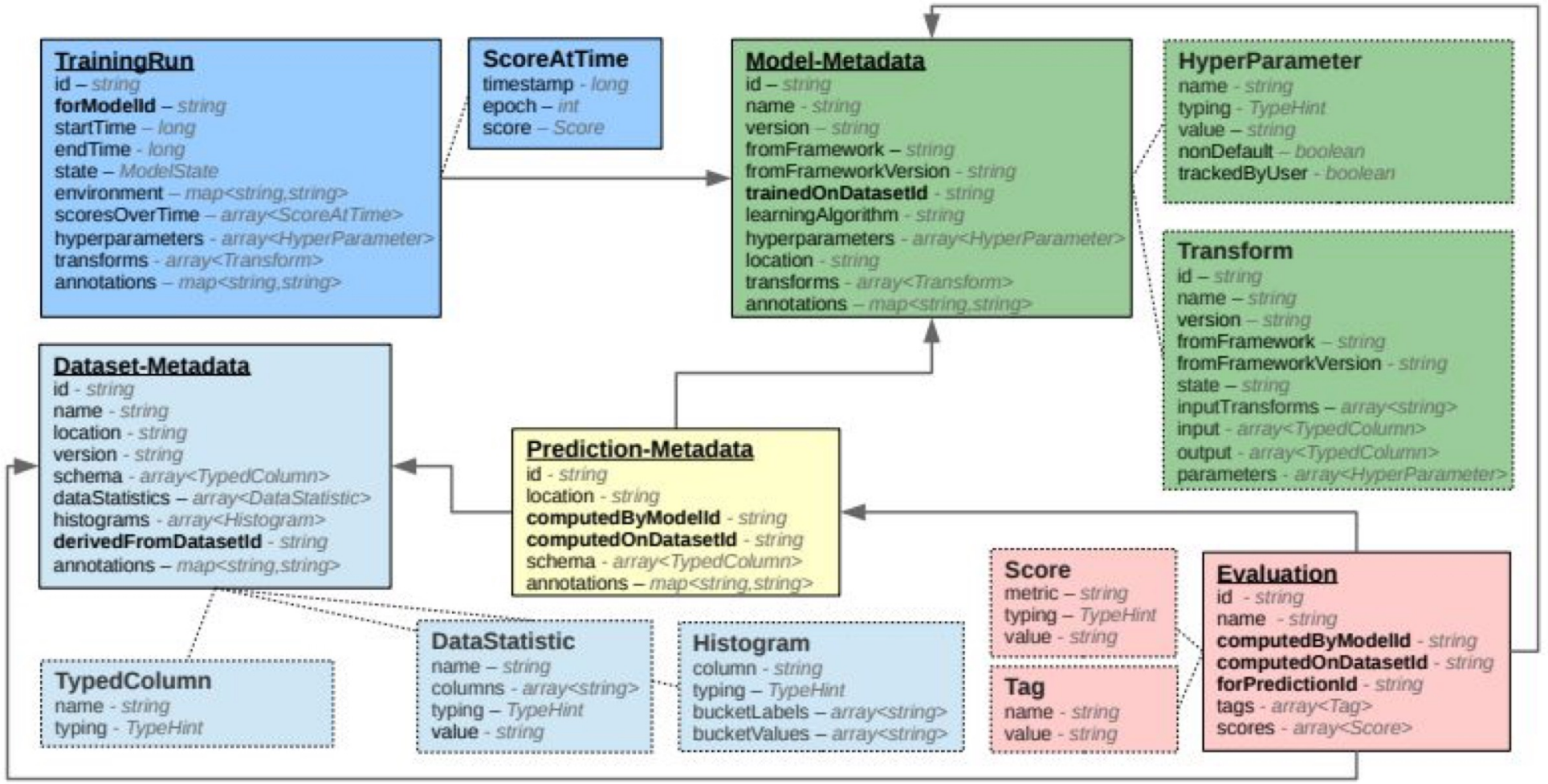
Graphs, metadata models, platforms that map information flow, …

- the specific requirements of the application
- the complexity of the data
- the level of security and privacy needed
- the desired level of transparency and accountability





Example: Storing Provenance Data



https://allprojects.github.io/SE4AI/assets/slides/sose2020/10_data_provenance_and_reproducibility_thursday.pdf

The Role of Provenance in Data Governance

Data Governance: rules and processes imposed on maintaining data in a company

- Who has ownership of the data?
- Who can access what data?
- What security measures are in place to protect data and privacy?
- How much of our data is compliant with new regulations?
- Which data sources are approved to use?

Provenance and Lineage in Cloud Services and Pipelines

- Data in the Cloud should be secure
- Provenance increases the quality and value of data in the cloud
- Trustworthy cloud provenance is a fundamental requirement

Recently

- Data quality, provenance, and lineage became part of data pipelines
- Apache Airflow started implementing Data Lineage methods

Summary

- Data Transformations and Data Storages
- Data Provenance and Data Lineage
- The Role of Data Provenance in Data Governance
- Provenance in Cloud Services and Pipelines

Data Engineering Project

Module 5
Data Transformations and Data Provenance

Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech