Data Engineering Project

Module 8
Security Issues
in Data Pipelines and the Cloud

Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech

Objectives

- Security Factors related to the Pipeline
- Security Factors related to Cloud (and Cloud Computing)
- Best Practices

How vulnerable is our Data Pipeline?

Components of the data pipeline are local and in the Cloud

- Different teams may work on the tasks
- Different teams will consume the results

- Data and processes in the pipeline should be secure
- Attacks should be prevented

Security factors related to Cloud Computing

A massive amount of data and resources are in the Cloud →

A massive concentration of risk

- loss from a single breach can be significantly larger
- the concentration of "users" (engineers and consumers) represents a concentration of threats

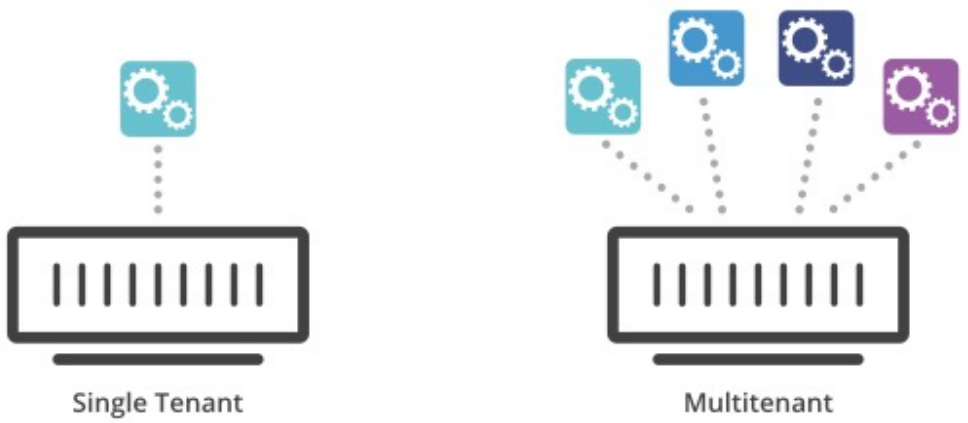
Many possible layers of access control

- e.g., access to the cloud, to servers, to services, to databases, to VMs, and to objects within a VM
- Some of these will be controlled by the provider and others by the "user"

The Concept of Multi-tenancy in the Cloud

Multitenancy: multiple customers of a cloud vendor using the same computing resources

- Cloud customers are not aware of each other
- Their data is kept totally separate



Cloud Computing brings New threats

Pros

- Lower cost
- Better use of resources

Cons

- Multiple independent users share the same physical infrastructure → an attacker can legitimately be in the same physical machine as the target
- Cybercriminals can take advantage of multiple access points to exploit systems vulnerabilities

Malicious Insiders

Client site

- Learn authentication information
- Gain control of the Virtual Machines

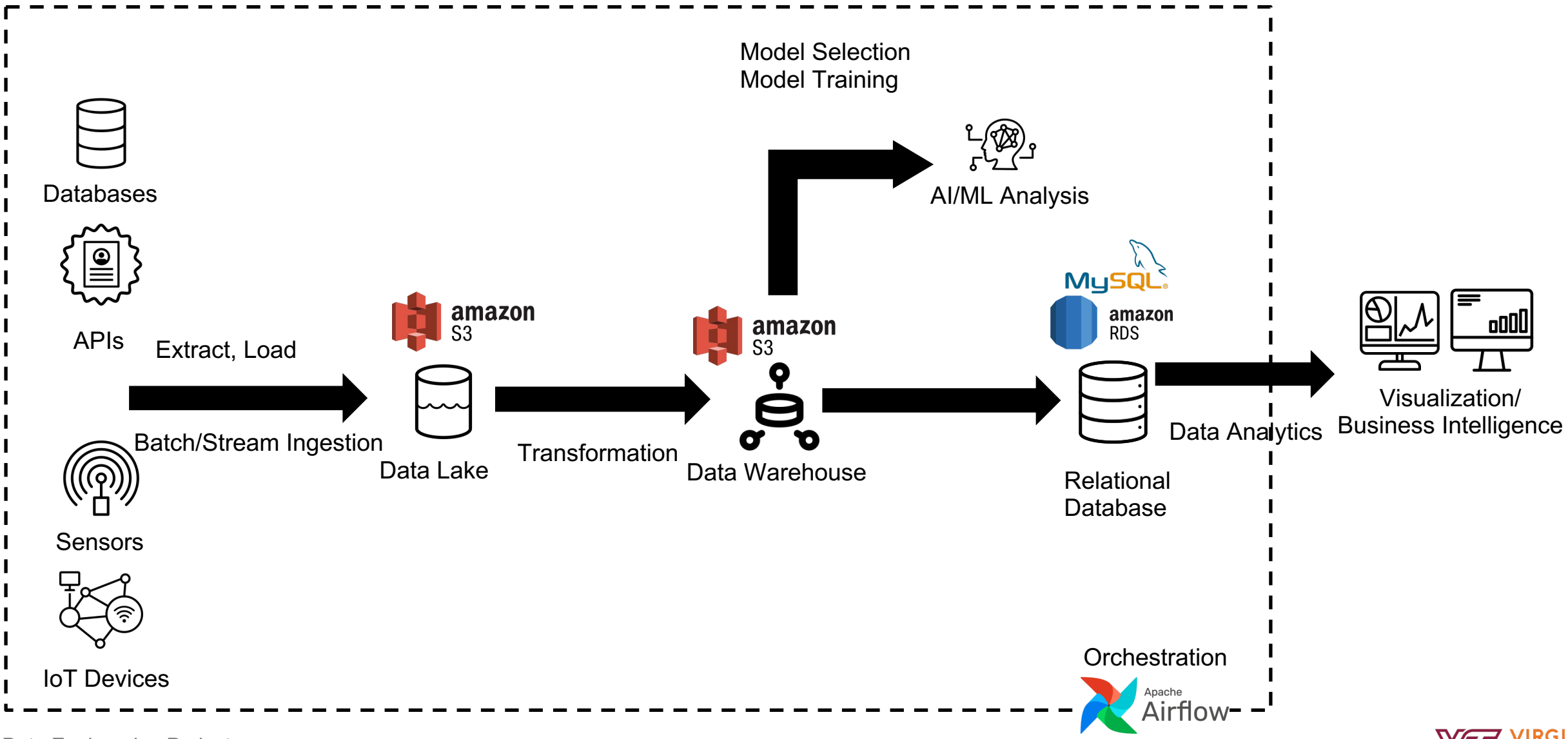
Cloud provider

- Can read unencrypted data
- Can possibly peek into VMs, or make copies of VMs
- Can monitor network communication, application patterns

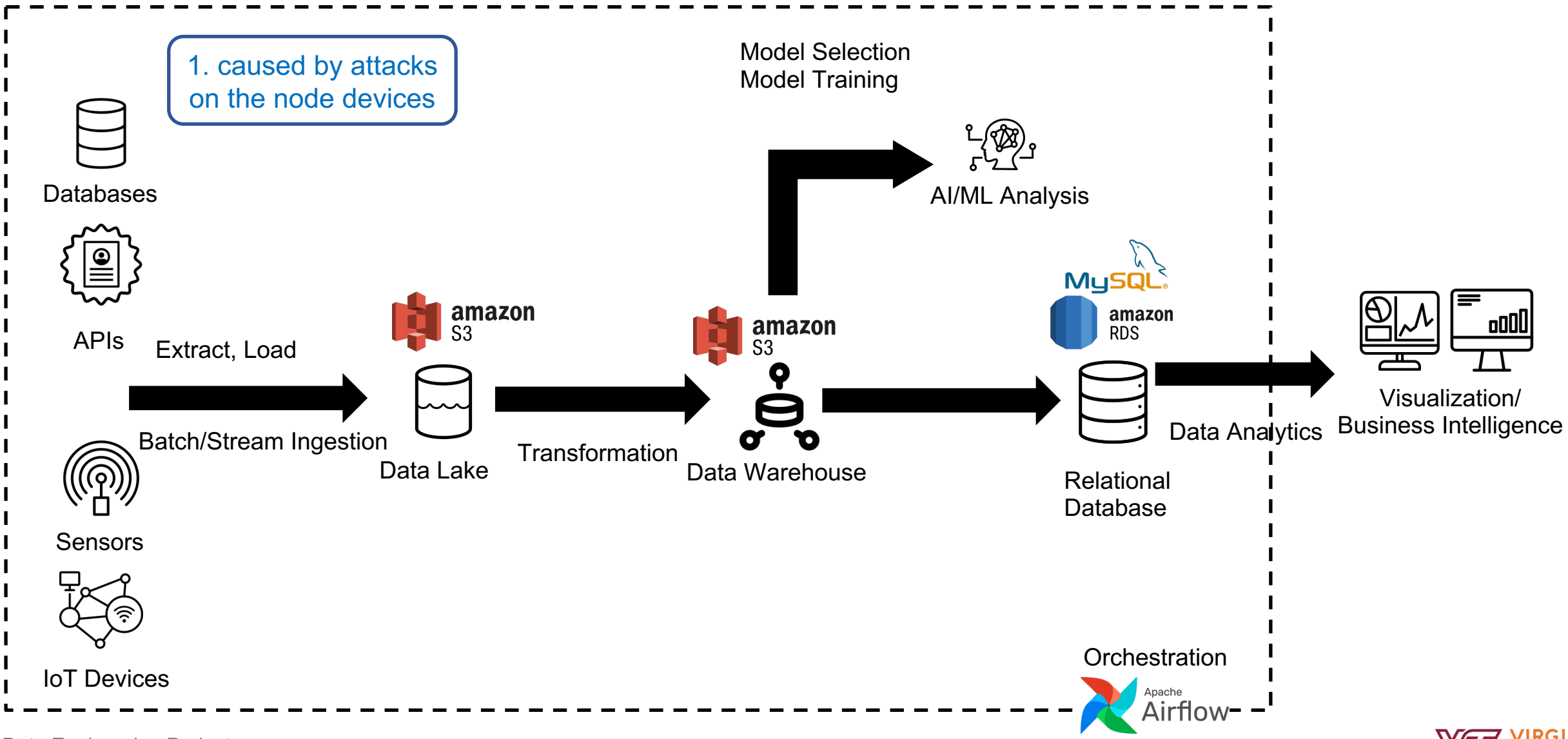
Outside Attackers

- Listen to network traffic (passive)
- Insert malicious traffic (active)
- Probe cloud structure (active)
- Launch Denial-of-Service

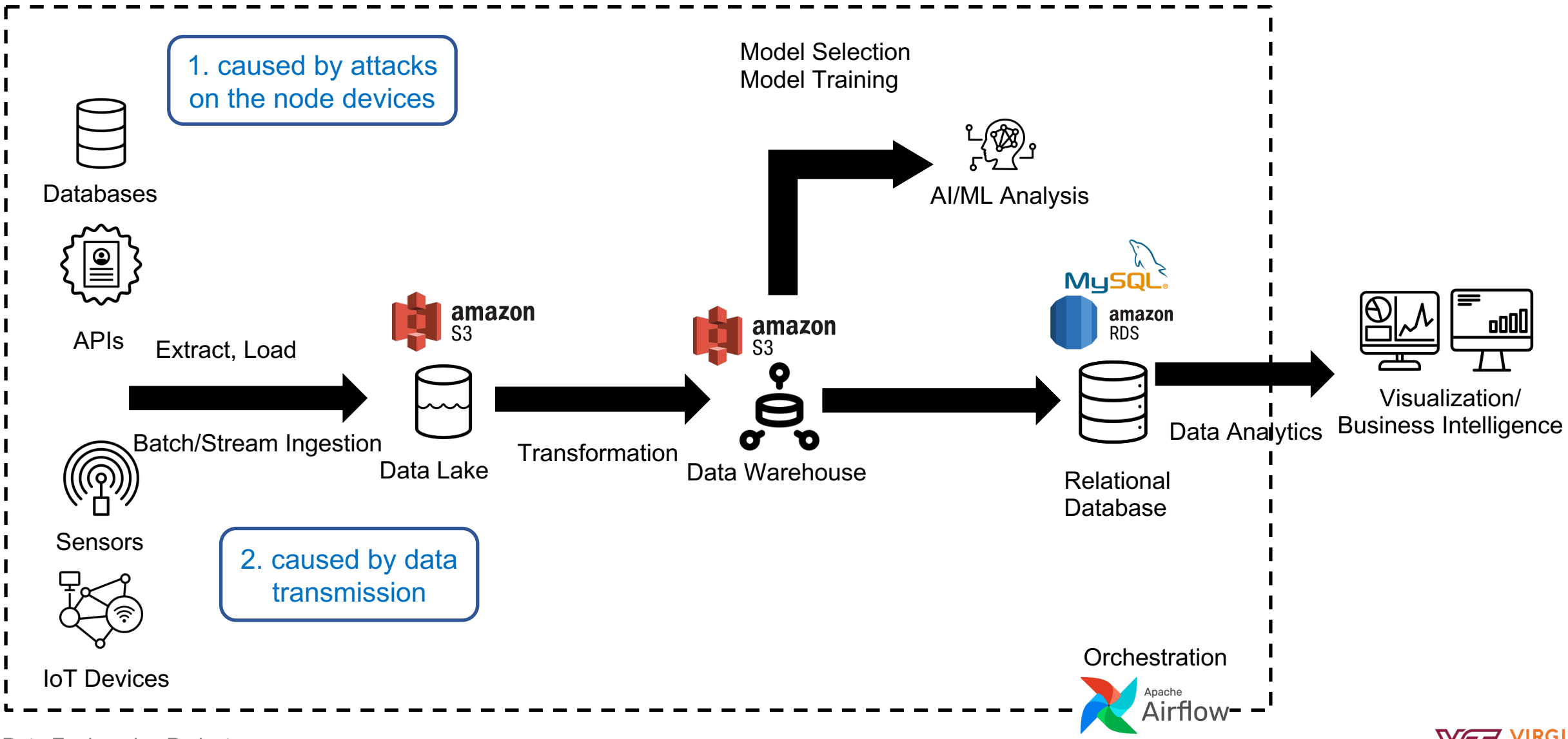
Security Issues in the Pipeline



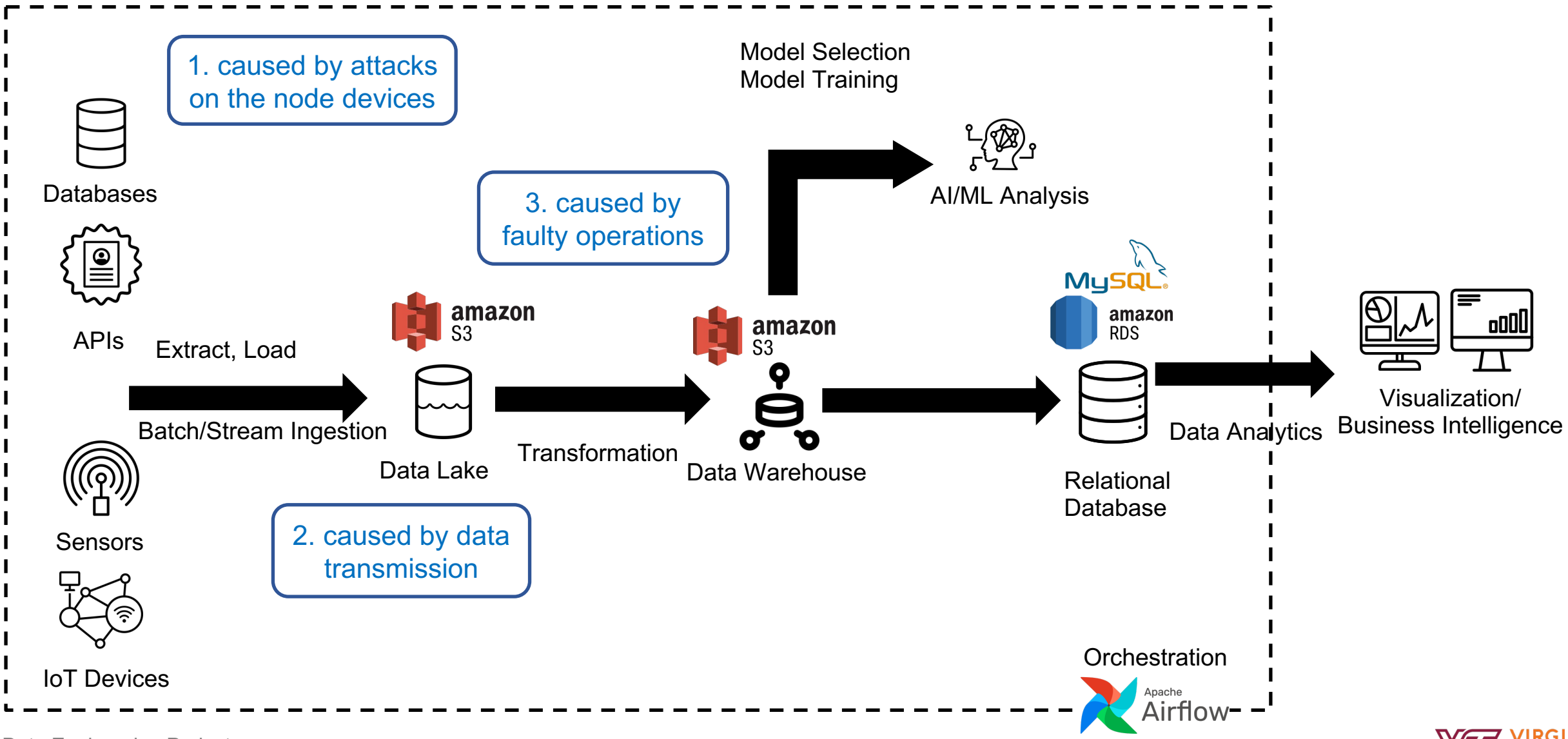
Security Issues in the Pipeline



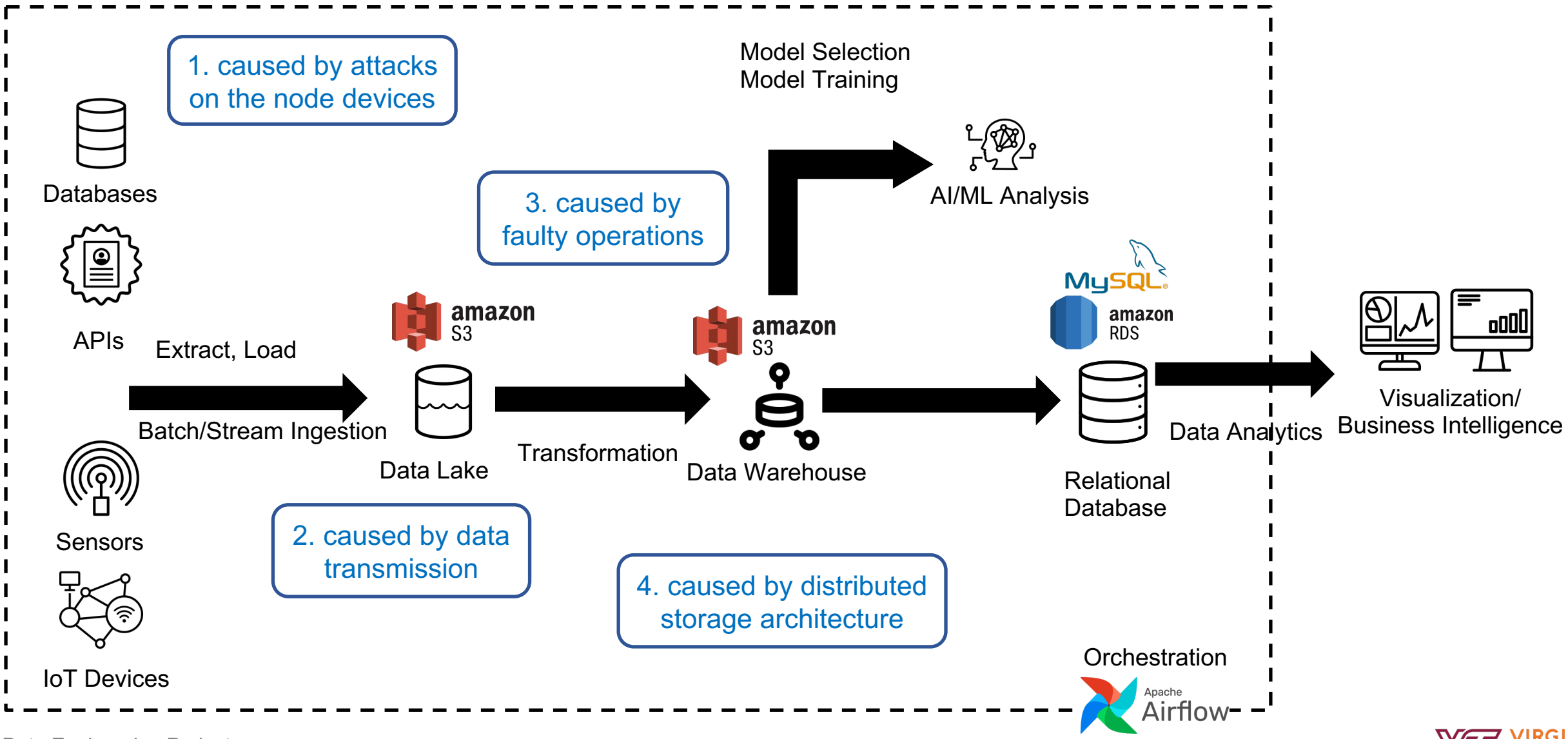
Security Issues in the Pipeline



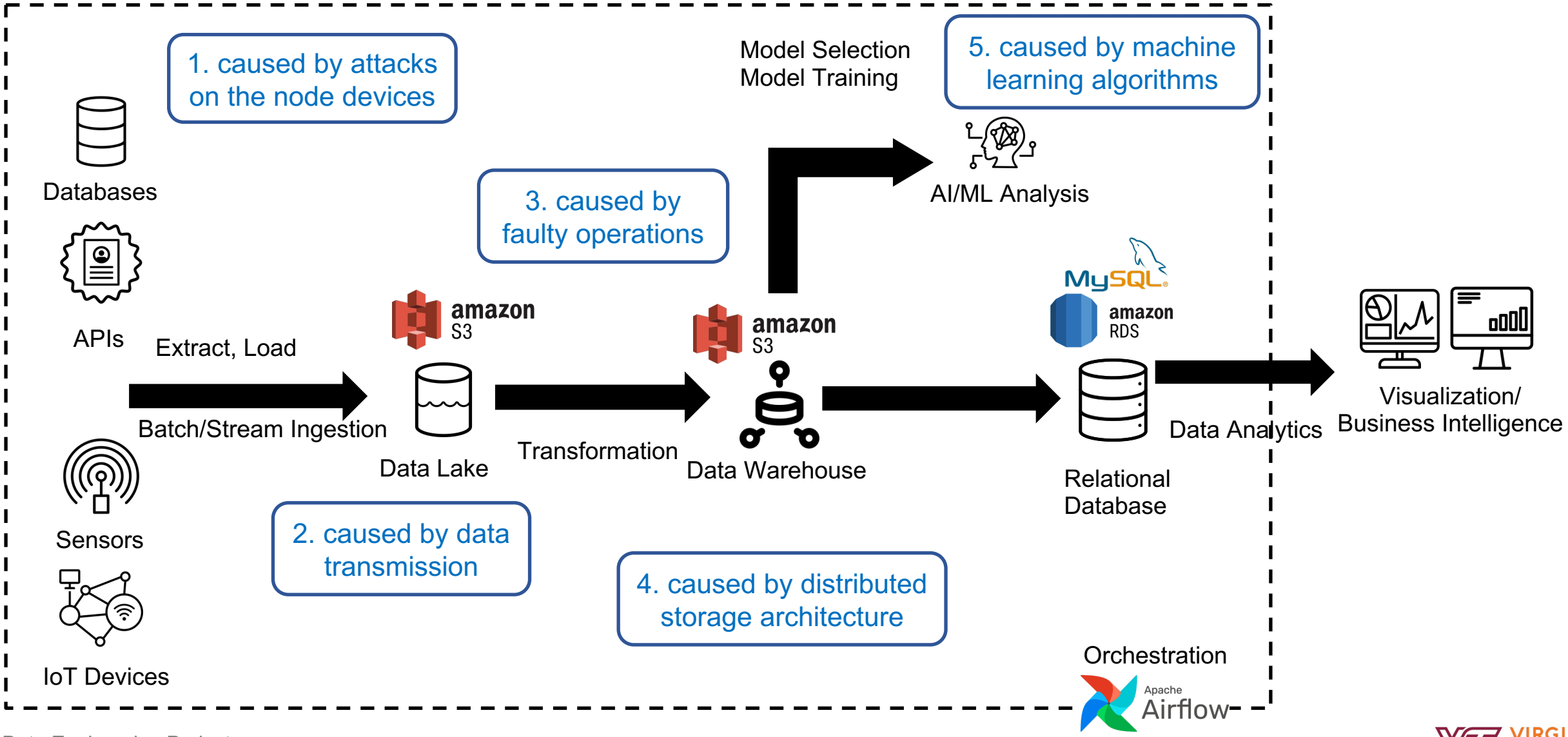
Security Issues in the Pipeline



Security Issues in the Pipeline

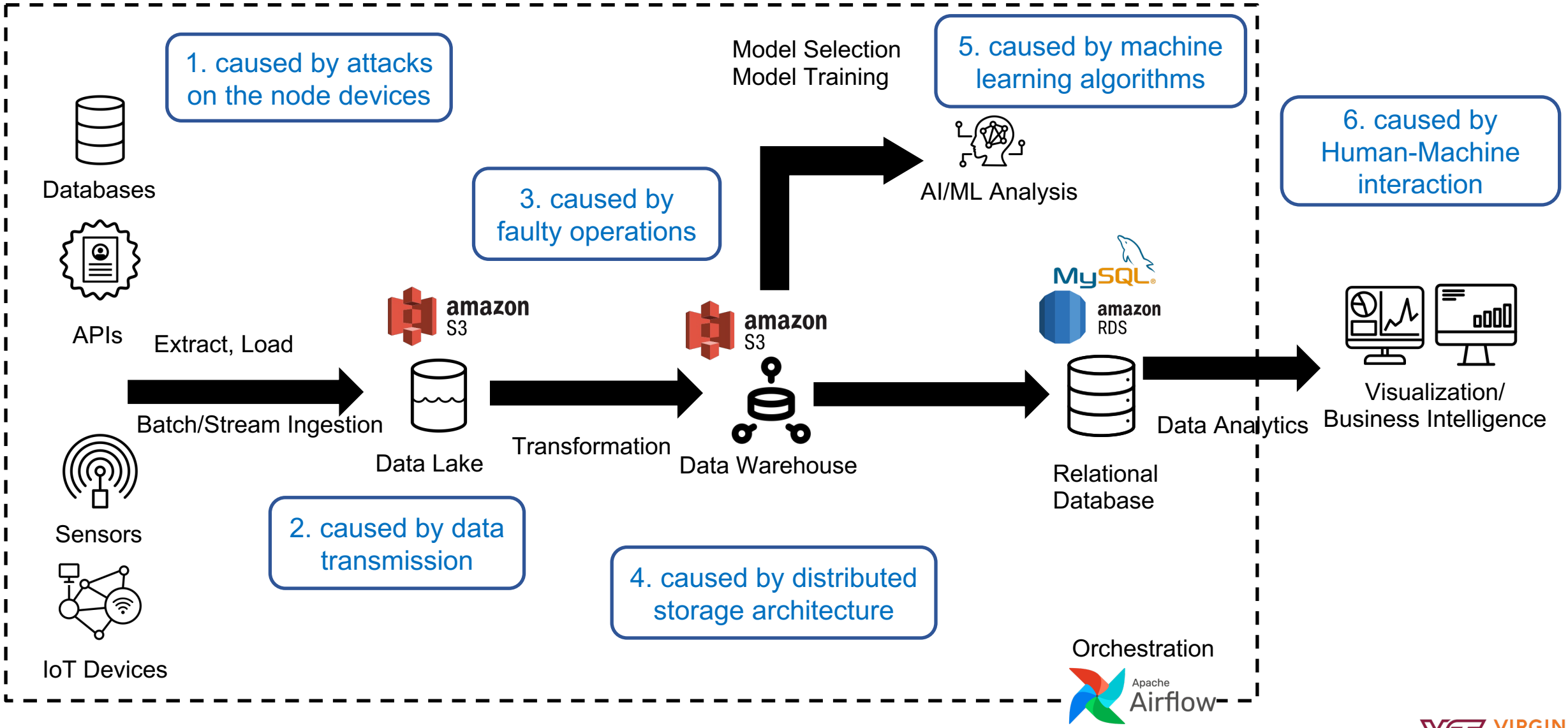


Security Issues in the Pipeline



Adapted from Nguyen, A. et al., (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. IEEE conf. on computer vision and pattern recognition.

Security Issues in the Pipeline



How to Secure our Pipelines – Best Practices

Applied research and commercial platforms solutions

e.g., AWS Glue supports encryption

How to Secure our Pipelines – Best Practices

Applied research and commercial platforms solutions

e.g., AWS Glue supports encryption

Applied research and commercial platforms solutions

e.g., a database running inside a private subnet

e.g., network segmentation

How to Secure our Pipelines – Best Practices

Applied research and commercial platforms solutions

e.g., AWS Glue supports encryption

Applied research and commercial platforms solutions

e.g., a database running inside a private subnet

e.g., network segmentation

Build or use existing mechanisms to authenticate the users

Establish a baseline of actions

Set up effective user permissions plan

How to Secure our Pipelines – Best Practices

Applied research and commercial platforms solutions

e.g., AWS Glue supports encryption

Applied research and commercial platforms solutions

e.g., a database running inside a private subnet

e.g., network segmentation

Build or use existing mechanisms to authenticate the users

Establish a baseline of actions

Set up effective user permissions plan

Version and trace code before deploying

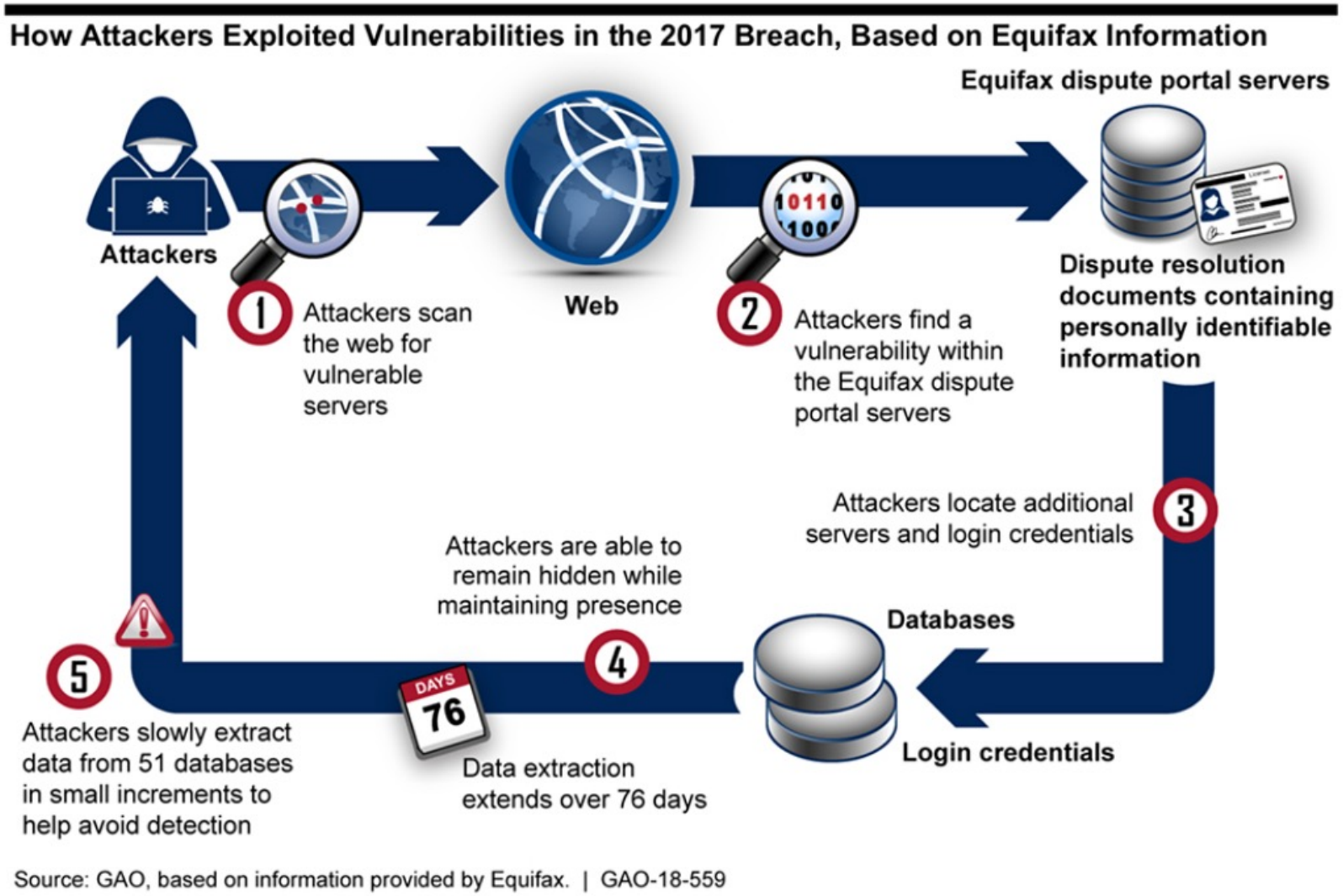
Avoid logging things that are sensitive

Careful with error handling

Create development environments for trying new things mirroring final production

Build towards and secure the whole system, not just the individual parts

The Equifax Data Breach



https://www.bankinfosecurity.com/postmortem-behind-equifax-breach-multiple-failures-a-11480

Summary

- Security Factors related to the Pipeline
- Security Factors related to Cloud (and Cloud Computing)
- Best Practices

Data Engineering Project

Module 8
Security Issues
in Data Pipelines and the Cloud

Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech