
# INTERNSHIP WITH COSS INDIA


# THANK YOU PRODEVANS FOR YOUR GUIDANCE

# DYNAMIC ETL ON ML PASS

# TRIBE -E

# TRIBE MEMBERS :

Teekshith kumar M

Ajay A

Sunitha nayak

Kajal Rajkumar Bawage

Virakthmath Shweta Satalingayya

Vaibhavi

Vaishnavi

1. Introduction
1. ETL
1. Data Warehouse
1. AWS
1. ML PaaS
1. Python
1. Code
1. Snapshots
1. Conclusion

# 1. What is ETL

Extract, Transform and Load (ETL) is a process in database usage and especially in data warehousing that involves:

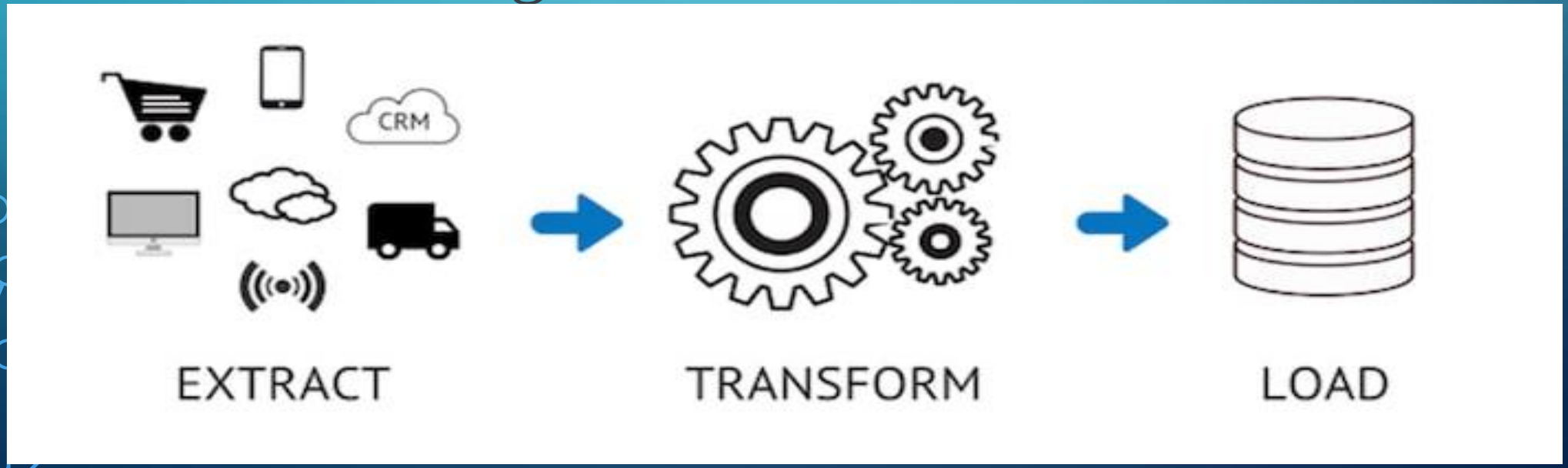
* Extracting data from outside sources.
* Transforming it to fit operational needs
* Loading it into target the end

# ETL TOOLS

- ETL tools enable data integration strategies by allowing companies to gather data from multiple data sources and consolidate it into a single, centralized location. ETL tools also make it possible for different types of data to work together.

# HOW ETL WORKS

- The ETL process is comprised of 3 steps that enable data integration from source to destination: data extraction, data transformation, and data loading.



# STEP 1: EXTRACTION

- Before data can be moved to a new destination, it must first be extracted from its source — such as a data warehouse or data lake. In this first step of the ETL process, structured and unstructured data is imported and consolidated into a single repository.
- Although it can be done manually with a team of data engineers, hand-coded data extraction can be time-intensive and prone to errors.

# STEP 2: TRANSFORMATION

- Transformation is generally considered to be the most important part of the ETL process. Data transformation improves data integrity — removing duplicates and ensuring that raw data arrives at its new destination fully compatible and ready to use.

# STEP 3: LOADING

- The final step in the ETL process is to load the newly transformed data into a new destination (data lake or data warehouse.) Data can be loaded all at once (full load) or at scheduled intervals (incremental load).

# ETL BENEFITS

- Visual flow
- Structured system design
- Operational resilience
- Data-lineage and impact analysis
- Advanced data profiling and cleansing
- Performance
- Big Data

# USES OF ETL

- ETL is commonly used to do the following

✓ Data warehousing

✓ Machine learning and artificial intelligence

✓ Marketing data integration

✓ IoT data integration

✓ Database replication

- Cloud migration

# ADVANTAGES AND DISADVANTAGES

# ADVANTAGES OF ETL:

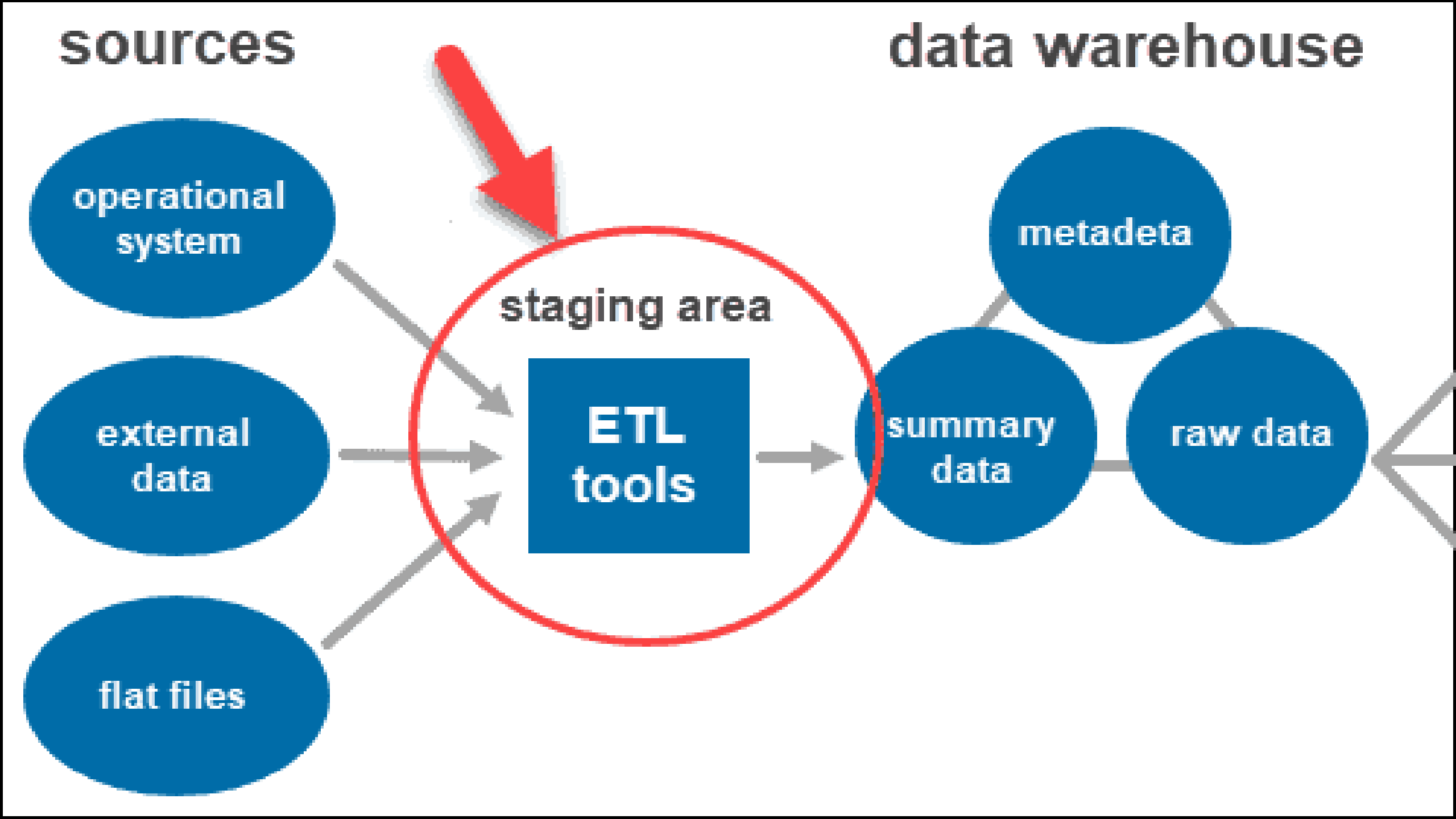
- Good for bulk data movements with complex rules and transformations.
- Makes maintenance and traceability much easier than hand-coding.
- Good for data warehouse environment

# DISADVANTAGES OF ETL:

- You must be data-oriented developer or database analyst to use
- Not ideal for near real-time or on-demand data access, where fast response is required
- Difficult to keep up with changing requirements.

# DATA WAREHOUSE

- A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouse are solely intended to perform queries and analysis and often contain large amounts of historical data.


sources
data warehouse
operational system
external data
flat files
staging area
ETL tools
metadeta
summary data
raw data

# AWS

- AWS stands for Amazon Web Services.
- The AWS service is provided by the Amazon that uses distributed IT infrastructure to provide different IT resources available on demand. It provides different services such as infrastructure as a service (IaaS), platform as a service (PaaS) and packaged software as a service (SaaS).

# AWS SERVICES

- • Infrastructure as a service (IaaS): Services in this category are the basic building blocks for cloud IT and typically provide you with access to networking features, computers (virtual or on dedicated hardware), and data storage space.
- ] • Platform as a service (PaaS): Services in this category reduce the need for you to manage the underlying infrastructure (usually hardware and operating systems) and enable you to focus on the deployment and management of your applications

Development Platform

Computing

Pay-As-You-Go

Networking


Database


Programming Models

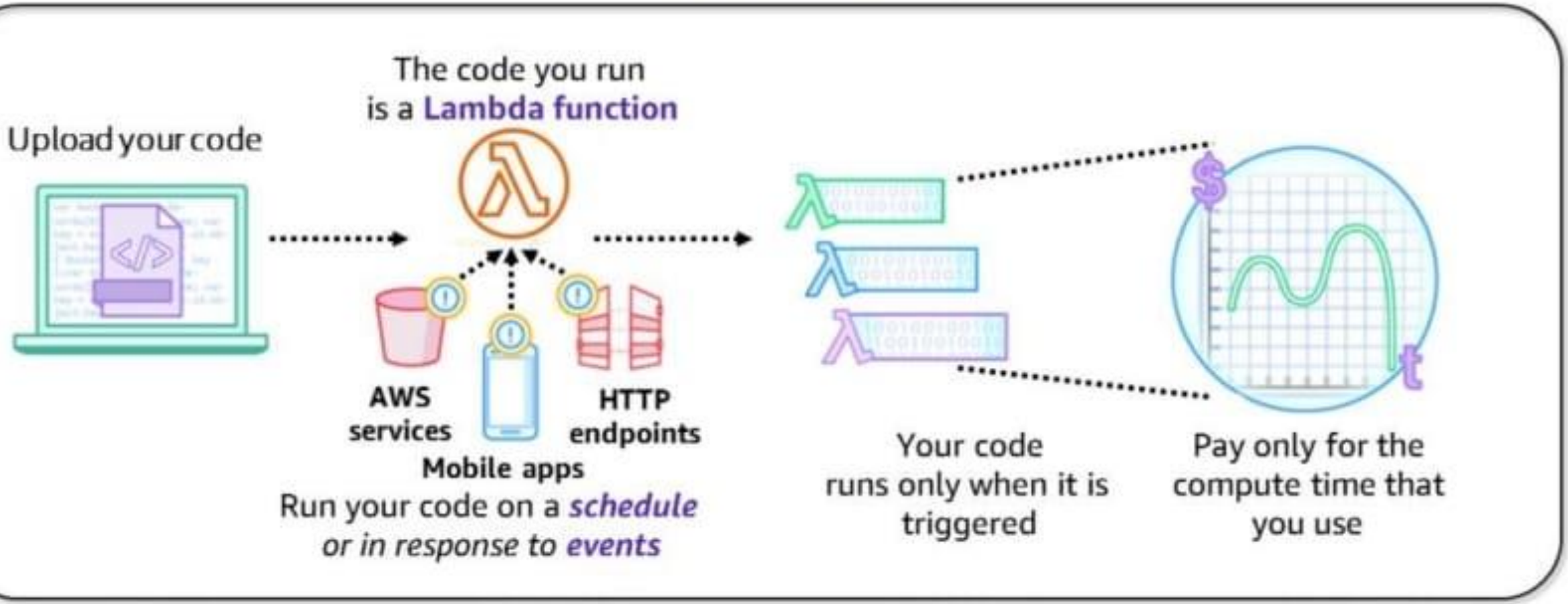
# AWS LAMBDA

- AWS Lambda is a serverless computing service that runs us code in response to events and automatically manages the underlying computing resources for us. We can use AWS Lambda to extend other AWS services with custom logic or create our back-end services. AWS Lambda can automatically run code in response to multiple events, such as HTTP requests via Amazon API

# AWS Lambda: Run code without servers

aws academy

AWS Lambda is a serverless compute service.

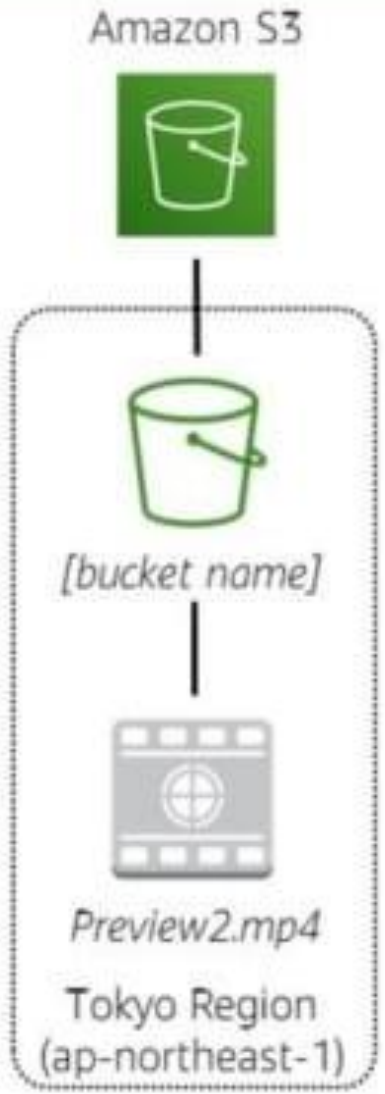


# AWS S3 BUCKET

- Amazon S3 is object-level storage, which means that if you want to change a part of a file, you must make the change and then re-upload the entire modified file. Amazon S3 stores data as objects within resources that are called buckets.

# Amazon Simple Storage Service (Amazon S3)

# Amazon S3 bucket URLs (two styles)




To upload your data:

1. Create a bucket in an AWS Region.
2. Upload almost any number of objects to the bucket.

Bucket path-style URL endpoint:

https://s3.ap-northeast-1.amazonaws.com/bucket-name

Region code — Bucket name

Bucket virtual hosted-style URL endpoint:

https://bucket-name.s3-ap-northeast-1.amazonaws.com

Bucket name — Region code

- To use Amazon S3 effectively, you must understand a few simple concepts. First, Amazon S3 stores data inside buckets. Buckets are essentially the prefix for a set of files, and must be uniquely named across all of Amazon S3 globally. Buckets are logical containers for objects.