University of Salford MANCHESTER

SCHOOL OF SCIENCE, ENGINEERING & ENVIRONMENT

Program: MSc of Data Science

Module: Big Data Tools and Techniques

Week 10

Big Data Use Cases and Your Solution to Some Real-World Scenarios

2025

Expectations

1. Choose a quiet place to attend the class and please concentrate during the lecture
2. Put your questions in Padlet and I will review them in the due time (Padlet link is in BB, week 10, Lecture folder for Q&A week10)
3. We will have 5 mins break after the first hour of the lecture (please remind me)
4. Jisc code will be shared during the break time

Learning Outcomes

1. To analyse Real-World Use Cases with Big Data.
2. To design Big Data Analytical Pipelines.
3. To evaluate Platforms for Big Data Analytics.

A few use cases

e-commerce

e-commerce …

Data Collection

Data Preparation

Data Analysis

Data Visualization

Storage

Data Ingestion

Data Retrieval

Transportation

Transportation …

Data Collection

Data Preparation

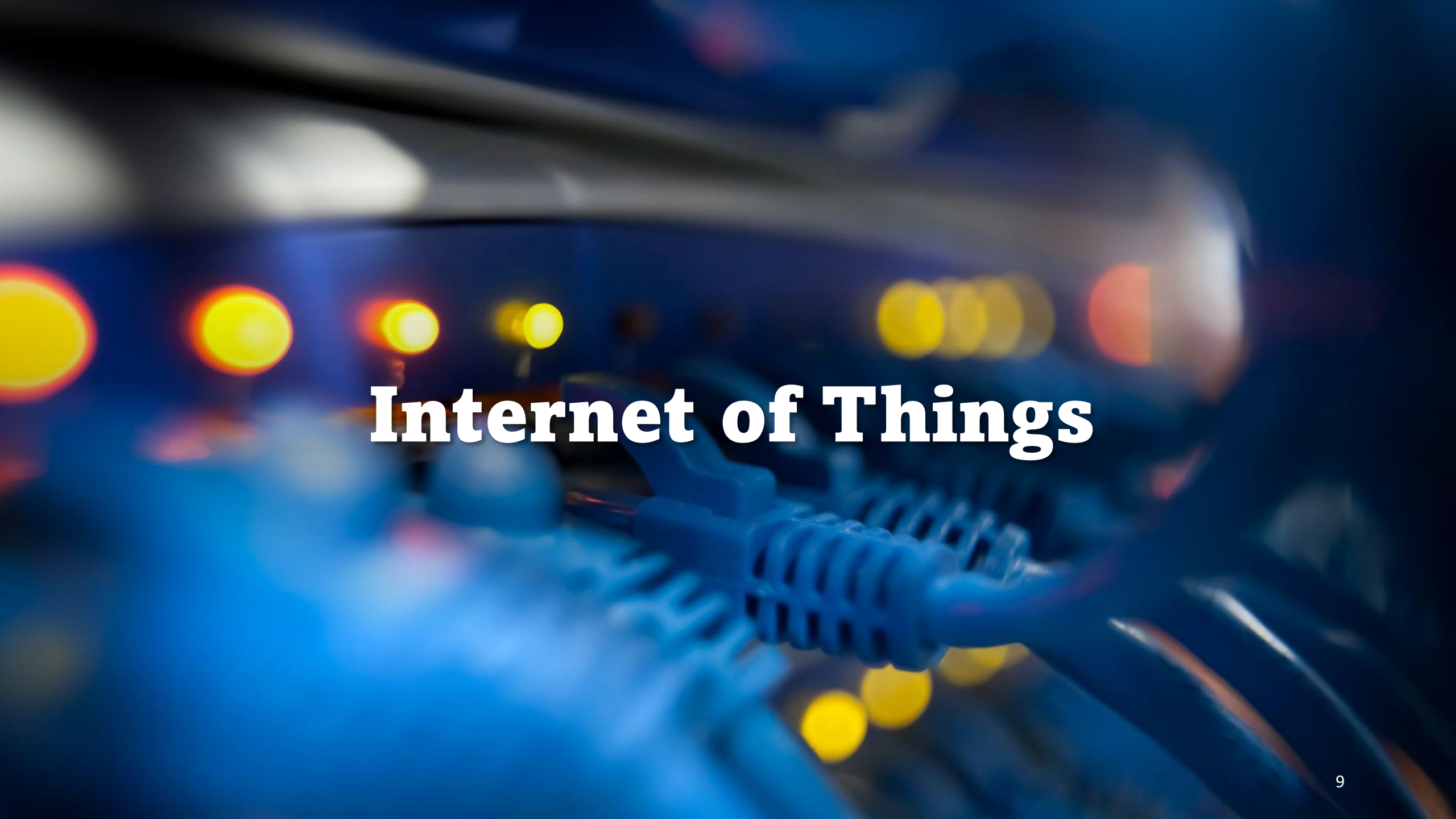
Data Analysis

Data Visualization

Storage

Data Ingestion

Data Retrieval

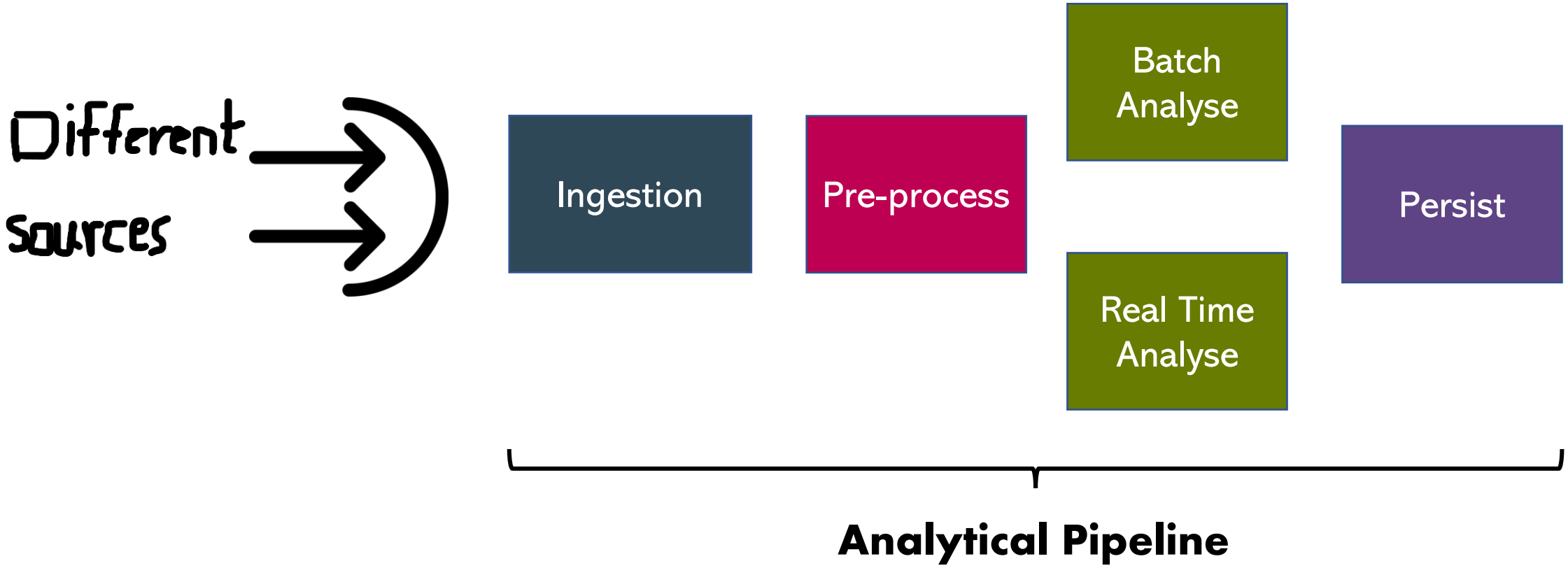


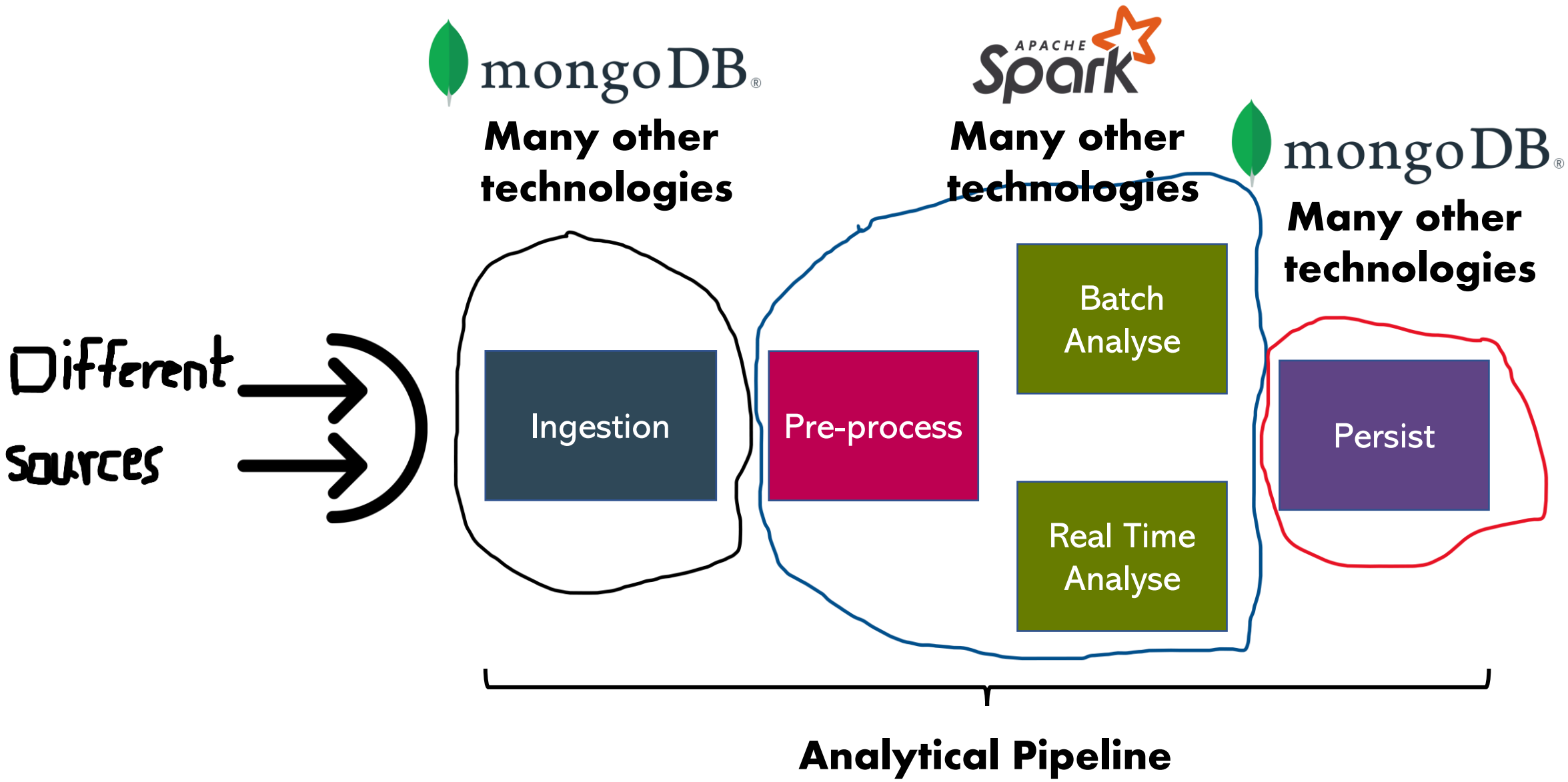
Internet of Things

IoT …

- Data Collection
- Data Preparation
- Data Analysis
- Real-time Analytics
- Storage
- Data Retrieval

A Big Picture


Different
Sources
Ingestion
Pre-process
Batch
Analyse
Real Time
Analyse
Persist
Analytical Pipeline


mongoDB
Many other technologies
APACHE Spark
Many other technologies
mongoDB
Many other technologies
Different Sources
Ingestion
Pre-process
Batch Analyse
Real Time Analyse
Persist
Analytical Pipeline

Activity 1

mymusic

At mymusic, we are more than just a music streaming application.

We are a music discovery platform.

Whatever your musical taste, we help you discover more of the music you love.

Connect with your favourite bands to purchase exclusive merchandise through their mymusic store.

Activity 1

- Think of some of the sources of data mymusic is likely to be working with
- Which is likely to be relevant – batch processing, stream processing, both?
- Think of some reasons they might have for analysing this data?
- What data processing would they need to carry out before using it for the above purpose(s)?
- Once transformation and analysis has been carried out, how could this data be persisted or used?

Activity 1: A Simple Example

- Application error logs (stream processing)
- Anomaly detection – e.g., identify anomalous error patterns which might suggest performance issues or incidents that need investigation
- Use stream processing to count error log entries over a window of time (perhaps grouped by type of error based on keywords in log entry.) We can use Spark Structured Streaming for this.
- Use this to identify spikes which need to be investigated
- Provide realtime dashboard for service team (e.g., using Dash)

Activity 1: Now it's your turn!

- Breakout sessions will last 30 mins
- In your groups come up with some suggestions for data analytics pipelines that could be relevant to mymusic, using the questions as prompts
- Then post your ideas in the Padlet using the template we've provided
- We will select a couple of groups to present their ideas (you may want to nominate a spokesperson before the 30 minutes is up)
- There isn't one 'right' answer! You can be as creative as you want, but think back to what we have covered in this module.

mymusic

Activity 1 Discussion

Activity 1: Some Related Links

- Netflix Recommendation Engine:

https://www.databricks.com/session/netflixs-recommendation-ml-pipeline-using-apache-spark

- The Netflix Tech Blog:

https://netflixtechblog.com/

- Big Data Processing at Spotify:

https://engineering.atspotify.com/2017/10/big-data-processing-at-spotify-the-road-to-scio-part-1/

Activity 2

Paymo

the online bank that works for you

Paymo is an online bank that takes the hassle out of banking.

Get instant approval on small loans to pay for the items you need. Benefit from our Protect+ fraud detection technology. Make and receive payments direct from your contacts on the app. Receive personalised recommendations in the app on products and services that might interest you.

Activity 2

- Think of some of the sources of data Paymo is likely to be working with
- Which is likely to be relevant – batch processing, stream processing, both?
- Think of some reasons they might have for analysing this data?
- What data processing would they need to carry out before using it for the above purpose(s)?
- Once transformation and analysis has been carried out, how could this data be persisted or used?

Activity 2: Now it's your turn!

- Breakout sessions will last 30 mins
- In your groups come up with some suggestions for data analytics pipelines that could be relevant to Paymo, using the questions as prompts
- Then post your ideas in the Padlet using the template we've provided
- We will select a couple of groups to present their ideas (you may want to nominate a spokesperson before the 30 minutes is up)
- There isn't one 'right' answer! You can be as creative as you want, but think back to what we have covered in this module.

Activity 2 Discussion

Activity 2: Some Related Links

- HSBC Databricks Case Study:

https://www.databricks.com/customers/hsbc

- Machine Learning at Monzo:

https://monzo.com/blog/2022/12/19/machine-learning-at-monzo-in-2022

https://monzo.com/blog/2022/04/26/monzos-machine-learning-stack

- Detecting Financial Fraud at Scale With Decision Trees and MLflow on Databricks – see Chapter 7 of The Big Book of Data Science Use Cases (will be uploaded to BB)