Welcome!

WWCode San Francisco - Backend Study Group

Jun 29, 2023

- We'll start in a moment :)
- We are **RECORDING** tonight's event
- We may plan to take screenshots for social media
- If you are comfortable, turn the video ON. If you want to be anonymous, then turn the video off
- We'll introduce the hosts & make some time for Q&A at the end of the presentation
- Feel free to take notes
- Online event best practices:
    - Don't multitask. Distractions reduce your ability to remember concepts
    - Mute yourself when you aren't talking
    - We want the session to be interactive
    - Use the 'Raise Hand' feature to ask questions
- **By attending our events, you agree to comply with our Code of Conduct**

WOMEN WHO CODE /san-francisco

Introduction & Agenda

• Welcome from WWCode!
• Our mission: Empower diverse women to excel in technology careers
• Our vision: A tech industry where diverse women and historically excluded people thrive at any level
• About Backend Study Group

Harini Rajendran
Presentor
Senior Software Engineer, Confluent
Lead, WWCode SF

Prachi Shah
Host
Senior Software Engineer, Unity
Director, WWCode SF

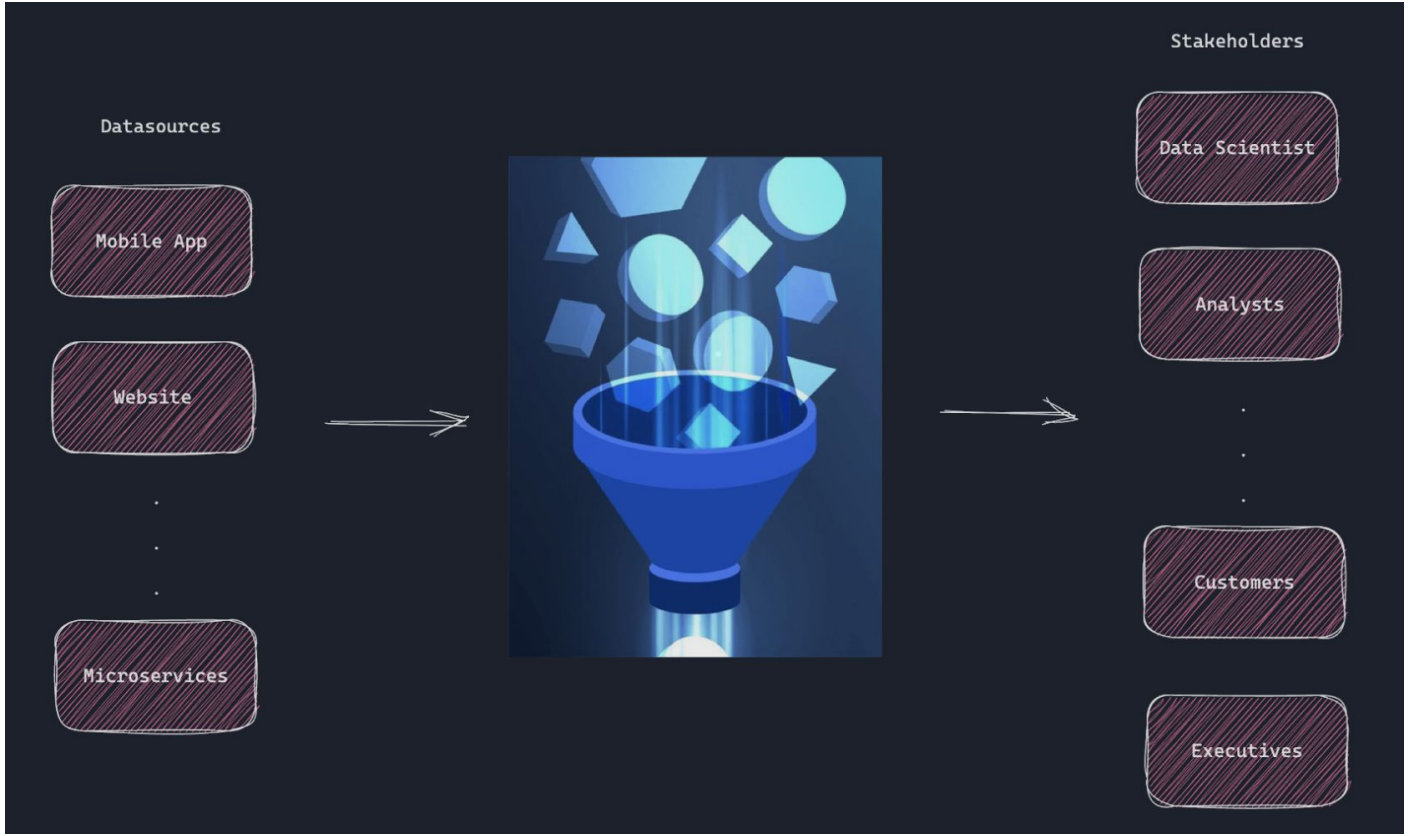
- **Data Pipelines 101**
  - **What are data pipelines**
  - **How to design data pipelines**
  - **Types of data pipelines**
  - **Aspects of data pipelines**
  - **Common technologies used**
  - **Examples of data pipelines**
  - **Q & A**

WOMEN WHO CODE /san-francisco

What are Data pipelines?

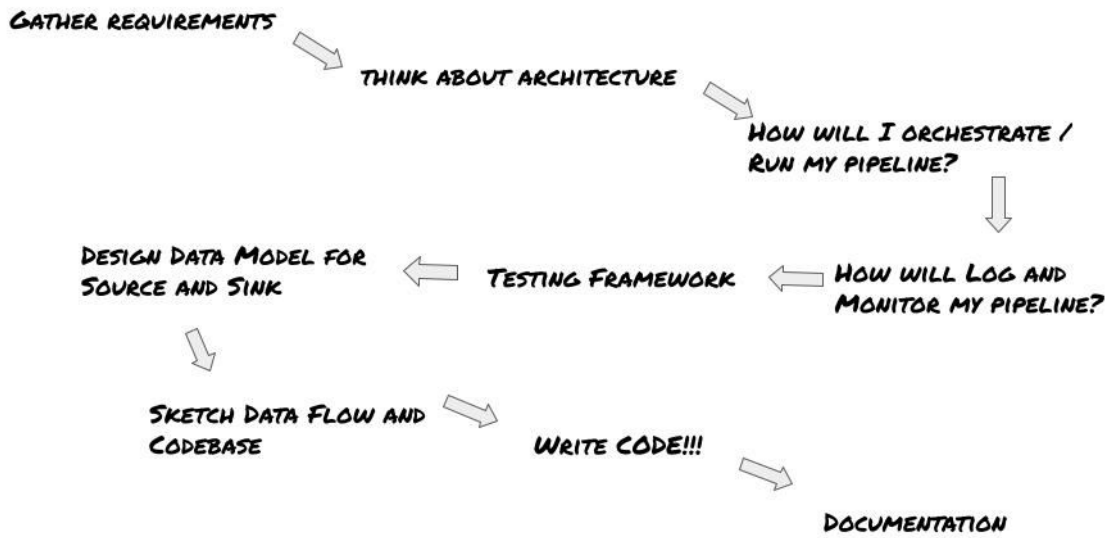
A pipeline is essentially a series of microservices/applications which performs certain actions that changes the raw data from various datasources to an understandable format so that we can store it and use it for data analysis and modelling by different stakeholders.

What are Data pipelines?



Steps in designing data pipelines

- Requirements
- Architecture
- Orchestration
- Logging and Monitoring
- Testing
- Data Modeling
- Sketch out data flow
- Implementation
- Documentation



Steps in designing data pipelines

- **Requirements**
  - Talk to the stakeholders and understand the problem
  - Collect all requirements
  - Decide on the scope
  - SLA and acceptable downtimes
  - What data is required and at what frequency and granularity
  - Preferred data format
  - What data columns/fields are needed in the output

  A 5 -10 minutes discussion can save days of development time
- **Architecture**
  - What are the different pieces?
  - How they fit together
  - Pros and cons of various options for each piece
  - How does it fit into the existing infra

Steps in designing data pipelines

- **Orchestration**
  - How is the pipeline run and who manages it? Run on a schedule vs manual trigger
  - Tools: Cronjob, Apache Airflow, etc
  - Pick the right orchestration tool
  - Monitoring and alerting of applications
  - Logging and debugging when things go wrong
- **Logging and Monitoring**
  - The more logging, monitoring and orchestration can work hand in hand, the better it is
  - Visual monitoring is better
  - Info on past runs, past failures, current runs and failures, etc. How are the apps performing?
  - Logging helps us debug what is wrong with different components
  - Never take logging and monitoring for granted

Steps in designing data pipelines

- **Testing**
  - Start thinking about testing before you write any code
  - All code should be unit testable
  - Choose technologies that makes testing easier
  - Think about how to test each piece while architecting them
- **Data Modeling**
  - Most important aspect
  - Data type and schema of source data and sink data
  - Relationship between different data entities
  - Data size and velocity
  - Data partition strategy

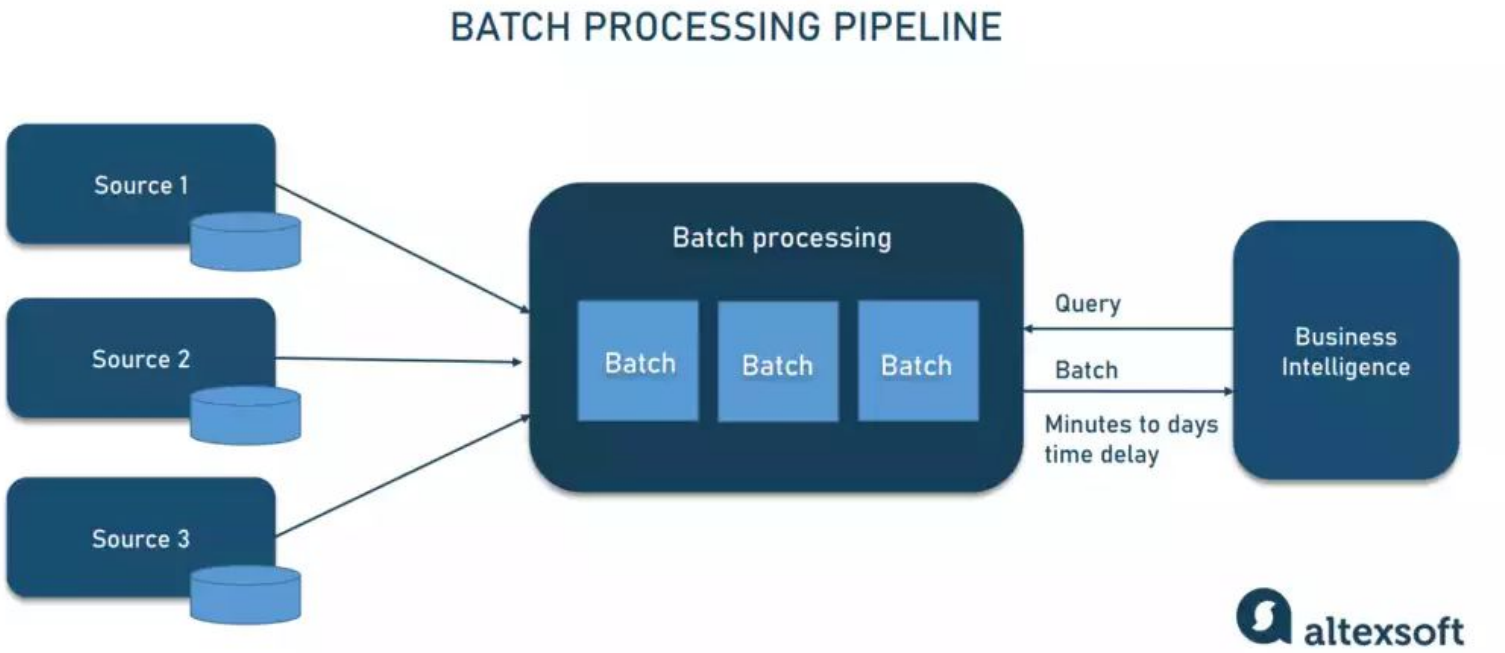
Steps in designing data pipelines

- **Sketch out data flow**
  - Write down the whole data flow (detailed design) for all the components
  - Helps to identify gaps and holes and bugs which we would have caught much later otherwise
  - Helps you to take a step back and rethink if something doesn't seem to work per your initial plan
- **Implementation**
  - Finally, write code
  - Write detailed comments that can be used for generating documentation
  - Following proper coding standards and best practices
  - Write easily maintainable and extensible code
- **Documentation**
  - Do not stop at implementation and skip this step
  - Could be a README page, Confluence page, google doc or whatever
  - Info on tech stack, testing, assumptions, data models, different sources and sinks, etc

Types of Data Pipelines

- **Batch Processing**
  - Traditional analytics is about making sense from data collected over a period (Historical data)
  - Batch of data is loaded into some repository periodically and jobs are scheduled during off-peak business hours
  - Works with high volume of data
  - Execution takes from few minutes to hours and even days
  - Used for non-urgent long term reporting like monthly accounting, daily summaries, etc
  - ETL pipelines (Extract, transform and load)
    - *Extract* – getting/ingesting data from original, disparate source systems
    - *Transform* – moving data in temporary storage known as a staging area. Transforming data to ensure it meets agreed formats for further uses, such as analysis
    - *Load* – loading reformatted data to the final storage destination

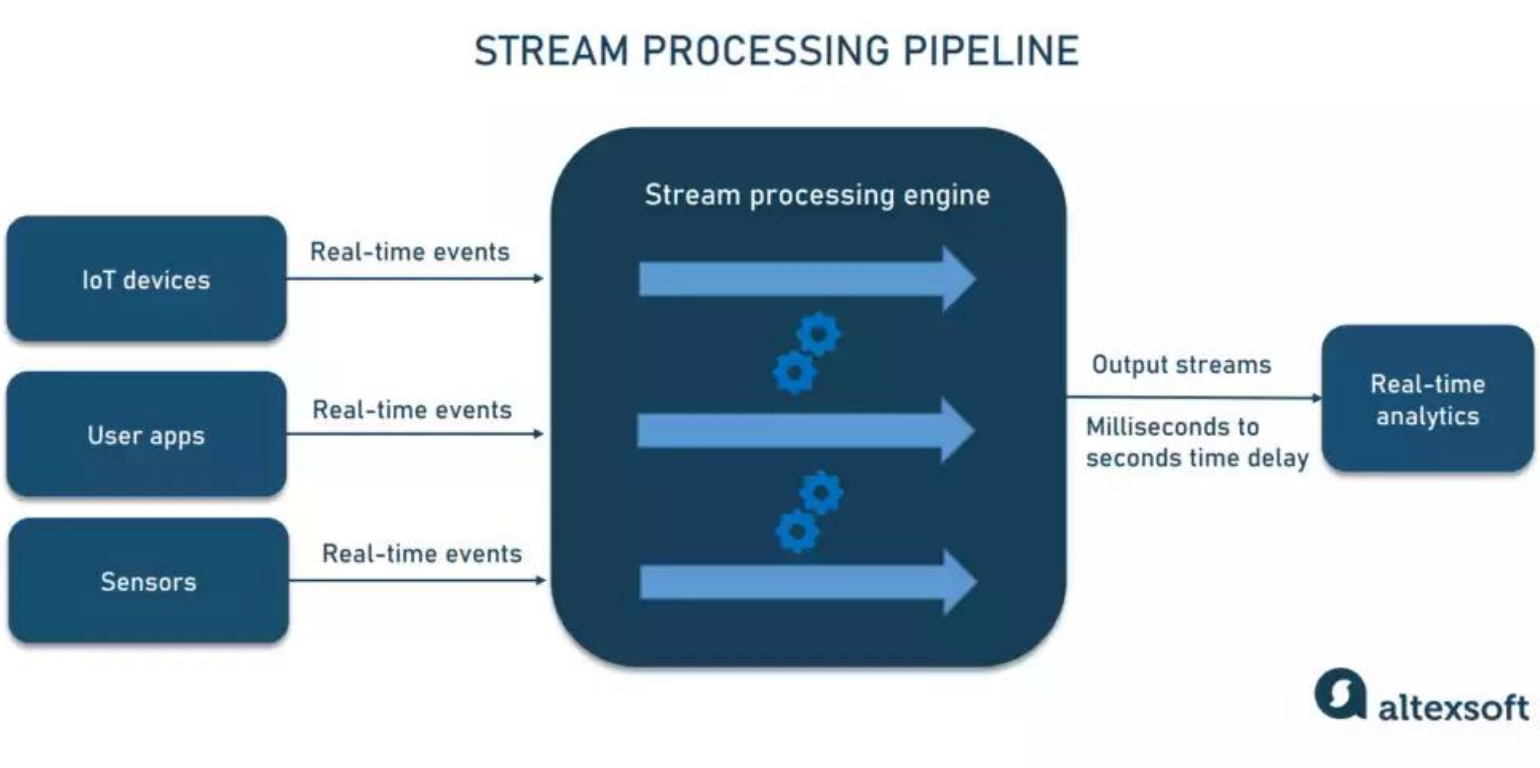
Types of Data Pipelines



Types of Data Pipelines

- **Stream Processing**
  - Real time analytics: Deriving insights from constant flow of data within seconds or milliseconds
  - Data is ingested in real time and insights are emitted in milliseconds to seconds delay
  - Queries are answered in milliseconds to seconds
  - Not as reliable as batch processing
  - Use cases
    - Used in real time operations monitoring that companies can use to react without a lot of delays
    - Used in real time analytics in trading
    - Autonomous vehicles operations
    - Real time inventory updates and historical sales data updates

Types of Data Pipelines



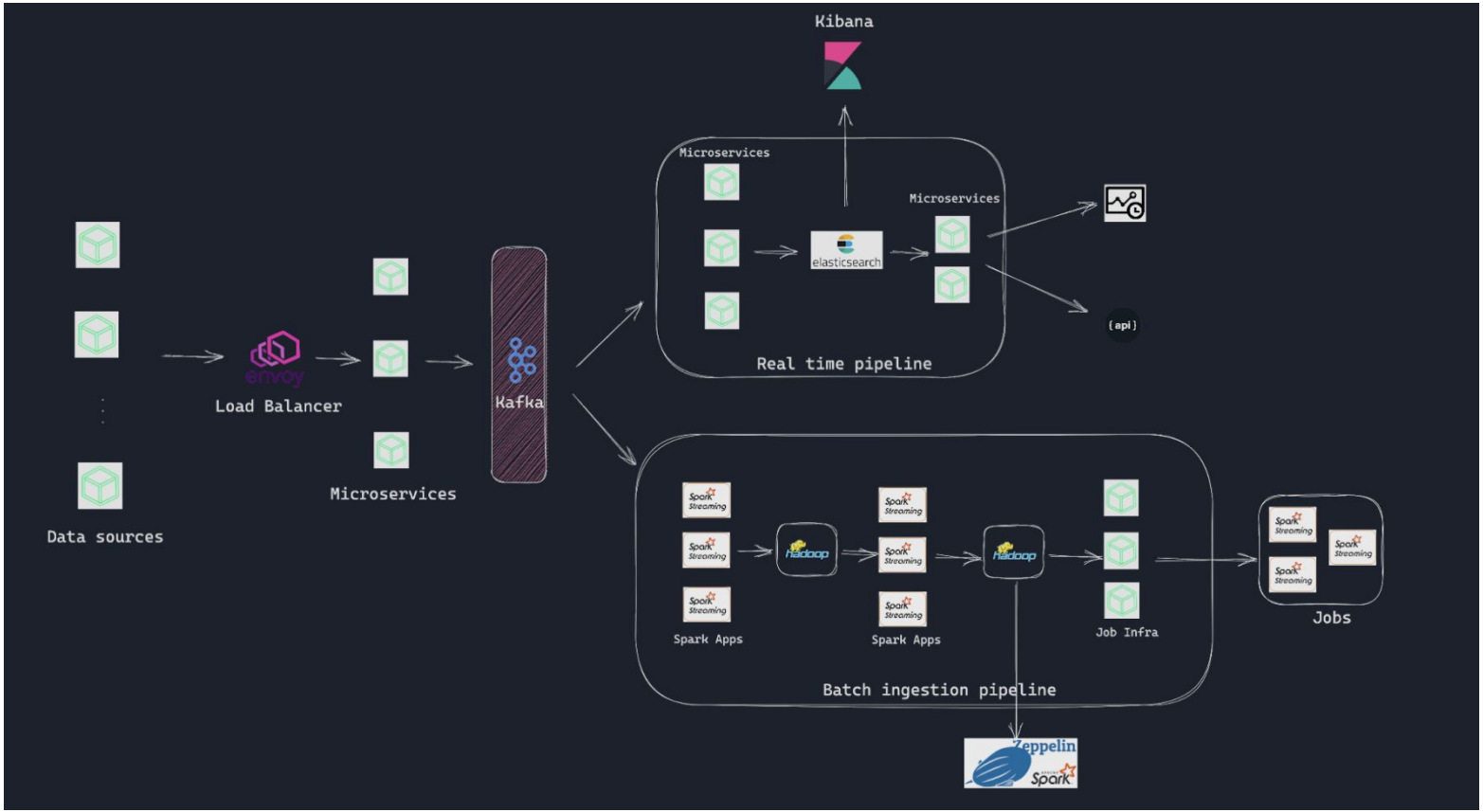
Aspects of Data Pipelines

- Source apps Instrumentation
  - Instrumenting the services (web-site, app, microservices, etc) to produce the required data
- Ingestion
  - Gathering all the data instrumented in step 1 in a centralized place. This is the first part of the big data pipeline. Mostly big data technologies like kafka, s3, etc are used here.
- Processing
  - Gathered data is cleaned, transformed and processed for easy and fast retrieval. Distributed data processing platforms like hadoop, spark, etc are used here
- Storage
  - The processed data is stored for efficient retrieval. This is where distributed storage systems like s3, NoSQL databases, etc comes into picture
- Access
  - Entities with relevant permissions should be able to access the data in an easy and straightforward manner. Data access should be fast and the retrieval time should mostly be in the order of milliseconds to under a few seconds based on the needs of the system. In case of batch data access, sub second latencies are not the norm.

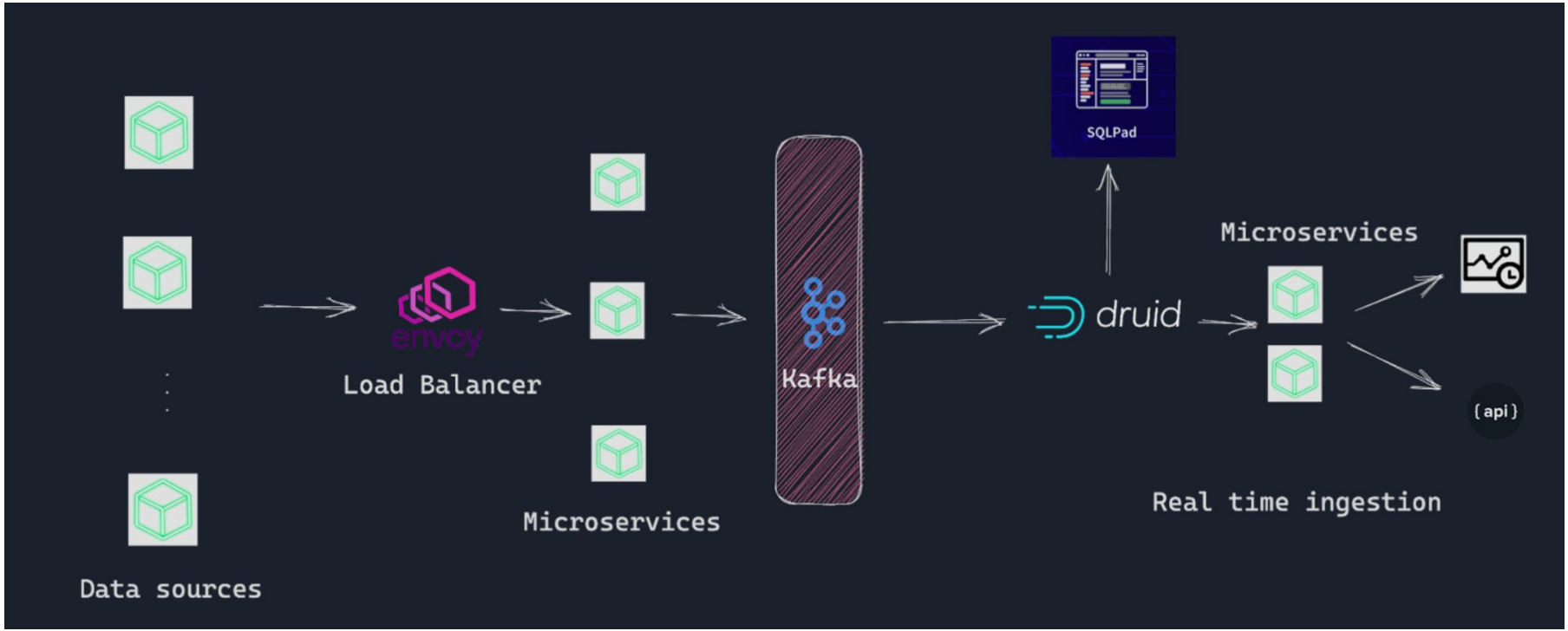
Common Technologies Used

Stream Processing	Apache Kafka, Amazon Kinesis, Apache Spark, Apache Flink
Batch Processing	Hadoop MapReduce, Apache Spark(Micro batching)
Databases	Relational DB: Sql, Postgres Datastore:Amazon S3, Azure blob, HDFS NoSQL DB: MongoDB, Elastic Search, Cassandra, Couch DB, Influx DB, Apache Druid

Example Pipeline 1



Example Pipeline 2



Backend Study Group

References:

- https://medium.com/international-school-of-ai-data-science/data-pipelines-101-a-brief-overview-32075c717d6c
- https://www.confessionsofadataguy.com/data-pipelines-101-the-basics/
- https://www.ibm.com/topics/data-pipeline
- https://www.altexsoft.com/blog/data-pipeline-components-and-types/
- WWCode Intro to Data Engineering slide deck

Backend Study Group:

- Presentations on GitHub and session recordings available on WWCode YouTube channel
- Upcoming sessions:
  - July 6th, 2023 - Ruby On Rails 101
  - Aug 3rd, 2023 - Domain Driven Design
  - Sep 21st, 2023 - Web Development 101

Women Who Code:

- Technical Tracks and Digital Events for more events
- Join the Digital mailing list for updates about WWCode
- Contacts us at: contact@womenwhocode.com
- Join our Slack workspace and join *#backend-study-group*!

You can unmute and talk or use the chat

WOMEN WHO CODE /san-francisco