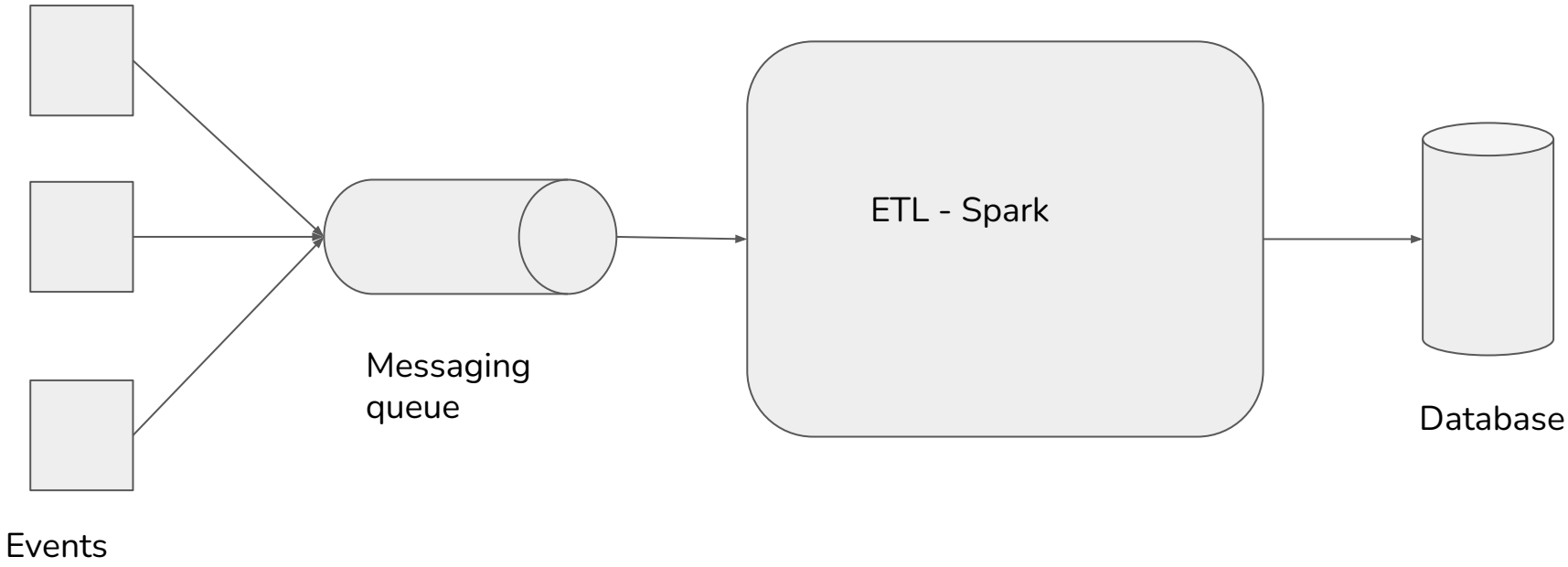
Spark Streaming

Agenda

- Spark Streaming
- Case Study

Introduction to Streaming

- So far we have focused on batch process - running all in one
- We want to start thinking of data pipelines



APIs - Application Programming Interface

- Websites and companies can create applications to access their data
  - APIs allow programmers to write programs that can access data programmatically.
  - Web APIs are very common
    - Google Maps
    - Yahoo Finance
    - NASDAQ
    - Weather data
- It relates to request responses
  - GET
  - POST
  - DELETE

ETL vs ELT

- ETL - Extract Transform Load
  - Based on Waterfall Approach
    - Extract - Get the data from the source
    - Transform - Design what's the best approach for the data cleaning and data wrangling
    - Load - Load data into the database
  - Slow turnaround
  - Less interactive
  - Traditional approach

ETL vs ELT

- ELT - Extract Load Transform
  - Based on Agile Approach
    - Extract - Get the data from the source
    - Load - Load the data into Staging Tables: Data is available right away so we can explore it
    - Transform - Transform the data from the Staging phase, and continuously work on improvement
  - Based on Continuous Improvement / Continuous Development (CI/CD)
  - Agile turnaround
  - Allows for multiple changes in the process.

Batch vs. Streaming

- Batch
  - We have been working with batch execution all along
  - Process and transformations are done at execution
  - Job is run once
  - Results are stored, job and memory is flushed away at the end
- Streaming
  - Jobs are run in time sequence
  - The process is also called mini-batches

Source: https://spark.apache.org/docs/latest/streaming-programming-guide.html

Case Study

Flights Prices and Delays

- You are the Data Engineer for a new startup that wants to disrupt the Flight + Hotel Business
  - Think Expedia, eDreams, Bookings, Hooper
- To start, you want to build a Proof of Concept
  - Can we use public API to determine flight prices?
  - What other information can we gather?
  - Can we visualize the results?
- After long research you decided to use a Web API
  - Amadeus!
  - Source: https://github.com/amadeus4dev/amadeus-python

Great Learning

Happy Learning !