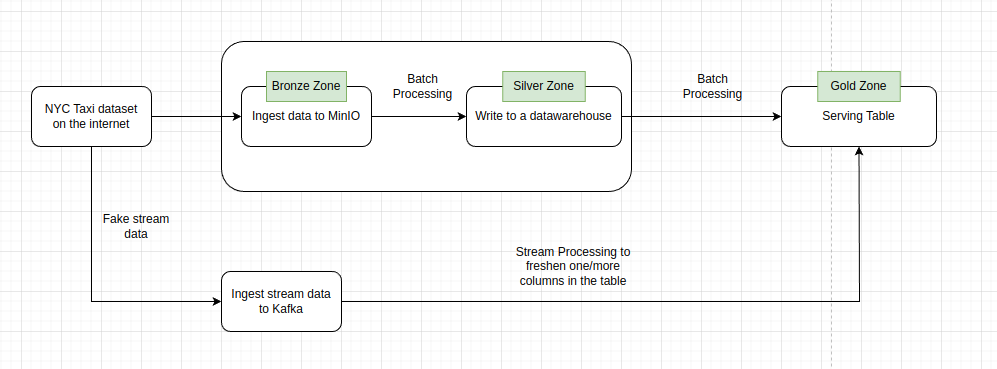
## **Mục tiêu**

* 1. Cấp chứng chỉ cho anh em đem đi khoe
  2. Ôn lại nội dung các bài học

## **Tiêu chí chấm điểm**

* 1. **Điểm bài giảng (70%):**   
     Trong dự án thực hiện cần sử dụng các công nghệ:
     1. MinIO (10%)
     2. Trino (10%)
     3. Data Lake table format (10%)
     4. Spark for batch processing (10%)
     5. Kafka or RabbitMQ or GCP PubSub for storing messages (10%)
     6. Flink or Spark for stream processing (10%)
        1. Dùng window: 10%
        2. Không dùng window chỉ được 5%
     7. GX or Deequ for data validation (5%)
     8. Pipeline Orchestration, i.e., Airflow, Prefect, or any similar tools (5%)
  2. **Điểm thực tế (20%):**
     1. Sử dụng GCE/GKE để deploy MinIO (5%)
     2. Sử dụng GCE/GKE/Dataproc để deploy Flink (lưu ý Dataproc không work với PyFlink) (5%)
     3. Sử dụng GCE/GKE/Dataproc để deploy Spark (5%)
     4. Sử dụng GCE/GKE/Cloud Composer để deploy Airflow/Prefect (5%)
  3. **Điểm dữ liệu to (10%): Dữ liệu lớn hơn 100 GB**  
     Một số kho dữ liệu có thể tham khảo
     1. [NYC Taxi & Limousine Commission - yellow taxi trip records](https://learn.microsoft.com/en-us/azure/open-datasets/dataset-taxi-yellow?tabs=azureml-opendatasets) từ [Azure Open Dataset](https://learn.microsoft.com/en-us/azure/open-datasets/dataset-catalog)
     2. [UC Ivrine ML Repository](https://archive.ics.uci.edu/datasets?skip=0&take=10&sort=desc&orderBy=NumInstances&search=)
     3. Tự generate dữ liệu

1. **Đề tài gợi ý:**Xây dựng hệ thống Feature Store với kiến trúc như sau



## Deliverables

* 1. Source code trên Github với diễn giải chi tiết cách cài đặt và cách chạy code ở README.md
  2. Video chứng minh các công nghệ sử dụng và demo full flow
  3. Một số ví dụ của học sinh khóa trước:
* Dữ liệu diabetes (tự sinh):
  + Đồ án feature store có kết hợp cả airflow pipeline để batch training, flink để xử lý data (data xử lý xong đẩy lại vào kafka và có 1 job kéo dữ liệu từ kafka gọi API predict): <https://github.com/dunghoang369/feature-store>
* Dữ liệu NYC Taxi:
  + Đồ án feature store có cả kiến thức của m1 (monitoring các kiểu): [https://github.com/luongphambao/nyc\_taxi\_feature\_store](https://github.com/luongphambao/nyc_taxi_feature_store?fbclid=IwAR3xVTaCPFhDk_GhSgy-RNl0u6v_vchjzmtLXkn4EkpAmyEsHmbILnLBbB4_aem_ATUe-Q0dDCPcCdDjvZCr202gpvAtobDdnJScnu0m1NIhGIE_KLq5oO7Pk1UuH0x_Izgma1kHvc8xrEexG3wY6VJJ)
  + Đồ án feature store khác đơn giản hơn: <https://github.com/HungNguyenDev1511/Capstone-Project-Data-Engineer>