Consult lại dự án

1 data project hoàn chỉnh sẽ có dạng như sau:

A diagram of data engineering lifecycle

Description automatically generated

\_ Vì vậy cần phải xem xét kĩ cho mỗi stage, chuẩn bị các bộ câu hỏi cho mỗi stage và câu hỏi tổng quát.

1. Câu hỏi tổng quát
   1. Bài toán này nhằm giải quyết vấn đề gì? (What)
   2. Bài toán này dành cho các đối tượng là ai? (Who)
   3. Làm sao để có thể giải quyết đưuocj bài toán này? (How)
   4. Các yêu cầu ngoài lề
2. Source System 🡪 nắm được kiểu dữ liệu 🡪 Chọn kiểu injestion và storage
   1. Data source bao gồm những kiểu dữ liệu gì?
   2. Data từ source system tồn tại bao lâu? Ngắn hạn hay dài hạn?
   3. Dữ liệu sinh ra với tốc độ như thế nào?
   4. Dữ liệu có được phép trùng lặp hay không?
   5. Schema của data là gì?
   6. Tần suất dữ liệu được lấy ra?
3. Storage -> Hot data hay cold data?
   1. Redis
   2. Elasticsearch
4. Injestion
   1. Nifi
   2. Kafka
5. Processing
   1. Spark
   2. Pentaho
   3. Logstash
   4. Kafka

Đối với RDBMS , phải có schema, tên bảng, sau đó fetch etl của bảng đó ra, load

Đối với nosql, load index, log

Nifi cho batch, kafka real time – messeing, logstash xử lí online

Log stash ghi vào đâu, ghi ntn, đóng vai trò gì, ghi chi tiết ra bảng dữ liệu đầu ra đàu vào nguồn

Yêu cầu, vẽ lại mô hình tổng thể, define chi tiết các chức năng ra