Interviewee Competency Test - Youtube Use Case

Assessment Criteria and Expectations

In this assessment we are looking to evaluate your expertise in the field of Data Engineering across solutioning, design, implementation, scalability and maintainability.

You are also encouraged to showcase any innovation as long as it's within the bounds of the assessment.

The outcome expected from this assessment is a functional pipeline with data ready for consumption by business analyst.

Scope and Implementation Direction:

- Your infrastructure solution, architecture and design should use only AWS Services where possible.
- Your data architecture should follow AWS Lakehouse methodology.
- Your code should by default use PySpark unless necessitated by your chosen technology solution.
- **Candidate is to use their own Google Project API Key for Youtube Data API.**

These are the sections you will be evaluated on:

- Data Pipeline Solution
  - How clearly you are able to plan the data flow for your ingestion pipeline
  - How clearly you are able to segregate your pipeline to be scalable and maintainable
- Data Architecture
  - Your understanding of how data should be structured and maintained
  - Where specific processing should take place in the data landscape
- Data Structure(s) and Model(s)
  - How you structure and partition your data
  - What data types and data formats you use
- Technical Knowledge, Approach and Algorithms
  - Your technical reasoning and considerations as you implement and address challenges

Submissions expected:

- Presentation Deck
  - Solution Overview (Architecture, Implementation and Data/Ingestion Flow)
  - Answers to the questions & How the answers were derived
  - Recommendations for Improvements
- Data Pipeline Code

- With comments on what each code block does

The agenda for the evaluation presentation is:

- Presentation - *walk us thru your solution and outcomes based on your presentation deck.*
- Showcase - *walk us thru the output of your working pipeline.*
- Q&A - *we will be asking you questions to evaluate your understanding of your solution and approaches.*

Key Points we would like to see (but not limited to):

- How you store and structure your data according to data management best practice(s).
- What file types or technology used to store your data.
- What technologies used to run and orchestrate your pipeline(s).
- How you manage the ingestion of data over time.
- Resourcefulness
- Your suggestions and potential solutions to challenges

The assessment is designed to be open ended to allow for Data Engineer candidates to showcase your strengths.

The **duration for this assessment is 1 week**, with no restrictions on the technology, coding language used or approach. **You will be allowed to ask (via email) for clarification on any item during this period** however we will not be addressing any "how-to" or "correct approach" questions.

Use Case

SPH Media deals with primarily audience data and these data may come from various social media sources such as facebook, instagram, youtube, etc.

SPH Media has many news publications under it's portfolio. Management would like to understand how well the publications are engaging with it's audience and the general performance of the videos that are published there. The publications Youtube Channels that need to be reported on are:

- Strait Times (▶ The Straits Times)
- Business Times (▶ The Business Times)
- ZaoBao (▶ zaobaosg)
- Tamil Marusu (▶ Tamil Murasu)
- Berita Harian (▶ Berita Harian Singapura)

The questions that management is asking are:

- How many subscribers are there for each channel?
- How many video's have been published for each channel?
- What is the trend for videos published by each channel over the last 12 months?
- Which are the most viewed videos?

You are to also highlight if there are any interesting findings per your analysis of these channels.