# Spark Technical Interview:

Exercise: **In the next interview we would like to discuss the following use case and understand the approach being followed to build a simple streaming application. We are not looking for a working example. - Use Case:**

1. Create a Spark Structured Streaming (Python or Scala) Pipeline to publish some data to Kafka
2. Using Spark Structured Streaming (Python or Scala) read the same data from Kafka and store it in HDFS Parquet - RAW Zone (use any sample XML with nested elements)
3. Reads data from RAW Zone using an Hourly scheduled Spark Batch process and loads the final parquet file – (Processed Zone)

## Output Data Requirement:

* Create a sample Project Folder Structure for Code i.e. scripts, logs, etc. to show how it will be organized
* Create sample scripts (pseudo code) and place them in corresponding folder
* Consider following in the code:
  + Consume Kafka (Offset Maintenance and De-duplication)
  + XML Parsing and flattening
  + Data Validation
    - Dynamic data validation
      * Schema Validation
      * Data Type Validation
      * Data formatting (trim, etc.)
    - Fault Tolerance for Application:
      * Error Handling
      * Continuous Streaming
      * Checkpoint Restarts from a specific restart
  + Partition the data based on a date field in the final Parquet file (Processed Zone)

## Expectation:

* + We expect to see pseudo code only
  + We would like to see the main Shell Script (spark submit) and the Python/Scala Spark program in the next meeting
  + Refer to github, google, etc. or any other Spark Source/documentation to complete this assignment.
  + We want to see the approach (pseudo code) and want you to do a code walk-through to understand what functions/methods you used and reasoning behind it. **We are not expecting a working code example**.

**Other inputs for consideration/discussion**

1. How will you make sure only delta/new records are pulled from RAW Zone to Processed Zone?
2. How will you move the old data from RAW Zone so that it does not become very large?
3. How will you run these programs – Cluster or Client
4. How will you decide how many Cores and Executors are needed for 1) Spark Stream Job 2) Hourly Spark Batch
5. How do you ensure we do not run into small file issue in the RAW and Processed Zone

Reach out if you have any questions.