# **PROJECT PROPOSAL**

## Problem Statement:

* 1. Context:

The objective of this project is to create a complete Data Engineering pipeline with a feasible size of StackExchange data set for Data Science enablement. It aims to load an initial dataset of StackExchange and integrate it with the data collected from API, do the necessary cleansing and make it in a format easily accessible and readable by power users.

* 1. Success Criteria:

The final product should be an end to end pipeline of data extraction, cleansing, transformation and loading with batch processing, scalability and logging built in.

Some sample use cases will be built as a nice-to-have feature if time permits to show the success of the product.

The final data output should be available for users to work on and build various analytics on top of it.

* 1. Scope:

The scope of this project is limited to the StackExchange data available in the archives. Also, the accessibility of API also needs to be worked on.

The scope of this project is limited by other factors like personal laptop power, technology availability for the non-commercial users, OS limitations if any etc.

## Data Sources:

* 1. Static/Bulk: Data as latest as Dec 2020 is available at https://archive.org/download/stackexchange

Datasets specific to certain domains are available. Ex: Data Science, Data Engineering, Travel etc. Within each domain, datasets available for various subcategories like:

Badges.xml

PostHistory.xml

PostLinks.xml

Users.xml

Comments.xml

Votes.xml

* 1. Incremental:

API available at https://api.stackexchange.com/docs

Applications should be registered on Stack Apps to get a request key. Request keys grant more requests per day and are necessary for using access\_tokens created via authentication.

Data is available in JSON format from API.

## Architectural elements and Technology choices

* 1. Make an initial study of the data available.

Build data models as applicable.

* 1. Extraction – The first step in the project is to extract the data.
     1. Python 3
     2. SQL - SQLite – Need alternatives to work on MAC OS. SQLite is not suitable for capstone projects in my view.
     3. Other technologies as discussed below.

Create empty tables in SQL that are relevant to the datasets downloaded. This will serve as a Staging area.

Parse the data from xml using Python and load some data into SQL DB.

Do the processing in bunches/blocks.

Build methods to parse JSON files and integrate with the above parsed data.

* 1. Transformation –
     1. Python 3
     2. SQL - SQLite – Need alternatives to work on MAC OS. SQLite is not suitable for capstone projects in my view.
     3. Other technologies as discussed below

Understand the relation between data tables

Update the data models

Build the facts and dimensional modeling if needed. Built data mart as needed.

Clean the data if needed

Build transformations.

Build some additional metrics to give the user a high-level perspective of the data so that they can see the relevance for their use case.

Use cloud technologies like Databricks, Azure Data Factory as applicable for data transformational logic.

If data is stored as raw files in Data Lake, make a study on ELT possibility if applicable.

* 1. Load

Load the transformed data into SQL Datawarehouse/data mart

Load the data files/data into Azure – Decide after going through the Cloud chapters in the curriculum.

Chose Azure Data Lake Store Gen 2 or Azure Synapse Analytics as applicable.

Build ways for users to access the data.

If needed, use Azure Polybase to migrate data from local to Azure for the first-time load.

* 1. Scheduled runs, scaling and other non-functional requirements:

Build the scheduled runs for ETL batches to integrate incremental data with initially loaded data and update the Datawarehouse

Build scalability as and where applicable

Build Orchestration using AirFlow kind of technologies as applicable.

* 1. Build and execute Unit Tests using Python or any other technology as applicable.
  2. Code /Build repository – Git, Github

Build logging along the way for every process.

## Sample use cases

Build some sample use cases and simple solutions as a proof of success.

The objective of this project is to create a complete Data Engineering pipeline with a feasible size of StackExchange data set for Data Science enablement. It aims to load an initial dataset of StackExchange and integrate it with the data collected from API, do the necessary cleansing and make it in a format easily accessible and readable by power users. An example use case for the data would be to observe a pattern of which time of the year, users are more active on a particular technology or what are the top buzz words used in ‘About Me’ in a particular dataset.