

<b>Project Name: START - Online Low-Intensity Conflict Research Library</b>	<b>Members: Alejandro Chacon, Amitesvarjit Broca, Matthew Chan, Ujwal Gupta, Tanishq Kaushik</b>
<b>Final Project Report</b>	<b>05/05/2023</b>

## Introduction

START is a university-based research and education center with scholarly researchers studying the causes of terrorism around the world. After pulling keywords from the articles and documents related to terrorism and counter-terrorism, the researchers categorize them into variables in their Excel dataset. Such variables include title, author, start and end date, geographic location, and more. By classifying the variables, START hopes that their users, such as academic researchers, government agencies and the general public, are able to search for these documents on their website using these keyword variables. Their goal of the website is to fill in the knowledge gap and understand what research that has already been done. Therefore, the research done by others will hopefully not be repeated and can instead be focused on other areas.

## Our Clients

Our clients are Elizabeth Radziszewski, who is a senior researcher and Project Lead, and Sean Thomas Doody, who is also a senior researcher and Tech Lead. We meet with our client Sean on a weekly basis and provide status updates on the project. The stakeholders involved in the overall START project include the Department of Defense (DoD) and the researchers who will be using the final product of this project.

## Problem

The problem our clients face requires finding a solution that would allow them to query data from their big data Excel sheet to their website.

## Solution/Results

### I. Understanding the Data and Formulating a Plan

At the beginning of our project, we recognized the need for a backend resource for our client to store and access their data. With this knowledge, we decided to create a relational database using SQL that would allow us to efficiently store and retrieve information. To store this instance, as well as relying on the fact that our client has an AWS subscription, we recommended the RDS product from Amazon that allows you to store a SQL server as a service in the cloud.

### II. Database Development

The dataset provided includes a lot of columns related to each article. Some articles had multiple rows assigned to them, and some of the columns had one-hot encoding and dummy variables to describe the articles attributes. In this step, we went through each of these attributes and noted the relationships between each of them and to the articles. With that information and our goal of normalizing the database to the third form, we began to draft an Entity Relation Diagram (ERD) schema that would satisfy all of our requirements.

### III. Transforming and Loading our Data

The Excel data was transformed into different subsets, and later tables, using Python scripts. The linking tables were created by carrying out a join among columns present in the form of dummy variables.

The non linking tables were created by extracting the unique values from the documentation provided to us.

To load the tables with the data into Amazon RDS, we wrote a Python insertion script that included host credentials. These scripts allowed us to establish a connection to directly populate the RDS instance.

#### IV. Solution Plan

Overall, for our solution plan, we had to first understand the relationship between elements in the database. This understanding in the relationships and normalizing allowed us to create an ERD on MySQL. This ERD was forward engineered into Amazon RDS to create the database instance. Then, we created Python scripts for generating the linking and non-linking tables. These tables were ultimately uploaded onto Amazon RDS using insertion scripts.

#### Next Steps/Future Plans

There are several possible future implementations for this project:

1. Implementing an update feature on the script for new entries
  - a. This will allow for new database entries to be entered. The back-end will continue to grow and have an easier process to handle new research.
2. Expand the database to contain all of their original columns
  - a. Originally, the scope of this project was narrowed down to a set amount of columns lower than the original column count. A future enhancement would be to add all the columns that remain from the original set of data.
3. Connecting the back-end database to the front-end website
  - a. Now that the back-end database has been designed and created, the next step is to have the front-end developer connect this database to their running site, allowing the end users to access and query the research data in the back-end.

#### Final Deliverables

Google Link: