# Weeks 11 & 12: Final Project Assignment

## Objective:

Leverage the technologies you’ve learned throughout the course to implement a data-focused solution of your choosing. You should document your work in a final report.

## Requirements:

1. **Use at least two components:**  
   Use the examples in the "Additional Resource” examples section to guide your component selection.
2. **Document your data source:**  
   Use your own dataset (you cannot use the same datasets from examples provided). Clearly outline where the data comes from and its format.
3. **Describe your data operations:**  
   Explain the operations or transformations you performed on the data (e.g., loading, cleaning, aggregating). Show how the data flows across your chosen components.
4. **Explain your component choices:**  
   Provide justification for selecting specific technologies for your use case.
5. **Include screenshots:**  
   Provide visual examples that demonstrate the usage of these components with your data.
6. **Two-page minimum report:**  
   Your report must summarize all steps, from data source to operations and integrations.

## Additional Resources:

You can refer to examples from the following repository: [DSC650 Week 11-12 Assignment Examples](https://github.com/bellevue-university/dsc650-updated/tree/main/week11-12/assignment/examples)  
  
You are encouraged to use these resources as guides, **but you cannot use the same datasets or libraries provided in these examples:**

**Please refer to the interactions in the examples below, as the Docker networking has already been configured and tested for seamless connectivity between the containers.**

* **Example\_Project.json:** This NiFi template downloads **customers-100.csv** from the assignment folder and puts it into HDFS in the /tmp directory. You can use your own data, load it to HDFS, and then put it in a Hive table. You may also explore Hive external tables, which allow Hive to read data from a directory without moving the data.
* **SparkAddLibraries.docx:** This guide walks through adding a Python library and running it with PySpark (e.g., using pip to install the requests library). You may explore other libraries in your project.
* **QueryHivePySpark.docx:** This document guides you through creating a Hive table with sample data and querying it using PySpark.
* **HdfsToHivePySpark.docx**: This document guides you through reading a CSV file from HDFS, creating a Hive table using PySpark, loading the data into the Hive table, and then querying the table using PySpark.
* **HiveQueryHbase.docx**: This document walks you through creating an HBase table, loading data into it, and using Hive to query the data stored in the HBase table.
* **SparkMachineLearning.docx:** This document walks you through creating a Linear Regression Machine Learning model with PySpark, Hive, and Spark MLib.

## Submission Guidelines:

1. Submit your report in PDF format.
2. Ensure all images and text are clear and easy to read.
3. Your report must be at least 2 pages.
4. Include links or references to any external datasets or sources you used.

## Evaluation Criteria:

* Complexity and relevance of your chosen use case.
* Effectiveness in leveraging the technologies learned in class.
* Clarity and organization of your report.