# MSBX5420 Final Project Design Document

V1.0

4/25/20

# Purpose

To create and deploy a series of applications utilizing several technologies utilized in the class over the course of the semester.

# Scope

Perform analysis of the COVID-19 dataset on the Leeds AWS cluster with data in an S3 bucket.

Provide several functionalities to do so, including:

1. Sentiment analysis
2. Word cloud
3. Word counts
4. Topic models

# Tools

* AWS EMR cluster
* AWS S3 cloud storage
* Jupyter Notebooks
* RStudio
* Github
* Python
* R
* Slack
* PuTTY

# Design Overview

## Requirements

1. The project shall be modular, with each piece of analysis in its own code file.
2. Data shall be stored on the S3 bucket for analysis
3. Each code file shall be able to be run on the EMR cluster, accessing data in the S3 bucket.
4. The code shall be horizontally scalable.
5. The code shall be hosted on Github.
6. Agile development methodology shall be adhered to during this project
7. Data must be ingested and saved as parquet format
8. Some analysis must be performed on the ingested dataset

## Process

1. Data preprocessing
   1. Using R, add additional information to the dataset for ease-of-use
   2. Using R, remove junk characters and clean the data
2. Analysis
   1. Ingest the preprocessed data
   2. Convert to parquet format
   3. Write modular code segments to perform analysis on the data
      1. Sentiment analysis
      2. Word cloud
      3. Word counts
      4. Topic models
   4. Unit test each code module
3. Deployment
   1. Convert code to Spark/deployable code
   2. Upload code and data to respective AWS location
   3. Run code on cluster

# Risks

* Have not used R with Spark, so learning curve there for integration
* Difficulties of performing more advanced analysis on Python and converting that to Spark
* Communication difficulties in a remote working environment

# Objectives

* Manage a project from start to finish using the Agile methodology
* Create, test, and deploy a project on AWS using Spark and other languages
* Learnt o use Github for project management
* Draw useful conclusions from analysis on project dataset
  + Visualize output
* Present findings upon completion

# Major Deadlines

1. Requirement phase – 4/12
   1. Requirements document deliverable
2. Design, Development, Test phase – 4/12
   1. SDD deliverable
   2. Code deliverable(s)
3. Deployment – 4/28
   1. Run code on cluster and get output
4. Presentation – 4/28
   1. Present findings to class