**MSBX 5420:** Unstructured Distributed Data Modeling/Analysis

**Team:** Mount Antero

**Members:** Alan Market, Alex McLaughlin, Matthew Hunter, Zhenkun Gao, Bonnie Tang

**Date:** 4/12/2020

**AWS Project Requirements**

**Project Dataset:** <https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26>

* This dataset contains details on news articles covering the current COVID-19 pandemic. The dataset includes seven variables for the title, authors, publishing date, description, main story text, and URL. Author refers to the news outlet from which the story was published contained in brackets and quotes. Title variables include the full text of the article title. Publish date includes both the date in YYYY-MM-DD format and time of day in HH:mm:ss format. Description includes a short description of the story while text includes the full text of the main story. URL provides the link to the article webpage.

**Functional Requirements:**

* Onboarding the data:
  + Loading the data into the class AWS cluster.
* Initial data analysis and exploration:
  + Checking the dataset for errors or missing fields that may need correction before analysis
* NLP:
  + Research and implement pyspark compatible NLP tools to calculate sentiment for each news article
* Produce final visualizations and reports:
  + Synthesize our project into a concise report and produce concise graphics for our presentation.

**Non-Functional Requirements:**

* Identify auxiliary datasets to enrich our primary analysis:
  + For our analysis, we are going to look for relevant data such as covid death rates to correlate with our news sentiment analysis to produce meaningful time series insights
* Develop our final report and presentation from our final analysis
* Maintain group collaboration
  + With the current restrictions on meeting in person, our group will be using tools such as Zoom and Google Drive to maintain remote collaboration

**Timeline:**

**April 12, 2020: Outline project requirements and create group GitHub repository**

**April 25, 2020: Submit design document to GitHub repository**

**April 28, 2020: Presentation and Deployment to large cluster**

**Dependencies:** Python, Pyspark, R, Sparklyr, Amazon Athena, AWS EMR, Git