The following shell commands can be used to run the job from a local machine without any setup. It will create a Docker container and submit the Spark job within the container.

- # Download my supplied zip file and unzip it
- # Open up terminal/shell and navigate within the unzipped folder
- **chmod +x build_and_run.sh**
- **./build_and_run.sh**

I chose Apache Spark as my data processing engine because it performs quickly and is especially helpful at scaling out for large sets of data ("linearly scalable"). Even if running on a single node, it will perform better than Pandas.

The PySpark script, **patient_analysis.py**, will run and achieve the following:

- Read in (ingest) the **100k_synthea_covid19_csv/conditions.csv** file as a Spark Dataframe and dynamically determine column names and data types
- Perform appropriate transformations to enable analysis of variance (ANOVA) to compare the most common symptoms for each of the following cohorts:
  - **A**: Covid-19 patients who a pregnant
  - **B**: Covid-19 patients who had asthma
  - **C**: Covid-19 patients who were smokers
  - **D**: Covid-19 patients who were not pregnant, did not have asthma, and were not smokers
- Write to TSV file under **output_data/symptom_comparisons/**

In order to interpret the results, please refer to the PDF, **results/symptom_comparisons.pdf**, I created based on the output TSV file.

- For all cohorts, the following are the overall most frequent symptoms to occur after a Covid-19 diagnosis:
  1. **Fever** (85-87% of patients in a given cohort)
  2. **Cough** (64-68%)
  3. **Loss of taste** (47-49%)
  4. **Fatigue** (37-38%)
  5. **Sputum** (32-33%)
- For smokers (cohort C), the following symptoms occurred with a higher percentage of patients than in other cohorts. In fact, this cohort included the most symptoms (8) with a 30% or higher chance following Covid-19 diagnosis.
  - **Hypoxemia** (33%)
  - **Pneumonia** (33%)
  - **Respiratory distress** (33%)
- Compared to the cohort D (not pregnant, no asmtha, and not a smoker), pregnant women (cohort A) were not significantly more at risk for various symptoms.