Use Apache Spark for analysis of your project data

Part A - Set up Apache Spark

- Sign Up for Databricks Community Edition https://accounts.cloud.databricks.com/registration.html#signup/community
- Read the Databricks User Guide https://docs.databricks.com/user-guide/index.html

or

- Download and run Apache Spark on a local machine or the cloud.

Part B - Structured Streaming (30 Points)

On *your project data*

- Set up a Structured Streaming analysis of your project data. https://docs.databricks.com/spark/latest/structured-streaming/index.html

or the below is an alternate to Structured Streaming (30 Points)

Part B - Alternate - Structured Streaming (30 Points)

Come up with a set of at least 10 SQL questions that involve joins, order by, group by and aggregate statements and implement them on your data using Spark SQL.

Part C - MLlib and Machine Learning (40 Points)

On *your project data*

- Apply a MLlib and Machine Learning analysis of your project data. https://docs.databricks.com/spark/latest/mllib/index.html

Part D - GraphX and GraphFrames or (30 Points)

On *your project data*

- Set up a GraphX and GraphFrames analysis of your project data. https://docs.databricks.com/spark/latest/graph-analysis/index.html

or the below is an alternate to MLlib and Machine Learning (40 Points) & GraphX and GraphFrames or (30 Points)

Part C & D - Alternate - Deep Learning with Apache Spark and TensorFlow (70 Points)

Implement a Deep Learning with Apache Spark https://databricks.com/blog/2016/01/25/deep-learning-with-apache-spark-and-tensorflow.html

Hyperparameter Tuning (30 Points): use Spark to find the best set of hyperparameters for neural network training.

Compare Apache Spark and TensorFlow to TensorFlow not on Apache Spark (40 Points).

Use Apache Spark to apply a trained neural network model using Apache Spark and not using Apache Spark. Write a report of the Pros and Cons