Part 1: The Machine Learning and Spark part

Question 1: Supervised Learning (4 marks)

You are going to train a [redacted] 'adult' dataset from Penn Machine learning Benchmarks:
https://epistasislab.gitl[redacted]ult.html data description

https://github.com/Epi[redacted]aster/datasets/adult the data (if you don't know how to open a gzip file, downl[redacted]or find whatever does that for mac)

Note that it's up to you [redacted]nto a training set, and a testing set

I suggest you do this with pandas and scikit learn

Then do it again on Spark (or presumably PySpark with pyspark.ml if you're sane), compare the results (hopefully obviously these should be pretty close to the same), and performance. Spark should be slower on a single machine, though not significantly, but I'd be interested to see here. If scikitlearn or whatever library you use isn't taking advantage of CPU cores properly Spark could be better.

Feel free to have a data preparation stage and an explanation of what you do to set everything up that's the same between both, and just the actual engine churning away at the analysis that's different.

Question 2: Unsupervised Learning (4 marks)

K-means clustering – You're going to do an unsupervised clustering algorithm on

http://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption

Again, just use scikit learn or the like. There are several papers based on this dataset so it shouldn't be too hard to understand.

Couple of notes. This dataset is reasonably large and it does have some missing/null entries. You need to （应该要先清理一下 dataset，如果 dataset 太大 可以 random sample 再 clustering）

As with Q1: Do this on Spark as well. （和第一题一样 要用 scikit learn library 然后 pyspark）

Question 3: Spark and SQL (2 marks)

You're going to rerun your SQL queries for the reddit data from A2, but this time load the data into spark and run the queries on spark.

You should try and do this two ways. 1: Managed table, where the data is in Spark and only in spark, and 2 and unmanaged table (you'll need spark + a database connector + an SQL database).

The point of this question is just to see how SQL queries in spark work, the actual queries themselves aren't very interesting. （看 sql 在 spark 上怎么 work 的）

Part 2: Choice

Do Only 1 of Qu

Question 4: Apache

(https://ci.apache.org/ s-stable/dev/datastream_api.html has a handy guide on how to do word count

Note: in the Workshop portion of the course we realised that the file output is happening on a on the worker thread. So you run the comman (flink run …. – output) the manager thread and the output file.

DataSets (e.g., filtering, mapping, joining, grouping)

1. Take a data set and break it apart at the file level (I suggest you just split it so that you have half of your columns in two different files), and using the dataset API re-merge the sets in Flink. Feel free to add a 'line number' (id) to both files to make the join easy.
2. Using the joined dataset from above, perform a simple reduction operation, but using the DataSet API on Flink.

Dynamic Queries

3. One of the cool things Flink can do is run SQL queries on streaming data. You can get the same results from SQL by repeatedly querying data, but the way SQL queries work means that will usually search entire databases, and or read the SQL data from disk, which is slow. With streaming you create a 'tumbling window' and only execute queries on that.
   So, what you're going to do, is load up your data into SQL on flink, and run a simple tumbling window query on it (examples might be find the average temperature in the last 10 seconds, or find the largest debt accumulated in the last 24 hours). Like in Spark, this just see how it works and the contrast with Spark, the queries aren't really important.

   https://medium.com/@mustafaakin/flink-streaming-sql-example-6076c1bc91c1 has a tutorial.

Question 5: Data security performance implications （如果选这题就只有 b 要 code，先 encrypt 再 decrypt，对比前后的 data size）

a) Describe how to properly encrypt tables in an SQL database to support both multiple admis, individual administrators not having access to all of the data, and encrypting data by row rather than by table (or the entire database). This is a theory question.
b) Design and perform an experiment evaluate the performance implications of encrypting and then decrypting data for use in a data pipeline. There are essentially two parts to this problem: 1 how do you encrypt the data using a reasonable encryption standard (there are libraries for

this, but you need to encrypt the data, then decrypt it, which takes CPU time) and then 2 how much data there is (encrypted data is usually larger than non-encrypted data).

c) There are a significant number of data breaches happening (either from failure to properly secure data or failure to have and follow data governance guidelines). https://tech.co updated-list for example has a tracker for the current year.

Examine at lea ch and discuss what failure occurred and what could have been done to p you will need to look at breaches that happened earlier in the year or last as been issued. There are several academic papers looking at this topic as is only work 3 marks total so this doesn't need to be a novel, but som rds would explain it well enough.