Data Engineer - Ingestion and Simple Analysis

Locate the Enron email data on AWS. The data is big (210 GB), so a challenge you need to solve is to work with this data without trying to download all of it to your own machine. Hints:

1. Create an EC2 instance with sufficient RAM and processing power for what you are trying to do

2. Attach an EBS volume sized 210 GB with the snapshot ID snap-d203feb5 while defining your EC2 instance.

3. Once you log in remotely to your EC2 instance (Linux command line) you can mount the EBS and you will find all the ENRON data there. You can choose the file format you find easier to work with: PST or XML.

4. It’s OK to test your code on sample data locally, but you should try to run your code on the entire data set.

Use AWS documentation if required.

Use Scala, Python or Java to answer the following questions:

1. What is the average length, in words, of the emails? (Ignore attachments)

2. Which are the top 100 recipient email addresses? (An email sent to N recipients would could N times - count “cc” as 50%)

Send us the code to produce the output via github, dropbox (or similar). We should be able to build and run and test your code. It may help to provide a readme file or other documentation.

If you need to make assumptions, make them explicit.