Nit

Brandon Haynes authored 17 hours ago

1b371ada

mini3.md 1.97 KB

Mini-Homework #3

Due date: December 10, 2019

Objectives:

Run queries on Apache Beam.

Assignment tools

Apache Beam installed on an EC2 node (best), your local machine, or on a Google cloud account.

What to Turn In

Submit answers to the questions listed below. Submit everything as a single markdown, notebook, or PDF.

How to submit the assignment

In your GitLab repository, you should see a directory called `Homeworks/mini-hw3`. Put your report in that directory. Remember to `git add, git commit, and git push`. You can add your report early and keep updating it and pushing it as you do more work. We will collect the final version after the deadline passes. If you need extra time on an assignment, let us know. This is a graduate course, so we are reasonably flexible with deadlines but please do not overuse this flexibility. Use extra time only when you truly need it.

Assignment Dataset

Environment

In this Assignment you will need access to a node running Apache Beam. You can do this on an EC2 instance (see the section notes), install Beam on your local system, or use a Google Cloud account.

Dataset

You will also need to download following data files:

- hamlet.txt
- muchado.txt

Ingest both of these datasets in your Beam system.

Questions (20 points)

- Count the *total number of words* in both the `hamlet` and `muchado` datasets.
- Count the *average number of letters per word* in both the `hamlet` and `muchado` datasets.

Submit the *time* required to generate the results for each question and dataset, as well as the results you obtained. Also indicate whether you used an EC2 instance (if so, indicate the instance type), your local machine, or a Google Cloud account.