MAS DSE 230
Scalable Analytics
Session 2 Assignments

Mai H. Nguyen

SESSION 2 ASSIGNMENTS

- Programming Assignment 2
  - Due Thursday 2022-04-28 at 11:59pm Pacific Time
  - 10 points
- Project Proposal Presentations
  - Presented in class during Session 3 (Friday 2022-04-29)
  - 10 points

PROGRAMMING ASSIGNMENT 2

- WordCount on Amazon Reviews
  - File on Canvas: PA2_Starter.ipynb
  - Use PySpark DataFrame (not RDD)
  - Find top 99 words starting with ‘s’ based on count
  - Find mean and standard deviation of execution times over 3 runs
    - Using 1 core, 2 cores, and 4 cores
- Submit
  - Jupyter notebook (.ipynb)
  - Word count results (.csv): Top 100 lines
    - Format: word, count
  - Execution times results (.csv)
    - # cores, time0, time1, time2, mean, stdev
- Due Thursday 2022-04-28 at 11:59pm Pacific Time

WORDCOUNT OUTLINE

- Read data into DataFrame
- Remove punctuations and convert to lower case
- Split data into words
- Put each word in a separate row
- Filter out words as needed
- Group rows by word to count the number of occurrences for each word
- Sort words by count
- Don’t forget
  - Copy data to HDFS
  - Copy results from HDFS

GETTING EXECUTION TIMES

- In notebook, execution time is printed out in cell before Spark session is stopped (next to last cell)
- Need to *restart the kernel* and run all cells without stopping to get accurate execution time:
  - Run -> Restart Kernel and Run All Cells
- Find mean and standard deviation of execution times over 3 runs for
  - 1 core, 2 cores, and 4 cores

Specify number of cores.
'*' uses all available cores

```
import pyspark
from pyspark.sql import SparkSession

conf = pyspark.SparkConf().setAll([
         ('spark.master',   'local[2]'),
         ('spark.app.name', 'PySpark WordCount')])
spark = SparkSession.builder.config(conf=conf).getOrCreate()
```

PROJECT PROPOSAL PRESENTATIONS

- Team sign-up sheet:
  - https://docs.google.com/spreadsheets/d/10UQRhEIlx7qctyVZ1DCkbArj6WFXMgeGe7Zr5R0RslM/edit?usp=sharing
  - Sign up by Monday 2022-04-18 at 11:59 pm Pacific time
  - 2-3 people per team
- To include in your presentation: problem to address, dataset description, analysis task planned, insights you hope to gain, and potential challenges with data and/or task
- All team members must present
- Q&A: At least 3 questions from audience
- Approx. 10 minutes: 7 minutes for presentation + 3 minutes for Q&A
- Presented in class during Session 3. Presentation order will be provided.
- See Project Description and Project Rubric on Canvas