CSE4/587 Data-intensive Computing Spring 2019

LAB 3: DATA ANALYTICS USING APACHE SPARK: B. RAMAMURTHY

OVERVIEW:

The hands-on practical learning components of the course comprises two types of activities: labs covering one or two knowledge units (skills, competencies) of data-intensive computing and a single term project serving as a capstone covering the entire data pipeline. In the first half of the course we learned data analytics and quantitative reasoning using R language. In the second half we focused on big data approaches for data analytics with Hadoop MapReduce and Apache Spark. In this lab, we will work on understanding concepts related to data analysis using Apache Spark [1].

In Lab1 we learned to analyze data using R language and R Studio. In lab2, we explored approaches that deal with big data, especially text data, using the Google's MapReduce algorithm. In Lab 3 we will explore processing graph data using Spark [1]. Here is a chance to show case your knowledge in Apache Spark data pipelines and big-data analytics. You can apply the model you build here, to numerous applications.

WHAT TO DO?

We are learning concepts of Spark model for data analytics. We are also exploring hands-on exercises on Apache Spark during lecture. In this third project, instead of coding a Spark project, you will write a problem statement identifying a use case and provide details like a answering an RFP (request for proposal). This document provides guidelines. You cannot wait till the last minute, like you did in the last labs since there is catch (see below).

LAB DESCRIPTION

In this lab work you will

1. **Explore** the Apache Spark framework and programming: spark context (sc), dataflow operations in **transformations, actions, pipelines and MLib.**
2. **All these without solving a problem. What, no coding?!!**
3. Since there is no coding this lab will count only 50% of the original weight of "tentative" grading guidelines given in the first day handout.
4. **You will install Spark in your computing environment.** It is already given in the VM pinned to Piazza. Please find a way to get an instance of Spark. **Do not wait till the last minute.**
5. You will study Spark using the references given, explore the code examples; execute the code examples given (copy and paste). Collect snippets of code (including MLLib use) that you tested in a folder, zip them and upload. **Do not wait till the last minute. Do not copy the code I have given, be creative and obtain one from your own exploration.**

   Example 1:

   ```
   text_file = spark.textFile("hdfs://...")
   text_file.flatMap(lambda line: line.split())
      .map(lambda word: (word, 1))
      .reduceByKey(lambda a, b: a+b)
   ```

Example 2:

```
pyspark
data= [1, 3,4,5,6]
distdata = sc.parallelize(data)
rdd =sc.parallelize(range(1,4).map(lambda x: (x, “a” * x))
rdd.saveAsSequenceFile(“mydata”)
sorted(sc.sequenceFile(“mydata”).collect())
```

Example 3: using MLLib

```
./bin/spark-submit examples/src/main/python/pagerank.py data/mllib/pagerank_data.txt 10
```

DETAILS:

1. Use case for Spark and MLLib: Decide on an appropriate use case for Spark by studying the literature, keep track of all your research material for reference in your submission. Requirement: choose a unique use case for full credit. You can check if yours is unique and not already taken by visiting your TA. You can lock up the use case by choosing it first. In technology, whoever who come to the gate first gets it. If you wait too long, say until the due date, and 10 teams have the same use case like “basketball” in Lab2, you share the score. 100/10= 10% only. So hurry up and lock up your use case. Treat this like a contest or hackathon.
2. Title: Decide a suitable title for your use case: Choose a representative title for your use case. This is what you will provide your TA when you visit him or her during recitation and office hours. You can meet me too! I can verify and help you lock up your use case for you.
3. Abstract: 100 word abstract. Executive summary: why, what, when, how Spark? You’ll need title and abstract minimally to lock up your use case.
4. Problem statement: What is the problem? How do you plan to solve it using Spark? What is its impact? What does it improve? 100 words approximately.
5. Solution (model) and design document: Architectural block diagram around Spark, explanation of your solution. Big data /data science pipeline HDFS, Spark, Tableau etc. Language you recommend for development. Any support platforms that you can suggest. Any vignettes that will be useful. What data source do you plan to use? 1-2 pages
6. Expected outcome: Charts and visualization concept diagrams. ½ page – 1page
7. Summary and takeaway points. ½ page max (use a list)
8. References

SUBMISSION DETAILS:

1. **Make sure you submit a single document contained all the details. Any outside information have to be attributed. Please enter your name and your team mate’s name with the person number, just below the title.**
2. **TAs and I will have running list of projects submitted (/taken/gone). Think of this like a contest, if you don’t get there first, you’ll have to work hard. All the cool problems will be TAKEN.**
3. Online submission: do not email me: (-10 points) if you email me the document.

DUE DATE: ONGOING, CONSULT FOR TOPIC/TITLE DURING RECITATION AND OFFICE HOURS FOR TOPIC LOCKUP.

FINAL DEADLINE: 5/5/2019 BY 5.00 PM. YOU MAY NOT RECEIVE FULL CREDIT IF YOU WAIT TILL THE LAST MINUTE. (IF SOMEBODY ELSE HAS THE SAME USE CASE, YOU AND THAT TEAM SHARE THE FULL CREDIT.)
Last week of recitations for review and grading and redoing if needed.

This is how it will work:

Step 1: Decide Use case, title and abstract: meet TA during office hours or recitation and get registered on the Google sheets we have. This locks up your topic. If another team chooses it after you register, they cannot since the topic is already taken. That team has to look for another topic.

Step 2: Work on the details required as specified in the Description and Details section, and submit it as pdf and a single document.

REFERENCES:

1. Apache Spark. http://spark.apache.org/, last viewed 2019.
2. Spark Programming guide: https://spark.apache.org/docs/1.2.0/programming-guide.html, last viewed 2019.
3. RDD Programming guide: https://spark.apache.org/docs/latest/rdd-programming-guide.html, last viewed 2019.
4. Spark Quickstart: https://spark.apache.org/docs/latest/quick-start.html, (interactive), last viewed 2019.
5. Cloudera Spark Manual: https://www.cloudera.com/documentation/enterprise/5-12-x/PDF/cloudera-spark.pdf, last viewed 2019.
6. S. Ryza, U. Laserson, S. Owen and J. Wills. Advanced Analytics with Spark. O'reilly, 2015. E-copy available in UB Library and online.

Addendum:

More details: What to submit and when?

Meet your TA during recitation or office hours or in lecture on May 2nd to get the topic approved. Once it is approved:

1. Redesign the **title** to be representative of your idea of topic that is going to be used for Spark analytics. **DO NOT COPY** VEBATIM any material online.
2. Write your team members names and person numbers.
3. Write an abstract of what are you solving, why Spark, where is data, how will you access data, what is the meta data; Don't say that your analytics will predict lung cancer, for example. **DO NOT COPY** VEBATIM any material online.

This is less than a page, about ½ page. Submit Lab3Abstract.pdf **on timberlake** (not on UBBox, or Google drive). Due date: 5/3 6pm. Hard deadline. Don't say your laptop crashed, bandwidth is poor. Don't pull any of these age-old excuses! No few minutes late, etc. This is a hard deadline.

Now prepare a report of how you are going to carry out what you claimed in your abstract.

1. Title: Same as above
2. Abstract: Same as above: 100 word abstract. Executive summary: why, what, when, how Spark?
3. Problem statement: What is the problem? How do you plan to solve it using Spark? What is its impact? What does it improve? 100 words approximately.
4. Solution (model) and design document: Architectural block diagram around Spark, explanation of your solution. **Data pipeline; algorithms used;** Big data /data science pipeline HDFS, Spark, Tableau etc. Language you recommend for development. Sample data; Any support platforms that you can suggest. Any vignettes that will be useful. What data source do you plan to use? What algorithms do you plan to use. Proof of concept for these: Look at the snippet example I have provided. Somebody asked on Piazza if they have to submit all the examples running: If anyone does that, they don't understand the problem: **They will get a 0**. If you need pagerank, but show wordcount as example, these people will also get a 0;
   Be creative; make sure your solution is feasible and novel; Don't ask TAs for topics!
   Expected length of this: about 2 pages. **DO NOT COPY** verbatim from any source.
5. Expected outcome: Charts and visualization concept diagrams. ½ page – 1page
6. Summary and takeaway points. ½ page max (use a list)
7. References

This report is expected to be approximately 4 pages. Submit Lab3Report.pdf **on timberlake** (not on UBBox, or Google drive). Due date: 5/5 6pm. **Hard deadline**. Don't say your laptop crashed, bandwidth is poor, etc. There is only one file. Don't say, I forgot this para or that. Common people, this is the first time since the last century I'm getting such bad excuses. Don't pull any of these age-old excuses! Not even few minutes late, is allowed. This is a hard deadline.