Examination 2: Study Guide

Last updated: Monday, October 11, 2021

Scope
All materials to date.
Suggestion: You are urged to study for the exam by constructing answers to each of the following questions. Consider it as a study guide. If you do not, you may not have time to complete the exam.

Summary

The example will cover an Spark ML, some cluster development including Docker, Hadoop's execution of tasks via YARN, HDFS, and some Map Reduce.

PySpark ML

1. What are the common phases of data mining program?`
    a. Specifically, what task (and in what order) are required to transform raw data into a helpful visualization that can guide a decision?
2. Be able to use a mix of magic commands (%sh, %sql) to review your data and then process it
    a. Use pyspark to read data
    b. Use SQL to review data
    c. Use pyspark to run ML
3. Be able to read a PySpark ML program
    a. E.g., Logistic regression, linear regression, etc.
    b. See the readings and the class notebooks
    c. Be able to read PySpark code with regression models
4. What are some example of ml.feature classes?
5. What is the purpose (i.e. context of use) for StringIndexer, OneHotEncoder, VectorAssembler, Bucketizer, Tokenizer, StopWordsRemover,?
6. In the general context of data mining, what is the purpose of creating features (feature engineering)?
7. Generally speaking, how can the accuracy of a mining model be improved?

Pipeline

8. What is a ML pipeline? What purpose does it serve?
9. What object types (classes) are in a Spark ML pipeline?
10. In a pipeline, what do these methods do: transform(), fit()?
11. On an estimator object, what methods are called by the pipeline, during a pipeline.fit()?
12. On an transform object, what methods are called by the pipeline, during a pipeline.fit()?
13. What is a parameter grid?
14. What is the purpose of CrossValidate?
15. For a given data set and one model (e.g., linear regression), how would I run the model with various parameters so that I could select the parameters for which the model performs the best? That is, how do I program hyperparameter tuning in PySpark?

a. Consider the above question, but now I would like to run multiple models, each with their own parameters, and then select the best model with the best parameters. How would I do such hyperparameter tuning across multiple models?

Docker

16. What is a Docker container?
17. Compare and contrast the architecture and resource demands of Docker vs virtual machines (such as VMware).
18. Define these Docker terms: docker client, docker host, docker registry, Dockerfile, Docker image, Docker container [available here]
19. How does Docker simplify building and maintaining applications?
20. How do Docker containers enable cloud computing? Given an example of a cloud computing service that relies on Docker containers. (You have used such a service in this course.)

HDFS (White book)

21. What are key features that the framework provides for parallel programming?
    a. Consider replication, fault tolerance, etc.
22. Terms: node, rack, block, split, task
23. What is the relationship between task and block?
24. What do these nodes do: Name node, data node?
25. What do these components do: Resource manager, Application manager?
26. What's a node container?
    b. See YARN notes.
27. By what steps does a client read blocks from the HDFS?
    c. See the diagram in the White book
28. How is the closest block determined (consider a metric)
29. What is block replication? Why is it done?
    d. Once a block is written, how does the data travel to the next node? See the diagram in the White book

MR Applications

30. Explain shuffle and sort
    e. What occurs on the Map node and on the Reduce node?

General questions

31. What kinds of computing problems are appropriate for the application of Hadoop and/or Spark? Compare and contrast the traditional application computing infrastructure with the Hadoop ecosystem as it applies to either: (1) processing transactions such as orders and their credit card payments, and (2) mining web pages for predicting, say the influenza or stock prices.
32. What distinguishes a Big Data application from a more traditional application?
33. You've been hired by a bank. They need to process loan applications. Would a Hadoop/Spark system be a good choice?
34. You've been hired by a bank. They need to build a system that can classify loan applications as potentially fraudulent. Would a Hadoop/Spark system be a good choice?

Databricks labs

35. Know how to use the following in a Databricks notebook:
    a. wget, spark.read.csv, show, summary().show(), printSchema, withColumn, randomSplit, RFormula, fit, transform, select, selectExpr
    b. drop, filter, where, count,

Homework 1 - 3

36. Be able to use withColumn to cast a data type
37. Be able to filter data using filter() or where()
38. Be able to join two DataFrames
39. Be able to write a UDF (for dataframes) and register a UDF for SQL
40. Know conceptually how regression works in PySpark. Specifically,
    a. What does RFormula do?
    b. What does LinearRegression do?
    c. What is contained in Pipeline?
    d. What is ParamGridBuilder used for?
    e. What is CrossValidator used for?

Homework 4 -5

It is assumed that you watched the videos and tried to run the labs.

41. What do these commands do (generally):
    a. docker build
    b. docker images
    c. kubectl logs
    d. gcloud auth login
    e. kubectl config use-context
    f. docker push gcr.io/spark8795/wrobinson/airbnb:1.0
42. When configurating a GCP Kubernetes cluster (GKE), can you
    g. Configure the kind of computer (CPUs, memory)?
    h. Configure the number of nodes?
43. Is it possible to run the same PySpark code, without modification, in these different environments?
    i. PySpark shell
    j. PyCharm Python terminal
    k. Docker for Desktop Kubernetes cluster
    l. GCP Kubernetes cluster