Examination 1: Study Guide

Last updated: Wednesday, October 27, 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.

Between 20 and 25 questions. Some multiple choice & matching and some open questions (short answer).

Summary

The example will cover an introduction to the role of Big Data in business, Data Engineer, Data Analyst, and introductory PySpark

Big Data Infrastructure Introduction

1. Compare and contrast two big data roles: data scientist and data engineer
2. Big data technologies (1) created to support web search, (2) process datasets too large or complex for traditional programming
3. What are the 3 V's of Big Data?
4. Define Big Data
5. Why is big data a recent phenomenon? What has enabled it?
    a. Consider the availability of open-source, commodity cloud computing is meeting the demand to analyze voluminous (IoT, social) data
6. ~~In a few sentences, describe a Big Data application, such as cross-selling~~
7. What do companies use big data?
    a. Productive, profitable, decrease expenses, innovation, etc.?

Apache Spark

8. What is a cluster? What are its elements?
9. What are the important differences between and RDD and a DataFrame?
10. In Spark execution,
    a. What is the role of the driver?
    b. What is the role of an executor?
    c. ~~What is the role of the cluster manager?~~
11. Why are PySpark UDF's slow to execute?
12. What is a narrow transformation?
13. What is a wide transformation?
14. What is lazy evaluation?
15. In Spark, what is an action? Give examples.

DataFrame

16. In Spark, what is a Data Frame?

    a. Be able to read and write simple statement about DataFrames
    b. Be able to join two DataFrames
17. What does the Spark Catalyst do?
18. Generally, which kind of code runs faster, RDDs or DataFrames?
19. Why are joins slow (especially in Spark)?
20. If the data is divided into two partitions, then how many Spark tasks will there be? What is the correspondence of partitions to tasks, if any?
21. Which is generally faster, groupByKey or reduceByKey? Why?

PySpark ML

22. 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?
23. Be able to use a mix of magic commands (%sh, %sql) to review your data and then process it
    b. Use pyspark to read data
    c. Use SQL to review data
    d. Use pyspark to run ML
24. Be able to read a PySpark ML program
    e. E.g., Logistic regression, linear regression, etc.
    f. See the readings and the class notebooks
    g. Be able to read PySpark code with regression models
25. What are some example of ml.feature classes?
26. What is the purpose (i.e. context of use) for StringIndexer, OneHotEncoder, VectorAssembler, Bucketizer, Tokenizer?
27. In the general context of data mining, what is the purpose of creating features (feature engineering)?
28. Generally speaking, how can the accuracy of a mining model be improved?

Pipeline

29. What is a ML pipeline? What purpose does it serve?
30. What object types (classes) are in a Spark ML pipeline?
31. In a pipeline, what do these methods do: transform(), fit()?
32. On an estimator object, what methods are called by the pipeline, during a pipeline.fit()?
33. On an transform object, what methods are called by the pipeline, during a pipeline.fit()?
34. What is a parameter grid?
35. What is the purpose of CrossValidate?
36. 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?

Databricks labs

37. Know how to use the following in a Databricks notebook: wget, spark.read.csv, take, show, summary().show(), printSchema, withColumn, randomSplit, RFormula, fit, transform, select, selectExpr
38. Know how to use the following in a Databricks notebook: ~~unionAll~~, spark.createDataFrame, drop, filter, where, count, explode, select star (e.g., select("complex.*"))

Computing Terms

39. Know these terms
    a. Bandwidth, latency, scale out, scale up, contention, database shard, throughput, response time, speed up, scale up, data skew
40. What factors limiting speedup and scaleup, making it sublinear?
41. ~~Know generally how does Spark work in terms of data and program distribution (See slide 51 of 2.2 Parallel computing)~~

Homework 1 – 3 (some repetition with above)

1. Be able to use withColumn to cast a data type
2. Be able to filter data using filter() or where()
3. Be able to join two DataFrames
4. Be able to write a UDF (for dataframes) and register a UDF for SQL
5. 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?
6. Using Pipeline, ParamGridBuilder, and CrossValidator is possible to train multiple models, each with their own parameters, and then select the best overall model?
    a. Conceptually, how is this done in PySpark