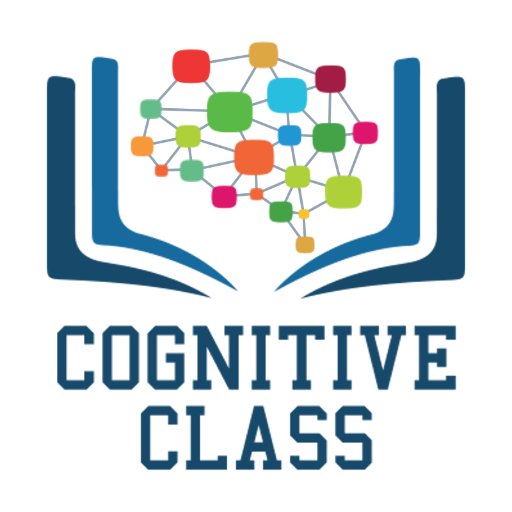
Supplement Manual Data Engineer and Data Scientist Part 2

Author: Peter Odenhoven

Version: 1.2

Date: Thursday, 17 January 2019

Course year : 18-19

Version control

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ver.** | **Status** | **Date** | **Author** | **Changes** |
| 1.0 | concept | 2018-10-01 | P.Odenhoven | Split the manual in Part 1 and Part 2 |
|  |  | 2018-10-04 | Z.Efendijeva | Mimimize assignment  Choice between   * Either parallel processing with Spark * Or Neural networks |
|  |  | 2018-10-06 | S.Robben | Minimize program, editorial comments |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Table of Contents**

[1. Examination and tests in block 2 4](#_Toc529285457)

[2. Test 3: Individual Assignment II 5](#_Toc529285458)

[3. Test 3: Rubric Data Engineer and Data Scientist individual assignment II 7](#_Toc529285459)

[4. Test 3: Checklist Report 8](#_Toc529285460)

[5. Test 4: Theoretical exam 9](#_Toc529285461)

[6. Weekly planning Part 2 10](#_Toc529285462)

[7. List of needed Software ( not at all trying to be complete …) 14](#_Toc529285463)

# Examination and tests in block 2

This chapter describes the examination and grading for the course Data Engineer and Data Scientist.

The course is rewarded with 8 European Credits ( studiepunten) if and only if the students has passed 4 test, with 4 separate grades:

Block 2

1. Data Engineer and Data Scientist: Test 3: Individual assignment II + report, 1
2. Data Engineer and Data Scientist: Test 4: Overall Theoretical exam 2 +

3

Important notes:

* All 4 separate grades for the tests should be sufficient, i.e. ≥ 5.5
* The overall theoretical exam will cover all the topics from Data Engineer and Data Scientist, including knowledge on SQL and NOSQL database, Hadoop, Map Reduce and Spark. In the weekly program all learning goals are explicitly listed. This list should be your checklist, when preparing for the theoretical exam.
* The individual assignments are individually assessed in a 10 minutes window at the end of each block. Students are only allowed to attend the assessment after the report is uploaded.

Important weeks for the examination/ tests

|  |  |
| --- | --- |
| Week 11-20, Block 2 | Description tests |
| week 18 | Assessment Individual assignment I ( = test 3) |
| week 19-20 | Theoretical exam ( = test 4) |
| week 19-20 | Resit Assessment Individual assignment I ( = test 1)  Resit Assessment Individual assignment II ( = test 3) |

# Test 3: Individual Assignment II

In this assignment we’ll use the same dataset as before. Simulating processing and analysing of a Big Data set on your machine can be done by using several libraries. The use of these libraries and their purposes will be the topic of each lesson during this block. Main goal of the second assignment can be stated as:

*“Demonstrating how machine learning can be done in a Big Data Environment”*

So, ee want you to use

* Either **all the data** from the Kaggle dataset on hotel reviews
* Or a dataset to be discussed with the teachers

The mandatory deliverables are

**1 Several R scripts or Python scripts**

1. To obtain an attractive visual representation of all the data in the dataset. With visual interactive elements to support the socalled Visualisation mantra:

* Overview: Gain an overview of the entire collection
* Zoom : Zoom in on items of interest
* Filter: filter out interesting items or filter in interesting items
* Details: on demand: Select an item or group and get relevant information accordingly

1. To simulate big data and RAM problems, additional libraries are used
   * In case of R , for instance the library FFBASE
   * In case of Python, for instance the library PyTable

After some initial selection cleaning the result should be written away as a Review\_pos.csv and Review\_neg.csv

1. All of the dataset is stored in a NOSQL database, for instance MONGODB. A live connection to filter data during the process of running the script should be implemented:
   * There should be a collection containing all of the data of the Kaggle dataset having the following structure
     1. Hotel\_Address text,
     2. Hotel\_Name text,
     3. Lat double,
     4. Lng double,
     5. Average\_Score double,
     6. Total\_Number\_of\_Reviews int,
     7. Additional\_Number\_of\_Scoring int,
     8. Reviewer\_Nationality text,
     9. Review\_Date text,
     10. Review text,
     11. Review\_Word\_Counts int,
     12. Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given int,
     13. Reviewer\_Score double,
     14. Tags text,
     15. Sentiment int, additional field indicating a positive Review 1, or a negative review 0
   * There should be a collection of balanced set of reviews, for instance a collection consisting of 10.000 positive and 10.000 negative reviews having a least the following structure
     1. Review text,
     2. Sentiment int
2. At least one more or less advanced feature should be implemented.
   * 1. Either the student simulates parallel processing power. For example, a sentiment analysis of the hotel reviews is solved by using Spark.
     2. Or the student demonstrates state-of -the-art algorithms. For example, a prediction on the sentiment of a review using the length of the review be built using Keras
        1. **A compact report** , meeting the following requirements , the report should be
3. In correct English or Dutch
4. Containing relevant screenshots of codes
5. Containing relevant screenshots of the visualisation
6. Clarify the process of Model Building
   1. What kind of Machine Learning algorithm was used?
   2. Why was it needed?
7. Communicating the results

# Test 3: Rubric Data Engineer and Data Scientist individual assignment II

|  |  |  |
| --- | --- | --- |
| Studentnumber: | Studentname: | Grading: |

Insufficient overall, if one of the following is missing:

* No interactive visualisation
* No separate Script to simulate RAM problems
* Not stored in MongoDB, no live aggregated query when making a selection
* No report that meets the standards, see checklist

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Insufficient 0 - 25 points | Marginal 26 - 50 points | Good 50 - 75 points | Excellent 75 - 100 points |
| *1.Visualisation* |  | Only one of the minimum requirements is implemented ( i.e. the mantra of Visualisation) | All of the minimum requirements are implemented ( i.e. the mantra of Visualisation) | Additional to *Good*:  Sophisticated interactive elements are implemented |
| 1. *Simulating Storage problems* |  | Some simulating is done, but only as an illustration | A simulation is done resulting in meaningful subsets to be used in the analysis | Libraries are used to upgrade performance of both memory and CPU time |
| *3. DB Storage NOSQL* |  | There is a live query for collecting data, however is a simple query, no aggregate is used. | There is a live query for collecting data, and the simple query is more or less complex, for instance an aggregate is used. | There is a live query for collecting data using the map reduce structure |
| *3.Simulating Parallel Processing / State of the art algorithms* | There is no simulation of parallel processing and the student has no idea about neural network model  to be used. | There is a simulation, data is copied to a spark source, however no machine learning is implemented. | There is a simulation, data is copied to a spark source, only one machine learning algorithm is implemented successfully. | There is a simulation, data is copied to a spark source, several machine learning algorithms are implemented successfully and compared |
| Student only knows to describe the neural network model involved in the script. | The student can explain model involved and some of the parameters involved. | Student has done some research on the topic, and is able to fine tune the model |
| *5.Coding* | Student cannot explain any of different statements in the code used to build a model. | Student can explain only the basic statements in the code behind the model. | Additional to *Marginal*:  Student knows how to explain all the ins and outs of the pieces of code involved. | Additional to *Good*:  Advanced tweaking of the parameters involved in the used classifiers has been used |

# Test 3: Checklist Report

* Title page
* Table of contents (incl page numbering)
* Summary/abstract
* Introduction
* Background
  + Contains theory about the models
* Methods
  + Can contain multiple subsections
  + Screenshots of code, only when relevant
* Results
  + Contain relevant plots
* Conclusion and/or recommendations
* Reference list
  + Choose a consistent reference style: APA or IEEE
* Optional: preface, footnotes, appendices, list of symbols, glossary)
* Report is written in understandable and correct Dutch or English

**Notes:**

This checklist is used to check the completeness of the report, not whether the parts are accurate.

**Only when your report is complete, you will be invited for the final assessment!**

This checklist is derived from the ‘Beoordelingsformulier Onderzoeksrapport research skills/stage’.

If you need advice on how to write a report: tips can be found via the course ‘Reseach skills’ and online via the internship- and graduation manuals. (Accessible via VLO or A-Z).

# Test 4: Theoretical exam

The theoretical exam involves all theory and obtained skills for the module Data Engineer and Data Scientist. It includes also the theory of the Mathematics lessons.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Learning objectives |  | Type vragen | | |  |  |
| Reproduction | Production | | |  |
| Understanding  Remembering | Applying | Analyzing | Evaluating | Creating |
| 1. Understand certain coding lines (for example a MongoDB query) | 5% |  |  |  |  | 5% |
| 1. Understand the mathematics involved in most of the data scientist algorithms |  | 10% |  |  |  | 10% |
| 1. Make motivated choices regarding the specific parameters within an algorithm |  | 10% | 5% |  |  | 15% |
| 1. Understand the outcomes of an algorithm used to analyse a dataset |  | 10% | 10% |  |  | 20% |
| 1. Make motivated choices regarding the use of several Data Scientist algorithms |  |  | 10% | 10% |  | 20% |
| 1. Advise on storage issues |  |  |  | 10% |  | 10% |
| 1. Advise on a Data Engineering and Data Scientist stack |  |  |  |  | 10% | 10 % |
| **Total** | 20 % | 25 % | 25 % | 20% | 10 % | 100% |

Moreover

* All questions will be open questions, no multiple choice.
* Questions are posed in English, students are allowed to answer in Dutch
* No calculations need to be done, so no calculators are needed
* Learning objective 5-7 are tested with a case study

# Weekly planning Part 2

Each lesson has an indication on the topic

* DE = Data Engineer , all learning goals are mandatory for the theoretical exam and some of them apply to the individual assignment
* DS = Data Scientist , all learning goals are mandatory for the theoretical exam and most of them apply to the individual assignment
* Workshop, handy hands-on practical for either the individual assignment or the project task

Learning goals not covered by slides on DataCamp are marked with an asterix (\*) and should be covered by additional slides.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Block 2 |  |  |
|  | Week 0 |  |  |
| **Workshop**  **DS** | Data Visualisation  DataCamp R   * + Building Web Applications in R with Shiny   DataCamp Python   * Interactive Data Visualisation with Bokeh | Hands-on exercise | Know how to plot variables into a   * Scatterplot * Barplot * Histogram * Pie chart * …   Know how to add   * Titles * Colors * Legend * …   Know how to add interactive elements like   * Filters * Hoovering effects * Drill down effects   … |
|  | Week 1 |  |  |
| **Workshop**  **DE** | Using R with Big Data and how about Python  DataCamp R   * Scalable data processing in R   Python   * PyTable | Hands-on exercise | Knowing   * The traditional limitations of R/ Python * How to surpass these limitations   \* https://www.analyticsvidhya.com/blog/2016/05/data-table-data-frame-work-large-data-sets/ |
|  | Week 2 |  |  |
| **Workshop**  **DE** | Storage NOSQL: MongoDB | A JVM is needed  Hands – on workshop | Knowing   * Why do we need MongoDB? \* * What is meant by scheme less? \* * How to getdata in and out of MongoDB   + Querying in a MongoDB shell   + Using MongoDb aggregate * How to make an connection with MongoDB with R or Python   \* 01 MongoDB Introduction 1819.pptx  \* 02 MongoDB MapReduce 1819.pptx |
|  | Week 3 |  |  |
| **Workshop**  **DE** | Hadoop  Cognitive Class   * Hadoop 101   + Module 1, Intro Hadoop   + Module 4, Components of Hadoop | Use  Cognitive Class account  Hands – on workshop Azure Machine Learning | Knowing   * The main components of Hadoop   + HFDS   + MapReduce   + HIVE   + PIG   + Sqoop   + Flume * How to explain the MapReduce concept * Why we need the MapReduce concept |
|  | Week 4 |  |  |
| **Workshop**  **DE** | Spark  Cognitive Class   * [Spark Fundamentals I](https://cognitiveclass.ai/courses/what-is-spark/)   + Module 1   + Module 2   DataCamp R   * Introduction to Spark in R using sparklyr   DataCamp Python   * Introduction to PySpark | Use  Cognitive Class account  Hands – on workshop | Knowing   * How to explain the purpose of Spark * The use of the RDD ( Resilient Distributed Dataset) * How to run some Spark examples * How to use various Spark libraries such as SparkSQL and MLlib |
|  | Week 5 |  |  |
| **Workshop**  **DS** | Neural Networks  Cognitive Class   * Deep Learning Fundamentals DeepLearning.TV ML0115EN   DataCamp R   * Tutorial: keras Deep Learning in R   DataCamp Python   * Deep Learning in Python | Use existing  Cognitive Class account  Demo | Knowing how to explain the concept of   * forward propagation * activation functions * weights * gradient descent * backward propagation   Know how to evaluate your model |
|  | Week 6 |  |  |
|  |  |  |  |
|  | Week 7 |  |  |
|  |  |  |  |
|  | Week 18-20 |  |  |
|  | Theoretical exam ( = test 4) |  |  |
|  | Resit Assessment of Individual assignment II ( = test 3) |  |  |

# List of needed Software ( not at all trying to be complete …)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Type software | Mandatory | requires | url | remark |
| 1. JAVA | JDK/JRE | YES | Oracle account | http://www.oracle.com/technetwork/java/javase/downloads/jdk10-downloads-4416644.html | Java Virtual Machine |
| 1. MySQL | DataBase | YES | Oracle account | <https://dev.mysql.com/downloads/> | Community download. On this page you can also find connectors and the workbench |
| 1. MySQL shell | shell | YES |  |  |  |
| Either | MySQL workbench | optional |  | <https://dev.mysql.com/downloads/> | Sometime s it crashes .. |
| Or | HeidiSQL | optional |  | <https://www.heidisql.com/> | Nice light weighted frontend |
| 1. R3.3 |  | YES | For most packages | https://www.freestatistics.org/cran/ | RStudio can R different R version |
| 1. R3.4 |  | YES | For some packages | https://www.freestatistics.org/cran/ |  |
| 1. RStudio | IDE for R | YES |  | https://www.rstudio.com/products/rstudio/download/ |  |
| 1. Anaconda (Python 3) | Data science platform | YES |  | <https://www.anaconda.com/download/>  <https://stackoverflow.com/questions/34097988/how-do-i-install-keras-and-theano-in-anaconda-python-on-windows> | Jupyter notebooks |
| 1. Python IDE | IDE for Python | Depends on your project |  | Check out : https://www.datacamp.com/community/tutorials/data-science-python-ide |  |
| 1. NO SQL Databases |  |  |  |  |  |
| 1. MongoDB | Database | YES |  | Check out :  <https://www.mongodb.com/download-center#community> |  |
| 1. Studio3T | MongoDB shell | YES |  | Check out:  <https://studio3t.com/> |  |