MASTER'S DEGREE EXAMINATION
Study major: Advanced Analytics – Big Data

1. The approach to data aggregation.
2. How to join multiple tables – describe possible methods.
3. What are the differences between single row and multiple row functions? When should they be used? What are the data types appropriate to be used by them?
4. Describe the single row functions classification.
5. Describe statements that can change the content of the table. What are the possible results of their execution? What is the possible scope?
6. The role of the Data Dictionary. Describe the methods of work with Data dictionary.
7. The database objects - their roles, purposes, methods of using.
8. The views. Why are they created? What are the possible clauses in a statement that create a view?
9. The syntaxes of set statements. What are the set operators and the results of their use?
10. The subqueries. Describe types of subqueries, possible clauses they may be used, possible operators.
11. Describe typical solutions Big Data provides in the area of data storage.
12. Describe the meaning of 3V and 5V in the context of Big Data.
13. Discuss ethical issues related to Big Data.
14. Evaluate capabilities and specific characteristics of analytical environments used in Big Data.
15. Please describe in detail one chosen algorithm used in Big Data analytics.
16. What is MapReduce and how does it work?
17. What is Deep Learning, give an example.
18. What are the typical characteristics of Big Data problems?
19. Discuss examples of pattern recognition techniques used in Big Data.
20. Define and describe distributed computing, in particular, in context of Big Data.
21. Describe a selected methodology describing a method of execution of development process of analytical models.
22. Outline key assumptions that are conditions of application of predictive models in support of decision making processes.
23. How a quality of a predictive model is measured?
24. Describe how usage of version control systems influences the effectiveness of analytical solution development process.
25. Explain what is meant by the term reproducibility of analytical process and why it is important in business.
26. Describe most important methods of ensuring reproducibility of analytical process.
27. Explain what does the term cutoff threshold mean in classification models and describe what are factors that influence its optimal value in case when such a model is used for supporting decision making.
28. Explain how regularization in used in the process of building of predictive models.
29. Explain the difference between observational, interventional and counterfactual reasoning.
30. Explain Simpson's paradox.
31. Economic gains from processing data in the cloud.

32. Serverless computing in gathering and processing data for analytics.
33. Storing big data in the cloud.
34. Scaling document-oriented databases in the cloud - the case of DynamoDB.
35. Scaling analytical processes in the cloud.
36. Function as a service - data processing model based on the Lambda architecture.
37. Creating and managing security of analytical platforms in the cloud for Python and R.
38. Managing security, users and access rights in the cloud - users, roles, policies and groups.
39. Managing a relational database in the cloud and applications for data analytics.
40. Data processing models for the cloud: IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service) and SaaS (Software-as-a-Service).
41. List and describe methodologies of data mining process.
42. Describe two main groups of data mining methods.
43. Describe the methods of feature selection and sampling for data mining modeling.
44. Data classification methods - present differences and similarities between them.
45. Describe decision tree models.
46. Describe random forest models.
47. Describe models of artificial neural networks.
48. Assessment of the predictive power of classification models.
49. Describe methods of data clustering.
50. Describe methods of transactional data analysis.
51. Discuss the data properties relevant to the data analysis process.
52. What is the importance of the context in data analysis.
53. What is data variability and how to take it into account in data visualization.
54. What is the uncertainty in data analysis and how can it be influenced.
55. What is the importance of metadata in data analysis.
56. Specify and discuss the coordinate systems used for data visualisation.
57. Specify and discuss methods for visualizing time series.
58. Specify and discuss methods for visualizing proportions.
59. Specify and discuss methods of relationship visualization.
60. List and discuss methods of visualization of spatial data.
61. Do logistic regression models belong to the class of generalized linear models? Justify the answer.
62. Estimation of logistic regression models.
63. Interpretation of the fitted logistic regression model.
64. Statistical inference in logistic regression models.
65. Assessment of goodness of fit of the logistic regression model.
66. Predictive power assessment of the logistic regression model.
67. Diagnostics of the logistic regression model.
68. Describe the proportional odds model.
69. Describe the multinomial logistic regression.
70. Compare the proportional odds model with the multinomial logistic regression model.
71. Economic and business benefits from analytics using Event History Analysis (EHA) models (survival analysis models).
72. Essence: philosophy - statistics - mathematics (including basic concepts) of the model of a single episode.

73. EHA models with discrete time versus EHA models with continuous time (including rules for the construction of databases for both types of models).
74. Traditional regression models versus EHA regression models - similarities and differences in theory, diagnostics and areas of application.
75. Basic procedures for estimating EHA models in SAS and available OPEN SOURCE software.
76. Theoretical foundations - applications - diagnostics and interpretation of the results of nonparametric models.
77. Theoretical foundations - applications - diagnostics and interpretation of the results of parametric models.
78. Theoretical foundations - applications - diagnostics and interpretation of the results of semiparametric models.
79. Semiparametric models of competitive risks (comparison of models: Cox and Fine-Gray).
80. Advanced EHA models & CLTV models. Prediction based on EHA models.
81. Data quality in business analytics. The meaning and assessment techniques.
82. Data imputation. The importance and meaning.
83. Define and describe the process of predictive modelling.
84. Name and describe selected measure of predictive power of a statistical model.
85. Explain what a distributed version control system is using Git as an example. Propose a typical simple workflow.
86. Discuss a selected data dimension reduction technique, its strong and weak points.
87. Discuss artificial neural network models using selected neural network topology.
88. Discuss the parallel computation concept and typical problems of parallel computations.
89. What is a robust estimator? Discuss using a selected example.
90. Discuss regularization techniques using a selected example, e.g., LASSO regression.
91. Methods of joining tables in SAS and SQL.
92. Advantages and disadvantages of data processing in SAS and SQL.
93. What is the macroprogramming in SAS?
94. What is the library in SAS System?
95. Examples of procedures in Base SAS and SAS/STAT units.
96. What descriptive statistics are robust on outliers?
97. What descriptive statistics are more adequate for not normal distributions?
98. How can be tested normality of distribution?
99. Advantages and disadvantages of analytical and transactional data structures.
100. What is PDV and sequential data processing in SAS?

Literature:

1. J. Price, Oracle Database 12c i SQL. Programowanie, Helion 2015;
2. J. Ullman, J. Widom, Podstawowy kurs baz danych Wyd. III, Helion 2011;
3. A. Alapati, D. Kuhn, B. Padfield, Oracle 12c. Problemy i rozwiązania, Helion 2014;
4. https://docs.oracle.com/database/121/SQLRF/toc.htm
5. Mayer-Schönberger V., Cukier K.: Big data :rewolucja, która zmieni nasze myślenie, pracę i życie : efektywna analiza danych; Warszawa : MT Biznes, 2017;

6. Surma J., Cyfryzacja życia w erze Big Data :człowiek, biznes, państwo /Warszawa : Wydawnictwo Naukowe PWN. 2017;
7. Inc, O.M., 2012. Big Data Now: 2012 Edition 2. wyd., O'Reilly Media;
8. Hand D., Mannila H., Smyth P,, Eksploracja danych / , WNT Wydawnictwa Naukowo-Techniczne, 2005;
9. White T., Hadoop :kompletny przewodnik : analiza i przechowywanie danych /; Gliwice : Helion, cop. 2016;
10. J. Gareth, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2013;
11. B. Kamiński: The Julia Express, http://bogumilkaminski.pl/files/julia_express.pdf;
12. B. Kamiński: Julia DataFrames Tutorial, https://github.com/bkamins/Julia-DataFrames-Tutorial;
13. M. Wittig, A. Wittig. Amazon web services in action, 2nd edition. Manning, 2018;
14. J. Baron, H. Baz, T. Bixler, B. Gaut, K. E. Kelly, S. Senior, J. Stamper. AWS certified solutions architect official study guide: associate exam. John Wiley & Sons, 2016;
15. Amazon (2016) Getting Started with AWS, wersja elektroniczna do pobrania za darmo w sklepie amazon.com;
16. Amazon (2009) The Economics of the AWS Cloud vs. Owned IT Infrastructure, do pobrania ze strony https://aws.amazon.com/whitepapers/
17. Amazon (2016) Amazon Elastic Compute Cloud (EC2) User Guide for Linux Instances, wersja elektroniczna do pobrania za darmo w sklepie amazon.com;
18. Introduction to AWS Economics, do pobrania ze strony https://aws.amazon.com/whitepapers/
19. Big Data Analytics Options on AWS, do pobrania ze strony https://aws.amazon.com/whitepapers/
20. Introduction to High Performance Computing on AWS, do pobrania ze strony https://aws.amazon.com/whitepapers/
21. Introduction to AWS Security, do pobrania ze strony https://aws.amazon.com/whitepapers/
22. Kamiński, B., & Szufel, P. (2015). On optimization of simulation execution on Amazon EC2 spot market. Simulation Modelling Practice and Theory, 58, 172-187;
23. D.T. Larose, Data Mining Methods and Models, Wiley, New York 2006;
24. J. Koronacki, J. Ćwik, Statystyczne systemy uczące się, WN-T, Warszawa 2005;
25. M. Lasek, M. Pęczkowski, Enterprise Miner: wykorzystywanie narzędzi Data Mining w systemie SAS, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2013;
26. R. Matignon, Data Mining Using SAS Enterprise Miner, Wiley, Hoboken, NJ 2007;
27. F. Provost, T. Fawcett, Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly, USA 2013;
28. T. Morzy, Eksploracja danych, Metody i algorytmy, PWN, Warszawa 2013;
29. N. Yau, Data points: visualization that means something, Indianapolis, Ind. Wiley, 2013;
30. N.C. Yau, Visualize this the FlowingData guide to design, visualization, and statistics, Indianapolis, Ind. Wiley 2011;
31. J. Maindonald, Data analysis and graphics using R': an example-based approach, Cambridge UK, New York: Cambridge University Press, 2003;
32. Frątczak E. (red.) Zaawansowane Metody Analiz Statystycznych, SGH, Warszawa 2012;
33. Allison P. D., Logistic Regression Using SAS: Theory and Application, Second Edition. Cary, NC: SAS Institute Inc., 2012;

34. Hosmer D. W., Jr., Lemeshow S., Sturdivant R. X., Applied Logistic Regression, Third Edition, John Wiley & Sons, 2013;
35. Kleinbaum D. G., Klein M., Logistic Regression: A Self-Learning Text, Third Edition, Springer, 2010;
36. Stanisz A., Modele regresji logistycznej. Zastosowania w medycynie, naukach przyrodniczych i społecznych. StatSoft Polska, Kraków, 2016;
37. Frątczak E., ,U.Sienkiewicz, H.Babiker, Analiza historii zdarzeń. Teoria, przykłady zastosowań z wykorzystaniem programów: SAS, TDA, STATA. SGH, Warszawa wyd. 2017;
38. Borucka J., Analiza i modelowanie ryzyka zachorowalności. Parametryczne i semiparametryczne modele przeżycia, SGH, 2017;
39. Allison P. , Survival Analysis Using SAS: A Practical Guide, Second Edition, 2010;
40. Broström G. Event History Analysis with R, Series: Chapman & Hall/CRC The R Series , CRC Press, 2012;
41. Crowder M., Multivariate Survival Analysis and Competing Risks. Chapman & Hall/CRC Texts in Statistical Science, 2012;
42. Elasthoff R.M., Gang Li, Ning Li. Joint Modelling of Longitudinal and Time to Event Data. Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 2016;
43. Xian Liu, Survival Analysis . Models and Applications. Wiley, 2013;
44. Korczyński A., Screening wariancji jako narzędzie wykrywania zmowy cenowej. Istota i znaczenie imputacji danych, Oficyna wydawnicza SGH, Warszawa, 2018;
45. Frątczak E. red. Zaawansowane Metody Analiz Statystycznych, SGH, Warszawa 2012;
46. Little A, Rubin D., Statistical Analysis with Missing Data. John Wiley & Sons: Hoboken 2002;
47. Malthouse E.C., Segmentation and Lifetime Value Models Using SAS, SAS Institute, 2013;
48. Svolba G., Applying Data Science. Business Case Studies, SAS Institute: Cary, NC, 2017;
49. W. Grzenda, A. Ptak-Chmielewska, K. Przanowski, U. Zwierz. Przetwarzanie danych w SAS, Oficyna Wydawnicza SGH, 2012;
50. SAS programming by example, Ron Cody and Ray Pass, SAS Publishing;
51. Zdzisław Dec, Wprowadzenie do systemu SAS, Wydawnictwo Editio, 2000;
52. Urszula Zwierz, Wstęp do systemu SAS wersja 8.1, Oficyna Wydawnicza SGH, 2002 ;
53. Jóźwiak J., Podgórski J.: Statystyka od podstaw, PWE, Warszawa.