**Федеральное государственное автономное**

**образовательное учреждение высшего образования   
"Национальный исследовательский университет   
"Высшая школа экономики"**

Факультет компьютерных наук

Департамент программной инженерии

**Рабочая программа дисциплины**

Распределенные базы и хранилища данных(на английском языке)

Distributed Databases and Data Warehouses

для образовательной программы «Программная инженерия»

направления подготовки 09.03.04 «Программная инженерия»

уровень - бакалавр

Разработчик программы

Брейман А.Д.,к.т.н., доцент, [abreyman@hse.ru](mailto:vpodbelskiy@hse.ru)

Одобрена на заседании департамента программной инженерии «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 2017 г.

Руководитель департамента Авдошин С.М. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Утверждена Академическим советом образовательной программы

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 2017 г., № протокола\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Академический руководитель образовательной программы

Шилов В.В. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Москва, 2017

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения-разработчика программы.*

1. **Course Description**
   1. Title of a Course: Distributed Databases and Data Warehouses
   2. Pre-requisites: Databases, Programming, Discrete Mathematics, Introduction into Software Engineering, Algorithms and Data Structures
   3. Course Type: Elective
   4. Abstract

Course presents a detailed introduction into distributed data processing, relational data warehouses, multidimensional OLAP tools and massive parallel data processing systems (Hadoop, Cassandra, MongoDB). Students will develop understanding in the design methodology for distributed databases and data warehouses. Practice studies include implementing databases and applications software in map/reduce paradigm and in several NoSQL data models.

The syllabus is prepared for teachers responsible for the course (or closely related disciplines), teaching assistants, students enrolled on the course as well as experts and statutory bodies carrying out as-signed or regular accreditations in accordance with

* educational standard of the federal state autonomous educational institution for higher education National Research University Higher School of Economics;
* B.Sc. curriculum ("Software Engineering", area code 09.03.04) 3rd year, 2017-2018 academic year;

The course is offered to students of the Bachelor Program «Software Engineering» (code 09.03.04) at Faculty of Computer Science of the National Research University Higher School of Economics (HSE).

This elective course belongs to professional cycle curricula unit (Б.ПЦ.Б.2 unit) of 2017-2018 academic year's working syllabus (3rd year of studies).

It is a two module course, which is delivered in modules #3 and #4 of the third academic year. Number of credits is 5. Total course length is 190 academic hours including 64 auditory hours (32 Lecture hours and 32 Practice hours) and 126 Self-study hours.

Academic control forms are two home assignments in form of a group project, one written test after module #3, and one written exam after module #4.

Good knowledge of English language (at least listening and reading, intermediate speaking) is mandatory since the course is completely given in English.

1. **Learning Objectives**

The objective of the course is to form professional competencies related to design and implementation of several kinds of distributed databases, including data warehouses, online analytical data processing and big data management tools. Students will get a grasp on strengths and weaknesses of wide spectrum of approaches to data storage, search and retrieval, resulting in informed choice of database model. This course studies different conceptual database models and their properties. The models that will be discussed are:

* Relational data warehouse;
* Multidimensional data warehouse;
* Online analytical processing;
* Map/reduce massive parallel data processing;
* Key/value, document, graph and wide columnar database models;
* Data stream processing.

For these conceptual models the course will concentrate on the following points: Why was the database model introduced? Which of the shortcomings of other models does it address? What are the most important concepts and notions for the database model? How is the model implemented? Which are the main techniques? The importance of understanding the internals of a particular database model cannot be overemphasized as it is closely connected to its limitations.

1. **Learning Outcomes**

After taking this course the student should have achieved the following objectives:

Knows the reference architectures of data warehouses and is aware of the basic functionality offered by available commercial and free data warehousing systems. Masters methods and tools for creating analytical database solutions. Can choose dimensions for multidimensional database, group them into hierarchies and define aggregates. Knows principles of data integration and extract-transform-load process design and implementation.

Knows the map/reduce approach and its Hadoop implementation. Can create map/reduce-based solutions to large-scale massive data processing problems.

Knows database models and tools of NoSQL class, including key-value stores, document databases , wide columnar databases and graph databases. Can implement systems and database applications using these tools.

Knows data stream management systems, its data models and concepts. Can compose queries for executing in these systems.

Students should be able to understand the language of studies models, choose and use appropriate models and programming languages, implement systems using chosen models, methods and tools.

The course contributes to the development of the following competencies:

| Competency | Code | Descriptors | Education forms and methods for competency formation |
| --- | --- | --- | --- |
| Universal | УК-1, СК-Б1 | Able to learn, acquire new knowledge, skills, including in the area other than the professional | To successfully complete homework assignment, students have to find information about the data sources and other domain knowledge for wide variety of possible areas – economics, transportation, manufacturing, medicine, etc. |
| Universal | УК-3, СК-Б4 | Able to solve problems in professional activities on the basis of the analysis and synthesis | This course is divided onto sections covering wide variety of methods and tools for data management including storage, retrieval, exchange, and others, that can be used for synthesizing a software solution to data processing tasks. This could be achieved by analyzing and dividing the given complex task (e.g. homework assignment) onto a sequence of simpler tasks. |
| Universal | УК-5, СК-Б6 | Able to work with the information: locate, evaluate and use information from different sources necessary for solving scientific and professional problems (including on the basis of the system approach) | To successfully complete homework assignment, students have to find information about the data they will use for their project, understand and properly model the data. They also need to find the appropriate ways (usually seek for on-line documentation) to manage the data (store, read, exchange, etc.) |
| Universal | УК-7, СК-Б8 | Able to work in a team | Teamwork is mandatory to complete homework assignment. |
| Professional, Scientific research activities | ПК-1, УК-1 | The ability to apply main concepts, principles, theories and facts, connected to computer science, in the process of scientific research and problem solving. | Lectures, Essay composing,  Home assignments, Self-study |
| Professional, Scientific research activities | ПК-2, УК-2 | The ability to formalize in its subject area within the constraints of methods of research | Lectures, Practical studies in data modeling,  Home assignments (data modeling part) |
| Professional, Scientific research activities | ПК-3, УК-3 | Preparedness to use research techniques and tools to study objects of professional activity | Lectures, Practical studies in data model assessment and evaluation,  Home assignments (solutions evalution part) |
| Professional, Analytical activities | ПК-6, УК-6 | The ability to formalize the subject area of a software project and to develop specifications the software product components | Lectures, Practical studies,  Home assignments |
| Professional,  Project activities | ПК-10, УК-10 | The ability to design, develop and test software products. | Lectures, Practical studies,  Home assignments |
| Professional,  Project activities | ПК-11, УК-11 | The ability to read, understand, and extract the main idea from source code and documentation | Practical studies,  Industrial cases reviews |
| Professional,  Project activities | ПК-12, УК-12 | Skills in modeling, analysis and use of formal methods of software | Lectures, Practical studies,  Home assignments |
| Professional, Technology-oriented activities | ПК-15, УК-15 | Skills in using the operating systems, network technologies, software interface development tools, languages and methods of formal specifications, database management systems | Lectures,  Practical studies,  Home assignments |
| Professional, Technology-oriented activities | ПК-16, УК-16 | Skills in using different software development technologies | Lectures, Practical studies,  Home assignments |
| Professional, Technology-oriented activities | ПК-17, УК-17 | The ability to apply core software development methods and tools | Lectures, Practical studies,  Home assignments |
| Professional, Technology-oriented activities | ПК-19, УК-19 | The ability to understand life cycle standards and models. | Lectures, Practical studies,  Home assignments |

1. **Course Plan**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Topic title | Total hours | Classroom hours | | | Self-study |
| Lectures | Seminars | Practice |
| Module #3, 3rd year | | | | | | |
| **14** | Data warehousing and Big Data Management | 11 | 2 |  | 2 | 7 |
| **15** | Data Warehousing Architectures and Models | 36 | 6 |  | 6 | 24 |
| **16** | Data Cleaning And Integration | 24 | 4 |  | 4 | 16 |
| **17** | Key/Value and Document Databases | 24 | 4 |  | 4 | 16 |
|  | **Module totals** | **95** | **16** |  | **16** | **63** |
| Module #4, 3rd year | | | | | | |
| **18** | Map/Reduce and Hadoop | 48 | 8 |  | 8 | 32 |
| **19** | Large-scale Distributed Databases | 24 | 4 |  | 4 | 16 |
| **20** | In-memory Databases | 12 | 2 |  | 2 | 8 |
| **21** | Data Streams Management | 11 | 2 |  | 2 | 7 |
|  | **Module totals** | **95** | **16** |  | **16** | **63** |
|  | **TOTAL** | **190** | **32** |  | **32** | **126** |

1. **Reading List**
   1. **Required**

* Vaisman A., Zimányi E. Data Warehouse Systems. Design and Implementation.

[Электронный ресурс] /Alejandro Vaisman, Esteban Zimányi. – Электрон. текстовые данные. – Springer, 2014. – 625 p. – 978-3-642-54655-6. — Режим доступа: https://proxylibrary.hse.ru:2184/book/10.1007%2F978-3-642-54655-6

* Vohra D. Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools. [Электронный ресурс] /Deepak Vohra. – Электрон. текстовые данные. – Apress, 2016. – 444 p. – 978-1-484-22198-3. — Режим доступа: https://proxylibrary.hse.ru:2251/toc.aspx?bookid=119892
* Hows D., Membrey P., Plugge E., Hawkins T. The Definitive Guide to MongoDB: A Complete Guide to Dealing with Big Data using MongoDB, Third Edition. [Электронный ресурс] / David Hows, Peter Membrey, Eelco Plugge and Tim Hawkins. – Электрон. текстовые данные. – Apress, 2015. – 376 p. – 978-1-484-21183-0. — Режим доступа: https://proxylibrary.hse.ru:2251/ toc.aspx?bookid=106761
  1. **Optional**
  + Golfarelli M., Rizzi S. Data Warehouse Design: Modern Principles and Methodologies, McGraw-Hill Osborne Media, 2009. — 480pp.
  + Celko J. Joe Celko's Analytics and OLAP in SQL, Morgan Kaufmann, 2006. — 208pp.
  + Redmond E. Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, Pragmatic Bookshelf, 2012. — 352pp.
  + Golab L., Oszu M. T. Data Stream Management (Synthesis Lectures on Data Management), Morgan & Claypool Publishers, 2010. — 80 pp.
  + Kimball R., Ross M. The Data Warehouse Toolkit, Wiley, 2002. — 447pp.
  + Kimball R., Caserta J. The Data Warehouse ETL Toolkit, Wiley, 2004. — 491pp.
  + Smith B.C., Clay C.R. Microsoft SQL Server 2008 MDX Step by Step, Microsoft Press, 2009. — 400pp.
  + Rodrigues F., Coles M., Dye D. Pro SQL Server 2012 Integration Services, Apress, 2012. — 636 pp.
  + White T. Hadoop: The Definitive Guide, 4th edition, O’Reilly, 2015. — 768pp.
  + Franks B. Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics (Wiley and SAS Business Series), Wiley, 2012. — 336pp.
* Rainardi V. Building a Data Warehouse: With Examples in SQL Server. Apress, 2008. — 541pp.
* Inmon W. H., Krishnan K. Building the Unstructured Data Warehouse: Architecture, Analysis, and Design, Technics Publications, 2011. — 216pp.
* Holmes A. Hadoop In Practice, Manning Publications, 2012. — 511pp.
* Miner D., Shook A. MapReduce Design Patterns, O’Reilly, 2012. — 232pp.
* Tiwari S. Professional NoSQL, Wrox, 2011. — 384pp.
* Bradberry, Russell, Lubow. Eric (2014) Practical Cassandra. A Developer’s Approach. – Addison Wesley, 2014. – 193 pp
* Plattner H. (2013). A Course in In-Memory Data Management, Springer, 2013. – 298 pp.

1. **Grading System**

Rounding procedure for grades (where applicable): up to an integer number of points.

Practice activity during practice hours is assessed by evaluating of student involvement into discussions as well as quality of exercise performance during practice. Practice activity grade Оclassroom uses a ten-point scale.

Students have to write an essay (on topic proposed by instructor at first lecture) once in each module (due 4th week of each module) with length of up to 2 pages. Grade *Оessays*  is an arithmetic averages of two essays grades (ten-point scale, rounding up to an integer number of points).

Students have to answer quiz questions in first 10 minutes of some lectures. Grade *Оquiz*  is an arithmetic averages of quiz grades respectively (ten-point scale, rounding up to an integer number of points). Average is calculated by division of sum of all student’s quiz answers grades on a total quizzes count in a year.

Students have to answer test questions on 7th week of 3rd module of 3rd year. Grade *Оtest* for that test is on ten-point scale.

Value of *Оhomework 1* (homework assignment 1) component of final grade formula is an integer value from interval [0,10] consists of the common score for the report and presentation (from 0 to 5; same score to all group members) and individual student score for the answers to the questions (from 0 to 5). If a student misses the project presentation because of some valid reason, s/he receives «absence» grade. If a student misses the project presentation because of any other reason, s/he receives grade based on individual score set to 0.

Value of *Оhomework 2* (homework assignment 2) component of final grade formula is an integer value from interval [0,10] consists of the common score for the report and presentation (from 0 to 5; same score to all group members) and individual student score for the answers to the questions (from 0 to 5). If a student misses the project presentation because of some valid reason, s/he receives «absence» grade. If a student misses the project presentation because of any other reason, s/he receives grade based on individual score set to 0.

Final written exam at the end of the course *Оfinal exam* is assessed on usual ten-point scale.

Cumulative grade for student’s current results is calculated using the following formula:

О*cumulative* = 0,7\* *Оcurrent* + 0,3\* Оclassroom

where

*Оcurrent* = *0,2·Оessays +0,2·Оtest+0,2·Оquiz +0,2·Оhomework 1*+*0,2·Оhomework 2*

Final grade for student is calculated using the following formula:

О*final* = 0,5 *Оcumulative*  + 0,5 Оfinal exam*.*

1. **Guidelines for Knowledge Assessment**

Home assignments 1 and 2 (resp. HA1 and HA2) has to be prepared by students in groups of up to 5. Students should identify a Big Data problem that they would like to work on and specify a difficult question to data and identify data sources that might help answering the question. Chosen Big Data problem should be approved by instructor.

HA1 has to be prepared in module 3 of 3rd year and includes design, implementation and testing of a data warehouse and application for online analytics in given subject area (chosen by group and approved by instructor or assigned by instructor). Results of HA1 should be presented in form of report that consists of design document, implementation description, results of testing. Mandatory appendixes are source code for application and database creation script. Report should be submitted to LMS not later that for 7 calendar days before assigned date of its presentation (on the last week of 3th module). HA1 should be presented and demonstrated by all group members. Each group member should demonstrate complete understanding of all project details and give correct answers to at least two questions of instructor.

Home assignment 2 (HA2) has to be prepared in module 4 of 3rd year by. Its result should contain Hadoop, MongoDB or Cassandra-based solution for problem identified in HA1. Solutions (DWH/OLAP-based from HA1 and NoSQL-based from HA2) should be compared, their relative strengths and weaknesses are revealed and described. Results of HA2 should be presented in form of project report that consists of design document, implementation description, results of testing and comparing of solutions. Mandatory appendixes are source code for both applications and data management scripts. Report should be submitted to LMS not later that for 7 calendar days before assigned date of its presentation (on the last week of module 2 of 4th year). Project should be presented and demonstrated by all group members. Each group member should demonstrate complete understanding of all project details and give correct answers to at least two questions of instructor.

Written test at the end of the first module (last week of module 3 of 3rd year of study) implies arrangement of the written test (in lecture room) for all students enrolled to the course. Topics covered by the test embrace first module material.

Written exam at the end of the second module (module 4 of 3rd year of study) implies arrangement of the written test (in lecture room) for all students enrolled to the course. Topics covered by the test embraces all course material.

1. **Methods of Instruction**

Course studies are organized in the form of lectures and practical studies. Besides traditional forms, some active and interactive forms are provided: discussion of real industry case studies; proposing and discussing group projects topics and its planned outcomes, using interactive simulators for database languages.

1. **Special Equipment and Software Support**

Projector for lectures and practical studies.

Software access: internal network, in accordance with license and contract.

* Microsoft Windows 7 Professional RUS
* Microsoft Windows 8.1 Professional RUS
* Microsoft Windows 10
* Apple Mac OS
* Microsoft Visual Studio 2015 Community (or later versions)
* SQL Server Management Studio