STUDIJŲ DALYKO (MODULIO) APRAŠAS

Dalyko (modulio) pavadinimas	Kodas
Duomenų saugyklų architektūra	

Dėstytojas (-ai)	Padalinys (-iai)
Koordinuojantis: Jurga Globienė **Kitas (-i):** dr. Gintautas Tamulevičius	Matematikos ir informatikos fakultetas Duomenų mokslo ir skaitmeninių technologijų institutas

Studijų pakopa	Dalyko (modulio) tipas
Pirmoji	Pasirenkamas

Įgyvendinimo forma	Vykdymo laikotarpis	Vykdymo kalba (-os)
Auditorinė, kontaktinė nuotolinė	7 semestras	Lietuvių / Anglų

Reikalavimai studijuojančiajam	
Išankstiniai reikalavimai: Duomenų bazių valdymo sistemos, duomenų struktūros ir algoritmai	**Gretutiniai reikalavimai (jei yra):** Duomenų bazių užklausų kalbos

Dalyko (modulio) apimtis kreditais	Visas studento darbo krūvis	Kontaktinio darbo valandos	Savarankiško darbo valandos
5	103	33	70

Dalyko (modulio) tikslas: studijų programos ugdomos kompetencijos		
Dalyko tikslas – įgyti žinias apie duomenų saugyklų architektūrą, veikimo principus ir duomenų struktūrų projektavimo taisykles. Kurso metu bus pristatyta dimensinio modeliavimo metodika, paaiškinanti, kaip turi būti struktūrizuoti duomenys duomenų saugyklose. Taip pat supažindinama su duomenų ETL procesais, jų automatizavimu. Modulio klausytojai gebės pritaikyti įgytas žinias ruošdami projektą realaus pasaulio užduočiai įgyvendinti.		
Dalyko (modulio) studijų siekiniai	**Studijų metodai**	**Vertinimo metodai**
Gebės paaiškinti esmines duomenų saugyklų sąvokas ir terminus, procesus, suvoks duomenų apdorojimo automatizavimo principus.	Literatūros analizė, probleminis dėstymas, praktinės užduotys, diskusijos.	Egzaminas
Gebės nustatyti reikalingus ETL procesų etapus, tokius kaip duomenų nukrovimas, valymas ir tikrinimas bei transformacija.		
Gebės bendrauti duomenų apdorojimo ir saugyklų tema valstybine ir užsienio kalbomis.	Grupinė projektinė užduotis, probleminis dėstymas.	Egzaminas, mentorystė bei projektinės užduoties metu paruošta ataskaita.
Gebės suprojektuoti duomenų saugyklą taikydami dimensinio modeliavimo metodiką, parengti reikiamą dokumentaciją.		
Gebės parengti duomenų saugyklos projektavimo dokumentaciją.		

Temos	Kontaktinio darbo valandos							Savarankiškų studijų laikas ir užduotys	
	Paskaitos	Konsultacijos	Seminarai	Pratybos	Laboratoriniai darbai	Praktika	**Visas kontaktinis darbas**	**Savarankiškas darbas**	**Užduotys**
1. Įvadinė paskaita: modulio turinys, modulio tikslas, studijų forma, taisyklės, žinių vertinimas. Supažindinimas su projektiniu darbu.	1						1		
2. Susipažinimas su duomenų analitinėmis sistemomis ir jų panaudojimu. Panašumai ir skirtumai nuo transakcinių sistemų. Duomenų saugyklos apibrėžimas.	1						1		
3. Duomenų saugyklų (DS) architektūra. Pagrindiniai DS tikslai ir veikimo principai. Pagrindines duomenų saugyklų dalys, jų atsakomybė bei procesai jose ir tarp jų. Restorano metafora architektūrai paaiškinti.	2						2		
4. Dimensinis modeliavimas. Kur ir kam jis naudojamas. Dimensijos, faktų lentelės ir žvaigždinės schemos apibrėžimai.	2				1		3	2	Projektinės užduoties srities pasirinkimas ir suderinimas
5. Pagrindiniai faktų lentelių modeliavimo metodai. Faktų lentelių detalumas, ryšys su dimensijomis, rodikliai bei indeksų svarba ir jų panaudojimas.	2				1		3	2	Projektinei užduočiai reikiamų faktų lentelių identifikavimas bei jų detalumo nustatymas.
6. Faktų lentelių tipai bei jų savybės ir duomenų atnaujinimo principai. Transakciniai, periodiniai snapshot, kaupiamieji, agreguoti bei akumuliuoti faktai.	2				2		3	8	Faktų lentelių tipo identifikavimas, loginis projektavimas bei jų aprašymas.

7. Pagrindiniai dimensijų projektavimo metodai: surogatiniai ir natūralūs raktai, duomenų denormalizacija, hierarchijos, nulinė dimensijos eilutė. Patvirtintų (conformed) dimensijų sąvoka.	2				1		3	4	Projektinei užduočiai reikiamų dimensijų identifikavimas bei jų detalumo nustatymas.
8. Lėtai kintančių dimensijų istorizavimas. Istorizavimo tipai. Dimensijų istorizavimo automatizavimas.	2				2		3	8	Dimensijų loginis projektavimas bei jų aprašymas.
9. Išplėstiniai faktų lentelių modeliavimo metodai: surogatiniai raktai, periodai kaip faktai, faktai įvairiais matavimo vienetais, išskirstyti, pelningumo faktai.	1				2		3	4	Detalus projektinės užduoties faktų lentelių projektavimas, indeksų, reikiamų laukų identifikavimas bei aprašymas
10. Išplėstiniai dimensijų modeliavimo metodai: outriger, daugiareikšmės dimensijos, tiltinės lentelės, agreguoti faktai kaip dimensijos dalis, dinaminiai rėžiai, skirtingos laiko zonos, skirtingo lygio hierarchijos.	1				2		3	4	Detalus projektinės užduoties dimensijų projektavimas, reikalingų laukų identifikavimas bei aprašymas
11. Duomenų nukrovimo (sourcing) būdai bei jų problematika. Duomenų savininkas. Duomenų nukrovimas failais, tiesiai iš DB, eilėmis (queues, data streaming).	2				0		2	0	
12. Duomenų patikra ir valymas. Duomenų tikrinimo reikalingumas. Pirminiai ir automatiniai duomenų tikrinimai. Duomenų tikrinimo tipai: laukų, lentelių struktūrų ir verslo logikos tikrinimai. Neteisingų duomenų apdorojimas bei taisymas.	2						2	0	
13. Duomenų „ežero“ (Data Lake) plėtiniai: nestruktūrizuoti duomenų tipai, pirminių (neapdorotų) duomenų analizė ir panaudojimas dirbtinio intelekto sistemoms, duomenų gavimas ir apdorojimas realiu laiku didžiųjų duomenų uždaviniuose. Domenų saugyklų ir duomenų „ežerų“ pasirinkimas ir naudojimas.	2						2	0	
16. Galutinės projektinės užduoties ataskaitos parengimas ir suderinimas.							2	18	Galutinės projektinės užduoties ataskaitos parengimas ir suderinimas
17. Pasiruošimas egzaminui.							0	20	
Iš viso	**22**				**11**		**33**	**70**	

Vertinimo strategija	**Svoris, proc.**	**Atsiskaitymo laikas**	**Vertinimo kriterijai**
Projektinės užduoties ataskaita	60	Paskutinioji semestro savaitė	Studentai dirbs grupėmis. Kiekviena grupė pasirinks arba gaus realią modeliavimo ir projektavimo užduotį. Darbas bus atliekamas savarankiškai su dėstytojos konsultacijomis ir pagalba laboratorinių paskaitų metu. Darbas bus suskirstytas į kelis etapus. Kiekvienas etapas bus užskaitomas bei vertinamas atskirai, atsižvelgiant į tai, kaip studentas / komanda dirbo to etapo metu. Tik pateikus galutinę projektinės užduoties ataskaitą darbas bus galutinai užskaitytas bei studentas gaus galutinį projektinės užduoties ataskaitos įvertinimą.
Egzaminas	40	Sesija	Egzaminas laikomas raštu. Kiekvienas studentas gauna 3-4 nedidelės apimties klausimus, kurie apims tiek teorines, tiek praktines įgytas žinias. Egzamino darbas vertinamas kaip teisingai atsakytų klausimų balų suma. Egzaminą laiko tik projektinę užduotį atlikę studentai. Pastarosios vertinimas egzamino vertinimui jokios įtakos neturės.
Perlaikymas		Sesija perlaikymui	Neatsiskaičius projektinės užduoties arba nesurinkus pakankamai balų iš užduoties bei egzamino, yra rašoma papildoma projektinė užduotis. Lygiai taip, kaip ir pagrindinės projektinės užduoties metu, yra pasirenkama tema ir yra ruošiama pilnos apimties projektinė ataskaita. Ataskaita yra derinama ir atsiskaitoma etapais, kiekvieną kartą gaunant patvirtinimą iš dėstytojos, kad galima judėti toliau. Maksimalus galimas surinkti balų skaičius perlaikymo metu yra 7. Pridavus ataskaitą, egzamino rašyti nebereikia.
Atsiskaitymas eksternu		Negalimas	Atsiskaitymas eksternu nėra galimas, nes pagrindinė galutinio balo dalis yra projektinis darbas, kuris yra atliekamas viso kurso metu su dėstytojos pagalba.

Autorius	**Leidimo metai**	**Pavadinimas**	**Periodinio leidinio Nr. ar leidinio tomas**	**Leidimo vieta ir leidykla ar internetinė nuoroda**
Privaloma literatūra				
Ralph Kimball, Margy Ross	2013 (3-asis l.)	The Data Warehouse Toolkit	-	Wiley
William H. Inmon	2005 (4-asis l.)	Building the Data Warehouse	-	Wiley
Papildoma literatūra				
Ralph Kimball, Joe Caserta	2004	The Data Warehouse ETL Toolkit	-	Wiley
Ralph Kimball	2007 (2-asis l.)	The Data Warehouse Lifecycle Toolkit	-	Wiley

COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Data Warehouse Architecture	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: Jurga Globienė **Other(s):** Dr. Gintautas Tamulevičius	Faculty of Mathematics and Informatics Institute of Data Science and Digital Technologies

Study cycle	Type of the course unit (module)
First	Optional

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Remote, face-to-face	7th semester	Lithuanian / English

Requirements for students	
Prerequisites: Data base management systems, Data structures and algorithms	**Additional requirements (if any):** Data base query languages, Data base design

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
5	103	33	70

Purpose of the course unit (module): programme competences to be developed		
The course goal is to explain data warehouse (DW) architecture, working principles and internal processes. Dimensional modelling methodology will be introduced by describing various data modelling techniques. ETL processes and their automation issues will be explained. Students will be able to use learnt material for designing real-world project use case.		
Learning outcomes of the course unit (module)	**Teaching and learning methods**	**Assessment methods**
Ability to explain the fundamental data warehousing concepts, definitions and processes, understand data processing and automation principles.	Literature analysis, problem-oriented teaching, tutorials, discussion.	Examination
Ability to identify required ETL process's types such as data loading, cleaning and verification and transformation.		
Ability to communicate in native and foreign language data processing and data warehousing topics.	Project assignment coursework, problem-oriented teaching	Examination, mentoring and final project assignment report preparation.
Ability to design data warehouse by using dimensional modelling technique and prepare corresponding documentation.		
Ability to prepare data warehouse design documentation.		

Content: breakdown of the topics	Contact hours							Self-study work: time and assignments	
	Lectures	Tutorials	Seminars	Exercises	Laboratory work	Internship/work placement	Contact hours	Self-study hours	Assignments
1. Introductory lecture: course content, goals, study form, rules, assessment. Introduction to project assignment.	1						1		
2. Introduction to data analytical systems and their need. Similarities and differences with transactional systems. Data Warehouse definition.	1						1		
3. Data Warehouse architecture (DW). DW goals and operational principles. DW architecture components, their purpose and responsibility. ETL processes within each component and between them. Restaurant metaphor.	2						2		
4. Dimensional modelling. Definition, purpose and DW component where it can be used. Dimension, fact table and star schema definitions.	2				1		3	2	Select and align use case for project assignment
5. Basic fact table modelling techniques. Fact table's granularity, measurements and descriptive attributes, relationship with dimensions. The purpose and usage of indexes for dimensional modelling.	2				1		3	2	Identify fact tables needed for selected use case. Specify their granularity.
6. Fact table types, their characteristics and data loading or renewal principles. Transactional, snapshot, factless, accumulated, aggregated and consolidated fact tables.	2				2		3	8	Create and describe logical design for selected use case fact tables

7. Basic dimension design techniques: surrogate and natural keys, data denormalization, hierarchies, null row. Conformed dimension definition.	2				1		3	4	Identify dimensions needed for selected use case. Specify their granularity.
8. Slowly changing dimensions historization. Historization types. Historization automation.	2				2		3	8	Create and describe logical design for selected use case dimensions
9. Advanced table modelling techniques: surrogate keys, lag/duration, multiple currency, multiple units, allocated, profit and loss facts.	1				2		3	4	Describe detail design for selected use case fact tables
10. Advanced dimension modelling techniques: outrigger, multivalued dimensions, bridge tables, aggregated facts as dimension attributes, dynamic value bands, multiple time zones, varied depth hierarchies.	1				2		3	4	Describe detail design for selected use case dimensions
11. Data sourcing techniques. Data owner. Data loading via files, DB to DB, queues or stream.	2				0		2	0	
12. Data validation and cleaning. Data validation purpose and benefits. Initial and automated data validation. Data validation types: column, structure, business logic. Dealing with incorrect raw data.	2						2	0	
13. Data Lake extension: unstructured data types, raw data analysis and usage for AI models, sourcing and processing data in real time on big data. When to use DL and when DW.	2						2	0	
16. Project assignment report finalization and alignment.							2	18	Project assignment report finalization and alignment.
17. Preparation for examination							0	20	
Total	**22**				**11**		**33**	**70**	

Assessment strategy	Weight, %	Deadline	Assessment criteria
Project assignment report	60	The last week of semester	Students will work in groups. Every group will be able to select real-world use case, model it and write detail project design report for it. Project assignment will be performed by students with lecture consultation and mentoring during laboratory lectures. Project assignment will be divided into several phases. Every phase will be aligned and evaluated separately, based on how student or group worked during it. Final evaluation will be given only then final project assignment report will be delivered and accepted.
Examination	40	Exam session	The examination is performed in written mode. During examination every student will receive 3-4 open questions which aim is to check theory and practice knowledge. The assessment of exam work is obtained as the sum of all correctly answered question points. Project assignment report evaluation will not be considered while evaluating exam, but only students with provided report will be examined.
Re-examination		Exam session for re-examination	In case student failed to provide project assignment report or collected less than minimum grade for final evaluation, there is an option to take additional project assignment. Similarly as with primary project assignment, the use-cases should be selected, described, and modeled. Assignment report should be aligned in phases and lecturer's approval should be received before continuing with another phase. Maximal evaluation grade for re-examination is 7. If additional project report is provided, no examination will be needed.
External students		Not possible	To participate in this course as external student is not possible, because a major part of final evaluation consists of project assignment, which is performed during all the course with the lecturer's help.

Author	Year of publication	Title	Issue of a periodical or volume of a publication	Publishing place and house or web link
Compulsory reading				
Ralph Kimball Margy Ross	2013 (3rd ed.)	The Data Warehouse Toolkit	-	Wiley
William H. Inmon	2005 (4th ed.)	Building the Data Warehouse	-	Wiley
Optional reading				
Ralph Kimball Joe Caserta	2004	The Data Warehouse ETL Toolkit	-	Wiley
Ralph Kimball	2007 (2nd ed.)	The Data Warehouse Lifecycle Toolkit	-	Wiley