БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

Утвърдил:
(акад. В. Дренски, Директор на ИМИ-БАН)

Учебна програма
за специализиран докторантски курс

Област на висше образование:	4. Природни науки, математика и информатика
професионално направление:	4.6. Информатика и компютърни науки
докторска програма:	Информатика
тема:	Методи и технологии за извличане на данни от оперативни бази данни
лектор:	Доц. д-р Десислава Панева-Маринова
данни за връзка с лектора (тел., имейл)	+359888894814, dessi@cc.bas.bg
хорариум:	20 часа лекции и 20 часа практически упражнения
кредити съгл. кредитната система на ЦО на БАН:	20

1. Анотация

Учебният курс цели запознаване и прилагане на методи и технологии за извличане и обработка на данни в оперативни бази от данни. Покриват се основно теми, свързани с концептуалния дизайн на процесите по извличане, трансформиране и зареждане (ETL) на данни при различни модели на организация и управление на процеса по събиране, верифициране, анализ на данни, подготовка за тяхното моделиране и трансформиране в зависимост от схемата на описание.

2. Необходими предварителни знания

Релационни бази от данни и СУБД

3. Компетентности, придобити в резултат на обучението

Знания и умения за прилагане на методи и технологии за извличане и обработка на данни от оперативни бази данни.

4. Тематично съдържание

тема	брой часове лекции	брой часове практически упражнения
Извличане и анализ на големи обеми от данни (Big Data). Методи и алгоритми за извличане на знания от данни.	2	
Методология и техники за верифициране на уеб данни. Извличане на полезни и коректни данни за анализ.	2	2
Методи за извличане, трансформиране и зареждане (ETL) на големи обеми от данни (Big Data). Парадигмата MapReduce (MR).	2	2
Основни функционалности на метода за извличане, трансформиране и зареждане (ETL) на данни. Промяна на въведени данни (CDC - Change Data Capture). Валидиране качеството на данните (DQV). Сурогатен ключ (SK).	2	2
Процес на извличане, трансформиране и зареждане (ETL) приложен при складове за данни (Data Warehouse systems). Извличане, интегриране, почистване и зареждане на данните.	2	2
Извличане на данни с SQL – агрегиране и сортиране на данни, извличане на данни от множество таблици, търсене на данни с различни типове оператори	2	2
Сравняване на ETL и Extract Load Transform (ELT). Приложение на двата метода за извличане на данни, Последователност на извличането и зареждането на данните при двата метода.	2	2
Извличане на данни от NoSQL модели на данните. Специфични за домейна езици (DSL) за извличане на данни (например Asterix Query Language, JAQL и др.)	2	2
Инструменти за анализ на данни в Облака – Spark, Mahout, Hunk	2	2
Извличане на данни от различни източници с похватите на R Programming. Изчистване и отделяне на приложими за анализ данни	2	2
Намиране и извличане на структурирани данни от уеб сайтове (HTML страници). Иинструменти и методология		2

5. Конспект

1. Извличане и анализ на големи обеми от данни (Big Data). Методи и алгоритми за извличане на знания от данни.
2. Методология и техники за верифициране на уеб данни. Извличане на полезни и коректни данни за анализ.
3. Методи за извличане, трансформиране и зареждане (ETL) на големи обеми от данни (Big Data). Парадигмата MapReduce (MR).
4. Основни функционалности на метода за извличане, трансформиране и зареждане (ETL) на данни. Промяна на въведени данни (CDC - Change Data Capture). Валидиране качеството на данните (DQV). Сурогатен ключ (SK).
5. Процес на извличане, трансформиране и зареждане (ETL) приложен при складове за данни (Data Warehouse systems). Извличане, интегриране, почистване и зареждане на данните.
6. Извличане на данни с SQL – агрегиране и сортиране на данни, извличане на данни от множество таблици, търсене на данни с различни типове оператори
7. Сравняване на ETL и Extract Load Transform (ELT). Приложение на двата метода за извличане на данни, Последователност на извличането и зареждането на данните при двата метода.
8. Извличане на данни от NoSQL модели на данните. Специфични за домейна езици (DSL) за извличане на данни (например Asterix Query Language, JAQL и др.)
9. Инструменти за анализ на данни в Облака – Spark, Mahout, Hunk
10. Извличане на данни от различни източници с похватите на R Programming. Изчистване и отделяне на приложими за анализ данни
11. Намиране и извличане на структурирани данни от уеб сайтове (HTML страници). Иинструменти и методология

6. Препоръчана литература:

1. S. Sakr, M. Medhat Gaber (eds.), Large Scale and Big Data - Processing and Management (Auerbach Publications, Boston, 2014)
2. S. Sakr, Big Data 2.0 Processing Systems (Springer, Switzerland, 2016)
3. A.Y. Zomaya and S. Sakr (eds.), Handbook of Big Data Technologies, Springer International Publishing AG 2017
4. D. Moody, “A practical methodology for the representation of large data models,” in Proceedings of the Australian Database and Information Systems Conference, 1991.
5. Ras. J, ETL - Extract, Transform, Load: Data Analytics Study Guide, Student Study Guides (August 5, 2018)

6. April Reeve, Managing Data in Motion: Data Integration Best Practice Techniques and Technologies (The Morgan Kaufmann Series on Business Intelligence), Morgan Kaufmann; 1 edition (March 15, 2013)
7. M. Bala, O. Boussaid, and Z. Alimazighi, Big-ETL: Extracting-Transforming-Loading Approach for Big Data, Int'l Conf. Par. and Dist. Proc. Tech. and Appl. | PDPTA'15
8. Abdullah, Marwah N., Alaa Hassan, and Nadia Naef., 2016, Knowledge-Based Analysis of Web Data Extraction, Proceedings of the Fifth International Conference on Informatics and Applications, Takamatsu, Japan, ISBN: 978-1-941968-41-3 SDIWC 26.
9. Grolemund, G., Hands-On Programming with R, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. 2014

7. Ресурсно осигуряване на обучението:

Не е предвидено специализирано ресурсно осигуряване.

8. Критерии за оценка

Изпитът се състои от две части – писмен и устен.
На писмения изпит докторантът развива своите идеи и концепции по два въпроса от конспекта.
На устния изпит докторантът отговаря на зададени от журито въпроси, свързани с темата на курса.
Крайната оценка е от 2 до 6 (с точност до 0.5).
Тя се формира на базата на следното съответствие:

Отличен (6)	Мн. добър (5)	Добър (4)	Среден (3)	Слаб (2)
Отлично владее материала. Изложението е изчерпателно, последователно, компетентно, логично и хармонично. Правилно обосновава предлаганите решения, знае как да обобщава и излага материала без да прави грешки. Притежава необходимите умения за изпълнение на практически задачи.	Познава материала. Излага го правилно без да допуска съществени неточности. Може правилно да прилага теоретични принципи и притежава необходимите умения за изпълнение на практически задачи.	Владее голяма част материала, но допуска неточности при изложението и отговорите на въпросите. Има известни неясноти при опитите за прилагане на материала в практически ситуации.	Владее само част от материала, но се затруднява в отделните детайли. Допуска неточности във формулировките и нарушава последователността при представянето на материал. Има затруднения при изпълнение на практически задачи.	Не познава значителна част от материала, допуска съществени грешки и с големи трудности изпълнява практически задачи.

Учебната програма е обсъдена и одобрена на заседание на секция „Математическа лингвистика“ на 28.02.2020.

Ръководител секция:
(доц. д-р Десислава Панева-Маринова)

Учебната програма е разгледана от Директорския съвет на ИМИ-БАН
на 12.03.2020 г. (протокол № 10).

Учебната програма е приета от Научния съвет на ИМИ-БАН
на 13.03.2020 г. (протокол № 4).