Syllabus Attività Formativa

Anno Offerta	2020
Corso di Studio	SM35 - DATA SCIENCE AND SCIENTIFIC COMPUTING
Regolamento Didattico	SM35-18-20
Percorso di Studio	SM35+3+ - Data Science and Applications
Insegnamento/Modulo	583SM - DATA MANAGEMENT FOR BIG DATA - DATA MANAGEMENT FOR BIG DATA
Attività Formativa Integrata	-
Partizione Studenti	-
Periodo Didattico	S2 - Secondo Semestre
Sede	TRIESTE
Anno Corso	1
Settore	INF/01 - INFORMATICA
Tipo attività Formativa	B - Caratterizzante
Ambito	50437 - Discipline matematiche, fisiche e informatiche
CFU	9.0
Ore Attività Frontali	72.0
AF_ID	291658

Tipo Testo	Codice Tipo Testo	Num. Max. Caratteri	Obbl.	Testo in Italiano	Testo in Inglese
Lingua insegnament	LINGUA_INS	3800	Sì	English.	English.

o					
Contenuti (Dipl.Sup.)	CONTENUTI	3800	Sì	The course consists of 3 parts. 1. Fundamentals of database systems. The students will learn, and practice, how to design, develop, populate, and manipulate (query and update) a relational database (data models, integrity constraints, normal forms, query and update languages, transactions, indexes). 2. Advanced database models, languages, and systems. The students will learn, and practice, advanced query processing techniques for relational databases. They will also be introduced to the basic elements of distributed and parallel database management systems that play a fundamental role in the management of big data. Moreover, alternative data models and languages (e.g., XML databases) are introduced. 3. Data analysis and big data. The students will learn, and practice, the main techniques and tools for data analysis and big data management. A special attention	The course consists of 3 parts. 1. Fundamentals of database systems. The students will learn, and practice, how to design, develop, populate, and manipulate (query and update) a relational database (data models, integrity constraints, normal forms, query and update languages, transactions, indexes). 2. Advanced database models, languages, and systems. The students will learn, and practice, advanced query processing techniques for relational databases. They will also be introduced to the basic elements of distributed and parallel database management systems that play a fundamental role in the management of big data. Moreover, alternative data models and languages (e.g., XML databases) are introduced. 3. Data analysis and big data. The students will learn, and practice, the main techniques and tools for data analysis and big data management. A special

				will be given to practical use cases, data warehousing, and methods and tools for big data. A number of key topics will be addressed, ranging from the MapReduce paradigm to time series and text analytics.	attention will be given to practical use cases, data warehousing, and methods and tools for big data. A number of key topics will be addressed, ranging from the MapReduce paradigm to time series and text analytics.
Testi di riferimento	TESTI_RIF	3800	Sì	- Fundamentals of Database Systems (7th Edition), Elmasri and Navathe, Pearson, 2016 - Basi di dati (5^ edizione o edizioni precedenti - in italiano), Atzeni, Ceri, Fraternali, Paraboschi, and Torlone, McGraw-Hill, 2018 - Database System Concepts (7th Edition), Silberschatz, Korth, and Sudarshan, McGraw-Hill, 2020 - Readings in Database Systems (online, http://www.redbook.io) - Principles of Distributed Database Systems (3rd Edition), Özsu and Valduriez, Springer, 2011 - Data Warehouse Systems - Design and Implementation, A. Vaisman, E. Zimányi, Springer, 2014 - Business Analytics: A Contemporary Approach, Thomas Jackson, Steven Lockwood, WHSmith, 2018 - SQL & NoSQL Databases - Models, Languages, Consistency Options and Architectures for Big Data Management,	- Fundamentals of Database Systems (7th Edition), Elmasri and Navathe, Pearson, 2016 - Basi di dati (5^ edizione o edizioni precedenti - in italiano), Atzeni, Ceri, Fraternali, Paraboschi, and Torlone, McGraw-Hill, 2018 - Database System Concepts (7th Edition), Silberschatz, Korth, and Sudarshan, McGraw-Hill, 2020 - Readings in Database Systems (online, http://www.redbook.io) - Principles of Distributed Database Systems (3rd Edition), Özsu and Valduriez, Springer, 2011 - Data Warehouse Systems - Design and Implementation, A. Vaisman, E. Zimányi, Springer, 2014 - Business Analytics: A Contemporary Approach, Thomas Jackson, Steven Lockwood, WHSmith, 2018 - SQL & NoSQL Databases - Models, Languages, Consistency Options and Architectures for Big Data Management,

				Andreas Meier, Michael Kaufmann, Springer, 2019 - Text Mining: Concepts, Implementation, and Big Data Challenge (1st Edition), Taeho Jo, Springer, 2019 - Temporal Data Mining, Theophano Mitsa, CRC Press, 2010. - Hadoop: The Definitive Guide (4th Edition), Tom White, O'Reilly, 2015. - The MongoDB 4.2 Manual, MongoDB, Inc., https://docs.mongodb.com/manual/	Andreas Meier, Michael Kaufmann, Springer, 2019 - Text Mining: Concepts, Implementation, and Big Data Challenge (1st Edition), Taeho Jo, Springer, 2019 - Temporal Data Mining, Theophano Mitsa, CRC Press, 2010. - Hadoop: The Definitive Guide (4th Edition), Tom White, O'Reilly, 2015. - The MongoDB 4.2 Manual, MongoDB, Inc., https://docs.mongodb.com/manual/
Obiettivi formativi	OBIETT_FORM	3800	Sì	The students must learn how to organize, manipulate, and analyze small and big data with a variety methods, techniques, and tools. Knowledge and understanding: the students must acquire the necessary knowledge to model, import, tidy, transform, query, visualize, and analyze data as well as to communicate the results of the analysis. We take into consideration relational data as well as semistructured and unstructured data. Applied knowledge and understanding: the students must learn	The students must learn how to organize, manipulate, and analyze small and big data with a variety methods, techniques, and tools. Knowledge and understanding: the students must acquire the necessary knowledge to model, import, tidy, transform, query, visualize, and analyze data as well as to communicate the results of the analysis. We take into consideration relational data as well as semistructured and unstructured data. Applied knowledge and understanding: the students must learn

				languages and tools for the manipulation, analysis, and visualization of data, such as, for instance, PostgreSQL and BaseX for the management of relational and XML data, R and RStudio environment for data analysis and visualization, Processing for the visualization of data, and R Markdown language for the communication of the results of the analysis. Making judgements: the students must be able to interpret the experimental results of the analysis and draw effective conclusions relevant to the domain of discourse. Communication skills: the students must be able to communicate effectively the results of the analysis. This includes both analyst-to-analyst communication and analyst-to-decision-maker communication. Learning skills: the students must demonstrate that they have learned the ability to choose a sufficiently rich row data set, to analyze the data to extract	languages and tools for the manipulation, analysis, and visualization of data, such as, for instance, PostgreSQL and BaseX for the management of relational and XML data, R and RStudio environment for data analysis and visualization, Processing for the visualization of data, and R Markdown language for the communication of the results of the analysis. Making judgements: the students must be able to interpret the experimental results of the analysis and draw effective conclusions relevant to the domain of discourse. Communication skills: the students must be able to communicate effectively the results of the analysis. This includes both analyst-to-analyst communication and analyst-to-decision-maker communication. Learning skills: the students must demonstrate that they have learned the ability to choose a sufficiently rich row data set, to analyze the data to extract

				meaningful information, to draw and to communicate conclusions	meaningful information, to draw and to communicate conclusions
Prerequisiti	PREREQ	3800	Sì	Some basic knowledge about programming, algorithms and data structures, logic, and statistics are desirable	Some basic knowledge about programming, algorithms and data structures, logic, and statistics are desirable
Metodi didattici	METODI_DID	3800	Sì	The course is organized in three parts and it is given by two different teachers; in addition, some advanced topics might be covered by invited experts in the field. Classes mainly consist in lectures given by the teacher. Students are also introduced to software resources to download, install, and run for the first time: the teacher will give a brief practical introduction to them.	The course is organized in three parts and it is given by two different teachers; in addition, some advanced topics might be covered by invited experts in the field. Classes mainly consist in lectures given by the teacher. Students are also introduced to software resources to download, install, and run for the first time: the teacher will give a brief practical introduction to them.
Altre informazioni	ALTRO	3800	Sì	Additional suggested books: - PostgreSQL: Up and Running (3rd Edition), Regina Obe and Leo Hsu, O'Reilly Media, 2017 - An Introduction to XML and Web Technologies, Anders Møller and Michael I. Schwartzbach, Addison-Wesley, 2006 - Building the Data Warehouse (4th Edition), W. I. Immon, Wiley Publishing, 2005 - Big Data: A Very Short Introduction, Dawn Holmes, Oxford, 2017 - The Design and Implementation of Modern Colum-Oriented Database Systems, Daniel	Additional suggested books: - PostgreSQL: Up and Running (3rd Edition), Regina Obe and Leo Hsu, O'Reilly Media, 2017 - An Introduction to XML and Web Technologies, Anders Møller and Michael I. Schwartzbach, Addison-Wesley, 2006 - Building the Data Warehouse (4th Edition), W. I. Immon, Wiley Publishing, 2005 - Big Data: A Very Short Introduction, Dawn Holmes, Oxford, 2017 - The Design and Implementation of Modern

				Abadi, Peter Boncz, Stavros Harizopoulos, Stratos Idreos, Samuel Madden, 2013 - What's Really New with NewSQL?, A. Pavlo and M. Aslett, ACM SIGMOD Record, Vol. 45, No. 2, pages 45-55, June 2016 - Column-Oriented Database Systems (slides), Stavros Harizopoulos, Daniel Abadi, and Peter Boncz, VLDB 2009 Tutorial, http://nms.csail.mit.edu/~stavros/pubs/tutorial2009-column_stores.pdf - Graph Databases (2nd Edition), Ian Robinson, Jim Webber, and Emil Eifrem, O'Reilly Media, 2015 - Big Data Management and NoSQL Databases - Lecture 7. Column-family stores (slides), Irena Holubova, https://www.ksi.mff.cuni.cz/~svoboda/courses/2015-1-NDBI040/lectures/Lecture-07-Column.pdf - Tutorial by Jeffrey Heer on Text Visualization (CSR 512 - Data Visualization), University of Washington - Introduction to Time Series Mining (slides), Keogh Eamonn - Temporal Data Mining, Theophano Mitsa, Taylor & Francis Ltd, 2010 - Apache Hadoop Online Documentation, Pig Latin Basics, https://pig.apache.org/docs/latest/basic.html	Colum-Oriented Database Systems, Daniel Abadi, Peter Boncz, Stavros Harizopoulos, Stratos Idreos, Samuel Madden, 2013 - What's Really New with NewSQL?, A. Pavlo and M. Aslett, ACM SIGMOD Record, Vol. 45, No. 2, pages 45-55, June 2016 - Column-Oriented Database Systems (slides), Stavros Harizopoulos, Daniel Abadi, and Peter Boncz, VLDB 2009 Tutorial, http://nms.csail.mit.edu/~stavros/pubs/tutorial2009-column_stores.pdf - Graph Databases (2nd Edition), Ian Robinson, Jim Webber, and Emil Eifrem, O'Reilly Media, 2015 - Big Data Management and NoSQL Databases - Lecture 7. Column-family stores (slides), Irena Holubova, https://www.ksi.mff.cuni.cz/~svoboda/courses/2015-1-NDBI040/lectures/Lecture-07-Column.pdf - Tutorial by Jeffrey Heer on Text Visualization (CSR 512 - Data Visualization), University of Washington - Introduction to Time Series Mining (slides), Keogh Eamonn - Temporal Data Mining, Theophano Mitsa, Taylor & Francis Ltd, 2010 - Apache Hadoop Online Documentation, Pig Latin Basics,

				- Hadoop Platform and Application Framework - Tutorial offered on Coursera by the University of California San Diego - MongoDB 4 Quick Start Guide, Doug Bierer, Packt Publishing Ltd, 2018 - Mastering MongoDB 3.x, Alex Giamas, Packt Publishing, 2017 - MongoDB Architecture Guide, MongoDB, Inc., http://s3.amazonaws.com/info-mongodb-com/MongoDB_Architecture_Guide.pdf - MongoDB Data Modeling, Wilson da Rocha França, Packt Publishing Ltd, 2015	https://pig.apache.org/docs/latest/basic.html - Hadoop Platform and Application Framework - Tutorial offered on Coursera by the University of California San Diego - MongoDB 4 Quick Start Guide, Doug Bierer, Packt Publishing Ltd, 2018 - Mastering MongoDB 3.x, Alex Giamas, Packt Publishing, 2017 - MongoDB Architecture Guide, MongoDB, Inc., http://s3.amazonaws.com/info-mongodb-com/MongoDB_Architecture_Guide.pdf - MongoDB Data Modeling, Wilson da Rocha França, Packt Publishing Ltd, 2015
Modalità di verifica dell'apprendimento	MOD_VER_APPR	3800	Sì	The exam consists of a written test and, possibly, an additional oral examination.	The exam consists of a written test and, possibly, an additional oral examination.
Programma esteso	PROGR_EST	3800	Sì	Part 1 - 3 cfu (24 hours): Fundamentals of database systems - Introduction to the DataBase Management Systems (DBM) - 2 hours - Data models - 2 hours + Conceptual models (Entity-Relationship / ER Model) + Logical models (Relational Model)	Part 1 - 3 cfu (24 hours): Fundamentals of database systems - Introduction to the DataBase Management Systems (DBM) - 2 hours - Data models - 2 hours + Conceptual models (Entity-Relationship / ER Model) + Logical models (Relational Model)

				- Design methodologies - 4 hours + Mapping ER schemas into relational ones + Functional dependencies and normalization - Data definition, update, and query languages - 8 hours + Relation algebra and relational Calculus + SQL - Transactions - 4 hours - Indexes - 4 hours Part 2 - 3 cfu (24 hours): Advanced database models, languages, and systems - Query processing and optimization - 6 hours + Query processing + Algorithms for the join operation + Cost-based optimization and heuristics - Distributed and parallel database architectures - 12 hours + An introduction to parallel and distributed DBMS + Design of distributed databases (fragmentation and replication) + Distributed query processing	- Design methodologies - 4 hours + Mapping ER schemas into relational ones + Functional dependencies and normalization - Data definition, update, and query languages - 8 hours + Relation algebra and relational Calculus + SQL - Transactions - 4 hours - Indexes - 4 hours Part 2 - 3 cfu (24 hours): Advanced database models, languages, and systems - Query processing and optimization - 6 hours + Query processing + Algorithms for the join operation + Cost-based optimization and heuristics - Distributed and parallel database architectures - 12 hours + An introduction to parallel and distributed DBMS + Design of distributed databases (fragmentation and replication) + Distributed query processing

				+ Optimization of distributed queries + Transaction processing in distributed databases: the two-phase commit (2PC) protocol + Parallel DBMS - Semistructured Data and XML - 4 hours + Definition of semistructured data in XML + Querying XML data (XPath and XQuery) + XML and relational DBMS + Native XML databases - Cloud computing and DBMS - 2 hours Part 3 - 3 cfu (24 hours): Data analysis and big data - Data warehousing - 4 hours - Data mining - 6 hours + Time series analysis + Text mining - Fundamentals of big data - 6 hours + Distinctive features, data science, and applications + Technologies for the management of big data + Hadoop and Map Reduce	+ Optimization of distributed queries + Transaction processing in distributed databases: the two-phase commit (2PC) protocol + Parallel DBMS - Semistructured Data and XML - 4 hours + Definition of semistructured data in XML + Querying XML data (XPath and XQuery) + XML and relational DBMS + Native XML databases - Cloud computing and DBMS - 2 hours Part 3 - 3 cfu (24 hours): Data analysis and big data - Data warehousing - 4 hours - Data mining - 6 hours + Time series analysis + Text mining - Fundamentals of big data - 6 hours + Distinctive features, data science, and applications + Technologies for the management of big data + Hadoop and Map Reduce

				- NoSQL systems - 8 hours + Key-value stores + Document stores + Column Family stores + Graph Databases + MongoDB and applications	- NoSQL systems - 8 hours + Key-value stores + Document stores + Column Family stores + Graph Databases + MongoDB and applications