Список дисциплин профиля Технологии обработки больших данных

Первый семестр (осень) (3 дисциплины)

1. **Introduction to Big Data**

This course covers fundamental issues in Big Data. The course examines issues related to data organization, storage, retrieval, analysis and knowledge discover at scale. This will include topics such as large-scale data analysis, data storage systems, self-descriptive data representations, semi-structured data models.

2. **Programming languages (Python)**

3. **Data Base**

This course discusses foundational concepts of database systems, explores the role of databases in data warehousing, teaches the fundamental principles guiding data warehousing and Cloud-based online analytical processing (OLAP). Various database models are discussed with emphasis on the relational model and relational database design. Students will become proficient in SQL. Topics also include requirement gathering for data warehousing, data warehouse architecture, data warehouse development methodology/issues surrounding the planning of the data warehouse, physical database design for data warehousing, and extracting, transforming, and loading strategies.

Второй семестр (4 дисциплины)

1. **Data Analysis Methods**

This course will provide fundamental techniques for data analytics, including data collection, data extraction, data integration and data cleansing. The students will learn how to manage and optimize the analytics value chain, including collecting and extracting the suitable values, selecting the right data processing processes, integrating the data from various resources, data governance, security and privacy for Big Data applications.

2. **Distributed Systems and Cloud Computing**

This course provides the student a basic grounding in designing and implementing distributed and cloud systems. Course curriculum combines hands-on experience in developing cloud services, with a firm grounding in the tools and principles of building distributed and cloud applications. The course covers best practices for architecting, scaling and operating services

that perform reliably and well. This course include fundamental algorithms for distributed system architectures, inter-process communications, data consistency and replication, distributed transactions and concurrency control, distributed file systems, network transparency, fault tolerant, distributed systems synchronization, and reliability. Topics covered include distributed computing, interactive services, collaborative computing, and peer-to-peer sharing. Various distributed frameworks and technologies will be explored.

3. Large Scale Data Bases

This course refers to a set of techniques that have been designed to efficiently find important information or knowledge in large amounts of data. This course will provide students with understanding of the industry standard data mining methodologies, and with the ability of extracting information from a data set and transforming it into an understandable structure for further use. Topics covered include decision trees, classification, predictive modeling, association analysis, statistical modeling, Bayesian classification, anomaly detection, visualization. The course will be complemented with hands-on experience of using advanced data mining software to solve realistic problems based on real-world data

4. Big Data Programming Tools

The course topics include: Big Data concepts; Hadoop core technologies & ecosystem; planning for and installing a Hadoop cluster; the Hadoop Distributed File System(HDFS); MapReduce; managing and scheduling jobs with FIFO Scheduler and Fair Scheduler; cluster administration and management; installing and managing other Hadoop projects such as Apache Pig and Apache Hive; and, use cases and best practices for processing Big Data with Hadoop.

Третий семестр (4 дисциплины + 4 альтернативных)

1. Machine Learning

This course covers the theory and practical algorithms for machine learning from a variety of perspectives. Topics include decision tree learning, parametric and non-parametric learning, Support Vector Machines, statistical learning methods, unsupervised learning, reinforcement learning and adaptive control. Students will have an opportunity to experiment with machine learning techniques and apply them to solve a selected problem in the context of a term project. The course will also draw from numerous case studies and applications, so that students learn how to apply learning algorithms to build machine intelligence.

// **Imaging: Data Analytics and Pattern Recognition**

This course will introduce the basic techniques for image data analytics and pattern recognition. Topics include image processing and analysis in spatial and frequency domains, image restoration and compression, image segmentation and registration, morphological image processing, representation and description, feature description, face

recognition, iris recognition, fingerprint recognition, image analysis topics, such as medical image analysis.

2. Big Data Analytics

Big Data (Structured, semi-structured, & unstructured) refers to large datasets that are challenging to store, search, share, visualize, and analyze. Gathering and analyzing these large data sets are quickly becoming a key basis of competition. This course explores several key technologies used in acquiring, organizing, storing, and analyzing big data. Topics covered include Hadoop, unstructured data concepts (key-value), Map Reduce technology, related tools that provide SQL-like access to unstructured data: Pig and Hive, NoSQL storage solutions like HBase, Cassandra, and Oracle NoSQL and analytics for big data. A part of the course is devoted to public Cloud as a resource for big data analytics. The objective of the course is for students to gain the ability to employ the latest tools, technologies and techniques required to analyze, debug, iterate and optimize the analysis to infer actionable insights from Big Data.

// Parallel Programming

Introduction to parallel computer architectures; principles of parallel algorithm design; shared-memory programming models; message passing programming models used for cluster computing; data-parallel programming models for GPUs; case studies of parallel algorithms, systems, and applications; hands-on experience with writing parallel programs for tasks of interest

3. Data Visualization

This course will introduce visualization techniques for data from everyday life, social media, business, scientific computing, medical imaging, etc. The topics include human visual system and perception, visual design principles, open- source visualization tools and systems, visualization techniques for CT/MRI data, computational fluid dynamics, graphs and networks, time-series data, text and documents, Twitter data, and spatio-temporal data.

// Knowledge Management Systems

This course is an introduction to the fundamental concepts of knowledge management (KM). The aim of the course is to give students a set of frameworks and concepts that can be used to manage individual and organizational knowledge. The contents of the course includes theories and basic concepts on knowledge management, different types of knowledge and KM, an important of such models in KM, procedures of KM including knowledge identifying, acquiring, restoring, retrieving, transferring, using, and theories and methods on knowledge innovation and evaluation of KM . Foundations and business applications of knowledge management and intelligent systems, organisational learning,

ontologies and the semantic web, knowledge-based and expert systems, knowledge discovery, acquisition and learning, representation and reasoning, intelligent agents over the Internet with e-commerce applications, document and content management systems.

4. **Web Data Mining**

Covers topics ranging from Data mining, Web mining, classification, numeric prediction, association rules, sequential patterns, Web crawling, retrieval and search engines, social network analysis, link analysis, ranking, Web personalization and recommender systems, to Hadoop based technologies, including MapReduce, Spark, MLlib

// Special Topics in Big Data

Advanced Big Data Analytics. Advanced Distributed Systems. Programming with Big Data in R.