

APPROVED BY

Director of *Institute of Cybernetics*
 / *Dmitriy M. Sonkin*

Big Data Programming Tools

Field of Study: Big Data

Programme name: 09.04.04 Software Engineering

Level of Study: Master Degree Programme

Year of admission: 2019

Semester, year: 2, 1

ECTS: 3

Total Hours: 108


Contact Hours: 32

- **Lectures:** 16
- **Labs:** 16
- **Practical experience:** 0

Assessment: exam

Department: Software Engineering

Head of Department

 / Sherstnev V.S.

Instructor(s)

 / Gubin M.Y.

Big Data Programming Tools

Course Overview

Course Objectives	The objective of the course is to prepare students to process Big Data using the Hadoop ecosystem.
Learning Outcomes	<p>After completing this course, the students should:</p> <ul style="list-style-type: none"> • Apply knowledge of approaches to big data processing to determine which ecosystem and in what configuration can be used to solve a specific big data processing task. • Apply the skills of administering large data warehouses to plan for the configuration, deployment, and administration of Hadoop repositories. • Be able to establish data exchange between Hadoop clusters and various data sources and consumers. • Conduct theoretical and experimental research, including the search and study of necessary scientific and technical information, mathematical modeling, analysis and interpretation of the results, in the field of big data processing. • Be able to use the Hadoop ecosystem toolkit (Pig, Hive, Spark) to process data using the Hadoop cluster.
Course Outline	The course consists of 8 lectures and 8 labs covering the following main topics: introduction to Big Data concepts; Hadoop ecosystem and its core technologies; introduction to planning and deploying a Hadoop cluster; using the Hadoop Distributed File System (HDFS); Hadoop cluster administration and management; installing, using and managing other Hadoop projects such as Apache Pig, Apache Spark, and Apache Hive; and, use cases and best practices for processing Big Data with Hadoop. After completing the course, the students are expected to be capable of planning and deploying a Hadoop cluster tailored to the data, managing and administering, it, importing data, processing it and getting the results in a form convenient for further use.
Prerequisites (if available)	Introduction to Big Data, Programming languages (Python)
Course Structure	The course consists of two parts. The first part, “Introduction to Big Data”, introduces students to the concept of Big Data, outlines specific challenges that need to be solved to process Big Data successfully, and explains basics of installation and configuration of the Hadoop cluster to accommodate the needs of Big Data processing in various example cases. The second part, “Hadoop for Big Data processing”, provides an in-depth explanation of advanced big data processing methods using the various tools available in the Hadoop ecosystem, such as Pig, Hive, and Spark.
Facilities and Equipment	<p>3 servers with Big Data processing software (HP DL385p Gen8, 2 processors 6320 (2.8GHz-16MB) 8-Core Processor Option Kit, 6 Memory modules 8GB 2Rx4 PC3L-10600R-9 , RAID controller P420i (512MB) FBWC RAID 0,1,1+0,5,5+0, 11 HDD 500GB SC 6G 7.2K LFF SATA HotPlug Midline Drive 1y war, Flash drive 120GB 6G SATA VE 3.5in SCC EV G1 SSD)</p> <p>Hadoop cluster (Pig, Hive, Spark)</p>

Grading Policy	<p>In accordance with TPU rating system we use:</p> <ul style="list-style-type: none"> - Current assessment which is performed on a regular basis during the semester by scoring the quality of mastering of theoretical material and the results of practical activities (labs). Max score for current assessment is 60 points, min – 30 points. - Course final assessment (exam) is performed at the end of the semester. Max score for course final assessment is 40 points, min – 25 points. <p>The final rating is determined by summing the points of the current assessment during the semester and exam (credit test) scores at the end of the semester. Maximum overall rating corresponds to 100 points, min pass score is 55 points.</p>
Course Policy	Class attendance will be taken into consideration when evaluating students' participation in the course.
Teaching Aids and Resources	Additional Readings: Kaggle (https://www.kaggle.com)
Instructor (-s)	Gubin Maksim, +79069507314, catnip@tpu.ru