Processing big data with cloud-based technologies

Lecturer: Anikin Yury Aleksandrovich, C.Sc.

Semester: 2 **Duration**: 16 weeks

Workload (h): 144 **Presence** (h + CH): 64 (8) **Self-Study** (h): 72

Contents: The course aim is to give a brief of modern cloud platforms of stream processing, adapted for parallel data processing, to demonstrate ability of processing in applied domain of Twitter data.

Background and relations to other courses: Basics of Statistics, Data Mining Tools & Languages.

Main topics and learning objectives:

1. Twitter as a data source for research
2. Twitter architecture
3. Introduction to Hadoop
4. Apache Pig
5. Twitter API for streaming: java, python
6. Trends exploration
7. Search in Twitter in real-time
8. Twitter as a social network, network analysis
9. Information distribution analytics
10. Safety in Twitter
11. Spark as a tool of interactive analytics
12. Splunk as a tool for data analysis

Assessment:

Formative: In interaction with lecturer and tutor during learning period. On site, skype, email are preferable. There are 4 task to be done during the course: (1) using Twitter API; (2) writing Pig queries; (3) capturing twitter data stream; (4) Twitter trends exploration via Spark or Splunk.

Summative:

Number and Type; Connection to Course	**Duration**	**Part of final mark in %**
Pass test	30 min	20%
Course Assignments	4 tasks	80%

Learning outcomes:

Academic: *to know* tools for stream data processing. To be able to process Twitter data with the help of Splunk, and other tools. To ba able to apply technology of big data processing and to use network analysis, to be able to use Splunk in stream data analysis.

Prerequisites for Credit Points: The credit points will be granted when the course has been successfully completed, i.e. all parts of the examination are passed.