# SYLLABUS

## 1. Information regarding the programme

| | |
|---|---|
| 1.1 Higher education institution | **Babeş-Bolyai University of Cluj-Napoca** |
| 1.2 Faculty | **Faculty of Mathematics and Computer Science** |
| 1.3 Department | **Department of Computer Science** |
| 1.4 Field of study | **Computer Science** |
| 1.5 Study cycle | **Master** |
| 1.6 Study programme / Qualification | **High Performance Computing and Big Data Analytics** |

## 2. Information regarding the discipline

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2.1 Name of the discipline | **Functional Parallel Programming for Big Data Analytics** | | | | | | |
| 2.2 Course coordinator | **Prof. Dr. Frédéric Loulergue/ Lect. Dr. Horea Grebla** | | | | | | |
| 2.3 Seminar coordinator | **Prof. Dr. Frédéric Loulergue** | | | | | | |
| 2.4. Year of study | **1** | 2.5 Semester | **1** | 2.6. Type of evaluation | **E** | 2.7 Type of discipline | **Compulsory** |

## 3. Total estimated time (hours/semester of didactic activities)

| | | | | | |
|---|---|---|---|---|---|
| 3.1 Hours per week | 3 | Of which: 3.2 course | 2 | 3.3 seminar/laboratory | 1 sem |
| 3.4 Total hours in the curriculum | 42 | Of which: 3.5 course | 28 | 3.6 seminar/laboratory | 14 |
| Time allotment: | | | | | hours |
| Learning using manual, course support, bibliography, course notes | | | | | 35 |
| Additional documentation (in libraries, on electronic platforms, field documentation) | | | | | 45 |
| Preparation for seminars/labs, homework, papers, portfolios and essays | | | | | 47 |
| Tutorship | | | | | 15 |
| Evaluations | | | | | 16 |
| Other activities: ..... | | | | | - |
| 3.7 Total individual study hours | 158 | | | | |
| 3.8 Total hours per semester | 200 | | | | |
| 3.9 Number of ECTS credits | 8 | | | | |

## 4. Prerequisites (if necessary)

| | |
|---|---|
| 4.1. curriculum | |
| 4.2. competencies | |

## 5. Conditions (if necessary)

| | |
|---|---|
| 5.1. for the course | • Students will attend the course with their mobile phones shut down |
| 5.2. for the seminar /lab | • Students will attend the seminar with their mobile phones shut down |

| | |
|---|---|
| activities | • Room with computers as needed; high level programming language environment |

**6. Specific competencies acquired**

| | |
|---|---|
| **onal competencies** | • Understanding the concepts, methods and models used in MapReduce and other cloud computing frameworks<br>• Understanding the principles, design and implementation of various analysis in a cloud computing environment<br>• Learning to conduct incipient original research in programming for cloud computing |
| **ersal competencies** | • The ability to apply the MapReduce model to scalable data analytics in solving real world problems.<br>• Responsible execution of lab assignments, research and practical reports.<br>• Application of efficient and rigorous working rules.<br>• Manifest responsible attitudes toward the scientific and didactic fields. |

**7. Objectives of the discipline** (outcome of the acquired competencies)

| | |
|---|---|
| 7.1 General objective of the discipline | • To introduce the student to MapReduce and other cloud computing frameworks for data analytics |
| 7.2 Specific objective of the discipline | • To present the field of as a novel research and application domain.<br>• To induce the necessity of scalable data analysis methods by studying some relevant practical applications<br>• To offer the student the instruments that will allow him/her to develop different cloud computing applications for data analytics. |

**8. Content**

| 8.1 Course | Teaching methods | Remarks |
|---|---|---|
| Part One - Introduction<br><br>Introduction to Big Data, Four Vs of Big Data<br>Big Data - technical landscape, industry landscape<br>Search, Indexing and MemoryEfficient storage of big data<br>Scalable querying and reporting on massive data sets;<br>Scalable and distributed hardware and software architectures<br><br>Part Two - NoSQL Systems<br><br>The differences between traditional relational databases and NoSQL databases<br>MapReduce, Column-store, Graph, Document-store<br>Big Data using a NoSQL approach: Use cases and best practices to process Big Data with NoSQL<br>Database workloads and technologies; how to select database technology<br><br>Part Three - Analytics | | |

Introduction to Big Data Analytics
Introduction to R language
Using R for Analysis of the Data; Visualization with R
Big Data algorithms
Analytics examples

- An Introduction to Programming in Scala
- High-Order Functional Programming in Scala
- The MapReduce model
- Programming MapReduce Applications with Hadoop and Scala
- Applications of MapReduce: Clustering
- Optimisation of MapReduce Applications
- In Memory MapReduce
- MapReduce for Parallel SQL Queries Evaluation
- Formal Methods and MapReduce Development

**Bibliography**


1. Martin Odersky, Lex Spoon, and Bill Venners. Programming in Scala. Artima, second edition, 2010.
2. Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Pro- cessing on Large Clusters. In OSDI, pages 137–150. USENIX Association, 2004. http://static.usenix.org/events/osdi04/tech/full_ papers/dean/dean.pdf
3. Tom White. Hadoop – The Definitive Guide. O'Reilly, second edition, 2010.
4. Makoto Onizuka, Hiroyuki Kato, Soichiro Hidaka, Keisuke Nakano, and Zhenjiang Hu. Optimization for iterative queries on mapreduce. PVLDB, 7 (4):241–252, 2013.
5. Cliff Engle, Antonio Lupher, Reynold Xin, Matei Zaharia, Michael J. Franklin, Scott Shenker, and Ion Stoica. Shark: fast data analysis using coarse-grained distributed memory. In SIGMOD, pages 689–692, New York, NY, USA, 2012. ACM. Doi:10.1145/2213836.2213934.
6. M. Al Hajj Hassan and M. Bamha. Semi-join computation on distributed file systems using the map-reduce-merge model. In ACM SAC, pages 406–413, New York, NY, USA, 2010. ACM. Doi:10.1145/1774088.1774174.
7. Mohamad Al Hajj Hassan, Mostafa Bamha, and Frédéric Loulergue. Handling Data-skew Effects in Join Operations using MapReduce. In ICCS. Elsevier, 2014, to appear
8. Kento Emoto, Sebastian Fischer, and Zhenjiang Hu. Filter-embedding semir- ing fusion for programming with MapReduce. Formal Asp. Comput., 24(4-6): 623–645, 2012. doi:10.1007/s00165-012-0241-8.
9. Kento Emoto, Frédéric Loulergue, and Julien Tesson. A Verified Generate- Test-Aggregate Coq Library for Parallel Programs Extraction. 2014, to appear


| 8.2 Seminar / laboratory | Teaching methods | Remarks |
|---|---|---|
| 1. Administration. Survey of the sources of information available on Internet and Intranet. Chosing the paper topics and scheduling the presentations. | • Interactive exposure<br>• Explanation<br>• Conversation | |
| 2. Delivery of theoretical report | • Interactive exposure<br>• Explanation<br>• Conversation | |
| 3. Delivery of theoretical report | • Interactive exposure<br>• Explanation<br>• Conversation | |

| | | |
|---|---|---|
| 4. Delivery of experimental report | • Interactive exposure<br>• Explanation<br>• Conversation | |
| 5. Delivery of experimental report | • Interactive exposure<br>• Explanation<br>• Conversation | |
| 6. Delivery of software project | • Interactive exposure<br>• Explanation<br>• Conversation | |
| 7. Delivery of software project | • Interactive exposure<br>• Explanation<br>• Conversation | |

**Bibliography**


1. Martin Odersky, Lex Spoon, and Bill Venners. Programming in Scala. Ar- tima, second edition, 2010.
2. Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Pro- cessing on Large Clusters. In OSDI, pages 137–150. USENIX Association, 2004. http://static.usenix.org/events/osdi04/tech/full_ papers/dean/dean.pdf
3. Tom White. Hadoop – The Definitive Guide. O'Reilly, second edition, 2010.
4. Makoto Onizuka, Hiroyuki Kato, Soichiro Hidaka, Keisuke Nakano, and Zhenjiang Hu. Optimization for iterative queries on mapreduce. PVLDB, 7 (4):241–252, 2013.
5. Cliff Engle, Antonio Lupher, Reynold Xin, Matei Zaharia, Michael J. Franklin, Scott Shenker, and Ion Stoica. Shark: fast data analysis using coarse-grained distributed memory. In SIGMOD, pages 689–692, New York, NY, USA, 2012. ACM. Doi:10.1145/2213836.2213934.
6. M. Al Hajj Hassan and M. Bamha. Semi-join computation on distributed file systems using the map-reduce-merge model. In ACM SAC, pages 406–413, New York, NY, USA, 2010. ACM. Doi:10.1145/1774088.1774174.
7. Mohamad Al Hajj Hassan, Mostafa Bamha, and Frédéric Loulergue. Handling Data-skew Effects in Join Operations using MapReduce. In ICCS. Elsevier, 2014.
8. Kento Emoto, Sebastian Fischer, and Zhenjiang Hu. Filter-embedding semir- ing fusion for programming with MapReduce. Formal Asp. Comput., 24(4-6): 623–645, 2012. doi:10.1007/s00165-012-0241-8.
9. Kento Emoto, Frédéric Loulergue, and Julien Tesson. A Verified Generate- Test-Aggregate Coq Library for Parallel Programs Extraction. 2014, to appear


**9. Corroborating the content of the discipline with the expectations of the epistemic community, professional associations and representative employers within the field of the program**

About Google's MapReduce and Hadoop, the content is similar to content found in some disciplines around the world. The approach of using a functional language is original and motivated by the fact that by essence the MapReduce framework is functional. Other content about the optimisation of MapReduce programs, the algorithms and implementations techniques for SQL queries evaluation with MapReduce and formal methods for MapReduce are original are related to state-of-the-art research results in this area, all of them being of real practical interest.

**10. Evaluation**

| Type of activity | 10.1 Evaluation criteria | 10.2 Evaluation methods | 10.3 Share in the grade (%) |
|---|---|---|---|
| 10.4 Course | • The correctness and completeness of the accumulated knowledge. | Written exam (in the regular session) | 30% |

| | | | |
|---|---|---|---|
| | • A theoretical research report on an agent based topic, based on some recent research papers should be prepared and presented | Evaluation of the research report (a written paper of about 10 pages and an oral presentation) | 20% |
| 10.5 Seminar/lab activities | • Class attendance | Grade awarded pro rata | 10% |
| | • A theoretical research report on an agent based topic, based on some recent research papers should be prepared and presented | Evaluation of the research report (a written paper of about 10 pages and an oral presentation) | 20% |
| | • A personal software project fully implemented, without using only existing a MapReduce environment. | Evaluation of the project (software implementation, documentation and demonstration) | 20% |
| 10.6 Minimum performance standards | | | |
| • Each student has to prove that (s)he acquired an acceptable level of knowledge and understanding of the MapReduce domain, that (s)he is capable of stating these knowledge in a coherent form, that (s)he has the ability to establish certain connections and to use the knowledge in solving different problems.<br>• Penalty points are awarded for delays in submission of proposed topic choices and submission of final reports.<br>• Successful passing of the exam is conditioned by the final grade that has to be at least 5; the written exam grade has to be at least 5. | | | |

| Date | Signature of course coordinator | Signature of seminar coordinator |
|---|---|---|
| 06.03.2014 | Prof. Dr. Frédéric Loulergue | Prof. Dr. Frédéric Loulergue |
| Date of approval | | Signature of the head of department |
| | | Prof. dr. Bazil Pârv |