The 10th International Conference on eLearning (eLearning-2019),
26 - 27 September 2019, Belgrade, Serbia

BIG DATA ANALYTICS: LECTURES FROM THE LAMBDA NETWORK

VALENTINA JANEV, DEJAN PAUNOVIĆ
University of Belgrade, Institute Mihajlo Pupin, valentina.janev@pupin.rs

DAMIEN GRAUX,
Fraunhofer IAIS, Enterprise Information System, Damien.Graux@iais.fraunhofer.de

HAJIRA JABEEN
University of Bonn, Smart Data Analytics, jabeen@iai.uni-bonn.de

EMANUEL SALLINGER, SAHAR VAHDATI
University of Oxford, Department of Computer Science, emanuel.sallinger@cs.ox.ac.uk

Abstract: *Big Data Analytics is crucial component of the Big Data paradigm and corresponds to the knowledge extraction from enormous amount of data. As the number of Big Data related methods, tools, frameworks and solutions is growing, there is a need to systematize the knowledge about the domain. Hence, in the LAMBDA project framework an effort was made to develop a new set of lectures based on the education materials and courses offered by the University of Bonn and University of Oxford. This paper provides an overview of the LAMBDA training infrastructure and the open education contents developed so far, as well as initial evaluation about the possibilities for adoption of lectures at high-education institutions in the West Balkan countries.*

Keywords: *E-Learning, Big Data Analytics, open education, platform, lectures*

1. INTRODUCTION

Big Data refers to data sets which have large sizes and complex structures. The data size can range from dozens of terabytes to a few Zettabytes and is still growing [1]. While more than 800,000 Petabyte (1 PB= 1015bytes) of data were stored in the year 2000, this volume will reach 35 Zettabytes (1 ZB= 1021bytes) by end of 2019 [2], and is expected to grow 61% and exceeds 175 zettabytes by 2025 as per International Data Corporation IDC expectations [3]. Big Data Analytics, hence, refers to the strategy of analysing large volumes of data that gathered from a wide variety of sources, including social networks, transaction records, videos, digital images and different kind of sensors.
Challenges [4] related to the European ability to exploit the potential of Big Data are fragmentation of the data ecosystem, due to different national policies, languages, and sectors involved; fragmentation of data research efforts and lack of effective exchange of results; shortage of highly skilled persons for data-related jobs; and the complicated process of updating legislation. In an attempt to support the European data economy policy [5], in the LAMBDA project framework, the LAMBDA consortium proposed a training approach and established the infrastructure for collaborative work of LAMBDA teachers/trainers (see Figure 1, providers) with PhD students and other interested parties (see Figure 1, Consumers).

This paper describes the infrastructure that was established to reinforce organizational learning and capacity building at the Institute Mihajlo Pupin (PUPIN) and to facilitate teachers-trainees cooperation in the larger network of experts in the field of Enterprise Knowledge Graphs (EKGs), Semantic Big Data Architectures (ARCH), and Smart Data Analytics (SBDA).

The paper is structured as follows. Section 2 introduces the LAMBDA project, the learning approach and the learning infrastructure. Section 3 provides an overview of the newly developed lectures on topics from the Big Data Analytics domain. Section 4 presents the first evaluation of the lectures from PhD students and professors perspective, while Section 5 concludes the paper.

2. LAMBDA OPEN EDUCATION APPROACH

LAMBDA Project

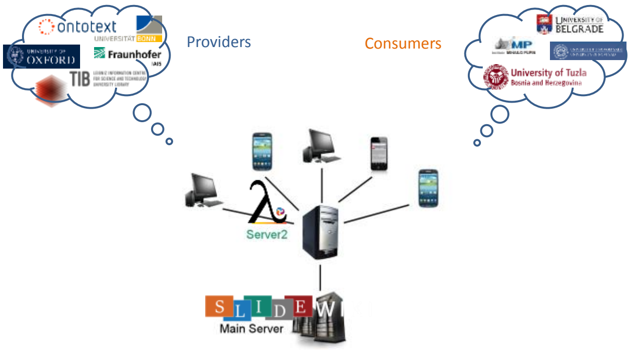
The overall objective of the project is to stimulate scientific excellence and innovation capacity, to increase the research capacities and to unlock the research potential of the biggest and the oldest R&D Institute in the ICT area in the whole West Balkan region, turning the Institute Mihajlo Pupin into a regional point of reference when it comes to multidisciplinary ICT competence related to Big Data analytics. In July 2018, the LAMBDA consortium started activities for improving the skills and competences for smart data management through a set of actions including:

- development of a Knowledge repository (Learning and Consulting Platform) that shall facilitate spreading excellence and exchange of learning materials and best practice between the international leading organizations and research institutions and Industry from the West Balkan countries;

- organization of international events (training, workshops, webinars, conferences) in the West Balkan countries for raising awareness about future trends in Big Data, Semantic Tools and Technologies, standards and applications (or adoption) in the industry;
- forecasting exercises about future trends of data services in Europe.

Organizational learning and capacity building approach

The capacity building approach [6] introduced into PUPIN with the LAMBDA project is based on the document *Strategic Capacity Building Plan*[1] that serves as a model and articulates how research institutions of low performing Member States and regions can reach the level of internationally-leading counterparts such as Fraunhofer Institute, University of Oxford and University of Bonn can support low performing Member States and regions can to reach the level of internationally-leading counterparts. Hence, in order to support the transfer of institutional knowledge and expertise to PUPIN staff, but also to other relevant stakeholders in the region, the *LAMBDA Learning and Consulting Platform* was established using the Drupal content management system (CMS). The platform facilitates collaboration including joined paper and deliverable writing, information sharing, and stakeholders' data-base management. The learning materials that were produced in the first project year are free, stored in the LAMBDA repository and are available online via the SlideWiki.org, an OpenCourseWare platform. The SlideWiki platform has been selected based on its collaborative features partially developed by LAMBDA researchers in SlideWiki framework.



LAMBDA Platform Customization

LAMBDA platform was configured for the needs of the future newcomers and professionals in the BDA domain. Currently, 3 different user roles have been defined:

- *Partner*, full access to the private pages of the portal.
- *Associated Partner*, full access to the Stakeholder database (restricted) and contents in the Knowledge Repository.
- *Administrator*, for managing the whole content management system.

The LAMBDA Knowledge repository (Learning and Consulting Platform) aims at facilitating the exchange of learning materials, tools, project results and best practice between the international leading organizations and research institutions and Industry from the West Balkan countries.

SlideWiki and LAMBDA platform integration

Initial activities related to providing access to the SlideWiki system on the LAMBDA platform were oriented towards analysing the most efficient approaches for the implementation of this feature on the LAMBDA CMS system. The following two approaches were considered and analysed:

- Embedding the SlideWiki within a Drupal page, on the LAMBDA platform, using an IFrame (Inline Frame). An Iframe represents a nested browsing context, embedding another HTML page into the current one.
- Creating a page in PHP, within the LAMBDA platform which would access the API of the SlideWiki micro-services, using the HTTPS secure protocol and display the results using the custom-made UI created in Drupal using PHP.

So far, the first approach has been fully implemented while we continued to work on the second approach. One very important feature of an OpenCourseWare system like SlideWiki is the ability to import and export data from/into different data formats [7]. The main data format used in SlideWiki is HTML. Instead of starting from scratch, SlideWiki users can import existing presentations, created using other applications and platforms, when creating new decks. Currently, SlideWiki supports direct import of content from .pptx files (mainly used by Microsoft PowerPoint) and .odp files (mainly used by LibreOffice/OpenOffice).

3. LAMBDA LECTURES

Systematizing the knowledge about the Big Data domain

Currently, main providers of lectures relevant for LAMBDA are the University of Bonn (UBO), The University of Oxford (UOXF), the German National Library for Science and Technology (TIB) and the Institute Mihajlo Pupin:

- UBO is working on the cutting edge technologies related to Big Data, Intelligent Analysis and Information Systems. The concerned team at the Smart Data Analytics (SDA) group is active in specializing applied research in intelligent data and knowledge analysis and teaching activities of the relevant topics.

[1] https://project-lambda.org/Action-Plan-2018-2020

- UOXF is working on cutting edge technologies related to Big Data and analytics. The concerned team at the VADA ("Value Added Data Systems") group is active in research and teaching activities with regard to these topics.
- TIB is working on development of cutting edge technologies semantic data processing, knowledge engineering and information systems in different domains.
- PUPIN is working on development of novel data analytics algorithms for different industrial domains including the energy sector[2].

Introduction to Knowledge Graphs

[EKGs-Lecture-1] This module introduces the topic of Knowledge Graphs, the similarities and differences between "world" Knowledge Graphs and Enterprise Knowledge Graphs, as well as theory and practice in the area. Knowledge Graph is [8]

- Fabric of concept, class, property, relationships, entity descriptions.
- Uses a knowledge representation formalism (RDF, OWL)
- Encompasses holistic knowledge (multi-domain, source, granularity):
  - **instance data** (ground truth), open (e.g. DBpedia, WikiData), private (e.g. supply chain data), closed data (product models),
  - derived, aggregated data,
  - **schema data** (vocabularies, ontologies)
  - **meta-data** (e.g. provenance, versioning, documentation licensing)
  - comprehensive **taxonomies** to categorize entities
  - **links** between internal and external data
  - **mappings** to data stored in other systems and databases.

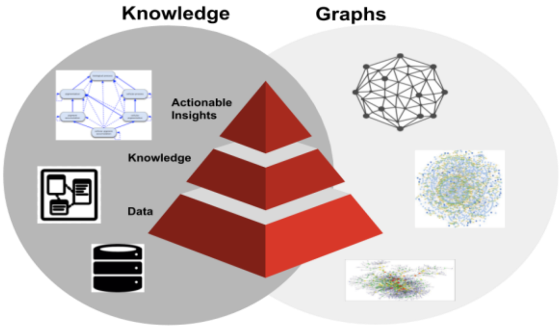


Image 1: Knowledge Graph (definition)[9]

Reasoning in Knowledge Graphs

[EKGs-Lecture-2] This module discusses reasoning in Knowledge Graphs. Reasoning is essential to gain value from Knowledge Graphs by deriving insights, and making available new implicit data from existing data. It covers the theory and practice of reasoning in Knowledge Graphs, and provides a number of easily accessible examples based on Oxford's Vadalog system [10].

Extraction for Knowledge Graphs

[EKGs-Lecture-3] This module discusses the topic of extraction for Knowledge Graphs. It focuses on web data extraction process that is essential for making the information available on the web accessible and usable by Knowledge Graphs.

Introduction to Big Data Architecture

[ARCH-Lecture-1] This lecture covers the existing advanced Big Data architectures following a bottom-up approach. In this lecture, the important knowledge to design and architect scalable solutions for challenging problems is introduced. The primary components in the architecture of such systems and their architectures is presented and discussed including "inter alia distributed kernels" and cluster managers, distributed file systems and storage systems.

Big Data Solutions in Practical Use-cases

[ARCH-Lecture-2] This lecture focuses on architecting Big Data solution. It discusses the role and importance of the components in realizing system architectures. The application of the introduced concepts and components is discussed in real-world example of practical use-cases.

Distributed Big Data Frameworks

[ARCH-Lecture-3] The "processing frameworks" are one of the most essential components of Big Data systems. There are three categories of such frameworks namely: Batch-only frameworks (Hadoop), Stream-only frameworks (Storm, Samza), and Hybrid frameworks (Spark, Hive and Flink). This lectures also one of the major Big Data frameworks, Apache Spark. It covers Spark fundamentals and the model of "Resilient Distributed Datasets (**RDDs**)" that are used in Spark to implement in-memory batch computation. Furthermore, essential parts of the important practical techniques are introduced such as Hadoop Distributed File System for the data resiliency, and the "lineage" property of "Directed Acyclic Graphs (DAG)" to achieve resilience for the computation resiliency, or use of catalyst for code optimization.

Distributed Big Data Libraries

[SBDA-Lecture-1] In the practical level, the Big Data frameworks use different APIs for graph computations and graph processing. In this lecture, the important libraries built on top of Apache Spark are covered. These include SparkSQL, GraphX and MLlib. The audience will learn to build scalable algorithms in Spark using Scala.

[2] https://project-lambda.org/Applying

Distributed Semantic Analytics I

[SBDA-Lecture-2] This module covers the needs and challenges of distributed analytics and then dive into the details of scalable semantic analytics stack (SANSA)[11] used to perform scalable analytics for knowledge graphs. It covers different SANSA layers and the underlying principles to achieve scalability for knowledge graph processing.

Distributed Semantic Analytics II

[SBDA-Lecture-3] This module covers the setup, APIs and different layers of SANSA. At the end of this module, the student is able to execute examples and create programs that use SANSA APIs. The final part of this lecture is planned to be an interactive session to wrap up the introduced concepts and present attendees some open research questions which are nowadays studied by the community.

4. EVALUATION

The possibilities for wider adoption of LAMBDA lectures was analysed with participants of the first LAMBDA Big Data Analytics School[3] that was organized in June 2019. Overall, more than 60 participants gathered at the PUPIN premises to exchange knowledge and expertise in Big Data technologies. The objective of the summer school was to give the PhD students and experts from Serbia and abroad and PUPIN researchers an opportunity to learn about the newest technologies and trends in this and related fields from respectable professors, as well as to hear about use cases from influential tech companies such as OntoText, SAS Institute, CISCO, Meltwater, and DeepReason.ai. The Feedback questionnaire was distributed to participants of the School and in-person interviews were conducted with selected participants. Twenty-tree (23) participants answered the questionnaire, for instance, parts of the answers are presented in Table 1. The feedback will be used to better plan the next-year summer school and to communicate the LAMBDA recommendations to relevant stakeholders. The *LAMBDA Foresight Exercise and Policy Recommendations* document will be prepared by June 2020.

5. CONCLUSION

The potential behind the exploitation of data (Open, Linked and Big) to boost economies and growth has been in the focus of many EU initiatives, the most recent of which is the Digital Single Market, which highlights the need to make sense of Big Data, since this is considered to be a fertile ground for innovation in both technology and development. Thus, the main topic of the European Union funded project LAMBDA (Learning, Applying, Multiplying Big Data Analytics) is Big Data Analytics and the semantics-based approach to processing data (Linked Data, Open Data, Big Data).

This paper presents some of the results achieved so far in the LAMBDA project, in particular, the ones related to the

- publishing of open educational resources (Big Data & Analytics lectures) via the LAMBDA portal and the SlideWiki OpenCourseWare platform;
- the organization of the Belgrade Big Data Analytics Summer School, for more than 60 participants from PUPIN, industry and other universities from Serbia and the West Balkan Region.

The exploitation of Big Data in various sectors has a potential socioeconomic impact far beyond the specific Big Data market. The foreseen activities in LAMBDA (open education, research-industry collaboration) will strengthen the digital skills of professionals and improve the technologies and services of the involved stakeholders (PUPIN and other stakeholders from West Balkan), thus contributing to national and regional sustainable development.

ACKNOWLEDGEMENT

The research presented in this paper is partly financed by the Ministry of Science and Technological Development of the Republic of Serbia (SOFIA project, Pr. No: TR-32010) and partly by the H2020 project “Learning, Applying, Multiplying Big Data Analytics” (GA No. 809965).

REFERENCES

[1] A. Chi Zhou and B. He, “Big Data and Exascale Computing,” in Sherif Sakr, Albert Y. Zomaya (eds) Encyclopedia of Big Data Technologies, Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-77525-8

[2] P. Zikopoulos and C. Eaton, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data,” 2011

[3] A. Patrizio, “IDC: Expect 175 zettabytes of data worldwide by 2025”, Network World, December 03, 2018, https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html

[4] F. G. Filip and E. Herrera-Viedma, Big Data in the European Union, The Bridge Vol. 44, No. 4, Winter 2014, https://www.nae.edu/Publications/Bridge/128772/129172.aspx

[5] European Commission, Building a European data economy, https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

[6] M. K. Nammous, C. Lange, V. Janev, LAMBDA Deliverable 2.4 Strategic Capacity Development Plan, 2018.

[7] K. Junghanns, D. Paunovic, A. Third, D2.4 SlideWiki Import/Export Module, 2016.

[8] S. Auer, “Towards Knowledge Graph based Representation, Augmentation and Exploration of Scholarly Communications,” Keynote at the 1st

[3] https://project-lambda.org/BDASummer-School-2019

LAMBDA Big Data Analytics Summer School (18-20.06.2019, Belgrade, Serbia

[9] M. E. Vidal, E. K.M., S. Jozashoori., G. Palma, "A Knowledge-Driven Pipeline for Transforming Big Data into Actionable Knowledge," In: Auer S., Vidal ME. (eds) Data Integration in the Life Sciences. DILS 2018. Lecture Notes in Computer Science, vol 11371. Springer, Cham

[10] Luigi Bellomarini, Emanuel Sallinger, Georg Gottlob (2018) "The Vadalog system: datalog-based reasoning for knowledge graphs," in Proceedings of the VLDB Endowment -Proceedings of the 44th International Conference on Very Large Data Bases (VLDB), Rio de Janeiro, Brazil, Volume 11 Issue 9, May 2018, Pages 975-987.

[11] J. Lehmann, et all "Distributed Semantic Analytics using the SANSA Stack," in Proceedings of 16th International Semantic Web Conference - Resources Track (ISWC'2017), 2017.

Table 1: Feedback collected during the first Big Data Analytics Summer School

Organization	Feedback collected
PhD student, University of Belgrade	*I would like to see the lectures presented at the School integrated in one of the master courses organized at the School of Electrical Engineering, for instance 'Data Mining and Semantic Web', 'Natural Language Processing' and other.*
PhD student, University of Niš	*The courses organized at the Electronic Engineering Faculty (AI, Knowledge Discovery, Data Warehouse) are related to the lectures presented at the School. What was interesting in the last 3 days were the Use Cases presented.*
Expert from public administration	*For someone with my interests and duties, this was in every sense an extraordinary experience (equally, the first, more general day, and the second day of a more technical nature). Probably all other participants, like me, are impatiently waiting for news from the consortium.*
Expert from Bank sector from Serbia	*I think that mentioned projects and Data analytics tools will be more useful to colleagues interested in the academic perspective to Data analytics and Big data in general, than to us who work in industry, but I believe that the information I get about them will be very important in the future.*
Professor from Romania	*The LAMBDA lectures are good candidates to be reused as support material within the curricula of Big Data master program: Introduction to Big Data and Architectures, Distributed Big Data Frameworks, Big Data Solutions in Practice. The courses that may benefit are: Data Warehouses and Big Data Technologies.*
Professor from Germany	*I would like to see more interactions between the teachers and participants.*