LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

Horizon 2020 Grant Agreement No 809965
Contract start date: July 1st 2018, Duration: 30 months

LAMBDA Deliverable 3.3
Semantic Big Data Architectures

Due date of deliverable: 31/12/2019
Actual submission date: 25/12/2019

Revision: Version 1.0

Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme, H2020-WIDESPREAD-2016-2017 Spreading Excellence and Widening Participation under grant agreement No 809965.

Author(s)	Hajira Jabeen (UBO), Damien Graux (IAIS), Dejan Paunović (PUPIN)
Contributor(s)	Jens Lehmann (UBO)
Internal Reviewer(s)	Valentina Janev (PUPIN)
Approval Date	
Remarks	

Workpackage	WP 3 Cooperation for Teacher and Student Training
Responsible for WP	Institute for Computer Science - University of Bonn
Deliverable Lead	University of Bonn (Hajira Jabeen)
Related Tasks	Task 3.2 'Train the Trainer' Lectures (UBO)

Document History and Contributions

Version	Date	Author(s)	Description
0.1	10.06.2019	Hajira Jabeen	Lectures
0.2	10.06.2019	Damien Graux	Lectures
0.3	14.06.2019	Jens Lehmann	Lectures - review
0.4	02.12.2019	Dejan Paunović	Toolbox
0.5	16.12.2019	Hajira Jabeen	Deliverable
0.6	23.12.2019	Valentina Janev	Internal Review

Executive Summary

One of the main objectives of LAMBDA (Learning, Applying, Multiplying Big Data Analytics) project is to develop a series of Big Data Analysis training session and foster knowledge exchange in this area. This deliverable reports the project period from month 1 till month 18. In this period we have held commenced the first LAMBDA summer school which has provided lectures on cutting edge technologies following different learning modes e.g., webinars, lectures, hands-on sessions and demonstrations. The new training material has been developed following the requirements from the consortium partners and attendees. The prepared material has been made available on different portals as OpenCourseWare (OCW) including the SlideWiki.org .platform. This report covers one of the three modules of the “Train the Trainer tasks” namely Semantic Big Data Architectures. In this report, we summarise the summer school and provide a brief overview of the topics covered in the first LAMBDA school. In addition, we provide a sketch of the planned activities from the coming period.

Table of Contents

Abbreviations and Acronyms

BD	Big Data
BDA	Big Data Analytics
HDFS	Network of experts Hadoop Distributed File System
NoSQL	Not only SQL
OCW	OpenCourseWare
PPT	Power Point File format/extension (Microsoft)
SQL	Structured Query Language
WP	Work Package

List of Figures

List of Tables

1. Introduction

Learning is a vital component of project LAMBDA (**L**earning, **A**pplying, **M**ultiplying **B**ig **D**ata **A**nalytics). This deliverable covers a set of lectures on "Semantic Big Data Architectures" as part of the online materials that have been produced for the 1st Belgrade Big Data Analytics Summer School 2019. The event was held at the PUPIN premises (part of the University of Belgrade) from 17-20 of June 2019. The LAMBDA portal provides more details about the organization website of the summer school[1], LAMBDA lectures[2] (see **Figure 1**) and topics discussed at the school.

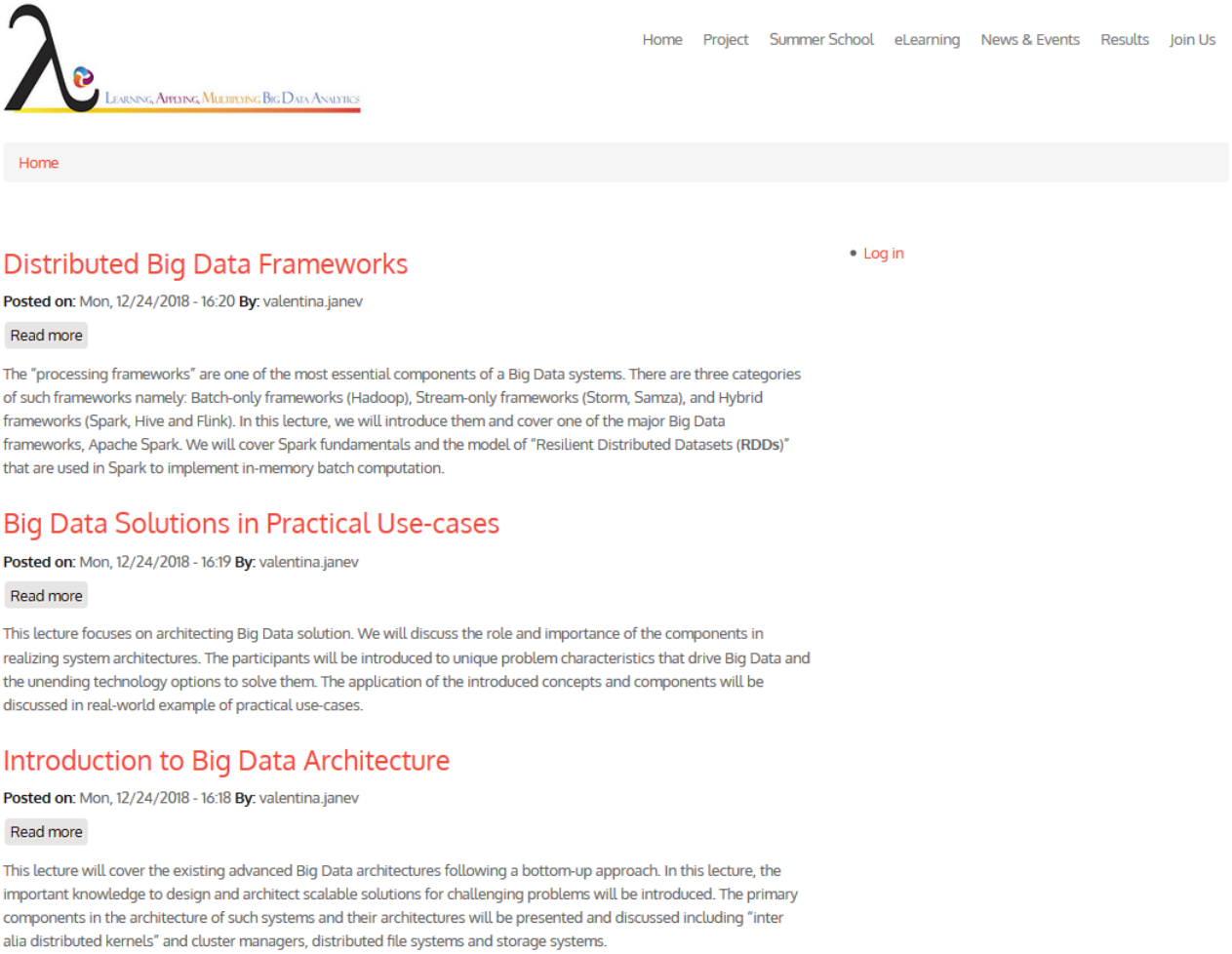

Figure 1. LAMBDA lectures on "Semantic Big Data Architectures"

The material (related to module 2 "Semantic Big Data Architectures") considered the latest advancements in the field. It was produced following international teaching standards and practices. The contents are aligned with the overall project objectives and demonstrate the scientific excellence and innovation capacity of the involved project partners.

Deliverable 3.3 Semantic Big Data Architectures reports about three lectures with the following titles

- Introduction big data & architectures[3] (55 slides)

[1] https://project-lambda.org/Summer-School-2019
[2] https://project-lambda.org/Knowledge-repository/Lectures
[3] https://project-lambda.org/ARCH-Lecture-1 , https://slidewiki.org/deck/128513/lecture-4/deck/128513

- Big data solutions in practical use-cases[4] (62 slides)
- Distributed big data frameworks[5] (56 slides)

The course material contains lecture slides that were also uploaded to the SlideWiki portal as well as hands-on tutorial sessions to execute the taught concepts. The material has also been uploaded to LAMBDA slide deck at SlideWiki[6].

1.1 Relation to Other Deliverables

Deliverable 3.3 Semantic BD Architectures (M18) is the second of three reports associated with Task 3.2 'Train the Trainer' Lectures. The set of LAMBDA lectures contain:

1. **Deliverable 3.2 Enterprise Knowledge Graphs**[7]:
2. **Deliverable 3.3 Semantic BD Architecture:**
3. **Deliverable 3.4 Smart Data Analytics**:

The summary of the LAMBDA school and all the lectures have already been covered in D3.2. Additionally, this deliverable is related to Deliverable 2.1[8] where the results of the first survey of the Big Data market were reported in M06 of the LAMBDA project.

Therefore, in this Deliverable, we focus on the above three lectures only.

1.2 Structure of the Deliverable

This deliverable presents a short description of the three lectures presented in this module (Section 2), as well as an introduction to the LAMBDA toolbox for experimentation with Big Data tools (Section 3).

[4] https://project-lambda.org/ARCH-Lecture-2 , https://slidewiki.org/deck/128515/lecture-6/deck/128515
[5] https://project-lambda.org/ARCH-Lecture-3 , https://slidewiki.org/deck/128514/lecture-5/deck/128514
[6] https://slidewiki.org/deck/128440-1/big-data-analytics-lectures
[7] https://project-lambda.org/sites/default/files/Deliverables/D3.2.pdf
[8] https://project-lambda.org/D2.1

2. Semantic BD Architectures (Lectures)

The main purpose of this module is to elaborate and provide a comprehensive overview of big data and AI landscape as shown in **Figure 2**.

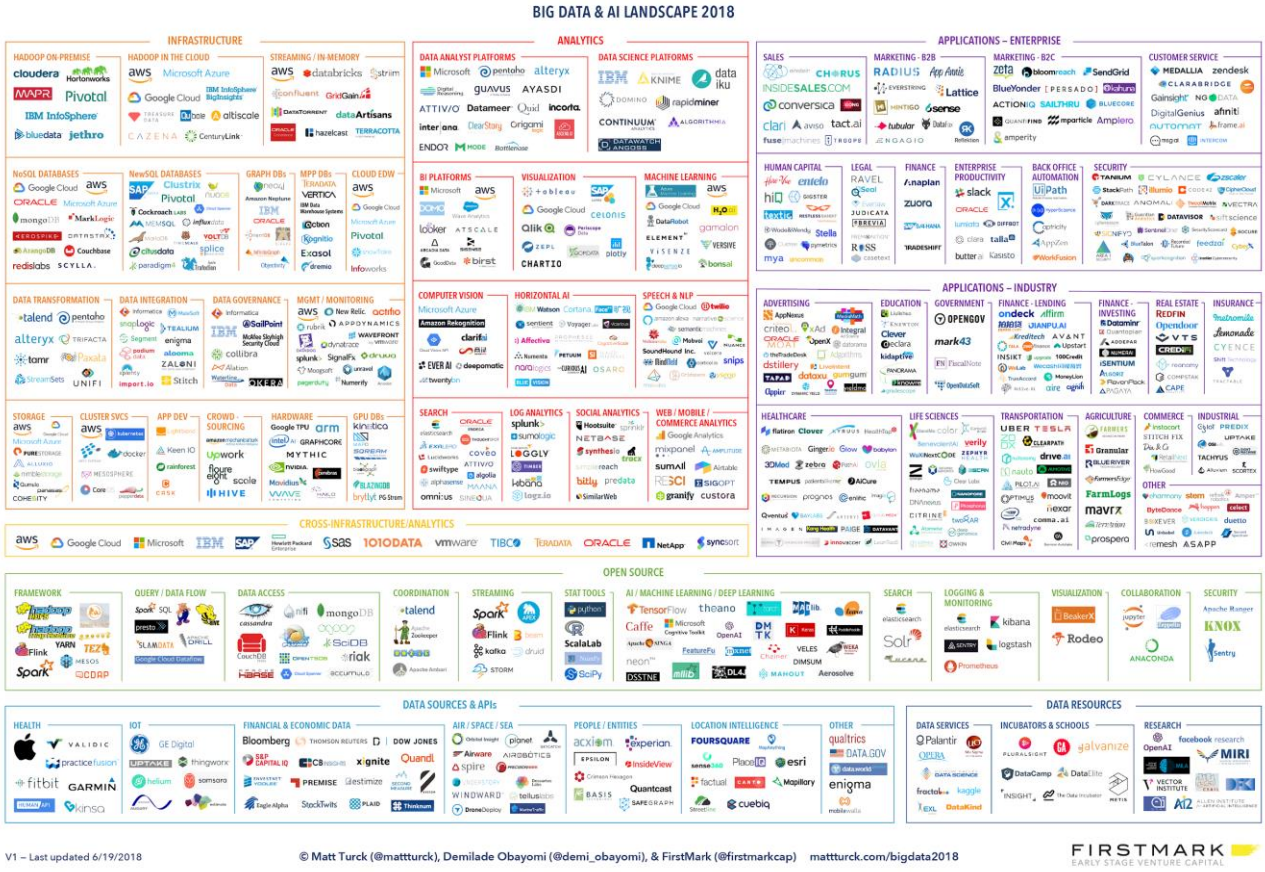

Figure 2. Big Data and AI landscape 2018

2.1 Introduction Big Data & Architectures [9]

This lecture covered in detail, the big data architecture components based on their functionalities and covered examples in each of the categories. **Figure 3** shows the types of big data architectures being used in practice. During the LAMBDA Summer School 2019, we also discussed the pros and cons of each architectural design methodology and talked about the real-life challenges and use-cases.

Depending upon the type of data and individual requirements of the organizations, the big data architectures are divided into three main types as below:

[9] https://project-lambda.org/ARCH-Lecture-1

Lambda Architecture:
The lambda architecture, first proposed by Nathan, addresses the issue of slow query results on batch data, and real-time data requiring fast querying. It combines the real-time results with the queries from batch analysis of older data.

Kappa Architecture: The kappa architecture was proposed by Jay Kreps as an alternative to the lambda architecture. Same as Lambda architecture, all data in Kappa architecture flows through a single path, however, it uses a stream processing system only. The kappa architecture focuses only on data stream processing, real-time processing, or processing of live discrete events.

Microservices based architecture : "Microservice Architecture" has sprung up over the last few years to describe a particular way of designing software applications as suites of independently deployable services.

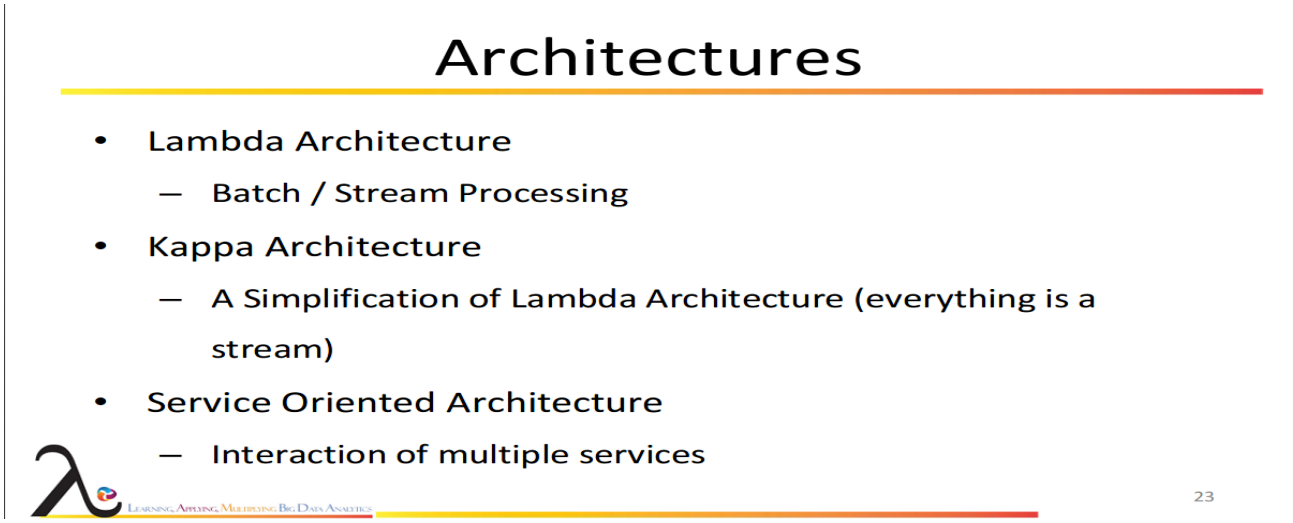

Figure 3. Three main types of Big Data Architectures

In addition, we covered the distributed Kernels and explained in detail their working and usage scenarios, see **Figure 4.**

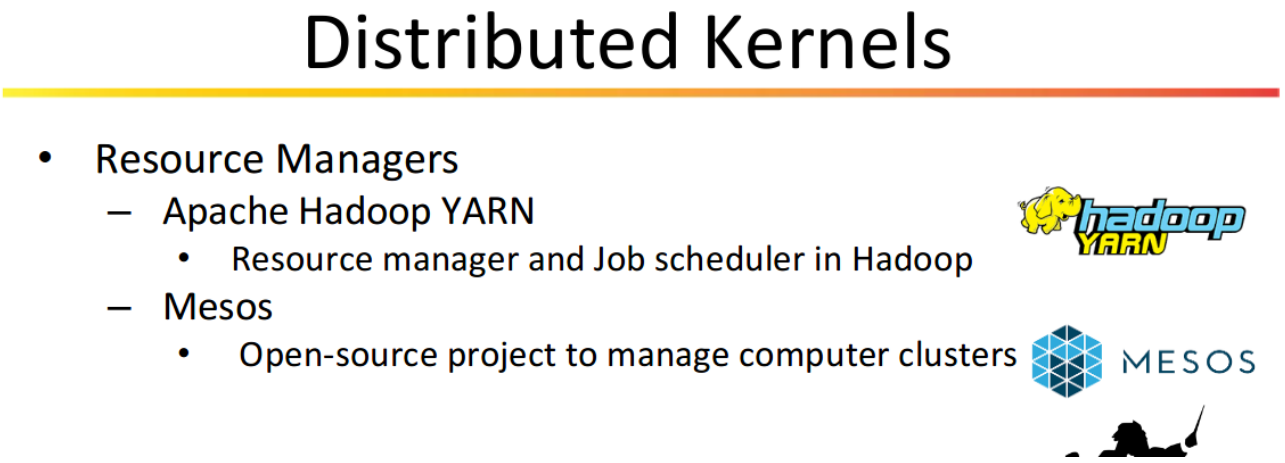

Figure 4. Distributed Kernels

Furthermore, we covered distributed file systems including the Hadoop Distributed File System (HDFS). Given that the HDFS is the defacto standard for big data storage, we covered in detail, its working model, storage model, architecture and discussed the design decisions that have led to a stable fault-tolerant system. The lecture covered name node, data node, replication factor, working of reads and writes on request.

2.2 Big Data Solutions in Practical Use-cases[10]

This session covers the basics of cloud architecture, their need, benefits and associated issues. We discussed the key enabling technologies, formal definitions **(Figure 5)**, different service models and deployment models in the cloud.

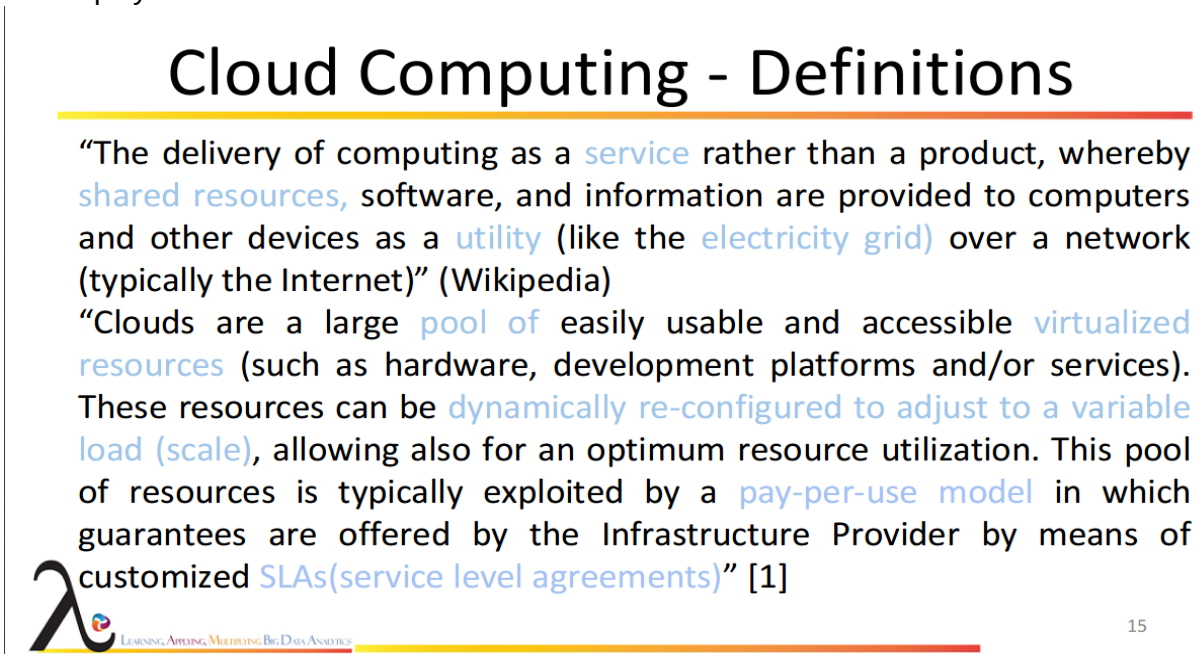

Figure 5. Cloud Computing (definitions)

After the discussion on the design concepts, we demonstrated the design of a cloud architecture platform. We talked about its design goals and how these goals were achieved using the technologies covered in the previous sections. This lecture was followed by a hands-on session which covered the installation and execution of the demonstrated platform.

[10] https://project-lambda.org/ARCH-Lecture-2

2.3 Distributed Big Data Frameworks[11]

This lecture covered different frameworks for data storage, NoSQL stores including Key-value stores, document databases, wide column stores, Graph Databases, Storage systems like MongoDB, Hive, Cassandra, Kafka for message passing Elastic search for indexing and for visualizing; Kibana. The lecture also covered analytical frameworks like Map-reduce, its drawbacks, followed by one of the general-purpose processing engine named Apache Flink.

The overall structure of the lecture is shown in **Figure 6.**

Big Data Ecosystem

File system	HDFS, NFS
Resource manager	Mesos, Yarn
Coordination	Zookeeper
Data Acquisition	Apache Flume, Apache Sqoop
Data Stores	MongoDB, Cassandra, Hbase, Hive
Data Processing	
• Frameworks	Hadoop MapReduce, Apache Spark, Apache Storm, Apache Flink
• Tools	Apache Pig, Apache Hive
• Libraries	SparkR, Apache Mahout, MlLib, etc
Data Integration	
• Message Passing • Managing data heterogeneity	Apache Kafka SemaGrow, Strabon
Operational Frameworks	
• Monitoring	Apache Ambari

Figure 6. Big Data Ecosystem (example)

These three lectures were delivered in close collaboration of UBO and Fraunhofer IAIS.
Figure 7 shows Dr. Hajira Jabeen from the University of Bonn and Dr. Damien Graux from Fraunhofer (at the time of tutorial) delivering their talks in close collaboration.

Figure 7. Dr. Hajira Jabeen (UBO) and Dr. Damien Graux (IAIS) giving lectures at PUPIN

[11] https://project-lambda.org/ARCH-Lecture-3

3. LAMBDA Big Data Toolbox

3.1 Categorization of Big Data Tools

Deliverable 2.1 entitled "Big Data Challenges and Analysis of Scientific and Technological Landscape" gives an overview of the Big Data concepts, outlines some of the relevant challenges in this domain and reviews and describes the current state of the art tools relevant to Big Data applications. For the purpose of the analysis, the "Big Data" landscape was divided into six segments as presented in the Table below (left side). Later on, as part of the **Task 4.3 LAMBDA Learning and Consulting Tools at PUPIN,** the tools analysed and tested were divided into ten groups (see right side of the Table below). Additional category was added for Cloud Marketplaces (7 software providers)

Table 1. Categorization of Big Data tools

Deliverable 2.1	**LAMBDA Toolbox**[12]	**# pages**
Big Data Frameworks	Hadoop as a Web Service / Platform	5
	Operational Database Management Systems	4
NoSQL Platforms and Knowledge Graphs	NoSQL/ Graph databases	10
Stream Processing Data Engines	Stream Processing Engines	7
Big Data Pre-processing	Library / API for Big Data	9
	ML Library / API for Big Data	5
Big Data Analytics	Analytics Software / System / Platform	9
	Data Analytics Languages	7
	Optimization Library for Big Data	2
Big Data Visualization Tools	Visualization Software / System	10

3.2 Toolbox Statistics

One of the objectives of Task 4.3 is focused on the integration of diverse open-source tools into a single environment (BDA Learning and Consulting platform, see **Figure 8**) for learning Big Data related algorithms, methods, tools and prototypes with the help of visiting scholars from the linked institutions. Based on the training materials provided in Task 3.2 and the survey of Big Data tools completed in Task 2.1 Science and Technology watch, the PUPIN team created a browsing environment for newcomers in the field. More than 70 tools have been studied and some of them already adopted for experimentation.

[12] https://project-lambda.org/tools-for-experimentation

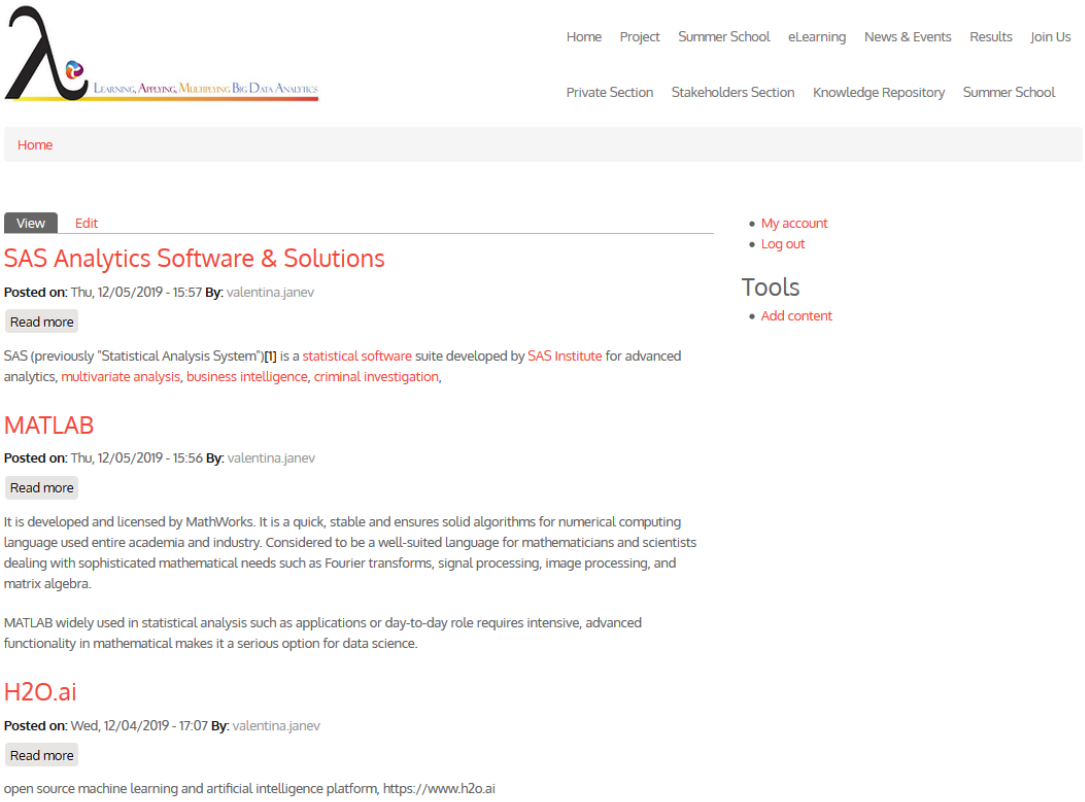

Figure 8. Links to Analytics Software / System / Platform via the LAMBDA portal

4. Conclusion

This deliverable provides a summary of the work carried out in the WP3 Task 2 (Module 2). The material for the summer school was prepared, presented and made available to the school attendees and associates. The consortium has put their best effort to incorporate the latest cutting edge technologies in the field of Big Data and Architectures into the delivered lectures. Furthermore, the lectures were augmented with interesting discussions on design issues, industrial use-cases, demonstrations, and applications. It is anticipated that the lectures have provided a meaningful background and head start to start with the covered big data technologies. In the next summer school, we plan to extend these lectures and cover additional topics.