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

LAMBDA Deliverable 3.4
Learning Material (Smart Data Analytics)

Due date of deliverable: 30/06/2020
Actual submission date: 30/06/2020

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)	Heba Mohamed (UBO), Valentina Janev (PUPIN)
Contributor(s)	All authors of LAMBDA lectures
Internal Reviewer(s)	Emanuel Sallinger (UOXF)
Approval Date	
Remarks	

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

Document History and Contributions

Version	Date	Author(s)	Description
0.1	1.06.2929	Heba Mohamed (UBO),	First draft
0.2	26.06.2020	Valentina Janev (PUPIN)	Update
0.3	30.06.2020	Emanuel Sallinger (UOXF)	Review
0.4			
0.5			

Executive Summary

One of the specific objectives of LAMBDA (**L**earning, **A**pplying, **M**ultiplying **B**ig **D**ata **A**nalytics) project is developing learning materials. In the last two years project partners developed a series of learning material which is made available at the LAMBDA portal, see https://project-lambda.org/Knowledge-repository/Lectures.

This deliverable provides a summary of the work carried out in the WP3 Task 3.2 'Train the Trainer' Lectures. It gives an overview of the different types of learning materials, such as the book - *Knowledge Graphs and Big Data Processing, LNCS 12072, 2020*, and learning lectures (38 lectures, 5 from invited external experts) that have been created for Big Data Analytics summer schools and other events. The lectures are divided into eight main modules: Artificial Intelligence (4), Survey (3), Foundations (3), Enterprise Knowledge Graphs (5, 1 external), Semantic Big Data Architectures (5), Big Data and Knowledge Graphs Tools (4), Smart Data Analytics (5), and Case Studies (9, 4 externals).

Table of Contents

Abbreviations and Acronyms

BDA	Big Data Analytics
NoE	Network of experts
OERs	Open Educational Resources
RDF	Resource Description Framework
WP	Work Package

List of Figures

List of Tables

1. Introduction

Learning is considered as one of the LAMBDA's (**L**earning, **A**pplying, **M**ultiplying **B**ig **D**ata **A**nalytics, http://www.project-lambda.org/) major building blocks. The main objective of **WP3 Open Education: Cooperation for Teacher and PhD Student Training** is

- to establish a highly effective learning environment that supports different learning modes (classroom learning, action-oriented learning, virtual training, work-based learning);
- to improve the existing training materials and their publication through the SlideWiki.org portal;
- to develop a series of Big Data Analytics training for graduate / PhD students and professionals on topics relevant for Big Data research, taking into consideration the 5Vs: Volume, Velocity, Variety, Veracity, and Value.

Aligned with the project's overall goals of stimulating scientific excellence and capacity for innovation, the consortium's initial consideration was to provide high quality and latest Big Data material. Consequently, a set of learning materials has been created for both Belgrade Big Data Analytics Summer Schools, 2019 and 2020, and other events. The LAMBDA website, https://project-lambda.org/Summer-Schools, provides more details about the organization and lectures presented at the summer schools, while the learning materials can be found on https://project-lambda.org/Knowledge-repository/Lectures.

1.1 Relation to other Deliverables

Deliverable 3.4 Smart Data Analytics is related to:

- Deliverable 3.1 The 'Trainers' Network' Infrastructure that describes the adoptions made on the LAMBDA platform (see https://project-lambda.org/) in order to facilitate teachers-trainees cooperation.
- Deliverable 3.2 Enterprise Knowledge Graphs: lectures that include formal conceptual frameworks for designing and maintaining knowledge graphs; such as strategies for the semi-automatic construction of such graphs from the combination of proprietary enterprise data and relevant public domain knowledge; opportunities and implications in terms of performance and access control.
- Deliverable 3.3 Semantic BD Architecture: lectures that include approaches for better supporting the variety dimension of Big Data comprising RDF, RDF-Schema and OWL knowledge representation formalisms, mapping standards such as R2RML, JSON-LD and CSVW, the SPARQL query language, etc. Integrating semantic and Big Data technologies can help to make Big Data architectures and applications more flexible, adaptive and their implementation more efficient.

1.2 Structure of the Deliverable

In order to categorize the lectures according to topics, a taxonomy was used that is presented in Section 2. Each lecture is described with a set of metadata that improves the search of the learning items. Lectures are accompanied with teaching material.

Section 3 gives an overview of all lectures registered in the LAMBDA Repository by June 2020.

2. Overview of Lectures

2.1 Categorization of Lectures

The lectures have been categorized into eight modules:

1. Artificial Intelligence (4 lectures),
2. Survey (3 lectures),
3. Foundations (3 lectures),
4. Enterprise Knowledge Graphs (4 lectures),
5. Semantic Big Data Architectures (7 lectures),
6. Big Data and Knowledge Graphs Tools (4 lecture),
7. Smart Data Analytics (5 lectures), and
8. Case Studies (5 lectures).

2.2 Searching the Repository

There are 2 possibilities to search the Lectures:

- by Module, see Figure 1
- by Event, see Figure 2.

Figure 1. Searching the lecture by Module

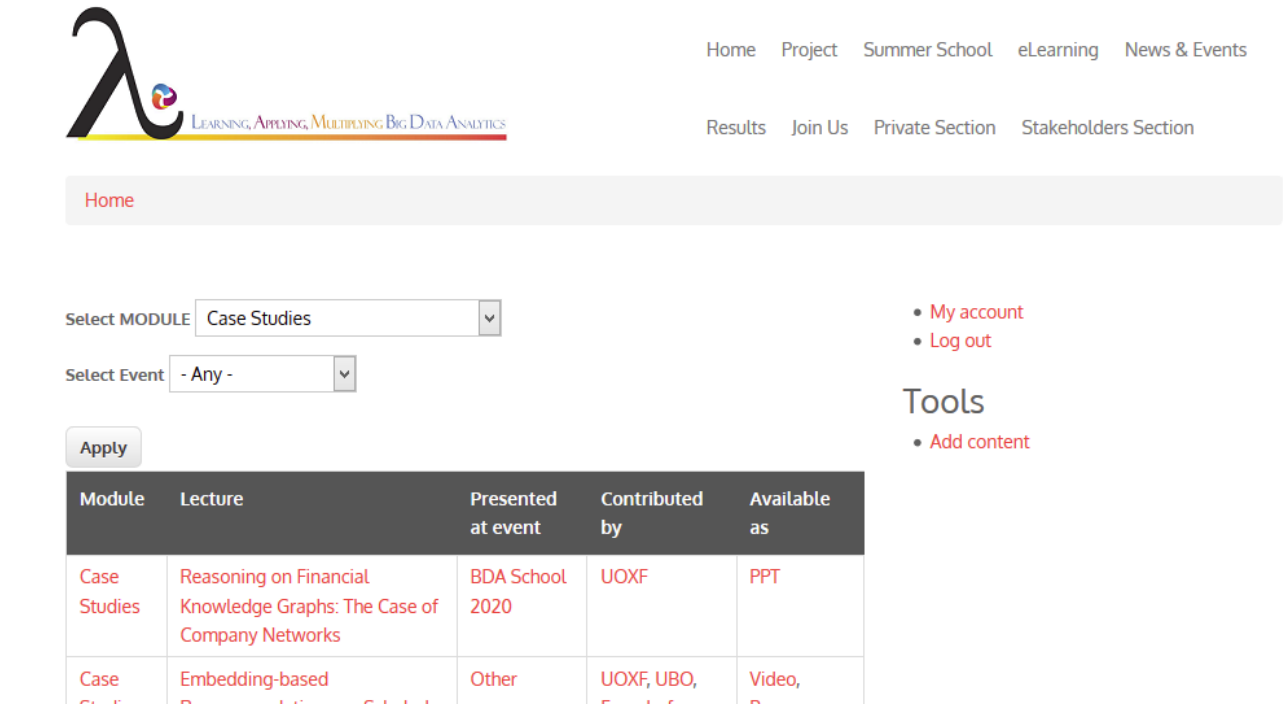
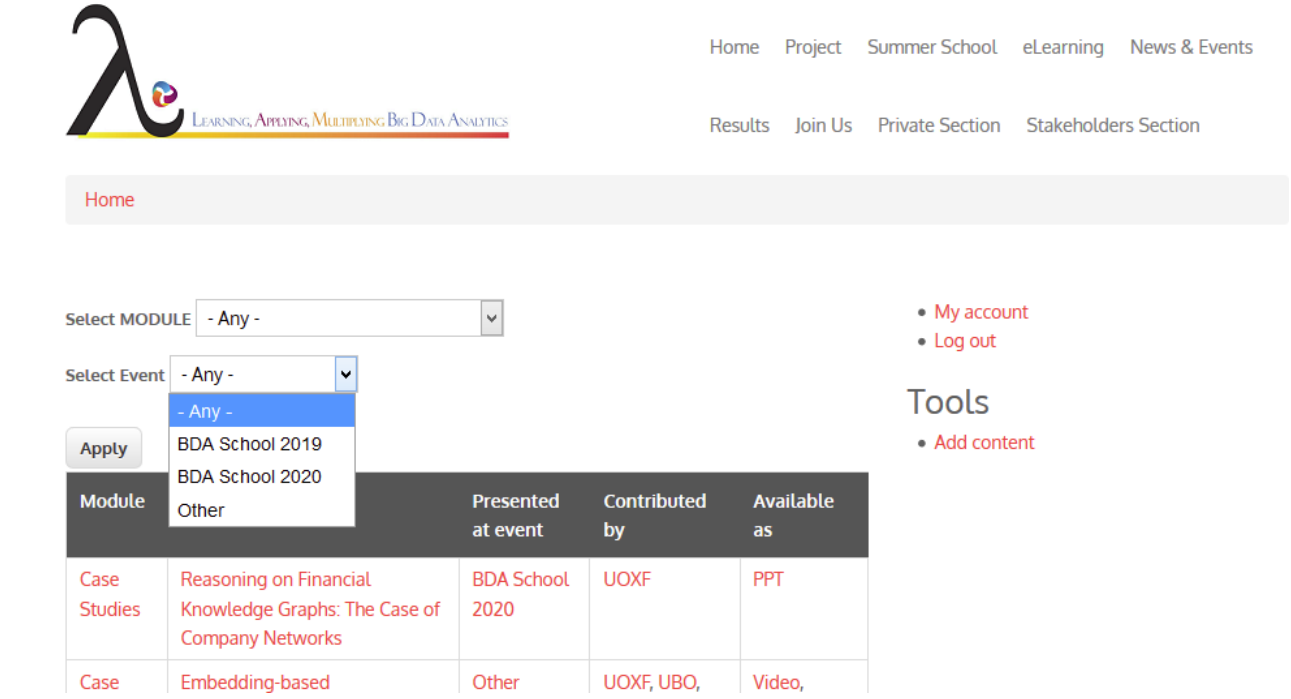


Figure 2. Searching the lectures by Event



2.3 Learning Material

The Book - **Knowledge Graphs and Big Data Processing, LNCS 12072, 2020, https://www.springer.com/gp/book/9783030531980**

Foundations

Chapter 1 Ecosystem of Big Data (Valentina Janev)

Chapter 2 Knowledge Graphs: The Layered Perspective (Luigi Bellomarini, Emanuel Sallinger, and Sahar Vahdati)

Chapter 3 Big Data Outlook, Tools, and Architectures (Hajira Jabeen)

Architecture

Chapter 4 Creation of Knowledge Graphs (Anastasia Dimou) - Invited

Chapter 5 Federated Query Processing (Kemele M. Endris, Maria-Esther Vidal, and Damien Graux)

Chapter 6 Reasoning in Knowledge Graphs: An Embeddings Spotlight (Luigi Bellomarini, Emanuel Sallinger, and Sahar Vahdati)

Methods and Solutions

Chapter 7 Scalable Knowledge Graph Processing using SANSA (Hajira Jabeen, Damien Graux, and Gezim Sejdiu)

Chapter 8 Context-Based Entity Matching for Big Data (Mayesha Tasnim, Diego Collarana, Damien Graux, and Maria-Esther Vidal)

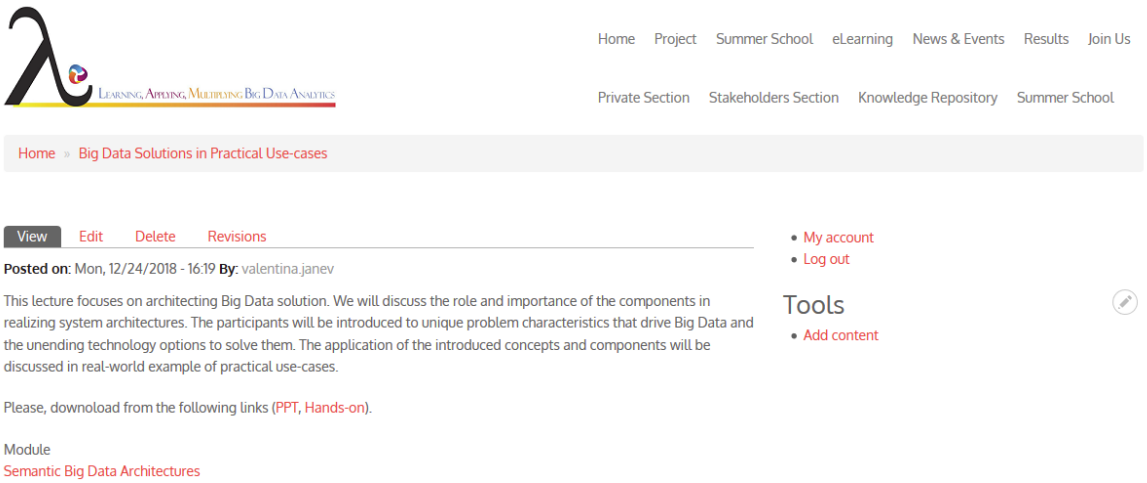
Applications

Chapter 9 Survey on Big Data Applications (Valentina Janev, Dea Puji´c, Marko Jeli´c, and Maria-Esther Vidal)

Chapter 10 Case Study from the Energy Domain (Dea Puji´c, Marko Jeli´c, Nikola Toma˘sevi´c, and Marko Bati´c)

The teaching materials are also mentioned on the Web page of the Lecture, see Figure 3.

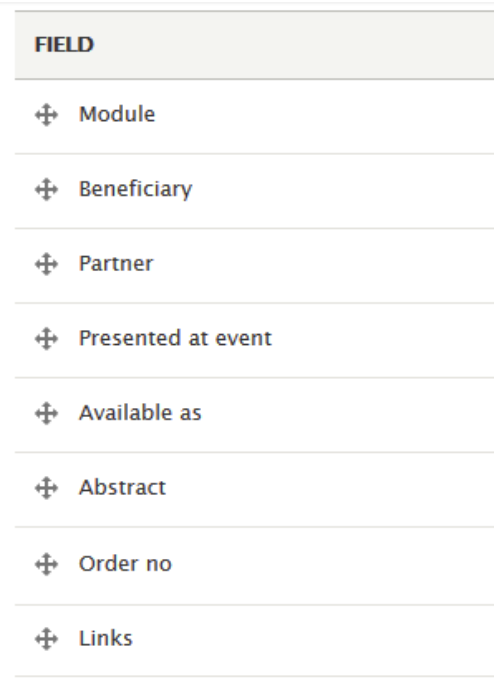
Figure 3. Abstract of one teaching material



2.4 Registering a new Lecture with the LAMBDA Platform

Authorized users can register a new Lecture in the Repository by entering the information about the Lecture, as is presented in Figure 4:

Figure 4. Registering a new lecture within LAMBDA platform



3. Description of Lectures (KPI = 38 Lectures, 5 from external partners)

University Level

Table 1. List of Lectures and its availability

#	Lecture	Institute	Contributed By	Availability
Artificial Intelligence				
1	Data for AI: Foresight	Fraunhofer	Simon Scerri	Video, PPT
2	AI and Knowledge Graphs	UOXF	Emanuel Sallinger	PPT
3	Conversational AI	Fraunhofer	Jens Lehmann	Video
4	The Revolution of AI	Fraunhofer	Stefan Wrobel	Video
Survey				
5	Survey on Big Data Tools	PUPIN	Marko Jelić, Dea Pujić	Paper, PPT
6	Overview and Comparison of Machine Learning Algorithms	PUPIN	Dea Pujić, Marko Jelić	Paper
7	Survey on Big Data Applications	PUPIN	Valentina Janev	Chapter-Book
Foundations				
8	Big Data Ecosystem	PUPIN	Valentina Janev	Chapter-Book
9	Introduction to Knowledge Graphs	UOXF	Emanuel Sallinger, Luigi Bellomarini, Sahar Vahdati	Chapter-Book
10	Big Data Outlook, Tools, and Architectures	UBO	Hajira Jabeen	Chapter-Book
Enterprise Knowledge Graphs				
11	What is Knowledge Graph?	Fraunhofer	Mikhail Galkin	Video
12	Introduction to Knowledge Graphs	UOXF	Emanuel Sallinger, Luigi Bellomarini, Sahar Vahdati	Chapter-Book
13	Creation of Knowledge Graphs			Chapter-Book
14	Extraction for Knowledge Graphs	UOXF	Tim Furche	Paper
15	Swift Logic for Big Data and Knowledge Graphs	UOXF	Georg Gottlob	Paper, Video

Semantic Big Data Architectures				
16	Reasoning in Knowledge Graphs	UOXF	Georg Gottlob	Video
17	Introduction to Big Data Architecture	Fraunhofer, UBO	Damien Graux, Hajira Jabeen	PPT
18	Big Data Solutions in Practical Use-cases	Fraunhofer, UBO	Damien Graux, Hajira Jabeen	PPT
19	Distributed Big Data Frameworks	UBO, Fraunhofer	Hajira Jabeen, Damien Graux	PPT
20	Data Lakes and Federated Query Processing	Fraunhofer	Damien Graux, Hajira Jabeen	Chapter-Book
Smart Data Analytics				
20	Distributed Big Data Libraries	UBO, Fraunhofer	Hajira Jabeen, Damien Graux	PPT
21	Distributed Semantic Analytics I	UBO, Fraunhofer	Damien Graux, Hajira Jabeen	Chapter-Book
22	Distributed Semantic Analytics II	UBO, Fraunhofer	Damien Graux, Hajira Jabeen	PPT, Other
23	SANSA - Scalable Semantic Analytics Stack	UBO	Hajira Jabeen, Damien Graux	Paper, Other
24	Scalable Knowledge Graph Processing using SANSA	UBO, Fraunhofer	Hajira Jabeen, Damien Graux	Chapter-Book
Big Data and KGs Tools				
25	Context-Based Entity Matching for Big Data	Fraunhofer	Diego Colarana	Chapter-Book
26	Vadalog System	UOXF	Emanuel Sallinger	Paper
27	Data Science with Spark and Hadoop	UBO	Hajira Jabeen	Video
28	Spark using Scala	UBO	Hajira Jabeen	Video
Case Studies				
29	Semantic Information Infrastructures from Business Information Delivery to Water Management		Mariana Damova	Video
30	Soft computing for Transparent synthesis of Geo Big Data		Gloria Bordogna	Video
31	Chronorobotics - Spatio-temporal models for		Tom Krajnik	Video

	social and service robots			
32	IntelliSys: Intelligent System for Road Safety		Debasis Das	Video
33	Reasoning on Financial Knowledge Graphs: The Case of Company Networks	UOXF	Luigi Bellomarini	PPT
34	Embedding-based Recommendations on Scholarly Knowledge Graphs	UOXF, UBO, Fraunhofer	Sahar Vahdati	Video, Paper
35	Open and Big Data – Utilization Perspective	PUPIN	Valentina Janev	PPT
36	Data Analytics for Energy Sector	PUPIN	Dea Pujić Marko Jelić	Paper, Chapter-Book
37	Predictive Analytics in Renewable Energy Systems	PUPIN	Dea Pujić Marko Jelić	

3.1 Artificial Intelligence (4)

Module	Lecture	Presented at event	Contributed by	Available as
Artificial intelligence	Data for AI: Foresight	BDA School 2020	Fraunhofer	Video, PPT
Artificial intelligence	AI and Knowledge Graphs	BDA School 2020	UOXF	PPT
Artificial intelligence	Conversational AI	Other	Fraunhofer	Video
Artificial intelligence	The Revolution of AI	Other	Fraunhofer	Video

Data for AI: Foresight: This lecture was delivered at the Big Data Analytics Summer School 2020 by Dr. Simon Scerri, Fraunhofer IAIS. Simon Scerri talked about the importance of data, the vision and actions from some industrial organizations point of view. The lecture lists some examples of technical solutions that have been suggested from secure and trusted data sharing.

AI and Knowledge Graphs: Knowledge Graphs (KGs) are one of the key trends among the next wave of technologies. Many definitions exist of what a Knowledge Graph is, and in this chapter, we are going to take the position that precisely in the multitude of definitions lies one of the strengths of the area. We will choose a particular perspective, which we will call the layered perspective, and three views on Knowledge Graphs. The importance of supporting implicit knowledge becomes central for KGs as well, especially when they are a component of an Enterprise AI application, to the point that intensional knowledge should be considered part of the KG itself.

Conversational AI: Prof. Jens Lehmann talked about Speech-to-Text, Question Answering via Knowledge Graphs and Text-to-Speech AI systems, and demonstrates the SPEAKER voice assistant platform, which is based on technologies of the Fraunhofer IAIS and IIS institutes. The SPEAKER project was awarded at the AI innovation contest of the BMWi. This lecture was held at last year's FUTURAS IN RES conference with the motto "What's the IQ of AI?".

The Revolution of AI: Prof. Stefan Wrobel talked about the major milestones in the history of intelligent systems and presents projects at Fraunhofer IAIS that could shape our future with AI. Besides technical research objectives, Prof. Wrobel proposes economic and ethical solutions such as the International Data Space, which aims at a global market standard for a sovereign use of data, or the AI certification project "the Bonn catalogue" for a credible and reliable AI. This lecture was held at last year's FUTURAS IN RES conference with the motto "What's the IQ of AI?".

3.2 Survey (3)

Module	Lecture	Presented at event	Contributed by	Available as
Survey	Survey on Big Data Tools	Other	PUPIN	Paper, PPT
Survey	Overview and Comparison of Machine Learning Algorithms	Other	PUPIN	Paper
Survey	Survey on Big Data Applications	BDA School 2020	PUPIN	Chapter-Book

Survey on Big Data Tools: This introductory lecture discusses the Big Data processing pipeline and the Big Data Landscape from the following perspectives: Big Data Frameworks, NoSQL Platforms and Knowledge Graphs, Stream Processing Data Engines, Big Data Preprocessing, Big Data Analytics, and Big Data Visualization Tools.

Overview and Comparison of Machine Learning Algorithms: Big Data Analytics is a crucial component of the Big data paradigm and refers to the process of extracting useful knowledge from large datasets or streams of data. Due to enormity, high dimensionality, heterogeneous, and distributed nature of data, traditional techniques of data mining may be unsuitable to work with big data. In this lecture, different Big data tools and machine learning algorithms are introduced, discussed and analyzed. Depending on the main learning algorithm, the machine learning algorithms can be categorized as supervised, unsupervised and reinforcement learning.

Survey on Big Data Applications: The goal of this chapter is to shed light on different types of big data applications needed in various industries including healthcare, transportation, energy, banking and insurance, digital media and e-commerce, environment, safety and security, telecommunications, and manufacturing. In response to the problems of analyzing large-scale data, different tools, techniques, and technologies have been developed and are available for experimentation. In our analysis, we focused on literature (review articles) accessible via the Elsevier ScienceDirect service and the SpringerLink service from more recent years, mainly from the last two decades. For the selected industries, this lecture also discusses challenges that can be addressed and overcome using the semantic processing approaches and knowledge reasoning approaches discussed in this book.

3.3 Foundations (3)

Module	Lecture	Presented at event	Contributed by	Available as
Foundations	Big Data Ecosystem	Other	PUPIN	Chapter-Book
Foundations	Introduction to Knowledge Graphs	BDA School 2019	UOXF	Chapter-Book
Foundations	Big Data Outlook, Tools, and Architectures	BDA School 2019	UBO	Chapter-Book

Big Data Ecosystem: The rapid development of digital technologies, IoT products and connectivity platforms, social networking applications, video, audio and geolocation services has created opportunities for collecting/accumulating a large amount of data. While in the past corporations used to deal with static, centrally stored data collected from various sources, with the birth of the web and cloud services, cloud computing is rapidly overtaking the traditional in-house system as a reliable, scalable and cost-effective IT solution. The high volumes of structures and unstructured data, stored in a distributed manner, and the wide variety of data sources pose problems related to data/knowledge representation and integration, data querying, business analysis and knowledge discovery. This introductory lecture serves to characterize the relevant aspects of the Big Data Ecosystem with respect to big data characteristics, the components needed for implementing end-to-end big data processing and the need for using semantics for improving the data management, integration, processing, and analytical tasks.

Introduction to Knowledge Graphs: Knowledge Graphs (KGs) are one of the key trends among the next wave of technologies. Many definitions exist of what a Knowledge Graph is, and in this chapter, we are going to take the position that precisely in the multitude of definitions lies one of the strengths of the area. We will choose a particular perspective, which we will call the layered perspective and three views on Knowledge Graphs: KGs as Knowledge Representation Tools, KGs as Knowledge Management Systems, and KGs as Knowledge Application Services.

Big Data Outlook, Tools, and Architectures: Big data is a reality and it is being generated and handled in almost all digitised scenarios. This chapter covers the history of Big data and discusses prominent related terminologies. The significant technologies including architectures and tools are reviewed. Finally, the lecture reviews big knowledge graphs that attempt to address the challenges (e.g. heterogeneity, interoperability, variety) of big data through their specialised representation format. This chapter aims to provide an overview of the existing terms and technologies related to big data. After reading this lecture, the reader can develop an understanding of the broad spectrum of big data ranging from important terms, challenges, used technologies, and their connection with large scale knowledge graphs.

3.4 Enterprise Knowledge Graphs (5, 1 external)

Module	Lecture	Presented at event	Contributed by	Available as
Enterprise Knowledge Graphs	What is Knowledge Graph ?	Other	Fraunhofer	Video
Enterprise Knowledge Graphs	Creation of Knowledge Graphs	BDA School 2020		Chapter-Book
Enterprise Knowledge Graphs	Extraction for Knowledge Graphs	BDA School 2019	UOXF	Paper
Enterprise Knowledge Graphs	Swift Logic for Big Data and Knowledge Graphs	Other	UOXF	Paper, Video
Enterprise Knowledge Graphs	Reasoning in Knowledge Graphs: An Embeddings Spotlight	BDA School 2020	UOXF	Chapter-Book, Video

What is Knowledge Graph?: Knowledge Graphs (KGs) are being used in a variety of applications including web search, answering questions, and for data integration. Knowledge graphs also target output for Natural Language Processing (NLP) and computer vision algorithms, and Machine Learning (ML) algorithms more generally. Knowledge graphs are a topic of a major program from Natural Science Foundation (NSF). This lecture summarized what is the KGs, how to create it, how to use it with modern artificial intelligence algorithms.

Creation of Knowledge Graphs: This Lecture introduces how Knowledge Graphs are generated. The goal is to gain: 1) an overview of different approaches that were proposed for creating a Knowledge Graph and find out more details about the current prevalent ones; 2) an understanding of the different solutions to generate Knowledge Graphs; 3) knowledge to choose the mapping language that suits best a certain use case. After reading this lecture, the reader should have an understanding of the different solutions available to generate Knowledge Graphs and should be able to choose the mapping language that best suits a certain use case.

Extraction for Knowledge Graphs: This lecture discusses the topic of extraction for Knowledge Graphs. We focus on web data extraction in this module. Web data extraction is essential to make information available on the web-accessible and usable by Knowledge Graphs. We provide a thorough introduction to the topic. This features both Oxford's Vadalog and OXPath systems.

Swift Logic for Big Data and Knowledge Graphs: Many modern companies wish to maintain knowledge in the form of a corporate knowledge graph and to use and manage this knowledge via a knowledge graph management system (KGMS). We formulate various requirements for a fully-fledged KGMS. In particular, such a system must be capable of performing complex reasoning tasks but, at the same time, achieve efficient and scalable reasoning over Big Data with an acceptable computational complexity. Moreover, a KGMS needs interfaces to corporate databases, the web, and machine-learning and analytics packages.

Reasoning in Knowledge Graphs: An Embeddings Spotlight: In this lecture, we introduce the aspect of reasoning in Knowledge Graphs. We give a broad overview focusing on the multitude of reasoning techniques: spanning logic-based reasoning, embedding-based reasoning, neural network-based reasoning, etc. In particular, we discuss three dimensions of reasoning in Knowledge Graphs. Complementing these dimensions, we will structure our exploration based on

a pragmatic view of reasoning tasks and families of reasoning tasks: reasoning for knowledge integration, knowledge discovery and application services.

3.5 Semantic Big Data Architectures (5)

Module	Lecture	Presented at event	Contributed by	Available as
Semantic Big Data Architectures	Reasoning in Knowledge Graphs	Other	UOXF	Video
Semantic Big Data Architectures	Introduction to Big Data Architecture	BDA School 2019	Fraunhofer, UBO	PPT
Semantic Big Data Architectures	Big Data Solutions in Practical Use-cases	BDA School 2019	Fraunhofer, UBO	PPT
Semantic Big Data Architectures	Distributed Big Data Frameworks	BDA School 2019	UBO, Fraunhofer	PPT
Semantic Big Data Architectures	Data Lakes and Federated Query Processing	BDA School 2020	Fraunhofer	Chapter-Book, Video, PPT

Reasoning in Knowledge Graphs: This lecture discusses reasoning in Knowledge Graphs. The reasoning is essential to gain value from Knowledge Graphs by deriving insights and making available new implicit data from existing data. We will cover the theory and practice of reasoning in Knowledge Graphs, and provide a number of easily accessible examples based on Oxford's Vadalog system.

Introduction to Big Data Architecture: 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 are presented and discussed including "inter alia distributed kernels" and cluster managers, distributed file systems and storage systems.

Big Data Solutions in Practical Use-cases: This lecture focuses on architecting Big Data solutions. We discussed the role and importance of the components in realizing system architectures. The participants are introduced to unique problem characteristics that drive Big Data and the unending technology options to solve them. The application of the introduced concepts and components are discussed in a real-world example of practical use-cases.

Distributed Big Data Frameworks: 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). In this lecture, we introduced them and covered one of the major Big Data frameworks, Apache Spark. We covered 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 the catalyst for code optimization.

Data Lakes and Federated Query Processing: Big data plays a relevant role in promoting both manufacturing and scientific development through industrial digitization and emerging

interdisciplinary research. Semantic web technologies have also experienced great progress, and scientific communities and practitioners have contributed to the problem of big data management with ontological models, controlled vocabularies, linked datasets, data models, query languages, as well as tools for transforming big data into knowledge from which decisions can be made. Despite the significant impact of big data and semantic web technologies, we are entering into a new era where domains like genomics are projected to grow very rapidly in the next decade. In this next era, integrating big data demands novel and scalable tools for enabling not only big data ingestion and curation but also efficient large-scale exploration and discovery. Federated query processing techniques provide a solution to scale up to large volumes of data distributed across multiple data sources. Federated query processing techniques resort to source descriptions to identify relevant data sources for a query, as well as to find efficient execution plans that minimize the total execution time of a query and maximize the completeness of the answers. This lecture summarizes the main characteristics of a federated query engine, reviews the current state of the field, and outlines the problems that still remain open and represent grand challenges for the area.

3.6 Big Data and Knowledge Graphs Processing Tools (4)

Module	Lecture	Presented at event	Contributed by	Available as
Big Data and KGs Tools	Context-Based Entity Matching for Big Data	BDA School 2020	Fraunhofer	Chapter-Book
Big Data and KGs Tools	Vadalog System	Other	UOXF	Paper
Big Data and KGs Tools	Data Science with Spark and Hadoop	BDA School 2019	UBO	Video
Big Data and KGs Tools	Spark using Scala	BDA School 2019	UBO	Video

Context-Based Entity Matching for Big Data: In the Big Data era, where variety is the most dominant dimension, the RDF data model enables the creation and integration of actionable knowledge from heterogeneous data sources. However, the RDF data model allows for describing entities under various contexts, e.g., people can be described from its demographic context, but as well from their professional contexts. Context-aware description poses challenges during entity matching of RDF datasets; the match might not be valid in every context. To perform a contextually relevant entity matching, the specific context under which a data-driven task, e.g., data integration is performed, must be taken into account. However, existing approaches only consider inter-schema and properties mapping of different data sources and prevent users from selecting contexts and conditions during a data integration process. We devise COMET, an entity matching technique that relies on both the knowledge stated in RDF vocabularies and a context-based similarity metric to map contextually equivalent RDF graphs. COMET follows a two-fold approach to solve the problem of entity matching in RDF graphs in a context-aware manner.

In the first step, COMET computes the similarity measures across RDF entities and resorts to the Formal Concept Analysis algorithm to map contextually equivalent RDF entities. Finally, COMET combines the results of the first step and executes a 1-1 perfect matching algorithm for matching RDF entities based on the combined scores. We empirically evaluate the performance of COMET

on testbed from DBpedia. The experimental results suggest that COMET accurately matches equivalent RDF graphs in a context-dependent manner.

Vadalog System: Over the past years, there has been a resurgence of Datalog-based systems in the database community as well as in industry. In this context, it has been recognized that to handle the complex knowledge-based scenarios encountered today, such as reasoning over large knowledge graphs, Datalog has to be extended with features such as existential quantification. Yet, Datalog-based reasoning in the presence of existential quantification is in general undecidable. Many e↵orts have been made to define decidable fragments. Warded Datalog+/- is a very promising one, as it captures PTIME complexity while allowing ontological reasoning. Yet so far, no implementation of Warded Datalog+/- was available. In this paper, we present the Vadalog system, a Datalog-based system for performing complex logic reasoning tasks, such as those required in advanced knowledge graphs. The Vadalog system is Oxford's contribution to the VADA research programme, a joint effort of the universities of Oxford, Manchester and Edinburgh and around 20 industrial partners. As the main contribution of this paper, we illustrate the first implementation of Warded Datalog+/-, a high-performanceDatalog+/- system utilizing an aggressive termination control strategy. We also provide a comprehensive experimental evaluation.

Data Science with Spark and Hadoop: This lecture briefly introduces the use of Apache Spark and Hadoop for Data Science applications.

Spark using Scala: This lecture introduces Apache Spark (Architecture, Libraries), the underlying data structures (Resilient Distributed Dataset) and an Example with Scala.

3.7 Smart Data Analytics (5)

Module	Lecture	Presented at event	Contributed by	Available as
Smart Data Analytics	Distributed Big Data Libraries	BDA School 2019	UBO, Fraunhofer	PPT
Smart Data Analytics	Distributed Semantic Analytics I	BDA School 2019	UBO, Fraunhofer	PPT
Smart Data Analytics	Distributed Semantic Analytics II	BDA School 2019	UBO, Fraunhofer	PPT, Other
Smart Data Analytics	SANSA - Scalable Semantic Analytics Stack	BDA School 2019	UBO	Paper, Other
Smart Data Analytics	Scalable Knowledge Graph Processing using SANSA	BDA School 2020	UBO, Fraunhofer	Chapter-Book

Distributed Big Data Libraries: 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.

Distributed Semantic Analytics I: This module will cover the needs and challenges of distributed analytics and then dive into the details of scalable semantic analytics stack (SANSA) used to

perform scalable analytics for knowledge graphs. It will cover different SANSA layers and the underlying principles to achieve scalability for knowledge graph processing.

Distributed Semantic Analytics II: This module will cover the setup, APIs and different layers of SANSA. At the end of this module, the audience will be 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.

SANSA - Scalable Semantic Analytics Stack: The size of knowledge graphs has reached the scale where centralised analytical approaches have become infeasible. Recent technological progress has enabled powerful distributed in-memory analytics that has been shown to work well on simple data structures. However, the application of such distributed analytics approaches to semantic knowledge graphs lags significantly behind. To advance both scalability and accuracy of large-scale knowledge graph analytics to a new level, foundational research on methods leveraging distributed in-memory computing and semantic technologies in combination with advancements in analytics approaches is indispensable.

This Lecture introduces SANSA Scalable Semantic Analytics Stack that provides support for: 1) efficient data distribution and semantics-aware computation of latent resource embeddings for knowledge graphs; 2) adaptive distributed querying; 3) efficient self-optimising inference execution plans; and 4) efficient distributed machine learning on semantic knowledge graphs of extremely large scale.

Scalable Knowledge Graph Processing using SANSA: The size and number of knowledge graphs have increased tremendously in recent years. In the meantime, the distributed data processing technologies have also advanced to deal with big data and large scale knowledge graphs. This lecture introduces Scalable Semantic Analytics Stack (SANSA), which addresses the challenge of dealing with large scale RDF data and provides a uni ed framework for applications like link prediction, knowledge base completion, querying, and reasoning. We discuss the motivation, background and architecture of SANSA. SANSA is built using general-purpose processing engines Apache Spark and Apache Flink. After reading this chapter, the reader should have an understanding of the different layers and corresponding APIs available to handle Knowledge Graphs at scale using SANSA.

3.8 Case Studies (9, 4 externals)

Module	Lecture	Presented at event	Contributed by	Available as
Case Studies	Semantic information infrastructures from business information delivery to water management	BDA School 2020		Video
Case Studies	Soft computing for Transparent synthesis of Geo Big Data	BDA School 2020		Video, PPT
Case Studies	Chronorobotics - Spatio-temporal models for social and service robots	BDA School 2020		Video
Case Studies	IntelliSys: Intelligent System for Road Safety	BDA School 2020		Video
Case Studies	Reasoning on Financial Knowledge Graphs: The Case of Company Networks	BDA School 2020	UOXF	PPT
Case Studies	Embedding-based Recommendations on Scholarly Knowledge Graphs	Other	UOXF, UBO, Fraunhofer	Video, Paper
Case Studies	Open and Big Data – Utilization Perspective	Other	PUPIN	PPT
Case Studies	Data Analytics for Energy Sector	BDA School 2019	PUPIN	Paper, Chapter-Book
Case Studies	Predictive Analytics in Renewable Energy Systems	Other	PUPIN	Paper

Semantic information infrastructures from business information delivery to water management: This is a Keynote Lecture delivered at Big data Analytics Summer School 2020 by Dr. Mariana Damova, Mozajka. The lecture introduces semantic information infrastructures where discusses the data indenced services. The lecture shows a variety of examples where these infrastructures can provide competitive advantages to the organizations that have adapted them.

Reasoning on Financial Knowledge Graphs: The Case of Company Network: The initial release of KGs was started on an industry scale by Google and further continued with the publication of other large-scale KGs such as Facebook, Microsoft, Amazon, DBpedia, Wikidata and many more. As an influence of the increasing hype in KG and advanced AI-based services, every individual company or organization is adapting to KG. The KG technology has immediately reached industry, and big companies have started to build their own graphs such as the industrial Knowledge Graph at Siemens. In a joint work for sharing ideas from large-scale industrial Knowledge Graphs, namely Microsoft, Google, Facebook, eBay and IMB, authors stated a broad range of challenges ahead of research and industry involving KGs. Despite the content-wise difference and similarities of those Knowledge Graphs, the discussions involve data acquisition and provenance problems due to source heterogeneity and scalability of the underlying management system. In this lecture we introduce the Enterprise Knowledge Graph of Italian companies for the Central Bank of Italy.

Embedding-based Recommendations on Scholarly Knowledge Graphs: The increasing availability of scholarly metadata in the form of Knowledge Graphs (KG) offers opportunities for

studying the structure of scholarly communication and the evolution of science. Such KGs build the foundation for knowledge-driven tasks e.g., link discovery, prediction and entity classification which allows providing recommendation services. Knowledge graph embedding (KGE) models have been investigated for such knowledge-driven tasks in different application domains. One of the applications of KGE models is to provide link predictions, which can also be viewed as a foundation for recommendation service, e.g. high confidence “co-author” links in a scholarly knowledge graph can be seen as suggested collaborations. In this paper, KGEs are reconciled with a specific loss function (Soft Margin) and examined with respect to their performance for co-authorship link prediction task on scholarly KGs.The results show a significant improvement in the accuracy of the experimented KGE models on the considered scholarly KGs using this specific loss. TransE with Soft Margin (TransE-SM) obtains a score of 79.5%Hits@10 for co-authorship link prediction tasks while the original TransEobtains 77.2%, on the same task. In terms of accuracy and Hits@10, TransE-SM also outperforms other state-of-the-art embedding models such as ComplEx, ConvE and RotatE in this setting. The predicted co-authorship links have been validated by evaluating the profile of scholars.

Open and Big Data – Utilization Perspective: Although each government in Europe with their public administration services can be treated as a big data ecosystem, the opportunities of interconnecting, integrating and processing the data on EU level presents a real challenge nowadays. Discussions on the public benefit of integrating and opening the data can be found in our previous work, where we examine the use of Linked Data Approach in European e-Government Systems. In the European data strategy, EU foresees that by 2025, 80% of the processing and analysis of data that currently takes place in data centres and centralised computing facilities will be processed in smart connected objects, such as cars, home appliances or manufacturing robots, and in computing facilities close to the user (‘edge computing’).

In this set of lectures, we discuss the potential and challenges of implementing the European data strategy in the West Balkan region, bearing in mind the needs and opportunities of SMEs and NGOs. The lectures were presented at the Open Data event organized in cooperation with the Chamber of Economy and the Ministry of Public Administration of Montenegro.

Data Analytics for Energy Sector: Big Data technologies are often used in domains where data is generated, stored and processes with rates that cannot be efficiently processed by one computer. One of those domains is definitely the domain of energy. Here, the processes of energy generation, transmission, distribution and use have to be concurrently monitored and analyzed in order to assure system stability without brownouts or blackouts. The transmission systems (grids) that transport electric energy are in general very large and robust infrastructures that are accompanied by an abundance of monitoring equipment. Novel Internet of Things (IoT) concepts of smart and interconnected homes are also pushing both sensors and actuators into people’s homes. The power supply of any country is considered to be one the most critical systems and as such its stability is of utmost importance. To that effect, a wide variety of systems are deployed for monitoring and control. Some of these tools are presented in this Lecture with a few from the perspective of end-users (Non-Intrusive Load Monitoring, Energy Conservation Measures and User Benchmarking) and a few from the perspective of the grid (production, demand and price forecasting).

Predictive Analytics in Renewable Energy Systems: With the aim of improving ecological interest, the share of renewable energy sources (RES) in energy production has to be increased. Nonetheless, that growth adversely influences the grid’s instability, as a result of the dependency between the RES production and weather conditions. Therefore, in order to provide a stable energy system, it is necessary to plan the consumption in advance with respect to the availability of RES production. This lecture is focused on comparing current SoA approaches for two different renewable energy sources, photovoltaic panels and solar thermal collectors.

4. Conclusion

This deliverable provides a summary of the work carried out in the WP3 Task 3. We provided the list of learning materials introduced by LAMBDA. First we gave an overview of the different types of learning materials, such as the book - *Knowledge Graphs and Big Data Processing, LNCS 12072, 2020.*, learning lectures (35 lectures) have been created for Big Data Analytics summer schools and other events. Afterwards, list the book chapters. Finally, gave a detailed overview of the 36 lectures (5 from external experts). The lectures are divided into eight main modules: Artificial Intelligence (4), Survey (3), Foundations (3), Enterprise Knowledge Graphs (5, 1 external), Semantic Big Data Architectures (5), Big Data and Knowledge Graphs Tools (4), Smart Data Analytics (5), and Case Studies (9, 4 externals).