12-10-2017

Teaching Big Data Management – An Active Learning Approach for Higher Education

Barbara Dinter
Chair of Information Systems, Chemnitz University of Technology, Chemnitz, Germany., barbara.dinter@wirtschaft.tu-chemnitz.de

Tobias Jaekel
Chemnitz University of Technology, tobias.jaekel@wirtschaft.tu-chemnitz.de

Christoph Kollwitz
Chemnitz University of Technology, christoph.kollwitz@wirtschaft.tu-chemnitz.de

Hendrik Wache
Chemnitz University of Technology, hendrik.wache@wirtschaft.tu-chemnitz.de

Teaching Big Data Management – An Active Learning Approach for Higher Education

Curriculum

Barbara Dinter
Chemnitz University of Technology
barbara.dinter@wirtschaft.tu-chemnitz.de

Tobias Jäkel
Chemnitz University of Technology
tobias.jaekel@wirtschaft.tu-chemnitz.de

Christoph Kollwitz
Chemnitz University of Technology
christoph.kollwitz@wirtschaft.tu-chemnitz.de

Hendrik Wache
Chemnitz University of Technology
hendrik.wache@wirtschaft.tu-chemnitz.de

Abstract

Since big data analytics has become an imperative for business success in the digital economy, universities face the challenge to train data scientists and data engineers on various technological and managerial skills. In addition to traditional lectures, active learning formats ensure a practice oriented education enabling students to handle novel big data technologies. In this paper, we present a big data management syllabus for master students in the field of big data analytics, which includes various hands-on and action learning elements. The course encompasses seven lectures and nine tutorials and takes place at Chemnitz University of Technology. It covers a broad range of big data applications and facilitates knowledge on various cognitive levels. The paper gives an overview of the course content and assigns learning objectives to lectures and tutorials using Krathwohl's revised taxonomy. Finally, we present the feedback, which we have received by the students over the years.

Keywords

Big data, analytics, higher education, syllabus, curriculum, learning objectives

Introduction

Big data analytics (BDA) is considered as a foundation for value creation and business success in a rapidly changing and digital economy (Manyika et al. 2011). As pointed out by Accenture (2014) a vast majority of companies expects that big data will dramatically change the way of doing business in the future. Consequently, the demand for experts in data analytics (so called data scientists) as well as for data architects and data managers grows worldwide and across all industries (Davenport and Patil 2012; Provost and Fawcett 2013). The required job profiles for big data professionals are complex and include competencies, such as in statistics, computer science and software development as well as in business administration and project management (Debortoli et al. 2014). Furthermore, the big data tool landscape is rapidly evolving and new technologies are continuously emerging making the selection of long-term valuable technologies challenging.

Universities have recognized the trend towards big data and have started to align academic education accordingly. Despite the development of information systems (IS) curricula and big data related adaptations, there is still a need for innovative teaching methods for big data. The paper at hand presents a comprehensive syllabus for Big Data Management (BDM). The syllabus was developed in the context of the master program "Business Intelligence & Analytics" at Chemnitz University of Technology, which was established in 2014. In terms of content the course focuses on providing up-to-date knowledge about BDA and related technologies. From a pedagogical perspective we use different teaching formats, such as hands-on exercises, and build on established theoretical concepts like the Revised Blooms Taxonomy (Krathwohl 2002) or active learning (Prince 2004).

The paper is intended to serve as inspiration and blue print for a practice oriented as well as scientific-grounded design of future syllabi in the field of BDA. It is organized as follows: First, we give an overview of the underpinning technical and pedagogical foundations, which we have used for the syllabus design. Then we explain the overall course setting and the role of our big data lab for the course. The content of the course as well as its classification regarding the educational and technical learning objectives constitute the main part of the paper. Our experiences with the course are summarized in the discussion. Finally, the paper is briefly concluded.

Foundations

To ensure a clear understanding of the terminologies and concepts used in this paper and to explain the pedagogical background, we introduce the terms big data and BDA as well as a framework for learning objectives. Additionally, we briefly present the state of the art of teaching BDA.

Big Data and Big Data Analytics

The term big data reflects two drivers of today's phenomena, the exponential growth and ever-increasing amount of data as well as the progressive digitization of the society, industry, and government (Manyika et al. 2011). Usually, big data is characterized by several "V"s. Depending on the source, three to five "V"s are utilized to describe the term big data (Chen et al. 2012; Kwon and Sim 2013; Manyika et al. 2011; Puget 2013). In this paper we rely on a five "V" definition.

The first "V" addresses the volume of today's data (Kwon and Sim 2013). As the world is digitizing, more digital data is generated with the result that we have already entered the zettabyte age, which goes beyond the human intuition. Velocity, the second "V", characterizes the continuous production of digital data and its increasing production and analysis speed (Kwon and Sim 2013). This is mainly driven by social networks, machine generated data and Internet of Things (IoT) technologies (Chen et al. 2014). The resulting data is not only structured, but also unstructured, semi-structured or event-based and systems have to deal with the whole range of data structure types – which is characterized by the third V, variety (Kwon and Sim 2013). The fourth "V" refers to the veracity of data (Puget 2013). Big data may exhibit a varying and unknown quality (Lukoianova and Rubin 2014) and always involves uncertainty and ambiguity resulting from inaccurate sensors, errors, and incompleteness. Finally, the fifth "V" is value (Chen et al. 2014) with respect to the analytics and usage of big data. The value of data is often constituted by the discovery of new insights, e.g. in form of relationships and correlations between data points. Therefore, analytics can increase the value of data and transform big data into smart data.

Following this understanding, BDA describes methods and technologies to analyze huge amounts of continuously produced data with varying structure and quality. This goes beyond traditional business intelligence (BI) methods that usually aim for structured data produced in an organizational context and for rather pre-defined queries (Russom 2011). Consequently, BI is mainly characterized by descriptive analytics while BDA focuses on predictive and prescriptive analytics. Chen et al. (2012) presented a framework to categorize emerging research topics in the field of analytics. They distinguish between big data, text, web, network, and mobile analytics. For example, BDA comprises machine learning or process mining while web analytics includes social media mining or cloud computing (Chen et al. 2012). Analytics also utilizes text mining algorithms to analyze unstructured data, network analytics to investigate relationships, or stream processing technologies for surveillance, alerting and predictive analytics.

As fast as the data volume and speed increase, new technologies and frameworks emerge to cope with the posed challenges. For instance, researchers introduced novel database models to store big data volumes in a decentral way (BigTable by Chang et al. 2008) and models to process compute-intensive algorithms by splitting and spreading the problem across several nodes for parallel computation (MapReduce by Dean and Ghemawat 2008). As memory capacity grows and becomes cheaper, data processing shifts from disk-based to in-memory-based technologies (Boncz et al. 1999). For example, the Hadoop ecosystem combines several technologies to store, process, analyze and visualize big data as well as to manage the underlying infrastructure (Landset et al. 2015).

Teaching Big Data Analytics

Previous literature about higher education – particularly in the field of IS – has addressed the rising phenomenon of big data by developing novel curricula and competency models (Cegielski and Jones-Farmer 2016; Topi et al. 2017; Wixom et al. 2014). In addition, it has been explored how existing teaching approaches from the field of BI can be extended and adapted to meet the new requirements of big data (Chen et al. 2012; Debortoli et al. 2014). Jacobi et al. (2014) propose a design model for IS curricula, taking the interdisciplinary nature of big data into account, while Gupta et al. (2015) suggest a multi-year curriculum for undergraduates, graduates and MBA students.

However, according to Wang (2015), there is still a lack of cutting edge teaching methods in the field of BDA dealing with real world scenarios and free online learning resources (e.g. from Teradata University Network or other academic alliances). Furthermore, existing research indicates a lack of specific knowledge about analytics software tools (Al-Sakran 2015) and in-depth knowledge about big data management (Chiang et al. 2012) in higher education.

Pedagogical Approach

Due to its interdisciplinary and complex nature teaching BDA challenges higher education. Therefore, traditional lectures, in which theoretical content is presented, should be supplemented by (1) a sensitization for the broad field of BDA as well as by (2) active learning formats (Asamoah et al. 2017; Dunaway 2017). Active learning provides the opportunity for students to learn through interaction with peers and/or the supervisor by working on exercises consisting of e.g. hands-on tutorials, real world case studies and experimental games (Asamoah et al. 2017). In order to organize such practice-oriented exercises, cooperation with software vendors or industry partners as well as the application of open source software and open educational resources are proposed in literature (Gupta et al. 2015; Schiller et al. 2015; Wixom et al. 2014). By the application of diverse teaching formats, different kinds of learning objectives can be achieved and in-depth knowledge can be transferred to the students.

To ensure the systematic development of a syllabus an underlying learning theory is required (Biggs and Tang 2011). In this context, a well-established approach widely used in the field of IS (Marjanovic 2012; Wixom et al. 2014) is the Revised Blooms Taxonomy (also known as Krathwohl's taxonomy), introduced by Krathwohl (Krathwohl 2002). The framework helps faculty to organize learning objectives and to guide students while they learn at different levels (Marjanovic 2012). Krathwohl distinguishes between the (1) knowledge dimension and the (2) cognitive process dimension (cf. Table 1).

	Cognitive process dimension					
Knowledge dimension	*Remember (Re)*	*Understanding (Un)*	*Apply (Ap)*	*Analyze (An)*	*Evaluate (Ev)*	*Create (Cr)*
Factual knowledge (FK)						
Conceptual knowledge (CK)						
Procedural knowledge (PK)						
Metacognitive knowledge (MK)						

Table 1. Revised Blooms Taxonomy (Krathwohl 2002)

On the one hand, the (1) knowledge dimension distinguishes the following levels (cf. Krathwohl 2002; Marjanovic 2012):

- Factual knowledge (fundamental terminologies and basic elements),
- Conceptual Knowledge (interaction and relationships between multiple basic elements),

- Procedural knowledge (methods, procedures, or techniques and their suitability in different contexts) and
- Meta-cognitive knowledge (knowledge about cognitive processes and one's own knowledge).

On the other hand, the (2) cognitive process dimension includes the levels (cf. Krathwohl 2002; Marjanovic 2012):

- Remember (retrieving knowledge from the long-term memory),
- Understand (understanding the meaning of instructions),
- Apply (using knowledge about procedures or methods),
- Analyze (disassembling problems in smaller parts and recognizing interrelationships between these parts),
- Evaluate (assessing content based on existing knowledge) and
- Create (assemble different elements to create a novel output).

For a detailed explanation of the two dimensions we refer to Krathwohl (2002). The paper at hand uses the Revised Blooms Taxonomy in order to organize the learning objectives of our BDM syllabus from a pedagogical perspective. Precisely, we apply this taxonomy to each lecture and tutorial and present examples how we have realized the learning objectives in class.

Course Setting

In this section, we give an overview of the basic setting of the course and explain the concept of the Big Data Lab (BDL), which serves as the underlying technical infrastructure in the tutorials.

General Setting

The master program Business Intelligence & Analytics comprises four semesters for full-time students. The goal of the program is to educate and train future BI professionals as well as data scientists and data engineers. The mandatory course Big Data Management, which is the main focus of the paper at hand, should be attended in the first semester. In addition, the course is open to other master programs, such as Value Chain Management, Customer Relationship Management and Computer Science master programs. The course aims at imparting basic knowledge in big data technologies and analytics for the students with a background in IS or computer science. As a first semester course it constitutes the foundation for more advanced courses in the master program, for instance courses in data mining, multimedia retrieval and database marketing as well as for subsequent capstone projects. As the students either attend the data mining course afterwards or have already attended in their undergraduate studies, this topic is not in the focus of the BDM syllabus. The same holds for typical data scientist programming skills, like in R or Python. As a master program course, BDM requires basic knowledge in database systems, data warehouses, and BI.

The course consists of biweekly lectures (90 min. each) providing the theoretical knowledge and a weekly tutorial (also 90 min each) as a hands-on training using tools from the BDA field. The lectures are complemented by a mandatory guest lecture at the end of the course to provide the students with a practical perspective on big data in general and big data projects from a management perspective.

Big Data Lab

The tutorials of the course take place in the BDL, which is operated and maintained by the professorship. This lab provides the required computing power and necessary software tools for the tutorials. Moreover, the lab is not only used for the BDM tutorials, but also for research, industry projects and further teaching purposes (e.g. master theses). Therefore, the lab is based on three pillars that are accompanied by four main application fields. A conceptual overview of the lab philosophy is illustrated in Figure 1.

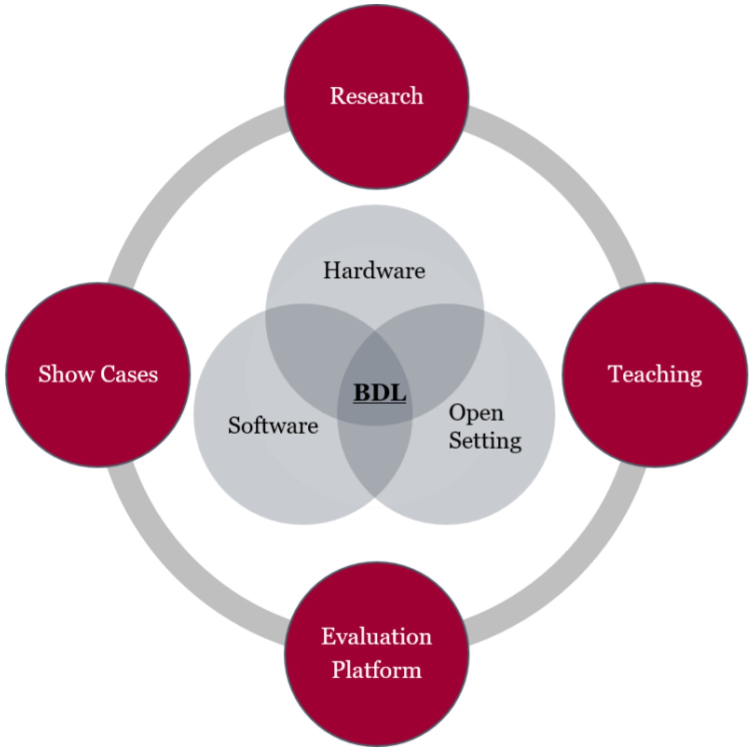

Figure 1. Conceptual Architecture of the Big Data Lab

The three pillars hardware, software, and open setting constitute the core of the lab. The hardware component encompasses the computing power back end as well as devices like Raspberry Pis and sensors. In detail, the back end is currently being extended by additional computing nodes. To provide a high level of flexibility with respect to the software and creativity, the hardware is abstracted by a particular infrastructure-as-a-service (IaaS) layer. Consequently, the use cases operate on virtual machines (VM) that can be deployed and undeployed remotely at any time with minimum effort.

The software component denotes the software tools required for all use cases, for instance Hadoop installations (cf. Section Software Tools) or IoT platforms for sensors. To provide a wide range of applications, different sets of preconfigured VMs are made available to students and researchers. These VMs can be combined to form a complex analytical infrastructure suited to specific problems or settings. Hence, a Hadoop cluster, for instance, can be easily set up by deploying the previously defined VMs for the master and slave nodes.

Finally, the open setting component refers to a creativity and innovation enhancing setting with no limits to thinking, doing, and testing in the context of big data and IoT. The lab will be extended in the near future by spaces for ideation sessions, walls suitable for writing, several single-board computers, and various sensors. Moreover, a flexible and start-up-like room layout supports the open setting of the BDL. This concept includes for instance multi-purpose tables or various multimedia front-ends for unsupervised usage and productive collaborations.

The three pillars constitute the basis for the four main use cases, namely academic research, teaching activities, evaluation platform, and show cases. Research refers to all academic and mission oriented research projects, including BDA, IoT settings, especially Industry 4.0, and data management. The teaching use case relates to all teaching activities in which the BDL is involved. The lab provides the required software tools for hands-on trainings as well as flexibility to set up additional software. Furthermore, the BDL serves as an evaluation platform for big data use cases and applications, like novel analytics platforms. Finally, the lab is used for show case scenarios to present big data and IoT opportunities to potential partners, regional companies, and interested students (e.g. at open university days).

Summed up, the BDL provides a flexible and creative environment to additionally support all kind of BDA scenarios in teaching and research.

Lecture Structure

Overview

The BDM course features seven lectures. The topics range from fundamentals of big data in general, over the Hadoop ecosystem and different database models, to BDA cases and data driven innovation. In Table 2 we present an overview of the lectures and the corresponding topics as well as a classification with respect to Krathwohl (2002) and list the addressed "V"s for each lecture. The classification according to the big data "V"s helps the students to understand which topics they would have to address in real-world scenarios if organizational data meets certain big data characteristics, and facilitates their learning processes by linking the lecture content with the well-known "V" classification. The classification with respect to the Krathwohl taxonomy is encoded in the following way: The first part refers to the respective knowledge layer whereas the second one relates to the cognitive process dimension (cf. Table 1). For example, CK-Un refers to conceptual understanding of a certain topic while FK-Re means remembering the basic factual knowledge.

Topics	Related tutorial	Classification by Krathwohl (2012)	Addressed big data Vs
Foundations o Understanding of big data o Big data applications o History of big data technologies	1	FK-Un, CK-Un	All "V"s
Hadoop (Distributed File System) and MapReduce o Hadoop Distributed File System (HDFS) o MapReduce o YARN o Apache Spark o Critical discussion of Hadoop	2, 6	FK-Un, CK-Un, CK-An	Volume, Variety
NoSQL and NewSQL database systems o Relational database model (repetition) o NoSQL database systems o In-Memory database systems o NewSQL database systems o NoSQL and NewSQL in Hadoop	3	FK-Un, CK-Un, CK-An	Volume, Variety, Velocity
Big Data Analytics 1 o Foundations (types of analytics, analytics process) o Data mining (short overview) o Text mining o Web mining o Network analytics (short overview)	4, 5, 8	FK-Un, CK-Un, PK-Ap	Variety, Value
Big Data Analytics 2 o Streaming analytics o Lambda architecture o Big data & IoT / digital manufacturing o Visualization	5, 7	FK-Un, CK-Un, PK-Ap	Velocity, Value
Big Data Management and Innovation o Data Warehousing / BI vs. big data o Data lakes and analytical ecosystems o Enterprise data management and governance o Data driven innovation	9	CK-Un, PK-An	Veracity, Value

Big Data & Law o Foundations: data privacy/protection regulations o Anonymization methods o Privacy by design	-	FK-Un, CK-Un	Veracity, Value

Table 2. Overview of the Lectures and Learning Objectives

Lectures Details

Below, we present the content of the lectures in detail. Afterwards, we explain the pedagogical classification, as represented in Table 2, with respect to Krathwohl's taxonomy and the aimed big data "V"s.

Lecture 01: Introduction and Foundations

The first lecture gives an introduction to the topic of big data from a high-level perspective. It introduces the concept of "Big Data" and explains the definition. The lecture specifically addresses "the three V's" (volume, velocity, variety) and demonstrates the extension of this definition, precisely the fourth and fifth "V". The second part of the lecture presents application fields and gives some real-life examples. Moreover, it is shown how big data technology extends existing domains and creates new applications and business models. The lecture concludes with the presentation of the two scaling methods (scale-out and scale-up) and the explanation of the CAP theorem.

This lecture aims for the basic factual and conceptual knowledge of the big data application landscape. It is achieved by the introduction and explanations of the big data "V"s. Moreover, the lecture addresses all big data "V"s.

Lecture 02: Hadoop (Distributed File System) and MapReduce

In the second lecture, the idea of distributed computing, Hadoop and its essential core elements are presented. Hadoop Distributed File System (HDFS), Hadoop MapReduce and YARN are covered in detail (structure and function). The historical development of Hadoop is also included.

With respect to Krathwohl's taxonomy (Krathwohl 2002), this lecture covers basic factual and conceptual knowledge, especially about the Hadoop ecosystem, but also aims for the students to understand and analyze concepts and their interrelationships. For instance, the MapReduce program model is taught by discussing a word counting example in detail. From a big data classification perspective, it addresses volume and variety.

Lecture 03: NoSQL and NewSQL database systems

The core of the third lecture is comprised by different database systems. The lecture introduces the various database technologies (traditional relational data model, NoSQL database systems and NewSQL database systems). The relational database model is primarily used to illustrate the main differences between the data models of more recently developed database systems. The students get familiar with NoSQL database systems and their different data models:

- Column-oriented databases
- Key-value databases
- Document-oriented databases
- Graph-oriented databases

Functionality, advantages and disadvantages, and typical application scenarios of the various technologies are discussed. Finally, the students learn how the CAP theorem can support the selection of the right database system type.

The lecture aims for basic factual knowledge with respect to the different database systems. Moreover, it focuses on conceptual knowledge about these systems, how they work and what their differences are. Later on, all database systems are briefly contrasted to each other. In this lecture the first three big data "V"s are addressed in particular.

Lecture 04: Big Data Analytics (1)

Lectures 04 and 05 focus on the different manifestations of big data analytics, as they have been suggested by Chen et al. (2012). Core concepts and terms are discussed in the fourth lecture before an abstract analysis process is presented. Since data mining is subject in other courses, only a brief overview is provided. Therefore, special emphasis is put on text mining and web mining (web content, web structure and web usage mining) and the respective performance indicators as well as different analytics tools and use cases. The topic of network analytics is only briefly discussed as it is taught in another mandatory course in detail. Finally, the students get familiar with a framework allowing them to select suitable analytics tools, depending on data volume and complexity.

The lecture aims for basic factual and conceptual knowledge on BDA. It provides the students with an overview of the different types of analytics approaches, like text and web mining, as well as use cases and tools. For example, web usage mining is explained by the functionality of a click stream analytics tool (applying procedural knowledge). The understanding of algorithms and the specific processes of the approaches is also included. Moreover, variety andvalue as the fifth big data "V" is the focus of this lecture.

Lecture 05: Big Data Analytics (2)

Streaming analytics as a further analytics type is covered in this lecture. Starting with a disambiguation, different types of data streams, common stream processing operators and application fields are illustrated. The students learn to differentiate between complex event processing and streaming analytics. With having an understanding of real-time processing then, the Lambda architecture can be introduced. As streaming analytics is frequently used in the context of IoT and digital manufacturing, these application fields are addressed, among others by means of a use case. The second part of the lecture provides an overview of visualization concepts and advanced visualization methods.

Lecture 05 has the same classification with respect to Krathwohl (2002) as lecture 04. For example, the conceptual knowledge for stream processing is presented by time series analytics for sensor data streams in combination with predictive maintenance. Additionally, we explain the relationships between the concepts and embed the stream processing within the lambda architecture model. Regarding the big data categorization, this lecture aims for the velocity and value.

Lecture 06: Big Data Management and Innovation

As the name of the Business Intelligence & Analytics master program suggests, the students gain in other courses a deep understanding of BI. Therefore, they should get familiar with the differences between BI and big data approaches as well understand how both worlds can be merged in analytical ecosystems. In addition, the lecture provides an overview how (traditional) data management has to be adapted and extended in the big data context. In particular, data quality management, metadata management, and data privacy & safety in the context of big data are presented. Data driven innovation constitutes the third part of the lecture. Students get familiar with the characteristics of data driven business models and innovation processes. Several real-life examples illustrate the potential value of new use cases for big data. Finally, the concepts of open innovation and open data are discussed.

From a pedagogical perspective the lecture targets the conceptual understanding of BDM as well as of data driven innovation. Additionally, it can be classified as applying the procedural knowledge, especially in the context of open data. The big data focus is set to veracity and value.

Lecture 07: Big Data and Law

The last lecture unit addresses aspects of law which are relevant in the context of big data. Firstly, basic terms and regulations, with regards to privacy, are presented to illustrate the challenges when realizing big data projects. The focus is set to privacy, anonymization, and copy right laws. Finally, privacy by design illustrates how a legal privacy framework can be implemented.

Moreover, the lecture addresses basic factual and conceptual knowledge about privacy in the context of big data. For instance, the conceptual knowledge with respect to privacy is illustrated by the Privacy by Design approach, which is also put into the legal context. With regard to the "V"s classification, this lecture addresses mainly value, because privacy aspects are of concern when data is analyzed.

Additional Teaching Resources

To prepare the class and to keep it up to date we draw back on several resources. On the one hand we build upon established educational textbooks and on the other hand on online resources, like communities, tutorials, and dedicated university networks.

Concerning educational textbooks, the course is based on (White 2012) and (Freiknecht 2014; in German). The books cover a broad range of big data fundamentals and technological knowledge, especially for the Hadoop ecosystem. To set up the tutorials we rely on community-based tutorials and resources, as novel and emerging technologies often lack established teaching material. In detail, for our tutorials we use the Apache communities (Noll 2011, Sako 2017, Apache Software Foundation 2017), the MongoDB community (MongoDB 2017), the RapidMiner tutorial (North 2012) and material from the EMC Academic Alliance. Furthermore, we always encourage students to use material and resources available online at Cognitive Class, an IBM community initiative (IBM 2017), and at the Teradata University Network (Teradata 2017).

Tutorial Structure Overview

To accompany the seven lectures, the course features nine tutorials with various purposes. On the one hand, the students act on the cognitive level of understanding and analyzing, and on the other hand the students apply certain technologies in dedicated hands-on trainings. The students are encouraged to work in small groups and to prepare the tutorial materials, which are available before the respective tutorial takes place. Table 3 presents an overview of all tutorials and categorizes each of them with respect to Krathwohl's taxonomy as well as the addressed big data "V"s. The content of each tutorial is presented afterwards. The taxonomy encoding is the same as in the lecture overview (cf. explanations for Table 2).

Tutorial	Content	Classification by Krathwohl (2012)	Addressed big data Vs
01	CAP theorem (Brewer's theorem), students play an interactive MapReduce card game to understand the concept of HDFS and distributed computing	CK-Un, PK-Un, PK-Ap	Volume, Variety, Velocity
02	Students present Hadoop ecosystem extensions to the class	CK-Un, CK-An	Volume, Variety, Velocity
03	Students set up and test a Hadoop cluster using VMs	CK-Un, PK-Un, PK-Ap, PK-An	Volume, Variety, Velocity
04	Students perform MapReduce jobs with MongoDB	CK-Ap, PK-Ap	Volume, Variety, Velocity
05	Students perform text analytics with RapidMiner	CK-Ap, PK-Ap	Variety, Value
06	Students analyze real-time temperature sensor data with IBM Bluemix (examplary IoT application)	CK-Un, CK-Ap, PK-Ap	Velocity, Value
07	Students analyze log data with Splunk	CK-Ap, PK-Ap	Volume, Velocity, Value
08	Students analyze usage patterns with Teradata Aster	CK-Ap, PK-Ap	Volume, Value
09	Students discuss big data case studies and participate in an interactive innovation workshop based on metadata descriptions	CK-Ap, CK-Ev, CK-Cr, PK-Ev, PK-Cr, MK-Un	All "V"s

Table 3. Overview of the Tutorial Units and Learning Objectives

Tutorial Details

The tutorial details are presented the same way as the lecture details. At first, we briefly describe the contents and learning goals. In a second paragraph we explain the categorizations with respect to Krathwohl's taxonomy and big data "V"s.

Tutorial 01: The CAP-Theorem & MapReduce Role Play

The first exercise repeats the CAP-Theorem and HDFS MapReduce, i.e. topics of the first and second lecture. As an individual exercise, the CAP-Theorem is repeated with an additional definition; the students explain the essential meaning of the components of the theorem in their own words. Afterwards the subject HDFS & MapReduce is repeated in individual work, with emphasis on the general concept of HDFS and the different stages of the MapReduce algorithm. Finally, a complex application example for the use of HDFS & MapReduce is solved in group work. The scenario is an online store whose portfolio is to be evaluated. To solve the problem, a role-play is performed by using cards of several card decks as key-value pairs. The students take on the roles of HDFS cluster nodes to simulate the MapReduce algorithm.

Pedagogically this tutorial aims for conceptual understanding as well as procedural understanding and analyzing. In particular, the interactive role playing game supports the procedural analyzing aspect. In the context of the big data "V"s classification, it corresponds to volume, variety, and velocity, which is demonstrated by the role play.

Tutorial 02: Brief Presentations - The Hadoop Ecosystem

The first part of the tutorial provides an overview of the Hadoop extensions and various Hadoop distributions. In the second part, student groups prepare short presentations about an extension of the Hadoop ecosystem and share their results with other students. Moreover, the groups are encouraged to classify their chosen extension within the Hadoop ecosystem and present relationships to other extensions.

From a pedagogical perspective this tutorial aims for conceptual knowledge about the Hadoop extensions, their conceptual understanding and a comparison with other extensions. This is achieved by an explanation of the Hadoop ecosystem and its extensions followed by the students research on specific Hadoop extensions and a subsequent presentation in front of the class. The addressed big data classification ranges from volume, over variety to velocity.

Tutorial 03: Designing and Testing a Hadoop Cluster with Apache Ambari

This hands-on unit includes the most complex and time-consuming task; about 4 hours are needed. Under instruction the students implement their own Hadoop cluster using Apache Ambari in the BDL and solve various tasks with this newly setup environment. In detail, the students create a VM, whereby the operating system needs to be configured correctly for the following steps. The cluster is created by an Ambari installation first, followed by the Hadoop cluster setup. Once the cluster is running, a word count test case is implemented and executed as a MapReduce job. All tasks are performed in groups of about three students.

The tasks in this tutorial aim for the conceptual understanding of Hadoop extensions and their interrelations. Furthermore, the students receive a procedural understanding by setting up such a cluster as well as procedural and analyzing knowledge by testing the setup. This is achieved by giving the students the task to set up a cluster by themselves under guidance of a tutor. Concerning the big data classification, the lecture serves volume, variety, and velocity.

Tutorial 04: NoSQL Databases (MapReduce with MongoDB & Neo4j)

The fourth tutorial includes the topic of NoSQL databases and therefore deepens the third lecture. MongoDB as a document-oriented database is presented. The students have to understand the basics of MongoDB on their own to solve a MapReduce task and perform full-text search in MongoDB. Afterwards Neo4j is presented as a representative for graph-oriented databases.

This tutorial aims for pure application skills of the underlying concepts and methods by having the students to create a database schema and to interact with it using basic queries as well as MapReduce jobs. Furthermore, it addresses volume, variety, and velocity in the context of big data.

Tutorial 05: Text Mining (RapidMiner)

Text analytics comprises the focus of tutorial 05. It is based on a case study taken from North (2012). In detail, the students have to load and process raw data in RapidMiner and perform certain analysis tasks on this data, for instance text analysis of customer complaints. Each analysis features specific algorithms, like for example k-means.

As the previous tutorial, it is a pure hands-on tutorial and aims for applying the acquired knowledge from a conceptual as well as a procedural perspective. This is achieved by drawing back to a text mining process introduced in the lecture and applying it on real world data to perform an analysis. With regard to the big data classification, volume and value are addressed in this tutorial.

Tutorial 06: IoT Applications and Analytics (IBM Bluemix)

This tutorial gives an introduction how to use big data technology within the context of IoT. As an example we rely on IBM Bluemix tutorials provided by IBM. At first the students get an introduction to IoT technologies and IoT platforms. They create their personal IoT setup within the Bluemix platform and analyze data generated by a (simulated) heat and humidity sensor. The students write a script to produce a Twitter tweet in case a certain temperature threshold is reached.

Based on Krathwohl’s taxonomy, this tutorial is classified as conceptual understanding and applying of IoT platforms as well as procedural application of a real scenario by applying the student’s knowledge about streaming data in order to monitor a simulated sensor. As the tutorial addresses an IoT setting, the focus lies on velocity and value.

Tutorial 07: Analyzing Log Data (Splunk)

At the beginning of this unit, the students become familiar with the concept of "Operational Intelligence" before they solve the case study "Telenor Gains Greater Insight for Incident Investigation, Troubleshooting and Improved Security" (Splunk n.d.) on their own. As a final task of this unit, the students deal with the hands-on exercise "Search Tutorial" (Splunk 2017). The exercise utilizes the documents provided by Splunk.

This tutorial is a pure hands-on exercise to develop conceptual and procedural application skills, by applying the knowledge gained in the previous tutorial to monitor performance indicators of a computer system. It addresses the big data characteristics volume, velocity, and value.

Tutorial 08: Big Data Appliance (Teradata Aster)

The goal of the tutorial is to experience BDA. For this purpose, first the basics of the BDA architecture of the discovery platform "Teradata Aster" are elaborated. This is followed by the implementation of selected use cases from the original Teradata big data analytics workshop "Query Grid”. This workshop covers the analysis of time series data as well as various visualization techniques for big data.

As the previous tutorial, this one features a hands-on training and aims for the same pedagogical goals, this time by detecting and visualizing patterns in data of TV watching and channel switching habits. However, only volume and value are addressed in this tutorial.

Tutorial 09: Big Data Use Cases and Innovation Workshop

In the first part the students solve two short big data use cases based on real world data. The use cases aim for simulating real-world scenarios in which organizations have to make a series of rather technical decisions when starting and implementing big data projects. For example, the students are asked to suggest the overall architecture, the database system type, the analytics types and the Hadoop extensions they would select in the concrete case. By that the student repeat and apply most content of the course in an overarching manner.

In the second part a data driven innovation workshop is conducted. Within the workshop, students participate in groups in order to create ideas for innovative products or services. As a resource for ideation, students are provided with short descriptions of various data sets, which are written on cue cards. These cue cards represent real world data sets, as available on open data portals (e.g. www.data.gov.uk). During the workshop, the students identify interesting data sets, combine them in an innovative way and develop concepts for products or services based on that data.

This unit concludes and summarizes the tutorials course by invoking the application of the knowledge taught throughout the lectures and tutorials. Overall, this leads to a broad range of learning objectives as well as a full coverage of the five "V"s, as seen in Table 3.

Software Tools

The tutorials focus on various use case scenarios in the field of BDA by providing special hands-on trainings. Consequently, the students get in touch with various software tools from different layers and vendors. In Table 4 we briefly classify these tools by their addressed layer and purpose. The tools may and should be adapted to emerging technologies and developments each year to ensure a state of the art education.

Layer / Application	Technology
Database layer	**Hadoop distributed file system** (HDFS) for Hadoop and MapReduce
	MongoDB as NoSQL document store
	Neo4j as NoSQL graph store
Management layer	**Apache Ambari** for cluster management and monitoring
	YARN for dynamic resource management
Programming layer	**Apache Pig** for high level MapReduce programming
	Optionally: Gnu R for mathematical / statistical computing
Visualization layer	**Gephi** for graph visualization
	Tableau for general data visualization
Application layer	**Splunk** for machine-generated data
	RapidMiner for data mining and text mining
	Teradata Aster for enterprise analytics
	IBM Bluemix as IoT platform
	Gephi for network analytics

Table 4. Overview of Software Tools for the Tutorials

We group the software tools into five categories. The first category represents the database layer, especially for big data scenarios. The corresponding representatives are HDFS, MongoDB, and Neo4j, covering a wide range of NoSQL database systems. The second category consists of two management tools for Hadoop systems. Apache Ambari represents the cluster management tools while YARN is introduced to demonstrate the resource management of a Hadoop cluster. To provide tools for the creation of analytical applications on the coding level, we use Apache Pig for high-level MapReduce programming and optionally Gnu R for statistical programming. The fourth category comprises two visualization tools, Gephi as a graph visualization tool and Tableau for data visualization in general. Finally, the application layer as the fifth category contains analytics applications. The students use Splunk to analyze machine-generated data (e.g. log data), RapidMiner for text mining in particular, Teradata Aster for an enterprise analytics scenario, IBM Bluemix as as an IoT analytics platform, and Gephi as a network analytics tool.

The software tools are open source or at least free of charge for academic use, which reduces the dependency on single vendors and gives the students the opportunity to continue their education on BDA with already known tools on their own by self-studying.

Discussion

The paper at hand presents a BDM syllabus for master students with a background in IS or computer science. The course is intended to convey the foundations of BDA as well as the underlying technologies and the ecosystem around it. Our approach combines traditional lectures with tutorials consisting of active learning formats and thus, addresses different levels of learning objectives. The lectures focus mainly on transferring factual and conceptual knowledge to achieve a fundamental understanding of big data, related concepts and their applications among the students. As a supplement, the exercises are designed to train the application and critical analysis of BDA and to reach an in-depth understanding of technological connections and dependencies (cf. Table 5).

The table shows how the lectures and tutorials fit into the Revised Blooms taxonomy. While *factual* and *conceptual* knowledge is mainly taught in the lecture units, procedural and, in some cases, metacognitive knowledge is conveyed primarily in the tutorials. With regard to the cognitive processes, remembering and understanding are the main objectives of the lectures. Thus, the lecture units serve as a foundation for the tutorials and enable the students to apply, analyze and evaluate the acquired knowledge as well as create a novel output (e.g. create data based services within the innovation workshop).

	Cognitive process dimension					
Knowledge dimension	*Remember (Re)*	*Understan-ding (Un)*	*Apply (Ap)*	*Analyze (An)*	*Evaluate (Ev)*	*Create (Cr)*
Factual knowledge (FK)						
Conceptual knowledge (CK)						
Procedural knowledge (PK)						
Metacognitive knowledge (MK)						

Learning objectives addressed in: lectures | tutorials | both

Table 5. Classification of Learning Objectives Using the Revised Blooms Taxonomy (Krathwohl 2002)

We have collected qualitative and quantitative feedback with regards to the tutorials from the students over two semester periods (n=35). Hereinafter, we summarize the most frequently mentioned issues, underpinned with quotes from students[1].

In general, the tutorials were evaluated with an excellent average grade of 1.52, where 1.0 is best and 5 the worst rating. Furthermore, the students appreciated the hands-on tutorials (in particular tutorial 3: Designing and Testing a Hadoop Cluster with Apache Ambari) as peculiarly positive and instructive:

- "Hands-on exercise was very informative!"
- "Practical approach should be kept"
- "Possibility to work independently during the exercise"

In addition, the wide range of different tools used in the tutorial was evaluated positively:

- "Many possibilities for self-studying and many insights in various technologies"
- "Working with tools dealing with various types of data"
- "The tutorials […] and the practical applications of various tools were good & very helpful for me, regarding my future career"

[1] Student quotes were translated from German by the authors.

Despite the frequent tutorial slots the broad range of topics and resulting tool exercise options causes time limitations. Therefore it is not surprising that the students would prefer more practical hands-on exercises.

- "Exercise should be offered 'more frequently' -> more 'hands-on units'"
- "More use cases and 2 additional practice hours, as these [exercises] were very helpful, but time too short"
- "Use case tutorial as long term self-study exercise (two weeks) -> Afterwards presentation and group discussion [in the class]"

From a faculty perspective the course runs very well. The students are more attentive and motivated than in other courses. The continuous extensions and changes in the field of big data, resulting in new content and new technologies & tools, require slight adaptations of the course each year. Main challenge in both, lectures and tutorials, is the limited time. Therefore, we consider to split the course in upcoming years and to teach it in two semesters.

Conclusion

As BDA is considered as a foundation for value creation and business success in today's digital economy, job profiles and the required skills, especially with respect to analytics in the context of big data, have changed. Consequently, universities have adapted their academic education towards these skills. The presented syllabus is embedded in a master program and serves as a basic course providing comprehensive knowledge on big data in general, big data technologies and their application scenarios, as well as big data analytics.

From a structural perspective, the course consists of seven lectures and nine tutorials designed as hands-on trainings. As we have presented the syllabus as it is currently taught, some parameters such as the number of lectures and tutorials, the name of the course and to a certain extent the content and the order of the units are subject to constraints of the master program at our university. Nevertheless, the overall course design might serve as a blueprint for similar courses at other universities. In such cases, the particular lecture order may vary, for instance the Big Data & Law unit (currently lecture 7) might be taught in the very beginning in order to additionally sensitize the students for this important topic. To provide a wide range of basics, the lectures address the foundations of big data, the Hadoop ecosystem, NoSQL / NewSQL technologies, and BDA methods, like text mining as well as streaming analytics. Additionally, the syllabus is completed by topics such as BDM and data driven innovation, and a classification of big data within the context of law and legal regulations. This knowledge is enhanced in the hands-on trainings in which the students apply various technologies, like Hadoop and Teradata Aster. As Table 5 clearly shows, the lecture units focus on remembering and understanding factual and conceptual knowledge, whereas the tutorials aim for applying, analyzing and evaluating conceptual and procedural knowledge.

The students' feedback for the tutorial evaluation was very positive. In sum, the special focus on hands-on exercises addresses the industry requirements for young professionals in data science or engineering by providing a deep insight in analytical approaches, technologies and various tools improving the technological and analytical skills of the students.

Acknowledgements

Special thanks go to Robert Krawatzeck for his initial work on the lectures and tutorials. Moreover, he collected the data for the quantitative and qualitative course feedback.

References

Accenture. 2014. "Big Success With Big Data," (available at https://www.accenture.com/us-en/insight-big-data-research.aspx).

Al-Sakran, H. O. 2015. "Development of Business Analytics Curricula to Close Skills Gap for Job Demand in Big Data," *Information and Knowledge Management*, (5:3), pp. 36–46.

Apache Software Foundation. 2017. "Pig Documentation," (available at http://pig.apache.org/docs/r0.17.0/start.html; retrieved September 15, 2017).

Asamoah, D. A., Sharda, R., Hassan Zadeh, A., and Kalgotra, P. 2017. "Preparing a Data Scientist: A Pedagogic Experience in Designing a Big Data Analytics Course," *Decision Sciences Journal of Innovative Education*, (15:2), pp. 161–190.

Biggs, J., and Tang, C. 2011. *Teaching For Quality Learning At University*, New York: Society for Research Into Higher Education.

Boncz, P. A., Manegold, S., and Kersten, M. L. 1999. "Database Architecture Optimized for the New Bottleneck: Memory Access," in *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 54–65.

Cegielski, C. G., and Jones-Farmer, L. A. 2016. "Knowledge, Skills, and Abilities for Entry-Level Business Analytics Positions: A Multi-Method Study," *Decision Sciences Journal of Innovative Education*, (14:1), pp. 91–118.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., and Gruber, R. E. 2008. "Bigtable: A Distributed Storage System for Structured Data," *ACM Trans. Comput. Syst.*, (26:2), New York, NY, USA: ACM, p. 4:1--4:26.

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data To Big Impact," *MIS Quarterly*, (36:4), pp. 1165–1188.

Chen, M., Mao, S., and Liu, Y. 2014. "Big Data: A Survey," *Mobile Networks and Applications*, (19:2), pp. 171–209.

Chiang, R. H. L., Goes, P., and Stohr, E. A. 2012. "Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline," *ACM Transactions on Management Information Systems*, (3:3), p. 12:1–12:13.

Davenport, T. H., and Patil, D. J. 2012. "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, pp. 70–76.

Dean, J., and Ghemawat, S. 2008. "MapReduce: Simplified Data Processing on Large Clusters," *Commun. ACM*, (51:1), New York, NY, USA: ACM, pp. 107–113.

Debortoli, S., Müller, O., and vom Brocke, J. 2014. "Comparing business intelligence and big data skills: A text mining study using job advertisements," *Business and Information Systems Engineering*, (6:5), pp. 289–300.

Dunaway, M. M. 2017. "Collaboration in a Data Analytics Curricula: An Active Learning Approach," in *Proceedings of Americas Conference on Information Systems (AMCIS)*, Boston, USA.

Freiknecht, J. 2014. *Big Data in der Praxis: Lösungen mit Hadoop, HBase und Hive - Daten speichern, aufbereiten, visualisieren*, Munich: Hanser.

Gupta, B., Goul, M., and Dinter, B. 2015. "Business intelligence and big data in higher education: Status of a multi-year model curriculum development effort for business school undergraduates, MS graduates, and MBAs," *Communications of the Association for Information Systems*, (36:1), pp. 449–476.

IBM. 2017. "IBM Cognitive Class," (available at https://cognitiveclass.ai; retrieved September 15, 2017).

Jacobi, F., Jahn, S., Krawatzeck, R., Dinter, B., and Lorenz, A. 2014. "Towards a Design Model for Interdisciplinary Information Systems Curriculum Development, As Exemplified by Big Data Analytics Education," in *Proceedings of the European Conference on Information Systems (ECIS)*, Tel Aviv, pp. 1–15.

Krathwohl, D. R. 2002. "A Revision of Bloom's Taxonomy: An Overview," *Theory into Practice*, (41:4), pp. 212–264.

Kwon, O., and Sim, J. M. 2013. "Effects of data set features on the performances of classification algorithms," *Expert Systems with Applications*, (40:5), pp. 1847–1857.

Landset, S., Khoshgoftaar, T. M., Richter, A. N., and Hasanin, T. 2015. "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *Journal of Big Data*, (2:1), p. 24.

Lukoianova, T., and Rubin, V. L. 2014. "Veracity Roadmap: Is Big Data Objective, Truthful and Credible?," *Advances in Classification Research Online*, (24:1), pp. 4–15.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. 2011. "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute*.

Marjanovic, O. 2012. "Using the Revised Bloom's Taxonomy To Scaffold Student Learning in Business Intelligence/Business Analytics," in *Proceedings of the European Conference on Information Systems (ECIS)*, Barcelona, Spain.

MongoDB. 2017. "The MongoDB Manual," (available at https://docs.mongodb.com/manual/; retrieved September 15, 2017).

Noll, M. 2011. "Running Hadoop on Ubuntu Linux (Multi-Node Cluster)," (available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/; retrieved September 15, 2017).

North, M. 2012. *Data Mining for the Masses*, Global Text Project.

Prince, M. 2004. "Does Active Learning Work? A Review of the Research," *Journal of Engineering Education*, (93:3), pp. 223–231.

Provost, F., and Fawcett, T. 2013. "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, (1:1), pp. 51–59.

Puget, J. F. 2013. "Big Data For Dummies," (available at https://www.ibm.com/developerworks/community/blogs/jfp/entry/big_data_for_dummies23?lang=en; retrieved September 15, 2017).

Russom, P. 2011. "Big data analytics," *TDWI best practices report, fourth quarter*, (19), p. 40.

Sako, Y. 2017. "Apache Ambari Quick Start Guide," (available at https://cwiki.apache.org/confluence/display/AMBARI/Quick+Start+Guide; retrieved September 15, 2017).

Schiller, S., Goul, M., Iyer, L. S., Sharda, R., Schrader, D., and Asamoah, D. 2015. "Build Your Dream (Not Just Big) Analytics Program," *Communications of the Association for Information Systems*, (37:1).

Splunk. 2017. "Search Tutorial," (available at http://docs.splunk.com/Documentation/Splunk/latest/SearchTutorial/WelcometotheSearchTutorial; retrieved September 15, 2017).

Splunk. (n.d.). "Telenor Gains Greater Insight for Incident Investigation, Troubleshooting and Improved Security," (available at https://www.splunk.com/pdfs/customer-success-stories/splunk-at-telenor.pdf; retrieved September 15, 2017).

Teradata. 2017. "Teradata University Network," (available at http://www.teradatauniversitynetwork.com/; retrieved September 15, 2017).

Topi, H., Karsten, H., Brown, S. A., Carvalho, J. A., Donnellan, B., Shen, J., Tan, B. C. Y., and Thouin, M. F. 2017. "MSIS 2016: Global Competency Model for Graduate Degree Programs in Information Systems," *Communications of the Association for Information Systems,* (40:1), Article 18.

Wang, Y. 2015. "Business Intelligence and Analytics Education: Hermeneutic Literature Review and Future Directions in IS Education," in *Proceedings of Americas Conference on Information Systems Proceedings (AMCIS)*, Puerto Rico, pp. 1–10.

White, T. 2012. *Hadoop: The Definitive Guide*, Sebatstopol, USA: O'Reilly Media.

Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., Kulkarni, U., Mooney, J. G., Phillips-Wren, G., and Turetken, O. 2014. "The current state of Business Intelligence in academia: The Arrival of Big Data," *Communications of the Association for Information Systems*, (34:1), pp. 1–13.