CHAPTER 1
Introduction

1.1 Motivation

In recent years, *big data* has emerged as one of the leading trends not only in computer science, but due to its potential, also in economy, science, and major branches of the industry. People realized that huge data sets have become a key asset which should be taken into account in evaluating business opportunities, company valuations, or product development. Several major mergers and acquisitions in recent years have been driven not only in order to gain synergies, customer base, or market access, but also to obtain access to valuable customer data. For example, Microsoft's acquisition of Linkedin gave it data on jobs, skills, career paths, and a contact network of millions of workers across the globe.

For technology vendors, consultancies as well as numerous startups, this rapid growth opened up huge new business opportunities. According to IDC, the market value of big data and business analytics is expected to grow beyond $200 Billion by the year 2020. Forbes [2017]. These forecasts have fueled huge investments in big data related research and development efforts, both in academia and in industry, leading to a wide range of proposed architectures, solutions, models, algorithms, as well as commercial products.

Large industry players have made the big data concept fundamental to their products, architectures, and strategies. Every day, new ventures emerge which concentrate solely on big data as an opportunity for innovation and growth. Those who failed to follow the trend early see the rising competition and disruption, even in well established and heavily regulated industries such as banking or insurance, as can be observed by the growing number of *fintech* and *insurtech* ventures.

Academia has been intensively updating curricula to educate the next generation of data scientists, big data engineers, DevOps, etc. The research areas and goals of computer science departments have followed suit. New dedicated MOOCs (Massive Online Open Courses) become available every month and gather thousands of attendants. The number of conference tracks and entirely new events around the subjects of analytics and processing of big data is growing rapidly each year.

While there is no single agreed on definition of *big data*, it is commonly regarded as a general move towards analytics and applications, which rely heavily on processing of extremely large data sets in order to provide intelligent, personalized services to the users and other services in the ecosystem. This trend has been largely supported by recent advances in parallel computing architectures, emergence of NoSQL databases, cloud computing technologies and continuous improvements in machine learning and other branches of Artificial Intelligence (AI).

Multi-Agent Systems (MAS) use the concept of the *agent* as a central entity for building systems. This is often confusing as the term is heavily overloaded even within computer science, not to mention its use in multiple other disciplines such as economy, sociology, cognitive science, etc. MAS however iterates specifically the properties an agent should implement. It should be autonomous, understood as making its own decision based on internal state, goals, and observations. An agent should be proactive, so it should act when it believes it is appropriate not only when explicitly called. Finally, it should be intelligent in the AI sense of intelligence, therefore capable of solving complex tasks and learning by past experiences. Building on such components, MAS tries to assemble complex systems in which agents communicate asynchronously and collaboratively solve given tasks.

Even though MAS emerged as a separate field of research much earlier than big data, it failed to achieve such wide adoption and popularity. We can identify several reasons for this. One is that, until recently, there were no advanced and mature architectures for efficient distributed asynchronous processing. Only in the last decade the limitations to Moore's Law increased the efforts towards parallel computations. Another reason is the radical approach to the distribution of control in MAS. Agents were proposed as highly independent, autonomous, proactive entities communicating with the use of "soft" protocols, e.g. negotiation, argumentation etc. These assumptions were not in line with available means for monitoring of such systems, and so were unacceptable for several practical industry applications, where strict control and risk minimization are key, e.g. energy grid management, financial systems, traffic monitoring, etc.

This publication argues that the fields of big data and MAS have a lot in common. If we track the evolution of the IT systems from monolith, through SOA to microservices and most recently cyber-physical systems, we can see that the elementary system components more and more resemble agents as proposed many decades back. We rely more and more on loosely-coupled components centered around some well defined functionality and capable of autonomous and flexible operations even if other components fail or are temporarily out of reach. Now that distributed, cloud based computing has become standard, database paradigms have shifted from a strict transactional

approach and physical objects obtain built-in intelligence, MAS approach no longer looks radical and unfeasible.

It seems we have arrived at the point where several research results achieved in both fields can be combined and benefit from cross-fertilization of ideas, tools, and architectures. Mobile agents for sensor networks can be applied for real time analytics in the fast growing area of the Internet of Things (IoT). Distributed machine learning algorithms can be coordinated with multi-agent cooperation protocols. Mobile and IoT cloud computing environments experience challenges related to resources and latency similar to the ones present in MAS especially for mobile agents.

On the other hand, modern big data environments offer unprecedented possibilities of performing large scale computations both in batch and streaming mode, which can greatly enhance capabilities of MAS. Cloud resources supporting mobile and IoT devices might well be used to empower intelligent agents located in the environment. On the lower level, modern distributed programming libraries (e.g. Scala Akka) can greatly improve performance of MAS, which often use less advanced environments, not capable of efficient thread and resource management.

1.2 Assumptions

While establishing the scope and focus of this book, several assumptions and compromises had to be made. Firstly, when describing a field such as Big Data, where new concepts and projects emerge on a daily basis, it is difficult to resist the temptation to include every new finding, so the book will be as up to date as possible at the time of publishing. On the other hand it is difficult to predict the future of freshly proposed solutions, before they become more mature and are hardened by real life applications.

Therefore, difficult choices have been made and some might argue that a particular important architecture, project, or framework has been left out. In general, I have been following the rule of writing about topics, which have some proven maturity, e.g. have become mainstream Apache projects, have been followed by highly cited publications, have been applied by at least one of the large and recognized industry players, etc.

Secondly, since the book title refers to big data architectures, the contents concentrate on large scale solutions capable of solving practical problems experienced in the industry. Therefore, specific tools applicable at particular points in the larger architectures are described only to the point where they are relevant from the point of view of the big picture they take part in, rather than in their internal and technical details. For example Hadoop, which is often regarded as a technological synonym for big data, is described

as a component for batch processing used in larger big data architectures. Map-Reduce, Hadoops', underlying algorithm, is presented as one of the generic computational models for processing extremely large data sets. Similarly, Spark is an example of stream processing and plays that role in larger big data setups.

In the field of MAS things have been somewhat easier, since the field is more mature in general and several comprehensive textbooks have been published to date, which summarize the research and development efforts in this area. Therefore, major agent models and architectures are described in line with the state of the art long established in the field. This is complemented with some more recent and more specific examples of applications of multi-agent paradigms in solving various big data problems.

1.3 For Whom Is This Book?

This book could be of interest to both researchers and practitioners from the fields of big data, analytics, machine learning, MAS, distributed computing, cloud computing, distributed artificial intelligence, as well as a number of other related fields.

The intention has been, for anyone from the fields mentioned above to see the current state of the art in distributed, asynchronous processing of massive data sets. As well as this it will be shown how various field and areas of research relate to each other by tackling similar issues and challenges from their respective perspectives.

For big data practitioners not familiar with MAS research it may come as a surprise how many relevant ideas and concepts have already been analyzed several years back. MAS researchers will find several big data environments, libraries, and tools very useful for taking their systems to the next level of efficiency.

In the end I hope that this book will initiate mutual discussion and exchange of ideas, which is to some extent already present but could become much more intense and fruitful.

1.4 Book Structure

The book is organized as follows. Chapter 2 discusses how major paradigms and concepts have changed over the last few decades, leading to the current landscape. Specifically we will analyze how the evolution of IT architectures influenced storage and analytics of the data. We will also look at the shift of paradigms in database systems, the growing role of the cloud, the Internet,

and the IoT. Also the concepts of an agent and an actor are introduced. We conclude by discussing how all these trends led to the rise of big data.

In Chapter 3 we look at where the data comes from in the big data setups. We start with the Internet as the most commonly available data source today. Then we iterate over various branches of science and industry looking at how much data they generate and what is specific about each of them. Finally, the IoT as the fast growing source of huge data streams is described.

Once we are familiar with the data sources, the book dives into specific tasks which need to be performed with the use of the data. Chapter 4 looks at the most important challenges that research and industry is working on in the big data area. This covers recommender systems, search, real time bidding, as well as multiple other topics.

Cloud computing is discussed in Chapter 5. It deserves a separate chapter as a major trend shaping the creation of the next generation of information systems. We look at the advantages and challenges of utilizing cloud resources and how it enables the building of scalable, distributed big data systems. The means for efficient cloud management both in VM and container based setups are described.

In Chapter 6 several big data architectures are presented. We start with fundamental computational models and move towards more complex setups. This includes among others Lambda and Kappa architectures, which have recently emerged as important design patterns for building scalable big data processing and analytics. A separate section is devoted to stream processing.

The means for data analytics and building machine learning models are the subject of Chapter 7. The role of SQL versus other forms of ad-hoc interaction with the data is analyzed. Tools and architectures for providing SQL capabilities in noSQL environments are analyzed. We look at frameworks and tools for efficient building, deploying, and testing of machine learning models.

Geographically distributed systems are the topic of Chapter 8. We will take a look at how the latest trends driven by mobile computing and the IoT led to the emergence of edge and fog computing as new paradigms for extending the cloud towards the distributed elements of the cyber-physical systems.

The work is closed by Chapter 9 with a summary and conclusions. References to the literature complete the volume.