Preface

Since the introduction of the Internet, we witness an explosive growth in the volume, velocity, and variety of data created on a daily basis. These data originated from numerous sources including mobile devices, sensors, individual archives, Internet of Things, government data holdings, software logs, public profiles in social networks, commercial datasets, etc. Everyday, more than 2.5 exabytes of data—as much as 90 % of the data in the world—are generated, and such data volumes are growing faster than they can be analyzed. Recently, large federated computing facilities have become prominent for enabling advanced scientific discoveries. For example, the Worldwide Large Hadron Collider Computing Grid, a global collaboration of more than 150 computing centers in nearly 40 countries, has evolved to provide global computing resources to store, distribute, and analyze 25 petabytes of data annually.[1] In geosciences, new remote sensing technologies provide an increasingly detailed view of the world's natural resources and environment. The data captured in geosciences domain enable more sophisticated and accurate analysis and prediction of natural events (e.g., climate change) and disasters (e.g., earthquake). Huge progress can also be observed in developments of sensor-based and mobile technologies supporting the new Internet of Things (IoT) systems.

The issue so-called "Data Intensive Computing (DIC)", or more generally "Big Data" problem, requires a continuous increase of processing speed in data servers and in the network infrastructure, and it becomes difficult for the analysis and interpretation with on-hand data management applications. Hundreds of petabytes of big data generated everyday need to be efficiently processed (stored, distributed, and indexed with a schema and semantics) in a way that does not compromise end-users' Quality of Service (QoS) in terms of data availability, data search delay, data analysis delay, data processing cost, and the like. The evolution of big data processing technologies includes new generation scalable cloud computing data centers, distributed message queue frameworks (e.g., Apache Kafka,[2]

[1] wlcg.web.cern.ch

[2] www.en.wikipedia.org/wiki/Apache_Kafka

Amazon Kinesis[3]), data storage frameworks (e.g., MongoDB,[4] Cassandra[5]), parallel processing frameworks (e.g., Hadoop,[6] Spark[7]), and distributed data mining frameworks (e.g., Mahout[8], GraphLab[9]). However, there is still a lack of effective decision making and data mining support tools and techniques that can significantly reduce the data analysis delays, minimize data processing cost, and maximize data availability.

Over the past decades, multi-agent systems (MAS) have been put forward as a key methodology to develop and tackle the complexity of self-managing large-scale distributed systems and can be applied in solving the problems with massive data processing and analytics and social components. The proactiveness of intelligent agents became a valuable key feature in the recent data intensive computing and big data analytics. The architecture of the multi-agent systems in such approaches provides agents with different competence levels that can adequately fit the computing capabilities of each system component, from client agents in mobile devices to more heavy-duty ones as nodes/servers in networks of highly distributed environments. Systems with such MAS support can run in a cooperative stance, with no need for central coordination.

This compendium herewith presents novel concepts in the analysis, implementation, and evaluation of next generation agent-based and cloud-based models, technologies, simulations, and implementations for data intensive applications. The general big data analytics problems are defined in "Bio-Inspired ICT for Big Data Management in Healthcare". Di Stefano et al. present the concept of bio-inspired approach to ICT systems, which is helpful in extraction of general knowledge from data, and make it available to those who have to design ICT interventions and services, considering the multitude of resources, in terms of data and sources, computational limits, and social dynamics. The authors demonstrate how to efficiently deploy the multi-agent framework to implement such bio-inspired mechanisms (social dynamics and biodiversity) for big data processing, storage management, and analysis. The illustrating application area is patient-centered healthcare, where big data analytics and storage techniques become crucial due to the large amounts of personal data related to patient care. The presented health mining paradigm allows to analyze correlations and causality relations between diseases and patients, which is the background of today's personalized medicine.

[3] www.aws.amazon.com/kinesis

[4] www.mongodb.org

[5] cassandra.apache.org

[6] hadoop.apache.org

[7] spark.apache.org

[8] mahout.apache.org

[9] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J.M. Hellerstein, "Distributed GraphLab: A Framework for Machine Learning in the Cloud," in VLDB Endow, 2012, pp. 716–727.

The remaining eight chapters are structured into two main parts:

1. *Agents*: A growing number of complex distributed systems are composed of customisable knowledge-based building blocks (agents). "Control Aspects in Multiagent Systems"–"Large-Scale Simulations with FLAME" discuss novel agent-oriented concepts and techniques that can be useful in big data analytics, from agents' organization, interaction and evolution models, to large-scale agents' simulators and demonstrators.
2. *Clouds*: Collectively known as computational resources or simply infrastructure, for which computing elements, storage, and services represent a crucial component in the formulation of intelligent cloud systems. Consequently, "Cloud Computing and Multiagent Systems, a Promising Relationship"–"Adaptive Resource Allocation in Cloud Computing Based on Agreement Protocols" showcase techniques and concepts for modernization of standard cloud mechanisms, to manage big data sets and to integrate them with multi-agent frameworks.

Agents

The first three chapters in this part focus on MAS organization, evolution, and management. In "Control Aspects in Multiagent Systems", Cicirelli and Nigro discuss the control methods in agent systems. They developed a scalable control framework for modeling, analysis, and execution of parallel and distributed time-dependent MAS with asynchronous agents' actions and message-passing. That framework has been implemented in popular JADE MAS environment and tested in the following two practical use cases: (i) MAS-supported help desk offering services to a company's customers and (ii) the schedulability analysis of a real-time tasking model under fixed priority scheduling. The provided experiments show high flexibility and practical usefulness of the proposed control methods, thanks to the separation of the agents' control structures from the business logic of MAS.

The analysis of agents' behavior and interactions remains a challenging research task, especially in big data approaches. Recently, Markovian agents (MAs) have been proposed as an effective MAS solution for data intensive problems, such as wireless sensor networks deployed in environmental protection, situation awareness, etc. This model is represented by a continuous time Markov chain (CTMC). Additionally, for Markovian agents, the infinitesimal generator has a fixed local component that may depend on the geographical position of the MA, and a component that depends on the interactions with other MAs. *Gribaudo and Iacono* in "A Different Perspective of Agent-Based Techniques: Markovian Agents" focus on the MAS performance evaluation problem and exploit decoupling between the MAS model representation with the software implementation and the performance evaluation technique used to study the overall behavior of the system. In their approach, MA is presented as a very valuable tool when the number of entities of a model is too big for a reasonable application of other state space-based techniques or accurate simulation approaches.

The implementation of scalable efficient agents' organization and interaction models is a challenging task also for engineers and practitioners. Gath et al. in "Autonomous, Adaptive, and Self-Organized Multiagent Systems for the Optimization of Decentralized Industrial Processes" provide a comprehensive critical state-of-the-art analysis of agent-based approaches in transport logistics and indicate the main limitations for their application in Industry 4.0 processes. The authors developed the *dispAgent* model with coordination and negotiation protocols for agents' synchronization, and a solver for the agents' individual decisions. They evaluated this model on two established benchmarks for the vehicle routing problem with real-world big data. The experiments demonstrated the difficulties in practical applications of theoretically developed agents' behavioral models.

Potentially, the discussed difficulties in MAS research may be partially solved through developments of novel implementation tools (new dedicated programming languages, ontologies) and scalable simulation platforms for large-scale MAS. Subburaj and Urban in "Formal Specification Language and Agent Applications" developed the extended Descartes specification language for the improvement of MAS implementation dedicated to concrete applications. The domains analyzed include information management, electronic commerce, and medical applications where agents are used in patient monitoring and health care. Coackley et al. in "Large-Scale Simulations with FLAME" present various implementations of FLAME simulation environment for execution of very complex data-intensive tasks in large-scale MAS. They justified the flexibility, efficiency, and usefulness of the developed software environments in a comprehensive application analysis in different areas (economy, biology, and engineering). A formal verification of the effectiveness of the proposed simulator was discussed for the case of a rule-based system, which can be formally analyzed using model checking techniques.

Clouds

Cloud computing gives application developers the ability to marshal virtually unlimited resources with an option to pay-per-use as needed, instead of requiring upfront investments in resources that may never be optimally used. Once applications are hosted in cloud resources, users are able to access them from anywhere at any time, using devices in a wide range of classes, from mobile devices (smartphones, tablets) to large computers. The data center cloud provides virtual centralization of application, computing, and data. While cloud computing optimizes the use of resources, it does not (yet) provide an effective solution hosting big data applications.

Large-scale distributed data-intensive applications, e.g., 3D model reconstruction in medical imaging, medical body area networks, earth observation applications, distributed blog analysis, and high energy physics simulation, need to process and manage massive data sets across geographically-distributed data centers. However, the capability of existing data center computing paradigms (e.g., MapReduce, workflow technologies) for data processing is limited to compute and store infrastructures within a local area network, e.g., a single cluster within a data center. This leads to unsatisfied Quality of Service (QoS) in terms of timeliness of results, dissemination of results across organizations, and administrative costs. There are multiple reasons for this state of affairs including: (i) lack of software frameworks and services that allow portability of such applications across multiple data centers (e.g., public data centers, private data centers, hybrid data centers, etc.); (ii) unavailability of required resources within a data center; (iii) manual approaches leading to non-optimal resource provisioning; and (iv) lack of a right set of programming abstractions, which can extend the capability of existing data processing paradigms to multiple data centers. "Cloud Computing and Multiagent Systems, a Promising Relationship"–"Adaptive Resource Allocation in Cloud Computing Based on Agreement Protocols" discuss in detail these limitations and present novel concepts in cloud computing for supporting big data analytics.

One of the most promising research trends in today's data intensive cloud computing is an integration of cloud environment with external scalable systems or models such as MAS. De La Pietra and Corchado in "Cloud Computing and Multiagent Systems, a Promising Relationship" applied multi-agent model for the management of the cloud infrastructure, services, and resources. They developed an agent-based cloud monitoring and resource management model which allows to support the local decisions of asynchronous agents and their interactions.

The effectiveness of big data storage and resource provisioning cloud technologies strongly depends on the data type and users' personal requirements or ethical procedures. Data protection and confidentiality have become one of the major challenging issues in cloud developments. Currently, private and public cloud data center providers support the following three storage service abstractions that differ in the way they store, index, and execute queries over the stored data: (i) Binary Large Object (BLOB) for binary files; (ii) Table-like storage based on key-value store; and (iii) RDBMS (Relational Database Management System), suitable for managing relational or structured data. Nevertheless, all the above technologies can only support shared-everything cloud application architectures, which can be clustered, replicated, and scaled on demand. There is a high inherent risk of data exposure to third parties or data tampering by a third party on the cloud or by the cloud provider itself. This has to be prevented, especially in emergency and health-related approaches. In "Privacy Risks in Cloud Computing", Del Mar Lopez Ruiz and Pedraza discuss the identification of legal and technical risks that threat user's data in public and hybrid clouds. The authors describe various aspects of privacy data protection including encryption techniques, privacy protocols, obfuscation, access control, resources separation, etc. Their analysis shows that the selection of security systems in cloud computing environment must take into account many complex factors such as cloud service model, results of privacy impact assessment, and the kind of data collected, stored and transferred. However, intelligent agent systems can be an effective cloud support due to its abilities of integration, interpretation, and classification of large amount of events, tasks, and unstructured data to provide personalized services in cloud environments.

Finally, in "Adaptive Resource Allocation in Cloud Computing Based on Agreement Protocols", Pop et al. address the classical scheduling problem in clouds for data intensive tasks, where negotiation of the resource provisioning, especially in distributed data centers, is the crucial issue. They implemented and tested novel negotiation strategies and protocols and evaluated the developed models on realistic big data sets.

We strongly believe that this book will serve as a reference for students, researchers, and industry practitioners currently working or interested in joining interdisciplinary works in the areas of data intensive computing and big data systems using emergent large-scale distributed computing paradigms. It will also allow newcomers to grasp key concepts and potential solutions in advanced topics of theory, models, technologies, system architectures, and implementation of applications in multi-agent systems and data intensive computing.

Kraków	Joanna Kołodziej
Lisbon	Luís Correia
Madrid	José Manuel Molina
June 2015	

http://www.springer.com/978-3-319-23741-1

Intelligent Agents in Data-intensive Computing
Kołodziej, J.; Correia, L.; Manuel Molina, J. (Eds.)
2016, XVIII, 216 p., Hardcover
ISBN: 978-3-319-23741-1