Chapter 1
Ecosystem of Big Data

Valentina Janev

Institute Mihajlo Pupin
University of Belgrade, Serbia

Abstract. 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 chapter 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.

1 Introduction

In 2001, in an attempt to characterize and visualize the changes that are likely to emerge in the future, Douglas Laney [269] of META Group (Gartner now) proposed three dimensions that characterize the challenges and opportunities of increasingly large data: Volume, Velocity, and Variety, known as the *3 Vs of big data*. Thus, according to Gartner

> "Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

According to Manyika et al. [295] this definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data. Along this lines, big data to Amazon or Google (see Table 1) is quite different from big data to a medium-sized insurance or telecommunications organization. Hence, many different definitions have emerged over time (see Chapter 3), but in general, it refers to "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" [295] and technologies that address "data management challenges" and process and analyze data to uncover valuable information that can benefit businesses and organizations. Additional “Vs” of data have been added over the years, but Volume, Velocity and Variety are the tree main dimensions that characterize the data.

The volume dimension refers to the largeness of the data. The data size in a big data ecosystem can range from dozens of terabytes to a few zettabytes and is still growing [480]. In 2010, the McKinsey Global Institute estimated that enterprises globally stored more than 7 exabytes of new data on disk drives, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks. While more than 800,000 Petabytes (1 PB= 10^{15} bytes) of data were stored in the year 2000, according to International Data Corporation expectations [344] this volume will exceed 175 zettabytes (1 PB= 10^{21} bytes) by 2025 [84].

The velocity dimension refers to the increasing speed at which big data is created and the increasing speed at which the data need to be stored and analysed, while the variety dimension refers to increased diversity of data types.

Variety introduces additional complexity to data processing as more kinds of data need to be processed, combined and stored. While the 3 Vs have been continuously used to describe big data, the additional dimensions of veracity and value have been added to describe data integrity and quality, in what is called the *5 Vs of big data*. More Vs have been introduced, including validity, vulnerability, volatility, and visualization, which sums up to the *10 Vs of big data* [137] (see Table 1). Regardless of how many descriptors are isolated when describing the nature of big data, it is abundantly clear that the nature of big data is highly complex and that it, as such, requires special technical solutions for every step in the data workflow.

2 Big Data Ecosystem

The term **Ecosystem** is defined in scientific literature as a complex network or interconnected systems. 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. Thus, large datasets – log files, social media sentiments, click-streams – are no longer expected to reside within a central server or within a fixed place in the cloud. To handle the copious amounts of data, advanced analytical tools are needed which can process and store billions of bytes of real-time data, with hundreds of thousands of transactions per second. Hence, the goal of this book is to introduce definitions, methods, tools, frameworks and solutions for big data processing starting from the process of information extraction, via knowledge processing and knowledge representation to storing and visualization, sense-making, and practical applications.

Table 1: Big Data Characteristics

3Vs	**Volume**	vast amount of data that has to be captured, stored, processed and displayed
	Velocity	rate at which the data is being generated, or analyzed
	Variety	differences in data structure (format) or differences in data sources themselves (text, images, voice, geospacial data)
5Vs	**Veracity**	truthfulness (uncertainty) of data, authenticity, provenance, accountability
	Validity	suitability of the selected dataset for a given application, accuracy and correctness of the data for its intended use
7Vs	**Volatility**	temporal validity and fluency of the data, data currency and availability, and ensures rapid retrieval of information as required
	Value	usefulness and relevance of the extracted data in making decisions and capacity in turning information into action
10Vs	**Visualization**	data representation and understandability of methods (data clustering or using tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone trees)
	Vulnerability	security and privacy concerns associated with data processing
	Variability	the changing meaning of data, inconsistencies in the data, biases, ambiguities, and noise in data

3 Components of the Big Data Ecosystem

In order to depict the information processing flow in just a few phases, in Figure 1, from left to right, we have divided the processing workflow into three layers:

- Data sources;
- Data management (integration, storage and processing);
- Data analytics, Business intelligence (BI) and knowledge discovery (KD).

Such partition will allow the authors of this book to discuss big data topics from different perspectives. For computer scientists and engineers, big data poses problems of data storage and management, communication, and computation. For data scientists and statisticians responsible for machine learning models development, the issues are how to get usable information out of datasets that are too huge and complex for many traditional or classical methods to handle. From an organizational viewpoint, business analysts are expected to select and deploy analytics services and solutions that contribute mostly to the organizational strategic goals, for instance, taking into consideration a framework for measuring the organizational performance.

Data Sources. In a modern data ecosystem, the data sources layer is composed of both private and public data sources – see the left side of Figure 2. The corporate data originates from internal systems, cloud-based systems, as well as external data provided from partners and third parties. Within a modern data

Table 2: Examples of Big Data Ecosystems

Facebook	Facebook (2018) has more than two billion users on **millions of servers**, running thousands of **configuration changes** every day involving trillions of configuration checks. [308]
LinkedIn	It takes a lot of horsepower to support LinkedIn's 467 million members worldwide (in 2017), especially when you consider that each member is getting a **personalized experience** and a **web page** that includes only their contacts. Supporting the load are some 100,000 servers spread across **multiple data centers**. [214]
Alibaba	The 402,000 **web-facing computers** that Alibaba hosts (2017) from China-allocated IP addresses would alone be sufficient to make Alibaba the second largest **hosting company** in the world today. [319]
Google	There's no official data on how many servers there are in Google's data centers, but Gartner estimated in a July 2016 report that Google at the time had 2.5 **million servers**. Google data centers process an average of 40 million searches per second, resulting in 3.5 billion searches per day and 1.2 **trillion searches per year**, Internet Live Stats reports. [388]
Amazon	... an estimate of 87 AWS datacenters in total and a range of somewhere between 2.8 and 5.6 million servers in Amazon's **cloud** (2014). [299]
Twitter	Twitter (2013) now has 150M worldwide **active users**, handles 300K queries per second (QPS) to generate timelines, and a firehose that churns out 22 MB/sec. Some 400 million tweets a day flow through the system and it can take up to 5 minutes for a tweet to flow from Lady Gaga's fingers to her 31 million followers. [196]

architecture, any type of data can be acquired and stored; however, the most challenging task is to capture the heterogeneous datasets from various service providers. In order to allow developers to create new applications on top of open datasets (see examples below), machine-readable formats are needed. As such, XML and JSON have quickly become the de facto format for the web and mobile applications due to their ease of integration into browser technologies and server technologies that support Javascript. Once the data has been acquired, the interlinking of diverse data sources is quite a complex and challenging process, especially for the acquired unstructured data. That is the reason why semantic technologies and Linked Data principles [50] have become popular over the last decade [221]. Using Linked Data principles and a set of agreed vocabularies for a domain, the input data is modeled in the form of resources, while the existing relationships are modeled as a set of (named) relationships between resources. In order to represent the knowledge of a specific domain, conceptual schemas are applied (also called ontologies). Automatic procedures are used to map the data to the target ontology, while standard languages are used to represent the mappings (see Chapter 4). Furthermore, in order to unify the knowledge representation and data processing, standardized hierarchical and multilingual schemas are used called taxonomies. Over the last decade, thousands of data reposito-

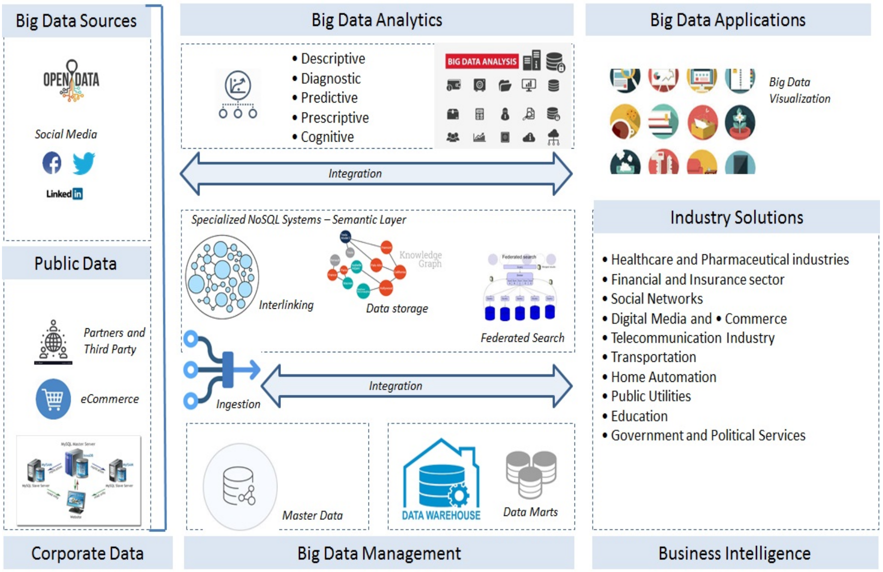

Fig. 1: From Data to Applications

ries emerged on the web[47] that companies can use to improve their products and/or processes. The public data sources (statistics, trends, conversations, images, videos, audios, and podcasts for instance from Google Trends, Twitter, Instagram, and others [297]) provide real-time information and on-demand insights that enable businesses to analyse user interactions, draw patterns and conclusions. IoT devices have also created significant challenges in many industries and enabled the development of new business models. However, one of the main challenges associated with these repositories is automatically understanding the underlying structures and patterns of the data. Such an understanding is a prerequisite to the application of advanced analytics to the retrieved data [142]. Examples of Open Data Sources from different domains are:

- **Facebook** Graph API, curated by Facebook, is the primary way for apps to read and write to the Facebook social graph. It is essentially a representation of all information on Facebook now and in the past. For more info see here[1].

- **Open Corporates** is one of the largest open databases of companies in the world and holds hundreds of millions of datasets in essentially any country. For more info, see here[2].

[1] https://developers.facebook.com/docs/graph-api
[2] https://opencorporates.com/

- **Global Financial Data**'s API is recommended for analysts who require large amounts of data for broad research needs. It enables researchers to study the interaction between different data series, sectors, and genres of data. The API supports R and Python so that the data can be directly uploaded to the target application. For more info, see here[3].

- **Open Street Map** is a map of the world, created by people free to use under an open license. It powers map data on thousands of websites, mobile apps, and hardware devices. For more info, see here[4].

- **The National Centers for Environmental Information** (NCEI) is responsible for hosting and providing access to one of the most significant archives on Earth, with comprehensive oceanic, atmospheric, and geophysical data. For more info about the data access, see here[5].

- **DBPedia** is a semantic version of Wikipedia. It has helped companies like Apple, Google, and IBM to support artificial intelligence projects. DBpedia is in the center of the Linked Data cloud presented in Figure 2, top-right quadrant[6]. For more info, see here[7].

Data Management. As data become increasingly available (from social media, web logs, IoT sensors etc.), the challenge of managing (selecting, combining, storing) and analyzing large and growing data sets is growing more urgent. From a data analytics point of view, that means that data processing has to be designed taking into consideration the diversity and scalability requirements of targeted data analytics applications. In modern settings, data acquisition via near real-time data streams in addition to batch loads is managed by different automated processes (see Figure 2, top-left quadrant presents an example of monitoring and control of electric power facilities with the Supervisory, Control and Data Acquisition Systems[8] developed by the Mihajlo Pupin Institute. The novel architecture [467] is 'flexible enough to support different service levels as well as optimal algorithms and techniques for the different query workloads' [424].

Over the last two decades, the emerging challenges in the design of end-to-end data processing pipelines were addressed by computer scientists and software providers in the following ways:

- In addition to operational database management systems (present on the market since 1970s), different **NoSQL stores** appeared that lack adherence to the time-honored SQL principles of ACID (atomicity, consistency, isolation, durability).

[3] https://www.globalfinancialdata.com/

[4] https://www.openstreetmap.org/

[5] https://www.ncdc.noaa.gov/data-access

[6] www.lod-cloud.net

[7] https://wiki.dbpedia.org/

[8] http://www.pupin.rs/en/products-services/process-management/

- **Cloud computing** emerged as a paradigm that focuses on sharing data and computations over a scalable network of nodes including end user computers, data centers (see Figure 2, bottom-left quadrant), and web services [23].
- The **Data Lake** concept as a new storage architecture was promoted where raw data can be stored regardless of source, structure and (usually) size. The *data warehousing* approach (based on a repository of structured, filtered data that has already been processed for a specific purpose) is thus perceived as outdated as it creates certain issues with respect to data integration and the addition of new data sources.

The wide availability of big data also means that there are many quality issues that need to be dealt with before using such data. For instance, data inherently contains a lot of noise and uncertainty or is compromised because of sensor malfunctioning or interferences, which may result in missing or conflicting data. Therefore, quality assessment approaches and methods applicable in open big data ecosystems have been developed [477].

Furthermore, in order to ensure interoperability between different processes and interconnected systems, the *semantic representation* of data sources / processes was introduced where a *knowledge graph*, from one side, meaningfully describes the data pipeline, and from the other, is used to generate new knowledge (see Chapter 4).

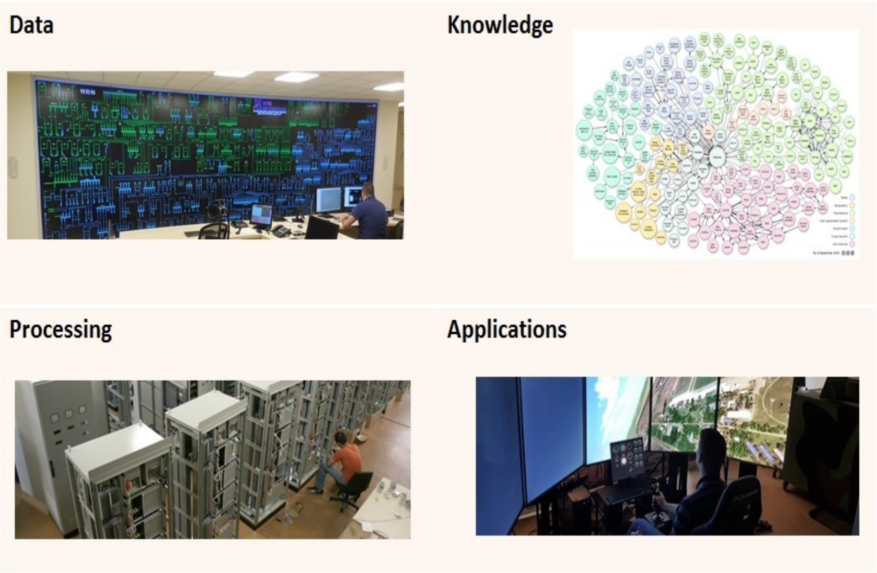

Fig. 2: Components of Big Data Ecosystem

Data Analytics. Data analytics refers to technologies that are grounded mostly in data mining and statistical analysis [75]. The selection of an appropriate processing model and analytical solution is a challenging problem and depends on the business issues of the targeted domain [220], for instance e-commerce [414], market intelligence, e-government [219], healthcare, energy efficiency [46], emergency management [307], production management, and/or security (see Figure 2, bottom-right quadrant, example of Simulators and training aids developed by the Mihajlo Pupin Institute). Depending on the class of problem that is being solved (e.g. risk assessment in banks and the financial sector, predictive maintenance of wind farms, sensing and cognition in production plants, automatic response in control rooms, etc.), the *data analytics* solution also relies on text / web / network / mobile analytical services. Here various machine learning techniques such as association rule mining, decision trees, regression, support vector machines, and others are used.

While simple reporting and business intelligence applications that generate aggregated measurements across different predefined dimensions based on the data-warehousing concept were enough in 1990s, since 1995 the focus has been on introducing parallelism into machine learning [433].

4 Using Semantics in Big Data Processing

Variety of Data Sources. In order to design and implement an adequate big data processing architecture, as well as volume and velocity companies also have to consider their ability to intercept the various available data sources. In addition to the existing enterprise resource management systems, data produced by a multitude of sources like sensors, smart devices and social media in raw, semi-structured, unstructured and rich media formats further complicate the processing and storage of data. Hence, different solutions for distributed storage, cloud computing, and data fusion are needed [284]. In order to make the data useful for data analysis, companies use different methods to reduce complexity, downsize the data scale (e.g. dimensional reduction, sampling, coding) and pre-process the data (data extraction, data cleaning, data integration, data transformation) [453]. The heterogeneity of data can thus be characterized across several dimensions:

- **Structural variety** refers to data representation; for instance, the satellite images format is very different from the format used to store tweets generated on the web;
- **Media variety** refers to the medium in which data gets delivered; for instance, the audio of a speech versus the transcript of the speech may represent the same information in two different media;
- **Semantic variety** refers to the meaning of the units (terms) used to measure or describe the data that are needed to interpret or operate on the data; for instance, a 'high' salary from a service in Ethiopia is very different from a 'high' salary from a similar service in the United States;

- **Availability variations** mean that the data can be accessed continuously; for instance, from traffic cameras, or intermediately, for instance, only when the satellite is over the region of interest.

Semantic variety and the need for standards. Attempts to explain the uses of semantics in logic and computing date from the middle of the last century. In the information processing domain, *semantics* refers to the "meaning and meaningful use of data" [468], i.e., the effective use of a data object for representing a concept or object in the real world. Since 1980, the Artificial Intelligence community has been promoting the idea of feeding intelligent systems and agents with general, formalized knowledge of the world (see also the panel report from 1997 *Data Semantics: what, where and how?*) [396]. In 2001, Sir Tim Berners-Lee, the Director of the Wide Web Consortium, outlined his vision for the Semantic Web as an extension of the conventional Web and as a world-wide distributed architecture where data and services easily interoperate. Additionally, in 2006, Berners-Lee proposed the basic (Linked Data) principles for interlinking linking datasets on the Web through references to common concepts [50]. The standard for the representation of the information that describes the concepts is RDF (Resource Description Framework). In parallel, the wider adoption of standards for representing and querying semantic information, such as RDF(s) and SPARQL, along with increased functionalities and improved robustness of modern RDF stores, have established Linked Data and semantic technologies in the areas of data and knowledge management. As part of the EC 'Interoperability Solutions for European Public Administrations' (ISA, http://ec.europa.eu/isa/ and https://ec.europa.eu/isa2/) program, with cooperation with W3C, core vocabularies have been adopted to represent high-value datasets relevant for boosting innovative services.

Knowledge engineering. Additionally, the scientific community has put a great deal of effort into showcasing how knowledge engineering [91], [220], [25] can take advantages from semantics-aware methods [221], which exploit knowledge kept in (big) data to better reasoning on data beyond the possibilities offered by more traditional data-instance-oriented approaches. With the announcement of the Google Knowledge Graph in 2012, representations of general world knowledge as graphs have drawn a lot of attention again [345].

To summarize, semantics principles can be used in big data processing for

- Representing (schema and schema-less) data;
- Representing metadata (about documentation, provenance, trust, accuracy, and other quality properties);
- Modeling data processes and flows, i.e., representing the entire pipeline making data representation shareable and verifiable.

The semantic representation of data in knowledge graphs (see Chapter 2), the semantic processing pipeline (see Chapter 3, Chapter 5, Chapter 8), reasoning in knowledge graphs (Chapter 6) and the semantic analysis of big data (Chapter 7) are the main topics of this book and will be explained in more detail in the subsequent chapters.

5 Big Data, Standards and Interoperability

Interoperability remains a major burden for the developers of the big data ecosystem. In its EU 2030 vision, the European Union has set out the creation of an internal single market through a standardised system of laws that apply in all member states and a single European data [84] space – a genuine single market for data where businesses have easy access to an almost infinite amount of high-quality industrial data. The vision is also supported by the EU Rolling Plan for ICT Standardisation [85] that identifies 170 actions organised around five priority domains — 5G, cloud, cybersecurity, big data and Internet of Things. In order to enable broad data integration, data exchange and interoperability with the overall goal of fostering innovation based on data, standardisation at different levels (such as metadata schemata, data representation formats and licensing conditions of open data) is needed. This refers to all types of (multilingual) data, including both structured and unstructured data, and data from different domains as diverse as geospatial data, statistical data, weather data, public sector information (PSI) and research data, to name just a few.

In the domain of big data, five different actions have been requested that also involve the following standardization organizations:

- CEN, the European Committee for Standardization, to support and assist the standardisation process and to coordinate with the relevant W3C groups on preventing incompatible changes and on the conditions for availability of the standard(s). The work will be in particular focused on the interoperability needs of data portals in Europe while providing semantic interoperability with other applications on the basis of reuse of established controlled vocabularies (e.g. EuroVoc) and mappings to existing metadata vocabularies (e.g. SDMX, INSPIRE metadata, Dublin Core, etc.);

- CENELEC (the European Committee for Electrotechnical Standardization) in particular in relation to personal data management and the protection of individuals' fundamental rights;

- ETSI (the European Telecommunications Standards Institute) to coordinate stakeholders and produce a detailed map of the necessary standards (e.g. for security, interoperability, data portability and reversibility) and together with CEN to work on various standardisation deliverables needed for the completion of the rationalised framework of e-signatures standards;
- IEEE has a series of new standards projects related to big data (mobile health, energy-efficient processing, personal agency and privacy) as well as pre-standardisation activities on big data and open data;

- ISO/IEC JTC1, WG 9 — Big Data, formed at the November 2014 in relation to requirements, use cases, vocabulary and a reference architecture for big data;

- OASIS, in relation to querying and sharing data across disparate applications and multiple stakeholders for reuse in enterprise, cloud, and mobile devices. Specification development in the OASIS OData TC builds on the core OData Protocol V4 released in 2014 and addresses additional requirements identified as extensions in four directional white papers: data aggregation, temporal data, JSON documents, and XML documents as streams;

- OGC, the Open Geospatial Consortium defines and maintains standards for location-based, spatio-temporal data and services. The work includes, for instance, schema allowing descriptions of spatio-temporal sensors, images, simulations, and statistics data (such as "datacubes"), a modular suite of standards for Web services allowing ingestion, extraction, fusion, and (with the web coverage processing service (WCPS) component standard) analytics of massive spatio-temporal data like satellite and climate archives. OGC also contributes to the INSPIRE project;

- W3C, the W3C Semantic Web Activity Group has accepted numerous Web technologies as standards or recommendations for building semantic applications including RDF (Resource Description Framework) as a general-purpose language; RDF Schema as a meta-language or vocabulary to define properties and classes of RDF resources; SPARQL as a standard language for querying RDF data: OWL, Web Ontology Language for effective reasoning. More about semantic standards can be found in [222].

6 Big Data Analytics

6.1 The Evolution of Analytics

Over the last 50 years, **Data Analytics** has emerged as an important area of study for both practitioners and researchers. The **Analytics 1.0** era began in the 1950s and lasted roughly 50 years. As a software approach, this field evolved significantly with the invention of Relational Databases in the 1970s by Edgar F. Codd, the development of artificial intelligence as a separate scientific discipline, and the invention of the Web by Sir Tim Berners-Lee in 1989. With the development of Web 2.0-based social and crowd-sourcing systems in the 2000s, the Analytics 2.0 era started. While the business solutions were tied to relational and multi-dimensional database models in the Analytics 1.0 era, the **Analytics 2.0** era brought NOSQL and big data database models that opened up new priorities and technical possibilities for analyzing large amounts of semi-structured and unstructured data. Companies and data scientists refer to these two periods in time as before big data (BBD) and after big data (ABD) [99]. The main limitations observed during the first era were that the potential capabilities of data were only utilised within organisations, i.e. the business intelligence activities addressed only what had happened in the past and offered no predictions about its future trends. The new generation of tools with fast-processing engines and NoSQL stores made possible the integration of internal data with externally sourced data coming from the internet, sensors of various types, public data initiatives (such as the human genome project), and captures of audio and video recordings. Also significantly developed in this period was the Data Science field (multifocal field consisting of an intersection of Mathematics& Statistics, Computer Science, and Domain Specific Knowledge), which delivered scientific methods, exploratory processes, algorithms and tools that can be easily leveraged to extract knowledge and insights from data in various forms.

The **Analytics 3.0** era started [23] with the development of the "Internet of Things" and cloud computing, which created possibilities for establishing hybrid technology environments for data storage, real-time analysis and intelligent customer-oriented services. Analytics 3.0 is also named *the Era of Impact* or *the Era of Data-enriched offerings* after the endless opportunities for capitalizing on analytics services. For creating value in the data economy, Davenport [99] suggests that the following factors need to be properly addressed:

- combining multiple types of data
- adoption of a new set of data management tools
- introduction of new "agile" analytical methods and machine-learning techniques to produce insights at a much faster rate
- embedding analytical and machine learning models into operational and decision processes
- requisite skills and processes to work with innovative discovery tools for data exploration
- requisite skills and processes to develop prescriptive models that involve large-scale testing and optimization and are a means of embedding analytics into key processes
- leveraging new approaches to decision making and management

Nowadays, being in the **Analytics 4.0** era or *the Era of Consumer-controled data*, the goal is to enable the customers to have full or partial control over data. Also aligned with the Industry 4.0 movement, there are different possibilities for automating and augmenting human/computer communications by combining machine translation, smart reply, chat-bots, and virtual assistants.

6.2 Different Types of Data Analytics

In general, analytic problems and techniques can be classified into

- **Descriptive** - What happened?
- **Diagnostic** - Why did it happen?
- **Predictive** - What is likely to happen?
- **Prescriptive** - What should be done about it?
- **Cognitive** - What don't we know?

Descriptive analytics focus on analyzing historic data for the purpose of identifying patterns (*hindsights*) or trends. While statistical theory and descriptive methodologies [7] are well documented in scientific literature, that is not the case for other types of analytics, especially observing the big data and cloud computing context.

Diagnostic analytics [362] discloses the root causes of a problem and gives *insight*. The methods are treated as an extension to descriptive analytics that provide an explanation to the question "Why did it happen?".

Predictive analytics-based services apply forecasting and statistical modelling to give insight into "what is likely to happen" in the future (*foresight*) based on supervised, unsupervised, and semi-supervised learning models.

Prescriptive analytics-based services [279] answers the question "What should I do?". In order to provide automated, time-dependent and optimal decisions based on the provided constraints and context, the software tools utilize artificial intelligence, optimization algorithms and expert systems approaches.

Cognitive analytics is a term introduced recently in the context of cognitive computing (see also *Deloitte Tech Trends 2019*). Motivated by the capability of the human mind, and other factors such as changing technologies, smart devices, sensors, and cloud computing capabilities, the goal is to develop "AI-based services that are able to interact with humans like a fellow human, interpret the contextual meaning, analyze the past record of the user and draw deductions based on that interactive session" [174], [176].

7 Challenges for Exploiting the Potential of Big Data

In order to exploit the full potential, big data professionals and researchers have to address different data and infrastructure management challenges that cannot be resolved with traditional approaches [71]. Hence, in the last decade, different techniques have emerged for acquisition, storing, processing and information derivation in the big data value chains.

In [402], the authors introduced three main categories of challenges as follows:

- **Data challenges** related to the characteristics of the data itself (e.g. data volume, variety, velocity, veracity, volatility, quality, discovery and dogmatism);
- **Process challenges** related to techniques (how to capture data, how to integrate data, how to transform data, how to select the right model for analysis and how to provide the results);
- **Management challenges** related to organizational aspects such as privacy, security, governance and ethical aspects.

Data, process and management challenges are interlinked and influence each other.

7.1 Challenges

The *3 Vs of big data* call for the integration of complex data sources (including complex types, complex structures, and complex patterns), as previously discussed. Therefore, *scalability* is considered to be a crucial bottleneck of big data solutions. Following the problem with processing, *storage* management is another unavoidable barrier regarding big data. Storing the huge quantity of data between its acquisition, processing and analysis requires gigantic memory capacity, thus rendering traditional solutions obsolete.

The inherent complexity of big data (*data complexity*) makes its perception, representation, understanding and computation far more challenging and results in sharp increases in the *computational complexity* required compared to traditional computing models based on total data. The design of system architectures, computing frameworks, processing modes, and benchmarks for highly energy-efficient big data processing platforms is the key issue to be addressed in *system complexity* [230]. Contemporary cloud-based solutions are also considered to be on the edge of feasibility since *responsiveness* can be a critical issue, especially in real-time applications, where upload speeds are considered the main bottleneck.

When simultaneously working with different data sources, the *reliability* of collected data will inevitably fluctuate with missed, partial and faulty measurements being unavoidable, resulting in serious potential trouble later on in the workflow, such as in the analytics stage. Hence, high-quality data management (i.e. data cleaning, filtering, transforming and other) actions are mandatory at the beginning of the process. Besides reliability, the *correctness* of the data is considered to be a key aspect of big data processing. High volumes, unstructured forms, the distributed nature of data in NoSQL data management systems and the necessity of near-to-real-time responses often lead to corrupted results with no method being able to guarantee their complete *validity*.

Other *quality* dimensions, that impact the design of a big data solution are *completeness*, *consistency*, *credibility*, *timeliness* and others.

For instance, in real-time applications (e.g. stock market, financial fraud detection and transactions parsing, traffic management, energy optimization etc.), quick responses are required and expected immediately because the retrieved information can be completely useless if it is derived with high latency with respect to the collected data.

An additional challenge from the human-computer perspective is the *visualization* of results. Although various ways in which the data can be displayed do not affect the data processing segment in any way, visualization is stated in the literature as a crucial factor because without adequate representation of the results, the derived knowledge is useless.

Depending on the type of data being processed, *security* can sometimes be a crucial component that requires special attention. When considering, for example, a weather forecast or public transport management use case, if a data loss or theft occurs, it can be considered practically irrelevant compared to a situation where personal information, names, addresses, location history, social security information or credit card PIN codes are stolen because in the latter case, data protection must be upheld at the highest possible standard.

7.2 Example: Analysis of Challenges and Solutions for Traffic management

Smart transportation is one of the key big data vertical applications, and refers to the integrated application of modern technologies and management strategies in transportation systems. Big data platforms available on the market contribute to a great extent to smart management of cities and the implementation of intelligent transportation systems. In order to showcase the usage of different type of data analytics and to strengthen the discussion on challenges, we will point to the traffic management system used for monitoring highways in Serbia [364]. Highways and motorways control systems generate a high volume of data that is relevant for a number of stakeholder,s from traffic and environmental departments to transport providers, citizens and the police. The figure below points to (a) the European corridors, and (b) the Corridor 10 that is managed in Serbia by the public enterprise "Roads of Serbia" using a control system provided by Institute Mihajlo Pupin. Its holistic supervisory function and control includes (a) toll collection and motorway and highway traffic control, and (b) urban traffic control and management. The main challenges on EU level are related to:

- **Interoperability** of tolling services on the entire European Union road network because the ones introduced at local and national levels from the early 1990s onwards are still generally non-interoperable;
- **Smart mobility** and the need of users to be more informed about different options in real-time;
- the need for efficient and effective approaches for **assessment and management of air pollution** due to improved ambient air quality.

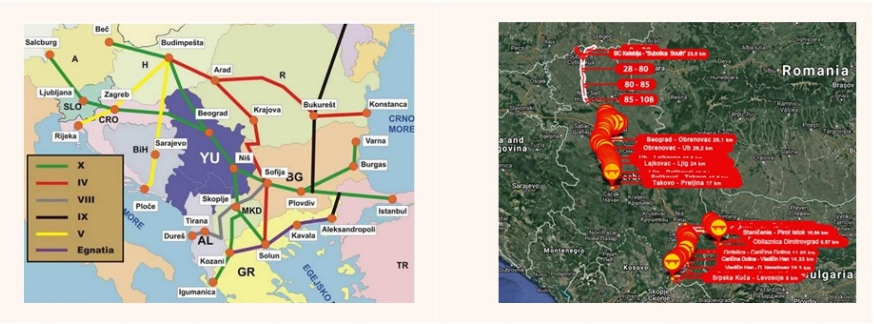

Fig. 3: Traffic Management

The main components of the traffic control system are:

- The toll **collection system**[9], which is hierarchically structured; it is fully modular, based on PC technology and up-to date real time operation sys-

[9] http://www.pupin.rs/en/products-services/traffic-control-systems/pay-toll/

tems, relational data base system and dedicated encryption of data transmission. Toll line controllers are based on industrial PC-technology and dedicated electronic interface boards. The toll plaza subsystem is the supervisory system for all line controllers. It collects all the data from lane controllers including financial transactions, digital images of vehicles, technical malfunctions, line operators' actions and failures. All data concerning toll collection processes and equipment status are permanently collected from the plaza computers and stored in a central system database. The toll collection system also comprises features concerning vehicle detection and classification, license plate recognition and microwave-based dedicated short-range communications.

- The Main **Control Centre** is connected through an optical communication link with the Plaza Control Centres. Also, the Control Centre is constantly exchanging data with various institutions such as: banks, insurance companies, institutions that handle credit and debit cards, RF tags vendors, etc. through a computer network. **Data analytics** is based on data warehouse architecture enabling optimal performances in near real time for statistical and historical analysis of large data volumes. Reporting is based on optimized data structures, allowing both predefined (standardized) reports as well as ad hoc (dynamic) reports, which are generated efficiently using the Oracle BI platform. Data analytics includes scenarios, such as

    - **Predicting and preventing road traffic congestion** analytics is used to improve congestion diagnosis and to enable traffic managers to proactively manage traffic and to organize the activities at toll collection stations before congestion is reached.
    - **Strategic environmental impact assessment** analytics is used to study the environmental impact and the effect of highways on adjacent flora, fauna, air, soil, water, humans, landscape, cultural heritage, etc. based on historical and real-time analysis. Passive pollution monitoring involves collecting data about the diffusion of air pollutants, e.g. emission estimates based on traffic counting. Passive pollution monitoring has been used to determine trends in long-term pollution levels. Road traffic pollution monitoring and visualization requires the integration of high volumes of (historical) traffic data with other parameters such as vehicle emission factors, background pollution data, meteorology data, and road topography.

Here, we have pointed to just one mode of transport and traffic management, i.e. the control of highways and motorways. However, nowadays, an increasing number of cities around the world struggle with traffic congestion, optimizing public transport, planning parking spaces, and planning cycling routes. These issues call for new approaches for studying human mobility by exploiting machine learning techniques [404], forecasting models or through the application of complex event processing tools [134].

8 Conclusions

This chapter presents the author's vision of a *Big data ecosystem*. It serves as an introductory chapter to point to a number of aspects that are relevant for this book. Over the last two decades, advances in hardware and software technologies, such as the Internet of Things, mobile technologies, data storage and cloud computing, and parallel machine learning algorithms have resulted in the ability to easily acquire, analyze and store large amounts of data from different kinds of quantitative and qualitative domain-specific data sources. The monitored and collected data presents opportunities and challenges that, as well as focusing on the three main characteristics of volume, variety, and velocity, require research of other characteristics such as validity, value and vulnerability. In order to automate and speed up the processing, interoperable data infrastructure is needed and standardization of data-related technologies, including developing metadata standards for big data management. One approach to achieve interoperability among datasets and services is to adopt data vocabularies and standards as defined in the W3C Data on the Web Best Practices, which are also applied in the tools presented in this book (see Chapters 4 to 9).

In order to elaborate the challenges and point to the potential of big data, a case study from the traffic sector is presented and discussed in this chapter, while more big data case studies are set out in Chapter 9 and Chapter 10.

Table 3: History of Big Data

Year	Description
1911	Computing-Tabulating-Recording Company was founded and renamed "International Business Machines" (**IBM**) in 1924
1929	The term "Super Computing" was first used in the New York World to refer to large custom-built tabulators that **IBM** had made for Columbia University
1937	Social security was introduced in the United States of America and the requirement arose for data management of 26 million residents
1945	John Von Neumann published a paper on the Electronic Discrete Variable Automatic Computer (EDVAC), the first "documented" discussion on program storage, and laid the foundations of computer architecture today
1957	A group of engineers established the Control Data Corporation (**CDC**) in Minneapolis, Minnesota
1960	Seymour Cray (**CDC**) completed the CDC 1604, one of the first solid-state computers, and the fastest computer in the world at a time when vacuum tubes were found in most large computers
1965	The first data center in the world was planned
1969	ARPANET set a message was sent from UCLA's host computer to Stanford's host computer
1970	Edgar Frank Codd invented the relational model for database management
1976	**SAS** Institute delivered the first version of the "Statistical Analysis System"
1977	**Oracle** Corporation was founded in Santa Clara, California, U.S.
1998	**Google** was founded at the Stanford University in California
1999	**Apache** software foundation was established
1989	The invention of the World Wide Web at CERN
2003	**Google** File System was invented
2004	World Wide Web Consortium (**W3C**), the main international standards organization for the Web was founded
2005	The start of development on **Apache** Hadoop which came into production in 2008
2007	The first publicly available dataset on **DBpedia** was published by the Free University of Berlin and the Leipzig University
2009	**Yahoo** released Pig and Facebook created Hive
2011	Start of real-time processing as opposed to batch processing with **Apache** Storm and Spark
2012	Creation of Kafka by **LinkedIn**, **Google** introduced its Knowledge Graph project
2013	The definition of the Lambda architecture for efficient big data processing
2014	The definition of the Kappa architecture and the beginning of hybrid data processing