Chapter 1

Big Data Technologies in DataBio

Caj Södergård, Tomas Mildorf, Arne J. Berre, Aphrodite Tsalgatidou,
and Karel Charvát

Abstract In this introductory chapter, we present the technological background needed for understanding the work in DataBio. We start with basic concepts of Big Data including the main characteristics volume, velocity and variety. Thereafter, we discuss data pipelines and the Big Data Value (BDV) Reference Model that is referred to repeatedly in the book. The layered reference model ranges from data acquisition from sensors up to visualization and user interaction. We then discuss the differences between open and closed data. These differences are important for farmers, foresters and fishermen to understand, when they are considering sharing their professional data. Data sharing is significantly easier, if the data management conforms to the FAIR principles. We end the chapter by describing our DataBio platform that is a software development platform. It is an environment in which a piece of software is developed and improved in an iterative process providing a toolset for services in agriculture, forestry and fishery. The DataBio assets are gathered on the DataBio Hub that links to content both on the DataBio website and to Docker software repositories on clouds.

1.1 Basic Concepts of Big Data

When we want to utilize data and computers to make raw material gathering more efficient and sustainable in bioeconomy, we will have to deal with vast amounts of heterogeneous data at high speeds, i.e. *Big Data*. This is because of the enormous

C. Södergård (✉)
VTT Technical Research Centre of Finland, Espoo, Finland
e-mail: Caj.Sodergard@vtt.fi

T. Mildorf
University of West Bohemia, Pilsen, Czech Republic

A. J. Berre · A. Tsalgatidou
SINTEF Digital, Forskningsvejen 1, Oslo, Norway

K. Charvát
Lesprojekt-slужby Ltd., Zaryby, Czech Republic

Fig. 1.1 Global data sphere grows exponentially. *Source* [3]

and all the time increasing flow of data from a variety of sensors and measurement devices, like cameras on satellites, aeroplanes and drones as well as measurement data from sensors in the air, in the soil and in the oceans. Moreover, the resolution and frequency of data acquisition from those sensors are exponentially increasing. Many industrial sectors benefit from Big Data, which were coined “the new oil” [1]. The term Big Data has been in use since 2001, when Doug Laney introduced the 3V characteristics: *volume*, *velocity* and *variety* [2]. The 3V’s have the following meanings:

Volume is the amount of generated data. The global data sphere grows exponentially (Fig. 1.1). IDC has predicted that it will grow from 45 ZettaBytes ($=10^{21}$ bytes) in 2020 till 175 ZettaBytes in 2025 [3]. This is mainly due to the growth in unstructured data, like multimedia (audio, images and video) as well as social media content. This puts a lot of pressure on Big Data technologies.

Velocity is the speed of generating and processing data. The development has gone from batch, periodic, near real time, to fully real time/streaming data, which requires a massive throughput.

Variety is the type of generated data (text, tables, images, video, etc.). Unstructured data is more and more dominating over semi-structured and unstructured data. The issue is to manage the heterogeneity of data.

Later, the Big Data concept has expanded with more V dimensions. Data has both social and economic *values*. Value is typically extracted from data with analytical methods, including predictive analytics, visualization and artificial intelligence tools. *Variability* refers to changes in data rate, format/structure, semantics, and/or quality that impact the supported application, analytic or problem [4]. Impacts can include the need to refactor architectures, interfaces, processing/algorithms, integration/fusion, storage, applicability or use of the data. Finally, *veracity* refers to the noise, biases and abnormality in Big Data. There is always the need of checking if the available data is relevant to the problem being studied.

Data *quality* is central in all processing. With low quality in the input data, we will get uncertain results out. *Metadata* (*=data about data*) allows identification of information resources. Metadata is needed for describing among other things

data types, geographic extent and temporal reference, quality and validity, interoperability of spatial data sets and services, constraints related to access and use, and the organization responsible for the resource.

In the DataBio project, the data handling specifically aimed at the following sectors:

- **Agriculture:** The main goal was to develop smart agriculture solutions that boost the production of raw materials for the agri-food chain in Europe while making farming sustainable. This includes optimized irrigation and use of fertilizers and pesticides, prediction of yield and diseases, identification of crops and assessment of damages. Such smart agriculture solutions are based on the use data from satellites, drones, IoT sensors, weather stations as well as genomic data.
- **Forestry:** Big Data methods are expected to bring the possibility to both increase the value of the forests as well as to decrease the costs within sustainability limits set by natural growth and ecological aspects. The key technology is to gather more and more accurate information about the trees from a host of sensors including new generations of satellites, drones, laser scanning from aeroplanes, crowdsourced data collected from mobile devices and data gathered from machines operating in forests.
- **Fisheries:** The ambition is to herald and promote the use of Big Data analytical tools to improve the ecological and economic sustainability, such as improved analysis of operational data for engine fault detection and fuel reduction, tools for planning and operational choices for fuel reduction when searching and choosing fishing grounds, as well as crowdsourcing methods for fish stock estimation.

1.2 Pipelines and the BDV Reference Model

When processing streaming time-dependent data from sensors, data is put to travel through *pipelines*. The term pipeline was used in the DataBio project to describe the data processing steps. Each step has its input and output data. A pipeline is created by chaining individual steps in a consecutive way, where the output from the preceding processing step is fed into the succeeding step. Typically in Big Data applications, the pipeline steps include data gathering, processing, analysis and visualization of the results. The US National Institute of Standards NIST describes this process in their Big Data Interoperability Framework [5]. In DataBio, we call these steps for a *generic pipeline* (Fig. 1.2). This generic pipeline is adapted to the agricultural,

Fig. 1.2 Top-level generic pipeline

Fig. 1.3 DataBio project structured technologies as vertical pipelines crossing the horizontal layers in the BDV Reference Model

forestry and fisheries domains.

In order to describe the Big Data Value chains in more detail, the Big Data Value (BDV) Reference Model was adopted in DataBio (Fig. 1.3). The BDV Reference Model has been developed by the industry-led Big Data Value Association (BDVA). This model takes into account input from technical experts and stakeholders along the whole Big Data Value chain, as well as interactions with other industrial associations and with the EU. The BDV Reference Model serves as a common reference framework to locate Big Data technologies on the overall IT stack. It addresses the main concerns and aspects to be considered for Big Data Value systems in different industries. The BDV Reference Model is compatible with standardized reference architectures, most notably the emerging standards ISO JTC1 SC42 AI and Big Data Reference Architecture.

The steps in the generic pipeline and the associated layers in the reference model are:

Data acquisition from things, sensors and actuators: This layer handles the interface with the data providers and includes the transportation of data from various sources to a storage medium where it can be accessed, used and analysed. A main source of Big Data is sensor data from an IoT context and actuator interaction in cyberphysical systems. Tasks in this layer, depending on the type of collected data and on application implementation, include accepting or performing specific collections of data, pulling data or receiving pushes of data from data providers and storing or buffering data. Initial metadata can also be created to facilitate subsequent aggregation or look-up methods. Security and privacy considerations can also be included in this step, since authentication and authorization activities as well as recording and maintaining data provenance activities are usually performed during data collection.

Cloud, high performance computing (HPC) and data management: Effective Big Data processing and data management might imply the effective usage of cloud and HPC platforms. Traditional relational databases (RDB) do not typically scale well, when new machines are added to handle vast amounts of data. They are also not especially good at handling unstructured data like images and video. Therefore, they are complemented with non-relational databases like key-store, column-oriented, document and graph databases [6]. Of these, column-oriented architectures are used, e.g. in the Apache Cassandra and Hbase software for storing big amounts of data. Document databases have seen an enormous growth in recent years. The most used document database recently is MongoDB, that also was used in the DataBio project, e.g. in the DataBio Hub for managing the project assets and in the GeoRocket database component.

Data preparation: Tasks performed in this step include data validation, like checking formats, data cleansing, such as removing outliers or bad fields, extraction of useful information and organization and integration of data collected from various sources. In addition, the tasks consist of leveraging metadata keys to create an expanded and enhanced dataset, annotation, publication and presentation of the data to make it available for discovery, reuse and preservation, standardization and reformatting, as well as encapsulating. Source data is frequently persisted to archive storage and provenance data is verified or associated. Optimization of data through manipulations, like data deduplication and indexing, can also be included here.

Data processing and protection: The key to processing Big Data volumes with high throughput, and sometimes, complex algorithms is arranging the computing to take place *in parallel*. Hardware for parallel computing comprises 10, 100 or several thousands processors, often collected into graphical processing unit (GPU) cards. GPUs are used especially in machine learning and visualization. Parallelizing is straightforward in image and video processing, where the same operations typically are applied to various parts of the image. Parallel computing on GPU's is used in DataBio, e.g. for visualizing data. *Data protection* includes privacy and anonymization mechanisms to facilitate protection of data. This is positioned between data management and processing, but it can also be associated with the area of cybersecurity.

Data analytics: In this layer, new patterns and relationships are discovered to provide new insights. The extraction of knowledge from the data is based on the requirements of the vertical application, which specify the data processing algorithms. Data analytics is a crucial step as it gives suggestions and makes decisions. Hashing, indexing and parallel computing are some of the methods used for Big Data analysis. Machine learning techniques and other artificial intelligence methods are also used in many cases.

Analytics utilize data both from the past and from the present.

- Data from the **past** is used for descriptive and diagnostic analytics, and classical querying and reporting. This includes performance data, transactional data, attitudinal data, behavioural data, location-related data and interactional data.

- Data from the **present** is harnessed in monitoring and real-time analytics. This requires fast processing many times handling data in real-time, for triggering alarms, actuators, etc.
- Harnessing data for the **future** includes prediction and recommendation. This typically requires processing of large data volumes, extensive modelling as well as combining knowledge from the past and present, to provide insight for the future.

Data visualization and user interaction: Visualization assists in the interpretation of data by creating graphical representations of the information conveyed. It thus adds more value to data as the human brain digests information better, when it is presented in charts or graphs rather than on spreadsheets or reports. In this way, users can comprehend large amounts of complex data, interact with the data, and make decisions. Effective data visualization needs to keep a balance between the visuals it provides and the way it provides them so that it attracts users' attention and conveys the right messages.

In the book chapters that follow, the above steps have been specialized based on the different data types used in the various project pilots. Solutions are set up according to different processing architectures, such as batch, real-time/streaming or interactive. See e.g. the pipelines for

- the real-time IoT data processing and decision-making in Chaps. 3 and 11,
- linked data integration and publication in Chap. 8,
- data flow in genomic selection and prediction in Chap. 16,
- farm weather insurance assessment in Chap. 19,
- data processing of Finnish forest data in Chap. 23.
- forest inventory in Chap. 24.

Vertical topics, that are relevant for all the layers in the reference model in Fig. 1.3, are:

- *Big Data Types and Semantics:* 6 Big Data types are identified, based on the fact that they often lead to the use of different techniques and mechanisms in the horizontal layers: (1) structured data; (2) time series data; (3) geospatial data; (4) media, image, video and audio data; (5) text data, including natural language processing data and genomics representations; and (6) graph data, network/web data and metadata. In addition, it is important to support both the syntactic and semantic aspects of data for all Big Data types.
- *Standards:* Standardization of Big Data technology areas to facilitate data integration, sharing and interoperability. Standards are advanced at many fora including communities like BDVA, and W3C as well as standardization bodies like ISO and NIST.
- *Communication and Connectivity:* Effective communication and connectivity mechanisms are necessary in providing support for Big Data. Especially important is wireless communication of sensor data. This area is advanced in various

communication communities, such as the 5G community as well as in telecom standardization bodies.

- *Cybersecurity*: Big Data often needs support to maintain security and trust beyond privacy and anonymization. The aspect of trust frequently has links to trust mechanisms such as blockchain technologies, smart contracts and various forms of encryption.
- *Engineering and DevOps* for building Big Data Value systems: In practise, the solutions have to be engineered and interfaced to existing legacy IT systems and feedback gathered about their usage. This topic is advanced especially in the Networked European Software and Service Initiative NESSI.
- *Marketplaces, Industrial Data Platforms (IDP) and Personal Data Platforms (PDPs), Ecosystems for Data Sharing and Innovation Support*: Data platforms include in addition to IDPs and PDPs, also Research Data Platforms (RDPs) and Urban/City Data Platforms (UDPs). These platforms facilitate the efficient usage of a number of the horizontal and vertical Big Data areas, most notably data management, data processing, data protection and cybersecurity.

1.3 Open, Closed and FAIR Data

Open and closed data

Open data means that data is freely available to everyone to use and republish, without restrictions from copyright, patents or other limiting mechanisms [7]. The access to closed datasets is restricted. Data is closed because of policies of data publishers and data providers. Closed data can be private data and/or personal data, valuable exploitable data, business or security sensitive data. Such data is usually not made accessible to the rest of the world. Data sharing is the act of certain entities (e.g. people) passing data from one to another, typically in electronic form [8].

Data sharing is central for bioeconomy solutions, especially in agriculture. At the same time, farmers need to be able to trust that their data is protected from unauthorized use. Therefore, it is necessary to understand that sharing data is different from the open data concept. Shared data can be closed data based on a certain agreement between specific parties, e.g. in a corporate setting, whereas open data is available to anyone in the public domain. Open data may require attribution to the contributing source, but still be completely available to the end user.

Data is constantly being shared between employees, customers and partners, necessitating a strategy that continuously secures *data stores* and users. Data moves among a variety of public and private storage locations, applications and operating environments and is accessed from different devices and platforms. That can happen at any stage of the data security lifecycle, which is why it is important to apply the right security controls at the right time. Trust of data owners is a key aspect for data sharing.

Generally, open data differs from closed data in three ways (see, e.g. www.opendatasoft.com).

1. Open data is accessible, usually via a data warehouse on the internet.
2. It is available in a readable format.
3. It is licenced as open source, which allows anyone to use the data or share it for non-commercial or commercial gain.

Closed data restricts access to the information in several potential ways:

1. It is only available to certain individuals within an organization.
2. The data is patented or proprietary.
3. The data is semi-restricted to certain groups.
4. Data that is open to the public through a licensure fee or other prerequisite.
5. Data that is difficult to access, such as paper records, that have not been digitized.

Examples of closed data are information that requires a security clearance; health-related information collected by a hospital or insurance carrier; or, on a smaller scale, your own personal tax returns.

FAIR data and data sharing

The FAIR data principles (Findable, Accessible, Interoperable, Reusable) ensure that data can be discovered through catalogues or search engines, is accessible through open interfaces, is compliant to standards for interoperable processing of that data and therefore can be easily reused also for other purposes than it was initially created for [9]. This reuse improves the cost-balance of the initial data production and allows cross-fertilization across communities. The FAIR principles were adopted in DataBio through its data management plan [10].

1.4 The DataBio Platform

An application running on a Big Data platform can be seen as a pipeline consisting of multiple components, which are wired together in order to solve a specific Big Data problem (see <https://www.big-data-europe.eu/>). The components are typically packaged in Docker containers or code libraries, for easy deployment on multiple servers. There are plenty of commercial systems from known vendors like Microsoft, Amazon, SAP, Google and IBM that market themselves as Big Data platforms. There are also open-source platforms like Apache Hadoop for processing and analysing Big Data.

The DataBio platform was not designed as a monolithic platform; instead, it combines several existing platforms. The reasons for this were several:

- The project sectors of agriculture, forestry and fishery are very diverse and a single monolithic platform cannot serve all users sufficiently well.
- It is unclear who would take the ownership of such a new platform and maintain and develop it after the project ends.

- Several consortium partners had already at the outset of the project their own platforms. Therefore, DataBio should not compete with these partners by creating a new separate platform or by building upon a certain partner platform.
- Platform interoperability (public/private), data and application sharing were seen as more essential than creating yet another platform.

The DataBio platform should be understood in a strictly technical sense as a *software development platform* [11]. It is an environment in which a piece of software is developed and improved in an iterative process where after learning from the tests and trials, the designs are modified, and a new circle starts (Fig. 1.4). The solution is finally deployed in hardware, virtualized infrastructure, operating system, middleware or a cloud. More specifically, DataBio produced a Big Data *toolset* for *services* in agriculture, forestry and fishery [12]. The toolset enables new software *components* to be easily and effectively combined with open-source, standard-based, and proprietary components and infrastructures. These combinations typically form reusable and deployable *pipelines* of interoperable components.

The DataBio sandbox uses as resources mainly the *DataBio Hub*, but also the project *web site* and *deployed software* on public and private clouds [13]. The Hub links to content both on the DataBio website (deliverables, models) and to the Docker repositories on various clouds. This environment has the potential to make it easier and faster to design, build and test digital solutions for the bioeconomy sectors in future.

The *DataBio Hub* (<https://databiohub.eu/>) helps to manage the DataBio project assets, which are pilot descriptions and results, software components, interfaces, component pipelines, datasets, and links to deliverables and Docker modules (Fig. 1.5). The Hub has helped the partners *during* the project and has the potential to guide third party developers *after* the project in integrating DataBio assets into

Platform offers resources for iterative Sandbox development

Fig. 1.4 Platform developed in DataBio consists of a network of resources for the interactive development of bioeconomy applications

Fig. 1.5 DataBio Hub provides searchable information on the assets developed in DataBio and helps the external developer to develop their own applications

new digital services for the bioeconomy sectors. The service framework at the core of the Hub is available as open source on [GitHub](https://github.com/digitalservices/hub/serviceregistry) (<https://github.com/digitalservices/hub/serviceregistry>).

We identified 95 components, mostly from partner organizations, that could be used in the pilots. They covered all layers of the previously mentioned BDVA Reference Model (Fig. 1.6).

In total, 62 of the components were used in one or more of the pilots. In addition, the platform assets consist of 65 *datasets* and 25 *pipelines* (7 generic) that served the 27 DataBio pilots (Fig. 1.7).

1.5 Introduction to the Technology Chapters

The following chapters in Part I–Part IV describe the technological foundation for developing the pilots. Chapter 2 covers international *standards* that are relevant for DataBio’s aim of improving raw material gathering in bioindustries. This chapter also discusses the emerging role of cloud-based platforms for managing Earth observation data in bioeconomy. The aim is to make Big Data processing a more seamless experience for bioeconomy data.

Chapters 3–6 in Part II describe the main *data types* that have been used in DataBio. These include the main categories sensor data and remote sensing data. Crowd-sourced and genomics data are also becoming increasingly important. The *sensor* chapter gives examples of in-situ IoT sensors for measuring atmospheric and soil

Fig. 1.6 Software components for use in DataBio pilots are in all parts of the BDV Reference Model, which here is presented in a simplified form. The number of components are given within the circles

Fig. 1.7 DataBio platform served the pilots with components, IoT and Earth observation datasets and pipelines to demonstrate improved decision-making

properties as well as of sensor data coming from machinery like tractors. The *remote sensing* chapter lists relevant Earth observation (EO) formats, sources, datasets and services as well as several technologies used in DataBio for handling EO data. The chapters on *crowdsourced* and *genomics* data give illustrative examples of how these data types are used in bioeconomy.

Data integration and modelling is dealt with in Chapters 7–9 in Part III. Chapter 7 explains how data from varying data sources is integrated with the help of a technology called *linked data*. Chapter 8 contains plenty of examples of *integrated linked data pipelines* in the various DataBio applications. Chapter 9 depicts how we *modelled* the pilot requirements and the architecture of the component pipelines. The models facilitate communication and comprehension among partners in the development phase. The chapter also defines metrics for *evaluating* the quality of the models and gives a quality assessment of the DataBio models.

Analytics and visualizing are the topics of Chaps. 10–13 in Part IV. *Data analytics and machine learning* are treated in Chap. 10, which covers the data mining technologies, the mining process as well as the experiences from data analysis in the three sectors of DataBio. Chapter 11 deals with *real-time data processing*, especially event processing, which is central in several DataBio pilots, where dashboards and alerts are computed from multiple events in real-time. *Privacy preserving analytics* is described in Chapter 12. This is crucial, as parts of the bioeconomy data is not open. The last chapter in Part IV is about *visualizing* data and analytics results.

Literature

1. Palmer, M. (2006). *Data is the new oil*. https://ana.blogs.com/maestros/2006/11/data_is_the_new.html. Accessed September 21, 2020.
2. Laney, D. (2001). *3D data management: Controlling data volume velocity and variety*. Gartner file No. 949. <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Accessed September 21, 2020.
3. Reinsel, D., Gantz, J., & Rydning, J. (2018). Data age 2025: The digitization of the world. *International Data Corporation*. <https://itupdate.com.au/page/data-age-2025-the-evolution-of-data-to-life-critical->. Accessed September 21, 2020.
4. ISO/IEC CD 205. (2019). Big Data overview and vocabulary. <https://www.iso.org/standard/68305.html>. Accessed September 28, 2020.
5. NIST Big Data Interoperability Framework: Volume 1, Definitions. (2019). <https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-definitions>. Accessed September 21, 2020.
6. Drake, M. (2019). *A comparison of NoSQL database management systems and models*. <https://www.digitalocean.com/community/tutorials/a-comparison-of-nosql-database-management-systems-and-models>. Accessed September 21, 2020.
7. Auer, S. R., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. The semantic web. *Lecture Notes in Computer Science* (4825, p. 722). https://doi.org/10.1007/978-3-540-76298-0_52. ISBN 978-3-540-76297-3. Accessed September 21, 2020.
8. Cambridge English Dictionary. (2020). Cambridge University Press. <https://dictionary.cambridge.org>. Accessed September 22, 2020.

9. Ayris, P., Berthou, J.-Y., Bruce, R., Lindstaedt, S., Monreale, A., Mons, B., Murayama, Y., Södergård, C., Tochtermann, K., & Wilkinson, R. (2016). *Realising the European open science cloud*. European Union. <https://doi.org/10.2777/940154>. Accessed December 9, 2020.
10. Habyarimana, E. et al. (2017). DataBio deliverable D6.2—data management plan. https://www.databio.eu/wp-content/uploads/2017/05/DataBio_D6.2-Data-Management-Plan_v1.0_2017-06-30_CREA.pdf. Accessed September 22, 2020.
11. Södergård, C. (2019). *DataBio platform*. Webinar: Big Data Breakthroughs for Global Bioeconomy Business. <https://es.slideshare.net/BDVA/bdv-webinar-series-caj-big-data-breakthroughs-for-global-bioeconomy-business>. Accessed December 9, 2020.
12. Chaabouni, K., Bagnato, A., Walderhaug, S., Berre, A. J., Södergård, C., & Sadovykh, A. (2019) Enterprise architecture modelling with ArchiMate. *CEUR Workshop Proceedings*, 2405, 79–84. https://ceur-ws.org/Vol-2405/14_paper.pdf. Accessed December 7, 2020.
13. Plakia, M., Rousopoulos, K., Hara, S., Simarro, J. H., Palomares, M. A. E., Södergård, C., Siltanen, P., Kalaoja, J., Hayarimana, E., Kubo, B., Senner, I., Fournier, F., Berre, A. J., Tsalgatiidou, A., Coene, Y., Auran, P. G., Kepka, M., Charvat, K., Charvat Jr., K., & Krommydas, S. (2019) *DataBio deliverable D4.4—service documentation*. https://www.databio.eu/wp-content/uploads/2017/05/DataBio_D4_4ServiceDocumentation_v1_2_2020_03_13_EXUS.pdf. Accessed December 7, 2020.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.