Introduction to Big Data and Its Applications in Insurance

Romain Billot, Cécile Bothorel, Philippe Lenca

Introduction to Big Data and Its Applications in Insurance

Romain Billot, Cécile Bothorel and Philippe Lenca

1.1. The explosion of data: a typical day in the 2010s

At 7 am on a Monday like any other, a young employee of a large French company wakes up to start her week at work. As for many of us, technology has appeared everywhere in her daily life. As soon as she wakes up, her connected watch, which also works as a sports coach when she goes jogging or cycling, gives her a synopsis of her sleep quality and a score and assessment of the last few months. Data on her heartbeat measured by her watch are transmitted by WiFi to an app installed on her latest generation mobile, before her sleep cycles are analyzed to produce easy-to-handle quality indicators, like an overall score, and thus encourage fun and regular monitoring of her sleep. It is her best night's sleep for a while and she hurries to share her results by text with her best friend, and then on social media via Facebook and Twitter. In this world of connected health, congratulatory messages flood in hailing her "performance"! During her shower, online music streaming services such as Spotify or Deezer suggest a "wake-up" playlist, put together from the preferences and comments of thousands of users. She can give feedback on any of the songs for the software to adapt the

upcoming songs in real time, with the help of a powerful recommendation system based on historical data. She enjoys her breakfast and is getting ready to go to work when the public transport Twitter account she subscribes to warns her of an incident causing serious disruption on the transport network. Hence, she decides to tackle the morning traffic by car, hoping to avoid arriving at work too late. To help her plan her route, she connects to a traffic information and community navigation app that obtains traffic information from GPS records generated by other drivers' devices throughout their journeys to update a real-time traffic information map. Users can flag up specific incidents on the transport network themselves, and our heroine marks slow traffic caused by an accident. She decides to take the alternative route suggested by the app. Having arrived at work, she vents her frustration at a difficult day's commute on social media. During her day at work, on top of her professional activity, she will be connected online to check her bank account balance and go shopping on a supermarket's "drive" app that lets her do her shop online and pick it up later in her car. Her consumer profile on the online shopping app gives her a historical overview of the last few months, as well as suggesting products that are likely to interest her. On her way home, the trunk full with food, some street art painted on a wall immediately attracts her attention. She stops to take a photo, edits it with a color filter and shares it on a social network similar to Instagram. The photo immediately receives about 10 "likes". That evening, a friend comments on the photo. Having recognized the artist, he gives her a link to an online video site like YouTube. The link is for a video of the street art being painted, put online by the artist to increase their visibility. She quickly watches it. Tired, she eats, plugs in her sleep app and goes to bed.

Between waking up and going to sleep, our heroine has generated a significant amount of data, a volume that it would have been difficult to imagine a few years earlier. With or without her knowledge, there have been hundreds of megabytes of data flow and digital records of her tastes, moods, desires, searches, location, etc. This *homo sapiens*, now *homo numericus*, is not alone – billions of us do the same. The figures are revealing and their growth astonishing: we have entered the era of big data. In 2016, one million links were shared, two million friend requests were made and three million

messages were sent every 20 minutes on Facebook [STA 16a]. The figures are breathtaking:

– 1,540,000,000 users active at least once a month;

– 974,000,000 smartphone users;

– 12% growth in users between 2014 and 2015;

– 81 million Facebook profiles;

– 20 million applications installed on Facebook every day.

Since the start of computing, engineers and researchers have certainly been confronted with strong growth in data volumes, stored in larger and larger databases that have come to be known as data warehouses, and with ever improving architectures to guarantee high quality service. However, since the 2000s, mobile Internet and the Internet of Things, among other things, have brought about an explosion in data. This has been more or less well managed, requiring classical schemes to be reconsidered, both in terms of architecture and data processing. Internet traffic, computer backups on the cloud, shares on social networks, open data, purchase transactions, sensors and records from connected objects make up an assembly of markers in space and/or time of human activity, in all its dimensions. We produce enormous quantities of data and can produce it continuously wherever we are (the Internet is accessible from the office, home, airports, trains, cars, restaurants, etc.). In just a few clicks, you can, for example, describe and review a meal and send a photo of your dish. This great wealth of data certainly poses some questions, about ethics and security among other things, and also presents a great opportunity for society [BOY 12]. Uses of data that were previously hidden or reserved for an elite are becoming accessible to more and more people.

The same is true for the open data phenomenon establishing itself at all administrative scales. For big companies, and insurance companies in particular, there are multiple opportunities [CHE 12]. For example, data revealing driving styles are of interest to non-life insurance, and data concerning health and lifestyle are useful for life insurance. In both cases, knowing more about the person being insured allows better estimation of future risks. Storing this data requires a flexible and tailored architecture [ZIK 11] to allow parallel and dynamic processing of “voluminous”, “varied” data at “velocity” while evaluating its “veracity” in order to derive the great

"value" of these new data flows [WU 14]. Big data, or megadata, is often presented in terms of these five Vs.

After initial reflection on the origin of the term and with a view to giving a reliable definition (section 1.2), we will return to the framework of these five Vs, which has the advantage of giving a pragmatic overview of the characteristics of big data (section 1.3). Section 1.4 will describe current architecture models capable of real-time processing of high-volume and varied data, using parallel and distributed processing. Finally, we will finish with a succinct presentation of some examples from the world of insurance.

1.2. How is big data defined?

> "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."
>
> Dan Ariely

It is difficult to define a term as generic, widely used and even clichéd as big data. According to Wikipedia[1]:

> "Big data is a term for datasets that are so large or complex that traditional data processing application software is inadequate to deal with them."

This definition of the big data phenomenon presents an interesting point of view. It focuses on the loss of capability of classical tools to process such high volumes of data. This point of view was put forward in a report from the consulting firm McKinsey and Company that describes big data as data whose scale, distribution, diversity and transience require new architectures and analysis techniques that can unlock new sources of value added [MAN 11]. Of course, this point of view prevails today (in 2016, as these lines are being written) and a universal definition must use more generic characteristics that

1 "Big Data", Wikipedia, The Free Encyclopedia, available at: https//en.wikipedia.org/wiki/Big_data, accessed 9th July 2017.

will stand the test of time. However, like many new concepts, there are as many definitions as there are authors on the subject. We refer the reader to [WAR 13] for an interesting discussion on this theme. To date the genesis of big data, why not make use of one of their greatest suppliers, the tech giant Google? Hence, we have extracted, with the help of the Google Trends tool, the growth in the number of searches for the term "big data" on the famous search engine. Figure 1.1 shows an almost exponential growth in the interest of people using the search engine from 2010 onwards, a sign of the youth of the term and perhaps a certain degree of surprise at a suddenly uncontrollable volume of data, as the Wikipedia definition, still relevant in 2016, suggests. However, articles have widely been using this concept since 1998, to relate a future development of data quantities and databases towards larger and larger scales [FAN 13, DIE 12]. The reference article, widely cited by the scientific community, dates from 2001 and is attributed to Doug Laney from the consultancy firm Gartner [LAN 01]. Curiously, the document never mentions the term big data, although it features the reference characterization of three Vs: *volume*, *velocity* and *variety*. "Volume" describes the size of the data, the term "velocity" captures the speed at which it is generated, communicated and must be processed, while the term "variety" refers to the heterogeneous nature of these new data flows. Most articles agree on the basic three Vs (see [FAN 13, FAN 14, CHE 14]), to which the fourth V of *veracity* (attributed to IBM [IBM 16]), as well as the fifth V, *value*, are added. The term "veracity" focuses on the reliability of the various data. Indeed, data can be erroneous, incomplete or too old for the intended analysis. The fifth V conveys the fact that data must above all create value for the companies involved, or society in general. In this respect, just as certain authors remind us that small volumes can create value ("small data also may lead to big value", see [GU 14]), we should not forget that companies, through adopting practices suited to big data, must most of all store, process and create *intelligent* data. Perhaps we should be talking about *smart data* rather than big data?

1.3. Characterizing big data with the five Vs

In our initial assessment of the big data phenomenon, it should be noted that the 3 Vs framework of volume, velocity and variety, popularized by the research firm Gartner [LAN 01], is now standard. We will thus start with this

classical scheme, shown in Figure 1.2, before considering other Vs, which will soon prove to be useful for developing this initial description.

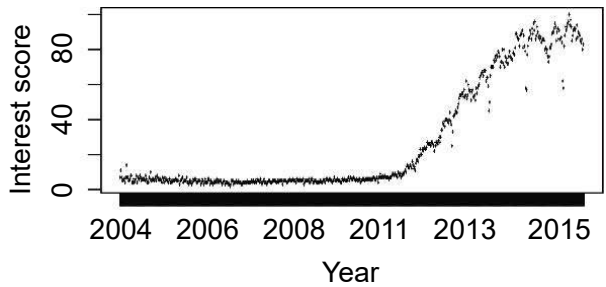

Figure 1.1. *Evolution of the interest in the term big data for Google searches (source: Google Trends, 27th September 2016)*

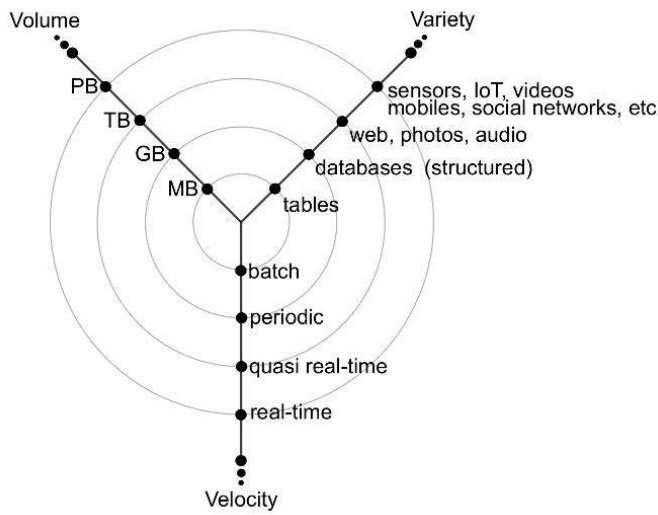

Figure 1.2. *The three Vs of big data*

1.3.1. *Variety*

In a break with tradition, we will start by focusing on the variety, rather than volume, of data. We refer here to the different types of data available today. As we illustrated in the introduction, data originates everywhere, for example:

– texts, photos and videos (Internet, etc.);

– spatio-temporal information (mobile devices, smart sensors, etc.);

– metadata on telephone messages and calls (mobile devices, etc.);

– medical information (patient databases, smart objects, etc.);

– astronomical and geographical data (satellites, ground-based observatories, etc.);

– client data (client databases, sensors and networked objects, etc.).

The handful of examples listed above illustrate the heterogeneity of sources and data – "classical" data like that seen before the era of big data, evidently, and also video signals, audio signals, metadata, etc.

This diversity of content has brought about an initial paradigm shift from structured to non-structured data. In the past, much data could be considered to be structured in the sense that they could be stored in relational databases. This was how client or commercial data was stored. Today, a large proportion of data is not structured (photos, video sequences, account updates, social network statuses, conversations, sensor data, recordings, etc.).

1.3.2. *Volume*

If you ask a range of different people to define big data, most of them will bring up the concept of size, volume or quantity. Just close your eyes and imagine the amount of messages, photos and videos exchanged per second globally. In parallel to the developing interest for the concept of big data on the search engine Google (Figure 1.1), Internet usage has also exploded in just a few years, as the annual number of Google searches bears witness (Table 1.1).

The explosion in Internet usage, and in particular mobile Internet as made possible by smartphones and high-speed standards, has led to an unstoppable growth in data volumes, towards units that our oldest readers have surely recently discovered: gigabytes, terabytes, petabytes, exabytes and even zettabytes (a zettabyte is 10^{21} bytes!), as shown in Figure 1.3.

Year	Annual number of searches	Average searches per day
2014	2,095,100,000,000	5,740,000,000
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000
2009	953,700,000,000	2,610,000,000
2008	637,200,000,000	1,745,000,000
2007	438,000,000,000	1,200,000,000
2000	22,000,000,000	60,000,000
1998	3,600,000	9,800

Table 1.1. *Annual Google statistics [STA 16b]*

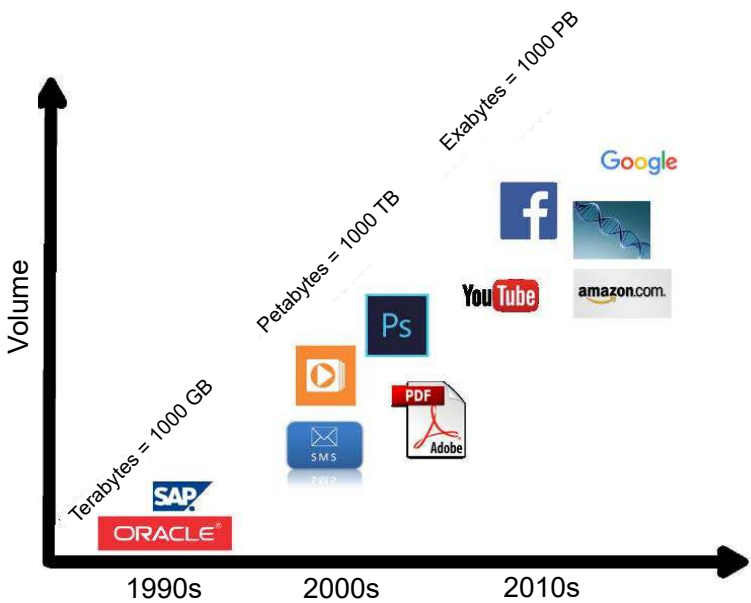

Figure 1.3. *Development of data volumes and their units of measure*

According to an annual report on the Internet of Things [GSM 15], by the end of 2015, there were 7.2 million mobile connections, with projections for smartphones alone reaching more than 7 million in 2019. This expansive volume of data is what brought forth the big data phenomenon. With current data stores unable to absorb such growth in data volumes, companies, engineers and researchers have had to create new solutions, notably offering distributed storage and processing of these masses of data (see section 1.4). The places that store this data, the famous data centers, also raise significant questions in terms of energetic consumption. One report highlights the fact that data centers handling American data consumed 91 billion kWh of electricity in 2013, equivalent to the annual output of 34 large coal-fired power plants [DEL 14]. This figure is likely to reach 140 billion in 2020, equivalent to the annual output of 50 power plants, costing the American population $13 billion per year in electricity bills. If we add to this the emission of 100 million metric tons of CO_2 per year, it is easy to see why large organizations have very quickly started taking this problem seriously, as demonstrated by the frequent installation of data centers in cold regions around the world, with ingenious systems for recycling natural energy [EUD 16].

1.3.3. *Velocity*

The last of the three historic Vs, the V for *velocity*, represents what would probably more naturally be called *speed*. It also covers multiple components, and it is intrinsic to the big data phenomenon. This is clear from the figures above regarding the development of the concept and volume of data, like a film in fast-forward. Speed can refer to the speed at which the data are generated, the speed at which they are transmitted and processed, and also the speed at which they can change form, provide value and, of course, disappear. Today, we must confront large waves of masses of data that must be processed in real time. This online-processed data allow decision makers to make strategic choices that they would not have even been aware of in the past.

1.3.4. *Towards the five Vs: veracity and value*

An enriched definition of big data quickly took shape with the appearance of a fourth element, the V of veracity, attributed to IBM [IBM 16]. The word

veracity brings us back to the quality of the data, a vital property for all data search processes. Again, this concept covers different aspects, such as imprecision, incompleteness, inconsistency and uncertainty. According to IBM, poor data quality costs on average $3.1 trillion per year. The firm adds that 27% of questionnaire respondents are not sure of the information that they input and that one in three decision makers have doubts concerning the data they base their decision on. Indeed, the variety of data flows, which are often unstructured, complicates the process of certifying data. This brings to mind, for example, the quality of data on the social network Twitter, whose imposed 140 character format does not lend itself to precise prose that can be easily identified by automatic natural language processing tools. Certifying data is a prerequisite for creating value, which constitutes the fifth V that is well established in modern practices. The capacity to store, understand and analyze these new waves of high-volume, high-velocity, varied data, and to ensure reliability while integrating them into a *business intelligence* ecosystem, will undoubtedly allow all companies to put in place new decision advice modules (for example, predictive analysis) with high added value. One striking example concerns American sport and ticket sales that are currently based on dynamic pricing methods enhanced by historical and real-time data. Like many other American sports teams, the San Francisco Giants baseball team has thus adapted its match ticketing system to make use of big data. They engaged the services of the company QCUE to set up algorithmic trading techniques inspired by airline companies. The ticket prices are updated in real time as a function of supply and demand. In particular, historical data on the quality of matches and attendances are used to adjust ticket prices to optimize seat/stadium occupation and the company's profits. On their website, QCUE report potential profit growth of up to 46% compared to the previous system.

Globally, big data represents a lucrative business. The McKinsey Institute has suggested that even the simple use of client location data could yield a potential annual consumer surplus of $600 billion [MAN 11]. The consulting group Wikibon estimates that the big data market, encompassing hardware, software and related services, will grow from $19.6 billion in 2013 to $84 billion in 2026 [KEL 15].

1.3.5. *Other possible Vs*

Skimming through the immense number of articles dedicated to the subject, the reader soon realizes that each author is tempted to add their own personal V, each making their own contribution to the various aspects of big data. Thus, the terms variability and validity, which relate directly back to the previous concepts of variety and veracity, can also be added to the list. The word variability focuses on the versatile (yet another V!) nature of data, which can change over time, whereas validity is a more explicit reference to a certification process of classical data. Finally, without degenerating into unhelpful one-upmanship, it seems worthwhile to mention one last V, for visualization. The V of visibility is sometimes tied in with this. Big data, with all of its characteristics as described so far, calls for new forms of visualization to make the data understandable and presentable for decision makers. This can range from simple reporting tools offering an overarching view of the main data characteristics to more advanced methods combining visualization and data analysis. For example, visualization techniques with graphs demonstrating the complex relationships between contributors on social networks, clients, communities or naturally forming groups, are now commonplace.

1.4. Architecture

The era of big data is persuading enterprises of all sizes to implement processes to help make decisions based on data analysis. Predicting what will satisfy a client, optimizing processes and, more generally, generating value from data have now become essential for any business that wants to remain competitive. Although these have always been central challenges for insurers, they are no less affected by the more complex environment of the data economy. Growing volumes of data, of various different natures, with variable lifetimes and of disparate quality, which we want to interrogate in real time, are influencing the tools used, which continue to evolve.

We will see in this section that the scientific and technical environment is becoming richer and more complex by the day. New algorithms are dreamt up to address problems, and new tools are created to test and apply them. In this context, the main task for companies is to incorporate these innovations

alongside existing tools in order to integrate new predictive data analysis processes with existing business procedures. This takes time and expertise, for the project to be defined, to get it running and then to maintain and update it.

1.4.1. *An increasingly complex technical ecosystem*

As has been mentioned already, the essence of the big data phenomenon lies in the limitation of "classical" tools and the need to upgrade them so that they can collect, store and analyze new types of ever greater volumes of data. As for data collection and storage, although all data combined together are usually high volume, each data source produces a "reasonable" volume that can still be managed by "classical" storage and analysis tools. An intelligent distribution of databases is often sufficient for the collection and storage of data in different physical servers, and if the need is felt to put them on the network, it is "sufficient" to use a distributed, robust and fault-tolerant storage system. Big data architectures are needed when each data source produces volumes incompatible with the analysis tools. We thus turn to parallelization, which expresses itself in two ways:

– data parallelism, where a single dataset is divided into subsets distributed over different machines;

– task parallelism, where the algorithms and different sub-procedures are executed concurrently on different processors.

Currently, the best-known big data architecture is probably Hadoop. Contrary to the myth attributing the creation of Hadoop to Yahoo, the project really started at Google. Doug Cutting was working on web content indexing there and needed a framework that would allow large numbers of operations to run in parallel over large collections of servers. The "MapReduce" principle of processing data spread over multiple servers, which is the programming model that Hadoop is based on, was published in 2004 by Google Labs. Doug Cutting joined Yahoo in 2008 and launched the first major Hadoop project, the Yahoo! Search Webmap, which runs on a cluster of 10,000 Linux cores. Today, Hadoop is an open source project managed by the Apache foundation [HAD 16], and its ecosystem is developing day by day with numerous projects optimizing or adding different components. In 2016,

the major Web actors like Twitter and Facebook stored and searched through their tens of petabytes[2] of data on Hadoop.

The Hadoop *framework* can be broken down into three main modules:

– the Hadoop distributed file system (HDFS): the system of files is distributed over different nodes of a cluster. These data nodes are machines networked using a master-slave model. The machines themselves can be relatively modest (and hence inexpensive) servers, it is the number of them that guarantees the big data capacity of the cluster. Every file is split up into blocks. The blocks are distributed across several machines, which allow large volumes of files to be stored, including volumes exceeding the storage capacity of each of the servers. One particular node, the name node, tracks the location of the different blocks and allows access to the data. Each block is replicated at least three times over three different data nodes to ensure redundancy. This principle of horizontal distribution (sharding) enjoys the advantage of being easily re-scaled, since more data nodes can be added to increase the data storage capacity. Overall, HDFS is an efficient, fault-tolerant and scalable file system, which undoubtedly contributed substantially to its success;

– the MapReduce data processing engine: a MapReduce job (a processing task) is completed in two stages, a mapping step that transforms raw data into a key/value format, and a reducing step that combines all of the values for each of the keys. Data handling generally gives rise to a chain of several MapReduce jobs;

– the YARN (Yet Another Resource Negotiator) resource manager: this module was introduced in the second version of Hadoop and allows the infrastructure management to be completely dissociated from the MapReduce data processing model. Thus, while MapReduce describes the data manipulation processes, YARN calls on the name nodes and deals with launching these processes on the different data nodes. At the simplest level, YARN orchestrates the parallel management of the different processes to optimize the distribution of the processing work over the different machines.

2 1 PB (petabyte) = 1,000 TB (terabytes) = 1,000,000 GB (gigabytes).

A range of projects supplementing these core modules enhance the services provided to users, some of which are shown in Figure 1.4. Examples of these services include database management (Hbase) and searches (Hive), real-time data flow processing (Storm), high-level data manipulation scripts (Pig), Web interfaces facilitating data processing (Hue) and, of course, data analysis and search libraries (Mahout).

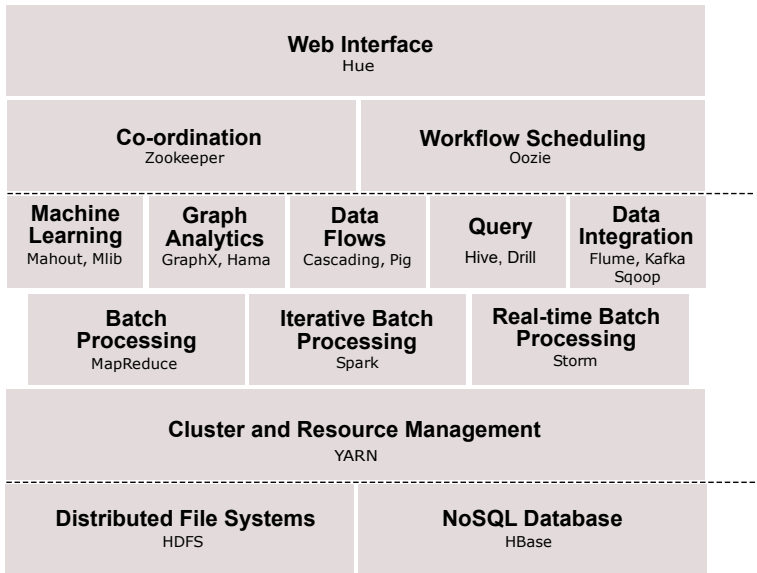

Figure 1.4. *Hadoop and its ecosystem (non-exhaustive)*

The Spark framework has been growing in reputation since 2014[3]. Originally developed in 2009 by AMPLab, from the University of California, Berkeley, the project became an Apache open source project in 2010. Spark, built on Hadoop and MapReduce, improves upon MapReduce by taking advantage of the nodes' random access memory when possible (via Resilient Distributed Datasets or RDD) and chaining together multiple processing steps

3 According to the Google Trends service, which statistically analyzes research subjects of interest to web users.

without systematically reading and writing to the hard disk as MapReduce does. This clever trick significantly speeds up the majority of data handling processes, such as sorting, word counting, unsupervised k-means classification or calculating PageRank centrality in a graph, by up to a factor of 5 [SHI 15]. Nevertheless, we note that according to [SHI 15], MapReduce performs better at managing the processes between the mapping and reducing phases. Furthermore, Spark comes with a complete environment, allowing (like MapReduce and Storm) real-time data flow problems as well as background (batch) tasks to be processed, for different types of data (text, graph, etc.). Applications can be written in Java, Scala or Python, and the MLib library (Spark Machine Learning Library), which comes from the data search library Mahout, from MapReduce, updates on the fly, all while offering an increasingly high-level data interface (RDDs have now been expanded into DataFrames, data displays that allow the data to be grouped in columns like in a table from a relational database).

Platforms specializing in decision-making solutions are also rapidly developing. They are offering more and more solutions for interfacing with open source tools. For example, SAS has offered SAS® Data Loader to interface with Hadoop, and since 2015 has clearly positioned itself with the main themes in the sector, such as cybersecurity or the Internet of Things. As another example, IBM is extending its IBM Cloud Bluemix platform with their Data Science Experience offering, based on Apache Spark. More specifically, this offering allows data scientists and developers access to 250 datasets, all powered by Spark and equipped with different open source software, like H2O, a Machine Learning solution. This data analysis software is not only compatible with big data platforms like Spark, but also claims to allow machine learning models developed in Python, Java or R to be easily deployed on these platforms. H2O is offered by a Californian start-up, H2O.ai.

According to KDnuggets [PIA 16], a site specializing in current affairs in business analytics, big data, data science and data mining, there are not many professionals who use only proprietary or indeed only open source solutions. A large majority of them use both families of tools. The dynamism of the open source community has made its technologies very popular to use. According to a 2013 survey run by O'Reilly, looking at data scientist salaries, the median

salary of a data scientist who uses open source tools is 130,000 $US compared to 90,000 $US for those who only use proprietary tools.

According to the same site, the use of tools in the "Hadoop/Big data" category is becoming more accessible. Almost half of professionals use these tools (39% in 2016 compared to 29% in 2015 and 17% in 2014). This development is primarily due to the growth of Apache Spark, MLlib and H2O (see Table 1.2).

Tool	2016	2015	2015 –> 2016
Hadoop	22.1%	18.4%	+20.5%
Spark	21.6%	11.3%	+91%
Hive	12.4%	10.2%	+21.3%
MLlib	11.6%	3.3%	+253%
SQL on Hadoop tools	7.3%	7.2%	+1.6%
H2O	6.7%	2.0%	+234%
HBase	5.5%	4.6%	+18.6%
Apache Pig	4.6%	5.4%	–16.1%
Apache Mahout	2.6%	2.8%	–7.2%
Dato	2.4%	0.5%	+338%
Datameer	0.4%	0.9%	–52.3%
Other Hadoop/HDFS-based tools	4.9%	4.5%	+7.5%

Table 1.2. *Usage statistics for big data tools according to a survey of 2,895 respondents from the data analytics community and vendors. The respondents were from US/Canada (40%), Europe (39%), Asia (9.4%), Latin America (5.8%), Africa/Middle East (2.9%) and Australia/NZ (2.2%). They were asked about 102 different tools, including the "Hadoop/big data tools" shown here [PIA 16].*

Continuing to look at the data from KDnuggets [PIA 16], R appears to be the preferred tool of data scientists for data analytics. Usually used on an office machine with datasets of reasonable size, this language originally designed for statisticians is perfect for exploratory analysis, because it comes with libraries rich in algorithms for machine learning, evaluation, producing graphs, etc. Combined with offers such as H2O (or Rserver), it is now transferrable to the big data environment. However, Python, a computer

programming language, is growing in popularity. Being flexible and open, and a generalist programming language, it is well suited to integrating analysis tasks with Web applications or with specific unconventional architectures. Its dedicated data science libraries make it a serious competitor to R.

1.4.2. *Migration towards a data-oriented strategy*

There are still very few companies who can boast of having migrated towards a data-oriented strategy. The specialist Internet press, informed by digital transformation consultants with a wide overview of these changes, agrees on four identifiable phases of big data adoption [DEM 16]:

1) experimentation with the big data platform;

2) implementation: developing first use cases;

3) expansion: deployment in multiple use cases;

4) optimization: integration with the business IT system.

The experimentation phase is when the potential of using a big data infrastructure is explored. The aim at this phase is to deal with installation and configuration. The main objective is to see how compatible the technology is with existing architecture. Such experimentation need not cost much because all that is required are a few bottom-of-the-range servers kitted out with open source software such as Hadoop/Spark. This experimentation phase very often results in the use of a data storage layer with pre-existing data, upon which a new layer of data handling is added, such as database queries.

Once the technical platform has been mastered, during the second implementation phase, the business tackles a use case that demonstrates the value of big data. This consists of developing a data processing chain for pre-existing data, then deploying this proof of concept in a production context. Common use cases at this stage include detecting fraud, log analysis for improved understanding of use patterns, predicting churn or, closer to the user experience, introducing recommendation systems. Data analytic libraries, such as MLib for Spark [SPA 16], have long lists of native (and optimized) algorithms for addressing these types of problems. The objective

here is to demonstrate the value added and the economic impact of setting up a big data architecture.

The third phase is of course generalizing use cases to different levels of the business's value chain. The teams in charge of big data will by now have examples of early successes to help convince the different stakeholders in the business, and the cost of developing a new use case will be reduced since the infrastructure already exists. This is where business applications see the light of day, each service seizing upon technology to optimize existing analysis, extending it, proposing new analysis or simply gaining a better understanding of their field. A financial service will seek to improve risk management or fraud detection, a health service will launch targeted prevention programs, aim to reduce readmission or analyze internal processes to improve their coordination.

Finally, the last phase consists of true integration of data analytics and its insights into the overall strategy of the business. The improvement in business procedures and/or economic benefits is turned into competitive advantages. Results from predictive analysis inform decision-making. At this stage, the decision makers consult someone with responsibility for data (the job title Chief Data Officer is starting to appear) and a dedicated data team maintains the infrastructure and sets about solving new, ad-hoc problems specific to the business. The data analyst, a specialist in statistics, helps to produce dashboards displaying the data and to make best use of data processing chains, whereas the data scientist, with expertise in mathematics, statistics and computing, produces new data processing chains and unlocks new opportunities, while also making sure to maintain real-time visualization of the company's performance.

1.4.3. *Is migration towards a big data architecture necessary?*

Companies are inevitably considering whether or not to migrate towards a big data architecture. Does the existing business intelligence (BI) system need replacing? As a simplification, this type of system consists of two main parts:

– the ETL process (extracting, transforming and loading data), which consists of extracting from the company's operational data sources all the (heterogeneous) data that could help respond to the decision makers'

questions. The data is then processed (cleaned, normalized, aggregated, etc.) and integrated so that it can be loaded into the data warehouse following predefined protocols;

– the data warehouse allowing all of a company's data to be consolidated and integrated and hence offering a cross-cutting and integrated overview of all aspects of the company's business. It can be made up of several subsets called datamarts which each characterize a defined business procedure. This data is structured in the form of multidimensional logical schemas allowing access to predefined indicators to be prepared, to fulfill a reporting requirement for example, while still allowing their analysis in several dimensions (for example, analyzing the "revenue" indicator "by region", "by period" or "by shop"). This modeling can be used to build multidimensional cubes (or hypercubes) on OLAP servers, allowing significant interactivity when searching. Graphical BI tools for analysis and reporting, like Excel, Table or Business Object, are often used to build dashboards and reports in consultation with the warehouse.

The arrival of big data has been accompanied by the emergence of new analytical processes (or workloads) that classical ETL or storage technologies would struggle to complete:

– exploratory analysis of raw, unmodeled and unstructured data;

– real-time processing, in contrast to ETL processes that run in batches;

– accelerated batch processing for large data volumes;

– agility and rapid data archiving, with the ability to rapidly repeat the processing necessary to update the warehouse data;

– complex analysis, such as the parallel application of many millions of scoring models on millions of bank accounts to detect fraud, for example.

The good news is that it is possible to bring the two worlds together and to use Hadoop as an efficient and scalable ETL solution for data that requires specific workloads. Once the data has been extracted and loaded in Hadoop, it can be subjected to complex transformations in batches by programming MapReduce or Spark jobs, or using high-level languages like HiveQL or Pig. It is possible to analyze (parse) the syntax of unstructured or semi-structured data, and to carry out calculations, joins and aggregations in order to integrate

data from diverse sources, or to structure them so that they can be inserted into data warehouses following classical business workflows.

Hadoop can also be used to build a flexible and scalable data warehouse and to interface it with classical BI tools, for reporting for example. However, the majority of data warehousing solution publishers such as Oracle or Teradata prefer to integrate Hadoop at the ETL level only, which allows their solutions to be augmented rather than replaced. Conversely, proponents of open source solutions champion workload management in which the distributed Hadoop environment plays the role of a data hub through which all the data in the company ecosystem transits, before being fed into multiple analytical platforms.

Analyzing all of these approaches is complex. Some authors have produced grids comparing the requirements of different technical choices, such as the properties of the data analysis algorithms [LAN 15], as well as their potential implications, for example, regarding skills and human resources [CHA 13].

1.5. Challenges and opportunities for the world of insurance

Data lies at the heart of insurance. It is the raw material for scoring models, allowing segmentation of premium holders, to know them better and offer them bespoke products, to better estimate their current and future risk and to make decisions. Big data and the digital transition are hence profoundly changing the insurance sector. As for all economic actors, insurers will of course face changes of organization, culture and competition. We will illustrate this development with two examples in which big data plays a central role: the first illustrates the impact of the development of the sharing economy and the second the impact of changing behaviors on segmentation.

Insurance is already part of the sharing economy [LAC 15]. New actors, not necessarily from the world of insurance, are creating communities of individuals with specific insurance needs in order to negotiate highly personalized contracts for them from insurers, and reducing costs as they do so. If community platforms are allowing individuals to articulate their needs, big data is allowing these new actors to be proactive in finding small groups of clients whose frustration accumulates online. Indeed, all that is required is to analyze search engine enquiries, blogs and social networks to determine

specific insurance needs. These new actors are thus changing the relationship between the insured and their insurers, but are also facilitating innovation since the (very) personalized solutions are either adaptations of existing contracts or completely new contracts. Although this type of market is still marginal, it seems likely for such a market of niches to be able to grow. This is particularly the case for the collaborative practices for sharing goods or services (carsharing, vehicle/apartment hire between individuals, etc.) which continue to develop. These are changing how risks are assessed and again specific, or even bespoke, warranties must be offered [INC 14]. Essentially, these practices are changing the paradigm from "one good for one owner" to "a multitude of users for one good". This shift from ownership towards usage is bringing about new types of risks and represents a challenge for insurers [LAC 15].

Big data also gives easy access to some of the information necessary for pricing and will gradually reduce the use of classical paper questionnaires. Hence, it allows faster decision-making. Even better, by giving access to previously inaccessible information, it will enable reduction in the existing information asymmetry [EWA 13] between the person being insured, who knows virtually all the information concerning them, and the insurer who has only partial information. Hence, big data allows greater knowledge of the insured and the risks associated with them, more precise evaluation of behavior and hence optimized selection of who to insure and fairer premium prices. Those being insured can, particularly if it is in their interest, give access to very private data about their way of life. The acceptability of such an approach, for consumers and regulators, is evidently critical [THO 15]. The slogan "pay as you live, drive etc." is already here, especially in automobile insurance. For example, connected driving allows precise analysis of driving style (speed, acceleration, braking, cornering, etc.), according to the road and weather conditions. This trend is also developing in health insurance with connected objects, allowing the physical condition (heart rate, sleep, etc.) and activity (number of steps taken, participation in sports, etc.) of the person being insured to be measured. The quality of their everyday environment can be evaluated using external and open data. However, "hyper-individualized" premium pricing could challenge the current model of segmentation and mutualization of risk [HOU 15], the underlying principle of how prices are set, and questions how risk portfolios will be structured [CHA 15]. The intrusion of insurers into the heart of individuals' private lives obviously

poses the problem of data protection. There are also questions regarding how new practices will develop and how they might impact society.

Through these two examples, we have demonstrated some of the opportunities offered by big data (new markets, innovation and reduction in information asymmetry). Improving the effectiveness of advertising campaigns and of targeting and reducing fraud are further examples. New challenges are appearing (the entrance of intermediaries, the fundamentals of insurance under question, data security, actuarial challenges) while ethical, security and legal questions are also being raised. Regulators may restrict the use of personal data or data that leads to segmentation considered to be discriminatory. Markets for fraudulent profiles could develop, and alert premium holders will create different profiles for private and public use, thus challenging the benefit of the reduction in information asymmetry. Finally, if big data represents a profitable investment, it risks destabilizing the whole insurance market. On the one hand, companies without the means to access big data and the necessary technologies and workforce skills will see their competitivity unravel. They therefore risk disappearing or being bought out. On the other hand, intermediary platforms, notably GAFA (Google, Apple, Facebook, Amazon), who control the whole data value chain (collection, the technology for storage and calculations, relevant expertise), could seek to take a significant proportion of the profits, or could even be tempted to become insurers themselves. Buying out weakened companies could thus allow them to enter the insurance market. A new form of asymmetry, of control over data, is probably already in place.

1.6. Conclusion

Big data is here. Without doubt, the flood of data should continue, if not grow. If properly stored, managed and exploited, big data offers numerous opportunities. Computing has laid down a gauntlet: new architectures and a new ecosystem have been developed and are continually evolving. Insurance has not been spared from this phenomenon. Big data will allow new opportunities to be seized and also brings new risks. The final three chapters of this book will shed light on these developments.

However, big data cannot do everything, all the time. One famous example, among others, is the failure of Google's flu forecasting system

(since abandoned) [LAZ 14]. Good predictions sometimes rely upon good understanding, and data science, despite inevitable changes to make and challenges to face, has bright days ahead of it. These issues as well as the main machine learning algorithms will be presented in the next two chapters.

1.7. Bibliography

[BOY 12] Boyd D., Crawford K., "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society*, no. 5, pp. 662–679, 2012.

[CHA 13] Chalmers S., Bothorel C., Picot Clémente R., Big Data – State of the Art, Report, Télécom Bretagne, 2013.

[CHA 15] Charpentier A., Denuit M.M., Elie R., "Segmentation et mutualisation, les deux faces d'une même pièce?", *Risques*, no. 103, pp. 19–23, 2015.

[CHE 12] Chen H., Chiang R.H., Storey V.C., "Business intelligence and analytics: from big data to gig impact", *MIS Quarterly*, no. 4, pp. 1165–1188, 2012.

[CHE 14] Chen M., Mao S., Liu Y., "Big data: a survey", *Mobile Networks and Applications*, no. 2, pp. 171–209, 2014.

[DEL 14] Delforge P., "America's data centers consuming and wasting growing amounts of energy", *Natural Resource Defence Council*, 2014. Available at: https//:www.nrdc.org/resources/americas-data-centers-consuming-and-wasting-growing-amounts-energy, accessed 18th April 2017.

[DEM 16] Demarest G., "Four Phases of Operationalizing Big Data", *CIOReview*, 2016. Available at: http://bigdata.cioreview.com/cxoinsight/four-phases-of-operationalizing-big-data-nid-15251-cid-15.html, accessed 18th April 2017.

[DIE 12] Diebold F., On the Origin(s) and development of the term "Big Data", Pier Working Paper Archive, Penn Institute for Economic Research, 2012.

[EUD 16] Eudes Y., "Visite exceptionnelle dans le data center de Facebook, en Suède", *Le Monde*, 2016. Available at: http://www.lemonde.fr/ pixels/article/2016/06/03/les-datas-du-grand-froid_4932566_4408996.html, accessed 18th April 2017.

[EWA 13] Ewald F., Thourot P., "Big Data, défis et opportunités pour les assureurs", *ENASS Papers 5, Banque & Stratégie*, no. 315, pp. 5–8, 2013.

[FAN 13] Fan W., Bifet A., "Mining big data: current status, and forecast to the future", *ACM SIGKDD Explorations Newsletter*, no. 2, pp. 1–5, 2013.

[FAN 14] Fan J., Han F., Liu H., "Challenges of big data analysis", *National Science Review*, no. 2, pp. 293–314, 2014.

[GSM 15] GSMA, Unlocking the Value of IoT Through Big Data, Report, GSM Association, 2015.

[GU 14] GU J., ZHANG L., "Some comments on big data and data science", *Annals of Data Science*, nos 3–4, pp. 283–291, 2014.

[HAD 16] HADOOP, Welcome to Apache™ Hadoop®!, available at: http://hadoop.apache.org/, accessed 18th July 2016.

[HOU 15] HOULLE O., "Le Big Data modifie le visage de l'assurance", *ENASS Papers 9, Banque & Stratégie*, no. 336, pp. 28–30, 2015.

[IBM 16] IBM, "IBM- What is big data?", 2016. Available at: https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, accessed 18th July 2016.

[INC 14] INC, "Consommation collaborative: quels enjeux et quelles limites pour les consommateurs?", *Colloque INC*, Ministère de l'Economie, de l'Industrie et du Numérique, Paris, France, 7th November 2014.

[KEL 15] KELLY J., Big Data Vendor Revenue and Market Forecast, 2011–2026, Report, WIKIBON, 2015.

[LAC 15] LACAZE O., "Le XXI^{e} siècle sera collaboratif, quid de l'assurance ?", *ENASS Papers 10, Banque & Stratégie*, no. 341, pp. 30–32, 2015.

[LAN 01] LANEY D., "3D Data managment: controlling data volume, velocity and variety", *Application Delivery Strategies*, no. 949, 2001.

[LAN 15] LANDSET S., KHOSHGOFTAAR T.M., RICHTER A.N. *et al.*, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem", *Journal of Big Data*, no. 1, pp. 1–36, 2015.

[LAZ 14] LAZER D., KENNEDY R., KING G. *et al.*, "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, no. 14, pp. 1203–1205, 2014.

[MAN 11] MANYIKA J., CHUI M., BROWN B. *et al.*, Big data: The next frontier for innovation, competition, and productivity, Report, The McKinsey Global Institute, 2011.

[PIA 16] PIATETSKY G., "R, Python Duel As Top Analytics, Data Science software – KDnuggets 2016 Software Poll Results", KDNUGGETS, 2016, available at: http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html, accessed 13th July 2016.

[SHI 15] SHI J., QIU Y., MINHAS U.F. *et al.*, "Clash of the titans: MapReduce vs. Spark for large scale data analytics", *Proceedings of the VLDB Endowment*, no. 13, pp. 2110–2121, 2015.

[SPA 16] SPARK, Spark Machine Learning Library (MLlib) Guide, MLlib: Main Guide - Spark 2.1.0 Documentation, available at: http://spark.apache.org/docs/latest/mllib-guide.html, accessed 13th July 2016.

[STA 16a] STATISTICBRAIN, Facebook Statistics, 2016, available at: http://www.statisticbrain.com/facebook-statistics/, accessed 13th July 2016.

[STA 16b] STATISTICBRAIN, Google Annual Search Statistics, 2016, available at: http://www.statisticbrain.com/google-searches/, accessed 13th July 2016.

[THO 15] THOUROT P., NESSI J.-M., FOLLY K.A., "Big data et tarification de l'assurance", *Risques*, no. 103, 2015.

[WAR 13] Ward J.S., Barker A., "Undefined by data: a survey of big data definitions", *arXiv preprint arXiv:1309.5821*, 2013.

[WU 14] Wu X., Zhu X., Wu G.-Q. *et al.*, "Data mining with big data", *IEEE Transactions on Knowledge and Data Engineering*, no. 1, pp. 97–107, 2014.

[ZIK 11] Zikopoulos P., Eaton C., *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media, New York, 2011.