What is Big Data?

1) A "marketing" approach derived from technology that the information technologies (IT) industry (and its associated players) comes up on a regular basis.

2) A reality we felt coming for a long time in the world of business (mostly linked to the growth of the Internet), but that did not yet have a name.

3) The formalization of a phenomenon that has existed for many years, but that has intensified with the growing digitalization of our world.

The answer is undoubtedly all three at the same time. The volume of available data continues to grow, and it grows in different formats, whereas the cost of storage continues to fall (see Figure 1.1), making it very simple to store large quantities of data. Processing this data (its volume and its format), however, is another problem altogether. Big Data (in its technical approach) is concerned with data processing; Smart Data is concerned with analysis, value and integrating Big Data into business decision-making processes.

Big Data should be seen as new data sources that the business needs to integrate and correlate with the data it already has, and not as a concept (and its associated solutions) that seeks to replace Business Intelligence (BI). Big Data is an addition to and completes the range of solutions businesses have implemented for data processing, use and distribution to shed light on their decision-making, whether it is for strategic or operational ends.

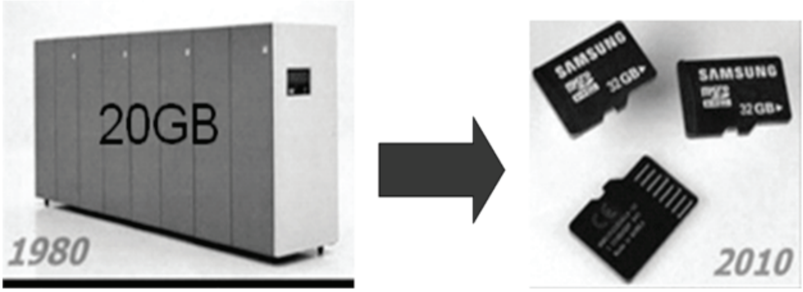

Figure 1.1. *In 1980, 20 GB of storing space weighed 1.5 tons and cost $1M; today 32 GB weighs 20 g and costs less than €20*

Technological evolutions have opened up new horizons for data storage and management, enabling anything and everything to be stored at a highly competitive price (taking into account the volume and the fact the data have very little structure, such as photographs, videos, etc.). A greater difficulty is getting value from this data, due to the "noise" generated by the data that has not been processed prior to the storage process (too much data "kills" data); this is a disadvantage. A benefit, however, is that "raw" data storage opens (or at least does not close) the door to making new discoveries from "source" data. This would not have been possible if the data had been processed and filtered before storage. It is therefore a good idea to arbitrate between these two axes, following the objectives that will have been set.

1.1. The four "V"s characterizing Big Data

Big Data is the "data" principally characterized by the four "V"s. They are Volume, Variety, Velocity and Value (associated with Smart Data).

1.1.1. *V for "Volume"*

In 2014, three billion Internet users connected to the Internet using over six billion objects (which are mainly servers, personal computers (PCs), tablets and smartphones) using an Internet Protocol (IP) address (a "unique" identifier that enables a connected object to be uniquely identified and therefore to enable communication with its peers, which are mainly smartphones, tablets and computers). This generated about eight exabytes (10 to the power of 18 = a billion) for 2014 alone. A byte is a sequence of eight bits (the bit is the basic unit in IT, represented by zero or one) and enables information to be digitalized. In the very near future (see Figure 1.2) and with the advent of connected objects (everyday objects such as televisions, domestic appliances and security cameras that will be connected to the Internet), it is predicted that there will be several tens of billions. We are talking somewhere in the region of 50 billion, which will be able to generate more than 40,000 exabytes (40,000 billion of billion bytes) of data a year. The Internet is, after all, full of words and billions of events occur every minute. Some may have value for or be relevant to a business, others less so. Therefore, to find out which have value, it is necessary to read them, sort them, in short, "reduce" the data by sending the data through a storage, filtering, organization and then analysis zone (see section 1.2).

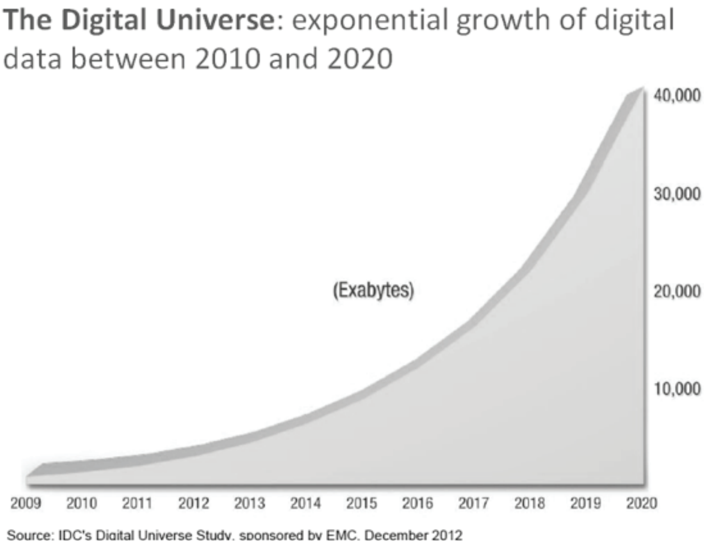

Figure 1.2. *Research by the IDC on the evolution of digital data between 2010 and 2020 (source: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf)*

The main reason for this exponential evolution will be connected objects. We expect there to be approximately 400 times the current annual volume in 2020.

1.1.2. *V for "Variety"*

For a long time, we only processed data that had a good structure, often from transaction systems. Once the data had been extracted and transformed, it was put into what are called decision-support databases. These databases differ from others by the data model (the way data are stored and the relationships between data):

– Transaction data model:

This model (structure of data storage and management) focuses on the execution speed of reading, writing and data modification actions to minimize the duration of a transaction to the lowest possible time (response time) and maximize the number of actions that can be conducted in parallel (scalability, e.g. an e-commerce site must be able to support thousands of Internet users who simultaneously access a catalog containing the products available and their prices via very selective criteria, which require little or no access to historical data). In this case, it is defined as a "normalized" data model, which organizes data structures into types, entities (e.g. client data are stored in a different structure to product data, invoice data, etc.), resulting in little or no data redundancy. In contrast, during the data query, we have to manage the countless and often complex, relations, joints between these entities (excellent knowledge of the data model is required, and these actions are delegated to solutions and applications and are very scarcely executed by a business analyst as they are much too complex).

In sum, the normalized model enables transaction activities to run efficiently, but makes implementing BI solutions and operational reporting (little or no space for analysis) difficult to implement directly on the transactional data model. To mitigate this issue, the operational data store (ODS) was put in place to implement some of the data tables (sourced from the transactional database) to an operational reporting database, with a more simple (light) data model. BI tools enabled a semantic layer (metadata) to be implemented, signaling a shift from a technical to a business view of the data, thereby allowing analysts to create reports without any knowledge of the physical data model.

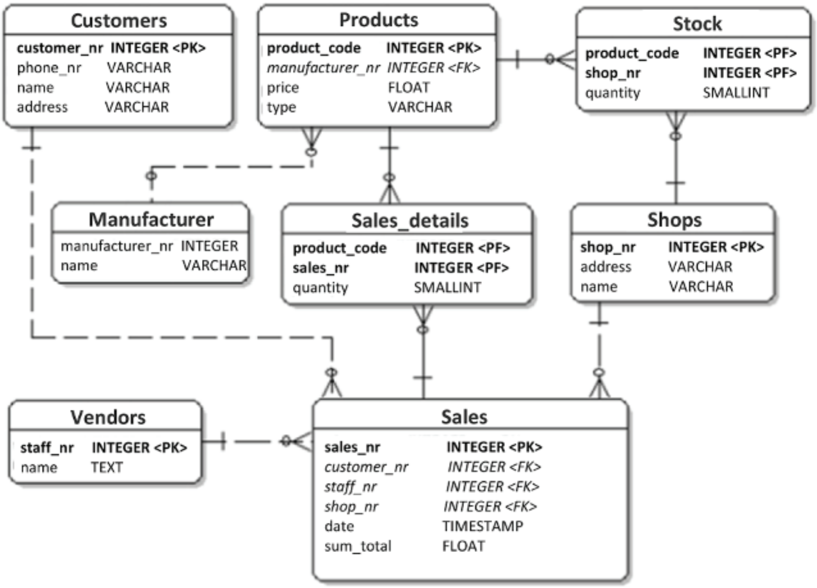

Figure 1.3. *(Normalized) transaction data model*

– Decision data model:

This model focuses on analysis, modeling, data mining, etc., which, the majority of the time, require a large volume of historic information: several years with much broader data access criteria (e.g. all products for all seasons). These restrictions have made the use of relational data models difficult, if not impossible (joints and relations between entities, associated with volume, had a huge impact on the execution time of queries). As a solution to this problem, denormalized data models were implemented. The structure of these models is much simpler (they are known as “star” or “snowflake” models, corresponding to the set of stars connected by their dimensions), where source data are stored in one structure containing all entities, for instance the client, the product, the price and the invoice are stored in the same table (known as a fact table), and can be accessed via analytical dimensions (such as the time, the customer, the product name, the location, etc.), giving the structure a star shape (hence the name of the model). This data model facilitates access (it has little or no joints beyond those necessary for dimension tables) and this access is much more sequential (though indexed). Conversely, there is a redundancy of data caused by the method information is stored in the “fact” table (there is therefore a larger volume to process).

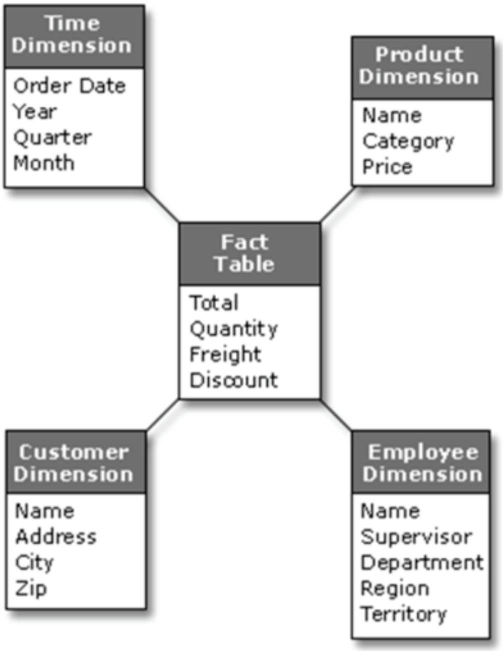

Figure 1.4. *“Star” data model (decision, denormalized)*

For several years, businesses have had to deal with data that are much less structured (or not structured at all, see

Figure 1.5), such as messaging services, blogs, social networks, Web logs, films, photos, etc. These new types of data have to be processed in a particular way (classification, MapReduce, etc.) so that they can be integrated into business decision-making solutions.

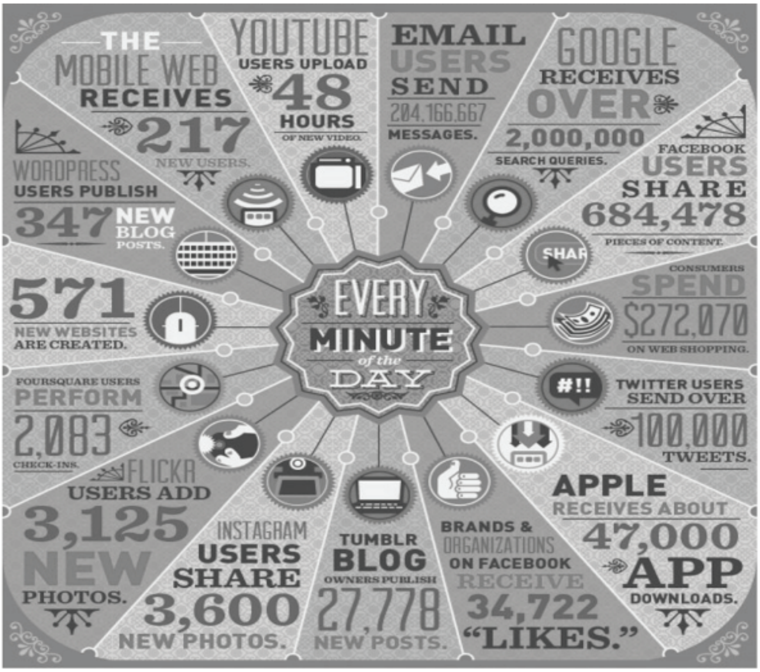

Figure 1.5. *Visual News study from 2012 gives an idea of the volume and format of data created every minute online (source: http://www.visualnews.com/ 2012/06/19/how-much-data-created-every-minute)*

1.1.3. *V for "Velocity"*

The Internet and its billions of users generate uninterrupted activity (the Internet never sleeps). All these activities (whether they are commercial, social, cultural, etc.) are generated by software agents – e-commerce sites, blogs, social networks, etc. – who produce continuous flows of data. Businesses must be able to process this data in "real time".

The term “real time” is still proving difficulty to define. In the context of the Internet, it could be said that this time must be aligned to the temporality of the user’s session. Businesses must be able to act and react (offer content, products, prices, etc., in line with their clients’ expectations, regardless of the time of day or night) in the extremely competitive context that is the Internet. A client does not belong (or no longer belongs) to one business or brand and the notion of loyalty is becoming increasingly blurred. Businesses and brands will only have a relationship with a client for as long as the client wants one and, in these conditions, meeting expectations every time is a must.

1.1.4. *V for “Value”, associated with Smart Data*

1.1.4.1. *What value can be taken from Big Data?*

This question is the heart of this topic/subject: the value of Big Data is the value of every piece of data. It could be said that one piece of data that would never have any value (and that would never be used in any way) will be reduced to a piece of data that has a cost (for its processing, storage, etc.). A piece of data therefore finds its value in its use. Businesses are well aware that they are far from using all the data at their disposition (they are primarily focused on well-structured data from transaction systems). Globalization associated with the (inflationist) digitalization of our world has highlighted this awareness: competition has become tougher, there are more opportunities and the ability of “knowing” before acting is a real advantage. Big Data follows the same value rule: it must be seen as an additional source of information (structured and unstructured) that will enrich businesses’ decision-making processes (both technical and human). It is from this “melting pot” that Big Data starts its transformation into Smart Data (see Chapter 2).

The example below (Figure 1.6) shows the results of an analysis into the number of tweets posted about the price of rice in Indonesia (it can easily be supposed that they are linked to purchases) and the price of rice itself (which is correlated with the tweet curve). Buyers with real-time access to this information will undoubtedly have an advantage (to be able to buy at the right moment, when the price is at its lowest) over others who do not.

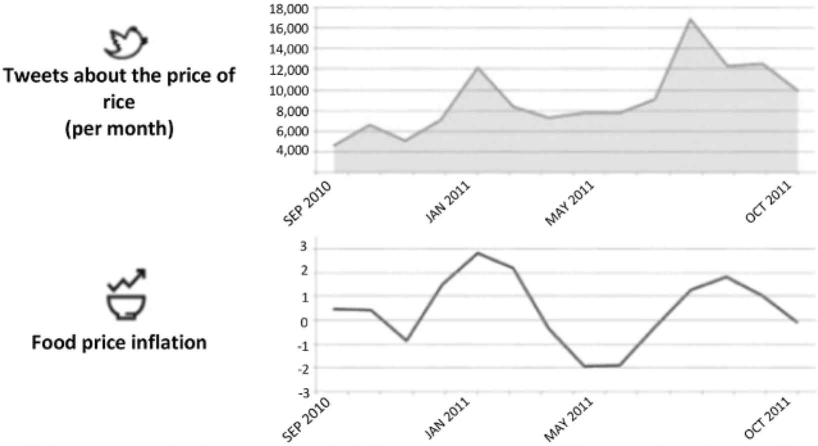

Figure 1.6. *UN Global Pulse study from 2012: correlation in Indonesia between tweets about the price of rice and the sale price of rice [UNI 14]*

Another valuable example is “cognitive business”, that is Web players’ (such as Google, Facebook, etc., which provide a certain number of free services for their users) ability to analyze the data they manage and store (provided to them free of charge by Internet users) to produce and sell it to economic players (information relevant to their activities).

1.2. The technology that supports Big Data

The technology was launched by Google (in 2004) to process huge volumes of data (billions of queries are made online every day on search engines). The technology was inspired by massively parallel processing solutions (MapReduce) used for large scientific calculations. The principle was to parallelize data processing by distributing it over hundreds (and even thousands) of servers (Hadoop Distributed File System) organized into processing nodes. Apache (Open Source) seized the concept and developed it into what we know today.

MapReduce is a set of data distribution processes and processing over a large number of servers (guaranteed by the "Map" process to ensure parallel processing). Results are consolidated (ensured by the "Reduce" process) to then feed the analytical follow-up where this information is analyzed and consolidated to enrich decision-making processes (either human or automated).

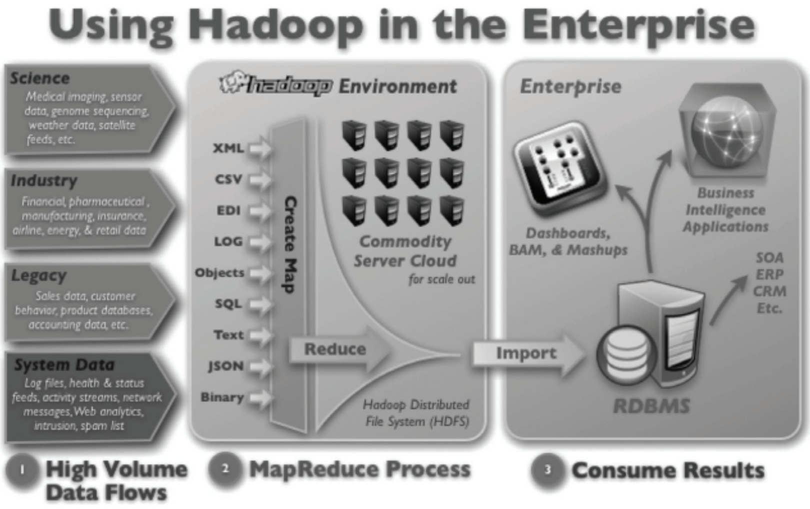
Using Hadoop in the Enterprise
Science
Medical imaging, sensor data, genome sequencing, weather data, satellite feeds, etc.
Industry
Financial, pharmaceutical, manufacturing, insurance, airline, energy, & retail data
Legacy
Sales data, customer behavior, product databases, accounting data, etc.
System Data
Log files, health & status feeds, activity streams, network messages, Web analytics, intrusion, spam list
hadoop Environment
XML
CSV
EDI
LOG
Objects
SQL
Text
JSON
Binary
Create Map
Commodity Server Cloud
for scale out
Reduce
Hadoop Distributed File System (HDFS)
Enterprise
Dashboards, BAM, & Mashups
Business Intelligence Applications
SOA
ERP
CRM
Etc.
Import
RDBMS
1 High Volume Data Flows
2 MapReduce Process
3 Consume Results

Figure 1.7. *Hadoop process & MapReduce*