International Journal of Engineering, Science and Mathematics
Vol7.Issue 3, March 2018,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com Double-Blind
Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

INVESTIGATING BUSINESS INTELLIGENCE IN THE ERA OF BIG DATA: CONCEPTS, BENEFITS AND CHALLENGES

H. EL BOUSTY*A
S. KRITB**
M. ELASIKRI*C**
M. KABRANE**D**
K. BENDAOUD***E**
K. KARIMI****F**
H. OUDANI*****G**

Abstract

Business intelligence suppose retrieving value from data floating in the organization environment. It provides methods and tools for collecting, storing, formatting and analyzing data for the purpose of helping managers in decision-making. At the start, only data from enterprise internal activities were examined. Now and in this turbulent business environment, organizations should incorporate analysis of the huge amount of external data gathered from multifarious sources. It is argued that Business Intelligence systems accuracy depends on quantity of data at their disposal, yet some storage and analysis methods are phased out and should be reviewed by academics and practitioners.

This paper presents an overview of BI challenges in the context of Big Data (BD) and some available solutions provided, either by using Cloud Computing (CC) or improving Data Warehouse (DW) efficiency.

Keywords:

Business Intelligence;
Big Data;
Data Warehouse;
Cloud Computing.

Author correspondence:

Hicham El bousty
Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Laboratory of Engineering Sciences and Energy, Agadir-Morocco

1. Introduction

*a,c,d,e,f,gPolydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Laboratory of Engineering Sciences and Energy, Agadir,Morocco

**bProfessor of computer sciences and Physics at Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Agadir,Morocco.

Data is a collection of unorganized raw facts. Putting this data together into a specific context generates information. Useful information constitutes knowledge which leads to better decisions-making. The cycle from collecting data to decision making is referenced as Business Intelligence. BI has attracted considerable attention as it gives rise to business value enhancement. Mining and statistical models need abundant data to perform better and return accurate results. In fact, large DWs with high velocity on which data are produced, provide ideal source for business analytics. Nevertheless, most of actual BI solutions cannot continue to monitor this rapid evolution.

This paper contributes to understand benefits and challenges of processing BD to strengthen BI analytics. In section two we present assets and challenges of incorporating BD in BI. Cloud BI which helps by expanding the physical resources needed to follow data explosion is introduced in section Three. The fourth section examines the impact of BD on data warehouses design and implementation; in section Five, conclusions and future investigation directions are discussed.

2. Big Data

Big Data is a concept that deals with formatting, storing and analyzing huge datasets. The first known use of the term Big data in the academic arena was in 1999 during an ACM communication "Visually exploring gigabyte data sets in real time," by Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes [1]. Authors used BD when referring to huge datasets which do not fit in a single memory. In term of volume, BD is not a static notion, what is considered BD today will not be in few years. In 2013 the generated data counts 4.4 zettabytes [2]. This amount was multiplied four times in 2016. By 2025, the datasphere will reach 163.1 zettabytes, as claimed by IDC [3]. Velocity and variety are also determinant criteria of BD Figure 1. Velocity is the cadence on which data are generated or transmitted. The greater the velocity, the more challenging it becomes to store, treat and analyze data. Variety, on the other hand, supposes multiple sources of data. This wide range of connected devices, PDAs, cameras, mobile phones, iPads and smart televisions involves different data forms. Retrieving knowledge and insights from data requires agile combination of all these forms.

Figure 1. Big data 3 Vs

2.1. Data evolution

Initially, websites were consulted for retrieving information about many subjects. It was a unidirectional interaction from the website to the internaut; however, now we are writing to the web too. Social media is the major example of user's contribution in creating web content. They share their own experience or learn about others' creations [4]. They also express their opinions, sentiment and judgments about created contents. According to Omnicore Agency, active twitter users attended 100 million per day, they share about 500 million tweets [5]. Technological companies as Google capture every action of every single user. The Information generated about

individual are greater than those written by this one [6], they are more powerful in terms of business value. This evolution was not possible without the reduction of storage costs [6]. A gigabyte of hard drive storage cost about 400 000$ in 1980. According to Statistic Brain Research Institute, the same budget allows storing 20 petabytes on 2016 [7] Figure 2.

Increasing number of connected devices (31 percent more devices on 2017 [8]) has participated actively in this data revolution. Cisco predicts that the internet of things will generate 600 ZB every year by 2020; however only a small amount is transmitted to data centers, about 6 ZB [9]. For example, connected plane transmit 0.1 percent of 40 TB generated every 1 hour of flying.

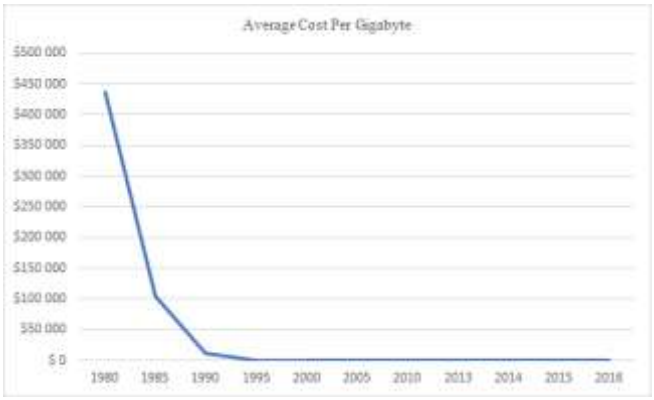

Figure 2. Evolution of hard drive costs

2.2. Big Business Intelligence

Sometimes BI is defined as the capacity of a business to use all available data [4], furthermore it is the ability to take advantage from technologies to exploit this data. In a nutshell BD is a BI booster.

BI has focused for long time on signals coming from inside the enterprise as production strategies, CRM systems and markets insights, in taking decision process. Analyzing data from outside has unlocked unparalleled business opportunities [4]. Different social fields are now seen as BD problems. In 2007, when the H1N1 flu struck in the United States, the Center for Disease Control tried to predict the spread of the disease. Forecast based on the transmitted data from local teams, engendered a week or two of delays compared to the real spread of the disease, due to infected persons consulting doctors days after their first flu symptoms and time needed for data collection and treatment. Later Google points out the need to develop a mathematical model based on the search engine. It correlates users' searches on google in a region and the high probability of infection. This model has proved highly accurate, and has helped health organizations to control the next H1N1 strike in 2009 [10].

For brands and businesses, BD permits fastest, reactions to both internal and external indicators [11]. Starbucks company surveyed blogs and social networks groups to collect customers' opinion about a new coffee product, the product tasted good, but they judged that it was too expensive [12]. The company lowered the price the same day and customers were satisfied. Considering this example, it is evident that BD enhance the dynamic and adaptive capability [13] for pricing, risk management, fraud detection and personalized customer services and measurement of satisfaction [14,15].

Today most managers are aware of the importance of BD in fields that were one dominated by intuition. Experienced managers are very well paid, since they have a very developed sense of prediction. However, the way managers are rewarded is about to change [4] if BD can overcome problems of small samples and data lack as claims McAfee and Brynjolfsson [16] "Big data leads to better predictions and better prediction yields better decision.".

2.3. Big Business Intelligence challenges

According to a research conducted by TDWI organization, most enterprises consider BD an opportunity. However, 30% still see BD as a "problem" [17]. The more data unlock new business opportunities, the more challenging storage and analytics become. Managers worries are mostly a result of confusion about the way to manage this huge mass of data. The large volume of data, the increasing speed at which data is generated and the wide variety of data types are all issues of concern to academics and professionals. Over the last 20 years, only 4 ZB capacity of hard drive was shipped which is equivalent to the data generated in 2013 [3]. Therefore, data will always surpass our capacity of storage. Data scientists and leaders suggest then focusing on the most useful data [3]. Some others have begun to question data fidelity.

No one argues the assets of BD, but the way information "is used, shared, archived and managed, "[6] is crucial for data quality. BD systems should ensure data integrity. Altering some information, halts system performance and accuracy. IDC emphasizes that only a part of critical data is secured. It claims that by 2025, 90% of created data should be secured, but only half will be protected [3]. Considering data fidelity, it is not only integrity that matters, for retrieving business insights. Objectivity is essential too [4]. Facebook likes, for example, could be irrelevant when thinking about people having more than one account and accounts that are owned by non-humans.

Human language tends to change over time. We are expressing things differently than decades ago. Sentiment analysis systems, focuses on human expressions about brands, products or even people. Efficiency of these systems supposes the right interpretation of human's words. Language evolution presents the main inconvenience to those systems. Moreover, the same situation is generally expressed in different words and images by people [4]. An emoji can express sadness for some, whereas it is just a sign of indifference for others.

3. Cloud Business Intelligence

3.1. Assets

One of the biggest challenges for organizations in the last decade is moving from on-premise conception, to a cloud solution. Moreover, most BI researches are based on the cloud. On premise software has been built on restricted data storage and limited power infrastructure [18]. Insufficient hard for stocking all data pushes organizations choose which data should be kept and which are not important for the organization progress. To stay competitive, an enterprise should consider data from all functions such as production, marketing, finance, besides this, experts advise taking advantages from unstructured data as well, like that provided from social media and connected objects.

Cloud computing is recognized as the evident answer for business computational elasticity worries. It provides the hardware, networking, security and software needed for immediate deployment of a BI solution [19]. Moreover, small businesses are now able to take profit from BI with a reduced capital expenditure. They pay only as the user go on the system [20]. Operational expenditure includes security management and assurance of services continuity; therefore, no support IT team is needed. BI editors guarantee money saving with their solutions. However, a return of investment indicator (ROI) should be calculated before adopting cloud or non-cloud BI. Lekha Menon et al.[21] emphasized that the ROI has two components: financial ROI shaped by revenue variation and the nonfinancial ROI which measures performance, user satisfaction and accuracy of results.

In addition to long term cost reducing [22], cloud computing showed tangible enhancements in terms of storage, performance and accuracy. Many experiments were conducted comparing cloud and non-cloud computational capacity. The major benefit of CC is scalability. Enterprise has no more to invest on huge technology infrastructure to scale up her capabilities. Furthermore, a series of experiments, conducted by Victor Chang et al. [22], comparing backup on both cloud and non-cloud storage systems of 10 TB data, demonstrates that

the execution time on the cloud is lower than non-cloud and there was a 99% consistency between the actual and expected execution time on the cloud and non-consistency on the non-cloud storage platform.

The 2008 crisis, forces scholars to reconsider financial risks analyses and forecasting systems. The Gaussian copula model makes some simplification allowing risks analyses running on desktop software on an acceptable time. The model underestimated risks on extreme conditions. The development of new model improvingaccuracy needs powerful resources. Hence, Victor Chang et al. [23] suggested a model based on the cloud technology, which achieves 95% accuracy as compared to the real stock index.

3.2. Challenges

It's obvious that CBI has many benefits, especially elasticity, scalability, mobility and immediate deployment as well as implementation costs reducing. However, organizations have some challenges to overcome when implementing CBI. Exchanging data over the internet with applications lodged on a third-party data center makes it subject to different attacks. Security is the major concern when implementing CBI. Adopting a hybrid cloud BI should minimize the risk of compromising enterprise security Figure 3. Highly sensitive data are stored and accessed locally, whereas low sensitive data are moved to the cloud. Encryption may also be used to secure data on cloud servers. Nevertheless, key management and performance drops are additional issues to consider when adopting an encryption approach.

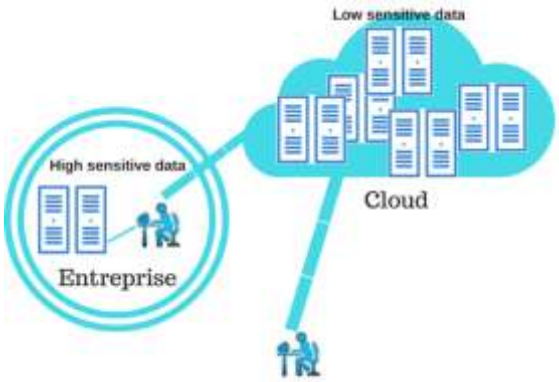

Figure 3. Hybrid Cloud Business Intelligence

Al-Aqrabi et al.[24] compared two cloud securing models, one is based on the Unified Threat Management (UTM) approach, and a second based upon cloud embedded security components. UTM cloud hosts the stateful inspection firewall, LDAP, anti-spam, intrusion detection and prevention systems and applications security components. Users' traffic is routed through all UTM servers before attending application servers Figure 4.a. On the embedded security components model, cloud UTM is eliminated and traffic is sent to cloud switches which distribute the load uniformly among application servers Figure 4.b. Further, security services are enabled on all servers. Hence, security management is a difficult task on this model and security breaches can easily be spotted by experienced hackers. However, the embedded security components model showed better performance compared to UTM. Al-Aqrabi et al. [24] suggested combining UTM and cloud embedded components for more efficient security. The move of applications security components from UTM cloud to BI decreases database queries traffic and improves performance.

Security is not the only challenge of CBI implementation. Availability and latency can also discourage organization from opting for a CBI. A brief suspension of a Forex BI service for example, can be costly for business. That is why organizations should give weight to select the suitable solution which maintains the quality of service required [18].

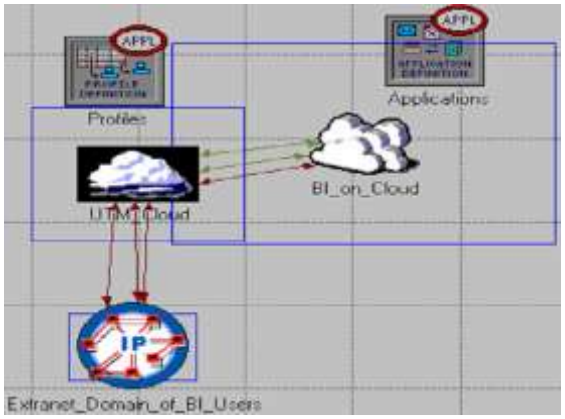

Figure 4.a. access of BI users through a UTM cloud [24]

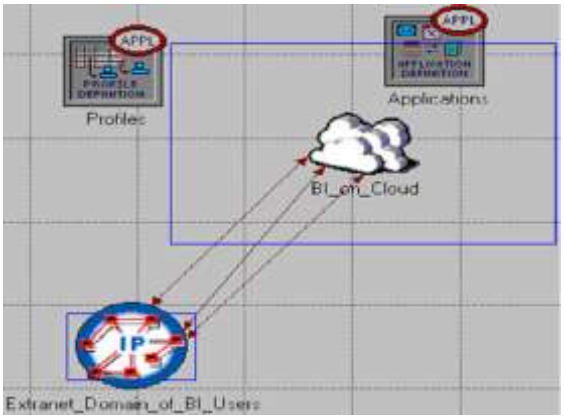

Figure 4.b. users directly connected to the BI application servers on the cloud [24]

4. Data Warehouse

Before any BI system can provide meaningful information for end users, stored data passes through several stages. First, data sources should be defined thoroughly. Data preparation operations are all organized under the Extraction, Transformation and Load (ETL) tools. Load step supports data integration in separate storage repository, commonly called Data Warehouse. The data interrogation and reports generation are performed by OnLine Analytical Processing applications (OLAP). These BI components experienced changes over time due to performance, accuracy and latency requirements.

Production databases are mainly designed to perform transactional operations related to the business. OnLine Transaction Processing (OLTP) are systems maintaining this operational activity. All users interact with the OLTP systems to undertake production tasks and missions. OLTP workloads are small write and read transactions (INSERT, UPDATE, SELECT). On the other hand, OLAP systems are designed for BI purpose. OLAP transactions are at most read queries involving aggregated data. OLTP and OLAP cannot use the same database efficiently, since the first contains details of current data, whereas OLAP deals with historical information [25]. Besides, the use of the same database for production and analysis harms system performance and user satisfaction, therefore DWs are used for storing reconciled data. Inmon [26] define DW as “A subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process.”. Furthermore, a well thought out DW schema may reduce the execution time of a query by 90% compared to the time needed in OLTP systems [27].

Today data sources are multiple, each produces outputs in different structures and formats. Moreover, companies are increasingly dealing with unstructured data (text files, audio and video), hence the necessity of restructuring and cleansing data. Extract Transform and Load (ETL) tools provide all functionalities needed for data preparation and integration. It is the major component of BI system [28]. Mainly, ETL are composed of 3 processes Figure 5:

Extract: Is the act of retrieving data from all sources and storing it in a staging area. The extraction process may harm the operational transactions and cause performance drop.

Transform: Data gathered are cleansed, aggregated and formatted to fit DW schema [29]. The more this data load, the more time and efforts needed for treatment are significant.

Load: Resulted data from the previous processing is loaded to the DW.

Figure 5. ETL process

4.1. Active Data Warehouse

One new trend for BI is the real time BI. Traditionally data is refreshed offline at a fixed time during the day, generally a period of downtime. Companies need to understand customers' expectations as fast as possible since too much latency may cause considerable damage for a business [30]. Besides, internet websites might be part of data sources for DW. Web datasets tend to disappear or morph over time. This trend begets new challenges for ETLprocesses. The ETL should refresh DW continuously to serve last data for analysis. In a such situation, the ETL become a consuming process and sources experience overhead problems steadily. To cope with source overhead, streaming regulators are used. In the case of multi-source, Data Processing Flow Regulator (DPFlowR) indicates which source is ready to transfer changes, captured by source flow regulators (SflowR) hosted on each source [29] Figure 6.a. The DPFlowR manages resources and priorities based on workload and administrator configurations. Sometimes a delay in transiting data may occur. It is therefore a Near Real Time Data Warehouse (NRTDW) configuration as DW is not updated the moment modifications take place. Despite the improvement of data refreshing in the NRTDW, short data disparity may cause considerable loss, particularly when it comes to critical data. High priority should be attributed to critical data. It should be propagated to DW the moment changes occur. Tanvi Jain et al. suggest a near real time refreshing model for handling critical data. Update significance is a parameter measuring the impact of an update [31]. In the banking field, a transaction exceeding a threshold can be considered as highly significant. Two queues are established one for critical data which is refreshed in the real time, and the other for less significant data transferred at a scheduled time. The difficulty with this proposition is identification of critical data which is a subjective assessment depending on the situation and analysis purpose. Moreover, the combination of some meaningless records may reveal some hidden business insights. Mischievous bankers exploit their clients' accounts to build a fortune. They make very small transactions from millions of accounts to their own. These transactions records will not be propagated in real time, since the value of transaction is negligible (update significance = 0). Hence, Fraud analyze algorithms will not detect bankers' attempts as it needs the latest data. The incorporation of Update Significance measure in the regulators ETL model may have better results with regards to maintaining acceptable performance and prioritizing refreshment of critical data Figure 6.b. Each data source regulator holds critical data queue that is refreshed at any record update. This model should minimize overhead problems while serving the DW with the most relevant information in a real-time manner. The rest of data is propagated as much as possible according to the workload.

Overload issue may also affect DW just as the way it affects data sources. OLAP application query DW to undertake users' requests. At the same time, new data are uploaded to DW. In practice, data are kept on a staging area during the transformation process. Propagation to DW should be controlled by Warehouse Flow Regulator (WFlowR) to curtail overload troubles Figure 7 [29].

In this era of BD, ETL becomes very demanding process in regard to resources. Any extension of CPU and/or RAM will be quickly overwhelmed by realistic workload. In recent years, there has been growing interest in parallel tasks thanks to the emergence of Map/Reduce paradigm. A comparison between commercial ETL and Map/Reduce based one, showed about 70% gain of

performance when using Map/Reduce ETL over commercial ETL. Authors in [32] confirmed this conclusion. They addressed an architecture of parallel ETL (P-ETL) based on Map/Reduce. Extracted data from a source are physically partitioned on several racks and loaded to Hadoop Distributed File System (HDFS). Logical partitioning is mandatory in the distributed environment. Each partition is handled by a mapper which performs a specific operation of cleansing and formatting (i.e. filtering, projection, conversion, concatenation). The reducer in P-ETL is responsible for the merging and aggregating operations. The implementation of P-ETL proved the assets of this architecture. Indeed, experiments showed that the processing time decrease when raising the number of tasks. However, adding new tasks has no leverage when memory and HDD are absorbed by those running.

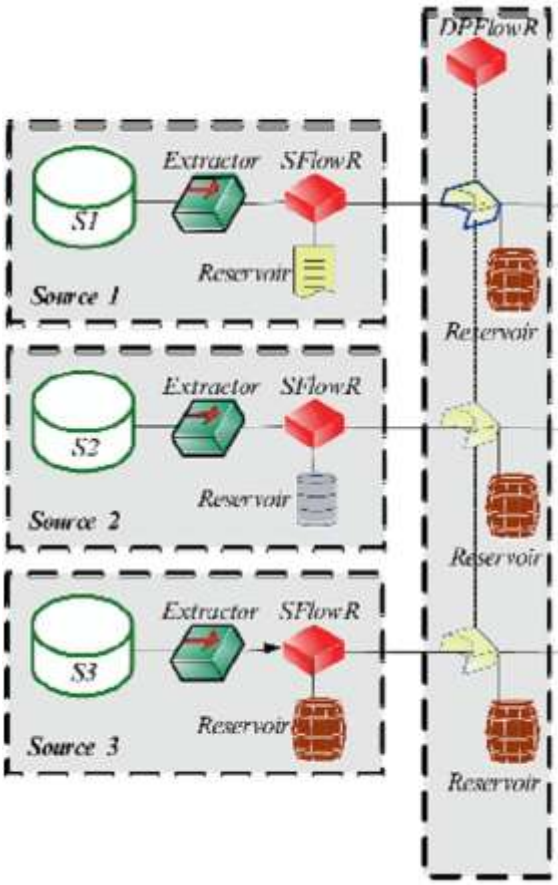

Figure 6.a. Extraction process of near real-time data warehouse [29]

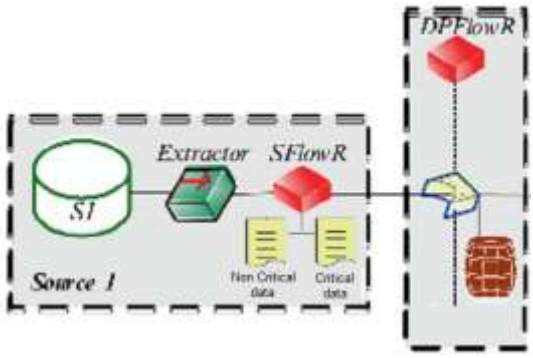

Fig6.b: Extraction process with two queues

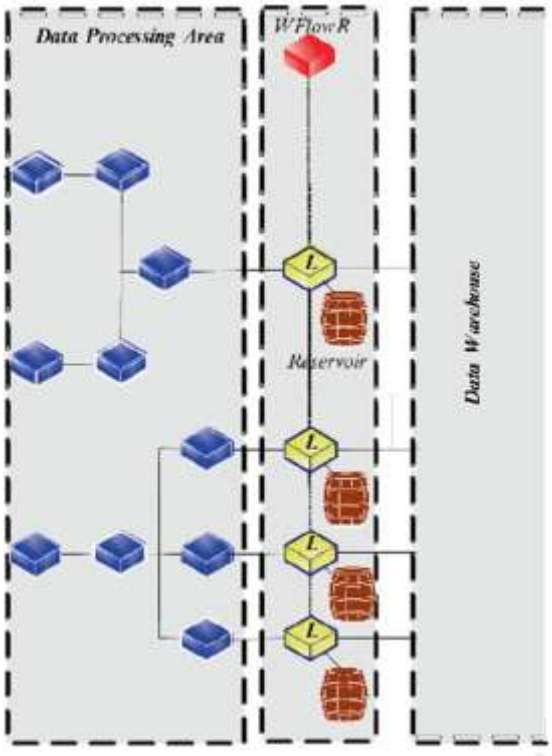

Figure 7. Load process of near real-time data warehouse [29]

4.2. Semantic Data Warehouse

Performance is not the only concern for ETL process, results accuracy has received much attention in the past decade. The inclusion of web sources raised new challenges for BI. Social data has generally complex or poor structure. Therefore, traditional ETL are unable to process this information. A new class of ETL has emerged, Semantic ETL. Most Semantic ETL projects are based on the ontology which defines a set of concepts for a specific domain of knowledge and how these concepts are linked [33,34]. Each concept has proprieties, the car concept for example has model year and license plate as proprieties. The concept car is also linked to the concept human, since every car is owned by a person. The query "who has a red car" would not be successful unless a set of cars and owners are stored in a structured manner. Semantic Web (SW) aims to retrieve data from unstructured and even unknown sources based on defined ontologies. SW encompasses all standards which allow describing, publishing, connecting and sharing data. Resource Description Framework (RDF), is a semantic web language, which models information in a web resource. RDF uses triples (subject, predicate, object) to represent an information over the web, for instance, “John owns a red car” expression is represented by a triple (John, owns, red car). Datasets in semantic ETL are mapped to triples according to an ontology data model which are loaded to an ontology based database Figure 8.

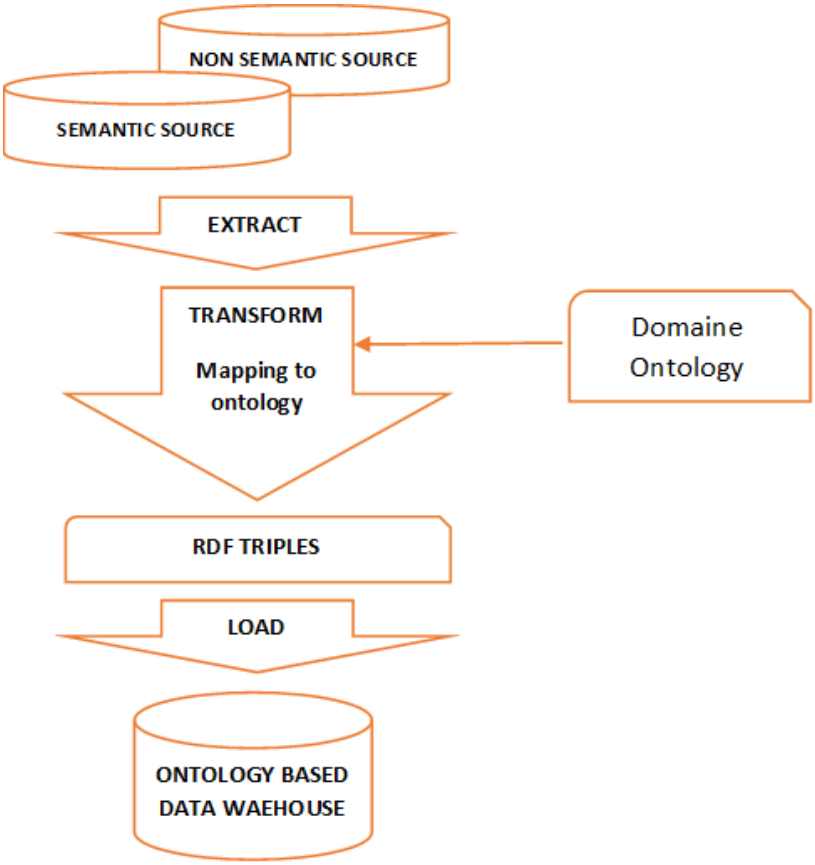

Figure 8. Semantic ETL process

Yet, some SW problems should be fixed before any efficient integration into the DW. Indeed, there are some web contents which are not annotated. Authors in [35] proposed an approach to partially annotating textual content from web sources. Besides, the utter variety of tools and languages for creating ontologies imposes enhancement of interoperability [36].

4.3. Data Warehouse query processing

For the purpose of increasing the QAS of BI applications, various techniques have been put forward. One obvious optimization method is the use of partitioning. Tables are split in vertical or horizontal way to speed up queries execution. In real time DW, data are growing faster so partitions. Instead of executing merge and split operations, partitioning algorithms such as Uniform Partitioning Algorithm (UPA), reinitialize the whole process which is a resource consuming task. Hence, the traditional partitioning is not suitable for the RTDW. Authors in [37] designed a partitioning algorithm, Near-uniform range Partitioning Algorithm (NPA), that calculates the minimum and maximum size of a partition. A partition is split when it exceeds the maximum size and merged with the next partition if it has low size. Experiments conducted unveil that NPA is four times faster than UPA. Issam Hamdi et al. [30] designed a tow level partitioning approach. The second level exploit the NPA algorithm to adjust partitions whereas the first level defines the number of clusters to be used for partitions based on probabilistic methods.

To perform analyze request, join queries are used frequently. The optimization of join algorithms may increase the DW performance. Thus, different join algorithms have been developed. M-HYBRIDJOIN is a stream based algorithm settled memory management problems of its predecessor (X-HYBRIDJOIN). Comparison of M-HYBRIDJOIN with some join algorithms (Meshjoin,R-Meshjoin,.) confirmed its high performance [38].

4.4 Column oriented Data Warehouse

In column oriented database, column items are stored sequentially one after another. OLAP queries need generally aggregated data of a single or few columns. Retrieving a column values from a row database, involves reading all attributes in each record. Therefore, performance drops due to excessive use of I/O. Unlike row stores, column oriented database reads only needed values for executing a query. It is more appropriate for running BI applications. Authors in [39], constructed a column database by partitioning vertically a row oriented database. This structure can be reached by using Binary Association Table. Each original row is mapped to a BAT which consists of two columns table (Row id, Attribute value) [40] Figure 9. Performing some queries on both original and mapped database, showed better performance of the column database.

id	name	balance
1	Alissa	100.00
2	Bob	200.00
3	Charles	300.00

(a) Row-Based Table `customer`

oid	int
101	1
102	2
103	3

(b) BAT `customer_id`

oid	varchar
101	Alissa
102	Bob
103	Charles

(c) BAT `customer_name`

oid	float
101	100.00
102	200.00
103	300.00

(d) BAT `customer_balance`

Figure 9. Customer Data in Row-Based and Column-Store (BAT) Format [40]

DWs are read optimized databases, however, to update or insert records, write operations are undertaken. Active DW has raised the challenge of improving write performance, since records insertion is performed continuously as much as read operations. Seeking Row Id in BAT, at each update or deletion operation, is time consuming. Qiming Chen et al. [40]., proposed a refinement of the BAT model in an out of core database environment. Timestamped BAT (TBAT) holds time of the update operation along with Row Id and Attribute Value. Each new update is appended to the end of TBAT. Changes are performed in an asynchronous manner, out of peak time moments. Thus, seeking and writing operations are reduced and only last updates are completed. This method is referenced as Asynchronous Out-of-Core Update.

Speed up the query processing is not the only advantage of using column oriented database, it also reduces the size of database storage as well. Dictionary based compression, sorts and stores distinct attribute values. Columns keeps only references to encoded data. Thus, compression algorithms perform better on column oriented approach, since there is high chance to have similar values in adjacent items [41]. Nonetheless, compression may introduce delay at each new records insertion. In fact, after an update task, new values might be inserted into dictionary and compression functions are recalled. For the purpose of keeping acceptable response time of write operations, a separate buffer structure is created. SAPs NetWeaver Business Warehouse Accelerator uses buffer delta, that allows high modification rates, to store new records [42]. Forbetter search performance, dictionary of the main storage should be updated. The main and delta storage are then merged.

4.5 Nosql Data Warehouse

Most DW solutions are implemented on RDBMS which are strictly structured and support the ACID properties (Atomicity, Consistency, Isolation, Durability). In the BD analytics context, reliability and integrity can be sacrificed to gain more performance. Yet, Relational databases are not suitable for unstructured and massive datasets. They don't scale out as they grow. Alternatives are referenced as Not only SQL (NoSQL) databases. they are schema-less model or have a lightweight schema. Besides, unnecessary RDBMS features have been omitted. Many experiments confirmed the efficiency of NoSQL on managing BD. Response time of insert, select, update and delete operations was lower in MongoDB (NoSQL) compared to PostgreSQL (RDBMS)[43]. These finding supports previously obtained results by Chieh Ming Wu et al. who compared Ms SQL to MongoDB [44].

It is clear that NoSQL database is better suited to BD environment. However, they are not as secure as RDBMS. Indeed, some NoSQL implementations have weak authentication between clients, and the server and no encryption of clients' communication or datafiles [45]. MongoDB uses JSON format for data and queries representation. It is well defined and provides better security measures, nevertheless its exposure to SQL injection is high as much as SQL databases [46].

4.6. In memory Data Warehouse

Performance is generally affected by read and write operations to disk storage. The number of I/O shapes the behavior of any system. In the context of real time, DW does no more tolerate delay introduced by I/O. Hence, in-memory DW is proposed as alternative to disk resident DW. Whole database data are stored in the physical main memory. The decrease of main memory costs Figure 10. and appearance of 64-bit processors, which allows more addressable memory, have fostered the use of main memory databases [27]. Several experiments showed the high performance of in-memory in running queries. Nevertheless, a comparison between FastDB, a main memory database, and MS-SQL showed that FasDB performance is lower than MS-SQL for heavy workloads [48]. Indeed, the decrease of available memory pushes the system to swap between main memory and disk, which affect main memory behavior. Scaling out memory may maintain steady response time of FastDB. However, the scalability of in memory database is expensive compared to resident disk database.

Volatility is another inconvenience of using main memory database. The data lifetime is as long as the power is applied. In principle, data should be backed in a nonvolatile media. However, the restoration process may take several hours, and the database is out of service. Non-Volatile Memory which is as fast as DRAM settle the volatility issue. NVM maintain data even after a power loss. Though, most of current DWs should be adapted to the NVM since they manage only volatile memory [49].

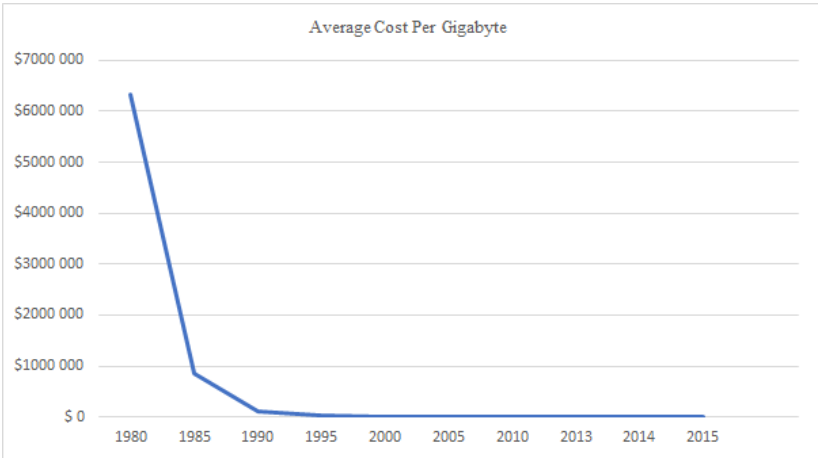

Figure 10: Evolution of RAM average costs [47]

5. Conclusion and future directions

BD has unlocked new business insights. However, this concept has uncovered limits of some actual technologies. To get benefits from all available data, servers' storage memory and processing capacity should often be extended. This evolving need, engender additional investments, which become affordable thanks to cloud computing. Many techniques and DW implementations have been developed over the last decade. A DW framework combining all presented techniques may be proposed in a subsequent research. Besides this, analytics techniques may also be reviewed in a future paper, in order to constitute with the present paper a complete reference in the field of BI and analytics.

References

[1] S. Bryson, D. Kenwright, M. Cox, D. Ellsworth, and R. Haimes, "Gigabyte Data Sets in Real Time," Commun. ACM, vol. 42, no. 8, pp. 82–90, 1999.
[2] E. D. U. with R. & A. by IDC, "Executive Summary Data Growth, Business Opportunities, and the IT Imperatives." [Online]. Available: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.
[3] D. Reinsel, J. Gantz, and J. Rydning, "Data Age 2025: The Evolution of Data to Life-Critical," IDC White Pap. Spons. by Seagate, no. April, pp. 1–25, 2017.
[4] C. Kimble and G. Milolidakis, "Big Data and Business Intelligence: Debunking the Myths," vol. 35, no. 1, pp. 23–34, 2015.
[5] O. Agency, "Twitter by the Numbers: Stats, Demographics & Fun Facts," 2017. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/.
[6] J. Gantz and D. Reinsel, "Extracting Value from Chaos State of the Universe: An Executive Summary," IDC iView, no. June, pp. 1–12, 2011.
[7] Statistic Brain Research Institute, "Average Cost of Hard Drive Storage," 2015. [Online]. Available: http://www.statisticbrain.com/average-cost-of-hard-drive-storage/.
[8] Gartner, "Newsroom announcment," 2017. [Online]. Available: https://www.gartner.com/newsroom/id/3598917.
[9] Cisco, "Cisco Global Cloud Index : Forecast and Methodology , 2014–2019," White Pap., pp. 1–41, 2014.
[10] K. C. Viktor Mayer-Schonberger, Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, 2013.
[11] H. Kościelniak and A. Puto, "50% BIG DATA in Decision Making Processes of Enterprises," Procedia Comput. Sci., vol. 65, no. Iccmit, pp. 1052–1058, 2015.
[12] H. J. Watson, "Tutorial : Big Data Analytics : Concepts , Technologies , and Applications Tutorial : Big Data Analytics : Concepts , Technologies , and Applications," vol. 34, no. January 2014, 2014.
[13] S. Erevelles, N. Fukawa, and L. Swayne, "Big Data consumer analytics and the transformation of marketing," J. Bus. Res., vol. 69, no. 2, pp. 897–904, 2016.
[14] S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," Electron. Mark., vol. 26, no. 2, pp. 173–194, May 2016.
[15] T. H. Davenport, "How Strategists use 'big data' to Support Internal Business Decisions, Discovery and Production," Strateg. Leadersh., vol. 42, no. 4, pp. 45–50, 2014.
[16] A. McAfee and E. Brynjolfsson, "Big Data. The management revolution," Harvard Buiness Rev., vol. 90, no. 10, pp. 61–68, 2012.

[17] P. Russom, "Big data analytics," TDWI Best Pract. Rep., pp. 1–35, 2011.
[18] R. Vidhyalakshmi and V. Kumar, "Challenges in Implementation of Cloud Analytics," no. June, 2016.
[19] R. Alsufyani and V. Chang, "Risk Analysis of Business Intelligence in Cloud Computing," 2015 IEEE 7th Int. Conf. Cloud Comput. Technol. Sci., pp. 558–563, 2015.
[20] "From Mainframe to Cloud Computing : A Study of Programming Paradigms with the Evolution of Client-Server Architecture From Mainframe to Cloud Computing : A Study of Programming Paradigms with the Evolution of Client- Server Architecture," no. February, 2015.
[21] L. Menon and B. Rehani, "Business Intelligence on the Cloud: Overview and Use Cases," pp. 25–30, 2011.
[22] V. Chang and G. Wills, "A model to compare cloud and non-cloud storage of Big Data," Futur. Gener. Comput. Syst., vol. 57, pp. 56–76, 2016.
[23] V. Chang, "The Business Intelligence as a Service in the Cloud," Futur. Gener. Comput. Syst., vol. 37, pp. 512–534, 2014.
[24] H. Al-Aqrabi, L. Liu, R. Hill, Z. Ding, and N. Antonopoulos, "Business intelligence security on the clouds: Challenges, solutions and future directions," Proc. - 2013 IEEE 7th Int. Symp. Serv. Syst. Eng. SOSE 2013, pp. 137–144, 2013.
[25] K. Kakish and T. a Kraft, "ETL Evolution for Real-Time Data Warehousing," Proc. Conf. Inf. Syst. Appl. Res., no. September, pp. 1–12, 2012.
[26] W. H. Inmon, "What is a data warehouse?," Prism Tech Top., p. 19, 1995.
[27] S. S. Baboo and P. R. Kumar, "Next Generation Data Warehouse and In-Memory Analytics," Int. J. Comput. Appl., vol. 69, no. 18, pp. 25–30, 2013.
[28] S. Misra, S. K. Saha, and C. Mazumdar, "Performance comparison of hadoop based tools with commercial ETL tools - A case study," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8302 LNCS, pp. 176–184, 2013.
[29] S. Kozielski and R. Wrembel, New Trends in Data Warehousing and Data Analysis, vol. 3. 2008.
[30] I. Hamdi, E. Bouazizi, S. Alshomrani, and J. Feki, "2LPA-RTDW: A Two-Level data Partitioning Approach for Real-time Data Warehouse," 2015 IEEE/ACIS 14th Int. Conf. Comput. Inf. Sci., pp. 632–638, 2015.
[31] T. Jain and S. Saluja, "Refreshing Datawarehouse in Near Real-Time," Int. J. Comput. Appl., vol. 46, no. 18, pp. 975–8887, 2012.
[32] [32] M. Bala, O. Boussaid, and Z. Alimazighi, "P-ETL: Parallel-ETL based on the MapReduce paradigm," Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA, vol. 2014, pp. 42–49, 2014.
[33] [33] L. Bellatreche, S. Khouri, and N. Berkani, "Semantic data warehouse design: From ETL to deployment à la Carte," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7826 LNCS, no. PART 2, pp. 64–83, 2013.
[34] [34] R. Pratap, D. Nath, K. Hose, and T. B. Pedersen, "Towards a Programmable Semantic Extract-Transform-Load Framework for Semantic Data Warehouses Presentation," pp. 1–35, 2015.
[35] [35] D. Sánchez, D. Isern, and M. Millan, "Content annotation for the semantic web: An automatic web-based approach," Knowl. Inf. Syst., vol. 27, no. 3, pp. 393–418, 2011.
[36] [36] I. Bajwa, "A Framework for Ontology Creation and Management for Semantic Web," Int. J. Innov. Manag., vol. 2, no. 2, pp. 116–118, 2011.
[37] [37] J. Song and Y. Bao, "NPA: Increased partitioning approach for massive data in real-time data warehouse," 2010 2nd Int. Conf. Inf. Technol. Converg. Serv. ITCS 2010, 2010.
[38] [38] S. Sudha and R. Scholar, "M-HYBRIDJOIN-AN ADAPTIVE APPROACH FOR STREAM BASED NEAR REAL-TIME DATA WAREHOUSING Address for Correspondence," Int. J. Adv. Eng. Technol. E Int J Adv Engg Tech, pp. 321–326, 2016.
[39] V. Shukla and R. Tiwari, "Column Oriented Database : Implementation and Performance Analysis," vol. 4, no. 9, pp. 2013–2015, 2015.
[40] F. Yu, T. Matacic, W. Xiong, M. A. A. Hamdi, W. Hou, and C. Science, "Data Cleaning in Out-of-Core Column-Store Databases : An Index-Based Approach," pp. 16–22.
[41] G. Matei, "Column-Oriented Databases, an Alternative for Analytical Environment," Database Syst. J., vol. I, no. 2, pp. 3–16, 2010.
[42] J. Krueger, M. Grund, C. Tinnefeld, H. Plattner, A. Zeier, and F. Faerber, "Optimizing write performance for read optimized databases," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5982 LNCS, no. PART 2, pp. 291–305, 2010.
[43] M. G. Jung, S. A. Youn, J. Bae, and Y. L. Choi, "A study on data input and output performance comparison of MongoDB and PostgreSQL in the big data environment," Proc. - 8th Int. Conf. Database Theory Appl. DTA 2015, pp. 14–17, 2016.
[44] C. Ming Wu, "Comparisons Between MongoDB and MS-SQL Databases on the TWC Website," Am. J. Softw. Eng. Appl., vol. 4, no. 2, p. 35, 2015.
[45] C. Nance, T. Losser, R. Iype, and G. Harmon, "Association for Information Systems AIS Electronic Library (AISeL) NOSQL VS RDBMS -WHY THERE IS ROOM FOR BOTH," 2013.
[46] A. Ron, A. Shulman-Peleg, and E. Bronshtein, "No SQL, No Injection? Examining NoSQL Security," arXiv Prepr. arXiv1506.04082, no. July, 2015.
[47] [1] Statistic Brain Research Institute, "Average Historic Price of RAM," 2017. [Online]. Available: https://www.statisticbrain.com/average-historic-price-of-ram/.M.

[48] Rahgozar, M. Siadaty, N. Razavi, and F. Raja, "A Comparative Study Of Main Memory Databases and Disk-Resident Databases," Citeseer, vol. 2, no. 2, pp. 128–131, 2006.

[49] J. Arulraj and A. Pavlo, "Let's Talk About Storage & Recovery Methods for Non-Volatile Memory Database Systems," Proc. 2015 ACM SIGMOD Int. Conf. Manag. Data, no. 1, pp. 707–722, 2015.