International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Review on Business Intelligence and Big Data

Erkan Sirin*[1],Hacer Karacan[2]

Accepted : 21/10/2017 Published: 30/12/2017

Abstract: Improvement of data generating, processing, storing and networking technologies has made storing, capturing and sharing of data easier and cheaper than before and has enabled organizations to handle the huge volume of data at high velocity and variety, named as big data. Big data offers many opportunities when the associated difficulties are addressed properly. Business Intelligence (BI) basically focuses on transforming raw data into usable, valuable and actionable information for decision-making. It can be classified as a kind of data-driven decision support system. Although big data related papers have increased for last fifteen years, there are not sufficient papers that directly overviews big data impact on BI. As data is growing exponentially, storage, process, and analytics tools and technologies become more important for BI solutions. With the advent of big data, BI's concept, architecture, and capabilities are meant to be changed. Unlike decades before, BI now is to be extract value from huge data ocean by using big data tools as well as classical ones. So, an interclusion has emerged between big data and BI. This paper overviews the current state of the art of BI and big data, and discuss how big data era affects BI solutions in general context.

Keywords: *Big data, business intelligence, data mining, machine learning.*

1. Introduction

Technological advancements of IT have led to storing more data at lower cost and drastically increased transmitting rates. Parallel computing has increased computing power as well by processing multiple cores simultaneously. It is hard to find any device that doesn't generate data like sensors, plane engines, online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions, science data, and mobile phones. All of these and their applications have begun to generate huge volume data at high velocity and variety which is impossible to store and process with classical technologies and programming paradigms. This kind of data is called big data.

International Data Corporation (IDC) reports that digital universe will continuously expand, be complex and interesting. The volume of data is expected to be 8 ZB by 2020 [1, 2]. Data generation speed is also increasing exponentially. It is estimated that world data will double [1] or triple itself every year. Therefore this tremendous increase has caused finding new technologies, processing techniques, programming paradigms, and analyzing tools, which are out of classical technology capabilities, to deal with and then extract value from it.

Many companies have been using big data for a while and investing in it. Google processes hundreds of PB, Facebook deals with hundreds of TB contents, every day tens of TB videos are uploaded to YouTube, for example on average 72 hours video is uploaded in a minute [3]. Twitter has more than 550 million active users and they produce 9100 tweets every second. 3 billion pieces of content are generated on Facebook every day. A Boeing jet engine can produce 10 terabytes operational information in 30 minutes [4].

There is increasing publication about big data for last decades. Initially, publications were drawing attention what big data was, how important it was, and its opportunities and challenges. Then more detailed papers were published about big data impact on many diverse areas,from science to retails. Although BI has in common with big data, there are not sufficient papers about how big data impacts on BI. To fill this gap this paper will overview big data and BI then argue big data effects on BI solutions.

2. Big Data

2.1. Defining Big Data

Shekhar and Sharma [5] categorize big data definitions in three types: attribute, comparative and architectural. Attribute definition is originated from IDS reports and emphasizes on 4Vs of big data mentioned below. The comparative definition depends on McKinsey's report and compares big data with traditional data. Architectural definition emphasizes on horizontal scaling for efficient processing. Although there are many diverse definitions of big data, some main aspects have in common in all definitions: Big data; "*is huge volumes of data generated with high speed, and has varying degree of complexity and ambiguity that can't be processed, stored, and managed with traditional technologies, processing methods, and algorithms but needs new technology platform and architecture which enables high-velocity capture, discovery, and/or analysis to extract value economically*" [2, 6, 7]. In attribute definition, it is overwhelmingly accepted that big data has three main characteristics: volume, velocity, and variety annotated as 3Vs of big data [8]. Volume means generation and collection of huge volume of data, which cannot be stored, managed and analyzed in traditional databases; velocity means generation and collection of data at high rate and should be processed and analyzed timely, like flood; and variety corresponds to the different kinds of data, mostly unstructured and semi-structured, such as logs, texts, videos, audios, webpages etc. Big data brings not only huge volumes of data together but also

[1] *Institute of Information Management Information Systems, Gazi University, Turkey*
[2]*Computer Engineering Department, Gazi University, Turkey*
* *erkan.sirin@gazi.edu.tr*

different kinds of data that was unimaginable to get together before. It is easier to handle and analyze structured data than unstructured one, for example using clustering functions on structured data is easier. Recently, the fourth and fifth V, even more, have been added to big data characteristics [8]. The fourth V stands for value which means discovering the values from big data [9]. Advances in storage and mining technologies have enabled to collect, analyze and mine huge volumes of data and yield valuable and actionable insights. The value of the restless accumulation of huge volumes of data resides not only in the quantity but also in new insights that enable decisions and actions to transform the economy and society [10]. The fifth V stands for veracity which refers to uncertainty and biases, noise and abnormality in data. Veracity in data analysis is one of the biggest challenges when compares to things like volume and velocity. Visibility can be added as sixth V, refers to properly presentation and abstraction of data in order to make an informative decision.

2.2. Challenges of Big Data

Although big data offers many opportunities, those who want to benefit from it have to confront its challenges. Traditional database systems, mostly rely on relational database management system (RDBMS), exploit structural data which is easier to handle than unstructured and semi-structured. Moreover, RDBMS has matured over years. But hardware requirement for RDBMS is getting more expensive to handle increasing performance expectations and growing datasets. RDBMS seems restricted in terms of data structure, volume, and heterogeneity of big data. Since traditional databases are restricted and come to an edge of hardware technology, both academia and industry sought to find new ways and paradigms to meet big data requirements.

Use of big data's full potential requires some issues to handle such as data policy, technology and techniques, organizational change and talent, access to data, and industry structure [11]. Big data challenges can be listed as following [12-15]:

- Data representation
- Redundancy reduction and data compression
- Data lifecycle management
- Analytical mechanism
- Data confidentiality
- Energy management
- Expendability and scalability
- Cooperation

2.3. Opportunities of Big Data

Big data brings new opportunities for discovering new values, helps to gain an in-depth understanding of the hidden values. Big data enables managers to make more informed decisions and companies to better understand their markets, customers, and suppliers respectively. Having big data, any organization can benefit from it by carefully analyzing to gain insights and depths to solve real problems. Since organizations are exposed to more data, decisions should no more based on instinctive, guesswork and stochastic events but clean, understandable, correct, and ready to use valuable information.

Having huge datasets is meaningless unless they produce value. One of the most important opportunities that big data offers is big data analytics (BDA). In order to determine research parameters, traditional statistics uses sampling that is supposed to represent the universe. This sample is analyzed within some probabilistic boundaries and offers an estimation of the universe parameters. But with big data, it is quite easy to work on whole universe data which leaves sampling unnecessary. Furthermore, it prevents sampling errors; hence wrong results.

There are many advantages in the business including increased operational efficiency, informing strategic direction, developing better customer service, identifying and developing new products and services, identifying new customers and markets, etc. For example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 [12].

3. Big Data Value Chain

To examine big data more methodically, value chain of big data is used which is divided into four consecutive phases: data generation, data acquisition, data storage and data analysis [5, 14].

3.1. Data Generation

It means how big data is generated from various sources. It is estimated that 90% of data today has been created in the last two years. This can give an idea how fast data is generated all over the world. Data generation is the first step of big data value chain. Search entries, posts, chatting records, sensors, videos, clickstreams, e-commerce data, Internet of Things (IoT), scientific research data are all example data sources. These datasets are distributed and generated at large scale and meaningless on their own unless accumulated in a big data repository and exploited. Moreover, these datasets don't fit in traditional IT infrastructures and need more computing capacity.

3.1.1. IoT Data

In IoT domain a lot of machine sensors deployed all around a wide network, depending on its function, generate various kind of data at different phases. Washing machines, lighting, alarm, GPS, mobile phone, fridges etc. can be given as example IoT sources. It is believed that IoT data will compose a great part of big data in the near future [14].

3.1.2. Enterprise Data

Enterprise data is the most prominent domain of big data. It was estimated in 2011 that volume of business data would double every 1.2 years [11]. Enterprise data is mainly structured data and managed by RDBMS and it holds every possible recordable action and activity of an enterprise like CRM, ERP, sales, financial, and production. For example, Walmart's 6000 stores all over the world generate approximately 267 million transaction data and in order to take advantage of this huge data, Walmart established petabyte-scale data warehouse (DW) [14]. As storage, processing, networking, and data management technology advances the data volume, variety and complexity increase. This increase requires more effective and advanced analysis techniques to infer decision making information.

3.1.3. Scientific Research Data

Scientific research domain is affected by big data phenomenon including particle physics, astronomy, bioinformatics, earth sciences, social simulations, medicine, genomics, biology, biogeochemistry, atmospheric-science, etc. In particle physics, sufficient computing power is needed for analysis of results and storage of data produced by the European Organization for Nuclear Research (CERN) Large Hadron Collider, which has started to generate data since 2009 and produced 13 petabytes of data in 2010 [16]. In astronomy, by taking pictures of thesky the scientists try to virtualize space and mostly work in this virtual space rather than real one [17]. In social science, big data is needed for analyzing, processing and storing of social and behavioral data.

3.1.4. Networking Data

The network is the base infrastructure for transmitting and sharing data and almost in every aspect of human life, people connect through the Internet or other private networks via wireless, wired or mobile connections. Even in distributed computing and storage systems, the network is used within the datacenter. Search, social networking services (SNS), websites, click streams etc. can be considered as network big data sources [3, 18]. Network sourced data are generated at very high speeds. For example, between 2002 and 2009 data traffic grew 56-fold while computing power grew only 16-fold [19]. Adoption of smartphones has been increasing massively for last few years. More than half of world population use a mobile phone and many of them are connected to the Internet whole day. Machines are connected to thenetwork as well as humans. Considering the growth of internet/network-connected devices, network data seems to grow exponentially and require better network technology.

3.2. Big Data Acquisition

Data acquisition consists of collecting and transmitting data towards big data storage. Data acquisition can be examined in three sub-steps: data collection (sensor, log file, web crawler), data transmission (IP backbone, datacenter transmission), and data pre-processing (integration, cleansing, redundancy). Pre-processing can be executed either before transmission or after transmission [18].

3.2.1. Data Collection

It is about retrieving data from data generating objects and environments. Log files are automatically created when some events like clicks, inserts, updates occur. Log generation is a widely used data collection method and they are generated in various formats. For example, website logs can be in public log file format (NCSA), expanded log format (W3C) or IIS format (Microsoft). Sensors are playing more role in social and industrial life. A sensor senses some physical events like vibration, pressure, humidity, velocity, sound, electric current etc., converts them into digital format and pushes through transmission media. This media can be wired fiber or twisted cable as well as wireless waves.

3.2.2. Data Transmission

Once the data is collected, it is transferred to a storage for processing or analyzing. Data can be transferred from data sources to the datacenter or within the datacenter. While data is transferred from sources to the data center, the physical network infrastructure is used. As network traffic grows exponentially, improvement in network technology is inevitable. Most widely used physical infrastructure is fiber optic systems. There is also network infrastructure in a datacenter which has various architectures. Datacenter networking fulfills the communication between hundreds of servers, server clusters, and storage units. There are some data center structures like fat-tree, two layers or three layers. Insufficiency of electronic packet switching restricts bandwidth increase keeping energy consumption low. Achievements in optical technology have made it prominent candidate technology for data centers due to high-throughput, low-delay, and low-energy consumption. With optical networking, even Tbps transfer rate can be achieved. Network virtualization is another option for efficient use of data center networks.

3.2.3. Preprocessing

Data is collected from different heterogeneous resources, in different formats and different quality. So there can be noisy, missing, and inconsistent data. "*Garbage in, garbage out*" stresses the importance of data quality. It is hard to get desired results from the analysis which uses low-quality data as input. Therefore, prior to data analysis, the preprocessing phase gains vital importance for achieving the desired level of information quality. According to ISO, ISO/IEC 25012:2008 data quality is categorized into fifteen attributes: accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability, recoverability. Pre-processing is considered the most time-consuming phase of data mining/analysis cycle. Some pre-processing techniques are integration, cleaning, conforming, redundancy elimination, transformation, and dimension reduction. The purpose of data integration techniques is to unify data and offer users a single view of data, single truth. In the traditional relational world, there is a mature technology which is known as ETL, the abbreviation of Extract, Transform, and Load. Extract selects needed data from sources such as a database, text files, XML, etc. Transform applies several transformations of selected data in order to improve quality and to reach to some target formats. Load delivers extracted and transformed data to destination storage, mostly a DW. Data cleansing deals with inaccurate, incomplete, unreasonable records, handles missing data, determines record usability, and erroneous data [20]. Kimball and Caserta (2004), most vigorous promoter of dimensional DW model, suggest cleansing and conforming are the most important stage of ETL process, because, most value is added into data at this stage.

3.3. Data Storage

Storage is identified as one of the three major IT infrastructure systems: computing, networking, and storage [21]. The growth of data is very high due to the easiness of data generation and ubiquity of data generating sources.

3.3.1. Data Collection

In RDBMS world disk units are used as storage resources for servers rather than local disks. But the storage systems designed for RDBMS are insufficient to store big data. New techniques and architectures are needed for big data storage. Big data storage is about storing and managing big data in persistent storages. A big data storage should handle very large amounts of data and keep scaling to keep up with growth and provide the input/output operations per second (IOPS) necessary for analytics tools [22].

With the CAP theorem, Brewer [23] suggests that any networked shared-data or distributed system can meet only two of following three requirements at the same time: Consistency (C) makes sure there is a single up-to-date copy of the data, but in distributed systems data is partitioned among servers in multiple pieces. C is for assuring multiple copies of data are identical. Anyone who reaches database will always see the latest state of data [24]. Availability (A) makes sure that users can reach, read and change data. Partition tolerance (P) is durability for network failures in distributed systems.

Not only Structured Query Language (NoSQL) supporters use CAP theorem as an argument against RDBMS. Having distributed nature, NoSQL database developers can't ignore partition tolerance, so P is the first one to choose. There remains C and A. While NoSQL developers generally prefer A to C; RDBMS developers use ACID (Atomicity, Consistency, Isolation, and Durability) which focus on C.

But in a recent article, Brewer [23]suggests that CAP theorem is misunderstood in some ways so that there is no need to totally sacrifice C or A. By explicitly handling partitions, it is possible to optimize consistency and availability, thereby achieving some trade-off of all three.

Classification of NoSQL stores according to CAP theorem concerns is as follows:

1. CA partly ignores partition tolerance and generally uses replication to ensure consistency and availability. RDBMS, Vertica, Aster Data and Greenplum are some example of CA concerned databases.

2. CP has weak support for availability and concerns about consistency and partition tolerance. CP concerning databases are distributed databases among many nodes. The foremost CP concerning databases are; BigTable (Column-oriented), Hypertable (Column-oriented), HBase (Column-oriented), MongoDB (Document), Terrastore (Document), Redis (Key-value), Scalaris (Key-value), MemcacheDB (Key-value), and Berkeley DB (Key-value).

3. AP concerns about availability and partition tolerance and achieves satisfying consistency. Some AP systems are Voldemort (Key-value), Tokyo Cabinet (Key-value), KAI (Key-value), CouchDB (Document-oriented), SimpleDB (Document-oriented), and Riak (Document-oriented).

3.3.2. Big Data File System

A file system is the methods and data structures that an operating system uses to keep track of files on a disk or partition; in other words, the way the files are organized on the disk. A distributed file system is a file system that allows access to files via a network. Distributed file system enables data sharing among hosts. GFS (Google), HDFS (Hadoop), Cosmos (Microsoft), Haystack (Facebook), TFS and FastDFS (Taobao) are featured distributed file systems.

Google designed and implemented its scalable distributed file system, Google File System (GFS), in order to meet rapidly growing demands of processing needs and data-intensive applications. Although Google is inspired by previous distributed file systems and shares same goals; GFS's design consideration is driven by its own needs. GFS provides fault tolerance besides it runs on inexpensive commodity hardware and delivers high aggregate performance to a large number of clients [25].

Some disadvantages of GFS such as single point of failure and poor performance for small files are overcome by Colossus[26]. After GFS, many open source initiatives and enterprises developed their distributed file systems to fulfill their big data storage requirements such as Apache's Hadoop Distributed File System (HDFS), Microsoft's Cosmos and Facebook's Haystack.

3.3.3. Big Databases

As relational databases fall short of big data requirements, there needed to seek alternative technologies. As an alternative to RDBMS, NoSQL databases are the most prominent candidates to meet big data requirements. NoSQL databases offer asolution for big data requirements which are hard to solve with RDBMS such as schema flexibility, complex queries, data update, and scalability. RDBMS is quite inflexible to take into account unplanned future needs. Especially updating schema for new design demands is not easy because of constraint, integrity, and normalization restrictions. But all of these don't mean RDBMS and NoSQL do the opposite things or RDBMS is totally useless.

NoSQL offers high storage size and performance and high availability at the expense of losing the ACID (Atomic, Consistent, Isolated, Durable) trait of RDBMS in exchange for the weaker BASE (Basic Availability, Soft state, Eventual consistency) feature [27].

Although there is no official classification of NoSQL databases [27] classification can make by following: (1) key-value stores, (2) document databases, (3) column-oriented databases, and (4) graph stores [14, 18, 28-30].

3.3.4. Big Data Processing Models

Users have widely manipulated data in databases via SQL queries for many years. But with the dawn of big data, SQL has fallen short to process data which has big data characteristics. To overcome SQL's inabilities new programming models were researched. Google worked on this problem and created MapReduce programming model for processing large multi-structured datasets[31, 32]. MapReduce model consists of two functions: map and reduce. Map functions load, parse, transform data while reduce functions handle asubset of map function output.

MapReduce is not a programing language, it is a programming model. The objective of the example MapReduce system in Figure 1 is to count letters. The system takes inputs and splits them into multiple pieces according to job and data nodes available. These pieces are mapped to multiple nodes, above split into three. Each map function run on its own nodes thereby computing occurs in parallel thus reduces processing time. Map function groups letters by sort. Thenthe system takes the output of each map function and merges the results. Reduce function takes these results and calculates totals of each letter. MapReduce can process and analyze large volumes of data. Indexing, search, sort, graph and text analysis, and machine learning are examples of MapReduce applications [32].

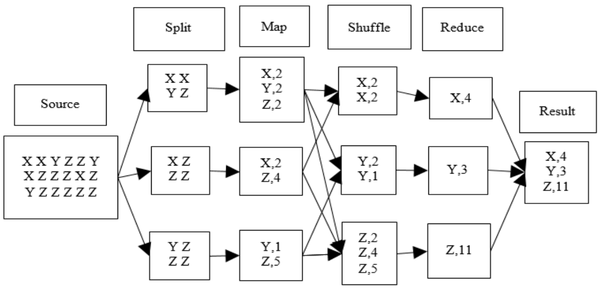

Figure 1. MapReduce

But as big data matures new technologies has come up. Improving MapReduce and curing some of its illness Spark emerged as a strong alternative for MapReduce and relies on three basic concepts Resilient Distributed Datasets (RDD), transformations and actions.

3.4. Big Data Analytics

Storing huge data sets is worthless unless they become meaningful. Big data has impacted on data analysis as well as data storage, processing, and management. As big data expands in all domains including science, engineering, business and healthcare, former data analysis techniques and architectures have become insufficient to analyze such an expansion. Therefore new data analyzing approach and technologies are needed to extract useful information from such large datasets[33-35]. Analysis of big data is named in different ways such as BDA, advanced analytics, analytics, large-data-set analytics, etc. [36]. In this paper,BDA will be used. BDA is the process of examining large data sets by using statistical models, datamining techniques, and computing technologies. This process includes discovering hidden business patterns and secret correlations[37], market trends, customer preferences and other useful business information [38], in data. For example; loyal customers account for the large part of revenue. So, to know whether a customer is loyal or not is very important for an enterprise. It is quite possible to predict that which customer will likely be a loyal one or leave enterprise using machine learning that can use all records no matter how big they are.

Although BDA offers many opportunities for organizations to benefit from, yet analyzing big data is a compelling task. X. Wu et al. [35] have analyzed the challenges at the data, model, and system levels. Data level challenge is developing a safe and sound information sharing protocol. The model level challenge is to design algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain the best model out of the Big Data. The system-level challenge is to design for linking unstructured data through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future.

3.4.1. Big Data Mining

Big data mining can be defined same as classical data mining, the process of a set of techniques and methods to extract valuable information from data or interesting, unexpected, and unknown patterns in data [39], to make better predictions and decisions. Alpaydın [40]also defines data mining as: "*Application of machine learning methods to large databases*". The difference between big data and classical data mining is the data characteristics (volume, variety etc.) on which they run. Data mining tries to give answers to the following kind of questions: What is in the data? What kind of patterns are hidden in the data? And how an organization can benefit from these patterns and relationships in the mess of data? [39].

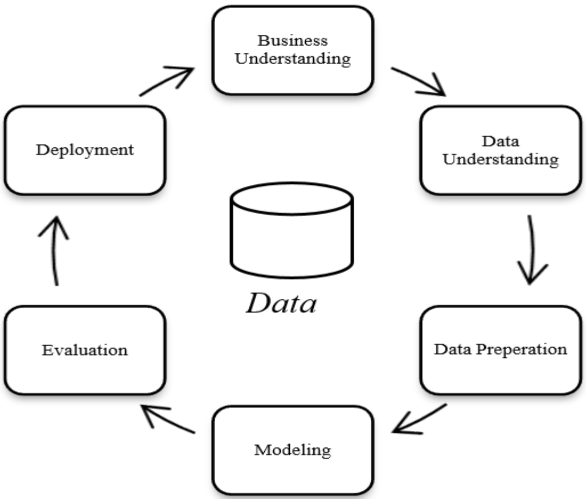

Figure 2. CRISP-DM

As seen from the definition above data mining is a process, defined by Cross Industry Standard Process for Data Mining (CRISP-DM) which is a process to pioneer data mining exertions [41]. CRISP-DM overviews data mining life cycle shown in Figure 2.

Business understanding is simply to understand the problem which can be sometimes simple and sometimes vague. Data understanding phase deals with raw material on which problem solving is based. Data preparation is the phase of getting data appropriate format, structure and quality to yield better results. Because most analytic tools require such a necessity. Modelling is the process of capturing the pattern in the data. Evaluation assesses the result whether they are consistent or reliable. Deployment is the use of data mining techniques in real world cases.

3.4.2. Machine Learning

Machine learning (ML) is situated the intersection of computer science, engineering, and statistics and as well as other disciplines [42]. ML learns from past and present data and uses it further future prediction and decisions. ML is used to solve complex problems that cannot be solved with standard algorithms. Standard algorithms are not able to solve every kind of problems like distinguishing spam e-mail from legitimate one [40]. To some degree, ML imitates human learning abilities and sometimes outperforms human being for handling complex tasks or problems. In order to adapt itself new situations without human interference and improve results, ML uses algorithms and iteratively learns from new data and old results. If an algorithm doesn't solve the problem and data are available, ML can solve. As data is increasing enormously, how to derive value from and discover hidden patterns in this huge data is a big problem. ML aims to automatically solve this problem using huge dataset, hence props up BDA. So, ML positions itself as a core component of BDA. All other components of big data platform serve for ML in some way [43].

ML algorithms can be broadly classified as supervised and unsupervised learning. In supervised learning, machine learns from training data. Supervised methods try to discover the relationship between inputs (independent variable) and target (dependent variable) in the context of a model [44]. If features and target variables of records are available supervised learning can predict accurately target variable given a new record. If there is no target but features then problem fits into unsupervised learning. When an enterprise wants to group his customers but doesn't sure how many groups should come out, it is better to use unsupervised learning. If this enterprise wants to know which customers will likely to purchase particular products and put these customers in a group, then supervised learning is in use.

Supervised learning uses different techniques than unsupervised one and produces better results [45]. Even so, both techniques are useful for solving different kinds of problems. Supervised learning has two main sub-branch: classification and regression. Classification is the task to predict a class label for a given unlabeled point [40]. A Bank can group their customers according to credit risk as high, medium and low risk, by using classification. Regression tries to predict dependent variable or target value using available input (independent variables) data. For a simple example; using patient's age and body mass index, cholesterol level can be predicted. Another example: Using some car attributes such as engine capacity, age, and fuel consumption we can evaluate the value of the car with regression method. The main difference between regression and classification is; while classification methods use categorical target, regression uses numeric one [46] and regression also requires numeric independent variables.

Unsupervised learning doesn't use training data for finding patterns from given data. Prominent methods for unsupervised learnings are clustering and associative rule mining. Clustering finds a natural grouping of instances given un-labeled data. Simply, it groups similar items. The cluster is a subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it. K-means clustering and apriori algorithm for association analysis, namely shopping basket analysis are widely used clustering algorithms.

4. Business Intelligence

For last decades there has been tremendous increase and easiness in regards to creating, capturing, collecting and transferring the data. By this technological development demand and willingness, extracting business value hidden within this huge data piles has also increased. Not much more than a few decades ago, reaching information in itself was a matter. But nowadays although information is abundant and ubiquitous, having the right and relevant information is a matter as much as reaching it beforehand. At this point, organizations need more sophisticated tools for their managers and analysts to retrieve valuable information from huge data sets. To meet this inevitable demand, BI Systems have come into action. BI (including any kind of software, hardware, and other tools) is a *system that collects and integrates raw data from different sources, makes it suitable to analyze, transforms it into valuable information, disseminates it timely and in expected format to those in need, regulates information sharing and flowing, makes sure security by imposing authentication, authorization, and access rights*. In short, BI can be described as the process of collection and transformation of data into valuable information and serving this information those who need.

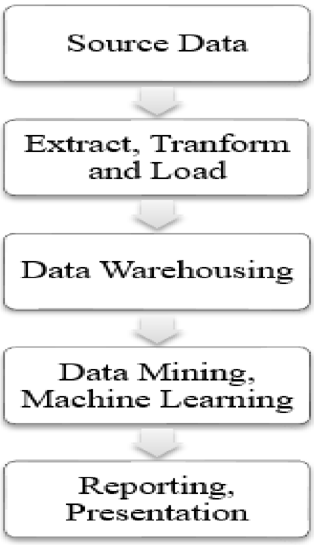

Figure 3. Data Flow in a Simple BI Architecture

BI is being used in wide range of sectors such as Business, Security, Finance, Marketing, Law, Education, Visualization, Science, Engineering, Medicine, Bioinformatics, Health Informatics, Humanities, Retailing, and Telecommunications. It has proved that it is an essential system for organizations to gain a competitive advantage in the global market [47] and is regarded as a key approach to increasing enterprise's value. Redwood Capital forecasts that global BI market is expected to reach $20.81 billion by 2018. This growth equals 8.28% of annual growth rate date back 2013 [48].

BI is just not used as a tool for better decision making for administrative purposes but also used by industries in order to extract useful patterns by outlier detection, process mining and clustering like improving combustion efficiency [49].

As BI main function is transforming data into information and further into knowledge, it is better to mention about these terms. The relationship between data, information, and knowledge is shown in Figure 4.

Data; is raw and by itself meaningless. It just exists in different formats.

Information; gives data a context, so data becomes meaningful. For example, 1300 is meaningless data but "engine volume: 1300 cc" means more. We can find answers "who", "what", "where", and "when" questions from information [50]. For example; a relational database provides information by establishing relations between data items [51].

Knowledge; is the concept of understanding information based on recognized patterns in a way that provides insight to information and can be gained by learning.

Wisdom; the quality of having experience, knowledge, and good judgment; the quality of being wise.

Carter [52] proposes that decision making would no longer be based on guess-work, hunches or random facts. Instead, it would be based on good, clean, valuable information which is ready to consume that is what BI tries to fulfill. Chaudhuri et al. [13]reviewed current BI technologies in their paper. They mentioned specific technologies related to BI including ETL tools, complex event processing engines, and relational database management systems, MapReduce paradigms, OLAP servers, reporting servers, enterprise search engines, data mining, and text analytic engines. They also depicted a typical BI architecture and the way data flows from data sources such as operational databases, text documents, web pages towards data movement and streaming engines, DWs, mid-tier servers like OLAP server or report server and front-end applications. Although BI approves its value to the business, it is rare to come across successful BI implementation and use. Olszak [47]asserts that although BI is anessential technology to be procured, implementation of many BI projects fails. She concludes her work by stating that managers are not fully aware of BI capabilities and are deprived of using its whole potential. She also concludes BI is treated as a technology or tool to collect and analyze data but not as a supportive instrument for effective decision making or improving business processes and performances.

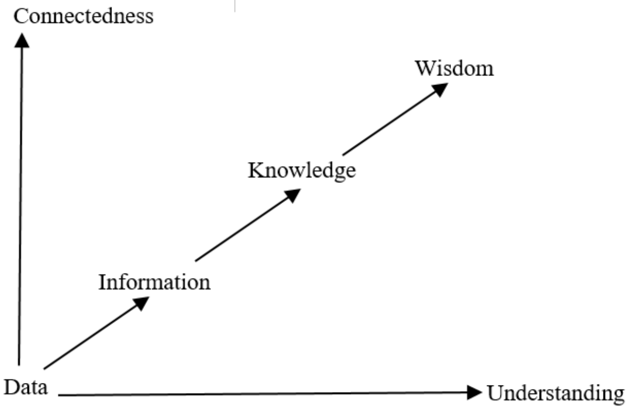

Figure 4. Relationship between data, information and knowledge (Cooper, 2010)

4.1. DW, Online Transactional Processing (OLTP), and Online Analytical Processing (OLAP)

It is widely accepted that DW is a core component of BI solutions [53]. A DW enables saving historical data and trend analysis as well as data mining on clean and integrated data. A DW is subject oriented, integrated, time-variant, and non-volatile collection of data that supports decision making [54]. It connects different databases one to another in order to produce more inclusive information for management information needs [55]. Online Transaction Processing (OLTP) systems are used for operational transactions. In the relational database world, OLTP databases are designed to optimize insert, update and delete queries whereas DW is designed to optimize select queries. So, OLTP is not suited for decision making due to poor select query performances. Besides using directly OLTP sources for business decision making will surely cause slow down business operations.

A traditional DW is mostly fed by organization's OLTP databases. The data is extracted, cleaned, confirmed, transformed and loaded into a DW. In other words, data quality is increased before loading to DW. Since its data quality is improved through ETL procedures, a DW is expected to have better data quality than OLTP database, and hence enables better data mining results. However, it is not easy to establish a DW which is consistent with OLTP database and to store high-quality data. There are some inconsistencies between DW and OLTP databases due to poor ETL processes and DW establishment and execution. Moreover, relational DWs have some limitations and disadvantages; they are not the optimal environment for non-structured data; it is hard to gain real-time insight due to change capture and ETL processes; they have additional costs, and organization's information culture and technology using skills may not benefit from DW potential.

In terms of DW architectures, among many architectures, two come to the forefront which is dimensional modeling advocated by Ralph Kimball and enterprise DW by Bill Inmon. Essential differences between two are development methodologies, data modeling, and DW architecture.

Table 1 Summary of Big Data and Business Intelligence Papers

Area	Papers
Big Data Introduction	[2], [5], [6], [7], [8], [9], [11], [12], [13], [14], [15]
Big Data Value Chain – Data Generation	[5], [11], [14], [16], [17], [18]
Big Data Value Chain – Data Acquisition	[19], [20]
Big Data Value Chain – Data Storage	[14], [18], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]
Big Data Value Chain – Big Data Analytics	[33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]
Business Intelligence	[13], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65]

Reviewing data from DW and gaining insight quickly is a quite cumbersome business. Especially when DW grows beyond a certain volume. OLAP enables examining quickly of huge relational datasets. OLAP uses data from DW, precomputes it and stores on its own format in order to provide faster access than DW. In another saying, data is aggregated, historical data, stored in multi-dimensional schemas thereby processing becomes faster.

While the size of data which is collected for BI in DW is rapidly growing, traditional DW solutions have reached its limits. Traditional DWs mainly depend on structured data and address smaller volumes, expects predictable and arranged data loads, use a single server, and need separate data repository [56]. To overcome this kind of limitations new approaches have come into sight. One of them is Hive. Hive is an open source Apache Hadoop subproject for scalable large DW and uses SQL like query language, HiveQL. Therefore transfers RDBMS users' SQL talents into MapReduce world.

5. Discussion

Both big data and BI area buzz words of nowadays. Two of them have some in common. BI has positioned itself at the core of data-driven decision support system. It also has tools like reports, dashboards, and tables to emit information those who need them, hence improve decision making, increase revenue and accomplish many other organizational goals. Big data offers new opportunities and challenges for not only a bunch of sciences and disciplines but also BI. BI had been using traditional technologies before the dawn of big data. Once the big data came out, BI began to exploit its offerings, like many others. With the help of big data, BI forms more attractive and largely relevant research topic [57] like BI and analytics, BDA, advanced analytics. Big data enhances BI capabilities by offering new technologies and tools. BI and big data break new ground as "business intelligence and big data analytics" [58-60].

To examine big data effects on BI, typical BI architecture is used which is a cyclical set of activities from raw data to valuable and consumer-ready information [13]. Stages of BI architecture are as follows: Data sources, data movement (ETL), data warehousing, mid-tier servers, and front-end applications.

- Data sources:

Although traditional BI is able to use unstructured and semi-structured data as a source, it has some limitations such as volume and speed. Before big data, data generating sources were not as much as now so, most of the data sources were composed of Online Transaction Processing (OLTP) which is based on RDBMS, like Customer Relationship Management - CRM, Enterprise Resource Planning System – ERP and any kind of enterprise information systems. But with big data, the composition of data sources and architecture of databases have changed due to unstructured and semi-structured data increase along with volume, velocity and variety change. NoSQL and NewSQL databases are gradually replacing or completing relational databases. Among the BI data sources, big databases and unstructured data sources are gaining weight. Thus, modern BI solutions need big data acquisition and storage capabilities which are two important big data value chain phases.

- Data Movement (ETL):

This stage integrates, cleanses and standardizes the data by ETL tools. By the big data, this stage becomes more challenging due to data source diversity and data volume, velocity, and complexity. ETL is not an easystage in classical technology, now in big data realm, it becomes more complex. To overcome such a difficulty requires novel approaches and paradigms [61]. New ETL tools are expected to process high volume, velocity, and variety data as fast as traditional ones. On the other hand, traditional RDBMS depended BI tools including ETL and DW solutions are in use. So new approaches for ETL should consider both RDBMS and big data challenges. Some of such novel big data ETL approaches are ETLMR [62, 63], and P-ETL [61]. ETL tools also need to mature to meet today's business expectations and have more user-friendly ETL tools which can handle big data as well as traditional one.

- Data Warehousing:

The main goal of DW is to establish single truth for the organization. In traditional BI, the raw data is loaded into new databases called data warehouse after ETL process. The change and new records in data sources are captured and reflected DW for an update. Mid-tier servers like OLAP uses DW as source databases because it offers cleaner and ready to consume high-quality data than raw data sources. In this stage, traditional relational databases used to play an important role but now they are insufficient due to scalability and efficiency. So, at this stage,big data tools are getting prevalent like Hive or hybrid data warehouses. Hive is popular SQL like MapReduce driven big data DW system. Users use SQL abilities and Hive converts these queries into MapReduce job for data manipulation. However, Hive lacks some of the traditional DW capabilities like slowly changing dimensions. A dimensional RDBMS DW offers slowly changing dimensions and updates, but Hive doesn't. Nevertheless, new frameworks like CloudETL [64] and ETLMR enables slowly changing dimensions ETL for dimensional DW.

Datalake is new DW like the term and has popped up by big data advent. Apart from DW or data marts, the data in a datalake is in its natural, uncultivated state mostly in big data repository like Hadoop. The data in datalake is accessed whenever needed.

- Mid-tier Servers:

This stage is consummate of data warehousing and adds more value to data on the route of the data-information-value journey. Mid-tier servers include OLAP server, search engines, data mining and analytics servers, and reporting servers. The most famous tool of this stage is OLAP, explained above. OLAP is an amazing tool for multidimensional data exploration. But big data has not yet matured as much as RDBMS in terms of integration of multidimensional data [65]. In this tier, more works are needed for scalable OLAP capabilities on large datasets.

Traditional BI mid-tier stage is focused on mainly descriptive information, answering "what was happened?" By big data predictive information answering "what will happen?" has gained more importance. Furthermore, prescriptive information answering "what will happen and what possible responses could be and what kind of precautions could be taken?" is what businessmen demand most because this one contributes more decision-making.

In terms of information currency, big data descriptive analytics refresh time have reduced in few seconds while traditional BI descriptive analytics is daily, at best a few hours.

- Front-end Applications:

At this stage business users consume information and then they make decisions, take actions or modify their processes. Consumption takes place via some applications like dashboards, spreadsheets, key performance indicators, ad-hoc query interfaces, inter-active data-mining tools, data exploration tools etc.

Visualization is a key factor for not only classical BI but also big data impacted BI. Visualization allows users to understand information quickly which derived large and complex datasets without making a great effort. Beyond 3V define of big data further Vs are asserted and one of them visibility stressing visualization of data.

- Workforce Skills:

Both BI and big data requires business understanding as well as technical skills. While traditional BI skills heap up theuse of commercial software like SAP, SAS and Microsoft and BI/DW theories; big data skills are more technical and code based like R, Python, and Scala. On that note; new BI workforce will require following skills; proficiency in BI software use, business understanding and field information, coding, technical, statistical data mining and machine learning skills.

6. Conclusions

As a core BI component, relational DWs have been widely used so far. But big data DW showed up a few years ago and proliferation of big data has triggered to change BI/DW solutions. This emergence doesn't mean usage of RDBMS will totally disappear and big data technology will replace it. Rather, new big data technology seems to complement existing BI/DW systems [66]. Big data has enhanced BI capabilities by enabling unusual new data sources, technologies and user skills [67]. In other words, by using big data, BI get many opportunities to function better and fulfill what is expected from itself. Big data seems to complete BI systems. Especially merging big data with advanced analytics is getting most profound trends in BI [36].

Big data can help BI within following aspects:

1. Enhance data sources and mining capabilities by acquiring and processing unstructured and semi-structured data,
2. Analysis of huge data sets that one machine and main memory can't handle, and new machine learning and data mining capabilities
3. New ways of data storage and management,
4. Near real-time or real-time analytics, (OLAP needs to reprocess after new data loads)
5. Stream processing.

Totally distinguish big data analytics and BI is an intricate job. It is hard to decide where first begins and latter ends. To our opinion analytics seems an intersection set of BI and big data. BI's outlook is broader than big data. A perceivable difference between big data and BI in a business context is that while big data focuses on handling with big data, BI deals with information flow, sharing, and needs within organization along with data processing. At the dawn of big data era, BI has to use big data technologies and analytics. BI will continue to use classical data handling tools for a while along with big data tools. But the proportion of big data will outperform classical tools. Organizations which arenot subject to big data will continue to use classical BI tools. But the organizations subject to big data are expected to use big data and advanced analytics in their BI solutions. We expect that the term BI and big data will merge into one term as "BI&BDA" in a near future.

7. References

1. Gantz, J. and D. Reinsel, *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east.* IDC iView: IDC Analyze the Future, 2012. **2007**: p. 1-16.
2. Gantz, J. and D. Reinsel, *Extracting value from chaos.* IDC iview, 2011(1142): p. 9-10.
3. Mayer-Schönberger, V. and K. Cukier, *Big data: A revolution that will transform how we live, work, and think.* 2013: Houghton Mifflin Harcourt.
4. Kambatla, K., et al., *Trends in big data analytics.* Journal of Parallel and Distributed Computing, 2014. **74**(7): p. 2561-2573.
5. Shekhar, H. and M. Sharma, *A Framework for Big Data Analytics as a Scalable Systems.* 2015.
6. Beyer, M.A. and D. Laney, *The importance of 'big data': a definition.* Stamford, CT: Gartner, 2012.
7. Krishnan, K., *Data warehousing in the age of big data*. 2013: Newnes.
8. Chen, M., S. Mao, and Y. Liu, *Big data: A survey.* Mobile Networks and Applications, 2014. **19**(2): p. 171-209.

9. Assunção, M.D., et al., *Big Data computing and clouds: Trends and future directions.* Journal of Parallel and Distributed Computing, 2015. **79**: p. 3-15.

10. Kalapesi, C. *Unlocking the value of personal data: From collection to usage*. in *World Economic Forum technical report*. 2013.

11. Manyika, J., et al., *Big data: The next frontier for innovation, competition, and productivity.* 2011.

12. Bertino, E., et al., *Challenges and Opportunities with Big Data.* 2011.

13. Chaudhuri, S., U. Dayal, and V. Narasayya, *An overview of business intelligence technology.* Communications of the ACM, 2011. **54**(8): p. 88-98.

14. Chen, C.P. and C.-Y. Zhang, *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.* Information Sciences, 2014. **275**: p. 314-347.

15. Labrinidis, A. and H. Jagadish, *Challenges and opportunities with big data.* Proceedings of the VLDB Endowment, 2012. **5**(12): p. 2032-2033.

16. Brumfiel, G., *High-energy physics: Down the petabyte highway.* Nature News, 2011. **469**(7330): p. 282-283.

17. Eisenstein, D.J., et al., *SDSS-III: Massive spectroscopic surveys of the distant universe, the Milky Way, and extra-solar planetary systems.* The Astronomical Journal, 2011. **142**(3): p. 72.

18. Hu, H., et al., *Toward scalable systems for big data analytics: A technology tutorial.* Access, IEEE, 2014. **2**: p. 652-687.

19. Mayer, M. *Innovation at Google: the physics of data*. in *PARC Forum*. 2009.

20. Maletic, J.I. and A. Marcus. *Data Cleansing: Beyond Integrity Analysis*. in *IQ*. 2000. Citeseer.

21. Poulton, N., *Data Storage Networking: Real World Skills for the CompTIA Storage+ Certification and Beyond*. 2014: John Wiley & Sons.

22. Adshead, A., *Big Data storage: Defining Big Data and the type of storage it needs.* Computer Weekly. ComputerWeekly. com. Published April, 2013.

23. Brewer, E., *CAP twelve years later: How the" rules" have changed.* Computer, 2012. **45**(2): p. 23-29.

24. Pokorny, J., *NoSQL databases: a step to database scalability in web environment.* International Journal of Web Information Systems, 2013. **9**(1): p. 69-82.

25. Ghemawat, S., H. Gobioff, and S.-T. Leung. *The Google file system*. in *ACM SIGOPS operating systems review*. 2003. ACM.

26. McKusick, M.K. and S. Quinlan, *GFS: Evolution on Fast-forward.* ACM Queue, 2009. **7**(7): p. 10.

27. Tudorica, B.G. and C. Bucur. *A comparison between several NoSQL databases with comments and notes*. in *Roedunet International Conference (RoEduNet), 2011 10th*. 2011. IEEE.

28. Han, J., et al. *Survey on NoSQL database*. in *Pervasive computing and applications (ICPCA), 2011 6th international conference on*. 2011. IEEE.

29. McCreary, D. and A. Kelly, *Making sense of NoSQL.* Greenwich, Conn.: Manning Publications, 2013.

30. Vaish, G., *Getting started with NoSQL*. 2013: Packt Publishing Ltd.

31. Dean, J. and S. Ghemawat, *MapReduce: simplified data processing on large clusters.* Communications of the ACM, 2008. **51**(1): p. 107-113.

32. White, C., *MapReduce and Data Scientist*. 2012, Teradata & Aster.

33. Begoli, E. and J. Horey. *Design principles for effective knowledge discovery from big data*. in *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on*. 2012. IEEE.

34. Fan, W. and A. Bifet, *Mining big data: current status, and forecast to the future.* ACM sIGKDD Explorations Newsletter, 2013. **14**(2): p. 1-5.

35. Wu, X., et al., *Data mining with big data.* Knowledge and Data Engineering, IEEE Transactions on, 2014. **26**(1): p. 97-107.

36. Russom, P., *Big data analytics.* TDWI Best Practices Report, Fourth Quarter, 2011.

37. Khurana, M. and P. Mehta, *Big Data Analytics And Technologies.* 2015.

38. Rouse, M., *What is big data analytics.* Definition from WhatIs. com.[online] Available at: http://searchbusinessanalytics. techtarget. com/definition/big-data-analytics [Accessed: 30 Mar 2014], 2012.

39. Ahlemeyer-Stubbe, A. and S. Coleman, *A practical guide to data mining for business and industry*. 2014: John Wiley & Sons.

40. Alpaydın, E., *Introduction to Machine Learning*. 2010: Massachusetts Institute of Technology. 579.

41. Shearer, C., *The CRISP-DM model: the new blueprint for data mining.* Journal of data warehousing, 2000. **5**(4): p. 13-22.

42. Harrington, P., *Machine learning in action*. 2012: Manning.

43. Wu, C., R. Buyya, and K. Ramamohanarao, *Big Data Analytics= Machine Learning+ Cloud Computing.* arXiv preprint arXiv:1601.03115, 2016.

44. Rokach, L. and O. Maimon, *Supervised Learning*, in *Data Mining and Knowledge Discovery Handbook*. 2009, Springer. p. 133-147.

45. Provost, F. and T. Fawcett, *Data Science for Business: What you need to know about data mining and data-analytic thinking*. 2013: " O'Reilly Media, Inc.".

46. Myatt, G.J., *Making sense of data: a practical guide to exploratory data analysis and data mining*. 2014: John Wiley & Sons.

47. Olszak, C.M. *Dynamic Business Intelligence and Analytical Capabilities in Organizations*. in *e-Skills for Knowledge Production and Innovation Conference*. 2014. Cape Town, South Africa.

48. Datacomy. *THREE FORECASTS YOU SHOULD KNOW: BIG DATA, BUSINESS INTELLIGENCE & ANALYTICS*. 2014 02 April 2016].

49. Duan, L. and L. Da Xu, *Business intelligence for enterprise systems: a survey.* Industrial Informatics, IEEE Transactions on, 2012. **8**(3): p. 679-687.

50. Cooper, P., *Data, information and knowledge.* Anaesthesia & Intensive Care Medicine, 2010. **11**(12): p. 505-506.

51. Bellinger, G., D. Castro, and A. Mills, *Data, information, knowledge, and wisdom*. 2004.

52. Carter, K.B., *Actionable Intelligence: A Guide to Delivering Business Results with Big Data Fast!* 2014: John Wiley & Sons.

53. Lim, E.-P., H. Chen, and G. Chen, *Business intelligence and analytics: Research directions.* ACM Transactions on Management Information Systems (TMIS), 2013. **3**(4): p. 17.

54. Inmon, W.H., *Building the data warehouse*. 2005: John wiley & sons.

55. Boateng, O., J. Singh, and G. Singh, *Data warehousing.* Bus. Intell. J, 2012. **5**(2).

56. Kune, R., et al., *The anatomy of big data computing.* Software: Practice and Experience, 2016. **46**(1): p. 79-105.

57. Winter, R., O. Marjanovic, and B.H. Wixom, *Introduction to the Business Analytics, Business Intelligence and Data Warehousing Minitrack.* IEEE Computer Society, 2012: p. 3767.

58. El-Gayar, O. and P. Timsina. *Opportunities for Business Intelligence and Big Data Analytics in Evidence Based Medicine*. in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*. 2014. IEEE.

59. He, X.J., *Business Intelligence and Big Data Analytics: An Overview.* Communications of the IIMA, 2016. **14**(3): p. 1.

60. Ong, V.K., *Business Intelligence and Big Data Analytics for Higher Education: Cases from UK Higher Education Institutions.* Information Engineering Express, 2016. **2**(1): p. 65-75.

61. Bala, M., O. Boussaid, and Z. Alimazighi. *P-ETL: Parallel-ETL based on the MapReduce paradigm*. in *Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on*. 2014. IEEE.

62. Liu, X., C. Thomsen, and T.B. Pedersen, *Mapreduce-based dimensional etl made easy.* Proceedings of the VLDB Endowment, 2012. **5**(12): p. 1882-1885.

63. Liu, X., C. Thomsen, and T.B. Pedersen, *ETLMR: a highly scalable dimensional ETL framework based on mapreduce*, in *Transactions on Large-Scale Data-and Knowledge-Centered Systems VIII.* 2013, Springer. p. 1-31.

64. Liu, X., C. Thomsen, and T.B. Pedersen. *CloudETL: scalable dimensional ETL for hive*. in *Proceedings of the 18th International Database Engineering & Applications Symposium*. 2014. ACM.

65. Cuzzocrea, A. *Analytics over big data: Exploring the convergence of datawarehousing, OLAP and data-intensive cloud infrastructures*. in *Computer Software and Applications Conference (COMPSAC), 2013 IEEE 37th Annual*. 2013. IEEE.

66. Russom, P., *Integrating Hadoop into Business Intelligence and Data Warehousing.* TDWI Best Practices Report, 2013.

67. Wixom, B., et al., *The current state of business intelligence in academia: The arrival of big data.* Communications of the Association for Information Systems, 2014. **34**(1): p. 1.