Business Intelligence Model to analyze Social Media through Big Data analytics

Kawtar Mouyassir *, *Mohamed* Hanine, and *Hassan* Ouahmane

National School of Applied Sciences, The Information Technology Laboratory, Chouaib Doukkali University - El Jadida, Morocco

Abstract. Business Intelligence (BI) is a collection of tools, technologies, and practices that include the entire process of collecting, processing, and analyzing qualitative information, to help entrepreneurs better understand their business and marketplace. Every day, social networks expand at a faster rate and pace, which sees them as a source of Big Data. Therefore, BI is developed in the same way on VoC (Voice of Customer) expressed in social media as qualitative data for company decision-makers, who desire to have a clear perception of customers' behaviour. In this article, we present a comparative study between traditional BI and social BI, then examine an approach to social business intelligence. Next, we are going to demonstrate the power of Big Data that can be integrated into BI so that we can finally describe in detail how Big Data technologies, like Apache Flume, help to collect unstructured data from various sources such as social media networks and store it in Hadoop storage.

1 Introduction

In recent years, the processing and analysis of large BI-oriented data have advanced. Traditionally, in very particular contexts, the most widely used methods have incorporated data warehouse (DW), multidimensional (MD) technologies, and online analytical processing (OLAP) [1], allowing the use of static and structured organizational data sets, all of which are entirely materialized and regularly stored for possible analysis in batch mode.

Today, BI decision-making processes are informed by social media patterns, the latter offering direct customer input on goods and services. Social networks are an essential aspect of the information infrastructure. These social media sites have attained an unparalleled degree of penetration for users, customers, and enterprises to provide the professional environment with a valuable information source, reduce costs, and offer quality support to multiple consumers.

In this context, Big Data technologies are one of the most powerful and widely implemented technologies these days that meet the challenges of business intelligence. Big Data will have a significant impact on Business Intelligence, especially for social business intelligence, which utilizes and analyzes enormous metadata generated in real-time by social networks. Besides, understanding and interpreting the semantics of Big Data is today a challenge for companies that aim to understand customers' behaviour [2] better. These analyses will lead and contribute to strategic decision-making.

2 Traditional BI versus social BI

In Business Intelligence or BI is a collection of computer technologies that facilitate data analysis and subsequent decision-making for decision-makers and business leaders.

On the one hand, traditional business intelligence is built on an On-Line Transaction Processing (OLTP) dataset, which is a software that can support transactional applications, provide atomicity, and handle regular orders. First, the raw data must be combined (data collection) from various sources (SQL Server, CSV files, Excel, etc.), consolidated, and cleansed using data integration software before being used in business intelligence applications [3]. This data will be imported into the analysis tools until processed. We will be able to pick the main performance metrics at this point to review and build visual analyzes and dashboards.

Then comes the analysis of the various sections of the company's data, such as the business analyst, data scientist, manager, or other related professions. In this stage, we will be able to study performance, monitor it over time, and produce conclusions. Finally, we will be able to make the required decisions and see the results of our subsequent analyses after reviewing the data using our decision-making method.

On the other hand, Social Business Intelligence is a similar process but going through different phases. Social Business Intelligence leverages data from social networks, allowing organizations to survey their impressions of the content and events. Furthermore, Social Business Intelligence relies on data from social networks, allowing organizations to probe the customers' feelings about the content and events associated with their products in real-time [4].

In Social Business Intelligence, companies aim to understand the sentiments and VoCs to understand better the thoughts and behaviours of consumers [5]. For this reason, social Business Intelligence relies on different social networks to provide competitive strategic information. Social media plays a significant role, which helps businesses overcome the limitations of old ways of collecting data. This generally includes data collection from several databases and extracting it using attempted analytical techniques.

3 Conceptual Framework

In today's modern world of technology, social networks have made an enormous impact. People will openly express and debate their opinions on a topic, making it a valuable source of knowledge. Sentiment analysis, also known as opinion mining, is a technique for identifying people's opinions or reactions to goods, programs, organizations, persons, and incidents.

Figure 1 shows that the system necessarily requires completing a series of critical steps to cover the life cycle of data accumulated from multiple social media sources [6]. This process aims to analyze this data according to several important components, such as data collection, classification of this data, content analysis, data storage, etc.

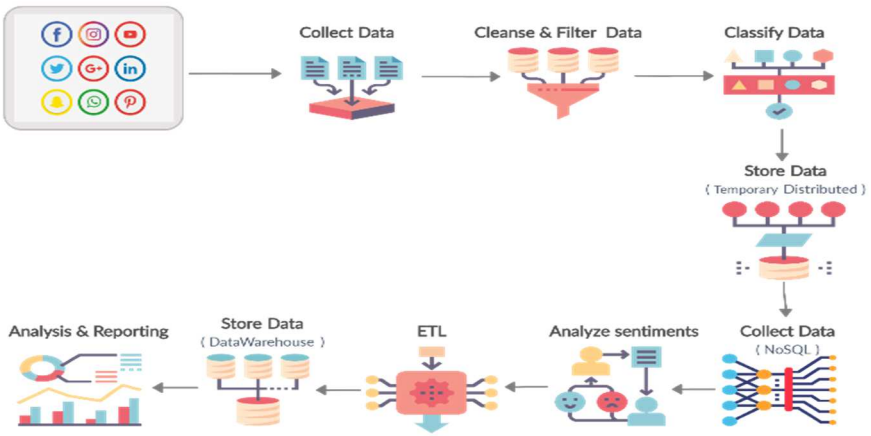

Fig. 1. Conceptual Framework.

3.1 Collect data

The first step in this process is data collection since it is a complex process that involves using tools to collect information from several sources.

Data management enables organizations to make better decisions, solve complex problems, gain a complete view of performance, improve workflow processes, and understand customer behaviour. Organizations can collect data through different social media platforms, like Facebook, Twitter, Instagram, etc.

In this article, we present Apache Flume as data collection tool. In the following, we will describe in detail how Big Data technology enables and facilitates the processing of this data, which is mainly collected from social networks by Apache Flume and stored in Hadoop storage.

3.2 Cleanse & Filter Data

In this second step, we will take care of the data cleaning, including several errors that require filtering and sorting, discarding irrelevant data, meaningless data, eliminating redundant data, etc.

3.3 Classify data

This third stage includes evaluating the quality of the social media data after it has been filtered, and using text classification.

Designating a collection of computerized methods for extracting, and quantifying information from textual documents, is a process known as text classification, categorization, Summarization, clustering [7], etc. This phase would allow the use not only of online data recovery and text cleaning tools but also of computational tools for constructive data use. There is a set of text classification algorithms used like SVM, Decision Tree [8], etc.

Many text mining tools have been developed to analyze the performance of social media platforms. These allow keeping track and interpret online texts from news, blogs, emails, and other outlets. Text mining tools can also help understanding how people are reacting to your brand and content on social media by measuring the number of messages, likes, and supporters.

3.4 Store Data (Temporary Distributed)

At this phase, the data will be processed in a Datawarehouse in a distributed and non-volatile method. Block, document, and object storage are the three types of storage that can relate to the distributed storage. The data would be sorted by documents in our process because it is dependent on the workload performed by the user of the system.

3.5 Collect Data (NoSQL)

In this stage, we will use the NoSQL language to collect all the data that has been deposited in a distributed manner.

We will be able to take advantage of an enormous amount of data by using NoSQL databases [9], which will conform to the massive amount of unstructured data coming from social networks, with fast reading speed, better flow, desirable performance, dynamic schema, and the ability to upgrade the machine's memory and storage, etc.

3.6 Analyze Sentiments

Sentiment analysis is a considerable phase since it focuses on the processing of the analysis of emerging functionalities related to customer sentiment. Consequently, it will be possible to process the customer's opinion, that is collected via its publications on social networks, to identify the customer's opinion on a product or service. This phase is about distinguishing the feelings and the emotions of the customer and give a clear vision of the VoC.

In the sentiments analysis phase, the detection system refers to the detection of unordinary consumer expression. The emotional state of gaining customers through social media reveals the value of organizational awareness and industry for decision-making [10]. The impact of ignoring or disregarding these patterns may be intense.

3.7 ETL

Since user opinions are so important in decision-making processes, many data warehouse solutions use the ETL (Extract, Transform, and Load) phase to incorporate opinions into a cleaning and integration process [11].

The ETL system seeks to make data more relevant so that users' views shared on social networks can be analyzed. The ETL aims to retrieve data first, then convert it using a series of processes such as error deletion, correction, data manipulation, counting of posts, analysis of comments, labelling of sentiments, etc.

Finally, we will go through the ETL's final process, that is responsible for loading the transformed data into a data warehouse. The ETL phase is critical since it allows the data to be structured so that it can be used by the tools in the following steps.

3.8 Store Data (Data Warehouse)

At this stage, the data already loaded earlier, then enter the ETL, is stored in the data warehouse.

3.9 Analysis & Reporting

Finally, reports are created as part of this pre-process to help the end-users understand the results. These end-users would be able to get a better understanding of consumer behaviour, allowing them to interpret the data and making it understandable [12].

Reporting is about transforming data into information, while analysis is the process of transforming information into knowledge.

4 Big Data integrated into Business intelligence

There are various reasons why we choose to integrate big data into BI to proceed toward Smart Data. The time trend focuses on the two keywords, "Business Intelligence" and "Big Data", which may be attributed to the massive amount of data generated on commercial apps, social networks, and so on...

Integration of Big Data technologies, including Hadoop, MapReduce, Flume, Spark, and others, makes data manipulation for BI easier, particularly for social media content that receives metadata. Therefore, it has piqued researchers' interest in social network data to use Big Data methods to benefit the BI tools. This study is based upon the idea that social media output is enormous and requires effective and scalable big data technology. For that reason, we introduce Apache Flume, which will facilitate the processing of massive amounts of data from social media, that must be correctly handled.

Flume is a software from the Apache foundation that represents a modular architecture based on continuous data flows. This design architecture allows the collection and analysis of log files, which can be programmed or triggered by events [13]. Flume is designed as a system that can handle the ingestion of a large amount of event-based data. Flume is fault-tolerant, reliable, scalable, and resilient with recovery mechanisms, even when multiple nodes fail.

Flume may also be used to transmit event data, including but not limited to network traffic data, data provided by social media websites, email messages, etc. The figure 2 represents the architecture of the Apache Flume, which is divided into three phases.

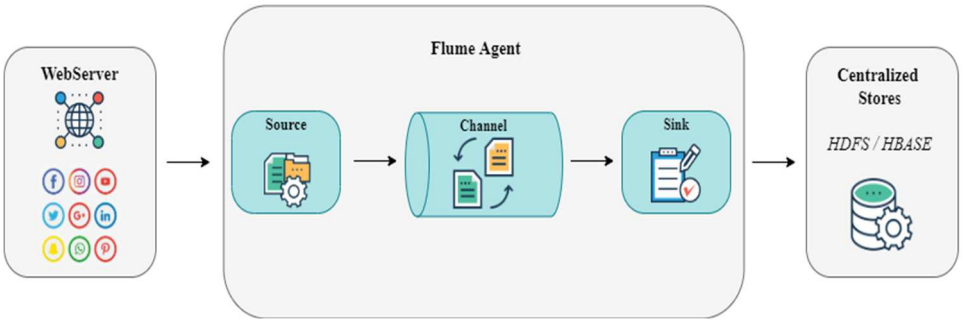

Fig. 2. Apache Flume architecture.

4.1 Source

First, Flume Data Source receives a series of events from external sources and allows the execution of the tail command to retrieve lines from a log file as it happens. The data, in our case, come in the form of events generated in social networks that are external sources.

4.2 Channel

In the second phase, Flume Data Channel receives an event from Flume Data Source and stores it in one or more channels to guarantee that no messages are lost in the event of an agent failure or restart. The channel serves as a storage location for the event before it is consumed sufficiently by the sink. This channel can use a local file system to store these events.

4.3 Sink

Finally, the Flume sink attempts to consume the logs of a channel in batch and write them to a destination, as well as handle a channel's event deletion and store it in an external HDFS / HBASE repository. There can be multiple channel agents, in which case the channel receiver forwards the event to the channel source of the next channel agent in the stream.

5 Conclusion and Future Work

In this article, we discussed the importance of business intelligence regarding the vast data volume of social media to provide real-time analysis of the content shared on social media. Then we presented an analysis of the applicability of social media in BI that helps businesses to get a global and well-defined perception of consumer's sentiments and emotions.

This research will continue by employing a series of technologies, like Big Data technologies, to extract the most significant amount of information from social networks such as Facebook, Instagram, Linked In, etc. Furthermore, using Big Data technologies to incorporate data warehouses and OLTP leads to a faster collection of distributed data, calling for real-time data visualization on dashboards. Our goal is to build a data warehouse system that will allow for automated calculations and data summaries, eliminating the need of counting manually the number of comments, likes, etc.

Also, we are targeting the implementation of a graphical model with a time dimension that will be integrated to transmit the reliability values of the users on the network in real-time. Then we propose deploying a trust inference algorithm that will be stretched over a series of processes based on constructing a metric that enables the convergence of the main attributes formulated.

References

1. P.B. Makeshwar, A. Kalra, N.S. Rajput, KP. Singh, Computational Scalability with Apache Flume and Mahout for Large Scale Round the Clock Analysis of Sensor Network Data, **6**, 3 (2015)
2. A. Hasan, S. Moin, A. Karim, S. Shamshirband, Machine Learning-Based Sentiment Analysis for Twitter Accounts, **15**, 3 (2018)
3. E. Gallinucci, M. Golfarelli, S. Rizzi, Advanced topic modelling for social business intelligence, **20**, 14 (2015)
4. R. Wrembe, *Data Warehouses and Olap: Concepts, Architectures and Solutions*, **332**, 217, December 2006, Paris, France (2006)
5. R. Berlanga, L. García-Moya, V. Nebot, M.J. Aramburu, I. Sanz, D.M. Llidó, SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence, **28**, 3 (2015)

6. I. Paik, T. Golfarelli, H. Tanaka, H. Ohashi, C. Chen, Big data infrastructure for active situation awareness on social network services, **2**, 1 (2013)
7. H. Han, W. Yonggang, C. Tat-Seng, L. Xuelong, Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, **36**, 6 (2014)
8. A. Meier, M. Kaufmann, *SQL & NoSQL Databases*, Models, Languages, Consistency Options and Architectures for Big Data Management, **218**, 169 (2019)
9. L. Schlesinger, F. Irmert, W. Lehner, Supporting the ETL-process by Web Service technologies, **17**, 9 (2005)
10. F. Laghaei, O. Bin Ibrahim, Using Social Network's Data By Extraction Transformation Loading (ETL), **11**, 4 (2017)
11. Y. Lu, F. Wang, R. Maciejewski, Business Intelligence from Social Media: A Study from the VAST Box Office Challenge, **11**, 5 (2018)
12. F. Sebastiani, Machine Learning in Automated Text Categorization, **47**, 34 (2002)
13. W. Sherchan, S. Nepal, C. Paris, A Survey of Trust in Social Networks, **33**, 19 (2013)