Big Data-as-a-Service Solution For Building Graph Social Networks

Siham Yousfi*
Computer Science Department, EMI School
Mohammed Vth University-Agdal
Rabat, Morocco
sihamyousfi@research.emi.ac.ma

Dalila Chiadmi
Computer Science Department, EMI School
Mohammed Vth University-Agdal
Rabat, Morocco
chaidmi@emi.ac.ma

Abstract—**Big Data analytics and Cloud Computing are the new trending that submerged the IT industry. In fact, Big Data technology is providing methods and tools for storing managing and analyzing a large amount of data, and cloud computing provides IT services in a scalable way via internet to a number of clients at low costs. While big data environment requires powerful cluster infrastructure, new ideas about combining this two paradigms were born to enhance business agility and productivity and enable greater efficiencies and reduce costs. Big data as-a-service (BDAAS) refers to common big data services provided as cloud hosted services. These services are intended to provide Big Data features in the cloud. The objective of our research is to describe a BDAAS solution based on Hadoop ecosystem that extracts data from social network and constructs a graph that could be used later for further analysis. As a prototype, we built a graph representing the feeling of a citizen toward a particular deputy. The analysis of the resulting graph will allow citizens and political parties identifying the most popular deputy by analyzing the most significant node.**

***Keywords*-Big Data; Cloud Computing, Big dataas-a-service; Hadoop, Flume, GraphBuilder, Twitter;**

I. Introduction

Recently the amount of data generated by individuals, organizations and academy has been increasing. These data are provided by different sources [1] (1) social networking and media like Facebook, Twitter, and Youtube etc. (2) Mobile devices: calls, texts, locations etc. (3) Internet transactions: purchases, banking activities .etc. (4) sensors: temperature, movement, humidity etc. These "Big Data" are different from "traditional data" according to five characteristics: volume (the data size), velocity (speed of incoming data), Variety (different data types), veracity (the truth of data), and value (the data added-value).

The Added-value of big data is to help companies understanding the customers' needs. For example, in marketing, companies usually pay for an expensive advertising campaign intended to all citizens knowing that a very small percentage would be interested in their products. Using big data tools, these companies may analyze the actions of Internet users across the web and access their personal information published on blogs and social networks to understand customers' needs and predict their preferences.

The environment of big data needs a high-level infrastructure of clusters that may handle the large volume, high velocity and different formats. Besides, data collected are enormous and are generally not ready for analysis. Therefore, there are a set of common steps, such as data extraction, cleaning, modeling that are performed by all companies [2]. Moreover, big data tools are increasing dynamically and users may need a flexible way to manage this environment.

Cloud computing is a new technology that allows clients to use a set of services related to infrastructure, platform or software. Users can create scalable virtual machines quickly and easily while maintenance and configuration tasks will be performed by services' providers. Being aware of its advantages, many IT organizations are looking to cloud computing as the structure to support their big data projects. Thereby, in order to deal with big data issues, a new cloud computing service was proposed: big data as-a-service. BDAAS offers services for managing large scale of data in the cloud in order to enhance efficiency and reduce costs.

The objective of the present paper is to describe and provide a prototype of BDAAS service based on Hadoop that extracts data from big data sources. Raw data will be filtered and processed semantically. The output data could be provided following different formats raw data, structured files, data in databases, graphs etc. we chose to provide a graph data for the following reason:

- Graphs are an intuitive way to represent the relationship between entities.
- Graph analytics provides answers to business problems like shortest path, centrality. etc.
- Graph analytics helps machine learning solutions by adding new parameters about relationships between objects.

In order to illustrate our proposition, we will consider a use case that builds a graph representing deputy reputation from tweets (see section II).

The remainder of this paper is organized as follows. First, the section II describes our use case. Then, we present an overview of big data and cloud computing paradigms in section III. Section IV introduces our proposition and the final section presents the conclusion and our perspectives.

978-1-4673-8149-9/15/$31.00 ©2015 IEEE

II. USE CASE

Social media are platforms of communication and sharing that allow users express their opinions and feelings about everything including services, products, political decisions etc. Analyzing publicly available opinions shared in the social media has become an increasingly popular method for studying socio-political issues that can help government managers to improve their services and activities.

The case of study we want to illustrate is about analyzing the reputation of deputies. The solution that we are describing allows building a graph representing the most popular MP on social media. We are using Twitter a microblogging service that allows users to communicate by publishing small texts called "tweets" [3].

On arrival of each new election, the members are selected based on the choice of citizens. Analyzing the popularity of deputies may be interesting for both parties and citizens.

- For parties: It would be interesting to know in advance their most popular members so as they would present them to the court of elections. Thus, they will have more chances to win the elections.
- For citizens: They may prefer compare their opinions to other citizens' views. And will be more likely to choose an honest deputy to represent them in the parliament.

To collect information from citizens, nothing is better than using social networks that bring together an important mass of data feeling about internet users.

We chose to offer a cloud service called big data as a service that allows data extraction from a set of heterogeneous sources like relational databases and big data. Our research will be in the first step on explicit patterns for example "I like *" "I appreciate *" where "*" is the deputy name already stored in the database. Data will be displayed as a directed graph where MPs and Internet users are nodes and edges model the feeling of appreciation expressed by a user in twitter. The analysis of such a graph will allow us later to identify who is the most popular deputy by analyzing the most significant node (The one with the largest number of incoming edge).

III. BIG DATA

A. Definition of BIG DATA

Today the amount of data that flows around the web is continuously increasing. In fact, the CGOC (Compliance, Governance and Oversight Council) mentioned in 2012 that the data generated by organizations doubles every 18-24 months and 90% of the existing data around the world has been created after 2010[4]. According to [5] a study made in the middle of 2013, the amount of data generated per a day is 2.5 exaoctet and most of these data are generated by social networks. However, with the increasing amount of stored information, the percentage of processed data is rapidly decreasing [6]. It became necessary to introduce new technologies to manage this large mass of data. These technologies are called "Big data".

During the past few years, Big Data, the new "buzz-word", has emerged drawing attentions from both science and industry. According to [7] Big Data represents the intensive data centric technologies. It indicates the continuity of the technology advancement on various individual activities including production, collaboration and consumption. For more information about the how much digital data are generated every minute please refer to [8].

However, the word big data doesn't refer only to a large volume of data. According to [6], big data is applied to the information that cannot be stored, processed or analyzed using the traditional tools. In a 2001, MetaGroup publication, Gartner analyst Doug Laney described Big data using the 3V's characteristics "volume, velocity and variety" [9]. IBM introduced a 4th V veracity through an infographic [10]. Other authors [7] attributed an additional characteristic "Value". The following section will describe these characteristics.

B. Caracteristics of BIG DATA

1) Volume

The first characteristic is volume that refers to data size. It is obvious that the volume of data is the most important challenge that requires specific optimization for traditional technologies. As mentioned in [7] the example of Low Frequency Array a radio telescope that collects 5 PB every hour. Moreover, [5] explained that social networks are closely contributing to this growing data size. Every minute, Facebook generated 350 GB of data, YouTube 72heure of videos and twitter 277000 of tweets and users execute 2million search query using google. These web giants have quickly realized the increase in the amount of data they regularly produce, store and analyze and have developed their own solutions.

2) Velocity

Velocity means how fast data are arriving and stored. In fact, effective management of big data, forces organization to process data while it has a value i.e. in near "real-time" for two reasons [6]: The first one is to gain a competitive advantage. For example, in e-commerce, every click builds the customer behaviors and allows organizations to guess the client's requirements. Thus, companies who can process a quick analysis of these requirements may attract clients by recommending similar products. The second reason why the processing speed must be near to data stream is to filter data before storing it in the Data Warehouse.

3) Variety

Variety represents different types of data. With the explosion of internet, several sources are providing data such as mobile phones, sensor's, health monitors etc. these data are most of time unstructured (images, videos) and semi structured (web pages, logs). As mentioned in [6] structured data from traditional relational databases represents only 20% of the existing data, the remaining 80% are unstructured and semi structured. Therefore, data storage and database design of traditional systems need to consider dynamic adaptation in order to meet the requirement of data variety [7].

4) Veracity

IBM declared Veracity as the 4th big data V that refers to the uncertainty of data wondering "How can you act upon information if you don't trust it?"[10]. In fact, the uncertainty of data may be caused by data inconsistency (statistics are not reliable) or data trustworthiness (subjective data like feelings and opinions). Therefore, big data solution needs to guarantee that data are trusted and secured as long as they persist by ensuring data integrity, data authenticity, availability, identification and evaluating data in its context.

5) Value

Unfortunately, the massive amount of data extracted leads to a situation where there is a lot of data noise. In facts, most of the existing data may be useless and companies need to filter these data in order to benefit from the true value that the extracted data could add to the intended activity.

Big data environment requires a powerful infrastructure of clusters and well configured tools. The implementation of such a solution by a non-specialist company induces a risk of spending a lot of money without guaranteeing a positive result. Moreover, Big Data analysis involves preliminary common steps like data extraction, formatting storage and processing. These operations are performed separately by many companies. Thereby, using a cloud computing platform helps companies to delegate the management of these steps to a specialized organization and ensures efficiency and reduces costs [11].

IV. CLOUD COMPUTING

"Cloud computing" is another buzz word that attracted recently the attention of both media and computer scientists especially in Information Technology and Economy Community. The primary objective of this new paradigm is to be flexible and responsive to customers' needs by providing IT services in a scalable way via internet to a number of clients at low costs. In fact, users will not care about managing various hardware and software installations, configuration and updates. All these operations will be performed by the providers of services [12]. Three categories of services can be offered by cloud computing:

A. Infrastruture as a service

The infrastructure as a service allows users rent customized virtual machines on which they can install a specific operating system and software. Customers pay only in terms of resources consumption e.g.: the volume disk used, CPU cycles per hour .etc. Amazon Elastic Compute Cloud [13] is an example of IAAS.

B. *Software as a service SAAS*

This layer allows clients to use directly applications from the cloud infrastructure. The end user can access to these software using a Web browser without caring about installing them and managing updates. Eg: Gmail.

C. Plateforme as a service PAAS

This layer aims to provide computing platform and the operating system so as other infrastructure tools like Databases are managed by supplier. However, the customers control the application layer they can develop new software using for example the framework already installed in the cloud platform.

1) Data as a service

Data-as-a-Service (DAAS) is a service that allows controlled access to a qualified data store (cleaned and enriched) in the cloud using a Web API provided by supplier. This service allows executing only basic queries. Google's public data service is an example of DAAS that provides public data and forecasts from a range of international organizations.

2) Storage as a service

More companies are manipulating large volume of data that needs huge storage capacity. Using the STAAS service, cloud suppliers allow users to access to a virtual memory space where they can store and share their data in the same infrastructure this improves analysis task [2]. Dropbox is an example of STAAS.

3) Database as a service

This service is a typical database run in the clouds that releases users from database administration tasks like creation, configuration and maintenance and offers database services like querying. Different types of databases are supported: relational databases, No SQL data stores, etc.

Many papers [14], [15] have discussed the opportunities of cloud computing such as helping companies saving money by handling variable business needs. In this case, they pay only for the resources they used. Cloud computing ensures also scalability and elasticity via dynamic allocation of resources and finally for the environment, it helps reducing electricity costs by sharing.

Nowadays, with the increase of big data, cloud computing paradigm is facing many challenges. In fact, according to [14], cloud infrastructure (storage capacity, bandwidth, and network) is not designed to support big data. Moreover, from one hand Big Data computational paradigms and tools such as Hadoop and MapReduce aren't optimal in the cloud and cause an increase of cost [16] and from the other hand the big data environment is dynamic and continues to improve.

Therefore, cloud computing needs to gradually converge to support big data by creating adequate services: Big Data-as-aservice.

D. Big Data-as-a-service

Big Data-as-a-service (BDAAS) refers to common big data services provided as cloud hosted service. These services are intended to create, store, maintain and analyze large scale of data in order to overcome cloud computing issues related to big data management and enhance efficiency and reduce costs.

According to [17] BDAAS encapsulates big data techniques into three layers Big data infrastructure-as-aservice, Big data platform-as-a-service BDPAAS and big data software-as-a-service BDSAAS. The paper [2] describes these architectures and challenges.

However, the BDAAS is very recent and adapted tools don't exist yet. For now, users can only rent the cloud infrastructure and manually install the big data tools needed. For complex distributed services, this can be a big issue.

V. SOLUTION DESCRIPTION

This section will introduce the architecture and the business workflow process of our solution in the context of our use case described in section II that builds a graph reputation for deputies from tweets.

A. Architecture of the solution

The architecture of our solution is mentioned in "Fig. 1"and based on Hadoop ecosystem. We are using java for easy integration. The components are described as follow:

- **Hadoop**: a Framework that allows for the distributed the processing of large data sets it provides cluster resource abstraction and fault tolerance.it is based on Mapreduce a parallel programming model for large scale of data. HDFS is the file system of Hadoop.
- **MySQL database**: Contains structured data that we can use to fulfill the graph properties. In our case MySQL database contains deputies' information (name, party…).
- **Apache Flume** is a Framework for data ingestion that allows collecting large amount of data into a series of events (e.g tweets) from a set of sources (e.g twitter Streaming API). The events are sent using channel to sinks that stores the event as a file in predefined location (e.g HDFS).
- **GraphBuilder** is a scalable graph construction library developed on java environment for Hadoop, which allows data scientists to construct graphs from semistructured or unstructured data sources. GraphBuilder provides for parallel graph construction, transformation, and normalization and partitioning.
- **ETL Framework**: based on properties retrieved from user interface this framework combines both data got from social media and RDBMS.
- **Web interface**: we built a web application that allows users to select the properties of the graph nodes and edges on the basis of the logical database schema and the structure of the retrieved file from Apache Flume.

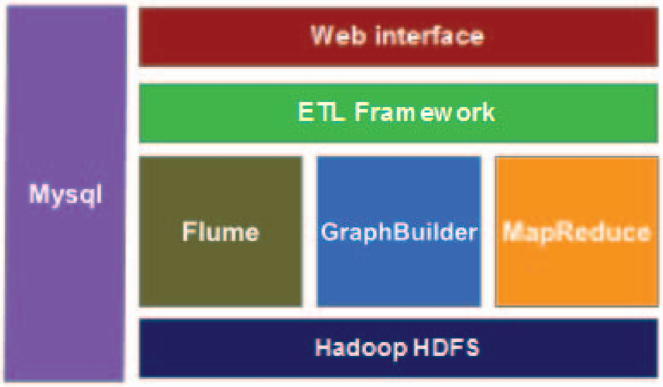

Figure1. Architecture of the system

B. Data flow

As illustrated in "Fig. 2" the workflow of our solution contains the following steps:

- **Generate Flume configuration file** (1): in our context we have two sources MySQL database that contains deputies' information and tweets that contain citizens' feeling. We need to use Flume framework to collect tweets talking about an existing member in our database. Therefore, the first step is to configure Flume in order to connect it to twitter service. "Config generator" is an application that generates the flume configuration file automatically where keywords are filled from the MySQL database.
- **Data gathering & storage** (2): In order to allow developers accessing twitter's data "Twitter" company provides the "Twitter Streaming API" that outputs a constant stream of tweets in a JSON format. Flume will collect every tweets talking about the keywords mentioned in the configuration file (that contains deputies' names) and stores the results automatically in HDFS.
- **Data formatting** (3): This step aims to format the retrieved data from tweeter so as to keep only plain text. Then we will proceed to a semantic analysis based on simple patterns such as "I like *" and "I appreciate *". The result of this step is a text file named "popularity.txt" containing user name and the deputy that he likes in the form of {user, deputy}.
- **Filling graph structure** (4): we built a web application that allows users to select the properties of the graph nodes and edges on the basis of the logical database schema and the structure of the "popularity.txt" file. This application returns two tables of objects called "NodeProperties" and "EdgeProperties".
- **Graph Data extraction** (5): based on "NodeProperties" and "EdgeProperties" we created a solution that selects data relating to nodes and edges from MySQL database, and the file popularity.txt.
- **Graph formation** (6): the program uses the GraphBuilder library to form a graph and perform other operations such as transformation, normalization and partitioning offered by GraphBuilder.

VI. RELATED ARTICLES:

Many authors especially blogger have tried to extract data from social media for example:[18] Shows the technical method of tweets gathering and analysis using Flume and Hive, our study proposes a solution that allows extending unstructured data got from social networks by structured data stored in RDBMS.

[20] Proposed a method for collecting data from a Facebook page like comments posts and status. However, the paper didn't talk about real time data extraction. Since social networks are sources of big data that are updated very quickly, they require a real-time collection and analysis. This real-time aspect has not been addressed by the authors.

Finally [21] introduced a solution based on Hadoop ecosystem and Cogito Intelligence API that analyses semantically the unstructured data in web pages. The result is presented as a graph. However authors are not talking about how to transform or partition the resulting graph.

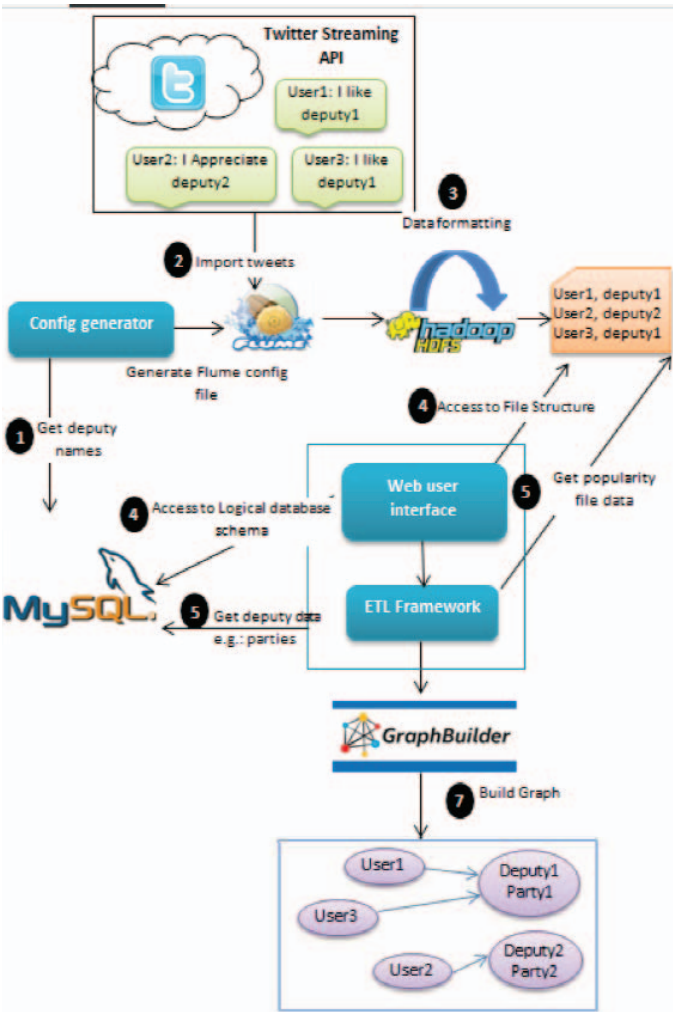

Figure 2. Data flow

VII. CONCLUSION AND PERSPECTIVES

This paper introduced a solution built on Hadoop ecosystem that collects data from social media and database tables and constructs graph using GraphBuilder Library. The use case that we presented is about creating a graph for deputy reputation based on tweets. In the next step we will try to improve the semantic analysis after checking the existing methods and finally we will also add a graph analysis step using GraphLab tool (e.g. determine the region where the deputy is popular).

REFERENCES

[1] J.Kelly, Big Data: Hadoop, Business Analytics and Beyond, [Online] 2014, http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_ Beyond, (Accessed 2015).

[2] Z. Zheng, J. Zhu, M. R. Lyu, Service-generated Big Data and Big Dataas-a-Service: An Overview, IEEE International Congress on Big Data, 2013

[3] A. Java, X. Song, T. Finin, and B. Tseng, “Why we twitter: understanding microblogging usage and communities,” in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 2007, pp. 56–65.

[4] D. Austin, “eDiscovery Trends: CGOCs Information Lifecycle Governance Leader Reference Guide,” ”http://www.ediscoverydaily.com”, May 2012.

[5] B.Wood, “Info-Graphic: How Big is Big Data?”, ,[Online] 2013, http://www.americanis.net/2013/info-graphic-how-big-is-big-data/ (accessed 2015)

[6] P. Zikopoulos, C. Eaton, and others, Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[7] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, “Addressing big data issues in scientific data infrastructure,” in Collaboration Technologies and Systems (CTS), 2013 International Conference on, 2013, pp. 48–55.

[8] J.James, “Data Never Sleeps 2.0”, (domo), [Online] 2014, http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/ (Accessed: 2015)

[9] D. Laney, “3D data management: Controlling data volume, velocity and variety,” META Group Research Note, vol. 6, 2001.

[10] 4-Vs-of-big-data,[Online], http://www- 01.ibm.com/software/data/bigdata/images/4-Vs-of-big-data.jpg (accessed 2015)

[11] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, “Big Data computing and clouds: Trends and future directions,” Journal of Parallel and Distributed Computing, 2014.

[12] L. Wang, G. Von Laszewski, A. Younge, X. He, M. Kunze, J. Tao, and C. Fu, “Cloud computing: a perspective study,” New Generation Computing, vol. 28, no. 2, pp. 137–146, 2010.

[13] D. Robinson, Amazon Web Services Made Simple: Learn how Amazon EC2, S3, SimpleDB and SQS Web Services enables you to reach business goals faster. Emereo Pty Ltd, 2008.

[14] H. Elazhary, “Cloud Computing for Big Data,” MAGNT Research Report, 2014.

[15] R. Moreno-Vozmediano, R. S. Montero, and I. M. Llorente, “Elastic management of cluster-based services in the cloud,” in Proceedings of the 1st workshop on Automated control for datacenters and clouds, 2009, pp. 19–24.

[16] P. M. Kasson, “Computational biology in the cloud: methods and new insights from computing at scale,” in Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2013, p. 451.

[17] “Big data-as-a-service: A market and technology perspective,” EMC Solution Group, Tech. Rep., 2012

[18] J.Natkins, How-to: Analyze Twitter Data with Apache Hadoop, [Online] 2012, http://blog.cloudera.com/blog/2012/09/analyzing-twitter-datawith-hadoop/ (Accessed 2015)

[19] C. Shah and others, “Politics 2.0 with Facebook–Collecting and Analyzing Public Comments on Facebook for Studying Political Discourses,” 2011.

[20] P. L. Puglisi, D. Montanari, A. Petrella, M. Picelli, and D. Rossetti, “From news to facts: An Hadoop-based social graphs analysis,” in High Performance Computing & Simulation (HPCS), 2014 International Conference on, 2014, pp. 315–322..