OLAP of the tweets: From modeling toward exploitation

Maha BEN KRAIEM, Jamel FEKI, Kaïs KHROUF
MIR@CL, University of Sfax
Airport Road Km 4, P.O.Box. 1088, 3018 Sfax, Tunisia
Maha.Benkraiem@yahoo.com, Jamel.Feki@fsegs.rnu.tn
Khrouf.Kais@isecs.rnu.tn.

Franck RAVAT, Olivier TESTE
IRIT, University of Toulouse,
2, Rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 9, France
{Olivier.Teste, Franck.Ravat}@irit.fr

***Abstract*—With the tremendous growth of social networks, there has been a growth in the amount of new data created every minute on these networking sites. Twitter acts as a great source of rich information for millions of users. Twitter messages, or tweets, are limited to 140 data characters. This limitation in length makes difficult their analysis. However, various accessible meta-data are associated with every message. Taking into account these meta-data, they can be very useful for analysis and making decisions. Applying OLAP (On-Line Analytical Processing) and data mining technologies on large volumes of tweets is a challenge that would allow the extraction of information and knowledge such as user behavior, new emerging issues, trends... This paper proposes a generic multidimensional model dedicated to the OLAP of tweets with some results and analyses for testing this multi-dimensional model on various data extracted from tweets.**

Keywords—twitter; tweets; constellation schema; OLAP.

I. INTRODUCTION

In recent decades, the Twitter social network has become more and more popular. Since Twitter is the most used microblogging website with about500 million users and 340 million tweets a day, it is a fascinating source of information and represents a new data provider. The messages, or *Tweets* in Twitter terms, are a very simple and effective way to share interests publicly. Tweets can be embedded, replied to, favorited, unfavorited and deleted[1]. Twitter distinguishes itself from other social media by the limited message size (maximum size of 140 data characters restricts users in their writing).The Twitter user should express information of interest to his/her subscribers, and try to unite others.

Since its appearance, Twitter set off a wave of research initiatives for analysis and knowledge discovery from data contained in a large volume of tweets. We notice that the majority of works provided in the literature of this domain (analysis of tweets) are intended to answer specific tasks or needs. For example, some researches have focused on the detection of real-time events as in [1] and [2]. Other studies have focused their efforts on monitoring trends or on the identification of Buzz (news searing). However, very few studies were interested in the multidimensional analysis of data from tweets so far. If it incorporates all the data issued from a tweet, this modeling could be a judicious opportunity to explore the tweets through an OLAP process [3]. That's why our objective is to develop a solution that promotes the multidimensional storage of tweets and the analysis of their content. We argue that tweets can be represented in a multidimensional way by considering all their data and meta-data. For this reason, we integrate tweets into a data warehouse as a tool for the storage and analysis of multidimensional data. Thus, it becomes possible to manipulate a set of measures according to different dimensions which may be provided with one or more hierarchies [4].Associated operators allow an intuitive navigation on different levels of the hierarchy [5]. OLAP tools provide means both to query and to analyze the warehoused information and then produce reports at different levels of detail. Moreover, data from tweets have particular specificities (e.g. inter-tweets relationships). Hence, the paper issue consists in integrating tweets into a multidimensional schema considering these features.

The remainder of this paper is organized as follows. Section 2 deals with the state of the art of related works. Section 3 describes the structure of a tweet. In Section 4, we introduce our generic multidimensional model, dedicated to the OLAP of the tweets, and then we describe the logical model, as well as its elaboration rules. In section 5, results and analyses for testing this multi-dimensional model on various data extracted from tweets are presented. We end this paper with conclusions and perspectives of improvement.

II. RELATED WORK

In the recent years the important developing of social network activity has led to a massive data volume generation, such as status updates, messaging, blog, and so on; as a result, a novel area for data analysis has emerged.

[1]https://dev.twitter.com/docs/platform-objects/tweets

Twitter, as a new data provider, has largely contributed to the appearance of new issues related to the modeling and manipulation of data. In this context, the analysis of textual content of tweets and their meta-data is a promised research topic that has attracted the attention of many researchers and has given birth to novel analysis areas, such as Social Network Analysis. Hence, the work related to this area can be subdivided into two major categories: Those addressing the storage of data from tweets (multidimensional modeling) while others are simply interested in the analysis of the contents of tweets and data mining.

A spectacular novel area of data analysis is that of the contents of tweets analysis. A pioneering work on this field was carried out by [6]. These authors use Twitter predicted users' personality types based on their Twitter activity and profile. They identified each user's type by their followers and subscription amounts and scored their personality based on how active they appeared to be on Twitter. Personality data was collected from 355 Twitter users and then used to study the relationship between user type and their personality traits. The researches could effectively predict users' personality types from their public Twitter data.

In 2007, [7] presented their observations of the microblogging phenomena by studying the topological and geographical properties of Twitter's social network. They came up with a few categories for Twitter usage, such as daily chatter, conversations, information and URL sharing or reporting news.

Other studies, with similar objectives, focused on the detection of events, sentiments and trends in real time, such as the works of [1], [2], [8] and [9].

In the work of [1], the authors propose to analyze the content of tweets in order to detect in real time alarms during the appearance of earthquakes. They equate every Twitter user with a sensor that is able to detect a target event and to achieve a probabilistic real-time reporting. Finally, for the detection of events and the location estimate, the authors have chosen two probabilistic models: a temporal model according to the date associated with each tweet and a spatial model (KalmanFilter).

Moreover, the authors of TwitterMonitor [2] developed a system for automatically extracting trends in data streams. Their system is based on four modules. A module called TwitterListener that accepts a volume of 1.2 M tweets/day, via a Twitter API. All these tweets are then transferred to a module called Bursty Keyword Detection which identifies words named Bursty. A keyword is identified as bursty when it is encountered at an unusually high rate in the stream. For example, the keyword NBA (National Basketball Association) may usually appear in 5 tweets per minute. However, this keyword may suddenly exhibit a rate of 100 tweets per minute. This sudden rise of the frequency of keywords is usually associated with a sudden popular interest in a particular subject and is often driven by emerging events. Hence, a sudden rise in the frequency of the keyword NBA may be related to a major NBA game in progress. TwitterMonitor treats bursty keywords as "entry points" for trend detection. Words (i.e., Bursty Keyword) detected in a relatively large number of tweets will be subsequently grouped through a module called Bursty Keyword Grouping. This module produces a set of trends (group of keywords) that will eventually be analyzed through the Trend Analysis module according to different factors (Region, Time ...).

A rather similar approach is proposed in the works of [8] where the authors considered RSS ("Really Simple Syndication") as a source for the extraction of information included in tweets in order to detect the various needs of twitter users. Thus, the authors of [10] developed a tool called EVARIST that allows a user, relying on a set of keywords (defined by the user himself), to visualize the most associated terms of Twitter, hence forming the searing news (Buzz) on the chosen topic. This tool is based on a five-step approach: i) Retrieving tweets containing keywords, ii) Cleaning the tweets (removing stop words, punctuations,…), iii) Stating the table of context with the tweets as objects and the words as attributes, iv) Building the corresponding Galois lattice (A Galois Lattice allows to group, in an exhaustive way, objects in classes, called concepts, using their shared properties), and v) Visualization of results.

[9] proposed MOA-TweetReader, a new system to perform twitter stream mining in real time. The input items of this system are the tweets obtained from the Twitter stream. These tweets are preprocessed and converted by a tf-idf filter to vectors of attributes. The second component of the system is a frequent item miner that stores the frequency of the most frequent terms. Finally, a change detector monitors changes in the frequencies of the items.

To our knowledge, these studies have not used the recent data storage technology, that is to say, neither the multidimensional modeling tweets nor the online analytical processing (OLAP: On-Line Analytical Processing) to analyze cubes of tweets. From the other hand, a warehousing system offers several merits; it allows manipulating (aggregate) data, generally quantitative (called measures) according to various dimensions representing axes of analysis [11]. We identified few researches that focused on the use of multidimensional tweets. Among these works, the one of [12] defined a multidimensional star model for analyzing a large number of tweets. However the proposed model was dedicated to a particular trend. In order to do this, the authors proposed an adapted measure, called "TF-IDF adaptive", which identifies the most significant words according to level hierarchies of the cube (the location dimension). Nevertheless, their case study deals with a specific area: the evolution of diseases, referring to the thesaurus MeSH (Medical Subject Headings) by adding to their multidimensional model a dimension called MotMesh (MeshWords).

[13]developed a system for warehousing Streams from Twitter. Their system lies on an architecture consisting of five layers: i) The data source layer is represented by the available Twitter APIs, ii) The ETL [14]layer (Extract, Transform and Load) for the extraction of data from tweets and processing in a suitable format for the target database, iii) The Data warehouse layer for the storage of data issued from tweets, iv) The Analysis layer dedicated for OLAP analyses of the tweets, and v) The Presentation layer of analysis results.

Other studies have simply focused on the automatic extraction of information when available, in order to supply a hierarchy, and then associate a tweet to a specific geographic location in order to facilitate multidimensional analysis. Among these works, we cite the approach proposed by [15] where the authors analyze in a first step the content of tweets in order to retrieve the relevant terms that might correspond to a specified location.

TABLE I. COMPAISON OF WORKS RELATED TO TWEETS

	ETL		**Storage**			**Restitution**			
	Technique used (JAVA, L4G…	***Real time***	***Model***	***Historisation***	***Genericity***	***Consultation of predefined report***	***Interrogation***	***OLAP analysis***	***Datamining***
[1]	Not Mentionned	+	Virtual storage	-	+	-	-	-	+
[2]	Not Mentionned	+	Virtual storage	-	-	+	-	-	-
[6]	Not Mentionned	-	Virtual storage	-	+	-	-	-	+
[8]	JAVA	+	Virtual storage	-	+	+	-	-	+
[9]	Not Mentionned	-	Virtual storage	-	-	+	-	-	+
[10]	Not Mentionned	-	Virtual storage	+	-	-	-	-	+
[12]	Postgresql 8.4 + Pentaho Mondrian 3.20	-	Multidimensional storage (Star schema)	-	-	+	+	+	-
[13]	BaseX + Microsoft SQL Server	-	Multidimensional storage (UML diagram)	-	+	+	-	+	-
[15]	Not Mentionned	-	physical storage	-	+	-	-	-	-

Then, the authors in [15] retrieve location information from the meta-data tweets and try to identify the relevant terms that might correspond to a specified location. This step is performed by using appropriate specific models. Then, the authors retrieve location information from the meta-data tweets and try to identify the geographic location from the location information extracted and, if such information is not available, they use the time zone to estimate the location.

In Table I, we present a summary of the works previously studied; the columns represent our evaluation criteria and the rows are the works studied. The + symbol indicates that the approach supports the corresponding evaluation criterion, whereas, the - symbol points out that the criterion is not supported.

Further to this study, we may conclude that most of these works ensure a special treatment of tweets but do not offer tools for the decision-makers to manipulate the information contained in the combined meta-data associated with their tweets.

In addition, we notice that very few studies have examined the use of cubes for tweets and the exploitation of their multidimensional potential. Hence, we aim at providing a generic multidimensional model supporting tweets i.e., independent of the special needs pre-defined a priori and taking into account the structural specificity and possibly semantic data. In order to do this, we start by studying the structure of a tweet.

III. Structure Of tweets

A tweet is a short message which contains less than 140 characters. On the opposite, the generated code for a tweet is a dozen-line length. In fact, a tweet is a data structure containing several information (User-Data and meta-data) that could be used in decision analyses. This structure is composed of mandatory fields and visible to twitter users, such as the author of the tweet or the tweet's creation date, but also other hidden fields, dedicated to certain features that allow to know whether the tweet is truncated, if used by the SMS services, its place of issue, or the number of followers, the tweet's unique ID, the number of followers... Hence, a tweet is not just a text but it can be assimilated to a complex structure including coded information and a collection of associated meta-data.

All the information of a tweet (including those hidden) can be divided into three parts:

- ***The tweet part*** containing the tweet's unique ID, the text of the tweet (140 characters), the tweet's creation date, the number of times that the tweet was re-tweeted, the application that sends the tweet (Web,…). If it is a tweet response, then it also contains the ID of an existing tweet that this tweet is in reply to, the screen name and the user ID of replied to tweet author.

- ***The User part*** describing the owner of the Twitter account; this is a set of information concerning the user (the author's user ID, the author's user name, the author's user screen name and the author's URL), other information regarding the account (the creation date and description of the account, the location that the account owner associated to their account, the time zone, offset in seconds and the user's selected language) and information concerning the profile (User profile's photo, Background image chosen by user for own twitter page, colors for page's characters and bars.).
- ***The place part*** characterizes the identification of the place associated with the Tweet, the URL to fetch a detailed polygon for this place, the printable names of this place, the type of this place – can be a "City or Neighborhood", the country in which the place is located the Bounding Box for this place.

IV. MODELING

A. Conceptuel modeling

Conceptual modeling provides a level of abstraction independently of technical aspects and focusing on decision-making needs (Rizzi et al., 2006).The multidimensional modeling consists in defining the subject of analysis to be analyzed as a point in a multidimensional space (Kimball, 1996). In fact, the data are organized in such a way to bring out the subject of analysis represented by the concept of fact, composed of measures corresponding to the additive information of the analyzed activity as well as the dimensions of this activity.

A dimension is composed of attributes expressing the characteristics according to which the measures of the fact are analyzed (i.e., activity). The attributes of a dimension can be organized into hierarchies, from the finer to the most general granularity. From the fact and the dimensions, it is possible to build different multidimensional models; the most popular one is called a star model. A star model is composed of one central fact surrounded by dimensions, whereas the constellation model consists in defining a set of facts that share common dimensions. The major drawback of these models is that they do not take into account the specificities of dynamic data such as those from tweets. Indeed, the dimensions contain not-null valued attributes, according to which are analyzed the measures of activity (i.e., fact). However, in practice, by examining the data from a tweet, we found that many of these data are missing (i.e., null-valued). Moreover, the specificity of the Tweet/Tweet-responses requires reviewing the principles used in the implementation of the OLAP cubes in order to reflect their characteristics. This led us to retain the concept of constellation for multidimensional modeling of tweets and for which we will make some extensions in order to reflect the specificities of the data from tweets.

Conventionally, a constellation is composed of inter-connected facts, by common dimensions.

- A constellation C is defined by(F ; D ; *StarC*) where:

  - $F = \{F_1, \ldots, F_n\}$ is a non-empty set of n≥1 facts,
  - $D = \{D_1, \ldots, D_m\}$ is a set of $m \geq 0$ dimensions,
  - *StarC*: $F \rightarrow 2^D$ associates each fact to the set of dimensions, according to which it can be analyzed.

We have extended the concept of fact to add a reflexive relationship (denoted R) between the instances of fact as follows:

- $\forall$ i $\in$ [1..n], a fact F_i is defined by ($NAME_{Fi}$; M_i ; INS_i ; R) where:

  - $NAME_{Fi}$ is the name identifying the fact F_i in the constellation,
  - $M_i = \{m_{i1}, \ldots, m_{ix}\}$ is a set of x measures,
  - $INS_i = \{ins_{i1}, \ldots, ins_{ij}\}$ is the set of j instances of the fact F_i,
  - $R : INS_i \rightarrow INS_i$, as $R(INS_i) = INS_i$.

- $\forall$ j $\in$ [1..x], a measure M_j is defined by ($NAME_j$; T_j ; F_j) where:

  - $NAME_j$ is the name of the measure,
  - T_j is the type of the measure,
  - F_j is a set of aggregation functions, compatible with the summarizability property (i.e. additivity) of the measure, $F_j \subset$ {SUM, AVG, MAX…},

In order to take into account the specificities of data extracted from tweets, we distinguish three types of measures: *numerical* measures, *textual* measures and measure composed of *list of elements*.

- A numerical measure has numerical values.
- A textual measure is a measure whose content is a string (one or more words).
- A measure composed of list of elements consists of a list of keywords, representing the most significant words of a tweet: hashtags in our case (a hashtag is a word or an unspaced phrase pre-fixed with the symbol # indicating the subject assigned to the message).

The OLAP environment offers many aggregate functions, depending on the type of measure. Some of these functions are adapted to the new type of measure presented in this work. Table 2 summarizes the possible aggregate functions by measure type.

TABLE II. MEASURE TYPES AND THEIR AGGREGATE FUNCTIONS

Type of measure	Aggregate Functions allowed
Numeric	Arithmetic functions (SUM, AVG, MIN, MAX,…), COUNT
Textual	TOP_KW ([16]), COUNT
Boolean	COUNT
List	AVG_KW([16]), COUNT

- $\forall$i $\in$ [1..m], a classic dimension D_i is defined by ($NAME_{Di}$; A_i ; H_i) where:
  - $NAME_{Di}$ is the name identifying the dimension in the constellation,
  - $A_i = \{a_{i1},..., a_{iz}\}$ is the set of z dimension attributes (parameters and weak attributes),
  - $H_i = \{h_1,..., h_{ip}\}$ is the set of p hierarchies showing the arrangement of the attributes of D.
- A hierarchy h_j is defined by ($NAME_{hj}$; P_{hj} ; $WEAK_{hj}$) where:
  - $NAME_{hj}$ is the name which identifies the dimension in the constellation,
  - $P_{hj} = \{p_{h1},..., p_{hy}\}$ is the set of parameters of the hierarchy,
  - $WEAK_{hj}$: $P_{hj} \rightarrow 2^W$ associates each parameter to a possible empty subset of weak attributes of the dimension of h_j.

Let us recall that the purpose of this work is to propose a multidimensional model dedicated to online analytical processing (OLAP) and to other more elaborate tweets treatments. Moreover, we aim to ensure that this model is generic; i.e., containing all the data from a tweet and which may be multidimensional concepts [17].

Hence, we examined all data of the tweets in order to judge those that could be potentially useful for OLAP analysis. Following this review, we excluded the following data which we considered a bit useful or even useless:

- Data describing the user profile (e.g., User profile's photo, Background image chosen by the user for his own twitter page, colors for characters and bars...).
- List of contributors of a tweet (i.e., a collection of brief user objects (usually only one) indicating users who contributed to the authorship of the tweet, on behalf of the official tweet author); however, we are restricted to the Boolean indicator ("Contributors-Enabled") to point out whether this account has enabled contributors.

We have identified two facts, a conventional fact called Activity-Twittos and a reflexive fact called Activity-Tweet.

- Activity-Twittos: Corresponds to observations on user accounts and allows the analysis of the user activity over time. It is composed of the following four numerical measures:
  - Fav_C: Number of favorites tweets this user has.
  - Sta_C: Number of tweets this user has.
  - Fre_C: Number of friends (users) this user is following.
  - Follow_C: Number of followers for this user.
- Activity-Tweet: it is a reflexive fact. It models links between a tweet and the person answered and then allows participants and other readers to easily follow the exchange of tweets (cf. Figure 1). Being reflexive, it links instances of the same entity. It is composed of a textual measure (the 140 tweet's characters), measure list of elements (Hashtags) and a numerical measure (Retweet-c) characterizing the number of times a tweet was re-tweeted and to indicate the degree of importance of the tweet exchanged.

The set of dimensions we have created for modeling tweets is as follows:

- USER: composed of elements of the User part of the meta-data of a tweet. This dimension has an identifier, six parameters (language, verified ...) and four weak attributes (name, screen-name, description and URL).
- PLACE: This dimension allows the identification (if the user allowed it during the configuration of his account), the name, the geographical address and phone number (coordinates), and other information about the place associated with tweets.
- TIME: has parameters going from the finest level (Minute) to the most general one (Year). For the fact Activity-tweet, this dimension plays the role of the date of creation of tweets.
- SOURCE: the digital application that sent the tweet (Web, Twitter for Android).

Fig. 1 depicts the extended multidimensional model for tweets. Indeed, the cardinality 0 of a reflexive fact is understandable by the fact that a tweet is not necessarily an answer to another tweet. The second specificity is relative to the possibility of having tweets without any associated locality (absence of the PLACE dimension). This aspect is taken into account by our model. Indeed, we defined a relation of type 1:0 between the fact Activity-Tweet and the PLACE dimension. This occurs when the user did not allow, during the configuration of his Twitter account, the identification of the place which he associated with tweets.

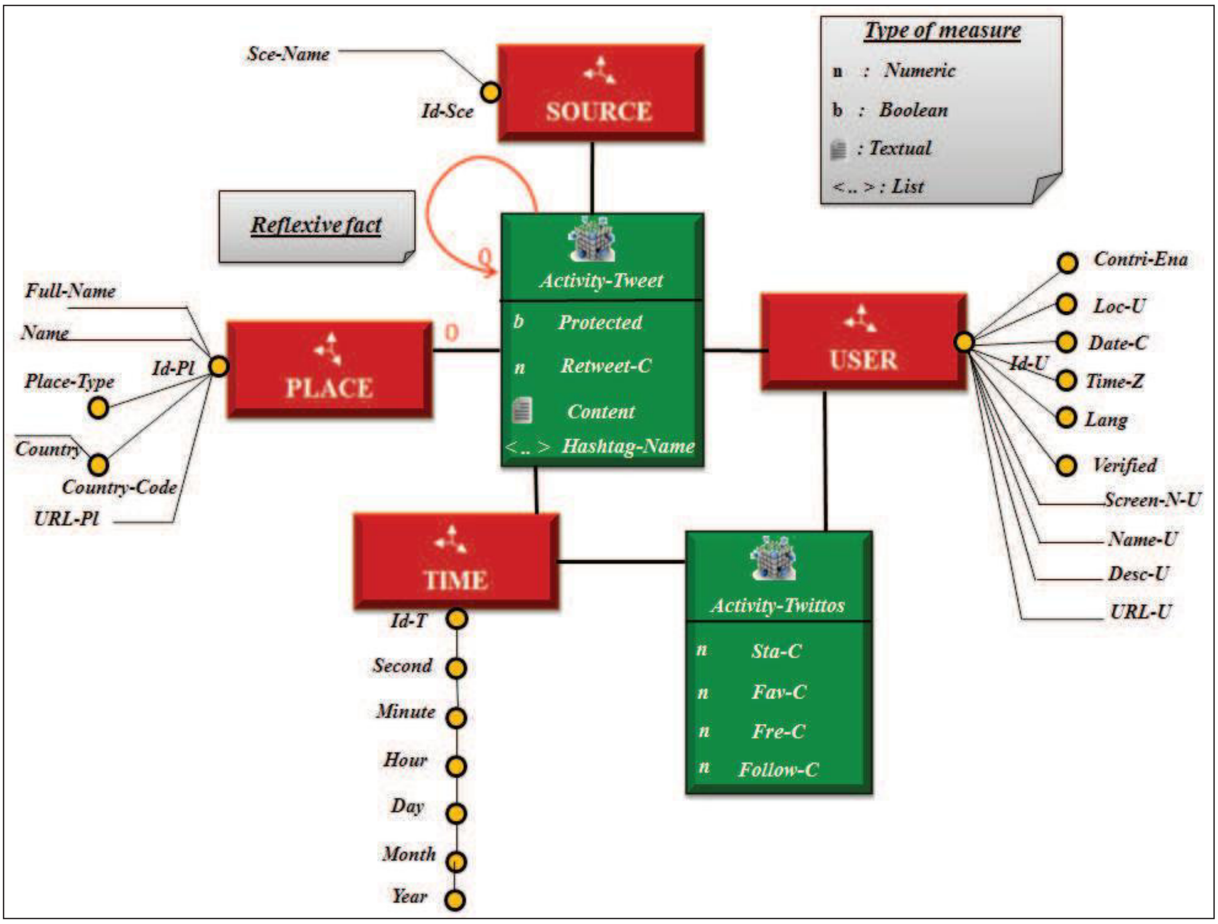

Fig. 1. Multidimensional constellation schema dedicated for the OLAP of tweets.

B. *Logical modeling*

Once the conceptual model defined, the logical model can be derived automatically by applying a set of rules. In this section, we present the main rules of transformation of a constellation into R-OLAP logical model. Although there are various types of R-OLAP model, we decided to detail the rules of transformation for the denormalized R-OLAP model. This model is the most used because few joins are needed during query execution.

We transform the proposed model into R-OLAP logical model according to the following set of three rules:

- Each dimension D is represented by a relation of the same name such that the primary key of the relation is the attribute of the finest level of granularity of D, and the attributes of the relation describe all aggregation levels of the dimension (the parameters and the weak attributes of D).
- Each fact *F* is represented by a relation of the same name composed of attributes representing the measures and the foreign keys referencing the dimensions connected to *F*. For a reflexive fact, the primary key contains an additional attribute (Id-Activity-Twt). The reflexive relationship is supported by the referential constraint. For a non-reflexive fact, the primary key is formed by the concatenation of its foreign keys.
- Each measure of type *list of elements* is transformed into a relational table, of the nameT-MeasureName containing the primary key of the corresponding fact table. The primary key of a T-MeasureName table is the concatenation of the primary key of the fact table and an additional attribute (Position of Hashtag in the tweet).

The processing result of the multidimensional constellation diagram is shown in Figure 2.

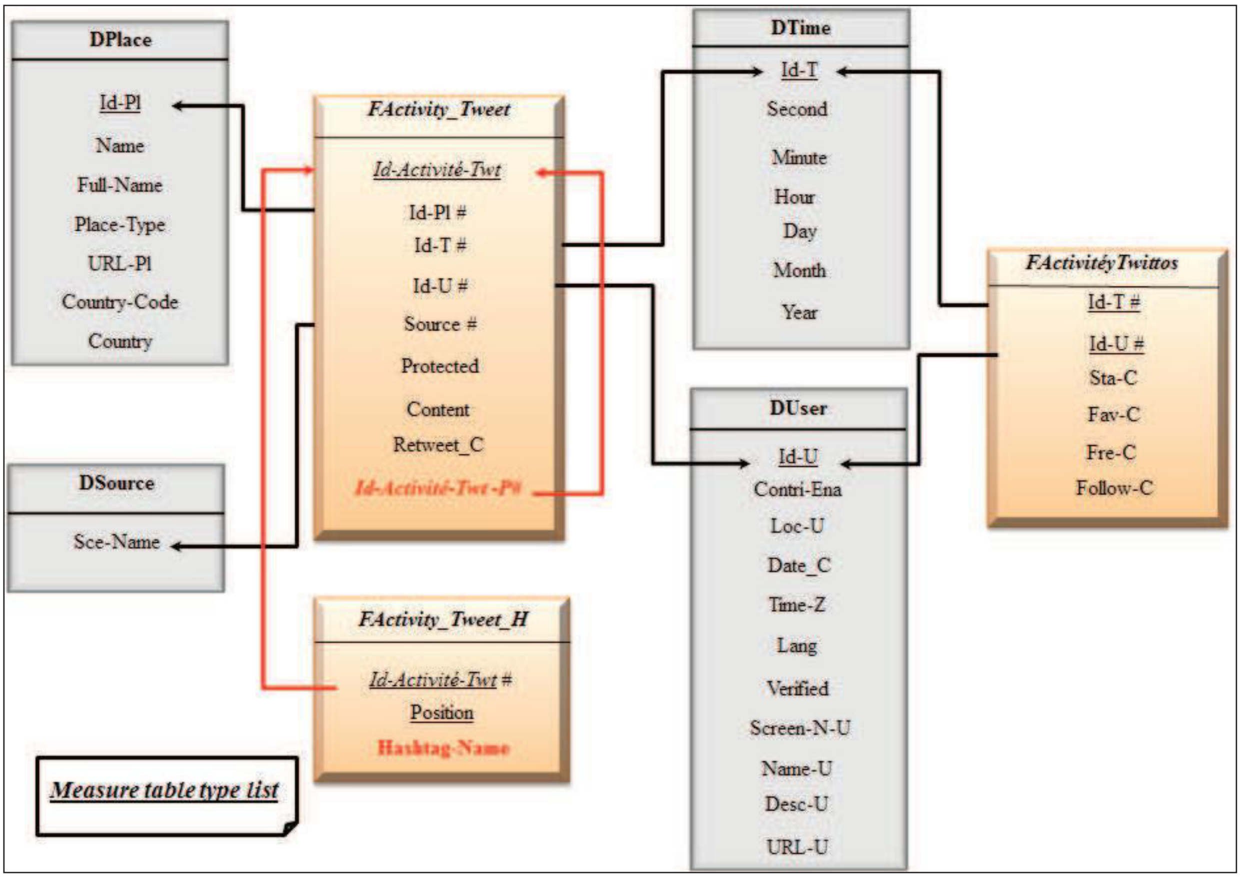

Fig. 2. Logical model R-OLAP

V. EXPERIMENTATION

A. *TweetOLAP developped tool*

In order to evaluate our approach we have developed a software prototype called *TweetOLAP*. Figure 3 depicts its architecture.

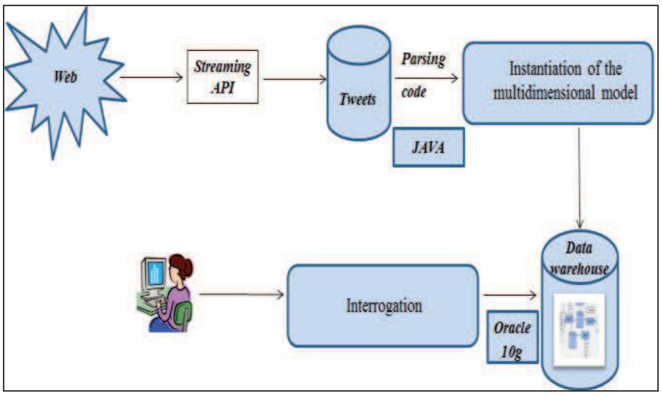

Fig. 3. Architecture of TweetOLAP

This architecture is composed of:

- Streaming API: The Twitter Application Programming Interface (API) currently provides a Streaming API and two discrete REST APIs. The Streaming API ([13]) provides real-time access to Tweets in sampled and filtered form. The API is HTTP based, and GET, POST, and DELETE requests can be used to access the data. In Twitter terminology, individual messages describe the "status" of a user. The streaming API allows near real-time access to subsets of public status descriptions, including replies and mentions created by public accounts. The dataset delivered by the Twitter Streaming API is semi-structured using the JSON (JavaScript Object Notation) as its output format. Each tweet is streamed as an object containing 67 data fields.
- Instantiating the multidimensional model: it consists in feeding automatically the various components of the multidimensional model (fact, dimensions, parameters), from the tweets by using Hibernate and Oracle 10g. The results of this stage are depicted in TABLE III.

TABLE III. DESCRIPTION OF THE DATASET

Table (Dimensions and facts)	Number of instances
DUser	63505
DPlace	741
DTime	65333
DSource	1169
FActivity_Tweet	65333
FActivity_Twittos	65333
FActivity_Tweet_H	13554

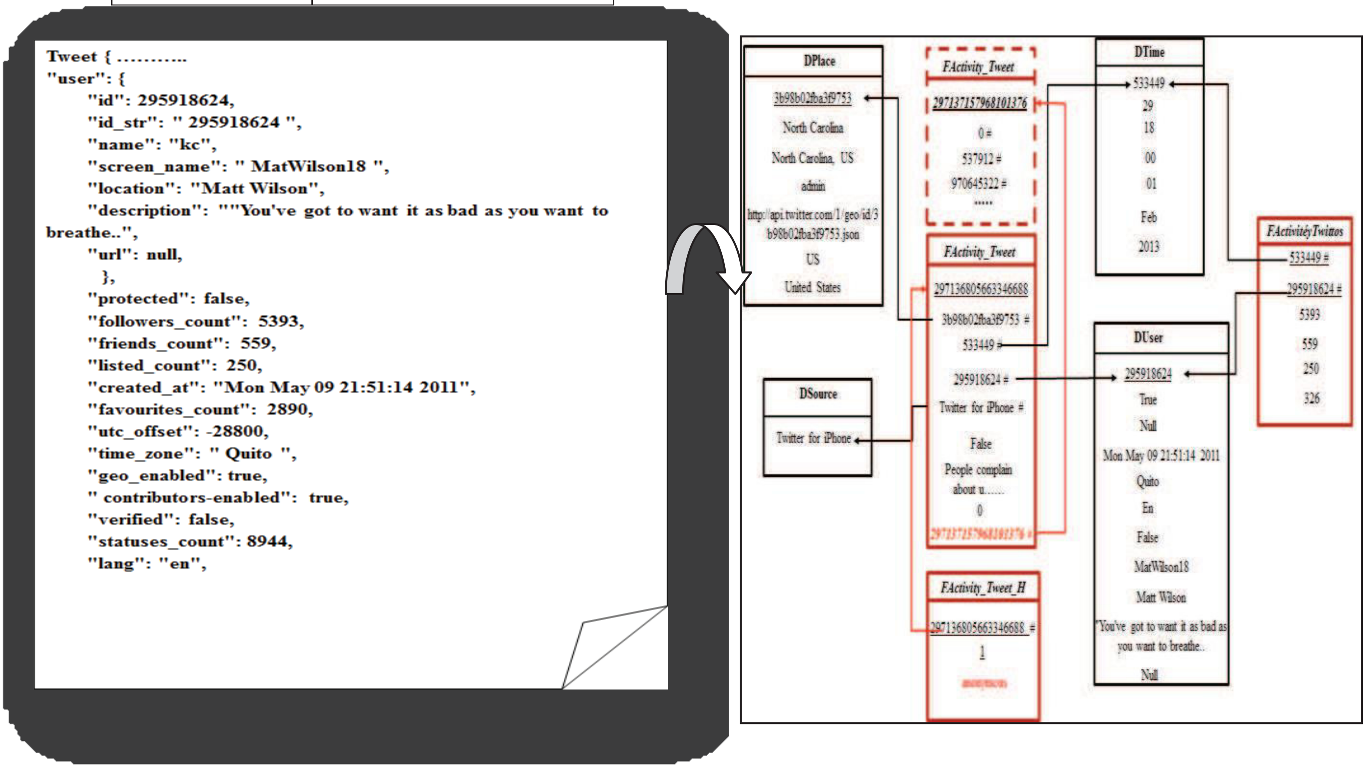

Fig. 4. Example of instantiation

- Once the multidimensional model is generated and fed, the decision maker can perform experiments OLAP analysis on tweets.

B. Experimental OLAP analyses

We present below some results of the OLAP analysis done on about 65333 tweets (cf. TABLE III) retained via the APIs Twitter. These tweets are written in different languages (cf. Fig.5), and collected from February 01 00:00:00 2013 to February 01 00:19:59 2013. We notice that among those tweets, only 1066 tweets were associated with a place and 13305 tweets present tweet-response.

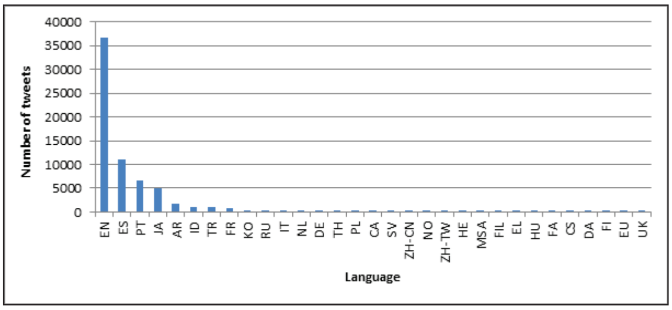

Fig.5. Distribution of tweets per language

First, we study the evolution of Twitter accounts created per year (the creation date for the twitter account) and language (Fact: Activity-twittos). We notice that since twitter was launched, the service rapidly gained worldwide popularity, in a way that the service quickly became popular and most users are from United States.

TABLE IV. DISTRIBUTION OF USERS' ACCOUNTS PER LANGUAGE AND YEAR

Language	Year						
	2007	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*
English	103	424	5584	5544	11315	11324	1399
Arabic	-	-	1	16	142	1288	293
Spanish	15	21	656	2521	3567	3574	424
French	3	3	37	79	218	422	46
Indonesian	-	-	44	93	385	517	46
Italian	-	1	20	27	51	91	13
Japanese	24	32	305	1090	1068	2070	227
Portuguese	2	5	510	1117	2313	1942	314
Russian	-	-	8	15	99	114	16
Turkish	-	-	40	142	350	464	73

Then, we study afterward the distribution of users analyzed by Source and Date (The UTC date time that the user account was created on Twitter: Dimension User). We notice that we have chosen the most relevant source for this analysis. The results presented in TABLE V leads to the following observations. The number of users' accounts is more and more important from 2007 to 2012 especially for the Web source.

TABLE V. DISTRIBUTION OF USERS PER SOURCE AND YEAR

Source	Year						
	2007	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*
web	25	106	1796	3094	5354	5738	977
Twitter for iPhone	14	90	1726	2083	5018	5460	575
Twitter for BlackBerry®	-	8	451	1154	2133	2529	317
Twitter for Android	1	22	812	1354	2997	4010	515
Mobile Web (M2)	1	6	73	198	411	524	80
Twitter for iPad	4	10	149	169	401	486	58
Facebook	1	8	121	116	231	244	26
Instagram	4	14	139	112	184	132	-
iOS	-	3	-	44	107	111	14
UberSocial for BlackBerry	-	-	179	226	240	104	-

Another study was conducted on the number of users by country (Dimension Place) and source (dimension Source), knowing that only 740 tweets were associated to a place in our dataset (65333 tweets). We restricted our analysis on some sources and countries. These results are depicted in TABLE VI.

TABLE VI. NUMBER OF USERS BY COUNTRY AND SOURCE

Source	Country				
	Brazil	*Mexico*	*Türkiye*	*United Kingdom*	*United States*
Twitter for Android	61	13	13	16	99
Twitter for BlackBerry®	-	-	-	3	-
Twitter for iPhone	45	-	-	54	197
Web	133	7	14	9	32

We are now interested in the study of the number of updates produced by language and year. The results of this study are presented in TABLE VI. Most tweets are written in English. These results are explained by the fact that the head office of Twitter is situated in the United States (in San Francisco), the initial interface of Twitter was in English and twitter became more and more popular. Since twitter was launched, the service experienced rapid growth. It had 103 tweets written in English posted in 2007. This grew to 11315 tweets posted in 2012.

TABLE VII. NUMBER OF TWEETS PER LANGUAGE AND YEAR

Language	Year						
	2007	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*
English	103	424	5584	5544	11315	11324	1399
Spanish	15	21	656	2521	3567	3574	424
Portuguese	2	5	510	1117	2313	1942	314
Japanese	24	32	305	1090	1068	2070	227
Arabic	-	-	1	16	142	1288	293
Indonesian	-	-	44	93	385	517	46
Fench	3	3	37	79	218	422	46
Turkish	-	-	40	142	350	464	73
Korean	-	-	6	51	88	152	32
Russian	-	-	8	15	99	114	16
Italian	-	1	20	27	51	91	13
Dutch	-	1	4	28	62	59	7

Each tweet is associated to a time-zone; we are now interested to the study of number of tweets per Time-zone and source. As usually, we only retained the most important sources and times-zones (cf. TABLE VIII).

TABLE VIII. NUMBER OF TWEETS PER TIME-ZONE AND SOURCE

Time-Zone	Source			
	Twitter for Android	*Twitter for BlackBerry®*	*Twitter for iPhone*	*Web*
Central Time (US & Canada)	925	332	1766	1214
Brasilia	334	126	178	2638
Eastern Time (US & Canada)	709	204	1710	1323
Santiago	213	133	164	1366
Quito	489	226	976	603
Greenland	197	-	134	837
Pacific Time (US & Canada)	454	556	788	566
Hawaii	340	224	426	447
Amsterdam	170	116	253	245
Atlantic Time (Canada)	377	-	782	500
Baghdad	119	161	-	224
London	150	136	420	310
Mountain Time (US & Canada)	221	138	418	312
Tokyo	206	-	353	134

The last study was conducted on the variation of the number of tweets produced by language and source. We notice that since the appearance of the phone service provider (Twitter for iPhone, twitter for Android…), the number of tweets produced using twitter for iPhone is more important using English language.

TABLE IX. NUMBER OF TWEETS PER LANGUAGE AND SOURCE

Language	**Source**			
	Twitter for Android	***Twitter for BlackBerry®***	***Twitter for iPhone***	***Web***
English	5907	3149	12068	6665
Spanish	1520	2108	745	4869
Portuguese	543	20	335	3998
Arabic	378	544	411	187
French	143	86	213	246
Japanese	732	2	920	256

C. Discussion

We have drawn the following lessons from these experimental results. In fact, the more data volume is important the more the execution time increases. This is due to the fact that the software tool used for querying the logical multidimensional model (cf. figure 2) is not initially designed to support huge data volumes. In order to alleviate this difficulty, we expect using recent technological tools dedicated to the Big Data paradigm. This will lead us to re-implement the logical model under the Hadoop [19] platform using the Map and Reduce functions. This will speed up the processing, reduce the response time and ensure scalability.

VI. CONCLUSION

The extended multidimensional model we provide is dedicated to the on-line analytical processing (OLAP) of data from exchanged tweets. We have ensured that this model took into account the specifics of data from tweets: links between tweets and tweets answers. For that purpose, we proceeded to an extension of the concept of fact by proposing a new type of fact called *reflexive fact*. This type allows connecting an instance of the fact table to one or several instances of the same table. This relationship will guarantee that every Tweet response added to the table corresponds to an existing Tweet.

Currently, we continue to perform other OLAP experiments on a larger number of tweets. We also intend to propose new OLAP operators which address the specificities of the dynamic data and those of the proposal model (reflexivity). Furthermore, we are interested to exploit the techniques of "Data Mining" in order to extract knowledge from tweets. Twitter as a rich source of social data, is a great starting point for social web mining because of its inherent openness for public consumption and well-documented API.

REFERENCES

[1] T.M .Sakaki and Y.Okazaki. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors",in IEEE Computer society, Vol. 25, Issue 4, April 2013.

[2] M.Mathioudakis and N.Koudas, "Twittermonitor: trend detection over the twitter stream",in Proceedings of 2010 International Conference on Management of Data ,SIGMOD 2010.

[3] S. Chaudhuri and U. Dayal, " Data Warehousing and OLAP for Decision Support". DOOD 1997, pp. 33-34.

[4] Vassiliadis, "A survey of logical models for OLAP databases", ACM SIGMOD Record, Volume 28 Issue 4, Dec.1999, pp, 64 – 69

[5] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, Algebraic and graphic languages for OLAP manipulations. International Journal of Data Warehousing and Mining, IGI Publishing, D. Taniar, Vol. 4, N°1, p.17-46, 2008.

[6] D. S. Daniele Quercia, M.Kosinski and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," IEEE International Conference on Social Computing, pp. 180–185, 2011.

[7] A.Java,X.Song,T.Finin,B.Tseng, "Why we twitter:understanding microbloggingusageandcommunities",in:Proceedingsofthe9th WebKDDand1stSNA-KDD2007WorkshoponWebMiningand Social NetworkAnalysis,ACM,pp.56–65.

[8] O.Phelan, K.McCarthy, B.Smyth, "Using twitter to recommend real-time topical news",in Proceedings of the third ACM conference on Recommender systems. ACM, 2009, pp. 385–388.

[9] A.Bifet, G.Holmes, B.Pfahringer, R.Gavaldà, "Detecting Sentiment Change in Twitter Streaming Data", in 2nd Workshop on Applications of Pattern Analysis, JMLR: Workshop and Conference Proceedings 17 ,2011,pp. 5-11.

[10] E.Cuvalier, M.A.Aufaure, " A Buzz and E-Reputation Monitoring Tool for Twitter Based on Galois Lattices",Conceptual Structures for Discovering Knowledge,Volume 6828, 2011, pp 91-103.

[11] R. Kimball, "The data warehouse toolkit: practical techniques for buildingdimensional data warehouses", John Wiley & Sons, ISBN 0-471-15337-0, 1996

[12] S.Bringay, A.Laurent, P.Poncelet, M. Roche, M.Teisseire, "Towards an On-Line Analysis of Tweets Processing", 22nd International Conference on Database and Expert Systems Applications, DEXA, Toulouse, France, 2011.

[13] N.Rehman, S.Mansmann, A.Weiler, M.H.Scholl,"Building a Data Warehouse for Twitter Stream Exploration.",ACM Fifteenth International Workshop on Data Warehousing and OLAP,DOLAP 2012.

[14] P. Vassiliadis, A. Simitsis, E. Baikousi, "A taxonomy of ETL activities.", DOLAP 2009, pp. 25-32.

[15] F.Bouillot, P.Poncelet, M.Roche, " How and why exploit tweet's location information ? ", Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, France, April, 24-27, 2012. ISBN: 978-90-816960-0-5.

[16] F. Ravat, O. Teste, R. Tournier, OLAP Aggregation Function for Textual Data Warehouse. 9th International Conference on Enterprise Information Systems (ICEIS'07), INSTICC Press, Vol. DISI, J. Cardoso, J. Cordeiro, J. Filipe, Funchal (Madeira, Portugal), p.151-156, juin 2007.

[17] Y.Hachaichi, J,Feki, "An Automatic Method for the Design of Multidimensional Schemas from Object Oriented Databases", in International Journal of Information Technology & Decision Making *(IJITDM*), Ed. World Scientific Publishing Company, vol. 12, n°6, November 2013, pp. 1223-1259.

[18] J.Kalucki, "Twitter streaming API", http://apiwiki.twitter.com/Streaming-API-Documentation, 2010.

[19] ApacheHadoop:http://hadoop.apache.org/.