An Architecture for Metadata Extractor of Big Data in Cloud Systems

Fathy E. Eassa, Hassanin Al-Barhamtoshy, Abdullah Almenbri, Osama H. Younis, Kamal Jambi

Abstract— In this research, we introduce an agent-based architecture for metadata extractor of big data, which is a part of a new framework for managing big data. In the architecture, many agents should be generated simultaneous to be migrated to existing big-data with different types: structured, semi-structured, and unstructured on the remote machines to collect the metadata and returned back to the user. The architecture includes the metadata storage that contains the metadata of existing big data: structured, semi-structured, and unstructured. These metadata describes the existing big data to be processed by software agents for knowledge discovery. The architecture consists of many mobile and stationary agents. The mobile agents migrate to remote machines that include big data to process and discover the required knowledge and return back to the main server. In the main server, the knowledge returned to the user by the interface agent. In this research, many techniques will be built for collecting metadata of different data such as natural language processing techniques, data mining technique, and lexical analysis technique, image processing technique, etc. The manager is built to manage big data on cloud distributed systems. In this paper, the technique of collecting metadata of tweets has been implemented and tested.

Index Terms— big data, mobile agents, metadata extraction, cloud, structured data, semi-structured data, unstructured data.

1 Introduction

Organizations with data warehousing solutions are facing problems of data outburst. Data sets being collected and analyzed for business intelligence are growing rapidly and size of the databases used in enterprises has been growing exponentially. Sources generate data from business processes, transactions, social networking sites, web servers, sensors that exist and collected as meta-data and remains in structured, semi-structured as well as unstructured form [1][3].

Big data in cloud is one of the major issues in computing today, detection of global weather patterns, social phenomena, or economic changes are examples of big data analysis tasks [4]. Cloud computing, poses an important impact on industry of information technology, and research communities [5]. Many of cloud services require users to share their data like health records for data mining and analytical process, taken into consideration privacy and security [6][7]. Enterprises are looking forward for applications having large scale, web-oriented, intensive features to be accessed from diverse devices including mobile devices. Extracting meaningful information, collecting, processing, and analyzing the huge amount of data as large datasets is a challenging task. The most popular open-source map-reduce implementation framework is Apache Hadoop [2], it has been used as an alternative to store and process extremely large data sets on commodity hardware. Scalability, unstructured data, accessibility, real time analytics, fault tolerance are various challenges faced in large data management. Variations means the amount of data stored in different sectors across domains, their data types structured or unstructured examples include graphs, messages, images, audio, video, text/numeric information or data. Data types vary across enterprises [3].

In this paper, we introduce an agent-based manger for extracting metadata of existing big data to be stored in metadata storage. The stored metadata will be used for discovering knowledge and information from big data. Extracting metadata and discovering knowledge based on agent technology solves the transportation and management challenges of big data.

2 Related work

Extracting meaningful knowledge from big data sets is of high importance for organizations. A business has to adapt to some flexible means in terms of infrastructure and software tools. Methodologies or techniques that are being applied to big data so far include data mining grids, massive parallel processing, scalable storage systems, cloud computing platforms, distributed file systems etc. Online transaction processing (OLTP) or Online analytical processing (OLAP), with time based event or transaction time knowledge delivery is the core characteristics in Big Data Analytics. Decision makers or various businesses always look forward for exceptional methods which can process large datasets within tolerable elapsed time. Their formulation uses statistical, mathematical, algebraic and economic operations. Currently computing infrastructures are using various methods and visualization techniques to perform Big Data Analytics [8].

Apache Hadoop is an open-source software solution for scalable and reliable distributed computing. Apache Hadoop's framework contains a library that allows for handling of large datasets across clusters with hundreds of nodes, using a simple programming model. It enables applications to work with thousands of computational independent computers and for petabytes of data. Hadoop was derived from Google's MapReduce and Google File System (GFS) [2].

Hadoop Distributed File System (HDFS) [9] in distributed environment provides fault tolerance and runs on commodity hardware. Data is chopped, stored and is scattered over numerous nodes. HDFS has one master node and multiple slaves' nodes. The name node stores the metadata and data nodes store data blocks.

- *Authors are faculty staff in King Abdulaziz University, faculty of Computing and Information Technology, Saudi Arabia. E-mails, respectively: feassa@kau.edu.sa, hassanin@kau.edu.sa, ammali@kau.edu.sa, osama.26@live.com, kjambi@kau.edu.sa*

All of this architecture resides on commodity hardware where each node/server provides local storage and computation. To store data Hadoop distributed file system (HDFS) uses multiple nodes across networks. On top of it map-reduce framework supports and is a mean to execute jobs across nodes. HDFS has master/slave architecture. Large data is automatically splitted across nodes to be stored and retrieved within hadoop clusters.

In the year of 2004, Google introduced a mapReduce programming framework for managing large data sets [10]. This framework provides ease in processing for distributed environments where data is divided into chunks and is spread over multiple clusters of thousands of nodes as map-reduce jobs. Google's model for executing large data set jobs uses a map function which in turn processes value pairs (key, index and further intermediate pairs) for data. Reduce function merges values, removes duplication for similar intermediate keys [11].

Focusing on the velocity of Big Data, a popular open-source stream processing engine (Storm) is used to perform real integration and trend detection on Twitter and Bitly streams [12]. Also, ClowdFlows platform with the real-time data streams is used to create specialized type of workflow component and a stream mining [13].

3 The High Level Architecture of Big Data Framework

The big data manager (shown in figure 1) consists of two sub-managers: metadata extraction sub-manager and knowledge discovery sub-manager. The metadata extraction sub-manager extracts and retrieves metadata of the big-data on the cloud machines and stores them in metadata storage. This sub-manager consists of multi-agent subsystems: multi-agent subsystem for unstructured data (USMASS-ME), multi-agent subsystem for semi-structured data (SSMASS-ME), and multi-agent subsystem for structured data (SMASS-ME). Each subsystem has many agents; the number of agents depends on the type of data that are processed by the subsystem. For example, the unstructured data includes text, videos, images, tweets, E-mails, and others. The construction of agents depends on extraction technique, type of data, and format of existing big-data.

The knowledge discovery sub-manager discovers the required knowledge or information that is needed by the user. The discovery sub-manager consists of three multi-agent subsystems: unstructured multi-agent sub-system(USMASS-KD), semi-structured multi-agent subsystem (SSMASS-KD), and structured multi-agent subsystem (SMASS-KD). Each subsystem consists of stationary and mobile agents. The task of mobile agent depends on the user query, type of big-data and the technique of discovery. For example, there is a technique for processing texts, another for processing E-mails, and so on. The discovered knowledge and information should be collected and delivered to the user.

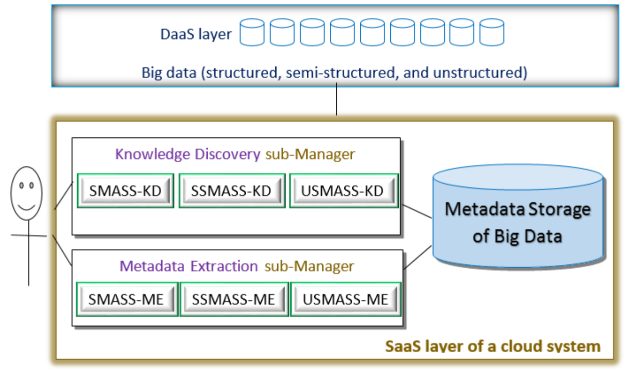

DaaS layer
Big data (structured, semi-structured, and unstructured)
Knowledge Discovery sub-Manager
SMASS-KD
SSMASS-KD
USMASS-KD
Metadata Storage of Big Data
Metadata Extraction sub-Manager
SMASS-ME
SSMASS-ME
USMASS-ME
SaaS layer of a cloud system

Fig. 1. High level Architecture of Big data manager of cloud systems.

3.1 The Agent-based Architecture of Metadata Extraction sub-Manager

Here, we propose the agent-based architecture of the big data Metadata Extraction sub-Manager, which consists of many agents: stationary and mobile agents, see figure 2.

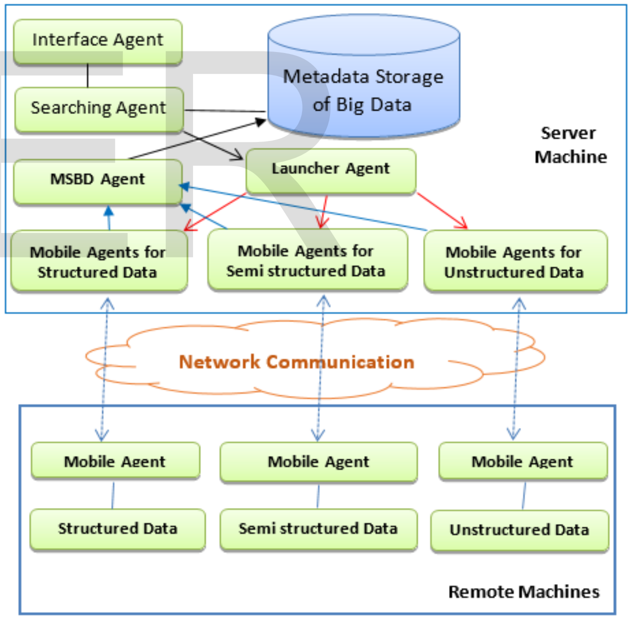


Fig. 2. Agent-based Metadata Extraction Manager of Big data.

Each mobile agent migrates from the big data server machine to a remote machine to collect the required data (metadata) from existing big-data. There are mobile agents for structured data, mobile agents for semi-structured data, and mobile agents for unstructured agents. The mechanism of extracting the required metadata depends on the type of the data. For example, the technique of collecting metadata from a text as unstructured data is different than collecting metadata from E-mail, Tweets, images, videos or other unstructured data. This means we have many different mobile agents for unstructured big data. Also, we have different agents for collecting metadata from semi-structured big data.

The details of some mechanisms, algorithms, and techniques of collecting metadata will be discussed here in this section. The interface agent receives the requests from the user of the big-data manager and displays the outputs. Based on the received requests, the launcher agent creates instances from different types to be migrated into remote machines in the cloud systems for extracting metadata. The agents return back to the main machine and send the extracted metadata to the Metadata Storage of Big Data (MSBD) Agent. The MSBD agent stores the returned metadata in the Metadata Storage of Big Data.

Figure 3 below shows the main activities that have been conducted extract the metadata from big data in a cloud distributed system. At the metadata server where all stationary and mobile agents are installed, many instances of mobile agents are created by the launcher agent. All instances of mobile agents migrate simultaneously into machines in the cloud system to analyze and extract the metadata of big-data.

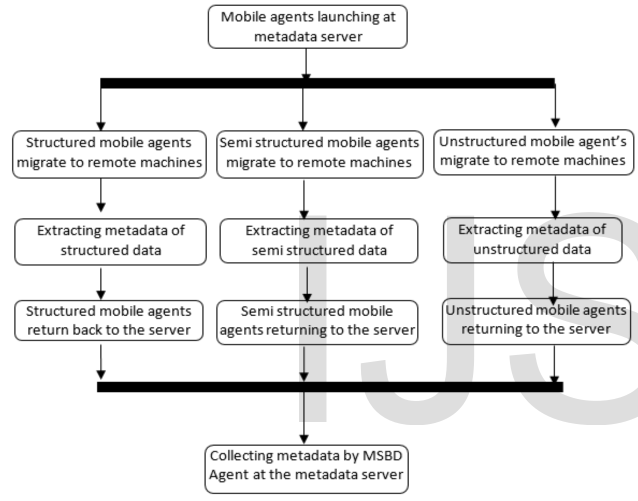

Mobile agents launching at metadata server
Structured mobile agents migrate to remote machines
Semi structured mobile agents migrate to remote machines
Unstructured mobile agent's migrate to remote machines
Extracting metadata of structured data
Extracting metadata of semi structured data
Extracting metadata of unstructured data
Structured mobile agents return back to the server
Semi structured mobile agents returning to the server
Unstructured mobile agents returning to the server
Collecting metadata by MSBD Agent at the metadata server

Fig. 3. Activity Diagram of the process of Retrieving Metadata of Remote.

The algorithms and techniques of analyzing big data that will be implemented in mobile agents depend on the varieties of big-data(structured, semi-structured, and unstructured) and type of data such as free-text, images, e-mails, chats , xml, html, or databases. The natural language processing algorithms and techniques, data mining algorithms, or statistical techniques can be used for analyzing and extracting metadata. Also, the lexical analysis, data mining and parsing techniques can be used especially for analyzing and extracting the metadata from semi-structured. After the analyzing and extracting metadata activity, all mobile agents return back to the metadata server for giving retrieved metadata to the MSBD Agent to be stored in metadata storage. In this paper the technique of extracting metadata from tweets has been introduced.

3.2 The Metadata structure

The proposed structure of the metadata of the framework is represented as a tree as shown in figure 4. The details of the metadata will be investigated in this research.

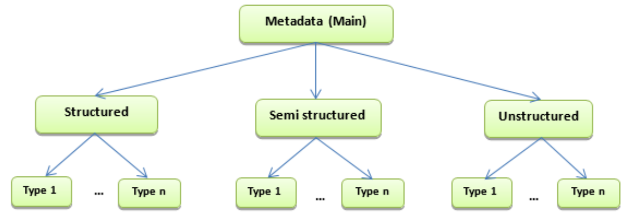


Fig. 4. The Structure of the Metadata Storage.

4 IMPLEMENTATION OF TWEETS METADATA EXTRACTOR

This section presents the way of extracting the tweets' metadata from the social media (Twitter). We have used API called twitter4j (version: 3.0.3) which allows to access and retrieve the tweet's information using Java programming language. It can be used to retrieve different information related to tweet and users such as user name, display name tweet text, date of tweet …etc, and it supports the searching about tweets. Figure 5 shows the elements of the implemented extractor of metadata of Tweets.

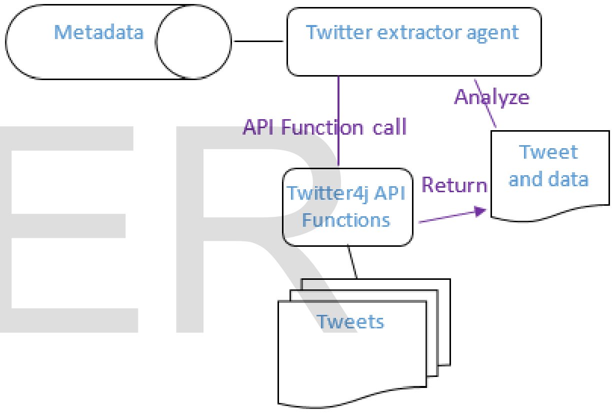


Fig. 4. The Structure of the Metadata Storage.

The process of extracting tweets metadata is as follows:

1- *TwitterExtractor* agent calls the Twitter4j API functions for retrieving the required tweets from the Twitter.

2- *TwitterExtractor* agent applies the algorithm of extraction of tweets metadata (Algorithm 1) to:

a) Retrieve the tweets by calling API function and passing the data in the domain table (table 1) as arguments to the called function.

b) Extract the required metadata from the retrieved tweets.

c) Store the extracted metadata in local data structure to be returned with the agent to the main machine. The returned metadata is stored by MSBD agent in Tweets table (Table 2) in the metadata storage of big data (figure 2)

TABLE 1
DOMAIN TABLE

Domain Table	
Domain-No	Is a Number representing the domain number (Primary key)
Domain	Is a text represent the required domain (hashtag, keyword, domain, topic…etc.)

TABLE 1
DOMAIN TABLE

TABLE 2
TWEETS TABLE

Tweets Table	
tweetNo	Is a serial number of tweet
tweenOwner	User who written the tweet
screenName	Screen name for user who written the tweet
tweetText	Tweet content
tweetDate	Created date of tweet
domain_NO	Domain number (foreign key for relationship with the Domain Table)

The following algorithm shows the steps of extracting metadata from tweets.

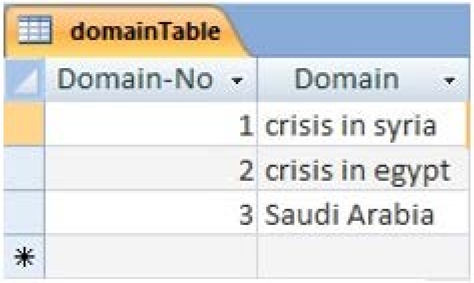
ALGORITHM 1
EXTRACTING METADATA OF TWEETS

```
Begin
   /* create object of ConfigurationBuilder */
   1: builder = new ConfigurationBuilder;
   /*set the four authorizations for this object)*/
   2: builder.SetOAuthAccessToken(AccessToken);
   3: builder.SetOAuthAccessTokenSecret(AccessTokenSecret);
   4: builder.SetOAuthConsumerKey(ConsumerKey);
   5: builder.SetOAuthConsumerSecret(ConsumerSecret);
   /*create OAuth authorization object*/
   6: auth=new OAuthAuthorization;
   /*assign the builder to auth*/
   7: auth=builder.build();
   8: tweet = new Twitter;
   /*assign auth to tweet*/
   9: tweet.authorization = auth;
   /*prepare required query*/
   10: queries[] = domainTableQueries;
   11: for each query in queries[] do
      11.1: queryResult = tweet.search(query);
   /*store tweets info to list of Status*/
      11.2: tweets[] new Status(queryResult.getTweets());
   12: end;
   13: for each tweet in tweets[] do
      13.1: tweetInfo[] = tweet.userName, tweet.screenName,
                        tweet.text, tweet.createdDate, tweet.domainNumber;
    /*store info to the DB*/
      13.2: tweetTable.insert(tweetInfo);
   14: end;
End;
```

5 TESTING THE TWITTER EXTRACTOR

Test cases have been conducted to test the functionality of the extractor. The data in the domain table (Table 3) used as arguments for retrieving tweets to be analyzed to extract the metadata. The results returned based on these queries are stored in different tables shown in the Tables 4, 5, 6.

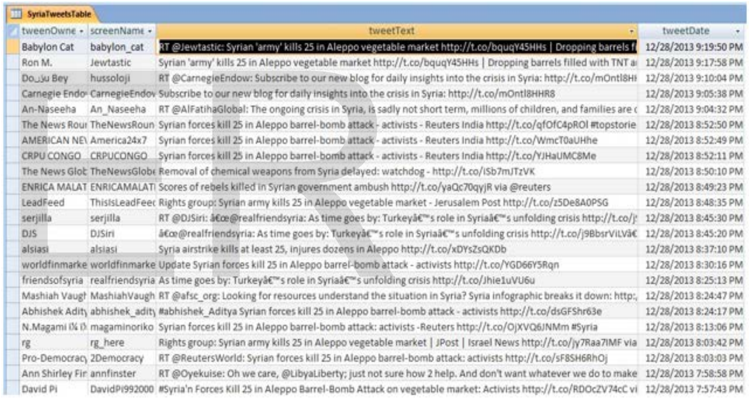
TABLE 3
DOMAINTABLE



domainTable

Domain-No	Domain
1	crisis in syria
2	crisis in egypt
3	Saudi Arabia
✱	

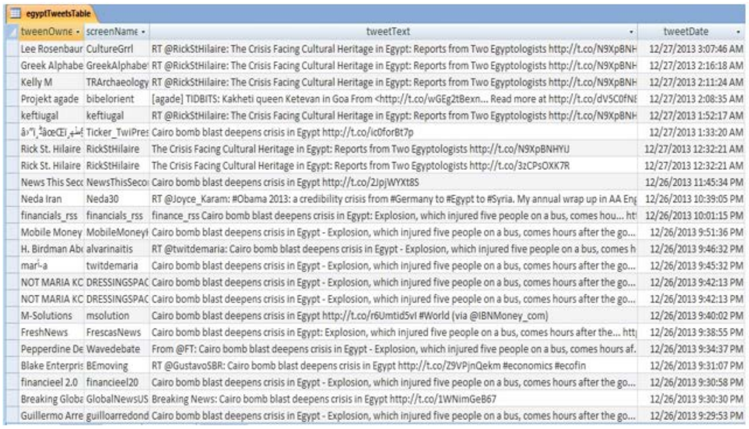
TABLE 4
THE RETRIEVED RESULTS FOR DOMAIN "CRISIS IN SYRIA"



SyriaTweetsTable

tweenOwne	screenName	tweetText	tweetDate
Babylon Cat	babylon_cat	RT @Jewtastic: Syrian 'army' kills 25 in Aleppo vegetable market http://t.co/bquqY45HHs | Dropping barrels f	12/28/2013 9:19:50 PM
Ron M.	Jewtastic	Syrian 'army' kills 25 in Aleppo vegetable market http://t.co/bquqY45HHs | Dropping barrels filled with TNT a	12/28/2013 9:17:58 PM
Do.¡.ıu Bey	hussoloji	RT @CarnegieEndow: Subscribe to our new blog for daily insights into the crisis in Syria: http://t.co/mOntl8HH	12/28/2013 9:10:04 PM
Carnegie Endo	CarnegieEndov	Subscribe to our new blog for daily insights into the crisis in Syria: http://t.co/mOntl8HHR8	12/28/2013 9:05:38 PM
An-Naseeha	An_Naseeha	RT @AlFatihaGlobal: The ongoing crisis in Syria, is sadly not short term, millions of children, and families are c	12/28/2013 9:04:32 PM
The News Rour	TheNewsRoun	Syrian forces kill 25 in Aleppo barrel-bomb attack - activists - Reuters India http://t.co/qfOfC4pROl #topstorie	12/28/2013 8:52:50 PM
AMERICAN NE\	America24x7	Syrian forces kill 25 in Aleppo barrel-bomb attack - activists - Reuters India http://t.co/WmcT0aUHhe	12/28/2013 8:52:49 PM
CRPU CONGO	CRPUCONGO	Syrian forces kill 25 in Aleppo barrel-bomb attack - activists - Reuters India http://t.co/YJHaUMC8Me	12/28/2013 8:52:11 PM
The News Glot	TheNewsGlobe	Removal of chemical weapons from Syria delayed: watchdog - http://t.co/iSb7mJTzVK	12/28/2013 8:50:10 PM
ENRICA MALAT	ENRICAMALATI	Scores of rebels killed in Syrian government ambush http://t.co/yaQc70qyjR via @reuters	12/28/2013 8:49:23 PM
LeadFeed	ThisIsLeadFeec	Rights group: Syrian army kills 25 in Aleppo vegetable market - Jerusalem Post http://t.co/z5De8A0PSG	12/28/2013 8:48:35 PM
serjilla	serjilla	RT @DJSiri: â€œ@realfriendsyria: As time goes by: Turkeyâ€™s role in Syriaâ€™s unfolding crisis http://t.co/j	12/28/2013 8:45:30 PM
DJS	DJSiri	â€œ@realfriendsyria: As time goes by: Turkeyâ€™s role in Syriaâ€™s unfolding crisis http://t.co/j9BbsrViLVâ€	12/28/2013 8:45:20 PM
alsiasi	alsiasi	Syria airstrike kills at least 25, injures dozens in Aleppo http://t.co/xDYsZsQKDb	12/28/2013 8:37:10 PM
worldfinmarke	worldfinmarke	Update Syrian forces kill 25 in Aleppo barrel-bomb attack - activists http://t.co/YGD66Y5Rqn	12/28/2013 8:30:16 PM
friendsofsyria	realfriendsyria	As time goes by: Turkeyâ€™s role in Syriaâ€™s unfolding crisis http://t.co/Jhie1uVU6u	12/28/2013 8:25:13 PM
Mashiah Vaugh	MashiahVaugh	RT @afsc_org: Looking for resources understand the situation in Syria? Syria infographic breaks it down: http:,	12/28/2013 8:24:47 PM
Abhishek Adity	abhishek_adity	#abhishek_Aditya Syrian forces kill 25 in Aleppo barrel-bomb attack - activists http://t.co/dsGFShr63e	12/28/2013 8:24:17 PM
N.Magami Ï¼ Ï	magaminoriko	Syrian forces kill 25 in Aleppo barrel-bomb attack: activists -Reuters http://t.co/OjXVQ6JNMm #Syria	12/28/2013 8:13:06 PM
rg	rg_here	Rights group: Syrian army kills 25 in Aleppo vegetable market | JPost | Israel News http://t.co/jy7Raa7IMF via	12/28/2013 8:03:42 PM
Pro-Democracy	2Democracy	RT @ReutersWorld: Syrian forces kill 25 in Aleppo barrel-bomb attack: activists http://t.co/sF8SH6RhOj	12/28/2013 8:03:03 PM
Ann Shirley Fir	annfinster	RT @Oyekuise: Oh we care, @LibyaLiberty; just not sure how 2 help. And don't want whatever we do to make	12/28/2013 7:58:58 PM
David Pi	DavidPi992000	#Syria'n Forces Kill 25 in Aleppo Barrel-Bomb Attack on vegetable market: Activists http://t.co/RDOcZV74cC vi	12/28/2013 7:57:43 PM

TABLE 5
THE RETRIEVED RESULTS FOR DOMAIN "CRISIS IN EGYPT"

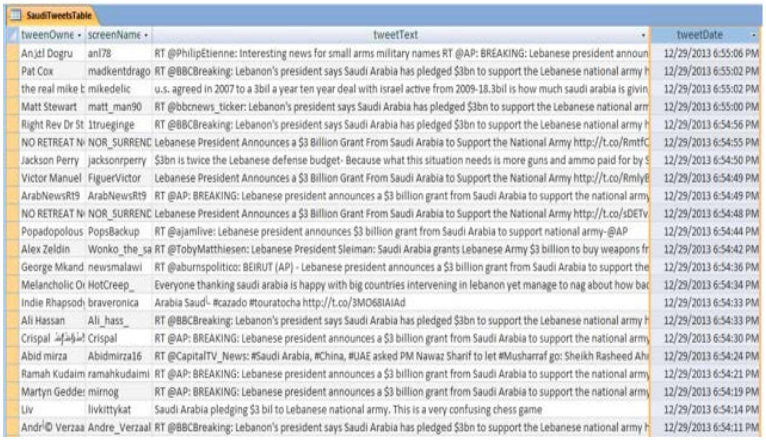


egyptTweetsTable

tweenOwne	screenName	tweetText	tweetDate
Lee Rosenbaur	CultureGrrl	RT @RickStHilaire: The Crisis Facing Cultural Heritage in Egypt: Reports from Two Egyptologists http://t.co/N9XpBNH	12/27/2013 3:07:46 AM
Greek Alphabe	GreekAlphabe	RT @RickStHilaire: The Crisis Facing Cultural Heritage in Egypt: Reports from Two Egyptologists http://t.co/N9XpBNH	12/27/2013 2:16:18 AM
Kelly M	TRArchaeology	RT @RickStHilaire: The Crisis Facing Cultural Heritage in Egypt: Reports from Two Egyptologists http://t.co/N9XpBNH	12/27/2013 2:11:24 AM
Projekt agade	bibelorient	[agade] TIDBITS: Kakheti queen Ketevan in Goa From <http://t.co/wGEg2tBexn... Read more at http://t.co/dVSC0fNE	12/27/2013 2:08:35 AM
keftiugal	keftiugal	RT @RickStHilaire: The Crisis Facing Cultural Heritage in Egypt: Reports from Two Egyptologists http://t.co/N9XpBNH	12/27/2013 1:52:17 AM
â›"ï¸¸âœŒï¸¸إلى	Ticker_TwiPres	Cairo bomb blast deepens crisis in Egypt http://t.co/ic0forBt7p	12/27/2013 1:33:20 AM
Rick St. Hilaire	RickStHilaire	The Crisis Facing Cultural Heritage in Egypt: Reports from Two Egyptologists http://t.co/N9XpBNHYU	12/27/2013 12:32:21 AM
Rick St. Hilaire	RickStHilaire	The Crisis Facing Cultural Heritage in Egypt: Reports from Two Egyptologists http://t.co/3zCPsOXK7R	12/27/2013 12:32:21 AM
News This Secc	NewsThisSeco	Cairo bomb blast deepens crisis in Egypt http://t.co/2JpjWYXt8S	12/26/2013 11:45:34 PM
Neda Iran	Neda30	RT @Joyce_Karam: #Obama 2013: a credibility crisis from #Germany to #Egypt to #Syria. My annual wrap up in AA Eng	12/26/2013 10:39:05 PM
financials_rss	financials_rss	finance_rss Cairo bomb blast deepens crisis in Egypt: Explosion, which injured five people on a bus, comes hou... ht	12/26/2013 10:01:15 PM
Mobile Money	MobileMoneyI	Cairo bomb blast deepens crisis in Egypt - Explosion, which injured five people on a bus, comes hours after the go...	12/26/2013 9:51:36 PM
H. Birdman Ab	alvarinaitis	RT @twitdemaria: Cairo bomb blast deepens crisis in Egypt - Explosion, which injured five people on a bus, comes h	12/26/2013 9:46:32 PM
mar¡-a	twitdemaria	Cairo bomb blast deepens crisis in Egypt - Explosion, which injured five people on a bus, comes hours after the go...	12/26/2013 9:45:32 PM
NOT MARIA KC	DRESSINGSPAC	Cairo bomb blast deepens crisis in Egypt - Explosion, which injured five people on a bus, comes hours after the go...	12/26/2013 9:42:13 PM
NOT MARIA KC	DRESSINGSPAC	Cairo bomb blast deepens crisis in Egypt - Explosion, which injured five people on a bus, comes hours after the go...	12/26/2013 9:42:13 PM
M-Solutions	msolution	Cairo bomb blast deepens crisis in Egypt http://t.co/r6Umtid5vI #World (via @IBNMoney_com)	12/26/2013 9:40:02 PM
FreshNews	FrescasNews	Cairo bomb blast deepens crisis in Egypt: Explosion, which injured five people on a bus, comes hours after the... htt	12/26/2013 9:38:55 PM
Pepperdine De	Wavedebate	From @FT: Cairo bomb blast deepens crisis in Egypt - Explosion, which injured five people on a bus, comes hours af.	12/26/2013 9:34:37 PM
Blake Enterpris	BEmoving	RT @GustavoSBR: Cairo bomb blast deepens crisis in Egypt http://t.co/Z9VPjnQekm #economics #ecofin	12/26/2013 9:31:07 PM
financieel 2.0	financieel20	Cairo bomb blast deepens crisis in Egypt - Explosion, which injured five people on a bus, comes hours after the go...	12/26/2013 9:30:58 PM
Breaking Globa	GlobalNewsUS	Breaking News: Cairo bomb blast deepens crisis in Egypt http://t.co/1WNimGeB67	12/26/2013 9:30:30 PM
Guillermo Arre	guilloarredond	Cairo bomb blast deepens crisis in Egypt - Explosion, which injured five people on a bus, comes hours after the go...	12/26/2013 9:29:53 PM

TABLE 6

THE RETRIEVED RESULTS FOR DOMAIN "SAUDI ARABIA"

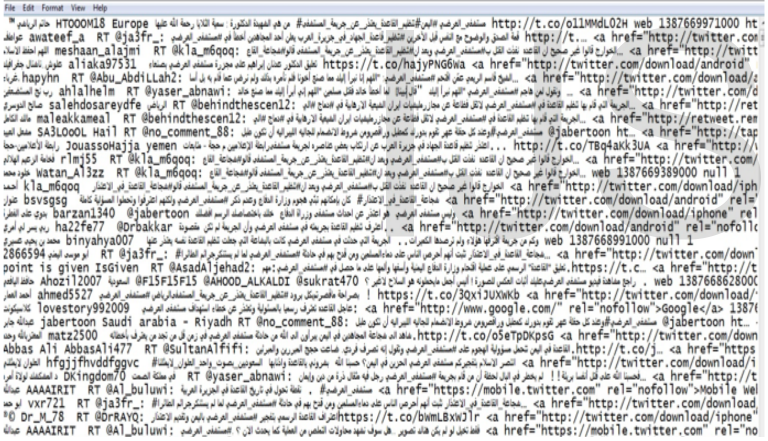


tweenOwne	screenName	tweetText	tweetDate
Anعtl Dogru	anl78	RT @PhilipEtienne: Interesting news for small arms military names RT @AP: BREAKING: Lebanese president announ	12/29/2013 6:55:06 PM
Pat Cox	madkentdrago	RT @BBCBreaking: Lebanon's president says Saudi Arabia has pledged $3bn to support the Lebanese national army h	12/29/2013 6:55:02 PM
the real mike b	mikedelic	u.s. agreed in 2007 to a 3bil a year ten year deal with israel active from 2009-18.3bil is how much saudi arabia is givin	12/29/2013 6:55:02 PM
Matt Stewart	matt_man90	RT @bbcnews_ticker: Lebanon's president says Saudi Arabia has pledged $3bn to support the Lebanese national arm	12/29/2013 6:55:00 PM
Right Rev Dr St	1trueginge	RT @BBCBreaking: Lebanon's president says Saudi Arabia has pledged $3bn to support the Lebanese national army h	12/29/2013 6:54:56 PM
NO RETREAT N	NOR_SURRENC	Lebanese President Announces a $3 Billion Grant From Saudi Arabia to Support the National Army http://t.co/RmtfC	12/29/2013 6:54:55 PM
Jackson Perry	jacksonrperry	$3bn is twice the Lebanese defense budget- Because what this situation needs is more guns and ammo paid for by S	12/29/2013 6:54:50 PM
Victor Manuel	FiguerVictor	Lebanese President Announces a $3 Billion Grant From Saudi Arabia to Support the National Army http://t.co/RmlyE	12/29/2013 6:54:49 PM
ArabNewsRt9	ArabNewsRt9	RT @AP: BREAKING: Lebanese president announces a $3 billion grant from Saudi Arabia to support the national army	12/29/2013 6:54:49 PM
NO RETREAT N	NOR_SURRENC	Lebanese President Announces a $3 Billion Grant From Saudi Arabia to Support the National Army http://t.co/sDETv	12/29/2013 6:54:48 PM
Popadopolous	PopsBackup	RT @ajamlive: Lebanese president announces $3 billion grant from Saudi Arabia to support national army-@AP	12/29/2013 6:54:44 PM
Alex Zeldin	Wonko_the_sa	RT @TobyMatthiesen: Lebanese President Sleiman: Saudi Arabia grants Lebanese Army $3 billion to buy weapons fr	12/29/2013 6:54:42 PM
George Mkand	newsmalawi	RT @aburnspolitico: BEIRUT (AP) - Lebanese president announces a $3 billion grant from Saudi Arabia to support the	12/29/2013 6:54:36 PM
Melancholic Or	HotCreep_	Everyone thanking saudi arabia is happy with big countries intervening in lebanon yet manage to nag about how bac	12/29/2013 6:54:34 PM
Indie Rhapsody	braveronica	Arabia Saudل #cazado #touratocha http://t.co/3MO68IAIAd	12/29/2013 6:54:33 PM
Ali Hassan	Ali_hass_	RT @BBCBreaking: Lebanon's president says Saudi Arabia has pledged $3bn to support the Lebanese national army h	12/29/2013 6:54:33 PM
Crispal [illegible]	Crispal	RT @AP: BREAKING: Lebanese president announces a $3 billion grant from Saudi Arabia to support the national army	12/29/2013 6:54:30 PM
Abid mirza	Abidmirza16	RT @CapitalTV_News: #Saudi Arabia, #China, #UAE asked PM Nawaz Sharif to let #Musharraf go: Sheikh Rasheed Ahr	12/29/2013 6:54:24 PM
Ramah Kudaim	ramahkudaimi	RT @AP: BREAKING: Lebanese president announces a $3 billion grant from Saudi Arabia to support the national army	12/29/2013 6:54:21 PM
Martyn Gedde:	mirnog	RT @AP: BREAKING: Lebanese president announces a $3 billion grant from Saudi Arabia to support the national army	12/29/2013 6:54:19 PM
Liv	livkittykat	Saudi Arabia pledging $3 bil to Lebanese national army. This is a very confusing chess game	12/29/2013 6:54:14 PM
Andrل© Verzaa	Andre_Verzaal	RT @BBCBreaking: Lebanon's president says Saudi Arabia has pledged $3bn to support the Lebanese national army h	12/29/2013 6:54:11 PM

In addition to the above tests, extracting metadata of tweets written in Arabic language has been conducted as shown in Table 7.

TABLE 7

THE RETRIEVED RESULTS FOR ARABIC DOMAIN (مستشفى_العرضي , البرد في سوريا)



```
[illegible] HTOOOM18 Europe [illegible] http://t.co/o1lMMdL02H web 1387669971000 ht
[illegible] awateef_a  RT @ja3fr_: [illegible] http://t... <a href="http://twitter.con
[illegible] meshaan_alajmi  RT @kla_m6qoq: [illegible] ... <a href="http://twitt
[illegible] aliaka97531  [illegible] https://t.co/hajyPNG6Wa <a href="http://twitter.com/download/android"
[illegible] hapyhn  RT @Abu_AbdiLLah2: [illegible] ... <a href="http://twitter.com/download/
[illegible] ahlalhelm  RT @yaser_abnawi: [illegible] ... <a href="http://twitter.com/d
[illegible] salehdosareydfe [illegible] RT @behindthescen12: [illegible] ... <a href="http://ret
[illegible] maleakkameal  RT @behindthescen12: [illegible] ... <a href="http://retweet.rem
[illegible] SA3LOOOL Hail RT @no_comment_88: [illegible] @jabertoon ht... <a href="http://tap
[illegible] JouassoHajja yemen [illegible] ... http://t.co/TBq4aKk3UA <a href="http://
[illegible] rlmj55  RT @kla_m6qoq: [illegible] ... <a href="http://twitter.com/
[illegible] Watan_Al3zz  RT @kla_m6qoq: [illegible] ... web 1387669389000 null 1
[illegible] kla_m6qoq  [illegible] <a href="http://twitter.com/download/iph
[illegible] bsvsgsg  [illegible] <a href="http://twitter.com/download/android" rel='
[illegible] barzan1340  @jabertoon [illegible] <a href="http://twitter.com/download/iphone" rel
[illegible] ha22fe77  @Drbakkar [illegible] .. <a href="http://twitter.com/download/android" rel="nofollc
[illegible] binyahya007  [illegible] .. [illegible] web 1387668991000 null 1
2866594 [illegible]  RT @ja3fr_: [illegible] ... <a href="http://twitter.com/down
point is given IsGiven  RT @AsadAljehad2: [illegible] https://t.c... <a href="http://t
[illegible] Ahozil2007 [illegible] @F15F15F15 @AHOOD_ALKALDI @sukrat470 [illegible] . web 138766862800(
[illegible] ahmed5527 [illegible] ! https://t.co/3QxiJUXWKb <a href="http://twitter.com/download/
[illegible] lovestory992009  [illegible]: <a href="http://www.google.com/" rel="nofollow">Google</a> 1387(
[illegible] jabertoon Saudi arabia - Riyadh RT @no_comment_88: [illegible] @jabertoon ht... -
[illegible] matz2500  [illegible] http://t.co/o5eTpDKpsG <a href="http://twitter.com/down
Abbas Ali AbbasAli477  RT @SultanAlfifi: [illegible] http://t.co/j... <a href="https
[illegible] hfgjjfhvddfggvc  [illegible] <a href="http://twitter.com/download/i
[illegible] DKingdom70 [illegible]  RT @yaser_abnawi: [illegible] !! [illegible] ... <a href="http://t
[illegible] AAAAIRIT  RT @Al_buluwi: [illegible] . [illegible] <a href="https://mobile.twitter.com" rel="nofollow">Mobile Wel
[illegible] vxr721  RT @ja3fr_: [illegible] ... <a href="http://twitter.com/download
°© Dr_M_78  RT @DrRAYQ: [illegible]https://t.co/bWmLBxWJlr <a href="http://twitter.com/download/iphone"
[illegible] AAAAIRIT  RT @Al_buluwi: [illegible] <a href="https://mobile.twitter.com" rel="no
```

6 CONCLUSION

In this research, an agent based sub-system for collecting the metadata of big data has been introduced. The sub-system is part of a big data manager that solves two big data challenges: big data transportation, and big data management. The transportation challenge will be solved by implementing part II of our manager. Part I that has been introduced collects the metadata of existing big data. The metadata of big data is different from type to other. Therefore, the design and techniques of mobile agents for collecting the metadata are different. In this paper we introduced the algorithm and technique of collecting metadata of tweets. The algorithm has been implemented as agent called Twitter Extractor Agent. . The agent has been tested and the collected metadata stored in the metadata storage. Our big data manager is scalable because many instances of agents can be created at the same time to be migrated to many machines. Therefore, the manager can manage any size of big data stored on any number of storage media and machines.

In the future, many extractors will be built to enhance the implementation of the manager. For example we will built extractor for E-mails, extractor for text, extractor for XML, and so on.

REFERENCES

[1] Impetus white paper, March 2011, "Planning Hadoop/NoSQL Projects for 2011" by Technologies, Available at: http://www.techrepublic.com/whitepapers/planninghadoopnosql-projects-for-2011/2923717.

[2] Apache Hadoop. [Online] Available at: http://wiki.apache.org/hadoop. [Accessed Jan 2014].

[3] A. Jacobs. (2009, Jul.) The pathologies of big data. [Online]. Available: http://queue.acm.org/detail.cfm?id=1563874 [Accessed Jan 2014].

[4] Chamikara Jayalath, Julian Stephen, and Patrick Eugster, (2014). From the Cloud to the Atmosphere: Running MapReduce across Data Centers, IEEE TRANSACTIONS ON COMPUTERS, VOL. 63, NO. 1, JANUARY 2014, pp. 74 - 87.

[5] Xuyun Zhang, Laurence T. Yang, Senior Member, Chang Liu, and Jinjun Chen, (2014). A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 2, FEBRUARY 2014, pp. 363 - 373.

[6] D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues," Future Generation Computer Systems, vol. 28, no. 3, pp. 583-592, 2011.

[7] L. Hsiao-Ying and W.G. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 6, pp.

[8] McKinsey Global Institute, 2011, Big Data: The next frontier for innovation, competition, and productivity, Available: www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx, Aug 2012.

[9] Apache Software Foundation. [Online] Official apache Hadoop website: http://hadoop.apache.org/, [Accessed January 2014].

[10] The Hadoop Architecture and Design. [Online], Available: http://hadoop.apache.org/common/docs/r0.16.4/hdfs_design.html, [Accessed Jan 2014].

[11] Hung-Chih Yang, Ali Dasdan, Ruey-Lung Hsiao, and D. Stott Parker from Yahoo and UCLA, "Map-Reduce-Merge: Simplified Data Processing on Large Clusters", paper published in Proc. of ACM SIGMOD, pp. 1029–1040, 2007.

[12] Thibaud Chardonnens, Benoit Perroud, (2013). Big Data Analytics on High Velocity Streams: A Case Study, 2013 IEEE International Conference on Big Data, pp. 784 – 787.

[13] Kranjc, Janez ; Podpecan, Vid ; Lavrac, Nada, (2013). Real-time data analysis in ClowdFlows, 2013 IEEE International Conference on Big Data, pp. 15 – 22.