Proposed Business Intelligence Systemthrough Big Data

Dr. Hasanen S. Abdullah
Computer Sciences Department, University of Technology/Baghdad
Saif Bashar Neama
Computer Sciences Department, University of Technology/Baghdad
Email: SAIF13LP@YAHOO.COM

Received on:7/10/2015 & Accepted on:9/3/2016

ABSTRACT

Every company or institution in the world has huge amounts of raw data. Since, we are living now in the era of data and data explosion, data is generated in an alarming rates. As the Big Data problem emerges, big data cannot be processed by traditional systems due to its huge size, complexity and rapid generation, and since data became the most important element in the business world to drive companies and institutions in the right direction. Business Intelligence Systems are built to serve that purpose. This paper introduces a system that will implement a Business Intelligence technique to handle Big Data problem through Hadoop Framework benefiting from its functionalities to provide a parallel processing environment by implementing a cluster of three nodes that will hold the data set and running queries on it in parallel using the functionality of Map Reduce algorithm. The system consists of four primary stages: first is the stage of loading the data into the cluster, second is the stage of constructing the data warehouse to become the source layer for the data analysis stage to extract the business insights. The third stage analyzes the data and answers the business problems. Finally the fourth stage is the data visualization stage where the answers that gathered from the previous stage will take the form of visual charts and graphs that will be contained in a unified business intelligence dashboard that will provide the overall look for the business operations.

Keywords: Business Intelligence, Big Data, Hadoop, and Big Data Analytics.

INTRODUCTION

Any company or an institution in the world has an enormous amounts of data, usually these data are out of tune flowing through the operational systems, the raw data in this form is useless for the companies' decision makers due to its size, so there must be a way to access, process and display the data as meaningful information to produces business insights that help the companies' decision makers to take crucial decisions that effects the company's future; that's when Business Intelligence Systems comes to picture, data is produced from transactions, log files, digital media, sensors, social media and other sources all have the ability to provide organizations with perceptions and a complete picture of customers and a stakeholders behaviors and ways to get a competitive benefits **[1]**.

Today with the fast growing and changing business environment that needs correct and just in time information is not only necessary for the success of a company but also is required for remaining in competition, business intelligence system does all that by analyzing the data through ad hoc querying, then translating the results of those queries in tovisualizations that are gathered in a single business intelligence dashboard**[2]**.

There are number of studies that are conducted on the Business Intelligence and Big Data topics, some of these studies are described below:

https://doi.org/10.30684/etj.34.4B.10
2412-0758/University of Technology-Iraq, Baghdad, Iraq

In 2010, Manoochehr Najmi, Mehran Sepehri and Spideh Hashemi, presented a work entitled "The Evaluation of Business Intelligence Maturity Level in Iranian Banking Industry" that address improving the maturity level of the business intelligence systems in the Iranian banks by improving data warehouse and data extraction capabilities in the stages of acquiring the data and the stage of analyzing the data, to add more value to their decision making processes [3].

In 2010, Charmaine Felder, presented work entitled "The Potential Role of Business Intelligence in Church Organizations" that shows the importance of business intelligence system implementation in a non-profit organizations and its benefits in addressing the informational needs and thereby improving decision making and the benefits in time saving, decreasing expenses and reaching wider audiences [4].

In 2012, Aditya B. Patel, Manashvi Birla and Ushma Nair, addressed a work entitled "Addressing Big Data Problem Using Hadoop and Map Reduce" the big data problem and how to solve it using Hadoop data cluster, HDFS and Map Reduce programming framework, as the needs of analyzing the data that grows at an exponential rates to improve business decision making and scientific applications **[5]**.

Business Intelligence

Business Intelligence or (BI) for short is the collection of techniques, methods and tools that is used to transform raw data into meaningful information to serve the business analysis purposes. Managing these amounts of data help to identify and create new strategic business opportunities. The main goal of a BI system is to allow the easy interpretation of these data by identifying new opportunities and implementing effective strategies based on insights that has an impact on the organization's operations by reducing cost, enhance sales, improve operations or any positive factor; Business managers and decisions makers are always under pressure to sustain their organizations' competitiveness and make the right decisions in challenging business environments. Business managers make use of BI tools to assist them make intelligent decisions and understand their business situations **[6]**.

Data Warehousing

A data warehouse is a single logical repository for an organization's transactional data, designed to separate the operational day-to-day operations that tend to be fast and responsive systems like the systems on the point of sale using Online Transactional Processing (OLTP), from the informational or analytical operations using Online Analysis Processing (OLAP), The data sources are connected to the data warehouse through a set of processes that are referred to as Extract Transform and Load (ETL) processes, the data warehouse will represent the BI Layer's source for the enterprise to enable strategic business decisions, Figure (1) describe the traditional BI system components [7].

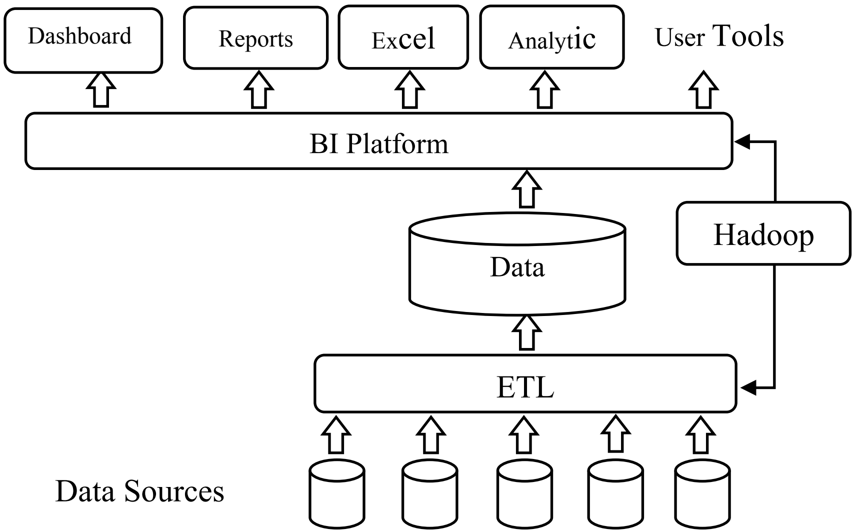

Figure (1), Traditional BI System Components

Big Data

By the wide spread use of the internet and the fact that it is growing at an exponential rate as IBM stated that everyday 2.5 quintillion bytes of data is generated, that's what now is called ***data explosion***, the massive generation of data is because of the various sources of data available nowadays; sources like sensor networks that gather climate information or surveillance information, RFID readers, machine generated data, internet websites log files, transaction systems, social media networks, GPS data, emails and other sources **[8]**, Another study conducted by EMC stated that from 2005 to 2020 the digital universe will grow by a factor of 300, from 130 Exabyte to 40.000 Exabyte which means more than 5200 gigabytes for every person on earth in 2020 as Figure (2) **[9]**.

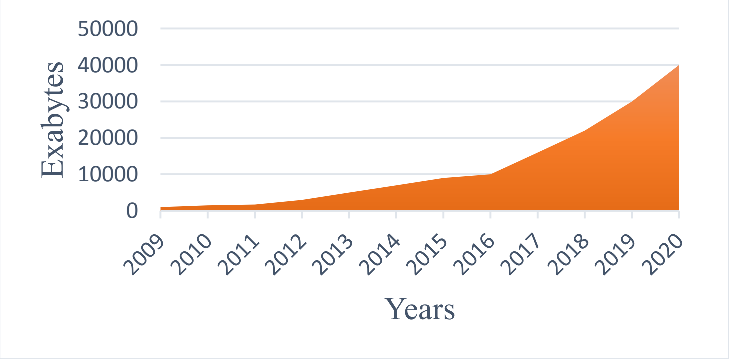

Figure (2), Data Explosion

The main three characteristics of big data is what IBM called the ***3Vs***, the first V is for ***Volume*** which is big, big data comes in huge volumes in terabytes and even peta bytes of data sizes, the second V is for ***Velocity*** which means that big data is generated continuously and in a fast rate; The third V is for ***Variety*** which means data has various formats and types there is structured data like the traditional relational databases that is confined in tables with a specific schema or structure applied to the data at the time of creation, and there is unstructured data like free text, images, audio and video files, and also there is the semi-structured data like Extensible Markup Language (XML) and Comma Separated Values (CSV) files**[8]**.

Hadoop

Hadoop is a Linux based framework programmed in Java, and built for distributed and parallel processing on a very large data sets across clusters of commodity hardware using a simple programming model called ***MapReduce***, Hadoop now is an ecosystem that contains many subsystems and projects but the main components of Hadoop or what's called Hadoop Common are:

- Hadoop Distributed File System (HDFS)
- MapReduce (The Programming Model)

Hadoop implements a Master/Slave architecture where the HDFS represents the storage part of Hadoop that will hold the data, HDFS is the file system of multiple nodes (slaves) in a Hadoop cluster these nodes are called ***DataNodes;*** DataNodes which are individual machines that hold the data in HDFS as fixed size blocks that are replicated across the nodes based on a predefined parameters; the master server is called the ***NameNode***it is a dedicated machine that is responsible for storing the metadata for the system and monitors the system's operations, the metadata contains the location of each data block on which Data Node across the cluster and also contains permissions and the file name; the NameNode manages the file system, regulate access to files by clients and oversees the health of all the DataNodes in the cluster**[10]**.

Another main components in the architecture of Hadoop are the ***Job Tracker***and the ***Task Tracker,*** The Task Tracker which resides on each of the DataNode, it is a daemon process that runs in the background and spawns child processes to perform the actual map or reduce tasks, The JobTracker which resides in the NameNode is responsible for coordinating all the activities across the slave TaskTracker Processes, it accepts MapReduce job requests from clients and schedules mappers and reducers on TaskTrackers to perform the work **[11]**.

Figure (3) describes the interaction between these components.

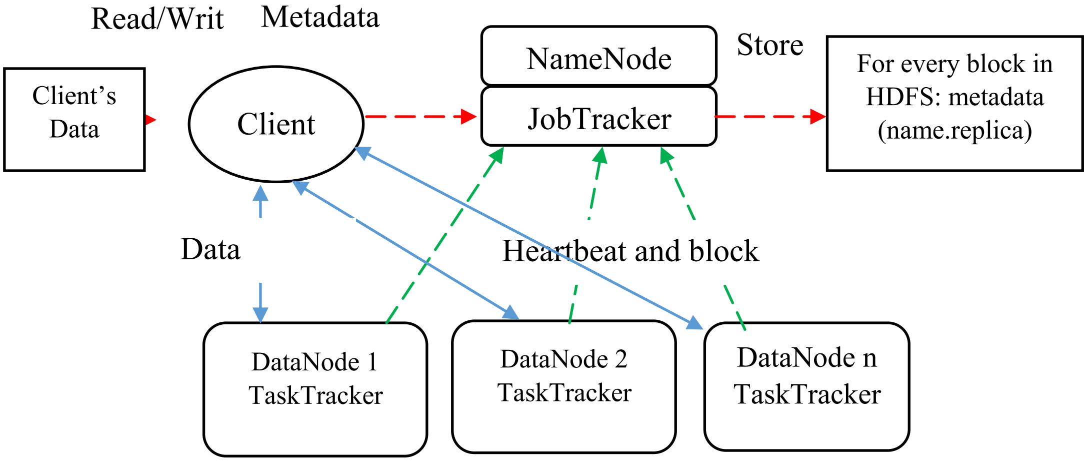

Figure (3), Hadoop Architecture [12].

The Proposed System

The proposed system implements a Hadoop cluster that contains three dedicated machines that are connected via LAN switch in a star topology connection in a fully distributed multi-node implementation and all the work is done under Linux environment.

The proposed system is used to work on a sample data set that exceeded 5.7 GB in size in CSV file format as a sample of semi-structured data with more than 36 million records that contains household level transactions for more than two years of transactions from a group of a 2500 households who are frequent shoppers at the retailer stores that are located in different locations and contains various types of commodities, this data set failed to be accessed and displayed in the traditional database management systems due to its large size and the fact that any software application has limitations in handling huge amounts of data, the proposed system has one machine is implemented as a dedicated master node (NameNode), the NameNode may have more powerful CPU and less storage since its main responsibilities are to monitor and coordinate the cluster's operations, and the other two machines are implemented as slave nodes (DataNode1 and DataNode2) they represent the storage part in Hadoop as HDFS is implemented on them as Figure (4) demonstrates.

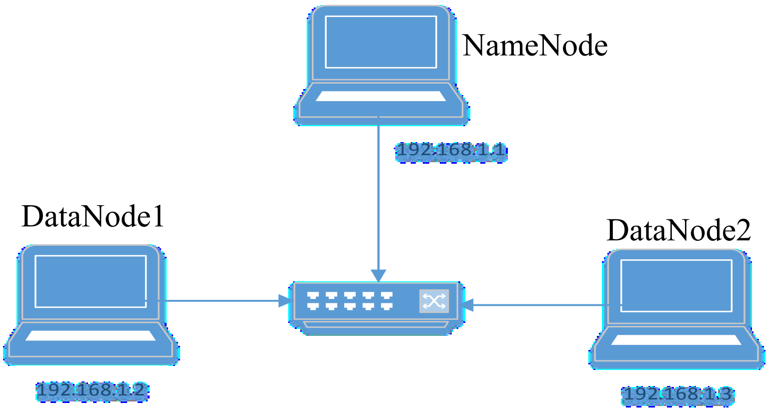

Figure (4), The Proposed System's Hadoop Cluster

Main System's Architecture

The general architecture of the proposed system contains several stages that starts with loading the big data set from the local system into Hadoop's HDFS where the data is divided into blocks which are distributed inside the slave DataNodes through the network, then the process of Extract Transform and Load (ETL) begins to extract the data from HDFSand perform some transformations on it then load the transformed data into a single file to construct the data warehouse by using the Pig engine with Pig Latin Language.

The data warehouse layer will serve as the source layer for the data analysis stage where the business intelligence and business insights are extracted, the data analysis stage using HiveQL language will produce the output as files that represent the answers to various business problems, these files are ready to be visualized to produce the desired business intelligence (BI) output in a form of business dashboard, Figure (5) demonstrates the flowchart for the proposed system's stages.

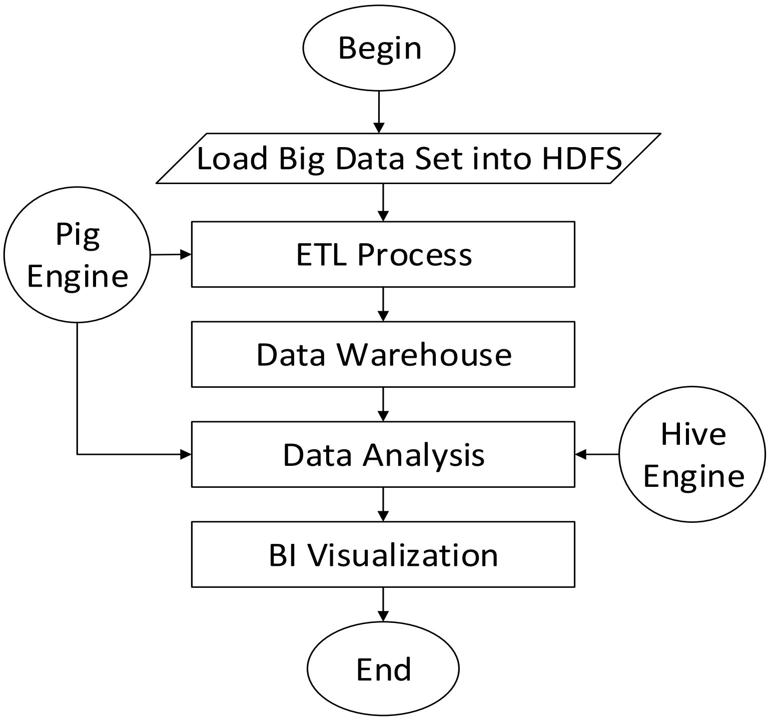

Figure (5), The Proposed System's Flowchart

The Proposed System's Stages

The proposed system consist of four basic stages, each stage leads to the next one, and when the processed data is entered to a particular stage is processed furthermore to became an input to the next stage and every stage contains specific details that will be explained in the following sub sections.

Loading the Big Data Set into HDFS Stage

In real life scenarios the data may resides in a specific local machine inside the company or there may be several sources for the data that flows from different departments. For the proposed system, the obtained big data set is from a free source from the internet as a used case, either ways the first step in the system is to load the data into the system's cluster or more specifically transfer the big data set into HDFS which represent the file system of the proposed system's cluster, in order to enable the system to access and process the big data in the upcoming stages.

Loading the data is performed in Linux shell bash script using Hadoop's specific commands, as the system will divide the data into splits of fixed size through dividing the total size of the input data set by the block size that is previously defined in the configuration files, then the data splits will be transferred sequentially from the local machine into HDFS and each split of data will be replicated inside the cluster depending on the replication factor parameter ***dfs.replication*** that was set in the system's configuration file ***hdfs-site.xml*** to avoid data loss in case of machine failure.

The replication factor was set to 2, the client machine that wants to write the data block (split) into HDFS consults the NameNode and receives a list of two DataNodes locations to copy the data block into them, then the client writes the block directly to the DataNode and the receiving DataNode will replicate the data block to the other DataNode, this process will be repeated until there is no blocks remaining in the input data set, when the two DataNodes receive a data block they will send a ***block received*** report to the NameNode and a ***success*** message to indicate the arrival of the data block and close the session, so the client will be ready to start the process again for the next data block and so on, the NameNode will keep track of each block and knows where to find it in which DataNode, this information will be stored in the NameNode as metadata, the process of loading the data set into HDFS will the same every time new data sets enter the system, this process can be described by the flowchart in Figure (6).

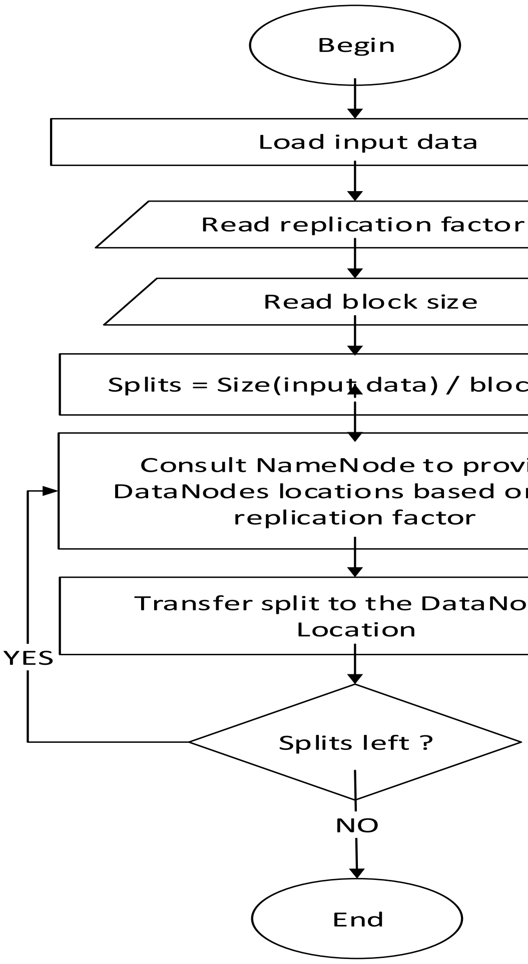
Begin
Load input data
Read replication factor
Read block size
Splits = Size(input data) / block size
Consult NameNode to provide DataNodes locations based on the replication factor
Transfer split to the DataNodes Location
YES
Splits left ?
NO
End

Figure (6), Loading Big Data into HDFS

And the Pseudo code (1) illustrates the stage of loading the input data into the cluster's HDFS as the following parameters are required to be read from Hadoop's configuration files:

- Block size
- Replication factor

Both are read for the ***hdfs-site.xml*** file

```
Input: Number of CSV files resides on local system
Output: CSV files resides on the cluster's HDFS
1: Begin
2: Load input data
3: block size = 64
4: replication factor = 2
5: splits = size (input data) / block size
6: For each split in splits Do
        Consult the NameNode to provide 2DataNodes locations
        Transfer split to these locations
Loop
7: End
```

Pseudo code (1), Loading Big Data into HDFS

Extract Transform and Load (ETL) Stage

After loading the data set into HDFS the next step is to build the data warehouse for the system, the data warehouse will be the main source for the data analysis stage (next stage) where queries will extract the business insights from the data warehouse directly, since the data sets in the proposed system contains multiple CSV files, queries in order to retrieve the desired output they may require join operations to combine the data from two or more files in HDFS, but join queries require a lot of processing and time specially in distributed environments, so building the data warehouse (by de normalizing the data in all the files and extract the one big data warehouse) will eliminate the overhead caused by multiple join queries in runtime, and that will produce a much faster queries; the data warehouse construction process is done using pig Latin language to extract the data that resides in multiple CSV files in HDFS as Figure (7) demonstrates, then load the data with the appropriate join operations based on primary and foreign keys defined previously in these files, finally store the result in a big file that represent the data warehouse that will be the source for the next system stage which is the data analysis stage.

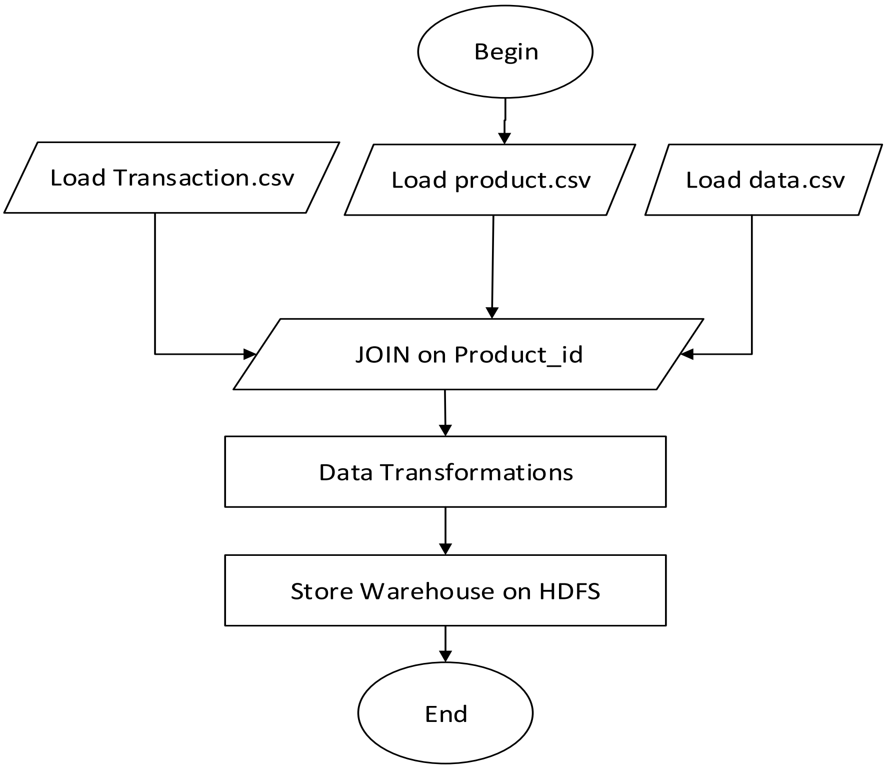

Figure (7), Building the Sample Data Warehouse

and the Pseudo code (2) illustrates the construction of the data warehouse.

Pseudo code (2), Constructing the Data Warehouse

```
Input: Number of CSV files resides on HDFS
Output: One big file resides on HDFS
Begin
Load input Transaction.csv into A
Load input Product.csv into B
Load input Data.csv into C
Join A, B, C on product_id on D
Store D on HDFS
End
```

Data Analysis Stage

The data analysis stage is basically a series of ad hoc queries issued on the data warehouse and to the other files in the sample data set inside the cluster, these queries represent answers to the specific business needs, and it depends on what the business manager or Chief Executive Officer (CEO) want to see or visualize and get insights from the company's big data set, As an example, the CEO may would like to understand the trend of sales in the past years on specific products in a specific stores, this trend would be helpful for him/her to decide any changes in that product, questions has been asked to the system in a form of queries constructed in Pig Latin language and HiveQL and the results are stored in files which are ready for the next stage of visualization and building the system's BI dashboard, many questions had been applied to the cluster and answered each answer will represent a graph placed on the dashboard, as an example of such questions:

Q1: Calculate for each of the 2500 household the total number of purchases over two years and order them in descending order?

The answer to this question using Pig Latin can be described by Pseudo code (3).

Pseudo code (3), Q1 Solved in Pig Data Flow Language

```
1: Begin
2: Load data warehouse file in HDFS into DW
3: Loadhousehold_key from DW
4: Group the result by each household_key in group
5: For each row in the group
Count the rows associated with each household_key
6: Order the result by the result of count in descending order
7: Store the result in HDFS on file q1
8: End
```

BI Visualization Stage

The output of a business intelligence system may come in various forms and types, it may be as reports, charts, tables and other forms, dashboards considered the most useful output to a business intelligence system because dashboards give the overview to the business situation and can be customizable and interactive GUI for the user; Dashboard implementation depends on the data visualization concept as there are many forms of data visualizations, the data analyst can benefit from building the system's BI dashboard.

RESULTS

The results of the proposed system are presented as a BI dashboard, all the implemented visualizations are answers to the questions asked in the data analysis stage, Figure (8) shows the proposed system's results in a form of BI dashboard; Questions that has been answered by the system as a used case are:

1.The total sales amounts for all households in two years visualized as a Scorecard.
2.Total Sales for each Store visualized as a geographical Map.
3.Total Sales for each product category visualized as a Pie chart.
4.Total brands as private or national visualized as a Donut chart.
5.Total purchases for each household over two years visualized as a Line chart.
6.Total Sales values by Store visualized as Tree Map.

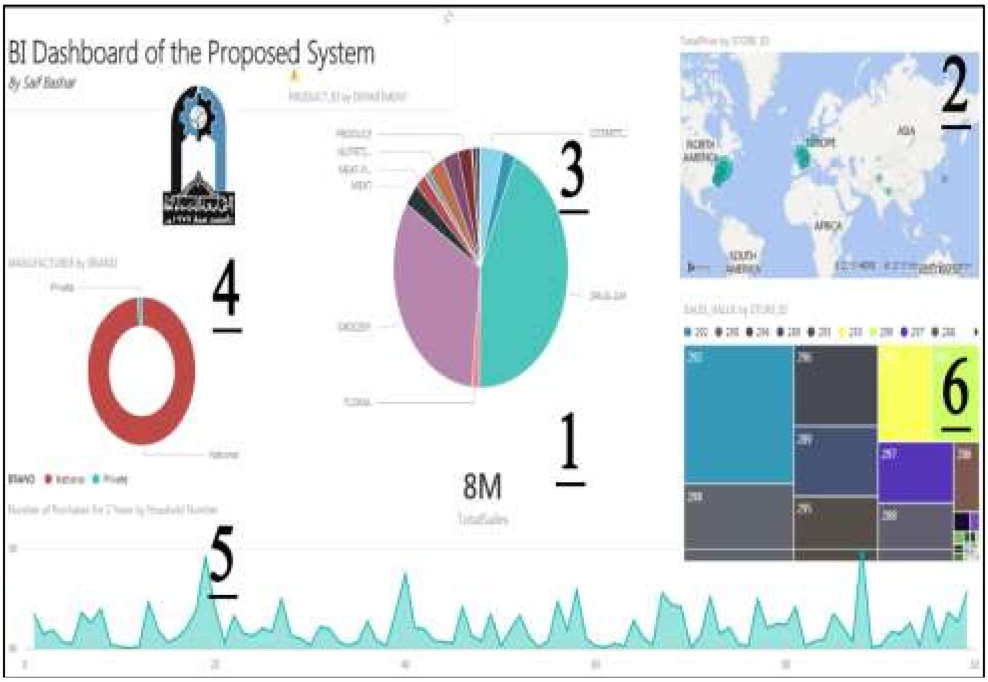

Figure (8), The Proposed System’s BI Dashboard

CONCLUSIONS

The proposed system introduced a way to extract business insights and built a business intelligence dashboardfor a sample of big data set that couldn’t access and analyzed through traditional systems due to its size.

Hadoop through its parallel processing capabilities in a cluster of machines provided environment to brake big data into smaller pieces and provided the processing model to process the data in parallel using the concept of moving computation to the data rather than moving data in the network which require a lot of resources and high network bandwidth, the choice of Hadoop framework is the optimal choice to process big data. The use of high level languages like (Pig and HiveQL)reduced the time and effort in submitting jobs to the cluster to construct the data warehouse and extract the BI insights.

The system’s dashboard with its interactivity provided the overview to the business operations based on the sample data set, the BI dashboard is web based to make it possible to be viewed on various platforms such as web, desktops, tablets and smart phone.

Future Works

Implementing new promising technologies can increase the system performance and decrease the job’s execution time and extract insight in real time environment. This is can be done by implementing new technologies or approaches such as STORM which is an Apache project that supports real time processing on large clusters.

1- Cloud computing technologies should be expanded to include new Hadoop implementations on the cloud, another recommendation is to use this efficient solution for companies, because it allows the companies to simply rent a cluster on the cloud rather than buying the cluster’s hardware and the networking infrastructure, big companies like Amazon’s Elastic MapReduce (EMR) and Microsoft’s HDInsights started to provide Hadoop as a service to perform big data analytics on the cloud.

REFERENCES

[1].Yusuf Bashir, “Next Generation Business Intelligence Software, Areas for Growth & Opportunities for Innovation”, Master Thesis MIT, 2011.

[2]. Saeed Rouhani, Sara Asgari and Sayed Vahid, “Review Study: Business Intelligence Concepts and Approaches”, American Journal of Scientific Research Issue 50, 2012.

[3]. Manoochehr Najmi, Mehran Sepehri and Spideh Hashemi, “The Evaluation of Business Intelligence Maturity Level in Iranian Banking Industry”, Industrial Engineering and Engineering Management (IE&EM), IEEE 17th International Conference, 2010.

[4]. Charmaine Felder, “The Potential Role of Business Intelligence in Church Organizations”, Thesis from Walden University of Technology, College of Management and Technology, 2010.

[5]. Aditya B. Patel, Manashvi Birla and Ushma Nair, “Addressing Big Data Problem Using Hadoop and Map Reduce”, Nirma University International Conference on Engineering, NUiCONE, 2012.

[6]. Robert T. Hans, Ernest Mnkandla, “Modeling Software Engineering Projects As A Business Intelligence Perspective”, IEEE AFRICON, 2013.

[7]. Paulraj Ponniah, “Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals”, John Wiley & Sons, 2001.

[8]. IBM Official Website: www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[9]. John Gantz and David Reinsel, “The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East”, EMC Corporation, 2012.

[10]. Tom White, “Hadoop the Definitive Guide”, O’Reilly, 3rd Edition, 2012.

[11]. Alex Holmes, “Hadoop in Practice”, Manning Publications, 2012.

[12]. Boris Lublinsky, “Professional Hadoop Solutions” Wrox A Wiley Brand, 2013.