International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 23923-23932

http://www.ripublication.com

An Approach For Big Data Analytics on Log Data of Medical Devices In Healthcare

Prajwal Kalmane and Jagadish S Kallimani
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology, Bangalore, India
prajwalk90@gmail.com and jsk_msrit@rediff_mail.com

Abstract

Medical devices (for ex: X-Ray machine, CT- scan etc) generate huge amounts of log every day. These device logs are in the form of text file. Such logs will make up to Terra bytes of data. In this new era of Big Data, such data has to be utilized to gain great business value to the organization. This paper provides an approach to extract useful patterns from huge amount of unstructured log data, find insights of data and present it in a graphical format so that it can be analyzed to bring huge business value to the organization. The approach specifies a design that uses Hadoop infrastructure to filter necessary data from log files, perform Extraction Transformation Loading (ETL) to extract the pattern and load it into relational database and finally visualize the information using reporting tool. Few use cases are portrayed which can bring great business value to the healthcare organization. Challenges that were faced during the implementation are also listed. Finally we conclude by depicting amount of data crunch at each stage of the design.

Keywords: Analytics, Big data, Hadoop, Healthcare, Logs, SQL, QlikView, Use cases, Workflow

Introduction

Every day 2.5 quintillion bytes of data are created— so much that 90% of the data in the world today has been created in the last two years alone [1]. Data in healthcare is also increasing exponentially. Reports say data from the U.S. healthcare system alone reached 150 Exabyte [2]. Five Exabyte (1018 gigabytes) of data would contain all the words ever spoken by human beings on earth. At this rate, big data for U.S. health care will soon reach zettabyte (10^21 bytes) scale and even yottabytes (10^24 bytes) not long after. One can hardly pick up any of today's health care publications without coming across a reference to "Big Data" and its growing impact on the industry. Healthcare systems use sophisticated electronic machines powered by software, to

drive the machines. Usually these systems log the events that occur, in the form of text files periodically. The log data that is obtained from such machines are semi structured in nature. These log files make up terra bytes of data and it is increasing exponentially. Storing and maintaining such data without getting any business information out of it just a loss of business value. These logs provide factual information. They are becoming increasingly valuable due to increased instrumentation of devices resulting into enriched logs data and customer demands from device manufactures to provide better predictive and proactive services [3]. These trends will only gather steam in this era of connected devices. By the definition big data from logs of medical devices refers to electronic data sets so large and complex that are difficult to manage with traditional software and/or hardware, nor can they be easily managed with traditional or common data management tools and methods [4]. By discovering patterns, understanding trends, cost can be reduced, better care can be provided and business can be improved.

This paper provides an approach that can be used by data analyst by extracting patterns, useful information from huge amount of unstructured/semi structured data of log files. This paper also suggests possible information that can be extracted from log data, which are depicted in the form of use cases. The approach uses a design where in flat log files are obtained from log repository. These files are preprocessed in Hadoop using Pig scripts. The results of Pig Scripts are stored in external Hive table. The preprocessed data present from hive table undergo Extraction Transformation and Loading (ETL). Transformation phase of ETL extracts patterns, structures the data, gives defined shape to the data and stores it in a database in the form of table. Then using visualization tools, insights into the data are provided. This paper also analyses the results and depicts the amount of data crunch that is attained at each step of the workflow. This paper also suggests challenges that one can face while extracting useful information from log files.

Related Work

There are few works that has happened in the area of log analysis. Leveraging analytics to reduce down time of medical devices [5], gives solution to monitor all medical devices, predicts and detects issues triages and recommends solution to the extent possible. Leveraging big data analytics to reduce healthcare costs [6], describes two novel applications that leverage big data to detect fraud, abuse, waste, and errors in health insurance claims, thus reducing recurrent losses and facilitating enhanced patient care. The results indicate that claim anomalies detected using these applications help private health insurance funds recover hidden cost overruns that are not detectable using transaction processing systems. Heterogeneous post surgical data analytics for predictive modeling of mortality risks in intensive care units [7], provides the prediction of mortality rates in Intensive Care Units (ICU) using patient-specific healthcare recordings.

The proposed system is more focused over business perspective. It aims at improving the business of health care organizations and to provide better care to the medical devices present in the hospital by providing the insights into the data. Several

use cases are selected and presented in the graphical way to provide the insights. The proposed work uses log data generated by healthcare systems for extracting information.

Design Approach

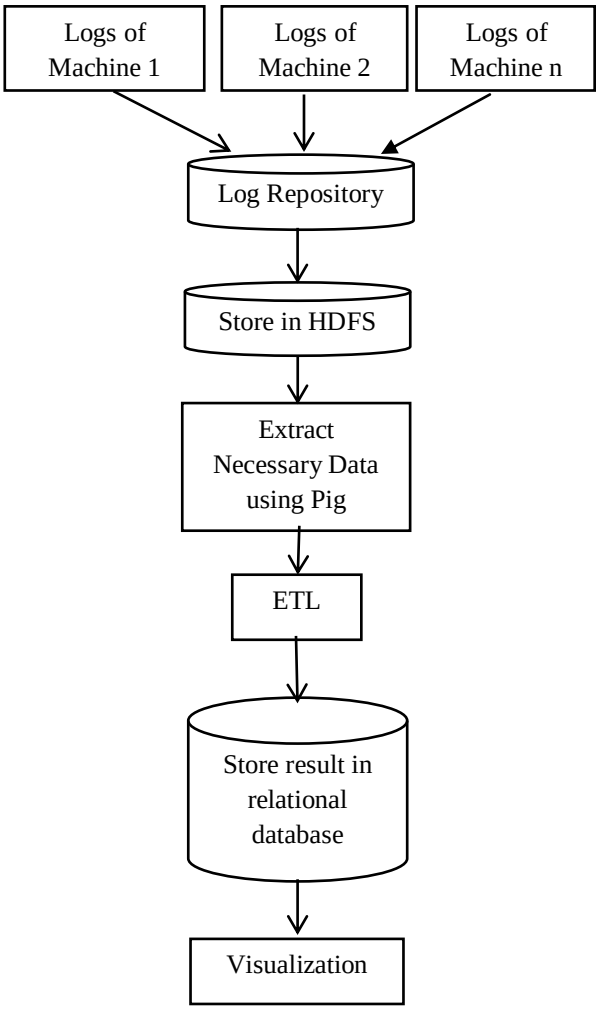

Figure 1: Design approach

Figure 1 shows the design approach. Design approach involves following steps:

- Collect the log files from various machines located in different hospitals and store it in a centralized repository.
- Zip the log files. Make each zip file as 128MB (default block size of Hadoop). Store the zipped files into Hadoop Distributed File System (HDFS).
- Filter those files using Pig scripts to extract data of our interest. Store the results into external hive table.

- Perform ETL on the data from Hive. Extraction phase extracts data from hive table, Transformation phase transforms the data into useful information. Finally load the result of transformation into database in the form of tables.
- Extract the structured data present in database and using some visualization tool perform analytics and present the information in the form of graph so that the user can drill down into the data to find its insights.

Implementation:

Experimental setup:
Following Figure 2 shows the components of our experimental setup.

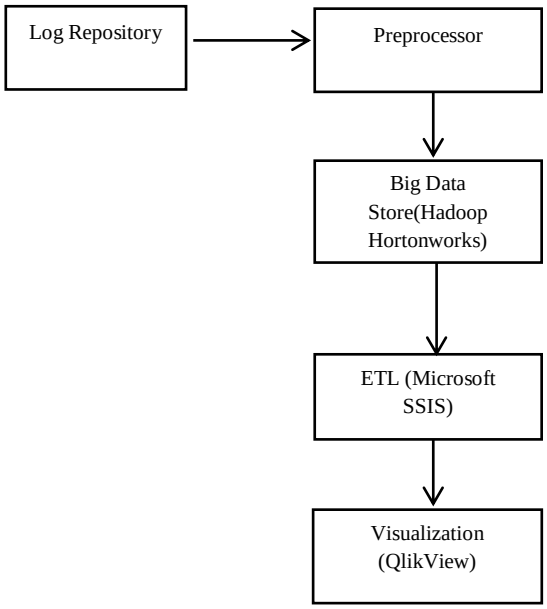

Figure 2: Components of experiment

Log Repository:
The log repository contains the event log data in the form of comma separated value (CSV) files.

Preprocessor:
Hadoop is a file based database that performs best with large CSV files (~128 MB compressed). But the log files considered are relatively small, ~6MB uncompressed. Hence these log files should be concatenated and zipped to make 128 MB. The preprocessor flattens, concatenates, adds metadata and compresses the event log files.

Big Data Store:
Unstructured data can be analyzed very fast with Hadoop Hortonworks. Hadoop is very scalable, with respect to performance and storage. New nodes can be added easily in case there is a lack of resources. Data can be filtered, aggregated and manipulated with MapReduce (MR) jobs. Developing a MR Job is time consuming; therefore tools are available to simplify this process, like PIG and HIVE. PIG scripts can be used to simplify the creation of MapReduce jobs. PIG is a scripting language that generates MapReduce jobs automatically. HIVE is a data warehouse on top of Hadoop that makes it possible to execute HQL (Hadoop Query Language) queries on the MapReduce results. HQL is very similar to SQL. From an external PC it is possible to execute HQL queries via an ODBC connection.

ETL:
ETL is the abbreviation of Extract, Transform, Load; the *extract* stage of an ETL process involves extracting the data from the source systems. The *transformation* stage applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. The *load* phase loads the data into the SQL database. ETL is performed using Microsoft Sql Server Integration Services (SSIS).

Visualization:
Visualizing information that is based on the large amounts of data must be done via the reporting tool. The currently used tool is QlikView (QV). QV can be used locally, but also runs on an access point environment via the web browser.

Detailed Implementation:
The implementation uses logs generated from X-Ray machine. These machine generated logs are in the form of text files containing different commands with respect to each event and are logged on daily basis. Preprocessor concatenates different log files to make each file 128MB of size, which is the default block size of the Hadoop environment. Log files are compressed using Linux gzip utility and are then stored in HDFS.

Pig scripts are executed to extract data of interest and filter the unnecessary data. The filtering is based on the pattern match over a predefined set of commands. Pig scripts generate Map Reduce jobs over Hadoop and the results of Map Reduce jobs are stored into external Hive table that contains necessary fields. The complete workflow in Hadoop is automated by configuring oozie (workflow manager available in Hadoop) script.

Next step in the implementation is Extraction, Transformation and Loading (ETL) using Microsoft SQL Server Integration Services (SSIS). Extraction phase extracts the data from the Hive table and provides as input to transformation phase. Transformation is the major phase of entire workflow where business rules can be incorporated to the data and the data can be turned to meaningful information, which can then be consumed by visualization tool to perform analytics. The transformation

rules are applied to the data based on specific keyword strings. The SSIS has two flows: Control flow and Data Flow.

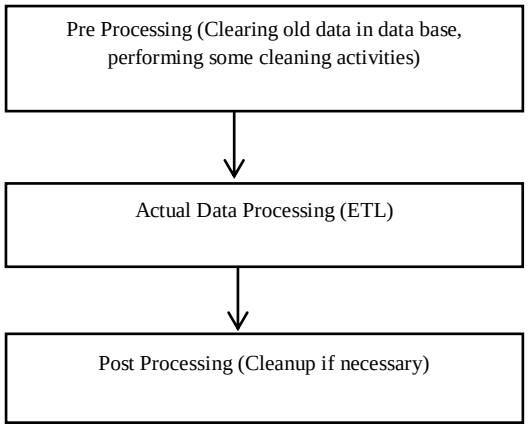

Figure 3: Control Flow in SSIS

A control flow task shown in figure 3 is an SSIS component that performs a high level of operations. It contains 3 components; one is preprocessor in which SQL statements are executed to clear the database. Another is container component containing data flow task which perform actual ETL. Last component is optional which is required again to clean up database or close some opened files. Data flow is shown in figure 4 below.

Data flow contains 3 components. (1) Source which extracts data from Hadoop hive using Open Database Connectivity (ODBC) connection. (2) Transformation component which extracts patterns based on the business rules or use cases and transforms the data into useful information. (3) Load component, into which the transformed data can be loaded into SQL tables with the help of Open Linking and Embedding Database (OLEDB) connection. The transformation phase extracts patterns and loads the information into multiple tables. In this case three related tables are shown containing specific information.

First table is exam details table; which contains high level information regarding each examination taken in the machine. It contains fields like, Examination date, start time, end time, duration, hospital name etc. Second table contains machine related information like, machine start date, stopped date, start time, stop time, on duration, idle time etc.

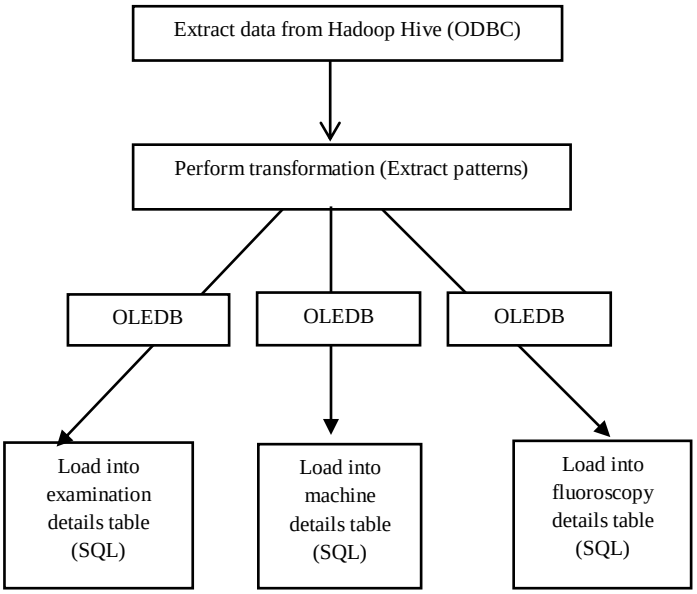

Figure 4: Data Flow in SSIS

Second table contains machine related information like machine start date, stopped date, start time, stop time, on duration, idle time etc.

Another table mentioned in the figure 4 is Fluoroscopy table which is related to Exam details table and contains detailed information regarding each examination. It contains fields like acquisition start time, end time, duration, number of images taken, shutter speed, voltage, current, dose information etc.

The information extracted can be put in the form of use cases which is very useful at the business level. Several use cases can be listed based on the business requirement. Certain use cases identified are put in the following section.

Use cases:

Use cases can be divided into three parts:

A. Exam related Information
B. Machine related information
C. Dose related information

A. Exam Related Information:

- *Time Before First Exam and Time After Last Exam:* This use case helps to identify the time interval between start of the system and the first examination conducted on the system each day. Similarly Time after last exam, tells us the time interval between end of the last exam and the shutdown of the system. This statistics helps to know on an average how long system is on before performing first examination and after conducting last examination in a day. These values can be calculated with the help of exam start time and exam end time.

- *Change over time:* This is the time between successive examinations. This information helps us to know on an average how much time system is on between successive examinations. This indicates how much time is required for preparation of patient for examination.
- *Idle time:* This is the time duration where the system is Idle. That is, it is average time where system was started but not doing any useful job. This includes time system was left on before first and after last exam.
- *Exam duration:* This is the total time taken for each examination. It is time difference between start and end of examinations. This statistics help to know on an average how much time does it take for each examination.

B. Dose Related Information:

- *Air Kerma:* Kerma stands for Kinetic Energy Released per milli Ampere. Air kerma means kerma in a given mass of air. The unit used to measure the quantity of air kerma is the *Gray (Gy).* This quantity is required for calculating dose. For estimating the air kerma (at a given distance, typically 1 meter) specific yield of the X-ray tube (mGy/mAs) is needed. The specific yield depends on the tube voltage and the beam filtration, but also on tube age, construction etc. So the best way is to measure the yield curve of your specific tube, as shown in Figure 5.

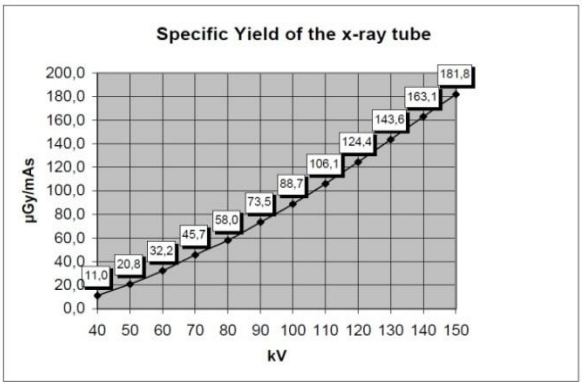

Figure 5: Specific Yield of x-ray tube

Air Kerma is calculated as:

$$\text{Basic Air Kerma for frame} = \text{TubeYield}_{\text{kv,FilterType}} * \text{mA} * \frac{\text{mS}}{1000} \text{----microGy} \quad (1)$$

$$\text{Air Kerma for run} = \left(\frac{\text{Basic Air Kerma}}{1000}\right) * (\text{No. of images}) \text{-----milliGy} \quad (2)$$

- *DAP (Dose Area Product):* DAP is a quantity used in assessing the radiation risk from diagnostic x-ray examinations and interventional procedures. It is defined as the absorbed dose multiplied by the area irradiated, expressed in graysquare centimeters (Gy*cm²) (sometimes mGy*cm² or cGy*cm²). DAP reflects not only the dose within the radiation field but also the area of tissue

irradiated. Therefore, it may be a better indicator of the overall risk of inducing cancer than the dose within the field. DAP can be calculated as follows:

$$\text{Dose Area Product for run} = (\text{Basic Air Kerma for frame}) * (\text{ShutterArea}) * (\text{No. of images})\text{----------microGymm}^2 \quad (3)$$

- *Comparison with newer version of system:* This is to project the values of Air Kerma and Dose Area Product if we used another new system with powerful features. This can be very helpful to customers to understand the difference in dose details that they would get if they used new system.

C. Machine Related Information:

- *System start and shut down time:* This data provides information regarding at what time the system started on a day and at what time system shut down.
- *Is shutdown on same day:* This is to check whether the system was shut down on the same day or it was simply switched on for multiple days. The statistics obtained from these is helpful to maintain the efficiency of the system and increase its performance.

Visualization using Qlik View

The information present in SQL is visualized using QlikView tool. QlikView allows creating application by extracting the data from the database. The extracted data are compressed and custom QlikView Document (QVD) are created. QVD acts as a data model for the applications developed in QlikView. Graphs are drawn in QlikView to represent the aggregated data which help in analysis. Dashboard is created which allows selection of particular hospital and drill down into the data. Once the data is available for users in a visual appealing way, data scientists or business analysts can look at the information, drill down into the root, get the insight of data and hence make business decisions to gain business value.

Result Analysis

The amount of data crunch in each level is shown in the Table 1 below. Final 250 MB of data provides useful information and helps to generate abundant business value to the organization.

Table 1: Amount of Data Crunch At Each Stage

Workflow Stage	**Size of the data**
Log files in HDFS	3 TB
Filtered data i/n HIVE	7 GB
After ETL in SQL	5 GB
In QVD of QlikView	1GB
In final QlikView application	250MB

Challenges

Despite the advantages that can be offered by the application of analytics, a number of barriers can stymie or slow adoption. Firstly, it takes significant amount of time to extract the data, and load it into SQL, as SQL is not a big data infrastructure. Secondly, since we entirely depend on the log files and if the log file does not contain events properly, our results may mislead. Thirdly, validating the result of analysis is very difficult and needs to manually look into each log file to verify whether the obtained results are true.

Conclusion

Big data analytics plays very important role in healthcare organization. Leveraging big data in healthcare helps to improve the business of healthcare industry. The proposed work effectively utilizes log data from the medical devices and generates business information to the healthcare organizations. The use cases specified enables to analyze the machine related information in more structured way by performing statistics on that information. The amount of data crunch indicates significant reduction in the data to be processed at each stage of the design.

References

[1] IBM: Bringing Big Data to Enterprise, http://www.ibm.com/software/in/data/bigdata/

[2] IHT2: Transforming Healthcare through Big Data, http://ihealthtran.com/wordpress/2013/03/iht%C2%B2releases-big-data-research-reportdownload-today/

[3] Dinesh Katiyar, Machine logs analytics – Next Frontier for Data Center Infrastructure Management (DCIM), http://www.glassbeam.com/machine-logs-analytics-next-frontier-for-data-center-infrastructure-management-dcim/

[4] Frost and Sullivan, Drowning in Big Data? Reducing information technology complexities and costs for healthcare organizations, http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technology-complexities-ar.pdf

[5] iGATE: Leveraging Analytics to reduce down time of the medical devices http://www.igate.com/documents/11041/147885/CS_Big_data_Analytics1.pdf/f72ddcb2-2175-4a50-b952-42c4d646d64b

[6] Srinivasan U, Arunasalam B, Leveraging big data analytics to reduce health care costs,2014 IEEE Vol. 15, Issue: 6, 2013, pp: 21-28.

[7] Yun Chen, Hui Yang, Heterogeneous postsurgical data analytics for predictive modeling of mortality risks in intensive care units, 36th IEEE Annual International Conference of the Engineering in Medicine and Biology Society (EMBC) 2014, pp: 4310 – 4314.