Challenges in implementing BI on big data

Palak Anand[1], Rachit Gupta[2] and Abhishek Nigam[3]

**[1] Cybage Software Pvt. Ltd.
Senior Software Engineer
Pune, Maharastra, 411014 India**

**[2] Yardi Software India Pvt. Ltd.
Programmer
Pune, Maharastra, 411014 India**

**[3] Cybage Software Pvt. Ltd.
Senior Software Engineer
Pune, Maharastra, 411014 India**

Introduction

As the amount of data continues to grow exponentially, compounded by the internet, social media, cloud computing and mobile devices, it poses both a challenge and an opportunity for organizations – how to manage and make use of the ever-increasing amount of data being generated. The big question is how organizations can unlock the economic value of big data through the adoption of big data analytics.

Even if an enterprise installed a big data application onto a perfect platform that integrated and connected all the different forms of data, there would be major issues. The truth is, you cannot suddenly analyze the unstructured data coming from heterogeneous sources. The data needs to be transformed in a way that it can be used and integrated with traditional BI system. Also, the unstructured data can't be directly converted into structured format however we can find a way to analyze the unstructured data. We can summarize Big data as, a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And big data may be as important to business – and society – as the Internet has become. So, there is a need to study the challenges that one can face while implementing BI on Big data.

Focusing on Big data challenges

While designing the BI system which uses Big data, we need to consider the big data challenges that would come across.

- **Large Volume**: Too much volume is a storage issue, but too much data is also a massive analysis issue.
- **Broad Variety**: Challenges in analyzing data coming from heterogeneous sources. Data coming from each source need to be processed and analyzed in a different way.
- **Rapidly changing Velocity**: Velocity means both how fast data is being produced and how fast the data must be processed to meet demand.

Big data Analytical Workloads

- **Analysis of data in motion:** The purpose of analyzing data-in-motion is to analyze events as they happen to detect patterns in the data that impact (or are predicted to impact) on costs, revenue, budget, risk, deadlines and customer satisfaction etc. When these occur, the appropriate action can then be taken to minimize the impact or maximize opportunity.

  Stream processing is unique because analysis of data needs to take place before this data is stored in a database or a file system.

- **Exploratory analysis of un-modeled multi-structured data:** The challenge with this type of data is that it can be very large in volume and may contain content in different languages and formats. It may also contain considerable amounts of poor quality data (e.g. spelling errors or abbreviations) and obsolete content. A key requirement for successful text analytics is to 'clean' the content before analysis takes place. However, many companies often have no mechanism to do this.

Content analytics goes beyond text analytics in that it can also handle audio, Video and graphics. Digital asset content e.g. sound and video is more difficult to parse and derive business value from because the content is not text. Deriving insight from this kind of content is more heavily reliant on sophisticated analytic routines and how well the content has been tagged to describe what the content is and what it is about.

- **Complex analysis of structured data**: This type of big data analytical workload may be on structured data taken from a data warehouse or from other data sources (e.g. operational transaction systems) for the specific purpose of doing complex analysis on that data. This may be needed so that power users can mine data to produce predictive models for use in every-day business operations

- **The storage and reprocessing of archived data:**
  - ✓ The need to store data for many years to remain compliant with legislation and/or industry regulations. This includes data associated with business transactions, master data, retired legacy applications, documents, emails and audio files
  - ✓ The need to store archived data on-line for auditing purposes
  - ✓ The need to collect and archive both structured and multi-structured data for the purposes of fending off future liabilities.
  - ✓ The need to manage cost and performance of data warehouses where data volumes continue to grow from increasing transaction volumes, more data sources and from business users demanding lower levels of detail. This is done by archiving older data whose value has diminished

- **Accelerating ETL and analytical processing of un-modeled data to enrich data in a data warehouse or analytical appliance**

Storage options for analytics on Big data

- **Analytical RDBMS**: Analytical RDBMS platforms are relational DBMS systems that typically run on their own special purpose hardware specifically optimized for analytical processing.

- **Hadoop solutions**: The Hadoop stack enables batch analytic applications to use thousands of computer nodes to process petabytes of data stored in a distributed file system.

- **NoSQL DBMSs such as graph DBMSs**: In addition to Hadoop HDFS, HBase, and Hive, there are other NoSQL DBMSs options available as an analytic data store. They include key value stores, document DBMSs, columnar DBMSs, graph databases and XML DBMSs. Some NoSQL databases are not aimed at big data analytics. Others are aimed at analysis of big data or for specific types of analyzes.

Analyzing Unstructured data

A. Analyzing Media Files: When track listening data is submitted to any application, it can undergo a validation and conversion phase, the end result of which will be a number of space-delimited text files containing the user ID, the track ID, the number of times the track was played, the number of times the track was listened to on the radio, and the number of times it was skipped. The table contains sample listening data, which is used in the following examples as input to the Track Statistics program.

UserId	TrackId	Played	Radio	Skip
11115	256	0	1	0
11113	225	1	0	0
11117	267	0	1	1
11115	287	1	1	0

These text files are the initial input provided to the Track Statistics program, which can consist of two jobs that calculate various values from this data and a third job that merges the results.

The Unique Listeners job calculates the total number of unique listeners for a track by counting the first listen by a user and ignoring all other listens by the same user.

The Sum job accumulates the total listens, radio listens, and skips for each track by counting these values for all listens by all users.

The final **"Merge" job** would be responsible for merging the intermediate output of the two other jobs into the final result. The end results of running the program are the following values per track:

- ✓ Number of unique listeners
- ✓ Number of times the track was listened to on the radio
- ✓ Number of times the track was listened to in total
- ✓ Number of times the track was skipped on the radio

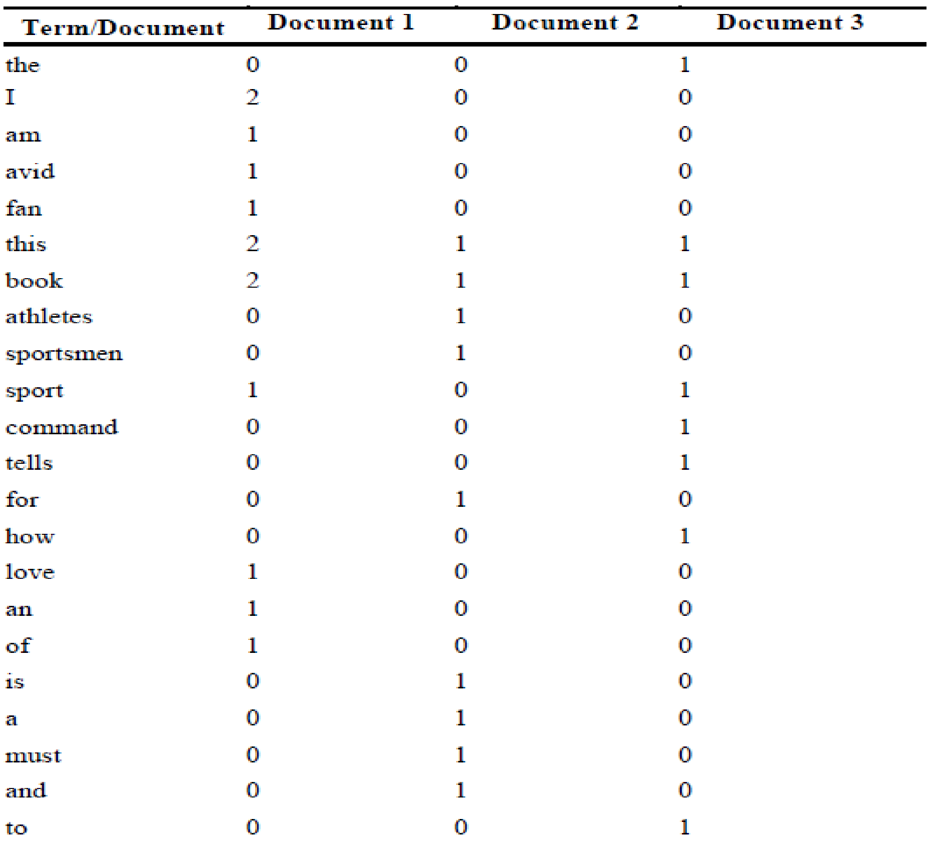
B. Analyzing Unstructured Text data: Most data explosion that is generated by social media network and other sources represents unstructured data that can be difficult to format and evaluate via data analysis. This includes unstructured data such as social media posts, recorded call center interactions between customers and agents, health records, and the bodies of email messages. Let us see how we can analyze unstructured text data. Below are steps that need to be followed during analysis.

- **Crawling, Searching and Indexing:** Extracting and searching the relevant documents from Big data systems. You can conduct exploratory analysis, such as extracting key topics or themes.
- **Content Categorization:** After collecting data from heterogeneous sources, we need to categorize that data. This would pace up our analysis.

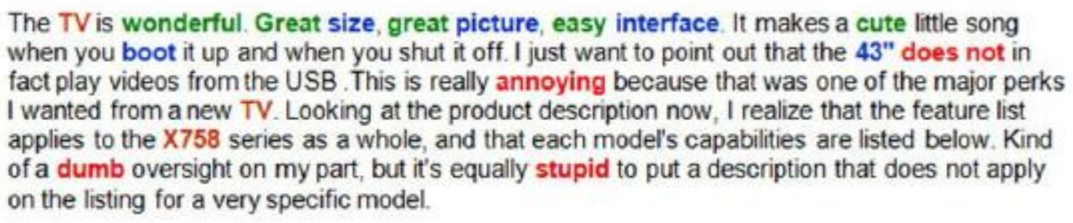
- **Text Mining:** Text Mining is done to process unstructured information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, we can analyze words, clusters of words used in documents, etc., or we could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining project.

- **Text Filtering:** We can find the occurrence of words in the different document and can store them in a form table that would pace up the search.

- **Sentiment Analysis:** Sentiment analysis referred to as various methods of examining and processing data in order to identify a subjective response, usually a general mood or a group's opinions about a particular topic. Data is often derived from social media services and similar user-generated content, such as reviews, comments, and discussion groups.



Term/Document	Document 1	Document 2	Document 3
the	0	0	1
I	2	0	0
am	1	0	0
avid	1	0	0
fan	1	0	0
this	2	1	1
book	2	1	1
athletes	0	1	0
sportsmen	0	1	0
sport	1	0	1
command	0	0	1
tells	0	0	1
for	0	1	0
how	0	0	1
love	1	0	0
an	1	0	0
of	1	0	0
is	0	1	0
a	0	1	0
must	0	1	0
and	0	1	0
to	0	0	1



The TV is wonderful. Great size, great picture, easy interface. It makes a cute little song when you boot it up and when you shut it off. I just want to point out that the 43" does not in fact play videos from the USB .This is really annoying because that was one of the major perks I wanted from a new TV. Looking at the product description now, I realize that the feature list applies to the X758 series as a whole, and that each model's capabilities are listed below. Kind of a dumb oversight on my part, but it's equally stupid to put a description that does not apply on the listing for a very specific model.

Let us consider an example where we have analyzed reviews for TV.

In the above example, the green color shows the positive sentiment, blue color indicates the attribute of the product and red color indicates negative sentiment about the product.

Finding meaning in people's comments is a Big Data analytics challenge

Many social media sites provide access to customer data via public API's; however, combing through data in a variety of formats and extracting comments relevant to the subject of interest is a significant challenge.

After extracting the data and determining the relevance of each comment, assessing the sentiment (positive, negative, or neutral) creates a further challenge. The volume of data, the rapid pace of social opinions, and the time value of information all place scale and latency demands on the process — the value of social data is directly proportional to how quickly a company acts on it.

Sentiment analysis is the most difficult task of all. Humans often have trouble understanding each other, even when speaking face to face. Without facial expressions or vocal clues, textual misunderstanding is common. That is why legal writing, for example, is excessively dense. It is an attempt to avoid misinterpretation. The problem of misinterpretation has challenged technology for years. In the past, enthusiastic experts were convinced of the power and sophistication that high-level math brought to bear on the problem.

Vendor challenges

An integrated BI and big data analytics platform are a different system. There exists build versus buy options from which to choose. We must consider existing systems, use cases, and the experience levels and competence of your staff. Some companies might want to build an entire open source system using nothing more than vanilla Hadoop (the Hadoop Distributed File System [HDFS] and MapReduce), Zookeeper, Solr, Sqoop, Hive, HBase, Nagios, and Cacti, whereas someone else might be looking for more support and try and build a system using proprietary products from other vendors. These products provide functionality to separate structured data and unstructured data and build a graphical user interface (GUI) layer for users, power users, and applications.

Integrating BI and big data analytics are no easy tasks. The goal for any data or analytical system is to make the data useful and available to as many users as possible. Big data appliances are one way to go. An open source Hadoop system is another way. Both require time, patience, and innovation. An open source system is far quicker and less expensive to implement, but you need a staff with that experience. If you are not experienced in working with big data, a big data vendor appliance might be the better choice, though it is more expensive. Remember, not everyone wants to be a software or hardware company. Sometimes, building an integrated BI and big data platform requires a little building and buying to get where you must go.

Conclusion

With growing data volumes, it is essential that real-time information that is of use to the business can be extracted from its IT systems, otherwise the business risks being swamped by a data deluge. Meanwhile, competitors that use data to deliver better insights to decision-makers stand a better chance of thriving through the difficult economy and beyond.

First of all, we need to overcome the bigdata challenges i.e. volume, velocity, and variety. There is a huge amount of data present which is in both structured and unstructured form. Structured data is easy to store and analyze, however, the unstructured data seems to be a big challenge. The unstructured data cannot be converted to structured data, but definitely it can be analyzed using various tools and techniques discussed in the paper.

There are various Hadoop stack projects available in the market using which we can analyze the big-data which require expertise and skills and, on the other hand, there are licensed products from vendors where huge cost are attached with it but offers a great deal of flexibility.

References:

1. Hadoop: The Definitive Guide, 4th Edition – Page 549
2. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

3. http://docs.aws.amazon.com/gettingstarted/latest/emr/getting-started-emr-sentiment-tutorial.html
4. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Palak Anand received her B.tech degree in Computer Science from Uttar Pradesh Technical University. She is employed with Cybage Software Pvt. Ltd. as a Senior Software Engineer. Her field of expertise is web development using Java and Big Data Analysis. She is currently working on a “proof of concept” for Morgan Stanley Bank.

Rachit Gupta has completed his Bachelor of Technology degree from Uttar Pradesh Technical University in 2012. He is a programmer, working with Yardi Software India Pvt. Ltd., Pune. His research interest lies in the area of big data, predictive analysis and machine learning. He has been part Athena health Inc. development team for data analysis.

Abhishek Nigam received his Bachelor of Technology degree from the Uttarakhand Technical University in 2012. He is a Senior Software Engineer, working with Cybage Software Pvt. Ltd., Pune. His research interest lies in the areas of Software Quality, Open-source library development, Cryptography and Business Intelligence. He is currently working on a product feature for US Bank.