European Journal of Science and Theology, October 2014, Vol.10, Suppl.1, 143-148

BIG DATA ANALYSIS

Andrej Trnka*

University of Ss. Cyril and Methodius, Faculty of Mass Media Communication, Nám. J. Herdu 2, 91701 Trnava, Slovak Republic

(Received 16 June 2014, revised 28 July 2014)

Abstract

The aim of this paper is to highlight the ever-increasing volume of data and methods of analysis. Phenomenon that is gaining prominence is called 'Big Data'. The amount of data is increasing over time and obtaining important data can take days. Problems of Big Data is mainly concerned with the non-user data (not inputed by the user), but loaded by various automated means (e.g. data from security cameras).

Keywords: HADOOP, church, volume, velocity, variety

1. Introduction to Big Data

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, we must choose an alternative way to process it [1].

The first definition of Big Data comes from Merv Adrian: "Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population" [2].

Another good definition is given by the McKinsey Global Institute: "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse" [3].

These definitions imply that what qualifies as big data will change over time as technology advances. What was historically big data or what is big data today won't be big data tomorrow. This aspect of the big data definition is that some people find unsettling. The preceding definitions also imply that what constitutes big data can vary by industry, or even organization, if the tools and technologies in place vary greatly in capability [4].

We should not be surprised that companies are tracking and analysing our data.

We rarely hear of theologians talking about what data can tell us about faith needs of parishioners. If big data is being used to guide us in our shopping habits, it could also be used to guide people to a deeper commitment to God,

*E-mail: andrej.trnka@ucm.sk

faith life and community [5]. In some churches that can leverage the analytical insights from data sets in the community and partner them up with the ongoing trends in their congregation, the larger data set can help in crafting new ministries and developing a strategic vision [*Church and "Big" Data*, http://proecclesia.net/2013/10/15/church-big-data/].

Big data is not a single technology but a combination of old and new technologies that helps companies gain actionable insight. Therefore, big data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction. Big data is typically broken down by three characteristics [6]:

- Volume - how much data,
- Velocity - how fast that data is processed,
- Variety - the various types of data.

These characteristics are called the three Vs of Big Data and a number of vendors have added more Vs to their own definitions.

Volume is the first thought that comes with big data: the big part. Some experts consider Petabytes the starting point of big data. As we generate more and more data, we are sure this starting point will keep growing. However, volume in itself is not a perfect criterion of big data, as we feel that the other two Vs have a more direct impact.

Velocity refers to the speed at which the data is being generated or the frequency with which it is delivered. Think of the stream of data coming from the highways' sensors in the Los Angeles area, or the video cameras in some airports that scan and process faces in a crowd. There is also the click stream data of popular e-commerce web sites.

Variety is about all the different data and file types that are available. Just think about the music files in the iTunes store (about 28 million songs and over 30 billion downloads), or the movies in Netflix (over 75,000), the articles in the New York Times web site (more than 13 million starting from 1851), tweets (over 500 million every day), foursquare check-ins with geolocation data (over five million every day), and then you have all the different log files produced by any system that has a computer embedded. When you combine these three Vs, you will start to get a more complete picture of what Big Data is all about [7].

Other authors explain the fourth V as Veracity.

Most of Big Data comes from sources outside our control and therefore suffers from significant correctness or accuracy problems. Veracity represents both the credibility of the data source as well as the suitability of the data for the target audience [8].

LinkedIn, Netflix, Facebook, Twitter, Expedia, national and local political campaigns, and dozens of other organizations are all generating enormous economic, social, and political value [9].

Some examples of Big Data [8, p. 2]:

- social media text,
- cell phone locations,
- channel click information from set-top box,

- web browsing and search,
- product manuals,
- communications network events,
- Call Detail Records (CDRs),
- Radio Frequency Identification (RFID) tags,
- maps,
- traffic patterns,
- weather data,
- mainframe logs.

2. Architecture for Big Data

Many companies already have large amounts of archived data, perhaps in the form of logs, but not the capacity to process it. Assuming that the volumes of data are larger than those conventional relational database infrastructures can cope with, processing options break down broadly into a choice between massively parallel processing architectures – data warehouses or databases such as Greenplum and Apache Hadoop-based solutions. Typically, data warehousing approaches involve predetermined schemas, suiting a regular and slowly evolving dataset. Apache Hadoop, on the other hand, places no conditions on the structure of the data it can process [1, p. 4].

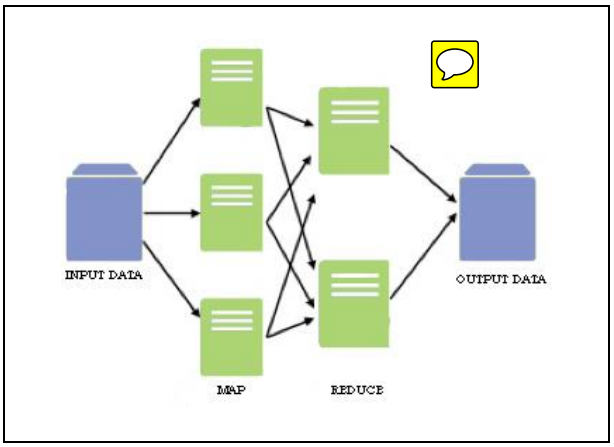

Figure 1. Hadoop with HDFS and MapReduce.

Hadoop is a platform for distributing computing problems across a number of servers. First developed and released as open source by Yahoo, it implements the MapReduce approach pioneered by Google in compiling its search indexes. Hadoop's MapReduce involves distributing a dataset among multiple servers and operating on the data: the 'map' stage. The partial results are then recombined: the 'reduce' stage. To store data, Hadoop utilizes its own distributed filesystem, HDFS, which makes data available to multiple computing nodes. A typical Hadoop usage pattern involves three stages (Figure 1):

- loading data into HDFS,
- MapReduce operations, and
- retrieving results from HDFS.

This process is by nature a batch operation, suited for analytical or non-interactive computing tasks. Because of this, Hadoop is not itself a database or data warehouse solution, but can act as an analytical adjunct to one. One of the most well-known Hadoop users is Facebook, whose model follows this pattern. A MySQL database stores the core data. This is then reflected into Hadoop, where computations occur, such as creating recommendations for you based on your friends' interests. Facebook then transfers the results back into MySQL, for use in pages served to users [1, p. 5].

3. Get results from Big Data

The first question that we need to ask ourselves before we dive into Big Data analysis is what problem are we trying to solve? We may not even be sure of what we are looking for. We have lots of data that we think we can get valuable insight from. And certainly, patterns can emerge from that data before we understand why they are there.

If we think about it, we must to have an idea of what we're interested in. For instance, are we interested in predicting customer behaviour to prevent churn? Do we want to analyse the driving patterns of our customers for insurance premium purposes? Are we interested in looking at our system log data to ultimately predict when problems might occur? Regarding the church, big data techniques can be used for storage of large amounts of data, for example audio or video demand, old church indexes (birth, marriage, death) and their scanned pictures. Big data analysis can be used for advanced search methods in this unstructured data. The kind of high-level problem is going to drive the analytics we decide to use. Alternately, if we are not exactly sure of the business problem we are trying to solve, maybe we need to look at areas in your business that needs improvement. Even an analytics-driven strategy – targeted at the right area – can provide useful results with big data [6, p. 142]. Table 1 shows types of Big Data analysis.

Table 1. Big Data analysis.

Analysis Type	**Description**
Basic analytics for insight	Slicing and dicing of data, reporting, simple visualizations, basic monitoring.
Advanced analytics for insight	More complex analysis such as predictive modeling and other pattern-matching techniques.
Operationalized analytics	Analytics become part of the business process.
Monetized analytics	Analytics are utilized to directly drive revenue.

Basic analytics can be used to explore your data, if you're not sure what you have, but you think something is of value. This might include simple

visualizations or simple statistics. Basic analysis is often used when you have large amounts of disparate data.

Advanced analytics provides algorithms for complex analysis of either structured or unstructured data. It includes sophisticated statistical models, machine learning, neural networks, text analytics and other advanced data-mining techniques. Among its many use cases, advanced analytics can be deployed to find patterns in data, prediction, forecasting, and complex event processing.

When we operationalize analytics, we make them part of a business process. For example, statisticians at an insurance company might build a model that predicts the likelihood of a claim being fraudulent. The model, along with some decision rules, could be included in the company's claims-processing system to flag claims with a high probability of fraud. These claims would be sent to an investigation unit for further review. In other cases, the model itself might not be as apparent to the end user. For example, a model could be built to predict customers who are good targets for upselling when they call into a call centres. The call centres agent, while on the phone with the customer, would receive a message on specific additional products to sell to this customer. The agent might not even know that a predictive model was working behind the scenes to make this recommendation.

Monetizing analytics can be used to optimize a business in order to create better decisions and drive bottom- and top-line revenue. However, big data analytics can also be used to derive revenue above and beyond the insights it provides just for one own department or company. We might be able to assemble a unique data set that is valuable to other companies, as well. For example, credit card providers take the data they assemble to offer value-added analytics products. Likewise, with financial institutions. Telecommunications companies are beginning to sell location-based insights to retailers. The idea is that various sources of data, such as billing data, location data, text messaging data, or web browsing data can be used together or separately to make inferences about customer behaviour patterns that retailers would find useful. As a regulated industry, they must do so in compliance with legislation and privacy policies [6, p. 143].

4. Big Data implementation

There are several ways to store larger amounts of data [10]. There is no singular method to deploy a business intelligence solution to answer unique company questions, but there is an approach to take advantage of Big Data which minimizes risk and increases the likelihood of a successful outcome.

Big Data projects are difficult and need know-how and experience to be successful. The implementation method of Big Data consists of 8 (or 9) steps [The Server Labs, *Implementing Big Data*, http://www.theserverlabs.com/big-data/implementing-big-data.html]:

1. begin with stakeholders and consider culture,

2. find data stewards,
3. set clear goals,
4. create the plan,
5. select the right strategy and tools,
6. establish metrics,
7. deploy the technology,
8. make big data little [CRMSearch. The Business Case for Big Data, http://www.crmsearch.com/big-data-implementation.php],
9. design for Continuous Process Improvement (CPI).

5. Conclusions

Eighty percent of the world's data is unstructured, and most businesses do not even attempt to use this data to their advantage. The trend of Big Data is generating new opportunities and new challenges for businesses across industries.

The churches have a lot of unstructured data. The data should be stored a analysed by Big Data techniques and method. This approach is helpful for better analyse.

Hadoop is scalable platform for ingesting Big Data and preparing it for analysis. Using Hadoop in Big Data can reduce time to analysis by hours or even days.

References

[1] ***, *Big Data Now*, O'Reilly Media, Sebastopol, 2012, 3.
[2] A. Merv, Teradata, **11(1)** (2011) 1, online at http://www.teradatamagazine.com/v11n01/Features/Big-Data.
[3] ***, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, New York, 2011, 1.
[4] B. Franks, *Taming the Big Data Tidal Wave*, New Jersey, John Wiley & Sons, New York, 2012, 336.
[5] M.D. Gutzler, Dialog: A Journal of Theology, **53(1)** (2014) 23-29.
[6] J. Hurwitz, A. Nugent, F. Halper and M. Kaufman, *Big Data for Dummies*, John Wiley & Sons, New York, 2013, 15.
[7] P. Zadrozny and R. Kodali, *Big Data Analytics Using Splun*, Apress, New York, 2013, 353.
[8] A. Sathi, *Big Data Analytics: Disruptive Technologies for Changing the Game*, MC Press, Boise, 2012, 4.
[9] K. Davis and D. Patterson, *Ethics of Big Data*, O'Reilly Media, city California, 2012, 1.
[10] R. Halenar, Appl. Mech. Mater., **229-231** (2012) 2125-2129.