Big Data: A New Era for Research

Sunil K Punjabi,
Assistant Professor, SIES-GST, Nerul, Navi Mumbai

Suvarna Kendre,
Assistant Professor, SIES-GST, Nerul, Navi Mumbai

Pranita Mahajan,
Assistant Professor, SIES-GST, Nerul, Navi Mumbai

ABSTRACT

Big data, a new way of managing and interacting with the massive data sets collected and stored by humans. The problem with the massive data collection and distribution systems is to manage this big data as large amount of data that is gathered from various domains of all sizes and types. Most of the captured data clutters lots of storage space because of which it has become a concern of individuals as awareness grows of breadth and depth of personal information being amazed in big data collection. Big data is a concern rather than precise term. In this paper we have discussed big data definitions with various aspects. Then followed by few case studies where in big data is being used. Smart-mall case study is discussed in detail in which customer behavior is analyzed to provide valuable feedback. Apart from that we have discussed issues such as fraud detections, loss of customers, customer behavior prediction etc.

Keywords:

Big data, Data Streams.

1. INTRODUCTION

Big data is more a concept than a precise term. Some apply the "Big data" label only to petabyte-scale data collections. (> one million GB). For others, a Big data collection may house only a few dozen terabytes of data. More often, however, Big Data is defined situationally rather than by size. Specifically a data collection is considered "Big Data" when it is so large and an organization cannot effectively or affordably manage or exploit it [1].

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data is popular term used to describe the exponential growth, availability and use of information, both structured and unstructured.

2. Definition of Big Data in various aspects:

2.1 The Original Big data

Big Data as the three Vs: Volume, Velocity, and Variety. This is the most venerable and well-known definition, first coined by Doug Laney of Gartner over twelve years ago. Since then, many others have tried to take it to 11 with additional Vs including Validity, Veracity, Value, and Visibility[5].

2.2 Big Data as Technology

Why did a 12-year old term suddenly zoom into the spotlight? It wasn't simply because we do indeed now have a lot more volume, velocity, and variety than a decade ago. Instead, it was fueled by new technology, and in particular the fast rise of open source technologies such as Hadoop and other NoSQL ways of storing and manipulating data[2].

The users of these new tools needed a term that differentiated them from previous technologies, and–somehow ended up settling on the woefully inadequate term Big Data.

2.3. Big data as Data Distinctions

The problem with big-data-as-technology is that (a) it's vague enough that every vendor in the industry jumped in to claim it for themselves and (b) everybody 'knew' that they were supposed to elevate the debate and talk about something more business and useful.

Here are two good attempts to help organizations understand why Big Data now is different from mere big data in the past:

- Transactions, Interactions, and Observations. Transactions make up the majority of what we have collected, stored and analyzed in the past. Interactions are data that comes from things like people clicking on web pages. Observations are data collected automatically.

- Process- Mediated Data, Human-Sourced Information, and Machine-Generated Data. This is brought to us by Barry Devlin, who co-wrote the first paper on data warehousing. It is basically the same as the above, but with clearer names.

2.4 Big Data as Signals

This is another business approach that divides the world by intent and timing rather than the type of data, courtesy of SAP's Steve Lucas. The 'old world' is about transactions, and by the time these transactions are recorded, it's too late to do anything about them: companies are constantly 'managing out of the rear-view mirror'. In the 'new world,' companies can instead use new 'signal' data to anticipate what's going to happen, and intervene to improve the situation.

Examples include tracking brand sentiment on social media (if your 'likes' fall off a cliff, your sales will surely follow) and predictive maintenance (complex algorithms determine when you need to replace an aircraft part, before the plane gets expensively stuck on the runway).

2.5 Big Data as Opportunity

Big data analysis was previously ignored because of technology limitations. Big data can be used as an opportunity with which analysis of large volume of data can be managed.

2.6 Big Data as Metaphor

In his wonderful book The Human Face of Big Data, journalist Rick Smolan says big data is "the process of helping the planet grow a nervous system, one in which we are just another, human, type of sensor."

2.7 Big Data as new term for old stuff

This is the laziest and most cynical use of the term, where projects that were possible using previous technology, and would have been called BI or analytics in the past have suddenly been rebaptized in a fairly blatant attempt to jump on the big data bandwagon. And finally, one bonus, fairly useless definition of big data.

The bottom line: whatever the disagreements over the definition, everybody agrees on one thing: big data is a big deal, and will lead to huge new opportunities in the coming years.

3. Digital data as the new data revolution

In 2012, approximately forty years after the beginning of the information era, all eyes are now on its basis: digital data. This may not seem very exciting, but the influx of various data types, plus the speed with which the trend will continue, probably into infinity, is certainly striking. Data, data and more data: we are at the centre of an expanding data universe, full of undiscovered connections. This is not abstract and general, but rather specific and concrete, as each new insight may be the entrance to a gold mine. This data explosion is so simple and fundamental that Joe Hellerstein of Berkeley University[6] speaks of 'a new industrial revolution': a revolution on the basis of digital data that form the engine of completely new business-operational and societal opportunities.

At present we are living in at least three periods with digital data as their basis: the information era, the social era, and the Big Data era. The explosive growth of data genuinely comes from all corners: from business transactions, mobile devices, sensors, social and traditional media, H-D video, cloud computing, stock-and-share markets, Web-clicks, etc., etc. All data is generated in the interaction between people, machines, applications and combinations of these.

4. Big Data case studies

4.1 Big data in Telecom: Loss of Clients

Until recently we were compelled to take random samples and analyze them. But how do we sample a network or a collection of sub-networks? If a telecom provider wishes to have insight into the circumstances under which a sub-network of friends and acquaintances suddenly switches to a rival company (it "churns"), we are probably dealing with a total of more than 10 million existing and recent subscribers, with information on their habits, their expenditures on services, and who their friends are: in other words, the number of times the phone is used for calls or sms messages, for example. We are dealing with tipping points: a part of the sub-network churns and the rest follow after a (short) time. In itself, this is rather predictable: if colleagues or friends have switched and are better off or cheaper out under a rival, then there is a social and economic stimulus to switch as well. A provider will, of course, attempt to prevent this situation arising and must take a hard look at all the data. For example, if a random sample is taken from a million clients, the circles of friends that formed the basis of the switch can no longer be seen as a unit, and therefore in this case the basis for accurate prediction crumbles. Therefore, sampling is not the appropriate method here. In order to obtain a good view of the tipping points we must examine all the data in their proper context and coherence. Then, on the basis of developing patterns, we can anticipate their churn at an early stage and apply retention actions and programs[7].

4.2 Big data for Detection of fraud

Another area for which we require a complete dataset is fraud detection. The signal is so small that it is impossible to work with random samples until the signal has been identified. Accordingly, all data must be analyzed in this field as well. It can justifiably be referred to as an evident case of Big data when the possibility of 'collusion is being examined: illegal collaboration that is directed toward impeding others as much as possible and of sabotaging them, as occurred in the casino world. Chum and fraud detection are examples of application possibilities of Big data Analytics[7].

4.3 Big data in Hotel Industry

Fast food restaurants generate a greater percentage of repeat business than any other section of the retail industry. Gift and loyalty programs help them maintain customer relationships, but payment card interchange fees greatly impact profit margins. One U.S.-based fast food convenience chain wanted to link loyalty, gift and prepayment cards to customer mobile phones to reduce costs while simultaneously increasing customer loyalty. our task, although simply defined, was daunting: find a way to unify retail outlets dispersed over a large geographic area conducting large numbers of small transactions so that mobile phones could be used for payment, customer appreciation and communication[14].

Challenges: Consultant's first challenge was to ensure that the new mobile solution reliably gave a customer access to personal loyalty, gift and prepaid accounts at any of the chain's fast food outlets at any time. Furthermore, payment management needed sufficient safeguards to prevent large-scale fraud. Suggestion was to use the food chain's existing point-of-sale infrastructure to interconnect outlets with customer payment account information. As a result, after a simple web or in-store sign up process, customers could have immediate access to their accounts at all locations within the chain, regardless of their preferred wireless network. Real-time payment processing and gift transactions became seamless to outlet employees and customers.

Results: From a revenue standpoint, prepaid accounts with larger payment card transactions resulted in savings of about $0.05 per individual small transaction. Perhaps more importantly, big data technology enabled extraction of useful information on customer behavior, habits and preferences, thereby improving the customer experience. Additionally, mobile phone features such as caller ID and text messaging provided the convenience chain with previously untapped customer communication avenues.

4.4 Big data in Retailer's Cloud computing Performance

Cloud computing options offer e-commerce retailers multiple benefits, ranging from platform scalability (particularly during seasonal shopping spikes or specialized sales) to significant infrastructure cost savings, but only if the cloud platform has a reliable notification system in place for consumer and service issues. A global retailer approached Alacer with a problem: how could it improve its cloud platform's incident responsiveness, which directly impacted the consumer experience and SLAs in place with multiple advertisers? The answer involved big data and using the retailer's existing information to not only respond to, but predict, issues impacting the company's overall revenues[14].

Challenges: All successful cloud platform monitoring programs have two aspects – reactive and proactive. To create the reactive portion of this solution, they needed to utilize big data to design a notification system that would deliver alerts in real time with a high level of fidelity. This required an understanding of the impact each anomaly would have on the ecosystem, as well as a more granular understanding of each specific issue that could impact in-place service level agreements (SLAs). Then, to move the client into a more proactive stance, they added options to the monitoring system that would identify and address problems diluting the end-user experience, thereby boosting customer satisfaction.

Results: By using the cloud with monitoring platform in place, the retailer reduced incident response time from one hour to mere seconds, positively impacting millions of users worldwide. This, in turn, enabled the company to off load the need for labor costs for system support personnel. Hundreds of thousands of dollars in SLA penalties from advertisers were also avoided

5. Big Data in Customer Behavior Analysis:

The Smartmall example below brief how big data can reduce cost of communication and gain customer satisfaction. The idea behind Smartmall is often referred to as multichannel customer interaction, meaning "how can one interact with customers that are in his brick-and-mortar store via their smartphones"? Rather than requiring customers to whip out their smartphone to browse prices on the internet, they would like to drive their behaviour proactively.

The goals of Smartmall are straightforward:

- Increase store traffic within the mall.
- Increase revenue per visit and per transaction.
- Reduce the non-buy percentage.

A picture speaks a thousand words, so Figure 1 shows both the real-time decision-making infrastructure and the batch data processing and model generation (analytics) infrastructure[9].

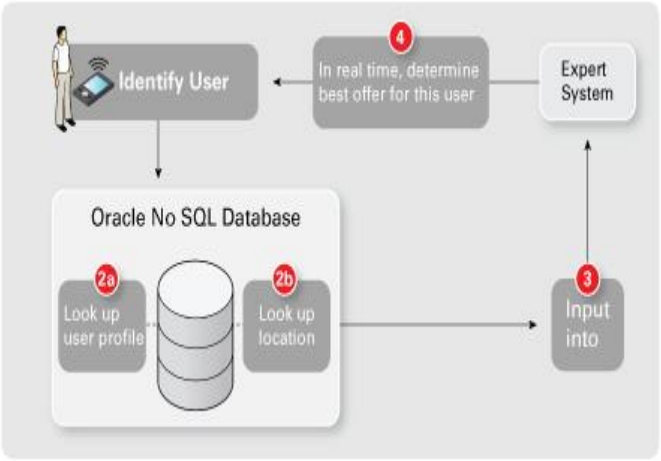

Figure 1: Example Infrastructure

The first—and, arguably, most important—step and the most important piece of data is the identification of a customer. Step 1, in this case, is the fact that a user with a smartphone walks into a mall. By identifying this, we trigger the lookups in step 2a and step 2b in a user-profile database.

We will discuss this a little more later but, in general, this is a database leveraging an indexed structure to do fast and efficient lookups. Once we find the actual customer, we feed the profile of this customer into our real-time expert system (step 3).

The models in the expert system (custom-built or COTS software) evaluate the offers and the profile and determine what action to take (for example, send a coupon). All this happens in real time, keeping in mind that Websites do this in milliseconds and our smart mall would probably be OK doing it in a second or so.

To build accurate models—and this where many of the typical big data buzz words come in—we add a batch-oriented massive-processing farm into the picture. The lower half of Figure 2 shows how we leverage a set of components that includes Apache Hadoop and the Apache Hadoop Distributed File System (HDFS) to create a model of buying behavior. Traditionally, we would leverage a database (or data warehouse [DW]) for this. We still do, but we now leverage an infrastructure before the database/data warehouse to go after more data and to continuously re-evaluate all the data.

A word on the data sources. One key element is point-of-sale (POS) data (in the relational database), which you want to link to customer information (either from your Web store, from cell phones, or from loyalty cards). The NoSQL

database with customer profiles in Figure 1 and Figure 2 show the Web store element. It is very important to make sure this multichannel data is integrated with our Web browsing, purchasing, searching, and social media data.

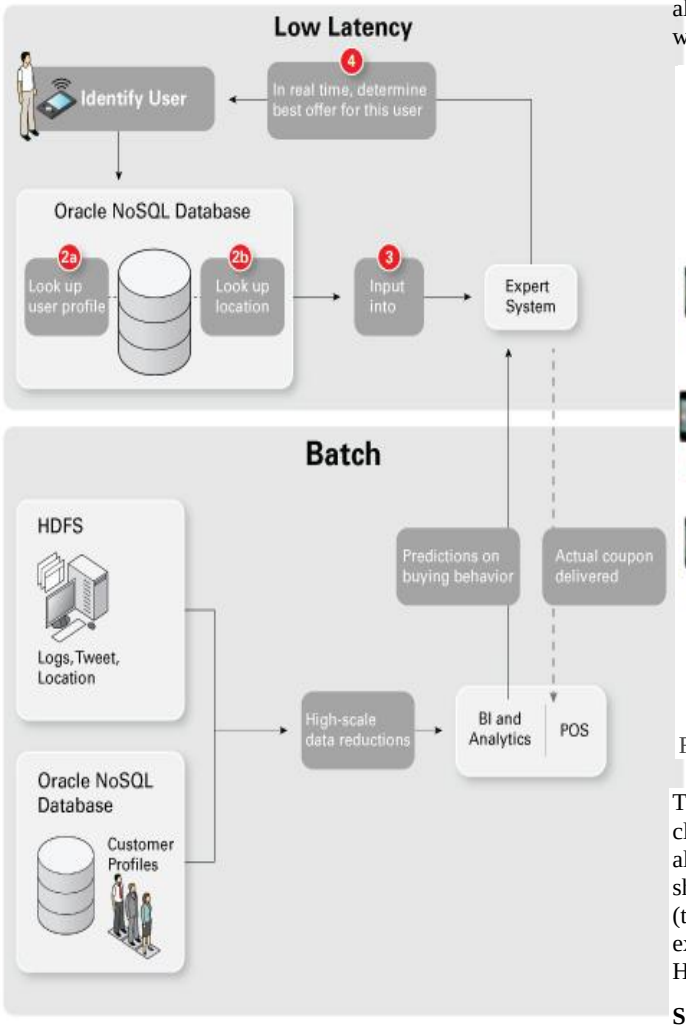

Figure 2: Creating a model of buying behaviour

Once the data linking and data integration is done, we can figure out the behaviour of an individual. In essence, big data allows microsegmentation at the person level—in effect, for every one of your millions of customers!

The final goal of all this is to build a highly accurate model that is placed within the real-time decision engine. The goal of the model is directly linked to the business goals mentioned earlier. In other words, how can we send a customer a coupon while the customer is in the mall that gets the customer to go to your store and spend money?

5.1 Detailed Data Flows and Product Ideas

Now, how do we implement this with real products and how does our data flow within this ecosystem? The answer is shown in the following sections.

Step 1: Collect Data

To look up data, collect it, and make decisions on it, we need to implement a system that is distributed. Because the devices essentially keep sending data, we need to be able to load the data (collect or acquire it) without much delay. That is done in the collection points shown in Figure 3. That is also the place to evaluate the data for real-time decisions. We will come back to the collection points later.

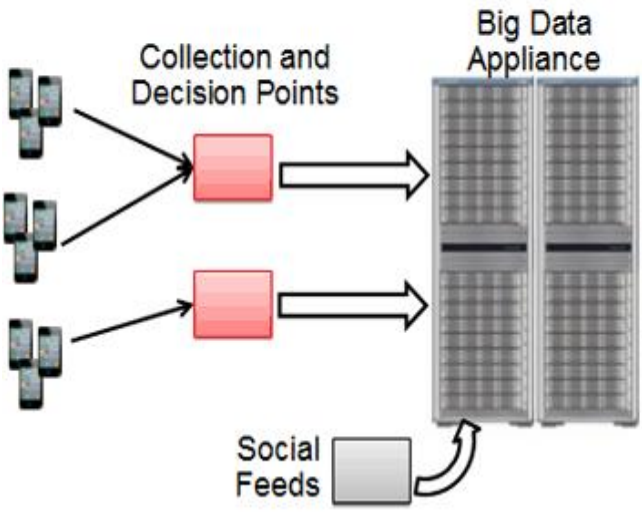

Figure 3: Collection Points

The data from the collection points flows into the Hadoop cluster, which, in our case, is a big data appliance. we would also feed other data into this appliance. The social feeds shown in Figure 3 would come from a data aggregator (typically a company) that sorts out relevant hash tags, for example. Then we use Flume or Scribe to load the data into Hadoop.

Step 2: Collate and Move the Data

The next step is to add data (social feeds, user profiles, and any other data required to make the results relevant to analysis) and to start collating, interpreting, and understanding the data.

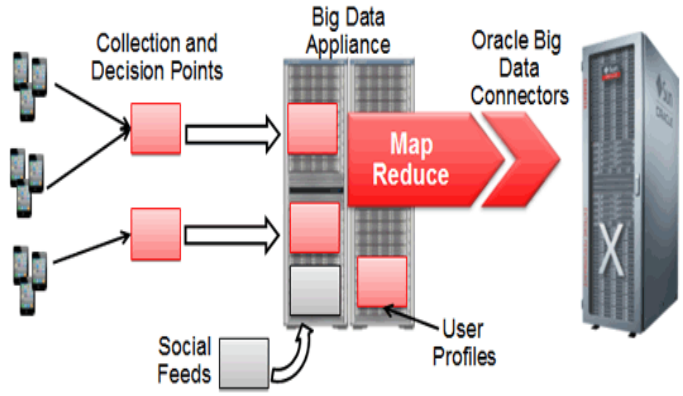

Figure 4: Collating and Interpreting the data

For instance, add user profiles to the social feeds and add the location data to build a comprehensive understanding of an individual user and the patterns associated with this user. Typically, this is done using Apache Hadoop MapReduce. The user profiles are batch-loaded from the Oracle NoSQL Database via a Hadoop InputFormat interface and, thus, added to the MapReduce data sets.

To combine all this with the POS data, customer relationship management (CRM) data, and all sorts of other transactional data, you would use Oracle Big Data Connectors to efficiently move the reduced data into the Oracle Database. Then we have a comprehensive view of the data that we can go after, either by using Oracle Exalytics or business intelligence (BI) tools or—and this is the interesting piece—via things such as data mining.

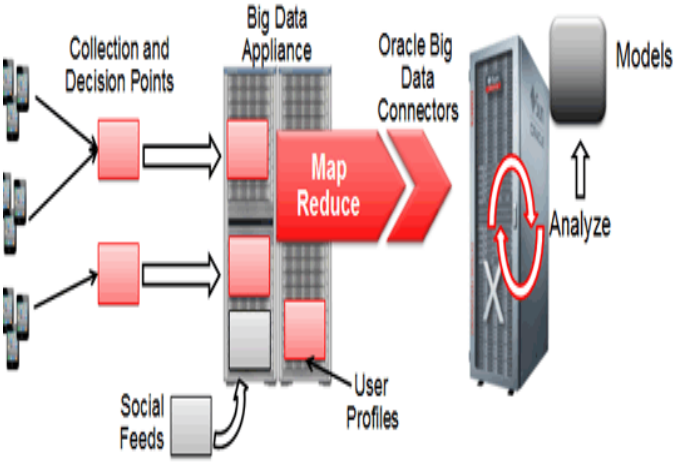

Figure 5: Moving the reduced data

Step 3: Analyze the Data

That last phase here called "analyze"— creates data mining models and statistical models that are used to produce the right coupons. These models are the real crown jewels, because they allow you to make decisions in real time based on very accurate models. The models go into the collection and decision points to act on real-time data, as shown in Figure 5.

In Figure 6, we see the gray model being utilized in the Expert Engine. That model describes and predicts the behavior of an individual customer and, based on those predictions, determines what action to take.

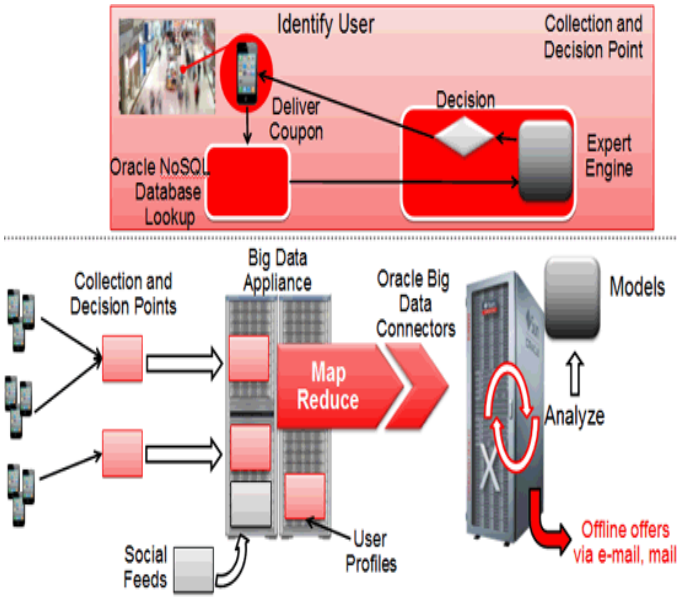

Figure 6: Analyzing the Data

Big data allows us to leverage tremendous amounts of data and processing resources to arrive at accurate models. It also allows us to determine all sorts of things that we were not expecting, which creates more-accurate models and also new ideas, new business, and so on.

We can implement the entire solution shown here using the Oracle Big Data Appliance on Oracle technology. Then we'll just need to find a few people who understand the programming models to create those crown jewels.

CONCLUSION:

The description above is an end-to-end look at "big data" and real-time decisions. Big data helps proactive managing customer experiences. Big data has provided economical ways of communication with end user satisfaction. It is difficult to manage large data using conventional tools. Various case studies discussed in this paper shows how big

data is a new data evolution for business and financial sectors. Finally, Smartmall case study shows how Big data appliance helps in doing so effectively and affordable managing the data based on customer behaviour.

ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the big data case studies. We sincerely thanks our colligues and friends. Special thanks to the review team for their valuable time and their valuable efforts and comments which has helped us in improving our work.

REFERENCES:

[1] P. Barnaghi, W. Wang, C. Henson, and K. Taylor, Semantics for the Internet of Things: Early progress and back to the future, International Journal on Semantic web and Information Systems, vol. 8, no. 1, pp. 1-21, 2012.

[2] P. P. Talukdar, D. Wijaya and T. Mitchell. Coupled Temporal Scoping of relational facts. In Proceedings of the fifth ACM International Conference on Web Search and Data Mining (WSDM) , Seattle, Washington, USA, February 2012.

[3] P. Boldi, M. Rosa, and S. Vigna. HyperANF: approximating the neighbourhood function of very large graph on a budget. In Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011, pages 625 – 634, 2011.

[4] P. Gilbert, L. P. Cox, J. Jung, and D. Wetherall. Toward trustworthy mobile sensing. In Proc. Workshop on Mobile Computing System and Applications, 2010.

[5] 5. A and E. Brynjolfsson, Big Data: The Management Revolution. Harvard Business Review, 2012 90(10): p 59-68.

[6] C. Bokermann and H. Blom. The Streams Framework. Technical Report 5, TU Dortmund University 12, 2012.

[7] Big Data vint research report : Creating Clarity with Big Data [Online].

[8] Visualizing the Petabyte Age, Nov 2010. [Online], Available: http://www.techwhizz.com/visualizing-petabyte-age-inforgraph.

[9] http://www.oracle.com/technetwork.

[10] Apache Hadoop, http://hadoop.apache.org.

[11] Laura Wilber "A Practical Guide to Big Data: Opportunities, Challenges & Tools" 2012 Dassult Systems.

[12] A. Rajaraman, J. Leskovec, and J. D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2010.

[13] J. Gama, Knowledge discovery from data streams Chapman & Hall/CRC 2010.

[14] http://www.alacergroup.com.