A Review on Big Data Mining

Ravita[1] , Joni Birla[2]
[1,2] Department Of Computer Science &Engineering,
Ganga Institute Of Technology And Management,
Kablana, Jhajjar, Haryana, India

***Abstract*--While "big data" has become a highlighted buzzword since last year, "big data mining", i.e., mining from big data, has almost immediately followed up as an emerging, interrelated research area. This paper provides an overview of big data mining and discusses the related challenges and the new opportunities. The discussion includes a review of state-of-the-art frameworks and platforms for processing and managing big data as well as the efforts expected on big data mining. We address broad issues related to big data and/or big data mining, and point out opportunities and research topics as they shall duly flesh out. We hope our effort will help reshape the subject area of today's data mining technology toward solving tomorrow's bigger challenges emerging in accordance with big data.**

I. INTRODUCTION

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.In a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Additionally, a new V "Veracity" is added by some organizations to describe

If Gartner's definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between big data and Business Intelligence, regarding data and their use.

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big data uses inductive statistics and concepts from nonlinear system identification [21] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships, dependencies and perform predictions of outcomes and behaviors.

Big data can also be defined as "Big data is a large volume unstructured data which cannot be handled by standard database management systems like DBMS, RDBMS or ORDBMS".

II. UNDERSTANDING A BIG DATA IMPLEMENTATION AND ITS COMPONENTS

It often get asked about big data, and more often than not we seem to be talking at different levels of abstraction and understanding. Words like real time show up, words like advanced analytics show up and we are instantly talking about products. The latter is typically not a good idea. So let's try to step back and go look at what big data means from a use case perspective and how we then map this use case into a usable, high-level infrastructure picture. As we walk through this all you will – hopefully – start to see a pattern and start to understand how words like real time and analytics fit…

III. THE USE CASE IN BUSINESS TERMS

Rather then inventing something from scratch I've looked at the keynote use case describing Smart Mall (you can see a nice animation and explanation of smart mall in this video).The idea behind this is often referred to as "multi-channel customer interaction", meaning as much as "how can I interact with customers that are in my brick and mortar store via their phone". Rather than having each customer pop out there smart phone to go browse prices on the internet, I would like to drive their behavior pro-actively.

The goals of smart mall are straight forward of course:

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

IV. WHAT DO I NEED?

In terms of technologies you would be looking at:

- Smart Devices with location information tied to an individual
- Data collection / decision points for real-time interactions and analytics
- Storage and Processing facilities for batch oriented analytics

In terms of data sets you would want to have at least:

- Customer profiles tied to an individual linked to their identifying device (phone, loyalty card etc.)
- A very fine grained customer segmentation
- Tied to detailed buying behavior
- Tied to elements like coupon usage, preferred products and other product recommendation like data sets

V. HIGH-LEVEL COMPONENTS

A picture speaks a thousand words, so the below is showing both the real-time decision making infrastructure and the batch data processing and model generation (analytics) infrastructure.

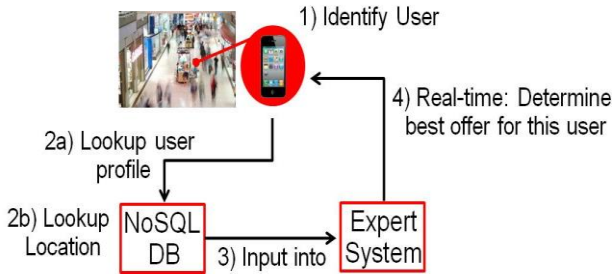

Fig 1. Highlevel Components

The first – and arguably most important step and the most important piece of data – is the identification of a customer. Step 1 is in this case the fact that a user with cell phone walks into a mall. By doing so we trigger the lookups in step 2a and 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 have found the actual customer, we feed the profile of this customer into our real time expert engine – step 3. The models in the expert system (customer built or COTS software) evaluate the offers and the profile and determine what action to take (send a coupon for something). All of 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.

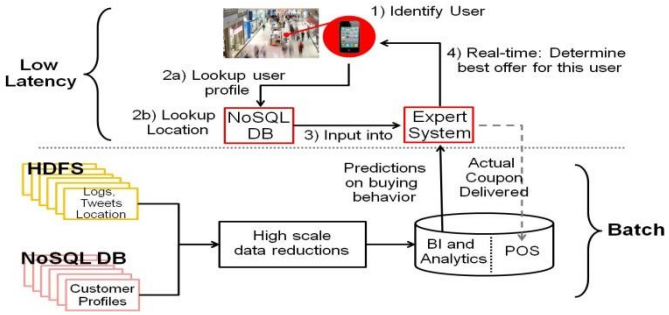

Fig 2. Backend Components Connection

To build accurate models – and this where a lot of the typical big data buzz words come around, we add a batch oriented massive processing farm into the picture. The lower half in the picture above shows how we leverage a set of components to create a model of buying behavior. Traditionally we would leverage the database (DW) for this. We still do, but we now leverage an infrastructure before that to go after much more data and to continuously re-evaluate all that data with new additions.

A word on the sources. One key element is POS data (in the relational database) which I want to link to customer information (either from my web store or from cell phones or from loyalty cards). The NoSQL DB – Customer Profiles in the picture show the web store element. It is very important to make sure this multi-channel data is integrated (and de-duplicated but that is a different topic) with my web browsing, purchasing, searching and social media data.

Once that is done, I can puzzle together of ebehavior of an individual. In essence big data allows micro segmentation at the person level. In effect for every one of my millions of customers!

The final goal of all of this is to build a highly accurate model to place within the real time decision engine. The goal of that model is directly linked to our business goals mentioned earlier. In other words, how can I send you a coupon while you are in the mall that gets you to the store and gets you to spend money…

VI. DETAILED DATA FLOWS AND PRODUCT IDEAS

Now, how do I implement this with real products and how does my data flow within this ecosystem? That is something shown in the following sections…

A. Collect Data

To look up data, collect it and make decisions on it you will need to implement a system that is distributed. As these devices essentially keep on sending data, you need to be able to load the data (collect or acquire) without much delay. That is done like below in the collection points. That is also the place to evaluate for real time decisions. We will come back to the Collection points later…

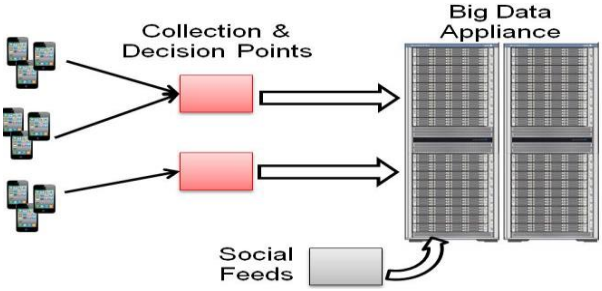

Fig 3. Collect Data

The data from the collection points flows into the Hardtop cluster – in our case of course a big data appliance. You would also feed other data into this. The social feeds shown above would come from a data aggregator (typically a company) that sorts out relevant hash tags for example. Then you use Flume or Scribe to load the data into the Hardtop cluster.

Next step is the add data and start collating, interpreting and understanding the data in relation to each other.

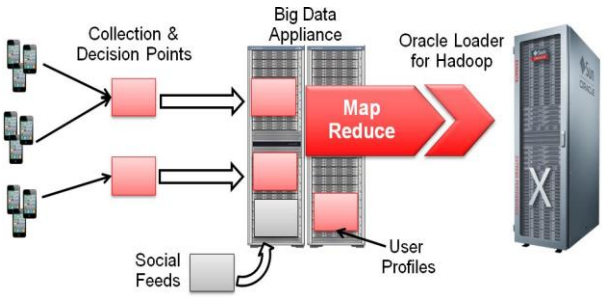

Fig 4. Oracle Loader for Hadoop

For instance, add user profiles to the social feeds and the location data to build up a comprehensive understanding of an individual user and the patterns associated with this user. Typically this is done using Map Reduce on Hadoop. The NoSQL user profiles are batch loaded from Mosul DB via a Hadoop Input Format and thus added to the MapReduce data sets.

To combine it all with Point of Sales (POS) data, with our Siebel CRM data and all sorts of other transactional data you would use Oracle Loader for Hadoop to efficiently move reduced data into Oracle. Now you have a comprehensive view of the data that your users can go after. Either via Exalytics or BI tools or, and this is the interesting piece for this post – via things like data mining.

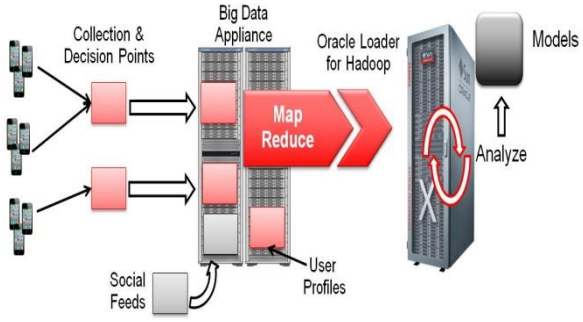

Fig 5. Map Reduce

That latter phase – here called analyze will create data mining models and statistical models that are going to be used to produce the right coupons. These models are the real crown jewels as they allow an organization to make decisions in real time based on very accurate models. The models are going into the Collection and Decision points to now act on real time data.

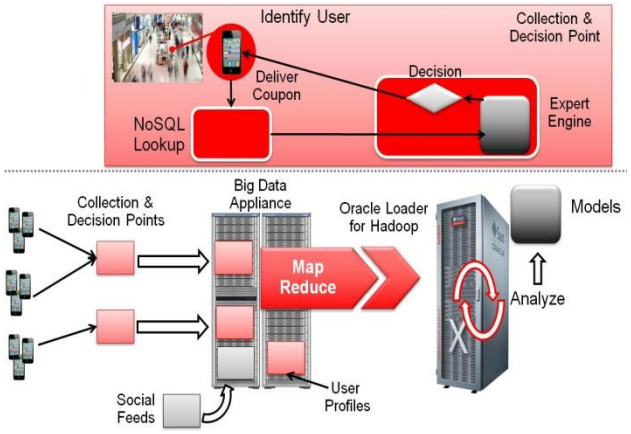

Fig 6. Map Reduce process

In the picture above you see the gray model being utilized in the Expert Engine. That model describes / predicts behavior of an individual customer and based on that prediction we determine what action to undertake.

The above is an end-to-end look at Big Data and real time decisions. Big Data allows us to leverage tremendous data and processing resources to come to accurate models. It also allows us to find out all sorts of things that we were not expecting, creating more accurate models, but also creating new ideas, new business etc.

Once the Big Data Appliance is available you can implement the entire solution as shown here on Oracle technology… now you just need to find a few people who understand the programming models and create those crown jewels.

VII. CONCLUSION

In this paper we have study about Big Data Mining

In different aspect like: Implementation and its component, high level component, data flow on work ideas.

REFERENCES

[1] "You Need A Weatherman To Know Which Way The Data Flows". *Media post .com*. Data and Targeting Insider. Retrieved 2 February 2015.

[2] "Data, data everywhere". *The Economist*. 25 February 2010. Retrieved 9 December 2012.

[3] "Community cleverness required". *Nature* 455 (7209): 1. 4 September 2008. doi:10.1038/455001a.

[4] "Sandia sees data management challenges spiral". *HPC Projects*. 4 August 2009.

[5] Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology". *Science* 331 (6018): 703–5. doi:10.1126/science.1197962. PMID 21311007.

[6] "Data Crush by Christopher Surdak". Retrieved 14 February 2014.

[7] Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". *Gigaom Blog*.

[8] Seagram, Toby; Hammerbacher, Jeff (2009). *Beautiful Data: The Stories behind Elegant Data Solutions*. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.

[9] Hilbert & Lopez 2011

[10] "IBM What is big data? — Bringing big data to the enterprise". Www.ibm.com. Retrieved 2013-08-26.