Use of Big Data in Marketing

Data Insight

Fall, 2013

Priyanka Deo(deo.priyanka02@gmail.com)

Dhananjay Apte(dhananjayapte@gmail.com)

Subhagini Chaudhary(subhagini.chaudhary@gmail.com)

Oren Gazit(oren.l.gazit@gmail.com)

Abstract

Use of Big Data in Marketing

Dhananjay Apte

Subhagini Chaudhary

Priyanka Deo

Oren Gazit

This project was designed to understand why our retail firm's revenues have been stagnant in the last few years. The Chief Marketing officer would like to invest in relevant and impactful technology to improve the sales program. We propose using the techniques involved in Big Data to analyze the various sources of information that the modern era provides, such as social media, store branded credit cards, and many others to help provide possible solutions to rejuvenate sales.

The paper introduces the history and concept of big data and how it came to be used in marketing. Although the concept of big data has been around for the last 20 years, its widespread use has been prominent in the last 5 years. We talk about various marketing techniques that have sprung up around big data, most notably targeted marketing but also other techniques such as dynamic pricing. The paper also discusses how big data could be used to improve store layout and otherwise capture missed opportunities. In fact, it has been found that intelligent usage of these big data techniques often increases sales by a large amount. Unsurprisingly, it turns out that many successful retailers, such as Wal-Mart and Amazon, are at the forefront of utilizing these techniques. Although the actual sales figures are proprietary, their sales are believed to have increased over the last few years.

Next, we discuss some of the more prominent technologies and techniques used in big data analytics, focusing on Hadoop and its associated distributions. We also discuss popular tools to visualize large data sets. We take a deeper look into target marketing and provide in-depth analysis. To explain the use of big data in targeted marketing, we provide a rundown of the various sources of data, such as company loyalty programs, and posts on social media(especially Facebook and Twitter). In addition, we talk about the various modules, APIs, and programs needed for big data analytics, such as sentiment and predictive analysis. We also give a description of the infrastructure and architecture required.

We look at Cloudera and Amazon EMR as possible distributions to use and discuss their pros and cons, allowing us to discuss which distribution to recommend under which circumstances. Our points of comparison include their various advantages and disadvantages, their cost, and more. In addition, we discuss the costs and advantages of developing an in-house solution, similar to what Wal-Mart has also done in the retail sector. As we must be realistic, we also discuss some of the challenges involved in adopting big data, such as a lack of trained personnel and other costs. Although the early adopters are not keen on disclosing how much it cost them to set up and implement their big data solutions, we use our best estimate, based off public data, such as the cost of servers, the cost of tools, the salaries of experienced data scientists and so on.

Finally, we conclude by suggesting a solution. In the end, we believe that not only should our company adopt big data techniques, using them to supplement our existing Data Warehousing and Business Intelligence, but should even look into expanding the number and type of big data techniques used in the future.

Table of Contents

1. Project Overview

The average basket size (purchase value) at the retailer "XYZ" has been stagnant over the last two years, in spite of an improved economy and general increase in consumer purchases at competitors.

The Chief Marketing Officer would like to understand the reasons for the stagnation and chart a path back to growth. The retailer has a website and mobile app that have good engagement from the customers. The retailer currently uses a traditional data warehouse (CRM, ERP, SCM) and BI for data analysis.

This report is intended to provide information on big data technologies and how the marketers for effective marketing can leverage these. Target marketing use case will be discussed in depth and viable solutions will be presented.

1.1. Why the topic is interesting

Traditionally companies have been using data warehouses in conjugation with Business Intelligence to help make business decisions. Data warehouses extract data from various sources like CRM, ERP,SCM, and others using Extract and Load tools into dedicated computers, where this information can be cleansed, reorganized and summarized. This data becomes a source of report generation, analysis and dashboards. (Andrews & Guerra, 2011)

Emergence of "Big Data":
"The term Big Data applies to information that can't be processed or analyzed using traditional tools and processes." (Eaton, Deroos, Deutsch, Lapis, & Zikopoulos, 2012, p. 3).

Big data is characterized by the volume, variety and velocity of data. These make traditional BI tools and data retrieval techniques insufficient to get any insight from this data.

Volume	Variety	Velocity
Organizations today are collecting petabytes of data from numerous sources like: ● social media ● blogs ● forums ● smartphone applications ● emails ● web log files ● call centers	A large proportion of data collected today is unstructured or semi structured (think of forum or social media posts). The variety of data imposes processing challenges as it is not possible to run statistical analysis on these data sets to reveal business insights. Traditional analytics	The velocity of data refers to the speed at which data is created and retrieved. Most data today from mobile and online channels is generated in real time at speeds multiple times faster than in the past and changes quickly. (Think of a user's location-based information)

- sensor data
- ERP, CRM, SCM data

platforms cannot handle this variety of data, thus creating a need for new tools and techniques.

Big data analytics and differences from traditional data warehouses: It is estimated that companies currently evaluate 20% of their data using data warehouses and remaining 80% data is lying untapped. It is not cost effective to evaluate all the data using a data warehouse. Big data analytics is performed on all or a huge amount of data unlike a data warehouse which uses a small set of data with perceived value. Exploratory analysis needs to be performed on this data over multiple iterations and these iterations need to be done quickly. The opportunity cost of this enormous amount of data is unknown unless it is retrieved, cleansed, converted into a standard form and evaluated. These issues are addressed by Hadoop. (Eaton et al., 2012, pp.3-63)

1.2. State of the art

Several companies have created platforms for implementing big data projects. These platforms extend Hadoop's capabilities and are bundled with tools for cluster administration, visualization, advanced text analytics etc. Some examples for these are:

IBM InfoSphere BigInsights and InfoSphere Streams -

InfoSphere BigInsights platforms is based on Hadoop and extending its capabilities. It performs analytics on static data. (Eaton et al., 2012, p. 17) On the other hand, Infosphere Streams performs analysis on data while it is in motion so that real time analysis can be performed for better decisions and outcomes.(Eaton et al., 2012, p. 123)

Some companies create their own in-house solutions that cater to their specific needs. For instance, @WalmartLabs for example created Muppet for analyzing fast data as Mapreduce cannot handle streaming data well. (Doan, Lam, Liu, Prasad, Rajaraman, & Vacheri, 2012)

When creating a big data solution from scratch multiple tools are available in the market which can be combined.

1) Data visualization tools display results of the real time analytics in various visual formats that are easy to understand while making it an interactive and compelling experience. ("Big Data Visualization: Turning Big Data Into Big Insights: The Rise of Visualization-Based Discovery Tools", 2013). Some popular tools are:

Jaspersoft BI Suite-

It is a open source tool that allow you to aggregate data from various sources and present it in the form of various interactive tables and graphs.(Wayner, 2012)

Pentaho Business Analytics-
It is a report generating engine; that is, similar to JasperSoft, it allows you to gather information from new sources. It can connect to most of the NoSQL databases, such as MongoDB and Cassandra.(Wayner, 2012)

TagCloud, D3.js, Pentaho Sunburst, Pentaho Zoom are a few other visualization tools.

2) Some of the tools enable you to use the development platform more efficiently. Below are a few such tools:

Karmasphere Studio -
It is an Eclipse-based specialized IDE that makes it easier to create and run Hadoop jobs.(Wayner, 2012)

Talend Open Studio-
This is also an Eclipse-based IDE for stringing together data processing jobs with Hadoop. Its tools help in data integration, data quality, and data management.(Wayner, 2012)

3) There are specialised tools that allow you to work on the data more effectively based on the business need.(Wayner, 2012)

Skytree Server-
This tool is used for running machine-learning algorithms on the data. (Wayner, 2012)

2. Project details

2.1. Big Data in Marketing

2.1.1. Various marketing techniques

Reports suggest that retailers that implement a big data strategy for marketing can achieve a 60% increase in their margin as well as improve employee productivity by 1%. This suggests that big data is important in today's age of marketing.("Big Data Offer Retailers the Chance to Stay Ahead of Their Competitors", 2013). A few of the marketing techniques are listed below:

2.1.1.1. Target Marketing

2.1.1.1.1. Understanding the customer profile and preferences:

Companies have access to a large amount of data about a customer from POS terminals, loyalty programs, websites logs, mobile apps, social media, email, customer center calls etc. Combining these disparate sets of information deepens a retailer's understanding of customers preferences. Wal-Mart created a project called "Social Genome" analyzing social media feeds and sentiment of customers. Walmart typically measures and combines this data with other customer data to provide better customer insights. (Manen, 2012)

2.1.1.1.2. Efficient, meaningful multi channel communication with users:

Customer profile provides retailers an understanding of how to best connect with customers to get incremental sales. Geo-tagged notifications, Bluetooth proximity based messaging inside a mobile app act as a contextual means to connect with the mobile savvy customer. Without big data, it would become challenging to engage a customer in a meaningful way. (Rijmenam, 2013)

2.1.1.1.3. Personalized shopping experience:

Multi channel information from online, mobile and in-store interactions and purchase history leads to creation of personalized product recommendations in real time, online and inside the store. Amazon is best known to study product purchase behavior at scale and recommend products to customers. These recommendations are constantly updated based on newer data and better algorithms. (Love, 2013)

2.1.1.2. Capturing missed opportunities

Retailers may not have the full picture if they rely only on in-store data collected from point of sales systems and CRM data, like loyalty programs. Augmenting online data on their websites with in-store sales data to reveal future opportunities. For example, while sales for a product in-store may not be very high, a lot of customers maybe visiting the website to search for that product but turning away. This could mean there are untapped opportunities if pricing, placement in the right stores or other factors were adjusted. (Banks, 2012)

2.1.1.3. Dynamic Pricing

Over 84% of the customer are using their mobile phone inside the store to search for more information. This is called “showrooming”, or comparative shopping to explore prices, discounts and product reviews. The advent of barcode readers for standard UPCs makes this even easier. The way stores typically could fight the advertisement is to adjust pricing of products both online and in store to keep the customer from finding a lower price option. Amazon regularly uses dynamic pricing. (“How Mobile is Transforming the Shopping Experience in Stores”, 2013 ; Mehra, 2013)

2.1.1.4. Store layout and product placement

Retails can optimize their floor plans but analyzing data from in-store video cameras, Wi-Fi, or Bluetooth data. Heat maps can be generated using the data to evaluate the foot traffic across the stores. This data, when analyzed along with the sales data, can provide information to optimize the store layout and product placement. (Rijmenam, 2013)

3. Industry adoption of Big Data

3.1. Usage statistics

According to a Gartner report of 2013, of the 720 organizations surveyed, 64% have either purchased or plan to invest in big data solutions. Of the 64%, less than 8% are using big data technologies.(Shu, 2013). According to IDC's Red Hat Hadoop Usage Survey(August 2013), Big Data and analytics will be a main constituent of infrastructure spending as businesses undergo a transformation to being a data driven.

Chart below the results of a survey conducted by the source written. They conducted survey on firms that are using Hadoop or are considering implementing Hadoop.

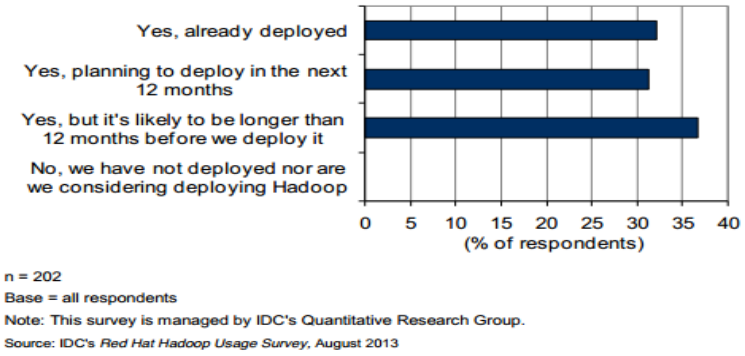


Figure1: Image from redhat.com (Redhat, 2013)

The results from the survey(refer the above chart) conducted on firms that are using Hadoop or are considering to implement Hadoop show that approximately 32% of firms have already deployed Hadoop and are using it for marketing. Primary use case for Hadoop is 'analysis transactional data from sales or point-of-sale systems' which constitutes for approximately 42%. The next being Analysis of online customer behavior data which is approximately 39%.

Companies which use Big data and analytics show increase in productivity rates and profitability and sales by 5-6% than peers which is an important aspect. According to McKinsey report, companies that put data in center and perform analytics on it for marketing and sales decisions improve their marketing Return Of Investment by 15- 20 %.This sums up to $150-$200 billion of additional value based on global annual marketing spend of an estimated $1 trillion which shows how data is important for marketing.(Gordan et al, 2013; Nadkarni, 2013).

3.2. Some examples of corporations using big data for target marketing

3.2.1. WalMart

Walmart was one of the first companies to recognize the potential in harnessing the vast quantity of customer data they possessed, and using it to analyze customers and organize the timing of sales for a given store location. (Bell, 2013) WalmartLabs actively implement various marketing techniques using big data. Walmart is using its big data collection to inform customers of their choices as a mobile marketing strategy. (Stokes, 2012)

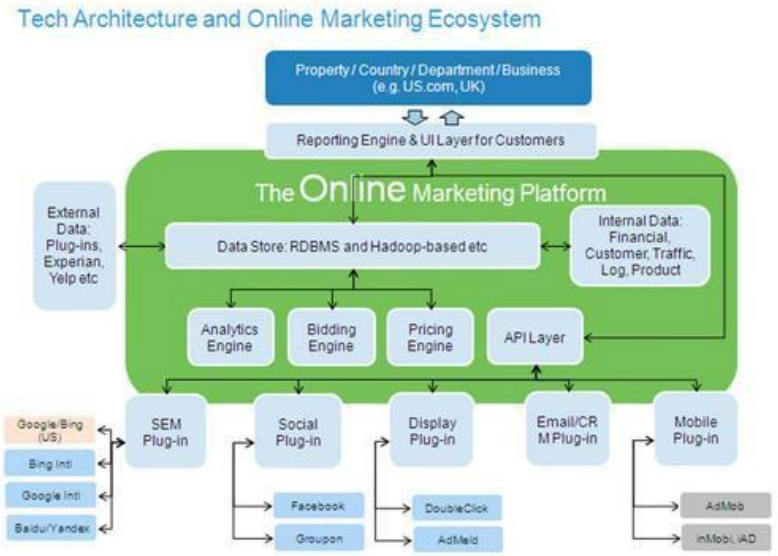


Figure2: Image from walmartlabs.com (Walmart, 2013)

Walmart uses techniques like machine learning, semantic analysis, optimization feedback loops, advanced algorithms, mathematics, and statistics, and scaling up through automation for its analysis. Walmart stores user information and then studies user behavior while shopping in order to provide personalized shopping experiences to the user. It is also used to give right recommendation in proper context. Walmart labs focuses on social, media and retail to create applications to promote marketing.

Walmart has developed 'The Shoppycat' for using big data in marketing. The Shoppycat is a social data based application that uses Facebook profile to suggest suitable products to the customers. It makes recommendations based on the interests and hobbies of your friends. This app has an interesting feature that if the product is out of stock in Walmart then it recommends the shops where that product can be found. (Zachary, 2012)
Walmart has developed mobile app with features like 'In store aisle location' and 'In store mode'. In store aisle location helps customers to locate products present on their shopping list on the app. Using this customers can select the products they want to buy beforehand and using this, the app will calculate the total amount that will be required based on the

local Walmart stores prices. In store mode is for customers who are shopping in retail stores. This includes features like scanning the product barcodes for price checking or extracting more information about products and special offers displayed in-store as QR Codes. It is also used for easy access to the Local Ad product list and general store information.

3.2.2. Amazon

Amazon.com is the largest online retailer in the United States. It uses its customer data and data analytics to build its recommendation engine. Amazon got high sales because of this engine nearly 35%.(Bell, 2013)
According to Hitwise Mobile, Amazon receives 59.36% of mobile department store visits. Also comparing Amazon receives 45.24% of desktop PC department store visits. The company reported a 29% sales increase to $12.83 billion during its second fiscal quarter, up from $9.9 billion during the same time last year. (Siwicki, 2013; Mangalindan, 2012; "About Data Mining", 2013, aboutdm)
Amazon uses big data for its big system namely the recommendation engine which is in real time on its website. This engine has 2 main parts: 'frequently bought together' and 'customers who bought this also bought'. Amazon uses customer tracking in which it uses consumer attributes and online behaviors to implement one-to-one marketing. This is called personalized marketing. It seeks to deliver the proper products to the best person at the right time. Another important aspect implemented by Amazon is multi-leveled e-commerce strategy.
Along with recommendation system Amazon uses email marketing. Amazon stores information on all its users and promotes marketing by sending email to the customers.

3.2.3. Others -

The other companies using big data include gaming industry which has been at the forefront of analytics. Gaming industry uses big data in analyzing the behavior of people and promote marketing accordingly. For example, Caesars Palace uses big data to detect if people are unhappy losing at slot machines. In this case the customers leave the game and blame it on Caesars in contrast to understanding the probabilistic nature of casinos. So they promote marketing by giving customers free meal at buffet and convincing them that things will be better when they come back. Results show that the likelihood that they return is 40% (as opposed to 20%).(Hanfield, 2013)

4. Technologies for Big Data

4.1. Big Data Solution

The big data landscape is continuously evolving with new technologies, products, tools etc. Here we will focus primarily on Apache Hadoop, the de facto standard for processing big data. Several products are available in the market ranging from the Apache Hadoop release, Hadoop distributions and Big data suites. The Hadoop distribution and Big data suites are offered by multiple vendors .(Wähner, 2013)

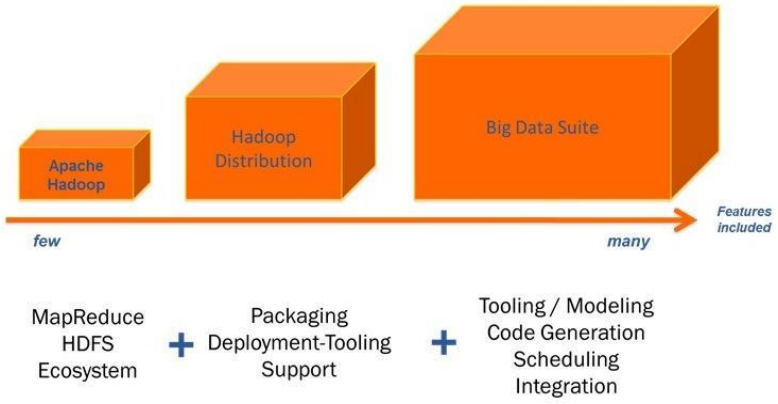


Figure 3: [Online image for Alternatives for Hadoop Platforms]

4.2. Hadoop

Hadoop is an Apache project written in Java. It was inspired by the Google File System(GFS) and uses MapReduce to operate on data stored in a distributed cluster system and processes the data in parallel. These clusters are created using commodity hardware and scale easily . Hadoop comprises of the Hadoop Distributed File System(HDFS) , MapReduce, Hadoop Common and Hadoop YARN. (Eaton et al.,2012, pp. 3-63 ; Wähner, 2013).

4.2.1. Map Reduce and HDFS

In HDFS, data is broken down into blocks and distributed in the cluster. Each block is replicated onto two additional servers. This provides for scalability and high availability, as community hardware has a high rate of failure. The default size for the block is 64MB. A dedicated node called the NameNode stores metadata related to information on where blocks are stored in the cluster. A backup of the NameNode is highly recommended. Hadoop uses the concept of data locality. When a Hadoop job is fired, the NameNode is contacted to find the location of the server that holds the data and then the application is

sent to run locally on those nodes, instead of the data being brought to the application.(Eaton et al.,2012, pp. 3-63)
MapReduce is a programming paradigm. It consists of two separate tasks called Map and Reduce, that the Hadoop program performs. The map task works on a set of data to convert it into key/value pairs, while the reduce task takes input from the Map task and creates a smaller set of the key/value pairs. An application sends a job to a special node called the JobTracker in the Hadoop cluster. JobTracker communicates to the NameNode to find the location of data and breaks down the job into map and reduce tasks.The status of each task is continuously monitored by a daemon called TaskTracker that informs the JobTracker in case any task fails so that it can be rescheduled elsewhere. (Eaton et al.,2012, pp. 3-63)

4.2.2. Hadoop Ecosystem

There are several Hadoop-related projects at Apache that comprise the Hadoop ecosystem. These are some of the few:

- To abstract the complexities of the MapReduce programming paradigm several languages have been developed. Some of these are *Pig, Jaql and Hive*.
- Apache Avro provides data serialization.
- *Cassandra* and *HBase* databases are popular for storing data in the Hadoop Cluster.
- Large Hadoop distributed systems can be monitored and managed using *Chukwa*.
- *Zooper* provides services for coordination of distributed application.
- *Mahout* is a library for data mining and machine learning.
- *Sqoop* is a tool for bulk data transfer between Hadoop and structured databases.
- Log data can be efficiently collected, aggregated, and moved using a distributed service called Flume.(Wähner, 2013)
- The Hadoop community is rapidly contributing towards the Hadoop ecosystem. The following diagram provides an overview:

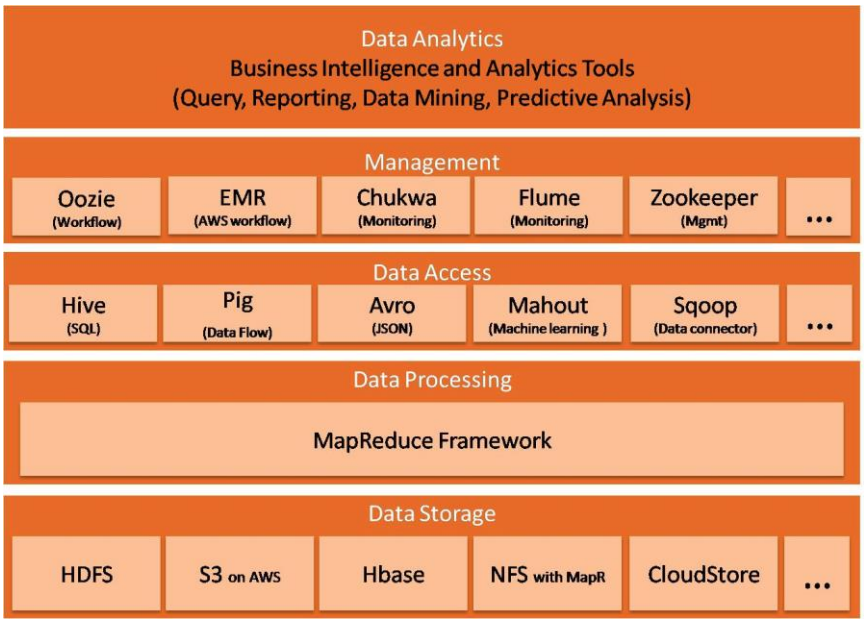


Figure 4: [Online image for Hadoop Ecosystem]

4.2.3. Limitations of Hadoop and the Hadoop ecosystem

Apache Hadoop has several known shortcomings. Some of these are listed below:

- The name node presents a single point of failure. Recovering from this can take a long time as all the metadata needs to be loaded back into the memory.
- Setting up the cluster is very complex. Cluster configuration and administration is laborious and complicated.
- Hadoop cannot process streaming data.
- Apache Hadoop has no commercial support, as it's an open source project.
- Hadoop related projects of the Hadoop ecosystem have to be installed manually. Care has to be taken while installation, as all the releases don't function properly together. Hence appropriate versions have to be selected to ensure compatibility. (Wähner, 2013)

4.3. Hadoop Distribution

The Hadoop distribution comprises of the different Hadoop-related projects (section 3.1.1.3) such that all related versions are compatible and work correctly together. New releases have updates from different projects.(Wähner, 2013)

In addition to packaging, the distributions include features for reducing complexity of cluster setup and monitoring. Graphical tools for cluster administration, deployment and monitoring are also included. The most popular vendors for Hadoop distribution are MapR, Cloudera and HortonWorks. Others such as Amazon Elastic MapReduce are

also popular. Cloudera is the most popular hadoop distribution in terms of number of deployments. These distributions have minor differences and vendors provide support for their products. These distributions can be used independently or can be combined with big data suites. (Wähner, 2013)

4.3.1. Limitations of Hadoop Distribution

Although the Hadoop distribution has many advantages, the code still needs to be written for MapReduce jobs and integration with different sources of data. This is not easy as it requires a high level of expertise. These issues are addressed by big data suites. (Wähner, 2013)

4.4. Big Data Suites

Big Data Suites operate on top of the Hadoop distribution or Apache Hadoop and add multiple features to it. Several vendors provide support for the different Hadoop distribution or have their down Hadoop solution. (Wähner, 2013). Some of the key features included in the big data suites are: tooling(use of an IDE like Eclipse), modeling (big data services can be modeled using graphical tools), code generation (MapReduce code is automatically generated), scheduling (defining and managing execution of MapReduce job) and integration (integrating data from different products and technologies like B2B products, social media, SQL and NoSQL databases). Several options exists between open source and proprietary vendors and most big software companies like IBM (InfoSphere BigInsights and InfoSphere Streams), Microsoft (HDInsight) provide big data suites. (Wähner, 2013)

4.5. NoSQL

Traditional relational databases technologies(RDBMs) lack the ability to store multi-structured data and cannot perform to big data scale. NoSQL(Not Only SQL) databases are best suited for big data applications. These are deployed over HDFS. A few of the NoSQL databases available in the market are:

- Cassandra
- HBase
- mongoDB
- MarkLogic
- CouchDB
- Riak
- DynamoDB
- Accumulo
- Aerospike. (Kelly,2013)

4.6. Techniques for Data Analytics

Advanced techniques for data analytics can be applied to the data processed from the MapReduce program. Application of these techniques can help gain meaningful insights , which can help make business decisions. Techniques for data analytics include predictive analysis ,natural language processing, text analysis, sentiment analysis, data mining and machine learning. Some of the statistical languages used by data scientists to perform this task are - R and SAS . ("What is Big Data Analytics?", n.d.; Kelly, 2013)

4.7. Visualization Tools

Results from big data analytics need to be displayed in ways that are easily to understand for the business users. Several visualization tools are available in the market that displays information in the form of charts and graphs. Some of these are Clustergram, Sencha, dojo, D3JS, Google Charts and Pentaho. (Vappalapati,n.d.)

4.8. Other Tools

There are several tools available in the market that generate reports from No-SQL databases, HDFS file data and HBase data- Jaspersoft BI Suite, Pentaho Business Analytics, Karmasphere Analyst. Other tools provide plugins for IDE's like Eclipse for running and creating Hadoop jobs and provide data management-Karmasphere Studio,Talend Open Studio. While others provide machine learning algorithms- Skytree Server. Tools like Tableau provide data visualization by slicing up data and mixing in a multiple ways to be examined differently. (Wayner, 2012)

5. Big Data solution for Target Marketing

5.1. Sources of data for target marketing

With the deep penetration of internet in a customer's life, a vast amount of data is being generated from sources like social media, online, and mobile retailing. For example, on average Facebook has 665 million active users daily.(Facebook. 2013) This daunting figure is self evident to the potential data available for analysis. Petabytes of data is generated from these sources.

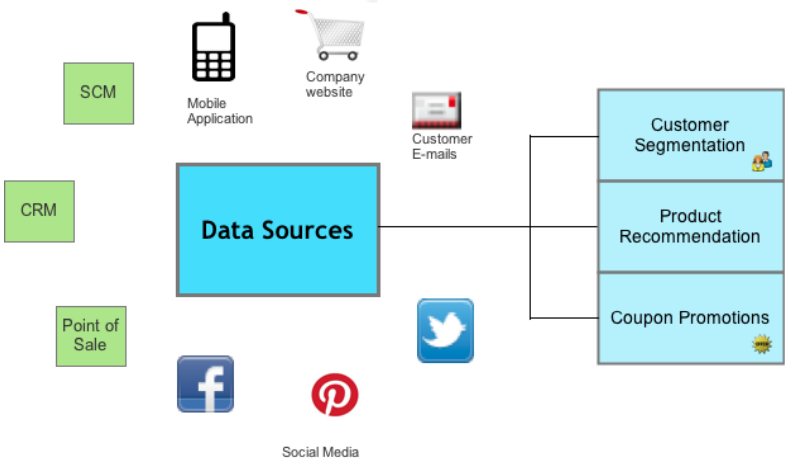


Figure 5: Sources of Big Data

All this data can be analyzed and used to create a customer profile. This customer profile is an ever evolving profile as it will be updated every time new data is gathered, The profile thus created can then be used to target customers based on their preferences and choices. It can be used to anticipate the customer purchases and send suggestions, offers and discounts to the customer to increase the sales. Below we explain these various data sources in detail.

5.1.1. Social media

Social media is helpful in gaining insights about the likes and dislikes of the customers. The opinion of the customer about specific brands or products. Social media can be effectively used to advertise new products. Gain insights on response and opinions on a product. Various social media vendors allow companies to utilize their customer information and provide various platforms and API to manipulate the various features of the social media to the benefit of the retailer/company. There are lot of different social media sites from where we can retrieve data. Some of the popular social media sites are discussed on the next page:

5.1.1.1. Facebook

Facebook is one of the most popular social media platforms where people connect with their friends and family. It has a lot of personal information like the education details, work details, location etc. It can be used to get insights on the customer's views, their opinions their likes and activities they participate in.
The opinion of a friend or family makes an impact on the choices of the customer. A customer may try out a certain product because his friend recommended it or because a lot of his friends like it. There are many such aspects of social life that can be leveraged to the benefit of the retailer.
The information gathered from Facebook can be used to refine the customer profile and can be used to effectively market the products that the customers like. For example, if it is known from the customer Facebook profile that he frequently goes for hiking, then the retailer can send him ads related to hiking or offer discounts on hiking products.

5.1.1.2. Twitter

With about 230+ million active users and 500 million tweets sent per day this microblogging and social media service is a huge source of data.(Twitter. 2013) The tweets made by the customers can be used for keyword and sentiment analysis. Twitter can be used by companies to engage with users and communicate with them. Users provide word-of-mouth marketing for companies by discussing the products on Twitter.
Twitter has users who follow set of people and set of people follow
him. Users can send updates which is seen by his followers and also users can retweet other's updates. Twitter tracks retweet counts for all tweets. Thus using Twitter data we can find out which product of our company is being discussed frequently. Also using sentiment analysis we can find whether the comments are positive or negative. (Strickland, 2013)

5.1.1.3. Pinterest

Pinterest has users who can pin things and store all of your pins on your account so that you can access them easily. Also users can follow friends on Pinterest and "repin" things that they have already pinned on their Pinterest boards or browse a live feed of items that are being pinned by strangers when you're searching for inspiration.
Thus companies can use this data to know the popularity of their products. Also company can know the feedbacks of their customers from this site. Using sentiment analysis, company can know whether they received positive or negative feedbacks. Pinterest has recently launched public APIs. It has started giving endpoints to some partners.(Pinterest, 2013)

5.1.2. Company website

As mentioned in the problem statement the company has web presence. This will allow customers to create their online accounts on the company website and buy various products. The company will be able to retrieve information like the birthdate, address, age, and other details about the customer. this will help us build a profile of every customer. This will be particularly helpful in targeting customers based on particular criteria. For example, using the age you can target customers between the age of 20-30 years.

The company website can also be used to keep track of the user clicks when he/she is logged in and determine what products the user has been searching for . Maybe the user does not buy a particular product because it is costly. In such a case the company can send a special discount coupon to the particular customer and thus will result in more sales. The user click data will also be helpful in determining what products are more popular and different insights could be drawn.

5.1.3. Loyalty programs

Loyalty programs can be used to target the more frequent customers. Loyalty programs are used to retain existing customers as retention significantly boosts profits. You can offer special discounts to regular customers. This increases customer satisfaction and in turn increases the profit.

5.1.4. Mobile Application data

The user data from the Mobile Application will also be used in a similar fashion. Apart from his purchases and browsing data, it can be used to track current location of the user. Based on this relevant offers can be targeted to that customer whenever they are in a stores vicinity.

5.1.5. Point of Sale data

Point of sale data can be used to determine the purchases made by the customer at the retail store. It can also be used to determine the returns of the products, reasons they were returned. In addition, it can be used to record customer suggestions and complaints.

5.1.6. Other sources

Other resources include the existing transactional data from the BI. Data of past customer history from BI. Data from SCM and CRM can be used to determine user and product information . Data from customer emails can be analyzed. Information from customer surveys can also be a source of data.

5.2. API summary

This section provides a brief summary of the various API's provided by the different social media discussed above.

5.2.1. Facebook:

Facebook provides a Facebook Platform that allows developers to develop various applications and features that will allow them to manipulate various Facebook features to their benefit and as per their requirements. It provides a set of API's each providing different set of capability to the developer. (Facebook. 2013)

A few of the Facebook APIs are listed below:

- *Graph API*-The Graph API is a simple HTTP-based API that allows you to build social applications by accessing the Facebook social graph. "Using the Graph API you are able to post new stories, upload photos, retrieve posts and a variety of other tasks that an app might need to do." (Facebook. 2013)

- *Open Graph*- This API allows you to post stories on the user's wall. Suppose a user likes a particular article in your app, you can post a story to the user wall that the user likes the article.

- *FQL*- FQL is Facebook Query Language, it allows you to query the Graph data to retrieve data.

- *Public Feed API*- This API allows you to read the status updates and comments of users that have a privacy setting as public.

- *Keyword Insights API*-"The Keyword Insights API exposes an analysis layer on top of all Facebook posts that enables you to query aggregate, anonymous insights about people mentioning a certain term." ("Facebook APIs",2013)

5.2.2. Twitter:

It is possible to use Twitter API which can be developed or predefined and Twitter Content in with the products or services to search, display, analyze, retrieve, view, and submit information to or on Twitter. Twitter provides many APIs using which we can retrieve and manipulate data.

- *Twitterlicious and Twitterific* -These are two applications which are used to access Twitter through desktop applications on PCs and Macs, respectively.

- *Tweet Scan*-This is an API which is used to search public Twitter posts in real time using either a customized search engine or Firefox's search box.

- *Twittervision*- This API integrates a Twitter feed into Google Maps. This is used to watch public posts go live through a world map.

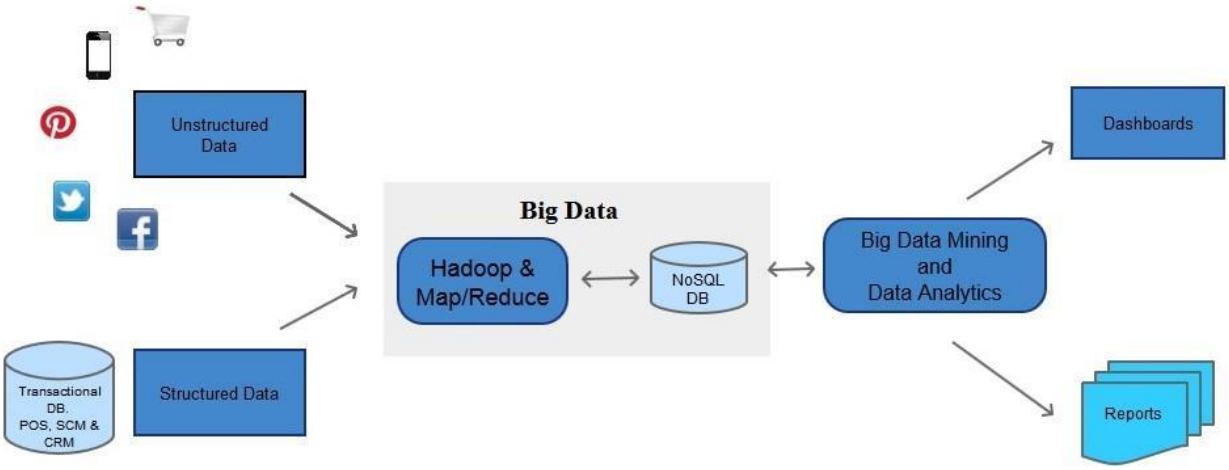
5.2.3. Pinterest:

Pinterest has released API that allows web developers to keep the most recent, trending and most-clicked pins (Gorman, 2013; Orisini ,2013)

- *Mapshape API*- This can be used to use Pinterest API
- *Top repin*- Top Repins Domain endpoint is used to display over a certain period of time, a collection of the most repinned Pins from your domain.
- *Most Recent Pins*- Most Recent Pins endpoint is used to fetch the most recently created Pins from your domain.
- *Related pins*- Related Pins API is used to display a list of Pins from your domain similar to a Pin you choose.
- *Domain search pin*- Domain Search API is used to see Pins from your domain that include appropriate search terms.(2013, “Pinterest APIs”)

5.3. Architecture

The sources of unstructured data are mainly through social media. We have shown a few sources in the diagram. Structured data can be obtained through sources like transactional DB, POS, SCM, CRM, etc. The data obtained from these sources is transferred into the Hadoop ecosystem. We can transfer relational data into Hadoop ecosystem by using tools like Apache Sqoop. Sqoop can be used to import or export data from all databases supporting the JDBC interface.(Apache, 2013)



Another tool used for data transfer is Apache Flume. It is a distributed, reliable, and available service used for collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming **data flows**. It is robust and fault tolerant. It uses an extensible data model that allows for online analytic application. FileChannel is a Flume channel that generally supports writing to multiple disks in parallel and encryption.

This data can be stored in NoSQL DB and retrieved whenever required. Data mining and analysis is done on this data. Various data mining and analysis tools are available in market. The statistics and results obtained can be displayed on dashboards. Also these results can be stored as reports. (Apache, 2013)

Below is an example of how Twitter data can be ingested into HDFS:

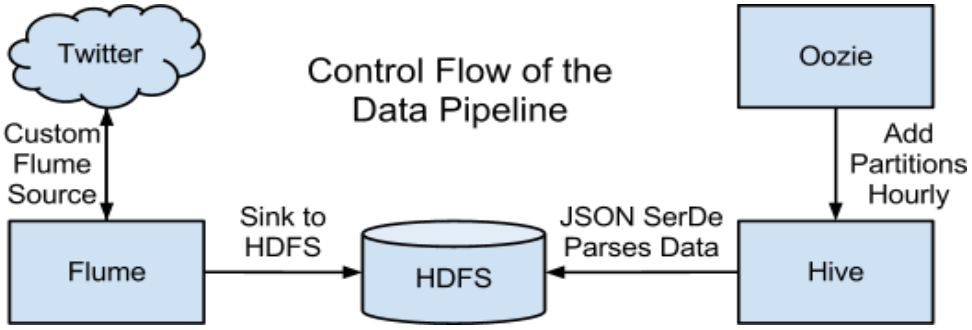


Image from Clourera.com (Cloudera, 2013)

The above diagram shows the complete data pipeline for using Twitter data for analysis. The data from Twitter is continuous stream of tweets. Flume is used as a data ingestion system. In Flume source of data is Twitter and sink is HDFS files. Hive is used to query this data which, is obtained in JSON format. This is similar to traditional SQL queries. As data is continuously streaming in, HDFS has to partition data. For this Oozie is used which is a workflow coordination system.(Natkins, 2013)
Thus, Twitter data can be obtained and information can be extracted from it.

5.4. Infrastructure

The infrastructure will require large number of commodity hardware. This can be either bought and installed by the retail company or there are many prominent cloud service providers that can provide it at an efficient cost. We would suggest that the company initially use the cloud platform so that the setup costs of the infrastructure and maintenance will be low. With cloud based infrastructure, the company can scale up and scale down as and when required.

5.5. Data Visualization

It is important that the data insights obtained from all the data be presented in an appropriate way. The managers or business users who will use the data should be able to draw meaning from the data. And data should be displayed in a manner convenient to the user of the data. For instance one manager may require certain data in an spreadsheet format while other person may want the data as a graph. These needs have to be satisfied for better output. A large number of data visualization tools are available to choose from. A few tools have been mentioned in Section 3.4.

5.6. Integration of Big Data with Data Warehouse/BI

Hadoop augments and complements Enterprise Data warehousing. The most common information source was online transaction processing (OLTP) with relational database for the BI/data warehousing architecture. It was based on structured data. But with the recent use of big data we can use unstructured or multi structured data as well. The basic architecture of BI/Data Warehousing is shown below:

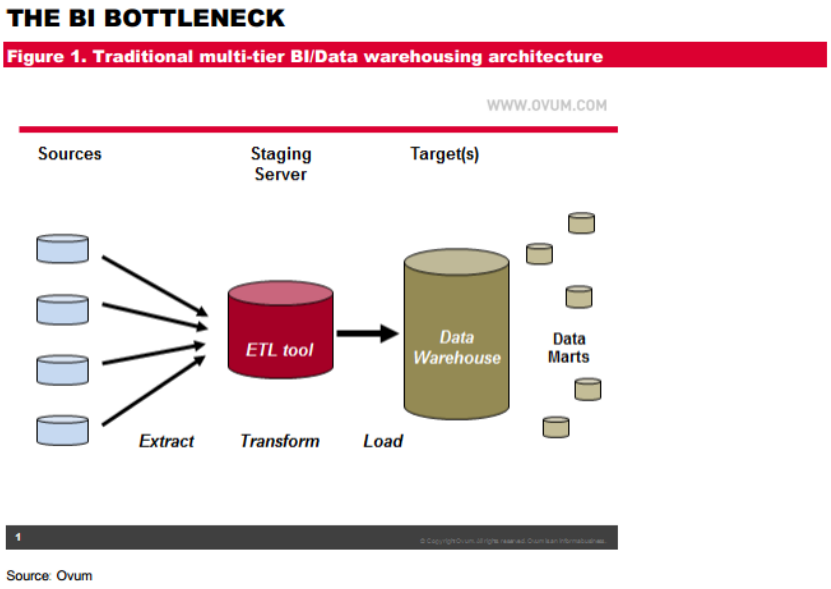
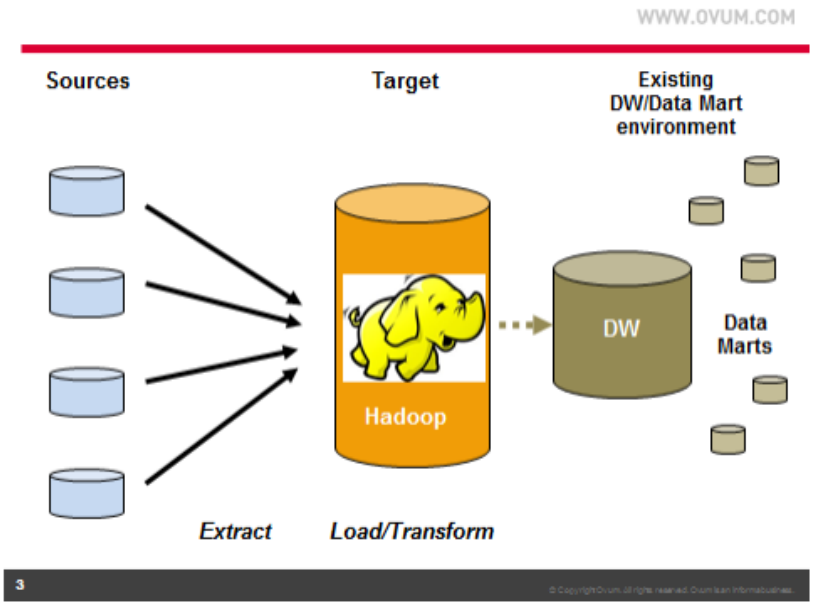


Image from Cloudera(cloudera.com)

Enterprise Data warehousing provides platforms for providing highly processed data into BI deliverables. EDW is needed for its expertise with relational and dimensional data. Also it has tough requirements for transformational processing and an audit trail. Hadoop is specially used for managing and processing huge file-based data. This data cannot be transformed and loaded into DBMS. Hence Hadoop and EDW are

complementary to each other. Hadoop can be used to analyze unstructured data, such as petabytes of Web log data in large Internet firms, social media data, call detail records in telecommunications, unstructured claims documents in insurance, XML documents in supply chain industries, and a wide variety of log data from machines and sensors. It provides a flexible, economic platform where data transformation cycles can be performed. This will reduce the burden from SQL systems.
The diagram below shows how Hadoop can fit into the traditional EDW system.

Figure 3. Hadoop as Data Transformation platform



Source: Ovum

Image from Cloudera(cloudera.com)

In the Data Warehouse architecture Hadoop can be used as a huge data staging areas and archive detailed source data. They also provide ease in analysis. (Russom, 2013; Ovum)

6. Comparison between different big data solutions

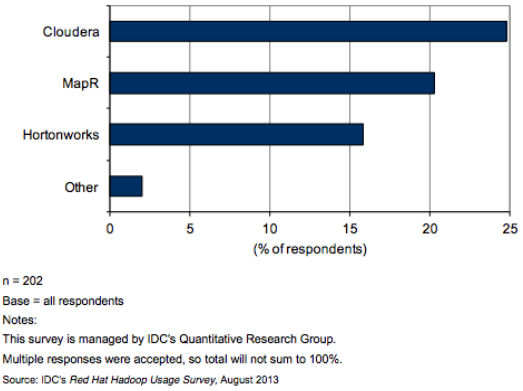
6.1. Current solutions in the market

There are a huge number of Big Data solutions on the market, ranging from Hadoop-based distributions to GFS and more. However, we will largely be focusing on the Hadoop family.

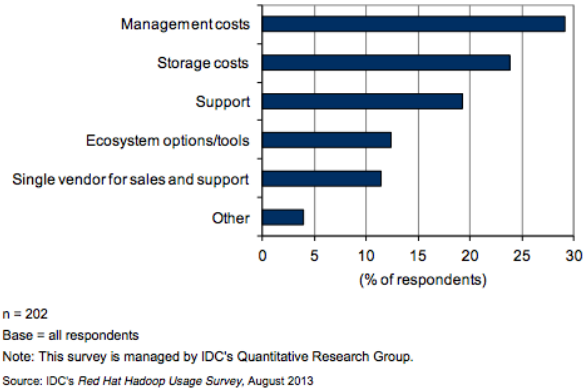
6.2. Comparison between two viable solutions

As per an IDC report, a survey conducted in August 2013 revealed that Cloudera was the most popular choice for the Hadoop distribution (Nadkarni & DuBois, 2013).

Most used Hadoop distribution



Primary Reason for Choosing Hadoop Distribution



Charts from a Red Hat paper: (Nadkami, 2013)

We will examine Cloudera and Amazon EMR as viable options for implementing the project for target marketing.

6.2.1. Cloudera

Cloudera is the most popular Hadoop distribution in the market. Its product offerings comprise the Standard Edition and a Enterprise Edition. The Enterprise Edition has a 60 free trial period, after which costs apply. The Hadoop distribution provided by Cloudera (CDH) is 100% Apache open source. Cloudera also provides the option to free download the CDH . (“Cloudera Enterprise”,n.d.; “Cloudera Product Comparison”, n.d.)

The features offered by Cloudera are as follows:

- *Hadoop Distribution - CDH*
  This comprises Apache Hadoop along with a number of other Apache open source projects that are a part of the Hadoop ecosystem.
- *Cluster Management - Cloudera Manager*

The Cloudera Manager helps in easily deploying, monitoring, managing and performing diagnostics on the cluster. Other features for disaster recovery and data management are available.

- *Centralized Data Management-Cloudera Navigator*
  It includes access management and data audits(Hive, HDFS and HBase).
- *Cloudera Support*.("Cloudera Enterprise",n.d.; "Cloudera Product Comparison", n.d.)

Features	Standard Edition	Enterprise Edition
CDH	Yes	Yes
Cloudera Manager	Limited, with features for deploying, managing , monitoring and perform diagnostics on the cluster.	Additional capabilities (RTD, RTQ, RTS, BDR) such as data management and disaster recovery can be added through add-on subscriptions.
Cloudera Navigator	No	Yes, with add-on subscription.
Cloudera Support	Limited. Support available for CDH and cloudera manager.	Comprehensive support available.
Cost	Free.	Costs apply.

** RTD(Real time Delivery), RTQ(Real-time Query with impala), RTS(Real time Search), BDR (Backup and Disaster Recovery). ("Cloudera Product Comparison", n.d.)*

Pricing information of the Enterprise edition is not available on the Cloudera website but the estimated yearly charges are approximately $4,000 per node (including subscriptions and support).This is similar to what MapR and EMC/Greenplum charge for their offerings.(Morgan,2011)
Below is a representation of the various subscription options available providing customization based on a customers requirement:

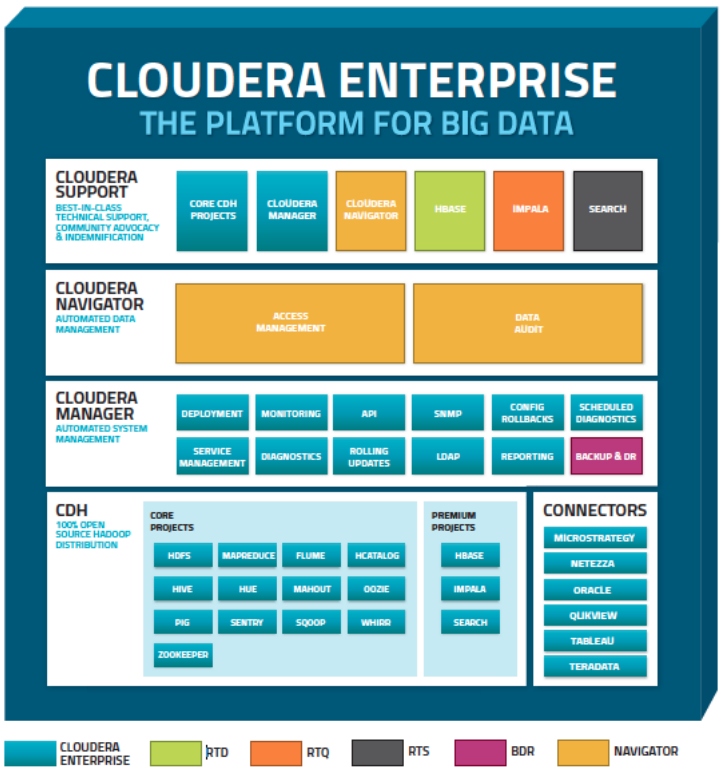


Image of Cloudera Enterprise Subscription Options (Cloudera, 2013)

Impala is a cloudera project that adds real-time querying capabilities to Hadoop, complementing the batch processing MapReduce. Impala does not use MapReduce and has been made open source. It uses the SQL like query syntax for data stored on HDFS and HBase using the Hue interface.Impala and Hive provides integration with some of the leading tools for BI .Cloudera provides JDBC and ODBC connectors for RDBMS systems. Most ELT tools provide integration for Hadoop and HBase.(Kornacker & Erickson, 2012)

Cloudera Search is another premium project at Cloudera that allows users to explore data from Hadoop and HBase using full-text data exploration and drill down navigation, using Hue. It is based on Apache Solr. (“Cloudera Search User Guide”, 2013)

Cloudera distribution can be hosted on premise or on the cloud. It is available on the cloud using Amazon EC2. The Hadoop\HBase clusters can be launched using Apache Whirr on EC2 using the Cloudera Manager Free Edition for up to 50 nodes. This can later be upgraded. (London, 2012)

Cloudera has recently partnered with several providers for cloud infrastructure. These include T-Systems, Softlayer, Verizon and Savvis.(Brandon, 2013)

6.2.2. Amazon EMR

Amazon EMR is also one of the more prominent Hadoop distributions. Unlike Cloudera, with Amazon EMR the distribution can be done on the Amazon servers. There are three major distributions: M3, which is free but has the fewest features, though it is a fully functional version; M5, which has more features but lacks several features, and M7, which has all the features, as seen on the figure below. We will focus on M7 and M3. Both M3 and M7 have the following features:

- A complete distribution of Hadoop
- The MapR Control System, or MCS, allowing for easier management of the cluster
- If done on the Amazon servers, M3 costs approximately ⅓ as much as M7. For example, in Northern California a single large M3 node costs approximately $0.06/hour, while an M7 node costs $0.17/hour.(AWS, 2013)

However, M7 also has many additional features:

- Better Data management
- Higher durability from increased mirroring and extra snapshots for both files and HBase
- Automated upgrades
- Increased automation in data management and other areas(AWS, 2013)

6.2.3. Comparison

Overall, EMR has advantages in the following:

- simple jobs, due to their web console
- multiple clusters/cluster sizing; thanks to the size of the Amazon servers, it's much easier to dynamically resize
- AWS integration; if we choose to use other AWS products, then EMR has a large advantage(Hammerbacher, 2010)

Cloudera has advantages in the following areas:

- Cost: EMR is estimated to be about 20% more expensive
- Open Source: Because Cloudera's distribution is open source, we can make changes to it on our own.
- Easier integration with Hive, Zookeeper, etc.
- Faster updates: Cloudera updates their Hadoop distribution faster than Amazon
- Non-proprietary cloud: If we wanted to run it on our own private cloud or on a different public cloud, we would be allowed to. (Hammerbacher, 2010)

6.3. Option of In-house solution development

As discussed in section 4.2.3 setting up the Hadoop cluster from scratch is complex and would require considerable time and money. In light of this, the option of in-house solution is currently not recommended. Instead Amazon EMR (M3) or Cloudera Standard edition be used, as both these software are free (only hardware cost \ cloud pricing would be incurred). Once a proof of concept is successful, further options can be evaluated.

7. Challenges in adoption of Big Data

7.1. Expertise

There is a dearth of professional who posses big data skill ranging from MapReduce to data Analytics and more. In an exhaustive analysis by Burtch Works, they found that the salaries of Big Data users are fairly high, with individual contributors receiving a median of $90000 per year, and managers receiving $145000 per year, both before bonuses.(Burtch, 2013) The reason for these high salaries is that Big Data is a new field which is also difficult to understand, leading to a situation where trained personnel, let alone good trained personnel, are very pricey.

7.2. Cost

7.2.1. Initial set-up cost

According to Winter Corp's exhaustive report: "Big Data - What does it really cost?", "A common estimate for the list price to acquire a Hadoop cluster, including hardware and open source software, is less than $1000 US per TB stored."(Winter, Gilbert, & Davis, 2013)

A survey of 202 companies by Red Hat found that the average cost for data migration is around $250,000, which is largely insignificant.(Nadkami, 2013)

However, applications need to be developed in order to use the data, and that is where the lion's share of the cost comes in; in a sample Warehouse from Winter Corp, out of the 9.3 million dollar startup cost, 7.2 million of it came in application development.(Winter, Gilbert, & Davis, 2013)

7.2.2. Annual cost

Annual costs come from two main sources: maintenance/rent, and staff. As discussed above, the average cost of staffing is about $90000/year for an employee and $145000 for a manager. If we assume the group consists of 20 engineers and three managers, then that is about 2.5 million dollars annually in salary. In addition, a survey by Red Hat shows that the median cost of maintaining a Hadoop node is between $1000 and $2000, and the median number of servers is around 50. As a result, we can assume a cost of around $75000 to maintain the servers, which is largely insignificant. In addition, we can expect a growth of about 20% per year(Nadkami, 2013), so we will slowly need to add more servers, but the numbers should stay relatively insignificant compared to the cost of salaries.

8. Conclusion/Suggested solution

Target marketing will help improve the companies business by enriching customers shopping experience. By gaining deep understanding for the customers the marketers can effectively cross sell and up-sell products. Better promotions can be launched product recommendations can be offered. This calls for analysing and storing data of varied structure and high volumes. Big data capabilities can be leveraged for this. Two viable solutions - Cloudera and Amazon EMR were discussed. An initial POC would be recommended using either Amazon EMR M3 edition or Cloudera Manager Free Edition on Amazon EC2 (which can scale up to 50 nodes) . This will help reduce the overall infrastructure cost and provide flexibility. Once successful further options can be evaluated.

In addition, other strategies for target marketing such as location based marketing can be evaluated. Coupons can be sent in real time depending on the users location. Other marketing techniques such as product assortment and store layout can be implemented in future versions of the project.

9. REFERENCES

Andrews, D., Guerra, J. (2011). Why You Need a Data Warehouse. In Rapid Decision. Retrieved Novmber 22 , 2013 from http://www.rapiddecision.net/pdfs/Why-You-Need-a-Data-Warehouse.pdf

Eaton, C., Deroos, D., Deutsch, T., Lapis, G., Zikopoulos, P. (2012) Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. NewYork: McGraw-Hill.

Shu, C. (2013). 64% of Organizations Have Invested in or Plan to Invest in Big Data Tech, But Only 8% Have Started Using it, says Gartner. In Techcrunch. Retrieved Novmber 22 , 2013 from http://techcrunch.com/2013/09/23/64-of-organizations-have-invested-in-or-plan-to-invest-in-big-data-tech-but-only-8-have-started-using-it-says-gartner/

Manen, T. V. (2012 , Sep 4). Fueling Sales with social data:The story of Walmart Labs & the social Genome. In Sogeti. Retrieved November 22, 2013 from http://vint.sogeti.com/fueling-sales-with-social-data-the-story-of-walmart-labs-the-social-genome/

Rijmenam, M. (2013, Sep 8). Big Data Offers the Chance for Retailers to Stay Ahead of Their Competitors. In Smart Data Collective. Retrieved November 22 , 2013 from http://smartdatacollective.com/bigdatastartups/145681/big-data-offer-retailers-chance-stay-ahead-their-competitors

Love, D. (2013, Jul 2). How To Use Big Data To Make A Fortune In The Retail Industry. In Business Insider. Retrieved November 22 , 2013 from http://www.businessinsider.com/big-data-in-retail-2013-6

Banks, B. (2012). Big Data for Retail is Flying Off the Shelves. In Forbes.com. Retrieved Novenber 22, 2013 from http://www.forbes.com/sites/sap/2012/05/11/big-data-for-retail-is-flying-off-the-shelves/

Mehra, G. (2013). 6 Uses of Big Data for Online Retailers. In Practical ECommerce. Retrieved November 22, 2013 from http://www.practicalecommerce.com/articles/3960-6-Uses-of-Big-Data-for-Online-Retailers

Wayner, P. (2012 , Apr 18). 7 Tools for Taming Big Data with Hadoop. In JavaWorld. Retrieved November 22, 2013 from http://www.javaworld.com/javaworld/jw-04-2012/120418-tools-for-hadoop.html?page=3

Nadkarni, A.,DuBois, L.(2013, Oct) .Trends in Enterprise Hadoop Deployments. In Redhat. Retrieved November 22, 2013 from http://www.redhat.com/rhecm/rest-rhecm/jcr/repository/collaboration/sites%20content/live/redhat/web-cabinet/static-files/library-assets/Trends%20in%20enterprise%20Hadoop%20deployments

Cloudera Enterprise.(n.d.).In Cloudera. Retrieved November 23, 2013 from http://www.cloudera.com/content/cloudera/en/products-and-services/cloudera-enterprise.html

Cloudera Product Comparison. (n.d.).In Cloudera. Retrieved November 23, 2013 from http://www.cloudera.com/content/cloudera/en/products/product-comparison.html

Morgan,T.P.(2011, Dec 08) Cloudera gets proactive with Hadoop management.In The Register.Retrieved November 23, 2013 from http://www.theregister.co.uk/2011/12/08/cloudera_enterprise_manager/

Cloudera. (2013). [Untitled image of Cloudera Enterprise Subscription Options]. Retrieved November 23, 2013 from http://www.cloudera.com/content/dam/cloudera/Resources/PDF/Cloudera_Datasheet_Enterprise_Subscription_Options.pdf

Vappalapati, C.(n.d.) .Big Data Presentation [PPT Document] .Retrieved from Lecture Notes Online Web site: https://sjsu.instructure.com/courses/1017137/files/29936593?module_item_id=6879279

Kornacker, M., Erickson, J. (2012, Oct 24). Cloudera Impala: Real-Time Queries in Apache Hadoop, For Real .In Cloudera. Retrieved November 23, 2013 from http://blog.cloudera.com/blog/2012/10/cloudera-impala-real-time-queries-in-apache-hadoop-for-real/

Cloudera Search User Guide.(2013, October 29). In Cloudera. Retrieved November 23, 2013 from http://www.cloudera.com/content/cloudera-content/cloudera-docs/Search/latest/PDF/Cloudera-Search-User-Guide.pdf

London,G.(2012, October 21).How-to: Set Up an Apache Hadoop/Apache HBase Cluster on EC2 in (About) an Hour. In Cloudera. Retrieved November 23, 2013 from http://blog.cloudera.com/blog/2012/10/set-up-a-hadoophbase-cluster-on-ec2-in-about-an-hour/

Brandon, J.(2013, Oct 29). Cloudera beefs up partnerships to bolster big data in the cloud. In Business Cloud News. Retrieved November 23, 2013 from http://www.businesscloudnews.com/2013/10/29/cloudera-beefs-up-partnerships-to-bolster-big-data-in-the-cloud/

Hadoop Ecosystem [Online Image].(2013, Jul 06). Retrieved November 23, 2013 from http://wishkane.wordpress.com/2013/07/06/hadoop-distributions-compared/

Weathington, J. (2013). Use Big Data for Marketing Accountability. In Tech Republic. Retrieved October 14, 2013 from http://www.techrepublic.com/blog/big-data-analytics/use-big-data-for-marketing-accountability/

How Mobile is Transforming the Shopping Experience in Stores. (2012). In Google.Retrieved October 14, 2013 from http://www.google.com/think/research-studies/mobile-in-store.html

Wähner, K. (2013, Jul 09). Spoilt for Choice – How to choose the right Big Data / Hadoop Platform?.In InfoQ. Retrieved November 19 ,2013 from http://www.infoq.com/articles/BigDataPlatform

[Online image for Alternatives for Hadoop Platforms]. Retrieved November 24 ,2013 from http://www.infoq.com/resource/articles/BigDataPlatform/en/resources/fig1large.jpg

[Online image for Hadoop Ecosystem]. Retrieved November 24 , 2013 from http://blog.briskgap.com/wp-content/uploads/2013/03/Hadoop-Ecosystem.jpg
Kelly, J. (2013, Sep 16). Big Data: Hadoop, Business Analytics and Beyond. In Wikibon. Retrieved November 22, 2013 from http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond

What is Big Data Analytics?. (n.d.). In IBM. Retrieved October 14, 2013 from http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html
Amazon EMR with the MapR Distribution for Hadoop. (2013). In AWS. Retrieved November 20, 2013 from http://aws.amazon.com/elasticmapreduce/mapr/

Winter, R., Gilbert, R., Davis, J.(2013). “Big Data: What does it really cost?” in AsterData. Retrieved November 20, 2013 from
http://www.asterdata.com/resources/assets/WinterCorp_Report_Big_Data_What_Does_it_Really_Cost.pdf

Burtch, L.(July 2013). “The Burtch Works Study: Salaries for Big Data Professionals” in Burtch Works. Retrieved November 21, 2013 from http://www.burtchworks.com/Burtch_Works_Study_Final.pdf

(2013) “Facebook APIs”. Facebook Developers. Retrieved November 22,2013 from https://developers.facebook.com/docs/reference/apis/

Thaploo, V.(March 11, 2013). “Hadoop Distributions Compared.” in Blazeclan. Retrieved November 22, 2013 from http://blog.blazeclan.com/252/

Doan, A., Lam, W., Liu, L., Prasad, S., Rajaraman, A., Vacheri, Z. (2012). Muppet: MapReduce-Style Processing of Fast Data. Very Large Data Base Endowment, Inc. Retrieved October 14, 2013 from http://vldb.org/pvldb/vol5/p1814_wanglam_vldb2012.pdf

Big Data Visualization: Turning Big Data Into Big Insights: The Rise of Visualization-Based Discovery Tools. Intel. (2013). Retrieved October 14, 2013 from http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-visualization-turning-big-data-into-big-insights.pdf

Hammerbacher. (2010, Nov. 9). “What Are the Advantages/Disadvantages running Cloudera’s distribution for Hadoop on EC2 instances rather than using Amazon’s Elastic Map Reduce service?” Quora. Retrieved November 23, 2013 from http://www.quora.com/What-are-the-advantages-disadvantages-running-Clouderas-distribution-for-Hadoop-on-EC2-instances-rather-than-using-Amazons-Elastic-Map-Reduce-Service?srid=sc

Nadkami, A. (October 2013) “Trends in Enterprise Hadoop Deployments” Red Hat. Retrieved November 23, 2013 from http://www.redhat.com/rhecm/rest-rhecm/jcr/repository/collaboration/sites%20content/live/redhat/web-cabinet/static-files/library-assets/Trends%20in%20enterprise%20Hadoop%20deployments

Robert Handfield (September 2013) “A Brief History of Big Data Analytics” Retrieved November 23, 2013 from http://iianalytics.com/2013/09/a-brief-history-of-big-data-analytics/

Jim Bell (September 2013) “Marketers Are Slowly But Surely Embracing Big Data Analytics” Retrieved November 23, 2013 from http://www.cmswire.com/cms/customer-experience/marketers-are-slowly-but-surely-embracing-big-data-analytics-022627.php

Bill Siwiki (March 2013) “ Amazon has a commanding lead among mobile department stores” Retrieved November 23, 2013 from http://www.internetretailer.com/2013/03/04/amazon-has-commanding-lead-among-mobile-department-stores

Jim Bell (September 2013) “Marketers Are Slowly But Surely Embracing Big Data Analytics” Retrieved November 23, 2013 from http://www.cmswire.com/cms/customer-experience/marketers-are-slowly-but-surely-embracing-big-data-analytics-022627.php

Philip Russom (Second Quarter 2013) “INTEGRATING HADOOP INTO BUSINESS INTELLIGENCE AND DATA WAREHOUSING” Retrieved November 23, 2013 from http://www.cloudera.com/content/dam/cloudera/Resources/PDF/TDWI%20Best%20Practices%20report%20-%20Hadoop%20foro%20BI%20and%20DW%20-%20April%202013.pdf

Ovum white paper for Cloudera “Hadoop: Extending Your Data Warehouse” http://www.cloudera.com/content/dam/cloudera/Resources/PDF/2013-04-02_Ovum_whitepaper_Hadoop_Extending_Your_DW.pdf

Jon Stokes (May 2012) “WALMARTLABS – TAKING BIG DATA INTO RETAIL ” Retrieved November 23, 2013 from http://www.freshminds.net/2012/05/walmartlabs-recommendation-big-data/

Bill Siwiki (March 2013) “Amazon has a commanding lead among mobile department stores ” Retrieved November 23, 2013 from

http://www.internetretailer.com/2013/03/04/amazon-has-commanding-lead-among-mobile-department-stores

JP Mangalindan (July 2012) “Amazon's recommendation secret” Retrieved November 23, 2013 from http://tech.fortune.cnn.com/2012/07/30/amazon-5/

Twitter. (2013) “About Twitter, Inc.”. Twitter About. Retrieved November 23, 2013 from https://about.twitter.com/company
Facebook. (2013). “The Graph API”. In Facebook Developers. Retrieved on November 14, 2013 from https://developers.facebook.com/docs/graph-api/

Michael Gorman (November 2013) “Pinterest's APIs let developers embed pins directly on their websites (updated)” Retrieved November 23, 2013 from http://www.engadget.com/2013/11/14/pinterest-embed-pins-search-api/

Facebook. May 1, 2013.”Investor Relations”. Facebook Reports First Quarter 2013 Results.Retrieved on November 23,2013 from http://investor.fb.com/releasedetail.cfm?ReleaseID=761090

Pinterest API (n.d.). In Programmableweb. Retrieved on November 23,2013 from http://www.programmableweb.com/api/pinterest

Lauren Orisini (November 2013) “Finally! Pinterest Unveils Its Public API” Retrieved on November 23, 2013 from http://readwrite.com/2013/11/14/pinterest-api#awesm=~oo66Thj6uyEFZ3

Jonathan Strickland “How Twitter Works” Retrieved on November 25, 2013 from http://computer.howstuffworks.com/internet/social-networking/networks/twitter2.htm

Jonathan Gordan, Jesko Perrey and Dennis Spillecke(July 2013) “Big Data, Analytics And The Future Future Of Marketing And Sales” Rerieved on November 24, 2013 from http://www.forbes.com/sites/mckinsey/2013/07/22/big-data-analytics-and-the-future-of-marketing-sales/

Apache 2013, Apache Flume, Retrieved on November 23,2013 from http://flume.apache.org/

John Natkins (September 2012), “How-to: Analyze Twitter Data with Apache Hadoop” retrieved on November 24, 2013 from http://blog.cloudera.com/blog/2012/09/analyzing-twitter-data-with-hadoop/

About Data Mining (January 2013). In aboutdm. Retrieved on November 23,2013 from http://www.aboutdm.com/2013/01/product-recommendation-by-amazon.html

Raven Zachary (May 2012) “Mobilizing our store customers” retrieved on November 23, 2013 from http://www.walmartlabs.com/category/apps/

Big Data Offer Retailers the Chance to Stay Ahead of Their Competitors (January 2013). In smartdatacollective. Retrieved on November 23,2013 from http://smartdatacollective.com/bigdatastartups/145681/big-data-offer-retailers-chance-stay-ahead-their-competitors