# Politecnico di Milano

*School of Industrial and Information Engineering*
*Master of Science in Telecommunication Engineering.*


## Agile Analytics: An exploratory study of technical complexity management

**Supervisor:** Prof. Chiara Francalanci
Ing. Paolo Ravanelli

**Candidate:**
Agnirudra Sikdar
Personal Code:10449950

**Academic Year 2014/2015**

# ACKNOWLEDGEMENT

Writing an acknowledgement is one of the very less traits which my Alma Mater, Polimi doesn't teach, but it makes me indebted to the colossal grandeur of the university till my last breath. I am extremely thankful to all my Professors for having given me this amazing opportunity. For having to bear my curious inquisitive minds. I want to thank the Department di Elettronica, Informazione e Bioingegneria of "**POLITECNICO DI MILANO***"* for giving me such a golden opportunity to help me reach till here.

I would like to express my deepest gratitude to my thesis supervisor "**Prof.Chiara Francalanci",** who encouraged me to set sail in this vast ocean of research. Her winds of constant, rays of motivation and sunshine of wisdom gave me ample support to strengthen my sails and redirected me to the unchartered territories. I would also like to thank my thesis co-ordinator **Ing. Paolo Ravanelli**, whose amazing expertise in my laboratory sessions gave me very strong insights into my area of research.

I would like to thank my colleagues, without whom life in Milan would have been pretty hard to come by. It was an amazing experience and I would cherish this friendship and moments till the day's end.

Last but not the least, I would like to be grateful to my parents, my grandparents, my aunt, uncle and my little cousin sister without their sacrifices, nothing would have been possible. Their constant support, motivation and encouragement acted like a guiding light in tough times.

Thank you everyone for the faith you entrusted on me. I hope I did not fail your expectation. I do want to express my profound gratefulness that I felt for being a part of this extended family and amazing journey.

# Contents

# *List of Figures and Tables*

# *CHAPTER I. INTRODUCTION*

Internet Technologies have come a long way from the quintessential Ethernet to Wireless Communication devices and much more. Along with this evolution, various technologies have evolved over time to satiate the needs of this immense network. With the advent of smartphones and the technologies associated to the Internet of Things, storing, managing and analyzing data has become a challenge.

Big Data Technologies is the answer to this challenge, and analytics is a part of this ecosystem to undercover the hidden secrets of this pool of Exabyte of data.

Agile Analytics is a great analytical methodology. Its approach is very lightweight and addresses the key concerns of improving efficiency, user experience, and performance while trying to reduce complexity. Agile analytics is emerging, mostly because this analytical techniques that caters to the present fast paced world where most of the companies compete against each other to reduce theirs cons and improve their pros. Agile also focuses on soft and managerial variables making it extremely flexible to work in both technical and non-technical domains. But unfortunately the technical guidelines into the implementation of Agile Analytics are still missing.

This thesis respects on an explanatory study, surveying public case studies to design an agile analytics method providing technical guidelines. Our aim is to fill the void of agile analytics into providing a methodology that can fulfill this objective.

We will be following an explanatory approach, studying multi case analysis to determine the root of the analytical problems and how it can be solved by implementing Agile. We will also try to determine from these cases, how many companies are using Agile and how it is being beneficial to them.

# *CHAPTER II. State of the Art*

## 2.1Definition of Big Data:

Big Data is typically referred to the following types of data:

- Traditional Enterprise Data- this includes customer information from Customer Relationship Management (CRM) systems, transactional ERP data, web store transactions, and general data from the ledger.
- Machine generated/ Sensor data- this includes Call Detail records (CDR), weblogs, smart meters, manufacturing sensors, trading systems data,
- Social Data- this includes customer feedback streams, micro-blogging sites like Twitter and social networks like Facebook.

FigThe McKinsey Global Institute has estimated the growth rate of data volume to be 40% per year, and it has been predicted to grow 44 times between 2009 and 2020. Although Volume of data is regarded as the most visible parameter, there are three more characteristics that define big data. **[1]**


*Fig 2.1 Big Data Tag Cloud*

**The four key characteristics of Big Data are as follows**:-

I. **Volume:** Data generated by machines have a much larger quantity in comparison to the non-traditional data.

II. **Velocity:** Social media data streams-is lesser in terms of quantity when compared to machine generated data. Produces a large influx of reviews and relationships, valuable to customer relationship management. Twitter for example handles data stream of over 8Tb/s.

III. **Variety:** Traditional data formats change slowly and tend to be well defined by data schema.

IV. **Value:** The economic value of various types of data vary significantly.

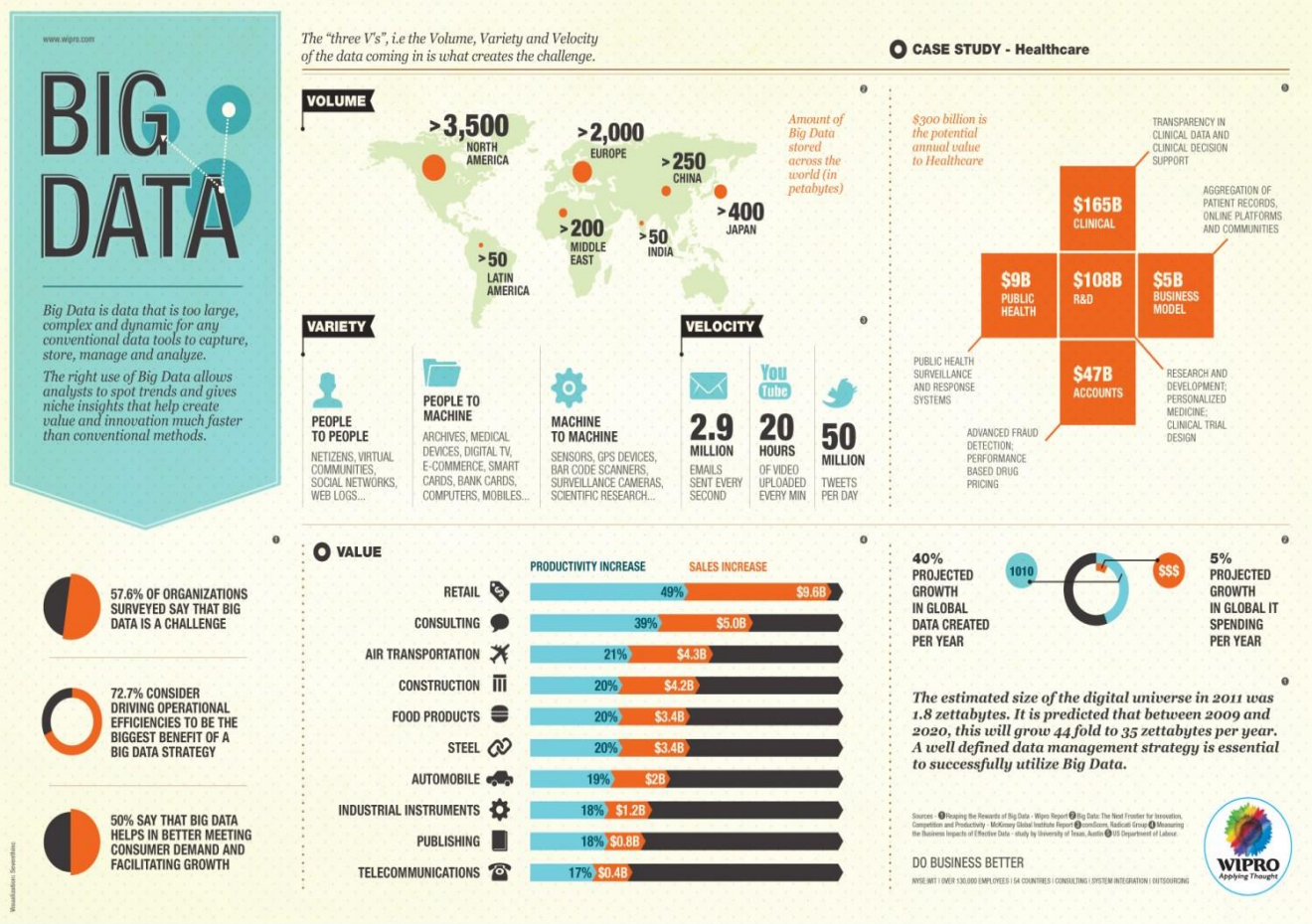


*Fig 2.2 Big Data charts.*

*"http://www.wipro.com/images/infographic.jpg"*

To make the most of these Big Data, companies must transform and evolve their IT Infrastructure in order to handle these large volume, high velocity, mammoth variety sources of data and integrate them with pre-existing enterprise data.

## 2.2 History of Big Data

More than 90% of the data that is available presently, has been created in the past three years. The term **Big Data**, was first coined in the year 2005, by Roger Mougalas from O'Reilly Media. Although the usage of Big Data has been lurking way before this period and the need to understand all available data has been much longer.

**The 20th Century**

The first major data big data project can be traced back to the time when the Social Security Act of USA became a law in the year 1937. The US government had to keep track of 26 million Americans. IBM got the contract for this project and developed punch-card reading machine for this massive bookkeeping project.

- The first data processing machine was developed by the British during the World War II to decipher Nazi codes. The device was called Colossus and it searched for patterns in intercepted messages.
- 1965 the US government decided to build the first data center to store 742 million tax returns and 175 million sets of fingerprints.

**The 21st Century**

- In 2005, Roger Mougalas from O'Reily Media coined the term Big Data, a year after they had defined the term Web 2.0
- In 2005 Hadoop was created by Yahoo! It was built on top of Google's MapReduce.
- In 2011, the McKinsey Report on Big Data stated that by 2018, USA alone will face a shortage of 140,000-190,000 data scientist along with 1.5 million data managers.

In the past years, there has been quite a massive increase of Big Data startups, all trying to tap into this hidden treasure trove of data. The migration to this new form of technology has been slow, but it is sure to pick up the pace and a new era of revolution will begin very soon. **[2]**

## 2.3 Big Data Technologies and Techniques

The techniques and technologies involved for the development of data science and BigData, primarily has the purpose of inventing more sophisticated and scientific methods for visualizing, analyzing, managing and exploiting informative knowledge from extremely diversified, distributed and huge datasets.

A paradigm shift is about to begin, as statistical techniques, coupled with mathematics and new data mining tools, along with the help of machine learning algorithms will help data scientists to analyze data and predict outcomes in a much efficient and faster time.

Some of the emerging technologies associated to big data, as stated by Dr. Satwant Kaur, regarded as ,"The First Lady of Emerging Technologies." According to her the following are regarded as the rising technologies in Big Data:

- Schema.
- MapReduce
- Hadoop
- Hive
- PIG
- WibiData
- PLATFORA
- SkyTree

Current tools focus on three classes, specifically, batch processing tools, interactive analysis tools, and stream processing tools. Most of the batch processing tools are based on Apache Hadoop architecture, like Dryad and Mahout. Google's Dremel and Apache

Drill are Big Data platforms based on interactive analysis. Multidisciplinary tools and methods are needed to explore and mine information from Big Data.

Big Data techniques involve quite a lot of disciplines from data mining, machine learning, neural networks, social network analysis, optimization methods, visualization approaches, statistics, signal processing and pattern recognition. **[3]**

**[4]**

| **Name** | Specified Use | Advantage |
|---|---|---|
| **Apache Hadoop** | Infrastructure and platform | High scalability, reliability, completeness |
| **Dryad** | Infrastructure and platform | High performance distributed execution engine, good programmability |
| **Apache Mahout** | Machine learning algorithms in business | Good maturity |
| **Jaspersoft BI Suite** | Business intelligence software | Cost-effective, self-service BI at scale |
| **Pentaho Business Analytics** | Business analytics platform | Robustness, scalability, flexibility in knowledge discovery |
| **Skytree Server** | Machine learning and advanced analytics | Process massive datasets accurately at high speeds |
| **Tableau** | Data visualization, Business analytics | Faster, smarter and ease of use dashboards |
| **Karmasphere Studio and Analyst** | Big Data Workspace | Collaborative and standards-based unconstrained analytics and self service |
| **Talend Open Studio** | Data management and application integration | Easy-to-use, eclipse-based graphical environment |

*Table 2.1 Big Data tools based on batch processing*

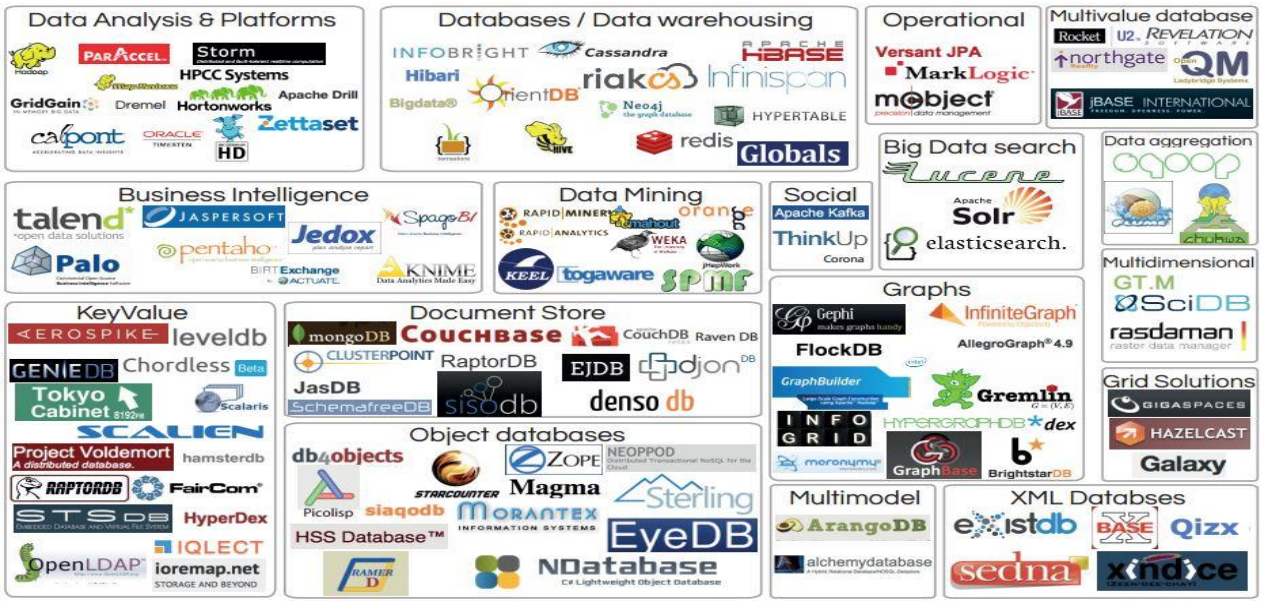


*Fig 2.3 Big Data Landscape*

## 2.4 Applications of Big Data

Big Data of late has experienced a mammoth growth in finding its potentials being applied in various domains. Applications of Big Data have increased significantly and it is slowly getting integrated into various domains of our everyday life. Its emerging potential has given the scope to exploit its features. Some of the noted applications of Big Data are stated as follows

1) **Government** uses Big Data analytics to process information that is efficient and beneficial to them in order to improve cost, productivity and innovation.
2) **Manufacturing** industry benefits from the transparency provided by its infrastructure. This helps to improve in supply planning and product quality.
3) **Retail.** Retail giants like Walmart, Alibaba, Amazon, Ebay are using Big Data to process and target their customers using their business intelligence in order to increase their profits.
4) **Information Technology** This vast domain has a lot of Big Data used cases from Internet of Things to E-commerce, and many others.

5) **Medicine.** IBM is using Big Data technologies to predict heard diseases. Analysis of electronic health record can reveal symptoms at an earlier stage. Apache Unstructured Information Management Architecture (UIMA) is used by IBM to extract signs and symptoms of heart failure from the data.
6) **Media.** From advertising to marketing industry to social networks, the presence of big data can be evidently seen.
7) **Science.** One of the most research intensive fields which demands huge data processing tasks in order to process the various kinds of data from Astronomy to Engineering and much more. Google's DNAStack is one of the applications using Big Data methodologies to study human genomes. **[5]**

**[6]**

| Company | Industry | Type | Purpose |
|---|---|---|---|
| **T-Mobile** | Communication | Optimize Funnel Conversion | T-Mobile uses various indicators like billings and sentiment analysis to identify customers to offer premium services. |
| **Master Card** | Finance | Behavioral Analytics | Master Card uses Behavioral Analytics to analyze their 1.8billion customer base. |
| **Walmart** | Retail | Customer Segmentation | Walmart combines social data, public data and internal data to monitor the reviews of customers and their friends. They use this data to target messages about specific products and also offer special discounts |

| | | | |
|---|---|---|---|
| **Morgan Stanley** | Finance | Predictive Support | Morgan Stanley uses real time data analytics to identify problems and prioritize which issues should be addressed. |
| **Etihad Airways** | Travel | Market Basket Analysis & Price Optimization | Etihad uses big data to determine which destinations should be added to maximize revenue. |
| **Amazon** | Online Retail | Predict Security Threats | Amazon has over 1.5 billion items in its catalogue. It uses cloud system S3, to predict which items are most likely to be stolen, in order to better secure its warehouses. |
| **Discovery Health** | Insurance | Fraud Detection | Discovery Health uses big data analytics to identify fraudulent claims. |
| **Shell** | Oil | Industry Specific | Shell uses sensor data to map its oil and gas wells to increase its output and improve the efficiency in its operations. |

*Table 2.2 Some of the used cases of Big Data are tabulated as follows.*

## 2.5 Introduction to Analytics

Analytics is defined as the discovery and communication of meaningful patterns in data. It is extremely valuable in areas rich with documented information. Analytics relies on documented simultaneous application of computer programming, operations research to statistics. Analytics is a multidimensional discipline. The insights obtained from data are used for recommending actions or to assist in decision making. It is not very much concerned with individual analysis steps, but with entire methodology. There is an increasing use of the term advanced analytics, mostly used to describe the technical aspects of analytics. For example, advanced analytics is used in machine learning technique like neural networks to perform predictive modelling. **[7]**

### 2.5.1 What is Big Data Analytics?

Big data analytics is the process of examining large data sets containing heterogeneous data to uncover hidden patterns, unknown correlations, customer preferences, market trends and other useful business information. The analytical results can lead to improved efficiency, competitive advantages over rival organizations, better customer services and many other business benefits.

The primary goal of big data analytics is to help companies make more informed business decisions by helping data scientists, predictive modelers and other analytics professionals to analyze data, or other untapped data by conventional business intelligence programs. Big Data can be analyzed with various software tools used mostly as a part of advanced analytics disciplines in data mining, text analytics, statistical analytics, etc. Mainstream business intelligence software and data visualization tools can also play a role in the analysis process.

The business cases for leveraging Big Data are compelling. For example, Netflix mined its subscribers' data to put the best recommendation upfront the subscriber. This made their recommendation model a big hit and allowed them to reap the benefits of Big Data analytics. **[8]**

Enterprises are increasingly looking to find actionable insights into their data. With the right set of tools, skills and expertise, enterprises can boost sales, increase efficiency, risk management and even customer service by the potentiality of Big Data Analytics.

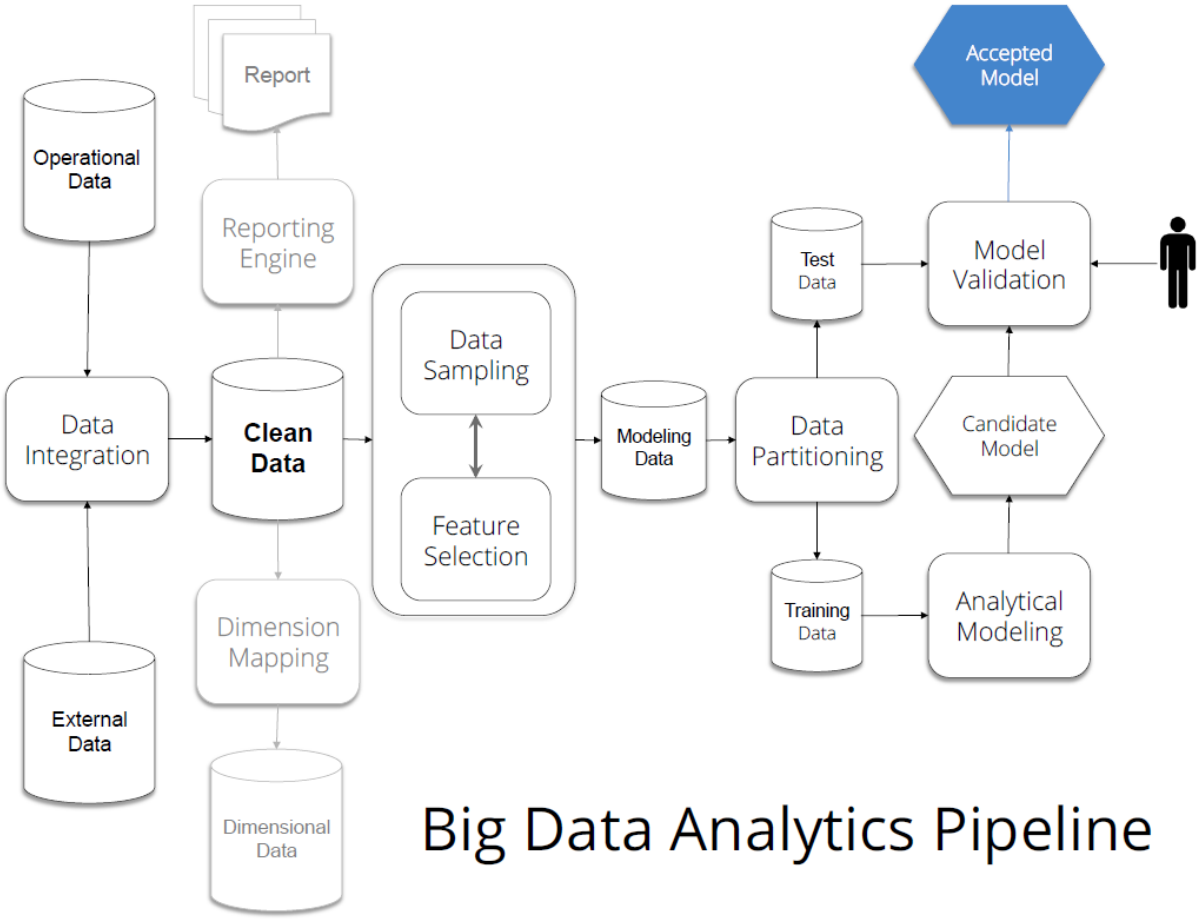


*Fig 2.4: Big Data Analytics Pipeline model*

## 2.6 Agile Analytics

## 2.6.1 What is Agile Analytics?

Agile Analytics includes practices for monitoring, management and project planning; for effective cooperation between stakeholders and business customers and also ensuring technical quality by the delivery team. **[9]**

Agile is a very specific word, reserved to describe a kind of development style. It means something very explicit. Sadly, it is often misused and misunderstood as a name for processes that are ad-hoc, lacking in discipline and slipshod. Agile depends more on discipline and consistency; although it's not a tough and a formal process despite several methodologists trying to code it with various approaches. It falls rather between flexibility and structuring. Agile is built on a set of meaningful values and principles, and requires high set of degree of discipline in order for it work properly.

Bringing agility into big data analytics has been a challenge to many talented data scientists and engineers. The motives are similar to the effort in adopting agile application software development: mostly in team dynamics and organizational culture. In case of agile analytics, often stakeholders proceed beyond IT solutions to include marketing and other domains.

**Agile analytics is all about failing fast.** This can be explained in terms of scientific context, where we first state a hypothesis and then seek to disapprove it using quantitative analysis of real data in rapid cycles. Agile focuses not on the data itself but on the insight and action that can ultimately be drawn from nimble business intelligence systems. Rather than beginning with investment and platform building, Agile analytics starts with learning and testing, so that companies can build their models and strategies based on solid answers to their most crucial business questions. The final solution is actionable insight and maximized value to the business. **[10]**

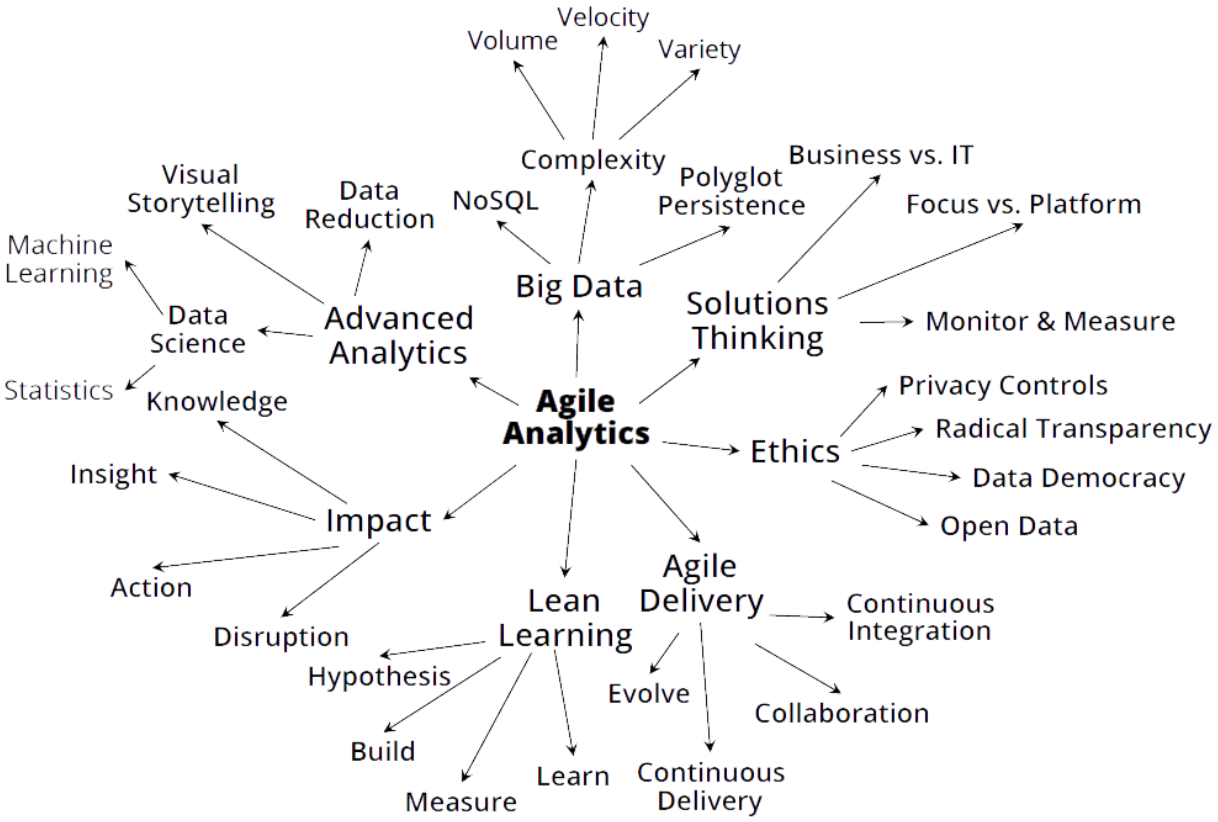


*Fig 2.5. Various influences of Agile Analytics*

## 2.6.2 How Agile Analytics is linked with the Scope of this Thesis?

The objective of this thesis is understand the how Agile Analytics can help to improve the efficiency in Big Data Analytics. To specify it more, the main purpose is to find the complexities in conventional Big Data analytic techniques and implementing Agile Methodologies to address these issues. The reason for which Agile is considered for the scope of this thesis, is because of its high flexibility to address the problems and restructure its modules in order to resolve them.

Data complexities arises from taking up various approaches and choice of selection of methodologies. With Agile, we will try to understand its flexibility and try to find out if taking

up this approach will improve the efficiency of the system. We will try to concentrate on Customer segmentation and clustering cases and will try to see what kind of complexities are involved. Following this, we will try to determine this solve the problems using Agile approach.

## 2.6.3 Practices of Agile Analytics

- **Adapt to Changing Conditions**

The core purpose of big data is to find key insights upon which an organization can pivot. In this way, big data by definition demands agility. Listen to what users, tests and business conditions are telling you, and work change into subsequent iterations.

- **Automate as Many Processes as Possible**

The greatest manpower should be saved for developing new features and collaborating across organizational and development teams. As such, it's important to automate as many regular processes as possible, from testing to administrative tasks so that developers can focus intensely on an iteration's set goals.

- **Foster Self-Organized Teams**

Hire talented, motivated individuals who can set their own goals for each iteration and function as effective self-managers. Then, trust them to do the job at hand, self-monitoring and adapting as they go.

- **Adapt Agile Methods to Individual Projects and Teams**

While Agile analytics has many guidelines, it is a style, not a process. More traditional tactics aren't antithetical to Agile if they're effective in achieving iterative goals. Choose the tactics that work best for each project and team rather than adhering to static rules.

- **Conduct Regular Reviews of Processes**

Agile systems development requires just as much discipline and rigor as the traditional waterfall method in order to stay on track. However, rigor should be applied not to adhering to rigid systems and static goals, but to constantly re-evaluating the effectiveness of the methods and styles at hand.

- **Constantly Learn**

Keep up-to-date with the best data warehousing and business intelligence practices and implement them fluidly into each iterative phase. This will substantially increase the development team's agility and keep the organization ahead of its competitors

## 2.6.4 Agile Methods [11]

| Methods/Frameworks | Description | References |
|---|---|---|
| **Adaptive software development (ASD)** | It grew from Rapid Application Development. Follow the principle that it's normal to adapt continuously the process to the work at hand. | "MESSY, EXCITING, AND ANXIETY-RIDDEN: ADAPTIVE SOFTWARE DEVELOPMENT" by Jim Highsmith |
| **Agile modeling** | It's a practice based methodology to perform effective modelling and documentation of software based platform. | http://www.agilemodeling.com/ |
| **Agile Unified Process (AUP)** | It's a simplified version of Rational Unified Process describing an easy way to | Scott Ambler |

| | | |
|---|---|---|
| | develop business application software | |
| **Business analyst designer method (BADM)** | It's a method for designing business change where there is significant change in IT system. | Heap, Tony. "Business Analyst Designer Method".http://www.its-all-design.com/. |
| **Crystal Clear Methods** | Methodology which prioritized for project safety. Describes the methodology of the lightest, and most habitable kind that will produce good results. | Alistair Cockburn |
| **Disciplined agile delivery** | Process decision framework that enables simplified process decisions around incremental and iterative solution delivery. | Scott Ambler |
| **Dynamic systems development method (DSDM)** | DSDM is an iterative and incremental approach that follows Agile Development principles, including continuous user/customer involvement. | Keith Richards, *Agile project management: running PRINCE2 projects with DSDM Atern.* OGC - Office of Government Commerce. The Stationery Office, 31 Jul. 2007 |

| | | |
|---|---|---|
| **Extreme programming (XP)** | It's a software development methodology with the intention to improve software quality and responsiveness based on the change of customer requirements, | Kent Beck |
| **Feature-driven development (FDD)** | FDD blends a number of industry recognized practices into a cohesive whole. It is also an Agile Method of Developing software and is lightweight in nature. | Jeff De Luca |
| **Lean software development** | It is a translation of lean IT principles and lean manufacturing. More popular with startups who want to test the market. | Mary Poppendieck and Tom Poppendieck |
| **Kanban (development)** | This method focuses on managing knowledge work and emphasizing on just-in-delivery. | David Anderson |
| **Scrum** | Scrum defines a flexible, holistic product development strategy where a development team works as a unit to reach a common goal. | Hirotaka Takeuchi and Ikujiro Nonaka in the *New Product Development Game*. |
| **Scrumban** | This methodology is more of a hybrid of Scrum and Kanban. | Corey Ladas |

*Table 2.3 Agile Methods and their description*

## 2.7. Conclusion

We can come to conclusion that Big Data is a pretty important technology with immense potential. The need for proper analytics to uncover this hidden data field is extremely important. The key to this Pandora's Box lies not in our hands but in the approach we follow to open it.

For Agile Analytics the managerial variables are clear while the technical ones are not yet addressed. But they should be, since the agile methodology typically involves both the components we will try to study about it in the next chapters with the primarily concern of trying to determine the technicalities which Agile Analytics can address and try to identify the different complexity problems that arises from reviewing of multi case studies and trying to propose methods by which we can improve the efficiency and solve these problems.

# *CHAPTER III. ANALYTICS & IT'S CASE STUDIES*

## 3.1 Introduction

Analytics is an integral part of the Big Data ecosystem, and it demands proper algorithms and modelling approaches to solve different issues. Efficiency at solving the problems and the importance of using the right set of approaches and tools determines the final outcome.

In this chapter we will try to focus on reviewing multiple case studies and would try to figure out the type of modelling approach they used. Also we will try to investigate the presence of Agile Methodology in all the cases.

## 3.2 A Detailed Description about Analytical modelling

**[12]** Why do we need Analytics in Big Data?

The production of data from conventional enterprise and non-enterprise sources are revealing innovative new channels of utilization from the confluence of various factors.

i) Explosion of a new source of raw and unstructured data.

ii) New prescriptive analytic technologies using very lower price-points compute-intensive resources.

iii) Innovation opportunities are opening up from this vast treasure trove of data which is yet to be mined for its benefits.

To design a Big Data system, with the correct recommendation technology, lots of considerations are needed to segment this vast voluminous data of varied structure Only

after this , proper action can be taken for every individual business scenario.

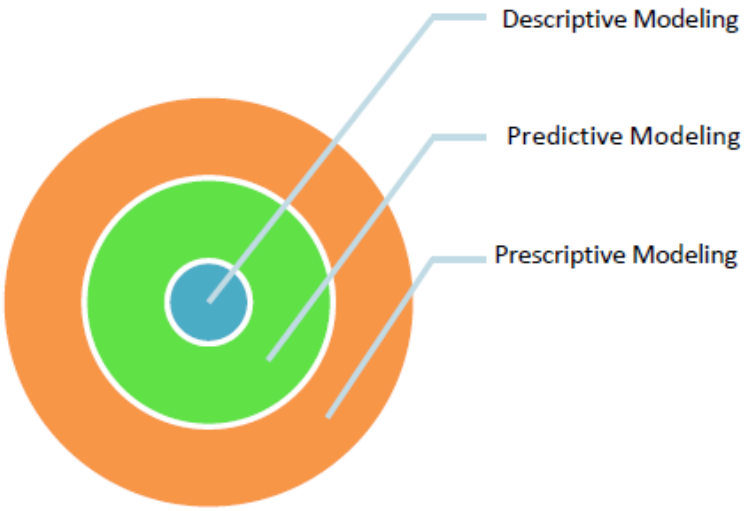


Fig.3.1 Components of Modelling

## 3.2.1 Descriptive Vs Prescriptive Modelling

Concepts of 'Descriptive' models versus 'Predictive' models are being used to design variables that define a client, systematically adequate to provide better predictive powers over the populations they represent. **Descriptive** analytics looks at past performance and comprehends that performance by mining past information to look for the motives behind previous success or failure. Nearly all structured reporting such as sales, marketing, operations, and finance, uses this kind of post-mortem analysis. Classical descriptive approaches (usually seen in data warehousing systems) like this are influential for describing the population but are limited in their predictive abilities with respect to the same population. **Predictive** models select major characteristics differently- i.e. while we may know a lot about our customers, we may not be able to precisely estimate what they will do next. In these instances applications of predictive models might help construct a list of clients for a study that will be more accurately target its lift ratio when applied to the real population.

Three elementary foundations of predictive analytics are:

- Predictive modeling
- Decision Analysis and Optimization

 Transaction Profiling

**Prescriptive** models though, take advantage of the data of descriptive models and the premise of predictive models and try to answer, not only what the client will do next, but why they will do so. This kind of cross analysis opens up investigations into questions not expected earlier in the design cycle, but need to be asked as the severe altering real-time requirements of the dynamic customer system.

## 3.2.2 Look-Alike Modeling

This is a specialized application of predictive modeling. It forces the sampling of a population such that the so-called 'salient' variables are selected as a means to represent the full population. Salient variables are those attributes of a customer that allows for accurate extrapolation with statistical rigor. Once a 'good' or behaviorally-predictable customer is identified, the objective is to find others like him. This is the basis of 'Look-alike' modeling where the target-marketed group (e.g. for a marketing campaign, product offering etc.) is an expanded list of members whose profiles look like the predictable member of the population. The considerations behind such modeling, is relaxing the variables to allow for a greater number of members into the 'Look-Alike' group without compromising the group's salient characteristics. A diverse set of techniques/disciplines that can be used include:

- Real-time Decision-Making
- Tree Map Analysis
- Cluster Analysis
- Variable Selection
- Multivariate Testing

## 3.3.1. Case studies of Analytical Data Solutions [13]

**Company:** Analytical Data Solutions are a data management consultancy company focused at delivering analytically efficient Customer-Market Intelligence solutions.

**Location:** London, United Kingdom

| CLIENT | BUSINESS ISSUE | SOLUTION | IMPACT ON BUSINESS | CUSTOMER INTELLIGENCE |
|---|---|---|---|---|
| US based high end luxury retailer of apparels. | 1) Identifying most valuable customers within existing customers, on basis of their buying patterns. 2) Assess the profiles of such potential youth customers. | **Customer Segmentation based on identification of demographic and behavioral sweet spot and relationship quotient.** | The case study helped to create not only the Best Customer Profile and characteristics, but also helped the retailer to create effective CRM program. | **Identifying Most Valuable Customer.** |
| Global Marketing Organization of leading personal Computer manufacturers | PC manufacturer wanted to use market sentiments and buzzwords in social media during early launch days to predict growth of products | **Data crawled from various websites like Facebook, Twitter, Amazon, Google, to create predictive indices around the market** | The solution helped them to have initial insights on assessing performance of a product and react quickly using potential corrective actions. | **Prediction of New Product Sales Trajectory using Social Buzz and Media Sentiments.** |

| | | | | |
|---|---|---|---|---|
| | | **sentiments to determine a potential trajectory for the sales of the product.** | | |

*Table 3.1 Case Studies of Analytical Data Solutions*

## 3.3.2. COLOMBUS FOODS: CASE STUDY [14]

**Company:** Colombus Salame (Colombus Sausage Company, Colombus Foods, Inc) is a food processing company based in America who specializes in Salami and other sorts of processed meats. It was founded in San Francisco in 1917.

**Location:** Hayward, California

**Business Issue:** Colombus food was finding it pretty difficult to manage its procurement, inventory management, logistics and manufacturing. The only solution to solve this problem was to acquire timely information to in order to make good decisions. They needed quick information in order to perform effective decisions. Operational data was expected to be extracted from various systems. **It would take upto two days to create a report to compare current inventory levels and production schedules against customer orders, and by the time information will be ready, it would be outdated.**

**OBJECTIVE:** Improve response on Operational Services during attending client demands.

**Approach:** IBM TM1 Cognos had been used to analyze financial data and support the forecasting, planning and budgeting processes. Applied Analytix team was outsourced to solve the problem of Colombus Foods, and they used IBM Cognos TM1 to address and solve the business issues of Colombus Foods.

INSTRUMENTED: Data from various business sectors like business, sales, procurement, manufacturing, inventory and logistics. These are collected and merged in IBM Cognos TM1.

INTERCONNECTED: Instant real time reports can be generated on stock levels, production schedules, order status, helping decision makers to improve their efficiency as per client demands.

INTELLIGENT: Forecasting the demands, preparing for delivery and production, and logistics to on-time delivery, order fulfillment and customer satisfaction.

Acceding to one of the employees of Colombus Foods, Dave Siegfried, " Cognos TM1 helps us to avoid inventory shortfalls by forecasting customer demand based on historical analysis of buying patterns and integrating information from out sales teams to predict how much each customer is likely to order in the coming months"

**CONCLUSION:** With a better understanding of demand volumes and production scheduling, the company was able to improve and optimize its inventory management, this reduced the risk of manufacturing surplus products which needed to be stored for a longer periods and before getting scrapped.

## 3.3.3. CORONA DIRECT Insurance Company. [15]

**Company:** Corona Direct is considered as the second largest direct insurance company in Belgium. It has over 150 employees and generates revenues of $69.3 million in 2009. It's a subsidiary of the Belgian-French banking and Insurance company called DEXIA. Corona Direct provides customers with products such as medical, fire, property and car insurance.

**Location:** Belgium

**Business Issue:** To sustain acquiring new customers in successive years from the revenues generated in the previous years to fund for the marketing campaigns. In the

past the cost of acquiring new customers exceeded by 50% of its revenues making Corona Direct's strategy at risk.

**Solution:** Corona Direct resorted to IBM SPSS predictive analytics software. The software helps them to efficiently create, optimize and execute their outbound marketing campaigns. Also using the predictive analytics, they could identify groups with the highest potential to respond to the campaigns. Along with this the company was able to perform profit-cost analysis-balancing growth targets against profit margins. Using all these points, Corona Direct was able to optimize its potential for growth.

**Results:** First year revenues covered campaign costs, enabling Corona Direct to adhere to its growth strategy.

1) Campaign costs reduced by 30%.
2) Long term customer profitability increased by 20%.
3) Product sales increased significantly.
4) Payback for the cost of implementation was achieved within six months.

The success of using IBM SPSS analytics has motivated Corona Direct to use predictive analytics to increase cross-selling within its call centers, thus creating profitable retention campaigns focusing on high-value customers. Also they plan to use past data to analyze and detect fraudulent claims, reduce false claim costs and eventually increase profits.

## 3.3.4. MEDIAMATH: CASE STUDY [16]

**Company:** Mediamath, first and largest, demand side platform (DSP) enabling company that enables advertisers to access and tap into 13 billion ad impressions daily. Software as a Service (SaaS) is provided in terms of automated media trading platform application, involving powerful analytical tools and simple workflow. They are the only Network Advertising Initiative (NAI) compliant DSP.

**Business Challenges:** Need best data analysts to enable the largest ad buyers to use all the information needed from every aspect for their ad campaigns. Requirement to select and deploy new solution within 3 months with minimal internal resources.

**Approach:** MediaMath used and outgrew MySQL. They tried with Oracle Standard Edition with 5TB of data and yet couldn't sustain the demand. According to one of the employees, Tom Craig, VP of Information Strategy at MediaMath, the code was fast but was difficult to maintain. They wanted something which was deployable very quickly as well as scalable to meet the demands of the business. From considering a set of software like Aster Data, Hadoop, Infobright, Oracle, Vertica, and Teradata, they finally decided to stick to Netezza.

This was because Netezza offered best return of investment with its ease of use. (The potential to use Agile Methodology was pretty much present so that the scalability in the future was absolutely prominent). Netezza provided multitude of rows and dimensions for analyzing the data.

**Applications:** MathClarity ™ interface was used for analytic insights, customer segmentation and audience targeting. . Some of the features of Netezza are mentioned as follows:

- Purchase funnel Analysis.
- Internal/Financial reporting.
- Dynamic interval reach and frequency.
- Fast decision making algorithms can be implemented.
- Near real time reporting and attribution
- Deep site analysis and classification.

**Conclusion:** With Netezza which powers the algorithmic trading, MediaMath was symbiotic enough to listen and act on more of these opportunities for improving the campaign performance. The two delivers the impression of real time level bidding. It sees and knows where every impression originated and thus helps the data scientists to create effective machine learning algorithms coupled with the best performance in this DSP domain.

## 3.3.5. MUELLER INC: Case Study [17]

**Company:** Mueller Inc, founded in 1930s is a leading retailer and manufacturer of pre-engineering metal roofing and metal building products. Their customer base is mostly in the southwestern United States from Texas, Louisiana and Okhlahoma.

**Location:** Ballinger, Texas, United States.

**Objective:** Metal Construction manufacturer Mueller wanted to shift their business strategy from manufacturing to retail led manufacturing. They wanted to adopt a more retail oriented business model.

**Approach:** Mueller Inc worked with IBM to deploy their analytics software, IBM® Cognos® Business Intelligence. IBM Cognos Business Intelligence gave a clear picture of the transformation path against a common set of Key Performance Indicators [KPIs], along with Cognos Metric Studio helped to understand the analytics to improve the performance analytics part. Using Metric Studio in Cognos Business Intelligence, an idea about the team's strategy performance was clearly visible.

The new Cognos gave the opportunity to customize interactive dashboard thus helping the analysts to have better insights into the business. Cognos TM1 gave a better understanding of the retail business, from ground zero. It helped Mueller Inc to understand what they needed to change in their immediate future to meet the growth targets. (**This shows signs of implementing Agile Implementation**) .

Presently the company plans to use IBM SPSS ® to tap into the Big Data Domain to mine enormous volumes of transactional data, which might reveal patterns and trends to help them predict the growth as well as expose unseen problems in the operations.

According to Mark Lack, in the past they needed to update a cube with new data by rebuilding it. With Dynamic Cube in Cognos Business Intelligence, they could build the cubes incrementally, so with each update only requires to lad new and changed data instead of rebuilding from the scratch. (**Again signs of Agile Implementation)**

**CONCLUSION:** The new IBM analytics software helped Mueller not only to adopt and be successful in their business model but also helped them to become a truly information

driven enterprise. Being a Small Medium Enterprise, but working smarter helped them to gain amazing insights for implementing their business strategies.

## 3.3.6 .TOP TOY Case Study [18]

**Company:** Top-Toy is the largest toy business in the Northern European market. They have an annual turnover of €500 million and 4,000 employees, TOP-TOY operates approximately 300 retail stores under its BR brand in the Scandinavian countries, as well as the licensed Toys “R”.

**Location:** Tune, Denmark

**Business Issue:** Top-Toy wanted that its warehouses would be stocked with the most popular toys but the long lead times made it difficult to efficiently manage the inventory. Since they ship products from China, it is required to plan about stocking the inventory months ahead of the busiest seasons. ERP was used, along with basic Spreadsheet applications, which failed to provide deep insights about filling up the stocks and addressing this business issue. The main objective was to predict and overcome the demand forecasting challenges.

**Solution:** To gain better and deeper insights into the predictability, Top-Toy decided to go for IBM Cognos Business Intelligence. Shifting to IBM® Cognos ® Business Intelligence helped the company to improve its inventory forecasting and made them align their orders on basis of their customer demands, several months in advance. (**Implementation of customer segmentation on basis of buying-selling patterns)**

With IBM Cognos , Top-Toy was able to create a data warehouse for its historical ERP, and connected Cognos to its forecasting-demand systems.

If Cognos showed that a product is a big hit with customers, the various space management teams might even provide various promotional benefits on it, thus improving the sales.

**Conclusion:** IBM Cognos Business Intelligence helped to resolve their business issue of meeting the customer demands and forecasting the demands for the inventory in the warehouses.

## 3.3.7. VAASAN FOOD GROUP CASE STUDY [19]

**Company:** VAASAN Group is one of the leading bakery operators in Northern Europe. Their business involves producing fresh bakery goods, crisp breads for sale in retail chains, restaurants and hotels as well various bake-off products. In 2011, VAASAN Group had employed 2,730 people, and achieved net sales of approximately €408 million.

**Location:** Helsinki, Finland.

**Business Issue:** VAASAN Group wanted to deliver the right products to the right customers in the proper time across the Nordic region. This required the company to predict and be dynamic to the rapid customer demands, adjust production and delivery schedules accordingly. VAASAN faced a business risk due to not having a standardized customer demand to predict the fluctuations.

**Approach:** VAASAN OY decided to create a solution for its demand planning that would analyze data from multiple sources after combining them, including the ERP system of the company. VAASAN Oy outsourced their business problem to Invenco Oy to help them build a solution. It involved sophisticated IBM Cognos Data Manager. Business Benefits were seen to rise by 30% in Sweden. (They tapped into rapid and effective response to market changes and was able to react quickly to the demands. **Agile Approach can be seen to have been used here)**

VAASAN team worked closely with their customers to gain insights into the location, size, sales history and used these variables to model their demand forecasting solution into their existing demand planning model.

The solutions generated gave VAASAN ample lead time to plan ahead and adjust the manufacturing schedule at its bakeries. This significantly increased their production volume. According to their Development Director, Nina Tuomikangas, if the market

situation changes, the modelled solution gives the production team an excellent understanding of changing the demand patters.

**Conclusion :** With the use of IBM Business Analytics solution, and implementing predictable models (with incorporating Agile Approach ), the company was able to solve their business issue of meeting the supplying demands on time.

## 3.3.8. CASE STUDY by Deloitte [20]

**Company:** Deloitte is a British Multinational company offering services in audit, tax, consulting, enterprise risk and financial advisory. With over 200,000 employees and a net revenue worth US$ 35.2 billion (2015), it is ranked as one of the best management consulting companies in the world.

**Location:** New York, USA.

**Client:** A major Telecom company.

**Business Issue:** The Company was struggling to process data of its customers for future analytical purpose. Previous attempts to migrate into Big Data technologies had failed. The company wanted a scalable big data and advanced analytics system in an efficient, cost effective way. The main challenge was to gain the fastest access into the data and deriving insights from it. :

**Approach:** The organization instead of migrating all of their systems into Big Data technologies at the same time, decided to select some Pilot projects and try out the potentiality of the system.

1) **Leadership Alignment**: The goal of the client was to align key stakeholders without getting caught up in details.

2) **Selection of Big Data Analytics Pilot Projects:** Deloitte helped to identify a mix of used business cases to demonstrate the big data insights and how it can improve the reduce costs.

3) **Design and Execution:** The teams adopted an **Agile implementation Process,** to meet and ensure the solutions that were being developed met the business needs.

**Results:** The team effectively implemented HADOOP as a solution. This enabled the team to easily process, analyze and even modify their data to access relevant information.

**Conclusion:** The pilot project were the first of a kind in the company. Now they are trying to develop advanced analytics program to help individuals much faster in order to take smarter business decisions.

## 3.3.9. Case Studies of PriceWater House Cooper (PwC) [21]

**Company:** PriceWater House Cooper is a multinational professional services company. Being one of the largest four auditors in the worlds. It has also been ranked as the most prestigious accounting firm in the world. Spanning over 157 countries, with more than 208,100 employees, PwC has total revenues worth $35.4 billion (2015).

**Location:** London, United Kingdom

**CASE I**

**Client:** A large software company with tens of millions of users.

**Business Issue:** The client wanted to improve its products and customers' online experience. The company earlier depended on 3rd party organization for various key data requirements, which were changing very fast .This unpredictable nature was the fundamental problem to modelling a predictable solution for keeping the products up-to-date and bringing important software updates on time.

Client wanted to balance their data storage and processing using cloud-based services since, it would reduce cost on infrastructure maintenance. (Being a seasonal company, most of their infrastructure was kept unused throughout the year).

Client's online product generated vast amount of data about the interactions with customers. They wanted to know ways to mine important information from these vast treasure trove of data to improve their product.

Client sought the help of PwC regarding the feasibility of migrating to Big Data Technologies and gaining benefits from it.

**Approach:** PwC realized after assessing from Pilot stage project that the client's data architecture model would be necessary for modifications, as well as they needed a proper technology roadmap. PwC also recommended using open-source tools and techniques, which would add flexibility, resulting in shorter software development cycles and faster release speeds to the market (**Agile Approach has been used here)**.

**Conclusion:** A 30% improvement in the speed-market. Improved technology infrastructure during peak hours. Lower costs in retiring of software licenses, and an improved user experience helped benefit a lot.

**CASE II [22]**

**CLIENT:** Canada's one of the leading pension firm.

**Business Issue:** to plan and support their information management system with the 5 year road path in consideration**.** The company was interested in making its data analysis and business intelligence more efficient and robust.

They identified 2 functional areas:-

1) Employee Productivity

2) Monthly members and Pensioner's report.

Company wanted to be able to explore and analyze data contained across systems, and view data on a time series basis, in order to see how metrics changed over time.

**Solution:**

PwC proposed **Agile analytics** to help the company address key issues related to the productivity measurement & monthly reporting. PwC provided the clients with working dashboards that addressed many of their immediate concerns.

a) Providing Business Staff with Self-serve raw data extraction capabilities.

b) Establishing common toolset for data preparation and data visualization tools shared across business.

c) Embedding technical data analysts within business units.

**Conclusion:** The client was convinced with the results and decided to migrate to Agile Analytics Methodologies. They also recognized the factors needed to address in order to implement a proper sustainable agile analytics function within their company. The agile methodologies gave the client. Using Agile enabled a low risk of, “Fast Fail,” approach to develop solutions. It also reduced financial risks associated with developing systems, allowed stakeholders to provide their active feedbacks and implementing prototypes instead of mockups. In short the client stuck jackpot with this. They accomplished as much in 3 weeks as they would have normally accomplished in 3 years.

## 3.3.10. Case study of Thought Works. [23]

**Company:** ThoughtWorks is a global technology company who specializes in providing software design and delivery as well as pioneering tools and consulting services. They are closely linked with the usage of agile software development and has contributed to a vast array of open source products.

**Client:** GAP

**Business Issue:** GAP wanted to improve their databases systems. They preferred migrating to Big Data Technologies, like MongoDB to help them resolve their analytics and data warehouse problems. They sought the help of ThoughtWorks to consult them in this whole process of changing into Big Data Systems.

**Approach:** This case had been taken up by Ryan Murray of ThoughtWorks. It showcased the effects of enabling of agile engineering with MongoDB for fast and efficient productivity and better business intelligence.

Keynotes of the case:

In relational databases, for instance, if we are making a database based on purchase order, we split this into various sub-lists (We tend to do this to normalize databases)

Developer tends to solve on how to design the relational database instead out with relational database. This tends to destroy the flow. (Impedance mismatch to be avoided while creating purchase orders for GAP, using MongoDB)

Introduction of micro service. Service dedicated to one business capability. Developers will be more focused into working on these services.

**Flexible Schemas:** 1) Data comes in different shapes and formats. We should be able to store them transparently in the databases. This is **Extreme Enabler,** for Agile Development. [Design-Implement-Release]. Using **Agile Methodologies**, we continue to evolve the system with every release.

Using MongoDB, setting up the databases on basis of business requirements was easy. MongoDB accepts whatever type of data and dataset you throw at it. Once the Client

(GAP Apparels in this case), visualizes it, gives their feedback, we incorporated it for the future releases with every iteration.

Flat data structure of BSON/JSON used in MongoDB is also easy to be understood by non-technical business consultants.

ThoughtWorks consultants had set up a 5 node cluster. It had 3 nodes for high availability and 2 disaster recovery nodes. These were all setup in a separate data center. Also they were able to setup a delayed replica set node. (If somebody fat fingered data, there was a real-time backup to retrieve and solve the mistake)

Final Output was: System in production was delivered in 25 months using MongoDB and implementing Agile Engineering methodologies. Constant scheme of migration wasn't needed with Mongo DB and Agile Engineering.

**Conclusion:** ThoughtWorks was successfully able to implement the big data system using Mongo DB and Agile Analytics. GAP experienced high efficiency in its data storage, data processing as well as its analytics field of its business. With the inclusion of Agile Methodology, the relationship with the client improved since the feedbacks were very quickly implemented with every iteration. ThoughtWorks was able to prove the power of adapting to this new technology.

| Company | Client | Business Issue | Approach | Technologies Used | Types of Analytics | Industry | Agile |
|---|---|---|---|---|---|---|---|
| **Analytical Data Solutions** | US Based Luxury Retailer | Identifying the most valuable customer. | Customer Segmentation. (Recency, Frequency, Monetary) | ----- | Predictive, Prescriptive. | Retail | no |
| | Global Marketing Organization of PC manufacturers | Using sentiments and buzzwords in social media to predict growth of products | Data Crawling from social networking websites. | ---- | Predictive | Electronics | Signs were present. |
| **IBM** | Colombus Foods | Improve response on Operational Services during attending client demands. | Instrumented, interconnected and intelligent approach was used | IBM Cognos TM1 | Descriptive, Predictive, Prescriptive | Food | No |
| | Corona Direct | To acquire new customers using successful marketing campaigns. | Using predictive analytics to identify the group with the highest potential to respond to campaigns | IBM SPSS Predictive Analytics software. | Predictive, Prescriptive | Financial Services (Insurance) | No |
| | Mediamath | Needed data analysts to enable the largest ad buyers for their ad campaigns. | Customer Segmentation, Audience Targeting, Deep Site Analysis, | Netezza, MathClarity ™ | Predictive, Prescriptive, Descriptive | Advertisement | Yes |
| | Mueller Inc | shift business strategy from manufacturing to retail led manufacturing & adopt a more retail oriented business model | Audience Targeting | IBM® Cognos® Business Intelligence. IBM SPSS ® | Predictive, Prescriptive | Manufacturing | Signs were present. |
| | Top Toy | Customer demand forecasting. | Implementation of customer segmentation on basis of selling patterns | IBM® Cognos® | Predictive | Retail | No |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Invenco Oy** | VAASAN | VAASAN faced a business risk due to not having a standardized customer demand to predict the fluctuations. | Modelled the data forecasting system using variables like location, size, history, sales. | IBM® Cognos® | Predictive, Prescriptive. | Food | Yes |
| **Deloitte** | A major Telecom company | Company was trying to efficiently migrate to Big Data Technologies. | Leadership Alignment, Selection of Big Data Analytics Pilot Project, Design & Execution | HADOOP | Prescriptive | Telecommunications | Yes |
| **PriceWater House Cooper** | A large software company | Client wanted to balance i) Data Storage and migrate to cloud based services. Ii) Client wanted to use advanced analytics to mine information from the vast data their customers generated. | Change in Data Architecture model, recommended using open source tools and techniques which would add flexibility | Open source tools | Predictive, Prescriptive | Information Technology | Yes |
| | Canada's leading pension firm | The company was interested in making its data analysis and business intelligence more efficient and robust | PwC proposed **Agile analytics** to help the company address key issues .they gave the clients with working dashboards that addressed many of their immediate concerns. | --- | Prescriptive | Finance | Yes |
| **ThoughtWorks** | GAP | GAP wanted to migrate to Big Data Technologies to address various business issues much efficiently and effectively. | ThoughtWorks proposed to migrate and implement Mongo DB as the best solution for Big Data systems. | Mongo DB | Predictive, Prescriptive | Apparel, Retail | Yes |

*Table 3.2 Comparison of case studies of various companies*

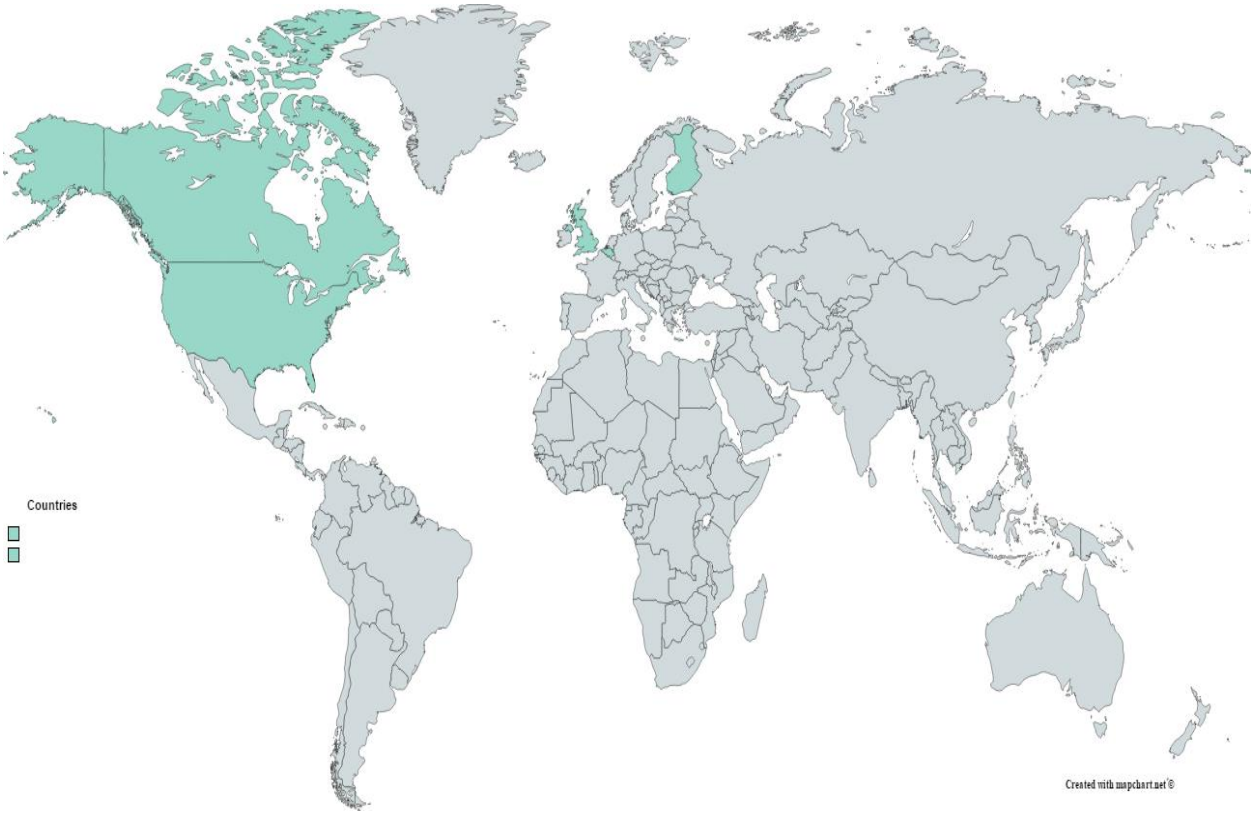


*Fig 3.2 Geolocations of the countries on which the case studies were carried out*

This is the heat map of the World. The highlighted portions denotes the areas where the case studies have been performed.

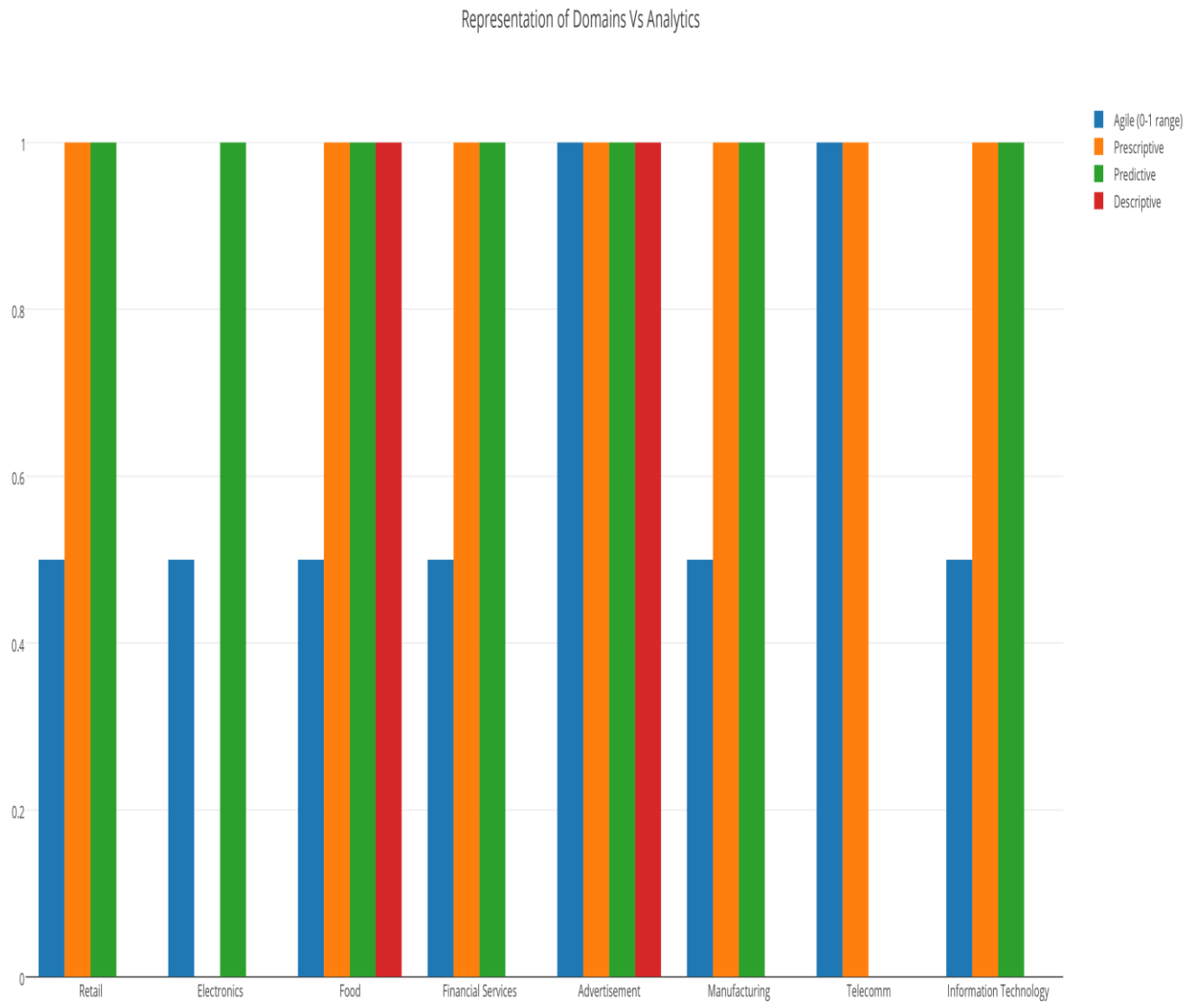


*Fig 3.3 A Graphical comparison of the different sectors along with the type of modelling followed*

The graph here represents a normalized comparison of the various industrial domains versus the types of analytics they are using in their business intelligence. To understand the context of the usage, I have taken into consideration (1 if present, 0 if absent for the cases of Prescriptive, Predictive & Descriptive, and for situations where Agile was used, have taken 1, and 0.5 if it was used in more than one case, and 0 if not used in the other.)

## 3.4 Conclusion

From the above cases, and also from the graph, we can see that the case study involving the domain Advertisement, has all the factors present needed for the research purpose of this thesis. Most of the cases have involved predictive and prescriptive approach, but a very less number of cases involved descriptive.

Also we saw that Agile Methodological approach wasn't implemented by all the companies and this makes us realize that companies do not want to explore this unchartered territory. The ones who explored and exploited this approach did gain substantially in their final outcome after modelling. Also we can clearly see that the modelling approaches depends on the domain of the industry.

Without further ado, we will try to focus into **Customer Segmentation and Clustering, and how the processes of Descriptive, Predictive and Prescriptive are involved in this scenario incorporating Agile Methodologies in the succeeding chapters.**

# *CHAPTER IV. CUSTOMER SEGMENTATION, CLUSTERING AND PROPOSED METHODOLOGIES*

## 4.1 What is Segmentation?

**[24]** The term “Market Segmentation” was first brought up by Wendell (1956), an American Market Researcher. This concept was further promoted by various companies as well as many scholars. Presently a large number of research study is being done on this field of market segmentation to provide deeper insights into its potentials.

Market Segmentation has numerous advantages over mass marketing. It frequently provides the opportunity to expand a market by better satisfying the specific needs or wishes of particular consumers (MacQueen et al., 1967). Also, it increases the profitability or effectiveness of the organization to the extent that the economic benefits gained by providing to the consumers exceeds the cost of the segmentation process. Kotler believed that the company can create a more fine-tuned product or service offering and price it appropriately for the target segment (Kotler, 2000). As a consequence of this, the company can easily select the best distribution and communication channels, and easily have a clear picture the market competitiveness and competitors. **[25] [26]**

## 4.2 What is Clustering?

Clustering is a useful technique for the discovery or identification of some knowledge from a dataset. Classification problems often resort to clustering for their solution because of its explanatory methods. When little or no information is present about the category structure in a data body, the appropriate usage of clustering can be prominently visible then. The main objective of clustering is to sample cases under consideration of the same

group, such that the member of the same group will have high degree of association while the association between members of other groups will be low.

Clustering is also called unsupervised classification, where no predefined classes are assigned (Tou & Gonzales, 1974).

## 4.3 Clustering Methods

The quality of a good cluster depends on the methods implemented to create them. A good clustering method is more likely to produce high-quality clusters with high intra-class similarities and low inter-class similarity. The quality of the clustering result depends on both the similarity measure used by the method and how it is implemented. It is measured by the ability of the system to discover and reveal new patterns from the datasets.

Clustering of datasets are based on two approaches, namely, Partitional and Hierarchical (Jain & Dubes, 1988; Jain, Murty, & Flynn, 1999). Over the last decades, there have been a high evolution in the domain of artificial intelligence and soft computing. This has led to a vast array of clustering methods being developed on various theories and techniques. **[27] [28]**

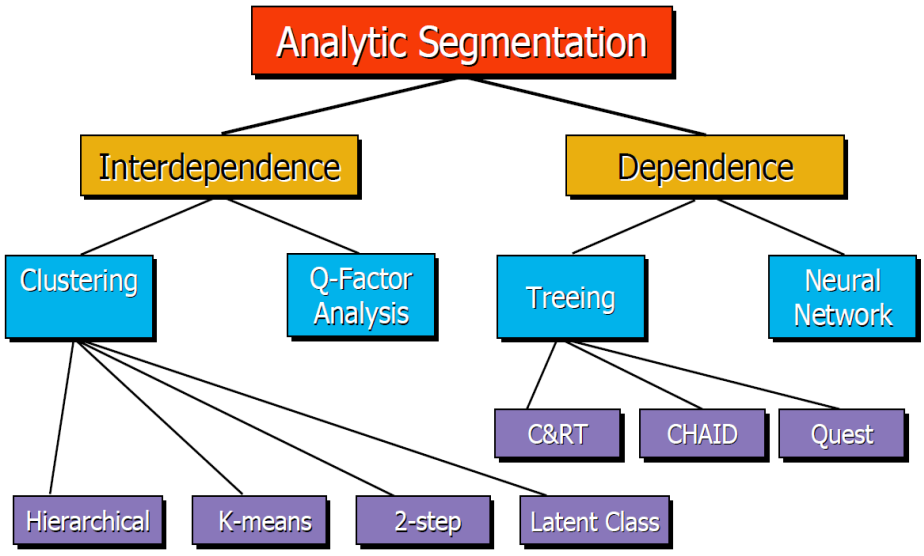


*Fig 4.1 Different Types of Segmentation Approaches*

**[29] [30]** In quant segmentation, there are two different approaches: “Interdependence” segmentation (or clustering) and “Dependence” segmentation (CHAID, or Chi-Squared Interaction Detection). Clustering is by far one of the most widely used interdependence procedures while CHAID method is subjected to individual preferences. In one way, the words “**strategy**” can be associated to the **interdependence segmentation** method, while the word “**tactical**” can be associated to the more **dependence segmentation** method.

“What do customers in this market want?”– Then a clustering approach is appropriate. In case of more tactical questions like, “What type of customers will explicitly purchase the product and how to reach them and prompt them to buy?” then CHAID is more fitting because of its dependence approach. So from this the practice that we can generalize is, for general strategy, we should consider interdependence segmentation and for specific tactics, we can go for dependence approach to segmentation.

## 4.4 K means Clustering & CHAID Clustering:

### 4.4.1 K means Clustering

**[31] [32]** K means clustering algorithm is one of the simplest algorithm used for solving clustering problems. In 1967, Mac Queen , proposed for the K- means algorithm for the first time.

During every iteration of the algorithm, each data is assigned to the nearest partition based upon some similarity parameters (as example Euclidean Distance measure). After the completion of every successive pass, a data may switch partitions, resulting in changing the values of the original partitions. Various steps of the standard K-Means clustering Algorithm are follows,

(1) The number of clusters is first initialized and accordingly the initial cluster centers are randomly selected.

(2) A new partition is then generated by assigning each data to the cluster that has the closest centroid.

(3) When all objects have been assigned, the positions of the K centroids are recalculated.

(4) Steps 2 and 3 are repeated until the centroids no longer move any cluster.

**[33]** The main objective of K-Means is the minimization of an objective function that determines the closeness between the data and the cluster centers, and is calculated as follows:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \left\| d(X_{i,} C_j) \right\|$$

where, $\left\| d(X_{i,} C_j) \right\|$ is the distance between the data $X_{i,}$ and the cluster center $C_j$. **The downside of K-Means algorithm is that, the result of clustering mostly depends on the initially selected centroids**. Spherical datasets cannot be efficiently clustered using K-Means.

## 4.4.2 CHAID:

**[34]** CHAID is a kind of decision tree based on adjusted significance testing. It is used for prediction. CHAID stands for Chi-Squared Automatic Interaction Detection. In application it is used the context of direct marketing to select groups of consumers and predict how their responses to some variables affect other variables. CHAID's advantages are that its output is easy to interpret and is highly visual. **Since it uses multiple splits by default, the sample sizes have to be large for effective functioning, else small sizes of segments can become too small for reliable analysis.** CHAID is recommended for dependence segmentation because amongst the treeing methods, it tends to be the most applicable approach. CHAID works with all great types of data, including missing data.

Two approaches, K method and CHAID differ significantly in the type of information they yield

## 4.5 RFM Customer Segmentation Procedure:

**[35]** RFM stands for **Recency, Frequency and Monetary. RFM** analysis is used to improve customer segmentation, by dividing the customers into several groups.to personalize specific services catering towards their taste and paste usage history. It is considered to be the easiest customer database segmentation procedure that finds its applications mostly in the field of reactivation campaigns, high valued customer programs, combating churn, etc. In recent years, data mining applications based on RFM concepts have been highly observed in different areas ranging from, computer security, automobile industry to electronics industry.

RFM is based on user activity data. Any kind of data from actual orders, website visits, app launches, etc are considered to be integral form of data for RFM computation. It can be applied to the activity-related data that has **measureable value** and is repeatable. Hosseini et.al (2010) combined weighted RFM model into K-means algorithm to improve Customer Relationship Management (CRM) for enterprises.

## 4.6 Definition of RFM Metrics: [36]

**Recency (R):** Time since last transaction.

**Frequency (F)**: Total number of transactions.

**Monetary (M):** Total transactions value.

This approach is called the TOTALS approach. Transactions can only increase customer value in the segmentation. Pros being it is easy to explain.

The AVERAGES approach contradicts in the Frequency and Monetary sector. Frequency is defined as Average time between transactions, while Monetary is defined as average transaction value. This approach can increase and decrease the customer value in the segmentation.

The RFM algorithm is described as follows:

Step 1: Calculate the RFM metrics for each customer

Step 2: Find the distribution for each metric and define the segmentation by splitting into bins. (The easiest way to split metrics into segments is by using quantiles)

Step 3: Add segment numbers to the RFM table. (This is called Segmented RFM table)

Use a stacked contingency table to count customers in each segment and compute the summary statistics. Use Recency Segmentation to identify customers at risk of churn (if survival Analysis is used for Recency segmentation, it works very well).

Use the frequency & Monetary segmentation to estimate customer value. (Example of customer segment names: Premium, Gold, Silver)

**[37]**

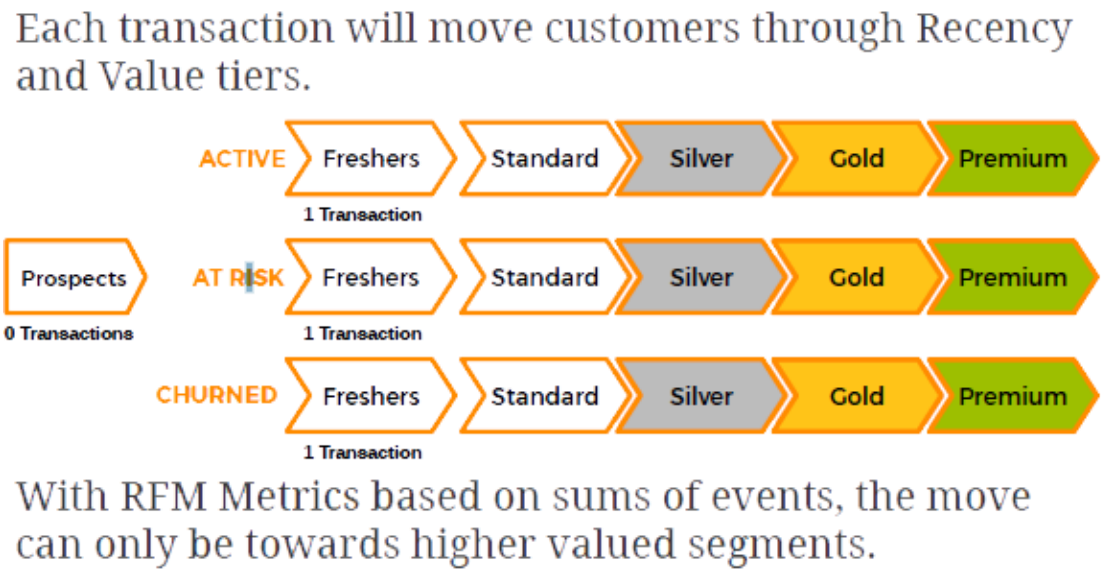

*Fig 4.2. An example of customer segmentation*

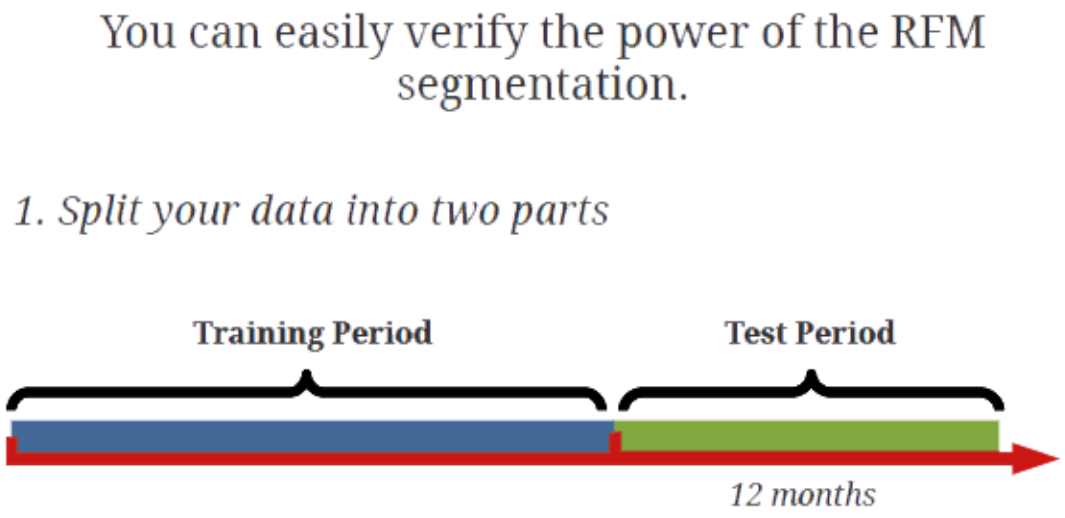


*Fig4.3 Step 1*

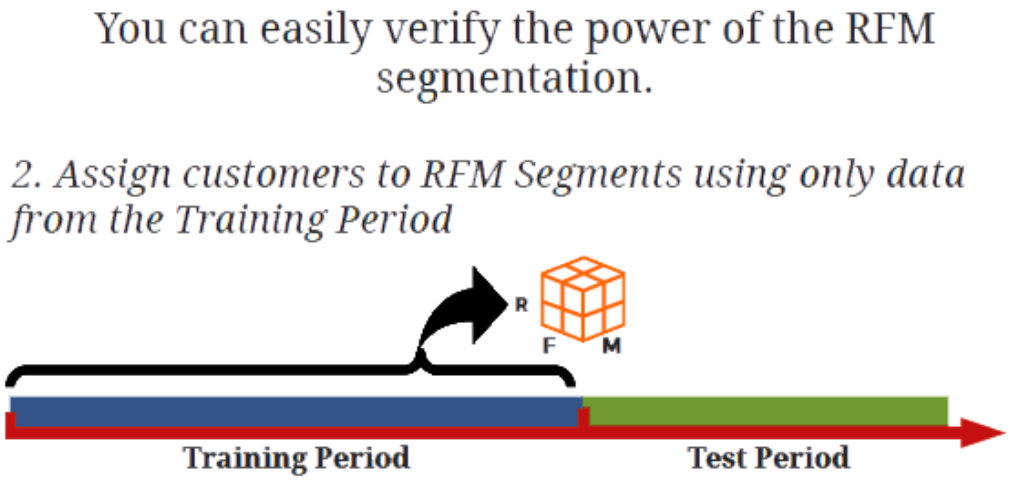


*Fig 4.4. Step 2*

3. Calculate the average value of customers in each RFM Segment over the Test Period

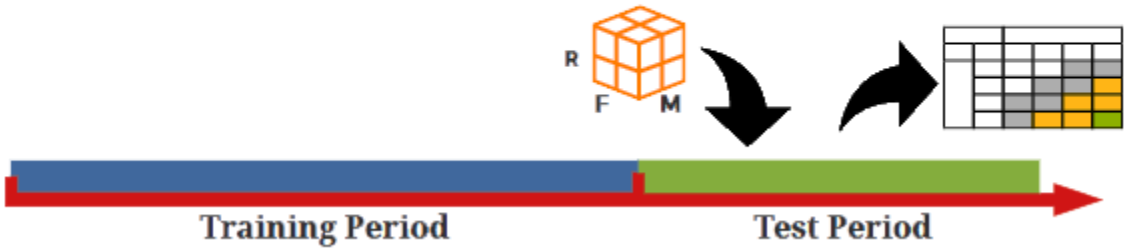


*Fig 4.5 Step 3*

You should see the average value increase consistently with segmentation

i.e. behaviour in the Training Period is a good predictor of value in the Test Period

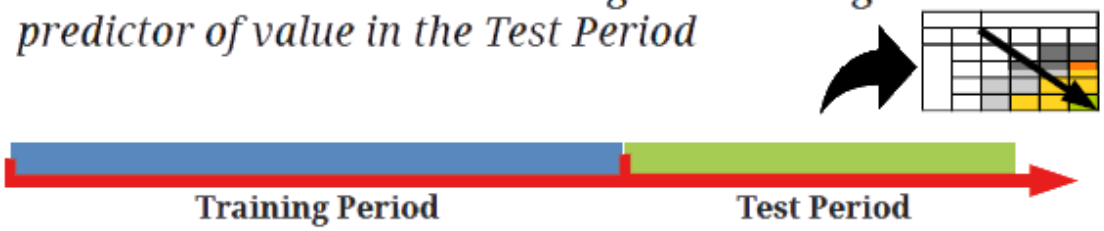


*Fig 4.6 Step 4.*

In short the flow chart of Data mining using RFM for segmentation is shown below

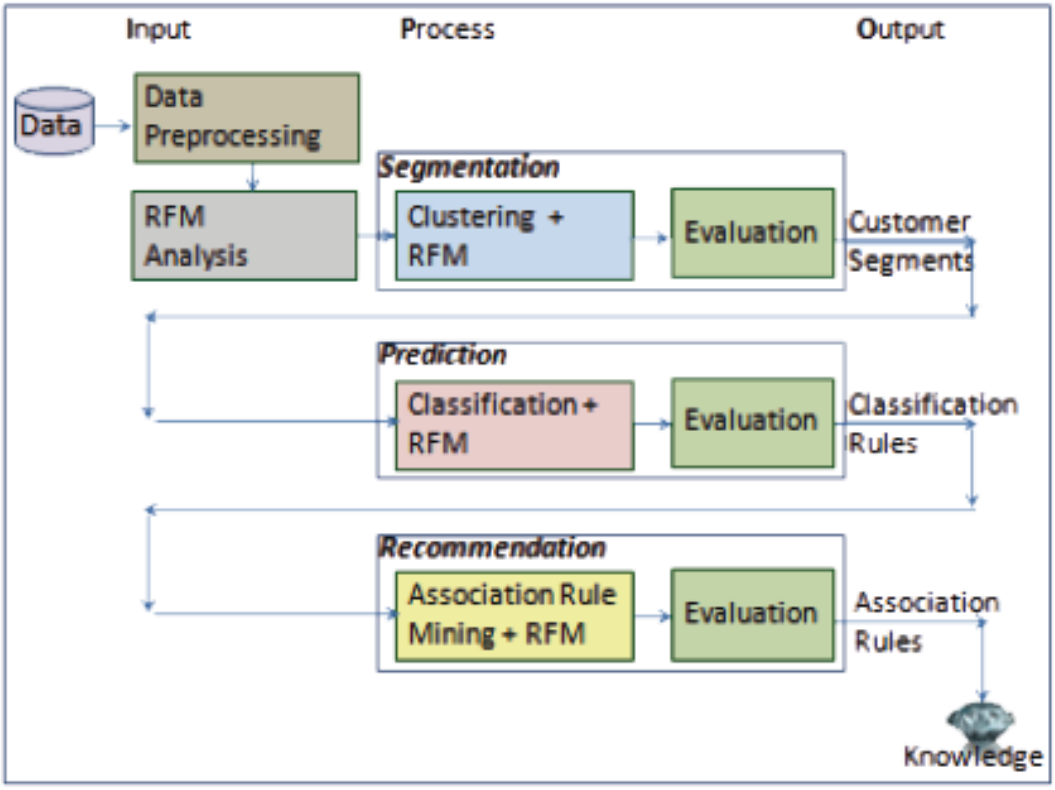


*Fig 4.7 A flowchart of Customer Segmentation procedure.*

Following are some case studies on Customer Segmentation using data mining techniques. (RFM model, k-means clustering)

## 4.7 CASE I:

**[38] Company**: Online UK based retailer, registered as non-store business with around 80 employees.

**Business Issue:** The merchant relied heavily on direct mailing catalogues. Recently they migrated to online platform and since then have been experiencing a steady growth of their customer base. They want to better understand the customer and **conduct customer-centric marketing** using customer segmentation and clustering.

**Approach:**

**Table 1:** Variables in the customer transaction dataset (4381 instances)

| *Variable name* | *Data type* | *Description; typical values and meanings* |
|---|---|---|
| Invoice | Nominal | Invoice number; a 6-digit integral number uniquely assigned to each transaction |
| StockCode | Nominal | Product (item) code; a 5-digit integral number uniquely assigned to each distinct product |
| Description | Nominal | Product (item) name; CARD I LOVE LONDON |
| Quantity | Numeric | The quantities of each product (item) per transaction |
| Price | Numeric | Product price per unit in sterling; £45.23 |
| InvoiceDate | Numeric | The day and time when each transaction was generated; 31/05/2011 15:59 |
| Address Line 1 | Nominal | Delivery address line 1; 103 Borough Road |
| Address Line 2 | Nominal | Delivery address line 2; Elephant and Castle |
| Address Line 3 | Nominal | Delivery address line 3; London |
| PostCode | Nominal | Delivery address postcode, mainly for consumers from the UK; SE1 0AA |
| Country | Nominal | Delivery address country; England |

*Table 4.1: Variables of interest chosen from the dataset*

The customer transaction dataset of the merchant has 11 variables. The variable PostCode is essential for the business. The time period of this case has been taken from 1st January, 2011 – 31st December, 2011. Over this period, there were 22190 valid transactions, with 4381 valid distinct postcodes. Only UK consumers have been analyzed.

**Data Preprocessing:**

1. Select variables of interest from the dataset. 6 variables were chosen. **Invoice, Stock Code, Quantity, Price, Invoice Data, & Post Code.**

2. Amount created as product of Quantity and Price.

3. Separate the variable Invoice Date into two variables Date & Time. This allows same day transactions to be treated differently due to their timings.

4. Filer out transactions with invalid Post Code.

5. Sort out dataset by Post Code and calculate Recency, Frequency & Monetary per postcode.

**Choice of Algorithm:**

K means clustering algorithm was used for this case. It is to be noted that K means clustering algorithm is very sensitive to dataset which contains anomalies or not normalized variables.

**Understanding the Clusters**

| | *Minimum* | *Median* | *Maximum* |
|---|---|---|---|
| *Cluster* 1 | | | |
| Recency | 8 | 9.8 | 12 |
| Frequency | 1 | 1.3 | 4 |
| Monetary | 3.75 | 361.20 | 7741.47 |
| First_Purchase | 8 | 11.1 | 12 |
| *Cluster* 2 | | | |
| Recency | 4 | 5.4 | 7 |
| Frequency | 1 | 2.3 | 13 |
| Monetary | 15 | 586.19 | 3906.27 |
| First_Purchase | 4 | 7.7 | 12 |
| *Cluster* 3 | | | |
| Recency | 0 | 1.5 | 3 |
| Frequency | 1 | 2.6 | 7 |
| Monetary | 20.8 | 685.71 | 4314.72 |
| First_Purchase | 0 | 5.3 | 12 |
| *Cluster* 4 | | | |
| Recency | 0 | 1.0 | 5 |
| Frequency | 3 | 8.3 | 16 |
| Monetary | 191.17 | 2425.09 | 7330.8 |
| First_Purchase | 1 | 1.0 | 12 |
| *Cluster* 5 | | | |
| Recency | 0 | 0.7 | 6 |
| Frequency | 3 | 17.7 | 28 |
| Monetary | 1641.48 | 5962.85 | 13110.02 |
| First_Purchase | 0 | 11.1 | 12 |

*Table 4.2 Clustering Results*

**Cluster 1:** 527 consumers, 14.4% of total population. **Least Profitable group. Frequency was also very low of 1.3**

**Cluster 5:** 188 consumers. **High Recency.High Frequency. High Monetary.** The customers in this cluster contributed 25.5% of total sales. This group was the smallest (5.05%) of total population but seemed **most profitable.**

**Cluster 4:** 627 customers. High value for frequency.

**Cluster 2:** 459 consumers. **Lower Frequency, smaller value of monetary. Indicating smaller amount spending per consumer.**

**Cluster 3:** Largest group with 1748 consumers. Have reasonable value of **Frequency. This group shows newly registered consumers. This group has certain level of uncertainty in terms of profitability.**

**Cluster 3,** is the most diverse cluster. To refine segmentation, a decision tree was used to create some nested segments internally.

**Conclusion:** The most valuable customers contributed to 60% of total sales, while least valuable made up only 4%. . The business can improve by better understanding associations among consumer groups.

## 4.8 Case II:

**[39] Company:** Four major Australian banks.

**Business Issue:** Clustering customers on basis of Electronic Funds Transfer at Point of Sale (EFTPOS) transaction data.

**Approach:** RFM analysis was used for data reduction. The size of the data was reduced from 130GB of raw transactional data to 80MB of RFM values of the retailers. The data obtained from the banks were received by Monash University for the purpose of this case study.

Data reduction processes were undertaken using hash tables and clustering algorithms used in retailer segmentation experiments needed to be optimized. The EFTPOS data contained **55 variables, these attributes, e.g., Retailer names, ids, credit/debit card numbers, pin numbers, billing address, etc. were masked for the sake of privacy.** Time period of the experiment was of 18 days, (19th September, 2013- 07 October-2013). **To reduce memory, smaller daily data files have been created by copying the values of relevant attributes, such as Time, Retailer Id, and Amount.**

****** The calculation of Recency was the most time consuming and resource intensive. In order to save time for the monetary calculation, and reduce processor load, each daily

transaction file is divided into 20 small subsets and the calculation is done on all the subsets (i.e. 20x18=360 subsets) in parallel.

Two types of clustering were used, a) K-Means Clustering b) Agglomerative Hierarchical Clustering.

**Results:**

Clustering model with 19 clusters were chosen to be optimum number of clusters. The values were divided into 5 regions, 1 being lowest, and 5 being highest. **Lower values of Recency suggest recent transactions. , low value of Recency coupled** with **high values** of both **Frequency and Monetary** is the feature of **a retailer** who **actively generating regular transactions**

Cluster 1,2,18 show these kind of characteristics. They make up 34.23% of the total retailers. This is the most profitable segment for the bank.

Clusters 5, 6, 13, 14, 15 consist of retailers with **moderately high Recency** and **Monetary values value with average Frequency value**. **These reta**ilers are **recently inactive** for some time **resulting** in the **moderate number of EFTPOS** transactions. Since **these retailers have the capacity to generate revenue**, the **bank may be want to employ a marketing strategy** to provide incentives for them to improve their business activities. Of the three attributes, Recency is seen as the most important indicator of customer loyalty. Clusters 3, 4, 7 to 12 and 17 have **retailers with high Recency value and average Frequency and Monetary values**. **This segment of retailers**, which constitutes 21% of the total retailers, has **been inactive for a while, hence they may be at risk of attrition**

**Conclusion:** The summary of the results of the clustering experiments demonstrates that even simple clustering of the retailers based on their RFM values can reveal the business behaviors of segments of retailers using the EFTPOS facilities.

A natural progression of this work is on finding ways to optimize the **Agglomerative Hierarchical Clustering (AHC)** algorithm for Big Data through parallelization. Splitting the data set into two subsets like what we have done in this paper is obviously not scalable.

The bottleneck in the **AHC algorithm** is in the **calculations of the distance metric between data points** to build the distance matrix and to update it upon the creation of a new cluster. **Time saving can be achieved by parallelization** of this part of the algorithm, e.g. the calculations of the distance between a set of data points with each other and with all the other data points are done on one processor. Hence a few processors will run in parallel, each is dedicated to a subset of the data set.
A Hadoop processing environment has recently been set up for this project which will enable parallel processing on very large unidentifiable intermediate data.

To gain more insight into segments of the retailers, we will perform rule inductions, association rules analysis and build causal models by using additional attributes, from the original data set as well as other exogenous attributes, like sociodemographic, advertising, social media data. Using additional attributes with large nominal values will introduce another Big Data challenge, i.e. Variety, into this project. Finding suitable attribute subset selection method for Big Data will be another exciting avenue in this research.

## 4.9 Issues Related to the following case studies

**Data complexities:** In processing large number of data sets, proper set of data reduction techniques must be implemented, along with the selection of right set of attributes.

**Algorithm Complexities:** Most of the customer segmentation was based on K clustering algorithms, though more efficient algorithms like K++ clustering and Agglomerative Hierarchical Clustering have been proved to work more efficiently. Time complexities vary in all of these algorithms, and choice of usage and implementation solely lies at the discretion of the analyst.

**Process complexity:** Process complexity lies in the fact about how the parallelization of the computer processing power will be used to compute data and analyze it.

## 4.10 How Agile Can be Helpful?

There are several key reasons why Agile methods are well suited to building analytical databases. Among them are business-driven vs. data-driven development, and reduced risk and complexity.

**Agile Methods Take a Business-Driven Approach.** Simply stated, creating analytical databases is complex, time-consuming and oftentimes overly expensive, especially when traditional data-driven methods are used. A fundamental truth of BI and data warehousing is that data integration and homogenization account for 70% to 80% of the project budget and an even higher percentage of the risk.

Let's take a simple example. Assume your organization has four internal OLTP (Online Transaction Processing) systems and one external data source. On average, each system has 30 database tables, and each table contains 30 columns. This means that:

(4 OLTP systems + 1 external system)
x 30 tables x 30 columns
= 4,500 data elements

In a data-driven approach, it's not uncommon to want to integrate and homogenize most—if not all—of the data before the first query or report can be written. This means that 70% to 80% of the project budget will be expended before any business value can be realized. Similarly, integrating thousands of fields can take upwards of 12 months to complete. What this means is that our goal should be to minimize the amount of effort associated with data integration and homogenization.

In an Agile delivery model, only data needed to answer specific business questions or to solve specific business problems is sourced. (These need statements are captured in a series of business "stories.") So instead of first trying to "boil the ocean" via a massive data integration effort, Agile practitioners work with the business community to define the hundred or so data elements that drive performance. This means that the business will be in a position to receive value much more quickly—in weeks or months rather than quarters or years.

**Agile Methods Reduce Risk and Produce Systems with High Adoption Rates.** Organizations that apply traditional waterfall methods to BI/ DW projects accrue unnecessary risk and may find out what they've created does not satisfy the business's needs. Waterfall methods mean it's all or nothing. In other words, design cannot begin until all of the requirements are defined, and coding cannot begin until design is complete. This means that the project takes on ever-increasing levels of risk and that business value is delivered at the end of the project.

It's not unusual to find that once an analytical database has been deployed using a data-driven and waterfall approach, it suffers from low adoption and usage. The primary culprit seems to be that business needs and priorities will have shifted between the time the requirements were originally defined and when the analytical database was deployed.

For the most part, the delivery team's efforts are seen as a black hole. What ends up being delivered is based on assumptions and interpretations of the requirements and may not contain the information the business really needs. Reality hits when that first report is written, oftentimes making the data integration effort completely moot.

When Agile methods are applied, value can be shown on a recurring basis. The key tasks of database design—data quality remediation, and data integration and homogenization— are broken into short, time boxed and scope-boxed delivery cycles, or "sprints," that generally last two to four weeks each. These data-focused tasks are paired with prototyping in the BI layer, allowing the business to interact with the data multiple times, helping to assure that the analytical database truly contains useful information.

The use of delivery sprints keeps business value at the forefront and drives project risk down to a mini-mum. At the end of each two- to four-week sprint cycle, the delivery team is required to demonstrate what they have produced, making their work much more visible to the business and allowing for midcourse corrections because another tenet of Agile is to "fail quickly." **[39]**

## 4.11. Proposed Methodologies

### 4.11.1 Data Acquisition:

**Acquisition Step**:

Data acquisition techniques are very important in the analytics domain. From our cases we have understood that the selection of the right set of variables from a proper dataset gives way to new discoveries. In cases of customer segmentation, we get to segment on basis of postal code (which is most of the time constant), amount of money spent, and product. This three variables are going to be very helpful for a company who prefers to segment customers for the purpose of marketing. But if the company started with the wrong set of variables, for instance, invoice, stock code, quantity, the desired analytical outcome would not have been possible.

**My proposed data acquisition step would be to select a set of 6 variables. 3 from the customer's perspective (location which can be in terms of postal code, gender so that gender specific clustering can be done, and age, also helps to understand the age group of the customers) and 3 from the company's perspective(product, price, quantity).** This approach can be considered a good starting point for companies who wants to cluster their clients for marketing purpose. This is because it gives way to the possibility to answer various questions like, "Which product is preferred by which age group?", or "how much money is spent by people of a particular group?" etc.

Agile methodology can be implemented in the data acquiring method by taking into consideration the most important variables for the first iteration, and then moving on to the less important variables from the datasets. This helps in focusing the main demands first and gradually narrows down with to a much specific purpose.

The fundamental reason for moving toward agile processes are

1) To respond quickly to immediate and pressing need
2) To engage with users throughout the development process.

This will help reducing timescales and cost of acquiring data from operational systems, enterprise applications. The sampling period is a very tricky question, owing to the two fundamental factors of Big Data, i.e, Velocity and Volume. Both of these issues have to be tamed in order to overcome this problem. The sampling period determines on two conditions,

a) The amount of data the organization is willing to sample.
b) Volume of the data.

The sample should be taken in such a way, so that the processing speed reduces and a particular outcome can be obtained quickly. **Agile Approach can be implemented here. We can increase the sampling period in an ascending manner with every iteration and can benefit from this outcome. Focus should be on the quickest sample sets and not the time consuming ones.** For sample size, we can consider around 400-600 minimum samples; more than 1000 as typical size. **Over-sampling should be avoided as much as possible.**

**For analytic segmentation, we should try to gather a large sample (more than 600).** Segments as small as 5% can often promise high profitability once they are weighted according to their spend levels. But generally, one would never want to make inferences from a subsample any smaller than around n=30. Thus, one would have had to start with an overall sample of 600, (30/5% = 600) in order to ensure that a segment as small as 5% could be reliably described.

We cannot know how well the data warehouse design matches the available data until we try to load it, nor how well it matches the stakeholder's actual BI requirements until they use it. This is why Agile Approach of early analytics and implementation can reduce the Data Warehouse/ Business Intelligence risk.

## 4.11.2 Selection of Tools for Analytics

Nowadays, since the explosion of Big Data, a demand for Analytical Tools have increased very significantly. It is extremely difficult to select the right set of tools for a specific purpose as most of them boast of their strengths. When choosing a data analysis software we need to answer a basic set of questions, like:

- Does it run natively on the computer?
- Does it handle large datasets?
- Does the software provide all methods we need?
- Is it affordable?
- Ease of use.

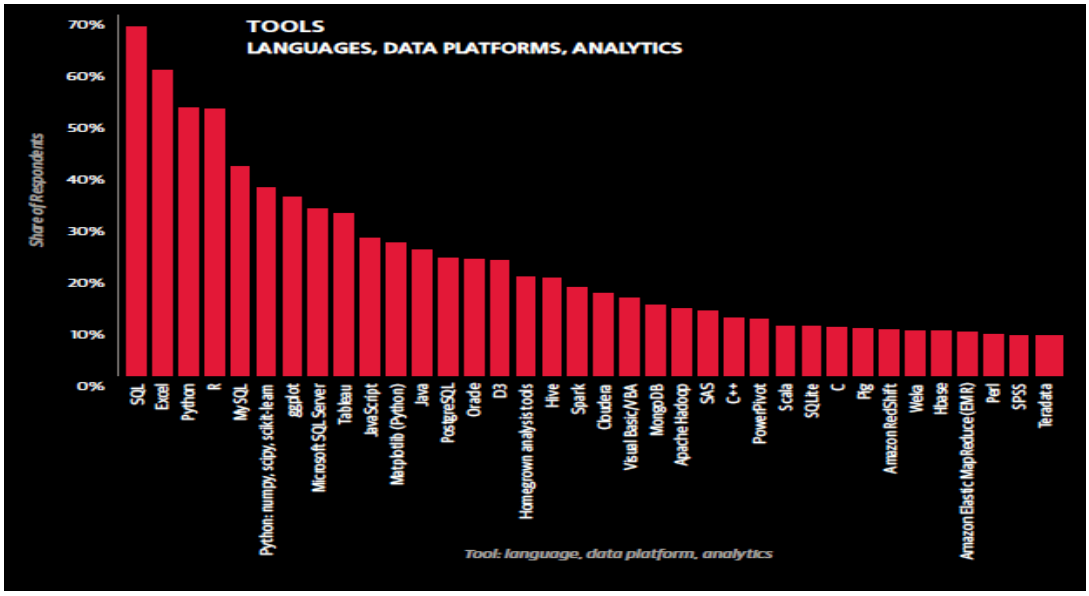


*Fig 4.8: Tools used by 2015 respondents to O'Reilly 2015 salary survey.* [http://r4stats.com/articles/popularity/](http://r4stats.com/articles/popularity/)

A lavastorm survey showcased a different story for the Advanced Analytics Domain.

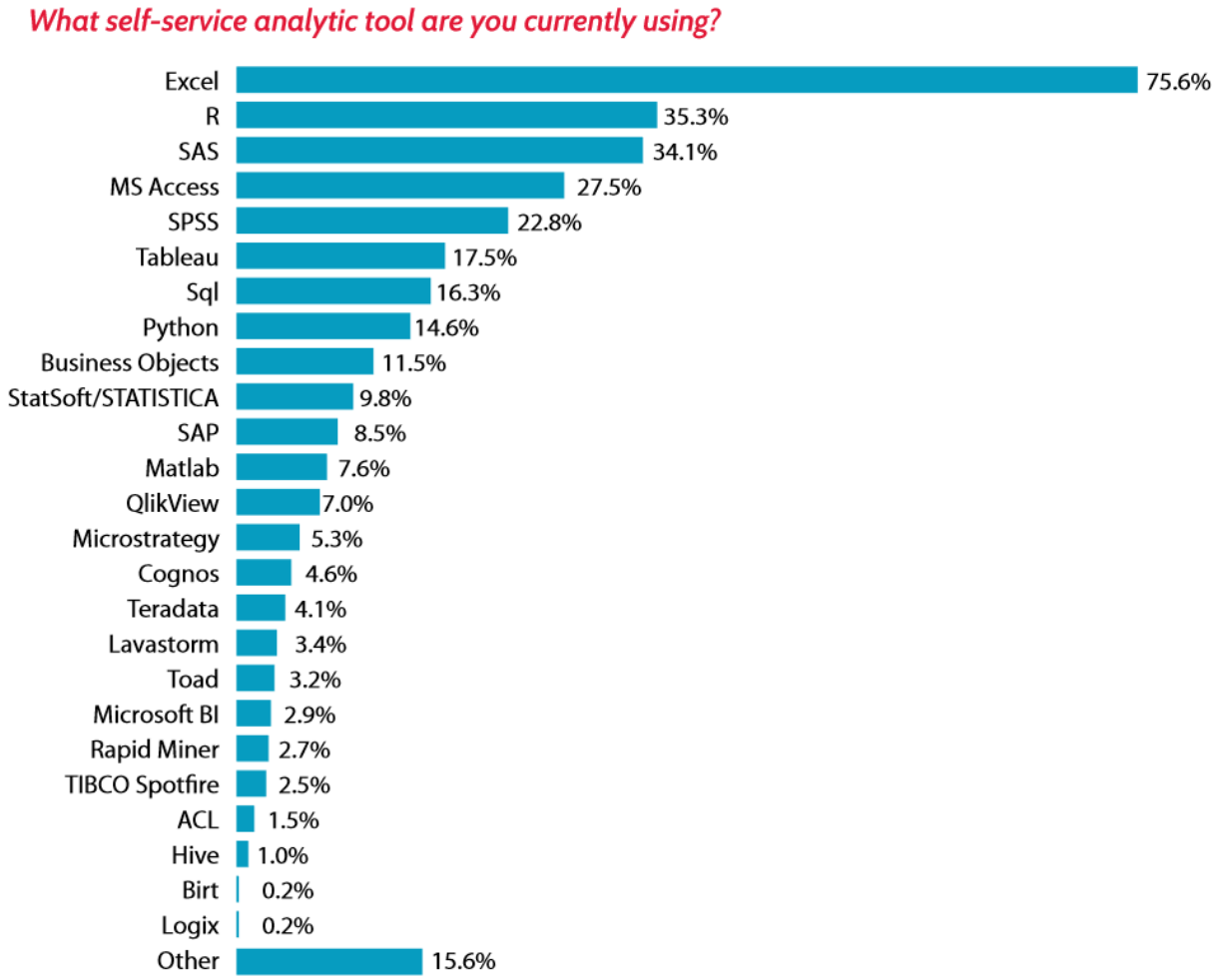


*FIG 4.9 Lavastorm survey of Analytics Tool*

*"http://r4stats.com/articles/popularity/"*

Although the ranking of each package varies due to the criteria used, yet we can see major trends. Among the software that tends has pre-written methods, R,SAS, SPSS, and STATA lie in the top, while R and SAS changing places frequently to hold to the pole position.

For software used in language analytics, C,C#,C++,Java, MATLAB, Python,R and SAS are always on the top.

For people who prefer workflow style control in their software, KNIME, Microsoft Azure Machine Language , RapidMiner , SPSS Modeler , SAS Studio are the leaders.

| Name | Advantages | Disadvantages | Open source? | Typical users |
|---|---|---|---|---|
| **R** | Library support; visualization | Steep learning curve | Yes | Finance; Statistics |
| **Matlab** | Elegant matrix support; visualization | Expensive; incomplete statistics support | No | Engineering |
| **SciPy/NumPy/Matplotlib** | Python (general-purpose programming language) | Immature | Yes | Engineering |
| **Excel** | Easy; visual; flexible | Large datasets | No | Business |
| **SAS** | Large datasets | Expensive; outdated programming language | No | Business; Government |
| **Stata** | Easy statistical analysis | | No | Science |
| **SPSS** | Like Stata but more expensive and worse | | | |

*Table 4.3 Comparison of Data Analysis Packages.*

**Although all of these software supports Agile Approach but the final selection choice lies on the user and for the purpose he wants to focus in.**

## 4.11.3 Selection of Algorithm

The choice of algorithm is extremely crucial to any form of analytics. It is because algorithms addresses two issues,

a) **Time complexity** (How long does it take to process the data using the specified algorithm)

b) **Space Complexity** (This is essentially the number of memory cells the algorithm needs).

In our cases of Clustering, we were focusing on Algorithms meant to sort and segment the data. When handling multi-dimensional data sets, we started off with K means clustering. This algorithm is an interdependence algorithm based on calculation of clustering distances. To improve the analytics, **I would propose to implement pre-clustering methods like Canopy Clusters. (Canopy Clusters can process huge datasets efficiently. It is an unsupervised pre-clustering algorithm).** This can be used as a pre-processing step for K-Means algorithm or the Hierarchical Clustering Algorithm. **This will speed up the clustering operations on large datasets.**

**With Agile Approach, we can also implement any of the following clustering methods with the increase of samples or data volume.**

- **Subspace clustering,** (Subspace clustering is the task of detecting *all* clusters in *all subspaces*.)
- **Projected Clustering** (Projected clustering seeks to assign each point to a unique cluster)
- **Hybrid Clustering.**
- **Correlation Clustering.**

For agglomerative algorithms, it is a bottom up algorithm. It builds up clusters from single objects. Divisive clustering algorithms on the other hand break up cluster containing all objects into smaller clusters. **Divisive Algorithms can minimize cut based cost in place of K means, as it the latter maximizes only similarity within a cluster and ignores cost of cuts.**

**Choice of CHAID will depend if it's a dependence approach addressing the following questions.**

1) **Who will buy the product and who will not?**
2) **How many segments are there with distinct propensities towards the product?**
3) **How to reach a specific target of customers.**

**Data Considerations for CHAID**

Stated purchase intent

Exaggeration-corrected, derived purchase intent (Assessor®)

Conjoint/choice derived

May be something different than purchase intent:

•Retention

•Advertising response

•Targetability = difference in predicted usage and actual usage

The long list of demographics, media use, and channel use become potential drivers.

**Some of the practices that I would like to propose for clustering are:**

a) Precede interdependence (clustering) segmentation with focus groups

b) To address the differential bias which arises from segmentation, (Differential Bias arises in issues of scaling). A short 5 point short scale can be proposed to address this kind of problem. I like to propose the following scale: 1=totally disagree, 2 = somewhat disagree, 3=neutral, 4= somewhat agree, and 5 = totally agree. On this scale it is also helpful for each item to be extreme in nature. It is much clearer to interpret the difference between a somewhat agree and a totally agree response to an item.

c) Conjoint or discrete-choice based utilities can be used as basis of clustering.

d) We should try to cluster only unique dimensions. We can get this unique dimensions by factor analyzation. (From case studies I have concluded that most brands cannot manage more than 7-10 segments). Once the desired number of segments is determined, we can perform K-means clustering.

e) We should try to focus on the reliability and accuracy of the dependent variable in CHAID.

f) Heavy data cleansing can be avoided in cases of CHAID.

g) Some technical best practices of CHAID are :
   (i) Smallest child node should be set at approximately 5% of total N, and parent node at twice the size of the child node.
   (ii) Alpha set at 0.5, with Bonferroni adjustment turned off. **[40]**

h) The best way to show CHAID segments is usually through a table which shows the definition of each segment very distinctly.

The choice of algorithm must address the issues like

- Cost Functions to optimize
- Similarity measures between clusters.

With the right set of Algorithms, we can have a very much efficient data processing to address the complexities that arise.

## 4.11.4. Different types of modelling approaches

**[41]Descriptive modelling** presents the main features of the data. Data generated from a good descriptive model will show same characteristics as the real data. Techniques and algorithms which fits the descriptive models to data are described in the following.

- **a)** We estimate probability densities using either parametric or non- parametric approach.
- **b)** Some of the types of parametric modelling techniques that can be followed are:
  - **I)** Mixture models
  - **II)** Mixture models and Expectation Maximization. (Time complexity O(Kp2n); space complexity O(Kn). Can be slow to converge; local maxima )
- **c)** Non-parametric density estimation doesn't scale well.
- **d)** For determining number of clusters, we can use K-Means clustering, Hartigan Criteria, Partitioning Around Medoids (PAM), Clustering Large Applications (CLARA), Hierarchical Clustering (Agglomerative versus divisive, Generic Agglomerative Algorithm ) .
- **e)** For identifying number of clusters, we can select the Method of Mojena.
- **f)** AGNES (Agglomerative Nesting), DIANA (Divisive Analysis).
- **g)** Clustering Market basket data based on ROCK (Robust Clustering using linKs)
- **h)** Clustering in QUEst (CLIQUE)
- **i)** Balanced Iterative Reducing and Clustering (BIRCH).
- **j)** Some of the defining decisions taken for clustering using descriptive approach are
  - **I)** Scalability
  - **II)** Ability to deal with various attributes
  - **III)** Discovery of clusters with arbitrary shapes.
  - **IV)** Interpretability & Usability
  - **V)** Minimal requirement for domain knowledge to determine input parameters.

**Predictive Modelling** term for customer segmentation is called clustering. With clustering, we create customer segments instead of marketers. It can be thought of a type of auto-segmentation. Algorithms helps to perform these segmentation based on various variables. The most used clustering algorithms based on predictive modelling are stated as follows: **[42]**

1) **Behavioral Clustering**
   This approach enlightens about how people behave while purchasing.
2) **Product Based Clustering/ Category Based Clustering**
   This algorithm discovers what different groupings of product people buy from.
3) **Brand based clustering**

   This clustering tells us what brands people prefer.

The other types of modelling techniques are used in Propensity Models for Predictions and Collaborative Filtering for Recommendations.

**Types of Propensity Models are as follows:**

**(i) Lifetime Prediction Value**
**(ii) Share of wallet predictions.**
**(iii)Propensity to engage.**
**(iv)Propensity to unsubscribe.**
**(v) Propensity to convert**
**(vi)Propensity to buy**
**(vii)Propensity to Churn**

**Types of Collaborative Filtering Models are as follows**

**A. Up Sell Recommendations.**
**B. Cross Sell Recommendations.**
**C. Next Sell Recommendations.**

**[43]**Some of the technical approaches in Predictive Modelling Techniques are Neural Networks, Support Vector machines (SVMs), Neural Network & logistic regression models.

## Prescriptive Modelling

Prescriptive analytics is the final stage in business analytics. It is still in its emerging form. This genre of analytical modelling demands skills in mathematical sciences business rule algorithms, machine learning and computational modelling techniques. Prescriptive analytics can help to optimize your scheduling chain design in order to deliver the right products in the right time using the most optimized means.

Since Prescriptive analytics is still in its infancy, a lot errors are encountered in its iterations. But **maintaining an Agile approach towards these failures,** has resulted this to be one of the strongest modelling techniques on the rise. **[44]**

Prescriptive analytics is about applying logic and mathematics to data, with the goal to specify a preferred course of action- unlike other type of analytics the output is a decision. . It's about trying to find the best decision, where best is defined by you, whether that's lowest cost, most efficient process, higher revenue, or one that meets customer needs. Fundamentally the focus begins with the business decision. **[45]**

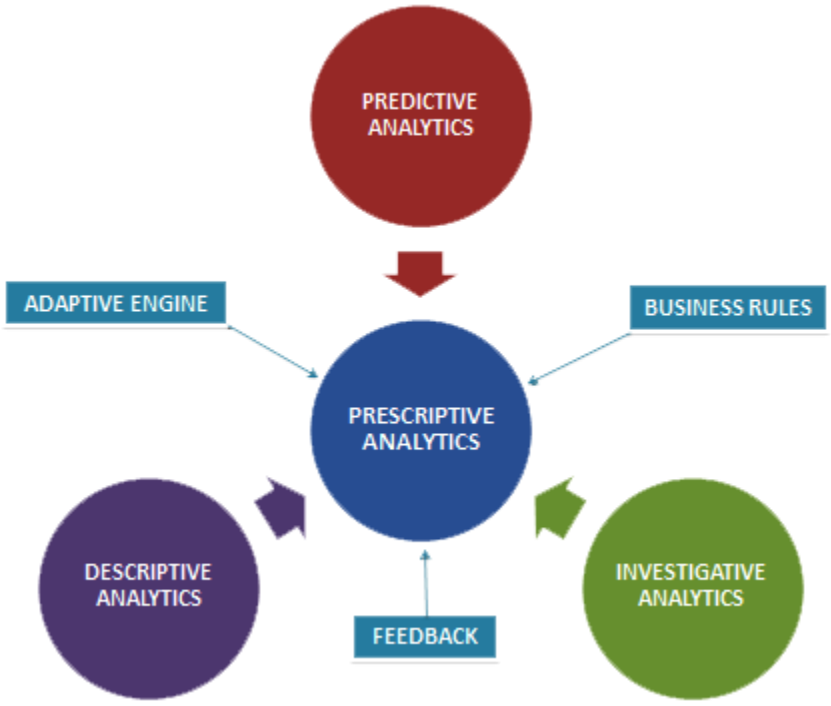


*Fig 4.10 Comparison of different modeling techniques.*

## 4.11.5 Outcomes and Benefits from Clustering.

The term cluster has been interpreted in various ways and there are many definitons present to define it. Some define clusters as a group of organisations in related industries having a common economic link, or, geographic concentrations of interconnected companies and institutions in a particular field. By extracting common themes from the case studies and literature reviews, I have noticed three underlying principles. These principles are present in different magnitues irrespective of their size, sector and structure. They are as follows:

**1. Commonality**,i.e,business functions in a common field.

**2. Concentration,** i.e, grouping of enterprises who can interact.

**3. Connectivity**,i.e,interconnected organizations with different types of relationships.

The reasons why clustering is cluster development policies can benefit from understanding process of clusteringare as follows:

1. Clusters are not static and do not have fixed boundaries.
2. Clusters have different stages of development. (**Agile approach can be implemented in developing a cluster as well as analyzing it.)**

Clustering helps to develop new products (based on research and development of customer analysis). A proposed four step process for tackling cluster development is given below:

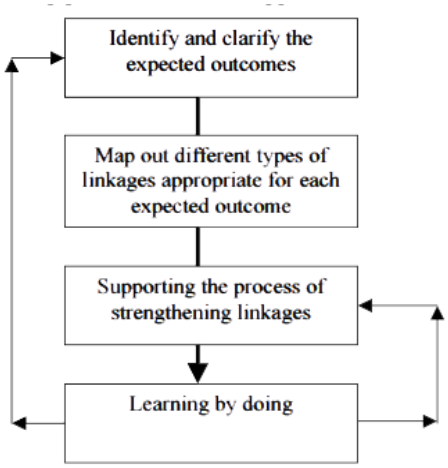


*Fig 4.11 Four Step Process for Cluster Support*

# *CHAPTER V. CONCLUSION*

From the study of various cases related to analytics, I have come to the conclusion that the data integration process in classical databases approach was very cumbersome and it poses a mammoth challenge to the present emerging Big Data Technologies. Along with the demand for intense analytical operations, the storage issue is slowly migrating to the cloud and NoSQL framework, allowing much faster deployment of analytics from anywhere in the world. This seamless procedure to integrate data is regarded extremely useful for the oncoming days of Big Data Challenges.

From analyzing the various cases in this thesis, we can conclude that there was quite a lot of commonality present, either in terms of Tool selection for data analysis, or selection of variables from datasets for clustering, or the choice of algorithms.

For the cases related to segmentation, most of them performed the same types of algorithm, for which time complexities, process complexities were present. Proposed steps to elucidate this issue has been described in details, from using pre-clustering methods like Canopy clustering. Also with the proper selection of algorithms, we will save time in performing analytics rather organizing the data. Algorithms have the ability to perform many operations together which would take a human, ages to process it.

Tool selection has also shown that it's not only the features that attracts an analyst, but the ease of use, scalability, modularity and ability to integrate with other software tools as well.

It has also been observed that Agile approach is very much applicable to most of the cases, and it can resolve the issues based on the demand of the customers. Agile approach can help to address the problems based on data acquisition, selection of segmentation variables, selection tool for analytics, modelling issues, and of course on the outcome of clustering as a whole. Although most of the domains are yet to integrate agile approach, it can expected that professionals in technical and non-technical domain taking methodology up is just a matter of time. So," Get Lean. Get Agile. Get Started "

# ***BIBLIOGRAPHY***


[1] http://www.gartner.com/it-glossary/big-data

[2] https://datafloq.com/read/big-data-history/239

[3] http://www.techrepublic.com/blog/big-data-analytics/10-emerging-technologies-for-big-data/

[4] http://www.techrepublic.com/blog/big-data-analytics/10-emerging-technologies-for-big-data/

[5] https://en.wikipedia.org/wiki/Big_data#Applications

[6] http://www.slideshare.net/Dell/big-data-use-cases-36019892

[7] https://en.wikipedia.org/wiki/Analytics

[8] https://www.thoughtworks.com/big-data-analytics

[9] http://data-informed.com/benefits-agile-analytics-development-right/

[10] https://www.thoughtworks.com/insights/blog/introducing-agile-analytics

[11] https://en.wikipedia.org/wiki/Agile_software_development

[12] http://www.oracle.com/us/technologies/big-data/bda-customer-segmentation-engines-2045188.pdf

[13] http://www.slideshare.net/gurmitcombo/case-studies-customer-marketing-analytics-for-retail

[14] http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=AB&infotype=PM&appname=SWGE_YT_YU_USEN&htmlfid=YTC03438USEN&attachment=YTC03438USEN.PDF

[15] ftp://public.dhe.ibm.com/software/be/pdf/CoronaDirect_Belgium_2010_SPSS.pdf

[16] http://www.ndm.net/datawarehouse/IBM/mediamath-case-study

[17] http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=AB&htmlfid=YTC03372USEN

[18] http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=AB&infotype=PM&appname=SWGE_YT_YU_DKEN&htmlfid=YTC03575DKEN&attachment=YTC03575DKEN.PDF

[19]http://public.dhe.ibm.com/common/ssi/ecm/yt/en/ytc03503fien/YTC03503FIEN.PDF

[20] http://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/too-much-of-a-good-thing.html

[21] http://www.pwc.com/us/en/advisory-services/case-studies/technology/bridging-big-data-divide.html

[22] http://www.pwc.com/ca/en/services/consulting/case-studies/case-studies-using-agile-analytics-to-drive-business-intelligence.html

[23] https://www.thoughtworks.com/insights/blog/enabling-agile-engineering-great-teams-mongodb

[24] http://www.sciencedirect.com/science/article/pii/S0957417408002212

[25] http://www.mrweb.com/mrt/seg10mar.htm

[26] http://www.consumerpsychologist.com/cb_Segmentation.html

[27] https://www.cs.rutgers.edu/~mlittman/courses/lightai03/jain99data.pdf

[28] https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Hybrid_Hierarchical_Clustering

[29] http://www.marcresearch.com/pdf/IIR05_Segmentation_FrankWyman_slides.pdf

[30] http://www.marcresearch.com/pdf/IIR_SegBP_FrankWyman_presentation_notes.pdf

[31] http://delivery.acm.org/10.1145/1410000/1409562/a16-peng.pdf?ip=131.175.28.198&id=1409562&acc=ACTIVE%20SERVICE&key=296E2ED678667973%2E7773E6D96819F65E%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=735224713&CFTOKEN=22912626&__acm__=1449150949_004d339a44a24449d0fbb7c3af911253

[32] http://delivery.acm.org/10.1145/2510000/2501982/a5-ramesh.pdf?ip=131.175.28.198&id=2501982&acc=ACTIVE%20SERVICE&key=296E2ED678667973%2E7773E6D96819F65E%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=735224713&CFTOKEN=22912626&__acm__=1449151066_a6256acf0715c63442cad2ac357ad513

[33] http://delivery.acm.org/10.1145/2350000/2345414/p106-mishra.pdf?ip=131.175.28.198&id=2345414&acc=ACTIVE%20SERVICE&key=296E2ED678667973%2E7773E6D96819F65E%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=735224713&CFTOKEN=22912626&__acm__=1449151190_89335b4120d799fc5451a6900e5ba04e

[34] https://en.wikipedia.org/wiki/CHAID

[35] http://cdn.intechopen.com/pdfs-wm/13162.pdf

[36] http://www.dataapple.net/?p=84

[37] http://www.slideshare.net/WhiteRavenPL/rfm-segmentation

[38] http://www.researchgate.net/publication/263329040_Data_mining_for_the_online_retail_industry_A_case_study_of_RFM_model-based_customer_segmentation_using_data_mining

[39] http://delivery.acm.org/10.1145/2650000/2644161/a19-singh.pdf?ip=131.175.28.198&id=2644161&acc=ACTIVE%20SERVICE&key=296E2ED678667973%2E7773E6D96819F65E%2E4D4702B0C3E38B35%2E4D4702B0C3E38B

35&CFID=735224713&CFTOKEN=22912626&__acm__=1449152630_6422ae3d0530a51cbe5e6c9a9fc8196b

[40] http://www.winsteps.com/winman/bonferroni.htm

[41] http://web.itu.edu.tr/sgunduz/courses/verimaden/paper/validity_survey.pdf

[42] http://www.stat.columbia.edu/~madigan/DM08/descriptive.ppt.pdf

[43] http://www.agilone.com/academy/the-definitive-guide-to-predictive-analytics-models-for-marketing/

[44] http://www.ibm.com/developerworks/library/ba-predictive-analytics2/

[45] http://gizmodo.com/how-prescriptive-analytics-could-harness-big-data-to-se-512396683

[46] http://www.information-age.com/it-management/strategy-and-innovation/123458977/insight-action-why-prescriptive-analytics-next-big-step-big-data