Table of contents

Table of contents

Introduction to Data

We live in a data rich, data driven world. Data is revolutionizing business in ways we never conceived. So much of what we do is being recorded and stored somewhere. Companies big and small, in traditional and non-traditional sectors, are using data to understand their customers better. Data is helping with better targeting and improved customer experiences. The insights gained from analyzing data is helping companies identify new growth areas and product opportunities, streamline costs, increase operating margins, make better human resource decisions and more effective budgets. Data is also impacting our world, our lives. Health care, the environment, travel….the list is endless.

Simply put, Data has begun a journey that will only grow in momentum for the next couple of decades. Experts predict that there will be a 4300% increase in annual data generated by 2030. By 2025, 10.4 million IT jobs globally will be created to support Big Data, generating 3.9 million IT jobs in the US. Every Big Data-related role will create employment for three people outside of IT, so over the next four years a total of 6 million jobs will be generated by the information economy in North America.

> By 2025, 10.4 million IT jobs globally will be created to support Big Data, generating 3.9 million IT jobs in the US.

At Splid Analytics we are pretty audacious. We will even go so far as to say that we are looking at a future where every facet of our lives will be driven by analytics and the success of any business will depend on the effectiveness and aggressiveness of its data initiatives.

But in spite of all this furor about Data and Data Analytics, there are many who still don't understand these data related buzz words. This book by Splid Analytics aims to give one an understanding of the many aspects of data, data analytics and the many popular tools and technologies used, while also explaining its application in various industries and across business functions. The book also talks about why Data Analytics is the hottest career of the 21st century and what the future holds in store for those who invest in gaining these all important data analysis skills. We also introduce you to the concept of Big Data and give you a host of resources that will enhance your learning.

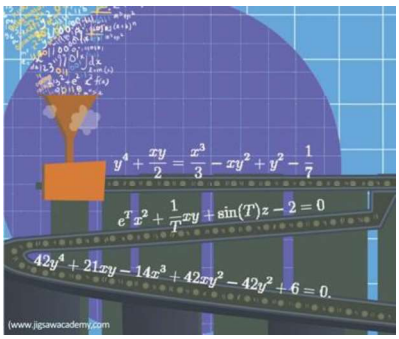
This book is also a useful companion to those of you enrolled in Splid's 'Analytics for Beginners' Course. You can use this book to compliment your learning and better understand the world of Data Analytics.

We hope you enjoy the introduction guide.
Data Analytics Team Splid Nigeria

CHAPTER 2

What is Data Analytics?

Data Analytics can range from a simple exploration into how many sales of a particular product were made last year to a complex neural network model predicting which customers to target for next year's marketing campaign.



In 'Competing on analytics', Thomas Davenport defines data analytics as "the extensive use of data, statistical and quantitative analysis, exploratory , predictive models, and fact based management to drive decisions and actions."

In layman's terms it can be defined as "the analysis of data to draw hidden insights to aid decision making".

There is a little bit of analyst in everyone. Analytics is an integral part of most businesses. You do not need to be an "analyst" to do analysis. Analytics is an essential skill for running any kind of business successfully. Common applications of analytics include the study of business data using statistical analysis in order to discover and understand historical patterns with an eye to predicting and improving business performance in the future.

There is a little bit of analyst in everyone. Analytics is an integral part of most businesses.

A call centre manager analyses his team's performance data for resource optimization. HR managers use employee data to predict attrition. Marketers use sales and marketing ROI data to decide the optimal allocation of the marketing budget.

In today's workplace, every manager and most contributors are leveraging analytics in one way or another.

2.1 Descriptive vs Predictive

There are essentially two kinds of Data analytics

1) Descriptive Data analytics
2) Predictive Data analytics.

Descriptive analytics describes what has happened in the past. Predictive analytics predicts what will happen in the future.

For example, a sales report of a company, say Pepsi. This report will tell you how many units of Pepsi were sold, where they were sold, what price and a lot of other things. All of this is information coming from the data. All you are doing is slicing and dicing the data in different ways, looking at it from different angles, along different dimensions etc. There is very little statistics involved in descriptive analytics and so you don't really need to be a statistical wiz to be able to do effective descriptive analytics.

While descriptive analytics is a very powerful tool, it is still giving us information about the past. Whereas, a business owner's primary concern is the future. If I run a hotel, I want to be able to predict how many of my rooms will be occupied next week. If I am a drug company, I want to know which of my under-test drugs is most likely to succeed. This is where predictive analytics comes in.

Predictive analytics works by identifying patterns in historical data and then using statistics to make inferences about the future.

Predictive analytics works by identifying patterns in historical data and then using statistics to make inferences about the future. At a very simplistic level, we try to fit the data into a certain pattern and if we believe that the data is following a certain pattern then we can predict what will happen in the future.

Let us take a couple of small examples to illustrate this point better. The first example we take is from the world of retail.

Retailers are very interested in understanding relationships

between products. They want to know if a person buys product A, is he also likely to buy product B or C. This is called product affinity analysis or association analysis and is commonly used in the retail industry. It is also called Market basket analysis. However, truly speaking, Market basket analysis or MBA refers to a set of techniques that can be applied to analyze a shopping basket or a transaction. Product affinity analysis is one of the techniques that form a part of MBA.

Take the example of fruits. We usually buy fruits in groups. That is, if I am buying one type of fruit, I am likely to buy other kinds as well. Thus it makes sense to always have all the fruits together in one place in a retail shop.

If I am buying bread, I am likely to buy eggs with it as well. Of course, many of these associations are fairly intuitive but once in a while we come across relationships that are not obvious and would not have been uncovered but for the data analysis.

We will talk about this technique in more detail later on in the course when we deal with the popular predictive modelling techniques. At this point, we are just using it as an example of how predictive analysis is used by businesses.

Let us take another example, this time from the telecom industry. Customer attrition (meaning customers who leave the telecom company's services – for example someone moving from Airtel to Hutch) is an important issues for telecom companies. They would love to be able to predict which of their customers are likely to leave their service in the future. Predictive modelling enables them to do that.

For example, if a customer's phone usage has gone down drastically in the last 1-2 months, it could be an indicator that they are likely to attrite. This is just one factor that we are talking about. In reality, a combination of such factors usually acts as a more effective indicator.

So there we have it – the 2 popular types of analytics – descriptive and predictive.

1) Descriptive analytics describes what has happened in the past.
2) Predictive analytics predicts what will happen in the future.

In addition to these 2, there is a third type of analytics which came into existence very recently (probably just a decade old). This is called prescriptive analytics. Prescriptive analytics goes beyond predictive analytics by not only predicting what will happen but also suggesting the most optimal decisions given what could happen and showing what will happen under different scenarios. It includes concepts like optimization and simulation.

CHAPTER 3

How is Data Analytics Used?

3.1 Application in Industries

Today Analytics is used for strategic, operational and tactical decision making across industry verticals such as Retail, E-commerce, Banking and Finance, Sports, Telecom, Manufacturing and Retail.

3.1.1 Retail

Today's customer is savvy, impatient and busy. They want instant gratification and excellent customer service. In order to compete and stay one step ahead, retailers need to have a 360-degree view of the customer. Retail analytics helps businesses get deep insights into customer behavior. It helps them understand their customer's requirements more precisely, while also helping them to bring in more of the right kind of customers.

It gives insights such as:

- How to increase margins at a product-level?
- Insights into your customer profile that helps answer questions like who they are and why they make certain purchases (Market Basket analysis)
- Identify items that are likely to be purchased together.
- Which marketing strategies work better than others?
- ROI of marketing spend
- Optimal Pricing
- What promotions and offers to employ in each store?
- Store wise product-mix
- Personalized offers
- Efficient stock strategy

3.1.2 E-commerce

Electronic commerce, or e-commerce, involves the sale of goods and services via electronic means. E- commerce analytics helps organizations convert data to insights, leading to better decision making for better business outcomes, resulting in maximizing revenue and profitability. With use of analytics, businesses can collect a wealth of information about their site, their visitors and where they came from, and use it to find new customers and increase conversions. E- commerce businesses primarily use analytics to understand:

- Acquisition - how your visitors and customers found and arrived at your site.
- Shopping and purchasing behavior: how users engage with your website, which products they view, which ones they add or remove from shopping carts; along with initiating, abandoning, and completing transactions.
- Economic Performance – how many products the average transaction includes, the average order value, refunds you had to issue.

3.1.3 Finance

The global financial analytics market is one of the fastest growing sectors of the data industry. Organizations big and small are investing in financial analytics tools and technologies to solve specific business problems, reduce costs, improve budgets and get insights into future financial scenarios.

Typically financial analytics includes

- Risk analysis
- Working capital management
- Fraud detection and prevention
- Shareholder metric analysis

3.1.4 Sports

In 2002, the Oakland Athletics Major League Baseball team won 103 games in the regular season, the same amount as the New York Yankees, despite its comparatively meagre resources. The secret was the use of sophisticated statistical models to analyze professional players' performance that allowed their coach to spot good players that others overlooked. Thus sports analytics was born and today coaches, players and sports managers have begun to use analytics from everything to scoring , signing contracts and preventing injuries

3.1.5 Others - Healthcare, Education, Telecom etc

Today analytics has begun to impact almost every industry including healthcare, education and telecom to name a few. Analytics can be used for evidence based medical care, improved patient care, predicting outbreaks of diseases and reducing hospital operating costs.

Analytics is also being used to improve teaching practices. It also enables teachers to better monitor student progress, personalize learning and improve educational institutions operational efficiencies.

In the telecom industry analytics is fast gaining much ground. Operators are using analytics to drive revenue, reduce churn and improve network performance.

3.2 Application in Business Functions

Apart from industries, Analytics can also have a huge impact on the key business functions of almost any organization.

3.2.1 Marketing

Understanding customers and how to find more people like them is the key to sustainable growth. Analytics can not only help companies do this but it can add value to other marketing functions as well, by gathering data across all marketing channels and consolidating it into a common marketing view. It helps measure, manage and analyze marketing performance to maximize its effectiveness and optimize return on investment (ROI).

Some of the key marketing questions that analytics can help answer are:

- How are our marketing initiatives performing today?
- Which of them are viable in the long run?
- How can we improve those which are not effective?
- How do our marketing activities compare with our competitors'?
- What can we learn from our competition?
- Are our marketing resources properly allocated?
- Are we using the right channels?

To get the most benefit from marketing analytics, companies need to use an array of analytic tools that can give a single window view of what was (initiatives in the past), what is (performance of current initiatives) and what will be(data driven predictions).

3.2.2 Sales

Davenport describes sales as one of the more conservative business functions where adopting analytics initiatives is concerned. Though sales analytics can help identify, model, understand and predict sales trends and outcomes we see very few companies realizing its potential to aid sales management. However the potential is huge and over the next several years sales analytics will be one of the most important domains for Data Analytics and Big Data.

What sales analytics can essentially do is:

- See what goods and services have and have not sold well.
- Determine optimal inventory
- Measure the effectiveness of the sales force and determine optimal sales force size
- Sales incentive cost analysis
- Competitor sales analysis

3.2.3 Supply chain management

Today new advanced analytic tools and technologies makes it possible to dig deeper into supply chain data in search of savings and efficiencies. Also several devices like the Radio Frequency Identification devices and GPS tracking devices are available at more reasonable costs and can help companies better monitor their supply chain. With the aid of such devices and the use of advanced analytics they can more easily identify supply chain problems and take corrective action in real time.

Supply chain analytics helps monetize and optimize:

- Current inventory status
- Forecasts
- Demand planning
- Sourcing
- Production
- Improved worker productivity measurement
- Transportation routing

3.2.4 HR

HR analytics helps managers by creating a single view of all relevant workforce and other HR related data. These insights can be used to make business decisions that drive business processes and initiatives and improve profitability.

Some of the key areas where workforce related data driven analytics can be used are:

- Talent acquisition and retention
- Attrition
- Headcount Management and Workforce Optimization
- Optimization of Compensation and Benefits
- Build Leadership

- Performance and Career Management
- Training and Development

3.2.5 Others - Finance, IT, Manufacturing and Strategy

As we earlier mentioned there is potential for analytics in every function in an organization. Other than the above mentioned ones, analytics can also improve the effectiveness of functions such as Manufacturing, Strategy, IT and Finance. From detecting fraud, risk management, data security, finding unexplored strategic opportunities, assessing manufacturing quality and using data from sensors on manufacturing equipment, analytics can provide insights that can help managers make better decisions that will improve the profitability of their businesses.

The key to real success lies in using an integrated approach to analytics. Experts say that the best decisions are founded on an integrated data environment that seamlessly assimilates relevant internal and external data sources and applications. Such an approach helps ensure that all available and relevant internal and external, structured and unstructured, data sources can be used to deliver insights that can be used to strengthen the company in the long term, while also improving profitability, employee satisfaction and customer delight.

CHAPTER 4

Careers in Data Analytics

Choosing a career is probably one of the most difficult decisions you ever make. Not only does it largely influence your happiness quotient but it also determines your earning power for the rest of your life. It is said that today there are over six hundred careers in the world to choose from, making the task of knowing just which one is the right one, all that more difficult. Wouldn't it be great if you could just do stuff that you were really good at and that you enjoyed and then get someone to pay you for doing it? Well if you are someone who naturally loves playing with numbers, are perceptive, intuitive, curious, inquisitive and persistent, then you just might be in luck!

It is said that today there are over six hundred careers in the world to choose from, making the task of knowing just which one is the right one, all that more difficult.

As the buzz words Data Analytics and Big Data consume the business world, data scientists continue to be in big demand and will continue to be in demand for a long time to come. The International Data Corporation (IDC) recently forecasted revenues for the global Big Data technology market reaching $23.8 billion by 2016. This will be the result of an annual growth rate of 31.7 % which is a staggering sevenfold of the rate of the entire ICT market. In the next couple of years the US alone will have a shortage of 1.4 to 1.8 lacs people with analytical skills. Data analysts or data scientists as they are more commonly referred to these days are being offered big salaries, faster growth paths and challenging and exciting work environments.

Yes, the business world will continue to create data at phenomenal rates. Imagine being a part of such growth? For those who do have the intrinsic qualities of a data scientist, now is the time to explore the field further. Learn to talk to data, invest in acquiring the necessary technical skills and begin a truly challenging and exciting career, one which can propel you towards a future that is both exhilaratingly satisfying and monetarily beneficial.

4.1 What does a data scientist do?

Data analysts and scientists perform a variety of tasks related to collecting, organizing, and interpreting statistical information. This would of course vary depending on the exact nature of the job and the sector or industry the analyst is working in, but simply put data scientists assign numerical values to different business functions, and are responsible for identifying efficiencies, problem areas, and possible improvements. Let's quickly list some of the basic tasks most data scientist performs:

1) Explore data, play with data, understand data and compile information from data Ask
2) lots of questions Define
3) and test hypothesis
4) Develop predictive models and algorithms'
5) Provide powerful Visualizations and create interesting business stories that will aid business decision making

Data analysts and scientists perform a variety of tasks related to collecting, organizing, and interpreting statistical information.

4.2 A day in the life of a data scientist

As part of report on niche analytics companies that we at Splid Nigeria published, we spoke to two data scientists and asked them to describe a typical working day. Here is what they said:

Eliza Matthen, Marketelligent

" On a normal day I come to office around 10am, I download the required data and process the data in the required format for which I use SAS or SQL. Then I create reports using Xcelsius or PowerPoint to make it visually appealing nd also provide insights/summary on the reports. This normally ends around 6.30 pm.

It all really depends on our client's needs. Some days it might go on a little longer as we have calls with our US based clients. We also have team meeting on some days where each team gives updates about their projects. There are also training/knowledge transfer that happens simultaneously with our usual work."

Kathirmani Sukumar, Gramener

“ Broadly speaking my day is divided between three tasks:

Analyzing data: This is typically done during the initial stage of a project. Usually we get sample datasets from the customer and with the help of the concerned domain experts from the customer side; I understand the nature of the data. Then using tools like Python, Gramex (our own visualization product), Excel and R, I analyze the data and derive insights. Analysis usually will be a mix of descriptive statistics, classifications, causal analysis. Toughest part of the work is creating a visual which conveys the insights effectively.

Interacting with the customer: After analysis and deriving insights, I am often seen at the customer premises, educating the customer about my analysis and usage of our visualizations. This has been the challenging part of my job, because generally the target audience is top level management and I have been successful in convincing them which is usually difficult.

Automating: Interesting part in a day is when I spend time automating all my analysis and visuals using our product and deploy it in the client's machine. Being incharge of the journey of raw data to the end product is a fascinating experience.”

4.3 How does one become a data scientist?

Today there are a plethora of institutions that offer a variety of courses, from foundation level to more advanced courses, certifications and degrees in data analytics. There are also several free resources available online that can aid your learning (see APPENDIX 1). But to begin with, if you are wondering how you can set the path for a career in analytics, here is what you can do to get started:

1) Develop an Integrated Analytical Skill Set.

It is essential for a Data Scientist to have expertise in diverse analytical tools. The language of SAS is one of the more popular analytics tools you can learn. However R and Hadoop are gaining momentum and today companies are using a combination of tools and technologies for data analysis. Hence it would be wise to develop expertise in a few of the popular ones. In addition a data scientist would also need a comprehensive and through knowledge of NPL software, languages like QL, Perl and Python.

2) Pick up Some Visualization Skills:

One of the core functions of a data analyst is to visually anatomize exploratory data, and then communicate their findings and insights using interesting and innovative visualization tools. As a data scientist your main objective is to bring insights to the management, so as to enable them to make better business decisions. What use are excellent data mining and modeling tools, if the results of an analysis are poorly visualized? It is thus imperative that data scientists are apt at the art of visual storytelling and are able to creatively and persuasively communicate, the stories their data tell.

3) Get exposure to Large Data Sets

As a data scientist you will need to work on sometimes extremely large data sets. Hence it's beneficial to get some exposure to working with large amounts of data, preferably even some data mining algorithms

4) Sharpen your Business Skills

In the world of data analytics, business skills like negotiation, persuasion, creativity and leadership are important to have. As a data scientist you need to be able to feel the pulse of the business, understand business terminologies, have good organization skills and be able to drive and influence change. Also amidst all the sorting, mining and visualizing data, skills like planning and organizing go a long way.

5) Keep Abreast of Innovations and News in the Analytics Industry

Lastly, it is imperative that you develop a keen understanding of the data analytics industry and keep abreast of the latest advancements. Take the time to engage and connect with the data analytics community. Subscribe to journals, download free ebooks and follow interesting blogs by analytics experts. Take advantage of the wealth of free but quality information out there, so that when you are ready to apply for those data analyst jobs, you are well prepared and truly an analytics expert in your own right.

4.4 Innate abilities

If you are someone who loves playing with numbers, then analytics is definitely a career to consider. Did you love playing book cricket as a child? Do you find patterns in number plates of cars that you drive by? Are you that someone who is always coming up with quirky data facts at a party? If yes then there is a data scientist lurking inside of you waiting to come out.

Data scientists also need to be curious. They must continuously explore and ask questions of their data. They need to stare at figures and facts and spot trends, do painstaking'what if' analysis and question assumptions and processes. They must be fueled by a desire to dig deep beneath the first layer to discover the root of the problem.

Here are some of the innate qualities most data scientists have:

1) An inherent aptitude for numbers
2) Curiosity
3) Born knack for storytelling
4) Good communication skills
5) Business savviness
6) Team skills

4.5 Technical skills

Data Science in the form it is today requires those who work in it to be familiar with popular analytics tools and technologies like SAS, R, Hadoop, Java, Python, SQL, Hive, and Pig. Of course basic programming skills and understanding of computer science in general will give one a firm foundation with regard to gathering data and dissecting it to find useful insights.

In a nutshell here are the technical skills a budding data scientist needs.

Statistical Skills, Algorithms, Machine Learning and Mathematics

At the very core, a data scientist needs to be able to use R, Excel, SAS, or other popular tools to piece together data and discover potential patterns and correlations through statistics. The mandate is clear- If you can't use the tools, you can't analyze the data. However a data scientist needs to know correlation, multivariate regression and other statistical aspects of modeling to be able to use those tools effectively.

Business Domain Expertise

A data scientist need to have a thorough understanding of the domain and business they are working in, across all divisions, from marketing, to sales, to distribution and supply chain, to operations, pricing, products, and finance. This will come from studying and asking lots of questions while working in that particular field.

Programming Expertise

A data scientist has to be is a programming expert. Even if you don't have a computer science , you need to be comfortable designing and programming in a variety of languages including Java, Python, C++ or C#. You need to be able to determine the right software packages or modules to run, be able to modify them or even design and develop new computational techniques to solve business problems (e.g., machine learning, natural language processing, graph/social network analysis, neural nets, and simulation modelling).

Visualization Skills:

The ability to visualize and communicate data is a critical function of the data scientist. Not only must data scientists be able to visually anatomize exploratory data, but they must also have the ability to visually communicate their findings. The data scientist should be able to take statistical and computational analysis and turn it into graphs, charts, and animations; create visualizations (e.g., motion charts, word maps) that clearly show insights from data and corresponding analytics; and generate static and dynamic visualizations in a variety of visual media.

4.6 Career path in analytics

As a data scientist, you career path depends on the company you are in. However invariably in an analytics services company, the career path would be similar to an IT services roles i.e BD./Project management/Vertical leadership. But in a traditionally structured company these are the progressive rungs of the career ladder you will find yourself climbing.

The Analyst

The entry level role in the field of analytics is that of an analyst. This role offers you a chance to work on projects in a team, building expertise in the domain as well as the analytical tools and techniques. Basic knowledge of statistics is mandatory. You need to be familiar with widely used statistical concepts like p-value, probability distributions, chi-square etc. In addition, knowledge of common analytical techniques like logistic regression, clustering and decision trees is highly desirable.

The Senior Analyst

You will spend anywhere from 18 to 36 months in your first role before moving to the next level i.e. a senior analyst. If you have a PhD, you could directly join at this level. In this role, you will be expected to work independently on projects, even leading some of the smaller ones. You will continue to learn new methodologies as well as build domain knowledge.

The Team Lead

After spending another 2 to 4 years in this role, you will move to a team lead or lead consultant level. At this stage, many people will move into a people management track where they get to demonstrate their leadership and people management skills. However, some people with a more technical bent may choose to become subject matter experts (SMEs) instead. As an SME, you will provide leadership in defining and executing analytic methodologies for specific projects as well as develop new techniques to improve current processes. At this level, you will also be expected to manage most of the communication with clients regarding the projects you are working on.

The Manager

The next role in your career will be that of a manager. As an analytics manager you will be responsible for a team of 5 to 30 people. You may have 1 or more team leads working under you. If you are on the SME track, you may not have a permanent team assigned to you but you would be expected to provide thought leadership on individual projects.

The Associate Vice President

The next role is that of a senior manager followed by the Associate vice president (AVP). Growth in analytics is largely merit-based. You would need to demonstrate strong quantitative skills and an analytical aptitude in the early years. However, as you move up, communication and people management skills become differentiating factors.

4.7 Salaries in analytics

All across the globe, the data analytics industry is fired up. Over the next few decades, we will continue to see data being used in ways we never imagined and data scientists will continue to be in big demand. As a lot more data analytics start ups raise huge sums of money, big corporates will consolidate their data operations and companies will set up a lot more offshore data centers in countries like India and China. The analytics job market will be more potent than it ever was and salaries for those working with data will only increase dramatically.

According to an InformationWeek salary survey conducted from November 2013 to February 2014 with 11,662 full-time information technology respondents across 23 categories, most IT skill sets were in demand, though most salaries were merely keeping pace with inflation. However, the study showed that Big Data practitioners and data scientists are two emerging categories that, while their job descriptions are not consistently defined, are two titles at the top of the pay scale.

The studies showed that the median staff salary for data scientists is $120,000, while the median salary is $87,000 for BI/analytics, $90,000 for Big Data and $100,000 for data integrating/warehousing.

The median management salary for data scientists is $160,000, while the median management salary is $110,000 for BI/analytics, $145,000 for Big Data and $120,000 for data integrating/warehousing. Comparatively, the median base salary for a CIO is $150,000, according to the InformationWeek survey.

This clearly shows that data scientists are leading the compensation package!

KDNuggets issued a 2014 salary survey that included some international variants. The data points weren't exactly the same as Burtch Works, but were in the same ballpark. Hands-on data scientists in the U.S. reported average salaries of $118,000; management level came in at $140,000.

According to Jigsaw's 2014 Analytics Salary Report, the average analytics salary in India is 11.1 lakhs per annum. Those professionals who are skilled in SAS, R and Hadoop earn a 10- 20% premium in salary, averaging about 11.6 lakhs per annum.

CHAPTER 5

Popular Analytics Tools

5.1 Paid Tools

SAS

SAS is a software suite that can mine, alter, manage and retrieve data from a variety of sources and perform statistical analysis on it. SAS provides a graphical point-and-click user interface for non-technical users and more advanced options through the SAS programming language. SAS programs have a DATA step, which retrieves and manipulates data, usually creating a SAS data set, and a PROC step, which analyses the data.

WPS

WPS can use programs written in the language of SAS without the need for translating them into any other language. In this regard WPS is compatible with the SAS system. WPS is a language interpreter able to process the language of SAS and produce similar results. It is sometimes used as an alternative to SAS as it is relatively cheaper.

MS Excel

Microsoft Excel is a spreadsheet application developed by Microsoft for Microsoft Windows and Mac OS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. It has been a very widely applied spreadsheet for these platforms, especially since version 5 in 1993, and it has replaced Lotus 1-2-3 as the industry standard for spreadsheets. Excel forms part of Microsoft Office.

Tableau

Tableau Software is an American computer software company headquartered in Seattle, Washington. It produces a family of interactive data visualization products focused on business intelligence. Tableau offers five main products: Tableau Desktop, Tableau Server, Tableau Online, Tableau Reader and Tableau Public.

Pentaho

Pentaho is a company that offers Pentaho Business Analytics, a suite of open source Business Intelligence (BI) products which provide data integration, OLAP services, reporting, dash boarding, data mining and ETL capabilities. Pentaho was founded in 2004 by five founders and is headquartered in Orlando, FL, USA. The Pentaho suite consists of two offerings, an enterprise and community edition. The enterprise edition contains extra features not found in the community edition.

Statistica

STATISTICA is a statistics and analytics software package developed by StatSoft. STATISTICA provides data analysis, data management, statistics, data mining, and data visualization procedures. STATISTICA product categories include Enterprise (for use across a site or organization), Web-Based (for use with a server and web browser), Concurrent Network Desktop, and Single-User Desktop.

Qlikview

Qlikview is a business intelligence software from Qlik. It helps its users understand the business in a better way by providing them features like consolidating relevant data from multiple sources, exploring the various associations in the data, enabling social decision making through secure, real-time collaboration etc.

Qlikview

Qlikview is a business intelligence software from Qlik. It helps its users understand the business in a better way by providing them features like consolidating relevant data from multiple sources, exploring the various associations in the data, enabling social decision making through secure, real-time collaboration etc.

KISSmetrics

KISSmetrics is a person-based analytics product that helps users identify, understand, and improve the metrics that drive their online business. They make it simple to get the information users need to make better product and marketing decisions.

WeKa

The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modelling algorithms implemented in other programming languages, plus data pre-processing utilities in C, and a Make file-based system for running machine learning experiments. This original version was primarily designed as a tool for analysing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research.

BigML

BigML is a Corvallis, Ore.based startup with a SaaS-based machine learning platform that allows everyday business users to create actionable predictive models within minutes.

5.2 Free Tools

R

R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. There are some important differences, but much code written for S runs unaltered. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++, and FORTRAN code can be linked and called at run time. Advanced users can write C, C++, Java or Python code to manipulate R objects directly.

Google Analytics

Google Analytics is a service offered by Google that generates detailed statistics about a website's traffic and traffic sources and measures conversions and sales. The product is aimed at marketers as opposed to webmasters and technologists from which the industry of web analytics originally grew. It's the most widely used website statistics service.

Python

Pandas (python data analysis library) is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. It provides tools for reading and writing data between in-memory data structures and different formats: CSV and text files, Microsoft Excel, SQL databases, and the fast HDF5 format and also intelligent data alignment and integrated handling of missing data: gain automatic label-based alignment in computations and easily manipulate messy data into an orderly form can be done using python

Spotfire

TIBCO Spotfire is an analytics and business intelligence platform for analysis of data by predictive and complex statistics. During the 2010 World Cup, FIFA used this software to give viewers analytics on country teams' past performances.

CHAPTER 6

Future of Data Analytics

The reliance on data driven decision making will continue to grow. Just like the widespread usage of metrics and reports today, companies will start expecting to see some predictive analytics insights as part of regular dashboards.

As analytics becomes more and more prevalent in the corporate consciousness, a basic awareness and understanding of analytical techniques will become a required skill for career growth at the middle to senior management tiers, irrespective of industry and function. There will also be an increased demand for some super specialized roles. These will require intensive expertise with programming and technology to support the actual analytics implementation.

In the next decade we will witness technological advances that will play an increasingly important role in the ability of companies to mine data for real time insights and actions in the context of the rapid pace of data produced and the variety of data that is being captured.

CHAPTER 7

Introduction to Big Data

Wild-raft is a 500 store retail chain that sells equipment for adventure sports such as trekking, climbing, kayaking etc. Ten years ago, Wild-raft implemented a loyalty program in its stores. This program enable Wild-raft to collect data on its customers – data that provided invaluable insights about their customers. Wild-raft used these insights to service their customers' needs better. They were thus able to out-perform their competitors and grow at a rapid pace.

After 6 years of expansion, their growth started to slow down four years ago and now they have a stagnant business. Their competitors have caught up with them in terms of analytic capability and have similar insights about the buyers. On top of that, they are facing increased competition from online retailers who are equipped with deeper insights about the customers because they have more comprehensive data on their online shoppers.

While Wild-raft's loyalty program has helped them greatly, it still accounts for only about 20% of their revenue. They have limited information about the remaining 80% of their buyers. Online retailers, on the other hand, are able to create detailed profiles about 100% of their buyers – even things like which category is a person interested in, what does he examine but not buy and many other pieces of information that are much easier to get if your store is virtual.

While traditional analytics has helped Wild-raft in the past, it can only go so far. Wild-raft now needs to take their customer understanding to the next level.

Need for Big Data

Wild-raft introduces video cameras in all their stores to capture the customers' buying behavior. Video cameras can track the behavior of the customers when they enter Wild-raft stores – where they stop, what they see, how long do they take to evaluate and so on.

Now the video content generated by these cameras is far more in volume than any human viewers can digest. It needs to be analyzed using tools that use machine learning algorithms to analyze and make sense of this data.

Further more, this information, while valuable in itself, can offer immensely more value when combined with other data sources that the company has access to. This includes –

- The internal databases which record customer purchases,
- The loyalty data that has customer information
- Social media data such as data from twitter and facebook and even forums and online communities where adventure sports lovers come together and share information

If the company can find a way to combine these disparate sources of data – from in-store videos to loyalty data to online text and image data – the combined power of this information could be enormous. It will give far more powerful insights than what can be gained from just the internal databases of the organization.

In order to meet its requirements, Wild-raft needs an analytics platform that has the capability to handle:

a) Massive amounts of data
b) Varied data such as video files to sql databases to text data
c) Data that comes in at varying frequency – from days to minutes

Big Data analytics platforms are designed to serve such needs of today's businesses. They have the ability to deal with all 3 of the above mentioned conditions and thus are able to offer businesses far more value than what a traditional analytics system can.

After Wild-raft implemented the Big Data analytics platform, they got new insights about their customers. They learnt about product features that are important to their customers and they were able to collect and analyze instant feedback from their customers through the social media data.

This helped Wild-raft offer better service to their customers and once again differentiate themselves from their competitors – further consolidating their position as the market leader.

This is an example of how Big Data is making an impact on businesses – giving them access to information they never had before, faster than they ever had before.

7.1 What is Big Data

It is now time to answer an important question – What is Big Data?

In simple terms, Big Data is data that has the 3 characteristics that we mentioned in the last section –

- It is big – typically in terabytes or even petabytes
- It is varied – it could be a traditional database, it could be video data, log data, text data or even voice data
- It keeps increasing as new data keeps flowing in

This kind of data is becoming common place in many fields including Science, public administration and business.

The ability to harness such data for better decision making is therefore in great demand in today's world.

7.2 Where is Big Data used?

Big Data is most prevalent in consumer-centric industries that typically generate large volumes of data. Examples of such industries are –

- Consumer products such as Proctor & Gamble
- Credit card and Insurance such as Capital One and Progressive Insurance
- E-commerce companies such as Amazon, Netflix and Flipkart
- Travel and leisure such as United Airlines and Caesars Casino
- Public utilities such as electricity companies

Big Data is also becoming increasingly important in industries such as –

- Telecom
- Media and Entertainment
- Education
- And healthcare

Within each of these industries, Big Data can be applied to various functions such as –

- Marketing – for example social media analysis to understand customer pulse
- Supply chain – for example better inventory management through GPS data analysis
- Finance – for example for fraud control
- Manufacturing – for example, to link manufacturing operations with the supply chain for better optimization

In this section, you have seen industries and functions where Big Data is making a significant impact.
Now let us get an overview of some of the technologies that are driving the Big Data revolution.

7.3 Big Data Technologies

'Big Data' as a term refers not only to massive data sets but also to the group of technologies that enable its analysis. Therefore, technology is an important part of 'Big Data'.

Perhaps this is why anyone looking to learn about Big Data will find themselves very quickly surrounded by a number of strange names referring to even stranger technologies. Big Data seems to have more than its fair share of languages, platforms and frameworks.

It is difficult for a beginner to understand the exact role each of these technologies play in Big Data analysis. Some of them complement each other, some are based on others and some are just substitutes of others.

In this section, we will familiarize ourselves with the various Big Data technologies and how they connect with each other.

MapReduce

To understand the beginning of Big Data technology, we will need to go back to 2004 when 2 Googlers – Sanjay Ghemawat and Jeffrey Dean wrote a paper that described how Google used the 'Divide and Conquer' approach to deal with its gigantic databases. This approach involves breaking a task into smaller sub-tasks and then working on sub-tasks in parallel, and results in huge efficiencies.

This approach which they called "MapReduce" forms the basis of some of the most popular Big Data technologies today. We will get a more comprehensive understanding of the "MapReduce" approach or framework in the next section.

Open source software enthusiast 'Doug Cutting' was one of the guys deeply inspired by the Google paper. Doug had been working on creating an open source search engine and had been struggling with scale issues for the last 2 years. He was able to scale his engine to process a couple of hundred million web pages but the requirement was for something 10,000 times faster than this. This is the computing power Google generates when it processes the trillions of webpages in existence.

Hadoop

Doug realized that the MapReduce framework was ideal for processing large amounts of data. For the next 2 years, Doug and his partner went about creating an Open source file system and processing framework that later came to be known as Hadoop. This formed the basis of their search engine "Nutch". While the original Google file system was based on C++, Doug's hadoop was based on Java. Doug and his partner were now able to put together 30 to 40 computers and run Hadoop on this cluster. Using Hadoop and its underlying MapReduce framework, Doug was able

to significantly enhance the processing capability of “Nutch”. So much so that it generated interest from another search engine giant “Yahoo”. Yahoo could see great potential in Hadoop and wanted to build out this open source technology. Doug wanted a chance to work on clusters that had tens of thousands of machines instead of his 40.

Doug joined Yahoo.

It took years of hard work not just from Yahoo but also from the global open source community to get Hadoop to where it is now – the most popular open source Big Data solution for businesses.
Over time, other companies such as Microsoft, Intel, Cloudera and EMC have all created their own versions of hadoop and offer customized solutions on these platforms.

Pig

As hadoop began to be implemented on a larger scale, Big Data specialists soon realized that they were wasting far too much time on writing MapReduce queries rather than actually analyzing data. MapReduce was long and time consuming to write. Developers at Yahoo soon came out with a work around – Pig. Pig is essentially an easier way to write MapReduce queries. It is similar to Python and allows for shorter and more efficient code to be written that can then be translated to MapReduce before execution.

Hive

While this solved the problem for a number of people, there were many who still found this difficult to learn. SQL is a language that most developers are familiar with and hence people at Facebook decided to create Hive – an alternative to Pig. Hive enables code to be written in Hive query language or HQL that, as the name suggests, is very similar to SQL. Thus, we now have an option – if we are familiar with Python, we can pick up Pig to write code. If we have knowledge of SQL, we can go for Hive. In either case, we get away from the time consuming job of writing MapReduce queries.

So far we have understood 4 of the most popular Big Data technologies – MapReduce, Hadoop, Pig and Hive.

Let us now get introduced to database technologies popularly used in Big Data.

We first need to understand the concept of NoSQL databases.

NoSQL

NoSQL refers to databases that do not follow the traditional tabular structure. This means that the data is not organized in the traditional rows and columns structure. An example of such data is the text from social media sites which can be analyzed to reveal trends and preferences. Another example is video data or sensor data.

As such data sources become more and more important, so will the importance and popularity of databases that can handle such data i.e. NoSQL databases.

There are a number of NoSQL database technologies that work well for specific data problems. Hbase, CouchDB, MongoDB and Cassandra are some examples of NoSQL databases.

Database technologies enable efficient storage and processing of data. however, in order to analyze this data, Big Data specialists require other specializd technologies.

Mahout

This is where technologies like Mahout come in. Mahout is a collection of algorithms that enable machine learning to be performed on hadoop databases. If you were looking to perform clustering, classification or collaborative filtering on your data, Mahout will help you do that.

E-commerce companies and retailers have a frequent need to perform tasks like clustering and collaborative filtering on their data and Mahout is a great choice for this.

Impala is another technology that enables analytics on Big Data. Impala is a query engine that allows Big Data specialists to perform analytics on data stored on hadoop via SQL or other BI tools. Impala as been developed and promoted by cloudera.

This was an overview of the popular Big Data technologies.

Next we will look at the role of a Big Data specialist and the skills required to become one.

7.4 Big Data Specialists

The most crucial ingredient in Big Data

One of the great things about Big Data is that almost every factor of Big Data is either very cheap or completely free. Most of the software is open source, the hardware has been commoditized by the likes of Amazon and is available at dirt cheap rates. And the data is usually already there in the organization or is easy to collect at no significant cost.

The one thing in Big Data that is hard to get hold of, is the people. Big Data specialists are in a lot of demand these days. And there aren't too many of them. There is a huge gap between the ever increasing demand and the lagging supply. Thousands of Big Data positions are going unfilled. So much so, that many companies are unable to even start their Big Data initiative because they do not have the people with Big Data skills.

So what does it take to become a Big Data specialist?

At first glance, the number of skills required to become a Big Data specialist can seem overwhelming. It makes you realize this field is not for everyone. However, the good news is that even if you pick up some of these skills, you will be rewarded handsomely.

Let us start with some innate skills required in Big Data. These are skills that an individual looking to enter this field should already have.

Quantitative aptitude – You don't need to be a Math genius to become a Big Data specialist. But you do need to be comfortable with numbers.

Logical thinking – This is the most important ability required in the field of Big Data. Most Big Data problems require logical thinking ability. One can argue that logical ability is required for almost any kind of work. However, the reason to include this here is to emphasize the importance of logic in analytics.

Good communication skills – Data scientists and Big Data specialists often play the role of influencers. They need to advise senior executives on important decisions. If there are intermediaries between the business and a data scientist, the senior executives are likely to lose some crucial information as it goes from the data scientist to the intermediary to the business.

Curiosity, impatience and action orientation - All of these have been clubbed together because they are complementary skills. People who are curious and impatient are often action oriented as well. This is an important trait for Big Data specialists who are often performing completely new and ground breaking tasks or are mastering new tools and technologies.

These are some of the important abilities a Big Data specialist should have. Now let us take a look at the technical skills needed to be a Big Data specialist.

Understanding of the MapReduce framework and the hadoop setup – The MapReduce framework is based on the 'Divide and rule' approach that most Big Data technologies are based on. The hadoop setup is also based on MapReduce and is currently the most popular setup for Big Data. An understanding of mapreduce and hadoop therefore is important to enter the field of Big Data.

Knowledge of Big Data language such as Pig or Hive – Languages such as Pig and hive have been created to manage and process large data sets. Commands written in these languages are translated to MapReduce before processing. The advantage these languages have is that they are easier to elarn and write than MapReduce. Therefore, rather than learning MapReduce itself, it is better to learn either Pig or Hive. Pig is easier to pick up for people who know Python. While hive is similar to SQL.

Knowledge of Big Data analytics technologies such as Impala and Mahout – Pig and Hive are used to manage large data sets. Big Data specialists need other technologies like Impala and Mahout to actually perofrm analytics on data sets.

Familiarity with NoSQL databases such as Hbase, Cassandra and Couch – Big Data often comes as unstructured data (discussed earlier) and therefore requires non-traditional databases that do not follow the table structure. These databases are called NoSQL databases and they are very commonly used in Big Data. Familiarity with NoSQL databases such as HBase, Counch and Mongo is thus important for a Big Data specialist.

In addition to these, the following analytical skills are essential as well –

Knowledge of statistical concepts and their application in analytics

- Statistical concepts usually form the backbone of most analytic techniques and therefore an understanding of how these concepts are applied in business situations is very important.

Analytics methodology – This is a step-by-step approach to performing any kind of analysis.

Predictive modeling techniques - Commonly used predictive modeling techniques such as regression, clustering and association rules are also an important part of a Big Data specialist's arsenal.

Command over analytics tools – There are specialized tools which help a data scientist in analyzing large amounts of data. A command over tools such as R, SAS or SPSS is important for a data scientist.

Conclusion

We hope you now have a broad understanding of analytics and the potential it has to transform business processes and impact profitability and productivity. However, as much as we are able to predict and anticipate the future of data analytics from where we stand today, we know that there will be more advanced tools and technologies that some very smart people are going to develop in the coming decade, that we cannot even imagine. Prepare to be awed, as newer, faster analytic tools and technologies come into the market, giving better insights, faster and using larger amounts of data.

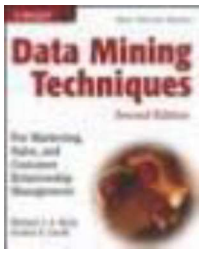
The future, many say belongs to those who embrace data. You have taken that first step… We now hope you get started at gaining the skills you need to join the data revolution. All the best!

APPENDIX

APPENDIX 1 - Resources on Data Analytics

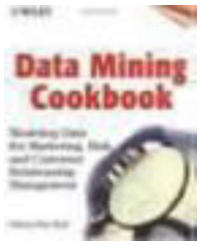
Books on Analytics

We have divided our list of favourite books into two sections.
The first section covers books on the fundamentals of analytics and business statistics.
We have also included books showcasing applications of analytics in various industries.



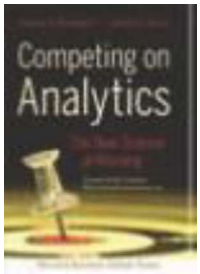
Data Mining Techniques by Michael Berry and Gordon Linoff

This is an excellent book on the most widely used analytic techniques. It starts off with defining data mining in the current business context and then summarizes some of the best practices in data mining. The book talks about some useful statistical concepts like p-value and chi-square as it takes the readers through the process of building a model. It explains analytic algorithms like Decision trees, market basket analysis, clustering, link analysis, clustering and survival analysis. The book is full of useful industry examples. This is the first book recommended for anyone with an interest in analytics



Data Mining Cookbook by Olivia Parr Rud

This book provides a detailed undestanding of the analytics methodology. It lists out several best analytic practices. The author primarily uses logistic regression as her technique and SAS and excels as the tools. The book has a very brief description of analytics so make sure you have some understanding of analytics beforeyou get to this book.



Competing on Analytics by Thomas Davenport

This book provides a detailed undestanding of the analytics methodology. It lists out several best analytic practices. The author primarily uses logistic regression as her technique and SAS and excels as the tools. The book has a very brief description of analytics so make sure you have some understanding of analytics before you get to this book.

Statistics for Management by Richard Levin and David Rubin

This is a great guide for statistics used in analytics. Starting from simple central tendency measures to probability distributions to decision theory, this book covers the essentials of business statistics. It is used as course book in most management institutes in India. This book is recommended for anyone interested in Statistics

Statistics for Management by Richard Levin and David Rubin

Freakonomics is an easy, interesting read, even for people who do not understand Statistics. The author uses the power of analytics to turn conventional wisdom on its head. With hypotheses like 'Legalization of abortions has led to crime reduction' the book throws up some very interesting questions for its readers. A must-read for everyone.

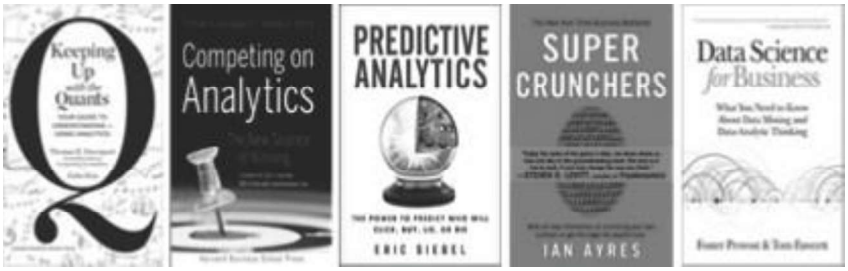
Predictive Data Mining by Sholom Weiss

This book has an easy style that will appeal to beginners. Some of the more complex topics may not be adequately addressed but a good book as an introduction.

Moneyball by Michael Lewis

This book studies the use of analytics in the sports industry. Itcase studies the Oakland Athletics, a US baseball team, with a payroll budget of less than a 3rd of some their rivals. Despite being small they have consistently been one of the best teams. The author shows how they leveraged analytics to get their advantage. OA analyzed metrics that were different from the ones traditionally looked at, but which they thought were more relevant to winning. A good example of analytics being used innovatively.

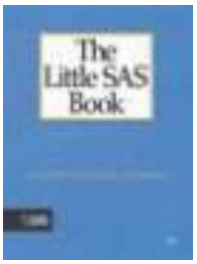
Other excellent books:



- "Keeping Up with the Quants: Your Guide to Understanding and Using Analytics" by Thomas H. Davenport and Jinho Kim
- "Competing on Analytics: The New Science of Winning" by Thomas H. Davenport and Jeanne G. Harris
- "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die" by Eric Siegel
- "Super Crunchers: Why Thinking-By-Numbers is the New Way To Be Smart" by Ian Ayres
- "Data Science for Business: What you need to know about data mining and data-analytic thinking" by Foster Provost and Tom Fawcett

Books on Analytics Tools

The second section consists of books around specific analytic tools or software.
These books cover tools like SAS, SPSS, R, SQL and excel.



Little SAS Book by Lora Delwiche

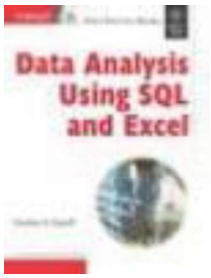
This is a good book to learn SAS. It is readable as it is composed of two-page articles. Each one focuses on a specific task or function of SAS. The book is divided in 10 chapters that go through reading data sets, building reports, combining data sets, writing macros, using graphics, debugging SAS programs, etc. The book is a good reference for simple tasks. Any simple task you don't know? Just look in the index and you will find the corresponding function. However, for more advanced topics, the book is a bit light. Well, that's what the title also says.

SAS Programming by Example by Ron Cody

It is one of the best books available for beginners in SAS. The book is simple and easy to read with many industry examples for better understanding. The book deals with the essentials of analytics and reporting using SAS.

Data Manipulation in R by Phil Spector

This thin book provides a solid introduction to many of the functions and packages for importing, manipulating and processing data in R. Using a variety of examples based on data sets included with R, along with easily simulated data sets, the book is recommended to anyone who wants to learn R.



Data Analysis using SQL and Excel by Gordon Linoff

A good book on understanding how tools like SQL and Excel can be leveraged to extract useful business information from relational databases. The book is organized around chapters that become increasingly complex in the use of SQL, Excel, and data mining concepts.

Analytics Blogs

These blogs can give you a range of useful perspectives and insights on statistics, data analytics, latest trends and innovations in data science as well as some interesting applications of analytics. These are all written by people who are in the industry and who are willing to share their experiences. There will be much you can learn from them.

1) Data Mining Blog http://www.dataminingblog.com/ - This blog by Sandro Saitta, a Swiss national, covers research issues, recent applications, important events, interviews with leading actors, current trends and book reviews in the field of analytics. The blog has some interesting book reviews as well as their own comprehensive list of analytics blogs.

2) Data Miners' Blog http://blog.data-miners.com/ - This blog is written by Michael Berry and Gordon Linoff, two leading figures in the field of analytics. They have written some of the most popular books in this field and their books are used extensively in data mining courses by Universities across the globe. Their blog is filled with excellent tips and knowledge nuggets taken from the writers' vast experience in analytics consulting and is a must-read for any business analyst.

3) Flowing Data - http://flowingdata.com/ - Through flowingdata, Nathan Yau, a PhD candidate in UCLA wants to make data useful for even those who are not very data savvy through effective use of visual techniques. It is basically a visualization and statistics site that shows different applications of data analysis.

4) Abbott Analytics - http://abbottanalytics.blogspot.com/ - The Abbot analytics blog has several Tips, tricks, and comments in data mining and predictive analytics, including data preprocessing, visualization, modeling, and model deployment. The blog is hosted by Dean Abbott, president of Abbott Analytics in California, USA. Mr. Abbott has over 21 years of experience applying advanced data mining, data preparation, and data visualization methods in real-world data intensive problems, including fraud detection, response modeling and survey analysis.

5) Occam's Razor-http://www.kaushik.net/avinash/- Occam's Razor is an anaytics blog written by Author, Digital Marketing Evangelist, and Google Co-founder, Avinash Kaushik. It includes useful posts and articles offering instruction and hands-on tips for practically implementing various analytics tools and applications, ways to extract the data you need and also the best ways to present and contextualize it.

6) Webtrends Blog-http://blogs.webtrends.com/category/analytics/- The Webtrends blog focuses on providing open access to data. It is a meeting place for analytics professionals, and a forum for information exchange. The blog is authored by executives and other employees of Webtrends.

7) Stats with Cats-http://statswithcats.wordpress.com/- The Stats with Cats blog is for those who can't solve life's problems with statistics alone. The blog is written by Charlie Kufs who has been crunching numbers for over thirty years, first as a hydrogeologist and since the 1990s, as a statistician. He is certified as a Six Sigma Green Belt by the American Society for Quality and is a Certified Professional Geologist in Pennsylvania. Charlie currently works as a statistician for the Federal Government.

8) BzST Business. Statistics. Technology-http://www.bzst.com/ The BzST blog is written by Galit Shmueli, Professor of Statistics at Indian School of Business, Hyderabad, India. It provides interesting insights on data analytics.

9) Jigsaw Academy- www.analyticstraining.com. Jigsaw Academy's very own blog that gives you the latest in analytics news and insights.

Other valuable SAS resources

- Books by SAS Press – http://support.sas.com/publishing/index.html The SAS Institute have a separate publishing division that focuses on the analytics ecosystem by generating and publishing a steady stream of books and literature on the SAS language. The SAS Bookstore is a great place to look for books at the beginner, intermediate and expert level at

- The SAS-L Email Group - the SAS-L email group is a supportive and helpful resource for programmers at any level when stuck in a particular program. The archives are available at http://listserv.uga.edu/archives/sas-l.html. People posting on the list are expected to paste a sample of the dataset structure, and a clear explanation of what they are trying to achieve.

- The SAS Community Website- SAS Users have an online wiki for collaboratively adding in code samples, programs and sharing tips. The site is available at www.sascommunity.org/wiki/Main_Page and it is quite useful in terms of daily tips to improve SAS language skills, blogs and papers.

- Blogs at SAS.com – This is a collection of blogs that deal with business analytics, and technical application in various domains and SAS related happenings. With almost 29 blogs, http://blogs.sas.com is one of the better locations to build your perspective in business analytics and not just the technical aspects of writing code. SAS blogs are exclusively maintained by the SAS Institute team of communication people and feature experts from across multiple businesses.

- SAS Online Document - SAS online document is the documentation available for everyone by web access. It is like a big book or HTML library and you can view the lat¬est version at http://support.sas.com It is very useful for fine-tuning your SAS language code and discovering extended functionality to your software as well as troubleshooting or debugging programs.

- Papers from SAS Global Users Conference Proceedings (SUGI) – This is an archive of papers presented at the annual conference for SAS Users. The conferences have been taking place since 1976, and the papers represent the best innovative uses in the language. They can be accessed online at http://support.sas.com/events/sasglobalforum/previous/online.html

Other Online Resources

The world wide web is really a great platform to enhance your analytics knowledge and gain some valuable insights. Here are some other popular free resources you can access online that will help keep you a well informed data analyst.

- Thearling's blog - http://www.thearling.com - Kurt Thearling's website is an excellent place to learn about data mining. Kurt is a thought leader and veteran in the analytics space with extensive experience in some of the leading analytic companies like Capital One, Dun & Bradstreet and Vertex. He has also authored several books on analytics and written a series of articles/papers that cover topics ranging from the basics of data mininging to advanced domain-specific techniques.

- Kdnuggets – www.kdnuggets.com – Established in 1997, this website is a comprehensive resource for anything related to data mining and analytics. Visit this site to learn about news, events and jobs in analytics. They also have links to datasets and training material on analytics.

- Analyticbridge – www.analyticbridge.com – It describes itself as the social network for analytics professionals. Use this site to connect and network within the analytics industry and stay updated on the latest news and events in the field.

- Analytics India Magazine-http://analyticsindiamag.com/Analytics India Magazine (AIM) serves the needs of Analytics professions and provides a focal point for all information related to analytics from the Indian context. It is a web-based magazine to promote, nourish and collect; ideas and thoughts on Analytics practice in India.

- Kaggle- http://www.kaggle.com/ Kaggle is the world's largest community of data scientists. It is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

- Khan Academy- https://www.khanacademy.org/. This is a free education portal for everyone. It has some valuable information and lessons on statistics, regression andother analytics insights.

- LinkedIn groups- LinkedIn has thousands of groups that are like virtual communities where people of similar interests can share knowledge. Some of the popular analytics related groups are :

    - Advanced analytics
    - Business analytics
    - Customer analytics group
    - Data Science Central
    - Global analytics network
    - India analytics network
    - Kdnuggets analytics and data mining
    - SAS analytics and BI
    - SAS and analytics users
    - Your analytics career

- YouTube. This social channel has a collection of insightful must watch videos on Analytics. Some videos you will find very interesting are:

    - How it works: Analytics: http://youtu.be/_HbjsNaUJ2A
    - A brief history of intelligence: http://youtu.be/yVIcIRcAhxc
    - What can Business Analytics Do for You? http://youtu.be/uP89kaDU40c
    - REvolutionAnalytics: This channel has many interesting videos on Big Data analytics using the open-source software R
    - SASsoftware: SAS's channel talks primarily about SAS but also has some good videos on analytics.
    - Googleanalytics: The official channel for all videos about and related to Google Analytics.
    - IBMbusinessanalytics: IBM's youtube channel focusing on analytics and its business application.
    - Jigsawacademy: Our own channel provides informative videos on analytics, interviews with industry leaders as well as snippets of our lectures.http://www.youtube.com/jigsawacademy

Books on Big Data

- "Big Data: A Revolution That Will Transform How We Live, Work, and Think" by Viktor Mayer-Schonberger and Kenneth Niel Cukier
- "Big Data at work: Dispelling the myths, uncovering the opportunities" by Thomas H. Davenport
- "Taming The Big Data Tidal Wave" by Bill Franks

Books on Hadoop

- "Hadoop: The Definitive Guide" 3rd Edition by Tom White
- "Hadoop in Practice" by Alex Holmes
- "Hadoop in Action" by Chuck Lam

Popular Blogs on Big Data and Hadoop

- Smarter Computing Blog http://www.smartercomputingblog.com/category/big-data/ - Maintained by IBM which includes articles around Big Data and cloud computing
- Planet Big Data http://planetbigdata.com/ - An aggregator of worldwide blogs about Big Data, Hadoop, and related topics.
- Big Data | Forrester Blogs http://blogs.forrester.com/category/big_data - An aggregation of blogs and articles from enterprise experts focusing on Big Data topics
- Hadoop Wizard http://www.hadoopwizard.com/ - A website dedicated to help people learn how to use Hadoop for "Big Data" analytics
- Yahoo! Hadoop Tutorial https://developer.yahoo.com/hadoop/tutorial/index.html - It includes free materials that cover in detail on how to use the Hadoop distributed data processing environment
- Hadoop 360 http://www.hadoop360.com/ - Exclusive Hadoop site maintained by data science central community
- Cloudera Developer Blog http://blog.cloudera.com/blog/ - Big Data best practices, how-to's, and internals from Cloudera Engineering and the community
- The Hortonworks Blog http://hortonworks.com/blog/ - Collation of articles around Hadoop related to latest releases, trends and updates from the expert team of Hortonworks

Video Resources for Hadoop

- Big Data University http://bigdatauniversity.com/ - An IBM initiative which offers free online courses taught by the leading experts in the field
- MapR Academy http://www.mapr.com/services/mapr-academy/training-videos - It provides few free training resources to help individuals and teams learn and use Hadoop
- Hadoop Screencasts http://www.hadoopscreencasts.com/ - A collection of good quality screencasts on installation and working with Apache Hadoop and the various components of the Apache Hadoop Ecosystem
- Hadoop Essentials http://www.cloudera.com/content/cloudera/en/resources/library/training/cloudera-essentials-for-apache-hadoop-the-motivation-for-hadoop.html - A six-part recorded webinar series offered for free by Cloudera about introduction and motivation for Hadoop
- Hortonworks Sandbox http://hortonworks.com/products/hortonworks-sandbox/#install - It is provided as a self-contained virtual machine by Hortonworks with hands on video tutorials and pre-installation of single node Hadoop cluster