VOLUME 11 | ISSUE 04

digit **Fast Track**

YOUR HANDY GUIDE TO EVERYDAY TECHNOLOGY

To BIG DATA

- WHAT'S BIG DATA?
- SCOPE OF BIG DATA
- THE NEED TO REINVENT
- EXPLORING A BIG DATA SYSTEM
- APPLICATIONS:
  - MANUFACTURING / GOVERNANCE
  - EDUCATION / MEDIA
  - SPORTS / HEALTHCARE
  - BANKING
  - SCIENCE
- HOW DOES NETFLIX USE BIG DATA
- CAREERS IN BIG DATA

A 9.9 Media Publication

Hello Helpdesk

FAST TRACK to

BIG DATA

powered by

digit

YOUR TECHNOLOGY NAVIGATOR

CHAPTERS

BIG DATA
APRIL 2016

CREDITS
The people behind this book

EDITORIAL
Executive Editor
Robert Sovereign-Smith

Managing Editor
Siddharth Parwatay

Technical Editor
Jayesh Shinde

Senior Reviewer
Mithun Mohandas

Copywriting
Arnab Mukherjee

Writers
Anand N.
Ashwin Krishnan
Dhinoj Dings
Ketsan C.
Kshitij Bedekar
Vishal More

DESIGN
Creative Director
Tharakaram G

Sr. Art Director
Anil VK

Associate Art Director
Anil T

Sr. Visualiser
Shigil Narayanan

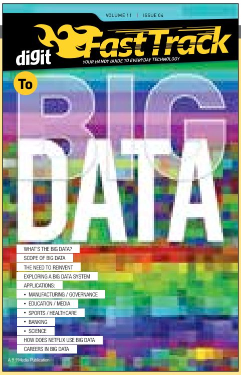
Visualiser
Baiju NV

Sr. Designer
Pradeep G Nair

digit.in

April 2016
Free with Digit. If you have paid to buy this Fast Track from any source other than 9.9 Mediaworx Pvt. Ltd., please write to editor@digit.in with details

Custom publishing
If you want us to create a customised Fast Track for you in order to demystify technology for your community, employees or students contact e-mail editor@digit.in



COVER DESIGN: ANIL T

The Big Deal

If each byte of data that we generate every day were to be a grain of sand, we could double the quantity of sand on earth in a mere 3 days. That is the extent to which the present day world of technology is driven by data. And most of that data, even up to 90% according to some studies, has been generated in the last couple of years. That comes as no surprise thanks to the explosive increase in the sale of smartphones, wearables and other smart devices. The world didn't take time to realise the potential of all this data and hence, the field of Big Data was born.

Big Data has been the buzzword of the information technology industry for some time now, yet there is little or no awareness regarding the actual internals and implications of this technology. Through this issue of Fastrack, we have attempted a create a basic guide to Big Data, after reading which you shall be ready to roll your sleeves up to pursue the field if you chose to do so.

Chapter 1 and 2 take you through the basics of Big data and clear out the jargon that the technology is accompanied by. The different types of big data and their sources are also explained. Most of us are aware that Relational Databases were already doing the job of dealing with data for enterprises. Chapter 3 highlights why exactly that failed with Big Data and how Big Data technologies evolved from the failure of Relational Databases.

Hadoop has been synonymous with Big Data for most of us, so much so that many of us can't differentiate between the two. Chapter 4 takes you from the emergence of Hadoop to it's future potential, while explaining what makes it the successful piece of Big Data technology that it is.

Chapters 5 through 9 highlight and showcase the various fields in which Big Data has been successfully used to achieve monetary gains as well as greater impact. Fields as diverse as education and banking have benefited by implementing Big Data technology into their systems. This Chapter also shows you the possible benefits that are yet to be reaped in all these fields.

Netflix is one of the most successful examples of a successful implementation of Big Data technology combined with computational algorithms to achieve customer satisfaction that would not have been possible otherwise. Chapter 10 takes you through this case study and explains how the entries from the Netflix prize competition played a pivotal role in creating the content suggestion algorithms that Netflix uses.

Finally, all of this information would not be of much use if we did not tell you how to enable yourself to be a Big Data professional. That is exactly what Chapter 11 is all about, along with highlighting the various resources that you will need on the way. So by the time you turn the final page, probably you will already be hunting for the next big job in Big Data. Best of Luck!

WHAT IS BIG DATA?

Big Data is the new buzz word in the industry and is cited as the technology that will change the way we do things. Large corporations and even governments are investing heavily in this new technology. And it is not just these corporations and governments that will reap the benefits of this technology, but we will as well on an individual level.

As the term implies, Big Data has a lot to do with large amounts of data and its collection, processing, analysing and generating actionable intelligence – but this would be oversimplifying things.

A simple example to get you going

Instead of diving straight into the technical definition first and trying to explain the nuances of Big Data, it would be easier to understand the whole thing once you have a simple analogy – your brain.

Big Data is the next level technology in the field of large scale Data

Think of it, you brain gathers a vast amount of data from different sources, processes it to make sense out of it and then generates intelligence from it on the basis of which you can take action.

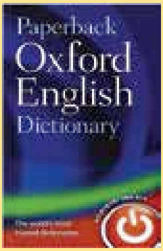
For example, you want to buy a new violin but are limited by budget. Your brain starts gathering fresh data and refers previously collected data. You remember contacts who have violins and contact them for advice on what type of of violin to buy. You remember someone talking about an upcoming discount season, an email last week from your bank about extremely low interest rate on personal loans, an unused shopping voucher gifted by friends for an online website. You process this data from different sources, analyse them as data about a violin and your brain uses this to generate intelligence about what type of violin to buy, from where to buy it and from what source and by using those resources which will be the most optimal for your budget. You get advice from friends, wait for the upcoming discount season, select the website on which the gift voucher is applicable to reduce the price, and use your bank's low interest loan to facilitate your purchase. What you have achieved in the end is an optimal buying situation of your violin, all because your brain processed information from different sources to give you the optimal condition where you ended up saving a lot of money.

The above explanation was just to give you a basic understanding about Big Data, it is much more vast with what it can achieve.

Getting into the details

While there may not be a consensus on the exact definition of Big Data (which is a technology that is constantly evolving), there is a common theme to it which basically states that Big Data is a collection of large sets of data – structured or unstructured –which can be stored, analysed, manipulated to reveal or discover patterns and trends.

OXFORD ENGLISH DICTIONARY INCLUSION



Big Data was incorporated in the Oxford English Dictionary in 2013 under Computing subcategory of Technology category as - data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.

Data formats

- **Structured:** This is the type of data which is based on a pre-defined model or structure. It is easier to manage structured data. Example of structured data are relational databases, CRM system data, XML files, etc.
- **Unstructured:** This type of data doesn't really follow a pre-defined model or structure. An example of this type of data can be word files, spreadsheet, video files, audio files, images, etc.

Characteristics of Big Data (3Vs (+ 2))

To understand what comes under the purview of Big Data we can characterise it by three shared characteristics, namely volume, velocity and variety, or more popularly known as the 3Vs.

Volume

One of the main characteristic of Big Data is that it is massive in size. From a few terabytes, it can go up to petabytes. The size of the data determines the insights that can be gained from it and the potential of the data.

Velocity

The speed at which the data is generated is also an important characteristic of Big Data. In today's world, massive amounts of data is generated real-time and hence needs to be worked upon real-time. Smart devices like sensors, meters, etc., generate a massive stream of information and hence need to be processed immediately.

Variety

Another important defining characteristic of a Big Data data platform is it's ability to process data from a variety of sources and formats. This data can be structured or unstructured. An examples of structured data could be the numeric data stored in traditional databases while unstructured data could be anything from emails and videos to stock tickers and market transactions.

Actually, two more dimensions can be taken into consideration when talking about Big Data characteristics.

BIG DATA
VOLUME
SCALE OF DATA
VARIETY
FORMS OF DATA
VELOCITY
ANALYSIS OF DATA FLOW
VERACITY
UNCERTAINTY OF DATA

Along with the 3Vs, Veracity is also an important characteristic of Big Data

Variability

In a Big Data system, the flow of data is not always consistent; there could be peak periods and lean periods. For example, the flow of data from social media will change during different days of the week, time of the day and different events.

Veracity

There is also an unavoidable uncertainty of the data's quality in a Big Data system. The variation of the quality of data can also affect analysis.

Big Data Analytics (the 6Cs)

Simply collecting data and processing it will not help much, unless you can generate something useful from the processed data. The 6Cs of Big Data Analytics systems focus on **Connections, Cloud computing, Cyber**

models, Content and context, Community sharing & collaboration, and Customization.

What does all this mean? Let's just take a look at an example then. Have you noticed that when you shop on some large e-commerce shopping website, you are given great suggestions the next time on what to buy; this is Big Data analytics in action. It just didn't track your shopping data and recommend the same products to you, it determined what you looking for as an end product and suggested options based similar to those. For example, if you are buying computer components online, a larger ecommerce website might suggest cabinets or some other product.

Big Data sources

There are several different sources of Big Data but they can be broadly classified under the following.

Enterprise Data

Enterprises generate a massive amount of data in formats like spreadsheets, office documents, pdfs, etc.This composition is referred to as Enterprise data.

Transactional Data

Software applications – whether it is web, mobile, CRM, etc. – used by an enterprise execute endless transactions throughout their existence. These transactions are recorded in a backend database. This type of data is called transactional data.

Almost every field that harnesses technology is a source for Big Data today

Social Media

This one doesn't require much explanation; social media platforms like Facebook, Twitter, Pinterest, etc., generate massive amounts of data every second, whether it is text, image, audio, video, etc. This type of data is usually unstructured,

Activity Generated

Machines generate a massive amount of data every second while going about their function (this is much more than the data generated by humans). Example of this kind of data can be, sensor data, satellite data, surveillance videos, Industrial machinery, medical devices, etc.

Public Data

Publically available data like weather data, published data by research institutes, government data, census data, Wikipedia, and other types of data which is freely available is classified as Public data.

Archives

This is historical data that is not required generally or data that is rarely used. It is archived by organisations, as against discarding it, for as long as possible due to the availability of cheap hardware. This kind of data can also include scanned documents like project documents, agreements, etc. These types of data sources that are less frequently accessed are called as Archive data.

THE ULTIMATE USER ACTIVITY

Big Data is all about user activity. As per statistics, people perform 40000 search queries every second on Google which accounts to an overall 1.2 trillion searches per year.

What is the difference between Big Data Analytics and Business Intelligence?

IT Consultant Eric D. Brown has summed up the answer to this question in a simple but accurate statement, as follows.

“Business Intelligence helps find answers to questions you know. Big Data helps you find the questions you don’t know you want to ask.”

At the most basic level it means that in Business Intelligence, you will get well defined reports and answers to the questions that you have asked, while in Big Data Analytics, all the data that undergoes analysis reveals new patterns and trends which you did not think of.

For example, using Business Intelligence tools on medical data you can find research which can help pharma companies determine the success of their products with very accurate details, or what modifications are required. Basically, you get the answers to what you ask. While, on the other hand, if you feed a vast amount of medical data to a Big Data system it could reveal

Big Data offers much more in depth capabilities than Business Intelligence

a pattern about medicinal conflicts, i.e:- you can find out a pattern where a combination of different medicines create side effects in a patient which had not been previously thought of before.

Mind boggling facts about Big Data

- Most of the data ever generated in human history was generated in the last few years.
- The rate of data generation is multiplying as we speak. By 2020, 1.7MB of data will be generated per person/per second.
- By 2020, out accumulated data will reach 44 trillion gigabytes.
- We are creating new data every second. On a top search engine alone we are generating around 1.2 trillion searches per year, which is again set to rise exponentially.
- A staggering number of one billion people used Facebook on a single day in August 2015.
- YouTube sees hundreds of hours of video upload every minute.
- More than a billion smartphones were shipped in 2015, all with sensors capable of generating a massive amount of data.
- In the next few years, there will be around 50 billion smart devices which will be connected and they will be generating and analysing a lot of data.

- Retailers can increase their operating margins by 60 per cent if they leverage the power of Big Data, while Fortune 1000 companies can get an additional 65 million and above income if they increase their data accessibility by 10 per cent.
- At the moment, of all the data being generated, only a minor fraction of it is analysed. Which means that there is a massive potential for what can be achieved when we achieve a state where most of the data is analysed.

The above are just a few facts when compared the enormous impact that Big Data can have on the whole for the human civilisation.

Issues of a different kind

Big Data from a technological perspective is a clear winner. However, when it comes to human civilisation, there are several other aspects to consider:

- **Legal:** The issues that could arise here would be related to intellectual property rights, privacy risk, data protection, licensing, contract, etc. The legal framework will have to catch up with technology if the power of Big Data is to be used to its full potential.
- **Social and ethical:** Unintended use of secondary data, re-use, transparency, profiling, tracking, can raise ethical issues as Big Data deals with human data on a large volume which could be an issue with social and moral codes.

The list of issues will keep on rising as this is a disruptive technology which will bring about a vast change and hence there is bound to be friction somewhere.

A few examples of Big Data usage

Improving business targeting: The most obvious place that Big Data is being used is in targeting the right customers by understanding their behaviour and preferences.

- **Making business process more efficient and optimised:** Companies that use Big Data are better able to optimise their product and its stock by analysing trends revealed by Big Data.
- **Quantified Self:** Quantifying personal data through sensors and trackers eventually leads to the formation of a pattern which can be revealed through Big Data. This can be used to improve lifestyle and/or performance.
- **Medical field:** Big Data systems can analyse DNA in a matter of minutes. When Big Data systems go through vast amounts of medical data, they

Today's connected world is harnessing Big Data at a rapidly growing rate

find patterns. These patterns could someday lead us to find cures to some of the most deadly diseases.

- **Science and Research:** Massive amounts of scientific data that is being generated every second can be used to reveal new scientific secrets. This could lead to a faster evolution of human technology.
- **National Security and Law:** It is no secret that sometimes anti-social elements manage to slip through security and wreak havoc on society. With Big Data implementation, the accuracy of determining terrorist strikes and law and order problems could be vastly improved. Thereby, this can lead to precision efficient and precision targeting of anti-social elements and events, ensuring better security.
- **Better Artificial Intelligence:** An AI system needs a vast amount of data to make sense of its environment. A "Machine" with Big Data analytics could decode vast volumes of information in different formats and in real-time for better processing and judgement – just like our brain (sounds familiar?).

Big Data is no doubt a disruptive force in the technological environment, and it has the potential to do what the wheel did for the human civilisation.

SCOPE OF BIG DATA

The only introduction to big-data-related jargon you need – we help you make sense of the kinds of big data and each of their relevant domains.

Let There Be Numbers

The oldest known system of counting is the tally mark system that came into being about 20,000 years ago. Back then; the average human beings were remarkably different, still coming to terms with all the changes that evolution had brought upon them. However, despite the plethora of modifications to our bodies and minds, there has been one thing that has stayed with us even today – the urge to collect data. The art of counting and collecting has evolved from being a way of keeping track of your cattle, to being the primary source of information for areas of study ranging from marketing to astrophysics.

In 2011, a paper published in the Science Express journal that the total amount of information that we stored from a period of 1986 – 2011 was close to 295 exabytes, with a 23% increase every year from 1986 to 2007. It is estimated that just Google, Amazon, Microsoft, and Facebook are in possession of a whopping 1,200 petabytes amongst the four of them. But

We generate incomprehensibly large amounts of data every day with every action we take – and it's only going to keep growing.

even this pales in comparison to the amount of scientific data being gathered on a daily basis. In 2013, scientists at CERN announced that in the course of 3 years, they had collected roughly 75 petabytes of data courtesy of the Large Hadron Collider.

All this goes to show that the mind-boggling amount of digital information we have available to us is increasing exponentially, and is going to keep doing so for the foreseeable future. This gigantic amount of data brings with it keys to a previously locked universe of patterns and trends that can be analyzed and re-analyzed to study subjects as unrelated as business trends, crime rates, rate of epidemic and so on and so forth. Utilized correctly, it could transform the very nature of statistical study. Any change so big, of course, leads to problems of its own. That is, how does one make sense of all this information?

In the mid-1960's, the Green Revolution changed the entire way of life for farmers all over India. The production of grain went through the roof, and it was estimated that most, if not all the starvation in the country would be wiped out. In the four decades that followed, it proved not to be the case. As of now, India is in the absurd position of having the largest population of hungry people in the world, all the while having enough food to feed them all. The problem? The storage and transportation of an unprecedented amount of grain. It has since been realized that while the actual production of grain was up, the facilities that were involved in the processing and distribution did not evolve at the same rate. This meant that this incredible leap in agronomy solved one problem, but only at the cost of creating several more.

In this same way, the recent surge in data could lead to a host of new problems. While the technological giants of the world do possess an adequate range of tools to collect, capture, process, and share all of it, they are still not sure how exactly to use all of it efficiently. Simply put, they have all the storage and transportation problems figured out, they just do not have the actual Green Revolution.

VIEW THE PULSE OF SOCIAL MEDIA

Nearly 32 million messages, 3 million videos and 50 billion photos are being shared every minute over Facebook with one billion people using it in a single day in August 2015. Over 350 million people use Twitter with 150 million unique monthly visitors via both mobile and PC.

Big data – a misnomer

When the term 'Big Data' was first used, it was a moniker given to every sort of data, ranging from food to health, to sales, to data that the government possesses, acquired through decades of paperwork involving passports and driving licenses. Over the years, it has also evolved to encompass all the schemes that proposed, and eventually used Big Data as their source of information.

Imagine if you will that Big Data as the digital equivalent of The Beatles. Now, The Beatles influenced musicians the world over, and consequently allowed them a different kind of freedom of expression. But to a person who was only vaguely aware of their existence, the best way to show them The Beatles' legacy would be to inform them of the most popular musicians to have been strongly motivated to do what they do because of them.

Just like the man who didn't have more than a vague idea of The Beatles, most people today don't know what Big Data is, and the only way to even partly understand this global phenomena and its staggering order of magnitude is to try to understand its most popular sub-parts. The most common terms bandied about by tech company executives on forums on the Internet are Smart Data, Identity Data, and People Data. These terms are neither all-inclusive, nor are they self-explanatory, but they seem to be the ones that are going to outlast the others that are thrown around. So, to understand the scope of Big Data and what/whom it affects and how, these are the terms that should be most carefully looked at.

Smart Data

Let's look at the most popular of the three terms to begin with – Smart Data. Smart Data is essentially the sub-set of Big Data that can immediately be put into use. As previously stated, Big Data comprises of the 3 V's. Typically, a Big Data problem consists primarily of Volume and Velocity, with Variety taking a back seat. However, in the real world, a large part of any information collected using an algorithm is just meta-data having no immediate practical purpose. Smart Data, on the other hand, is more purposeful in its problem solving skills, focusing on the third V (that is, Veracity and Value) to sort out valuable, actionable information.

The sorting of Big Data traditionally calls for professionals having a high level of knowledge regarding data collection and interpretation, and takes a lot of time and effort to pre-process, clean, and accurately segment before it becomes valuable to a particular company. The beauty of Smart Data is that it allows for both qualitative and quantitative sorting of data. While this would still require an expert to make the Smart Data platform, it doesn't need their service any more than that, as trained marketing and research employees could further use the data. For example, if you were to add baby diapers and soy milk to your online supermarket cart, they could immediately process this data and, in the future use it to target baby-product advertisement to your IP address.

A smaller section of Smart Data is Fast Data – instantaneous results from a live input stream. Basically, it is the ability of a machine to give instantaneous results on the basis of a continuous stream of information provided to it. To better understand it, you could ask yourself whether you'd rather have an inch-perfect traffic report an hour after you asked for it, or a slightly less accurate version almost immediately. Obviously, the answer is the latter – and this sums up Fast Data. It is extremely fast analysis of data in a way that, if asked to choose, picks speed over accuracy.

Identity Data

The postmodern understanding of the human identity, if paraphrased, states that every human being is the product of the intersection of all his or her societal interactions. While this has several philosophical implications, this essentially means that every step an individual takes and every decision he makes can be predicted – as long as there is a sufficient amount of information available to analyze. This is the essence of Identity Data. This particular sub-type is the driving force behind the ability of a machine to learn and predict the way a human will behave. Predictive modeling, as a whole, emerges from the analyses of Identity Data.

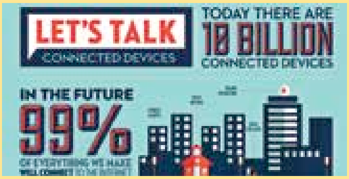
With big data, everything is quantifiable. Your shopping habits are your online fingerprints. We are defined by the data that we produce.

It is also the part of Big Data that is most important to security. When hackers stole identity information from AshleyMadison.com, or breached Target's database, the biggest problem both of them faced was the loss of identity data. This was because, to a supermarket like Target, this identity data, coupled with shopping patterns, credit card usage, and social media behavior, could very accurately predict your lifestyle choices, and enable them to advertise more efficiently to you. In the same way, AshleyMadison.com, a cheaters' dating website, could use all this information, plug it into their dating algorithm, and find a better match for their prospective consumers. Websites such as Wired and Slate also use identity data to predict what kind of articles their subscribers are likely to enjoy, and to try and improve on how they bring them the news.

People Data

People Data is the Internet equivalent of a call-center. This is because it is the type of data that is used by companies to treat their customers like they are real human beings. People Data is generally collected through the social interactions of a customer including what social media sites they prefer, where they got the recommendation to first stumble upon the website of said company, if they stayed on the website or closed it immediately, and on and on. Because of the subjective nature of such data, People Data differs from Big Data by having much smaller, less accurate, less stable, and increasingly shifting data sets, therefore entailing the most difficulty in analysis out of the three kinds of big data. However, it is essential to take People Data into account, especially in concert with its more reliable brethren, and that is evident from a cursory glance at the banking sector. While stable for the most part, this particular service has nevertheless been constantly plagued by scams and subterfuge. A major cause of such happenings could be that it is an industry set up entirely on the value of Smart Data, and not People Data.

MANAGE DATA WITH CONNECTED DEVICES



By the end of 2020, there will be 50 billion smart connected devices in the world to collect, analyze and share data for various activities. The Hadoop market arena is expected to grow at an annual growth rate of 58% surpassing $1 billion by 2020.

While it is apparent that the lines between the three subsets are blurry at best and indistinguishable at worst, it is important to remember that all three are simply refined versions of raw, large-scale data. With increasingly sophisticated techniques that help us make sense of incredible amounts of data, it is inevitable that the three will not only intersect, but may at times completely overlap as well. We are going to encounter these terms much more frequently as time passes, as they will be the cornerstones of most customer-based industries, scientific endeavors, and basic Internet-based human interaction.

Keeping this basic understanding of the terminology in mind, we are now ready to delve further down into the big data rabbit-hole.

REINVENTING THE MODEL: A NECESSITY RATHER THAN AN OPTION

Why relational databases are incapable of handling high data loads?

The relational model of data, based on which a relational database exists, was first proposed by the English Computer Scientist, E.F. Codd back in 1970 while he was working for a certain company called IBM.

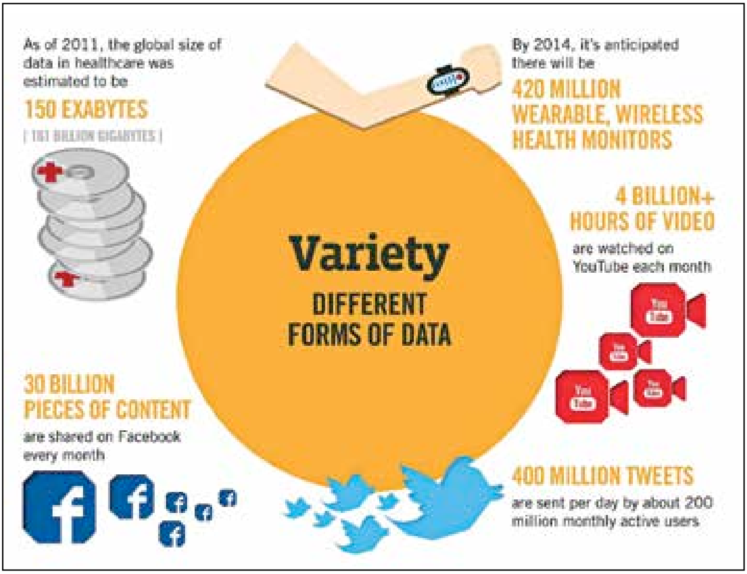
In Codd's model, data is arranged in one or multiple tables-otherwise called relations with rows and columns. For each row, there is a unique key to identify it. The general mode of data storage follows that while one relation represents an entity type-such as a company's customer, the corresponding column would have a value related to the entity type-perhaps the address of the customer.

Due to the relative simplicity and high efficiency of the model, relational databases are rather ubiquitous in many organizations. But the efficiency falters when it comes to handling high volumes of data.And in this day and age-when multiple systems and petabytes of data have to be analysed for many a business to function well, that is posing to be a big issue. There are multiple reasons why relational databases don't rise to the challenge of handling big data. And here we list the most important of them.

They are not adept at handling variety

Aside from the large volume of the data itself, one of the key parameters that define big data is the variety in terms of the data types. Many of the data could be so disparate in nature that finding a connect between them would be a major task in itself-requiring a lot of hours to be expended-both by humans and machines.

And as far as structured data goes-it only forms a small portion of data which organizations have to deal with. This means that relational databases will have to grapple with unstructured data for the most part when it comes to big data. Now, it isn't that relational databases cannot be configured to process data variety, but the changes can be brought about only with


As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Relational Databases are not built to handle the variety of data that is generated today strenuous efforts, and that too may not help derive the complete value from the data.

They are not designed for change

As mentioned before, in relational databases data is arranged in columns and rows-with each row having a unique entry and every column attributing unique value to the entry. This being the case, the data modeling has to be done in advance and sometimes this could mean months or even years, the duration dependent on the system.

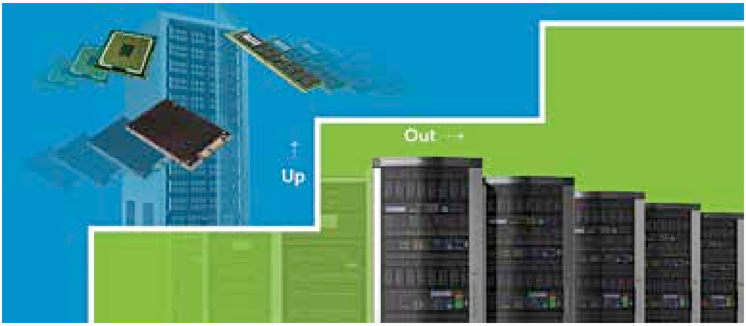
Once the modeling is done, to bring in significant changes in the structure is again a time and resource consuming endeavour. Big data, by default is consistently appended, or the values may keep changing for some of the parameters. This demands a high level of flexibility from the database. Something, the relational structure obviously isn't suited for.

THE JOY OF WINNING ELECTIONS WITH BIG DATA

India's current ruling party led by Prime Minister Narendra Modi had effectively used Big Data analytics in 2014 to fetch the taste and interests of people living in various parts of the country. This enabled them to identify potential issues from over 800 million voters and find remedial solutions.

For scalability,elasticity and resilience, relational databases are far from the best

Scaling up databases is an inevitability for many organizations if they are to keep functioning. This scenario spurred by an overwhelming rise in the volume of data has been noted by relational database vendors. And to address the problem they have devised such 'solutions' as in-memory processing, better use of replicas and distributed caching, even shared storage. In the right place though their heart may be, these re-engineering ideas from the vendor won't suffice to scale up relational databases to the desired levels. And more often than not, the only thing that's resulted from forcing a scale-up is accumulation of more expensive hardware without gaining any significant advantage.



Unscalable RDBMS are not fit for cloud based applications

'Mixed workloads' is not possible with them

Mixed workloads is the ability to handle both operational and analytical workloads. It was in the mid-1990s that a divide happened between databases that are optimized for operational workloads and those optimized for analytical workloads. This led to a multitude of distinct data warehouses, reference data stores, archives and data marts being created as a result of which the IT departments of organizations are having a tough time grappling with all the complexity. They require simpler solutions which would help deliver information in its various forms to different users whenever the information is required.

UNLEASHING HUGE AMOUNT OF DATA WITH E-COMMERCE

Flipkart analyzes around 25 million rows of inventory data every day to enable the relevant product teams to perform data-driven decision-making. Moreover, other online retailers also generate up to 40% of orders via Big Data tools.

Modern app development is not feasible with them

To build contemporary apps, object oriented programming languages are used. These languages consider data structures as objects which encapsulate both data and code. This approach is quite different from how relational databases handle data. As a way around this a method called object-relational mapping is used by developers. In this, app developers use business rules and logic to create data views that are most sensible from their perspective. The trouble is that object-relational mapping brings in complexity which in turn leads to performance degradation and higher chance of error in code.

RDBMS fails to fully support App development

It's not possible to track time-varying data with them

'Temporality' as it is known, is the property of a data by which it varies over time. Methods for managing time with relational databases could vary depending on the vendor and the development of such databases can be inhibitively complex. Also, managing bi-temporal data, which requires tracking when events happened as well as the time they were recorded is a greater challenge for relational databases. Whereas organizations may require precise histories of their data for analytics or regulatory compliance, relational databases owing to their rigid structure limit the possibility for the same.

They aren't effective for search based on relevance

Whereas search engines such as Google does and excellent job of getting you relevant data against your search parameters, a relational database would fail pathetically in the same. For with a relational database value-information that are contained in unstructured free texts are automatically ignored. The eschewing of such information prevents the user from getting the information that they need. And the more complex the question is, the harder it is to get the right results from a relational database. For instance, whereas a query like the number of people who have registered for blood donation camps since 2003 may return the right results, a search for the number of people registered for blood donation camps since 2003 and who have a medical history that put them in a high risk category for blood infection may not yield the right results.

The Enterprise-Class Features of relational databases are ineffective for Big Data workloads

What made relational databases the darling of organizations for decades are the high level of security, reliability and data management that they offered. However, such enterprise-class features become practically irrelevant since they cannot be applied to big data. Also, feasible options that cater to big data while maintaining the enterprise features crucial for businesses are available in the market. In other words, relational databases are hardly in the game anymore.

Big Data is needed to unlock the restrictions of Relational Databases

With these many impediments standing in the way of putting relational databases at the service of big data, the future of data management doesn't seem to have relational databases in it, at least not the future as it is envisioned now. As for alternatives, one of the most successful models being used by establishments these days is NoSQL databases. The SQL databases have the advantage of not having any rigid predetermined schemas as with relational databases. Even though NoSQL databases aren't used exclusively for big data they are particularly suited for it. And companies have a good number of options when it comes to choosing NoSQL database providers-Couchbase, Cassandra, MongoDB, MarkLogic, Datastax and Basho are just some of the leading vendors in the arena. d

THE RISE AND RISE OF HADOOP

The technology that is synonymous with Big Data, Hadoop is in high demand.

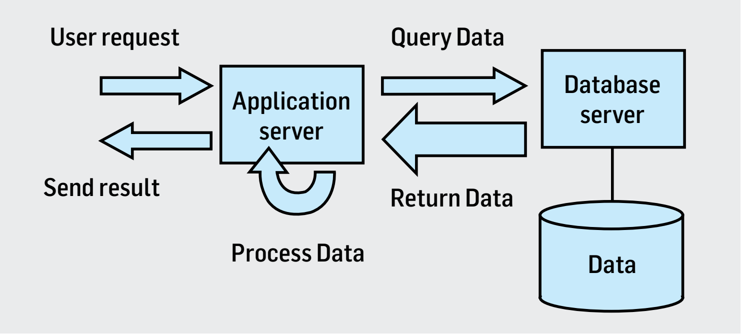
Introduction

Imagine, if you ever wanted to store a file having a size larger than your PC's storage capacity. There is no way you could do that, right? What if we told you that you could store files bigger than what can be stored on one particular node or server? With the advent of concept like Big Data and platform like Hadoop, it is possible to store very, very large files and many, many files.

The world, as we know today, is turning digital and with this digitization the amount of data being created and stored is exploding. This data, structured or unstructured, can be garnered from various sources like social media, data from internet-enabled devices (including smartphones and tablets), machine data, video and voice recordings etc. Almost 2.5 Quintillion (that's 18 Zeros) Bytes of data is generated daily. This would roughly be a stack of Blu-ray discs approximately of the height of Eiffel Tower back and forth, twice. Let alone on Facebook, 300 million photos are being uploaded and 4.75 billion pieces of content are shared daily. Big Data came into picture to address this massive volume of data processing and storage.

Traditional Approach of Data Storage and Usage

Traditional data systems, such as data warehouses and relational databases, have been around since more than four decades. They have been the primary method for enterprises to store and analyze their data. Regardless of the existence of other data storage technologies, the major chunk of enterprise data can be found in these traditional systems. However, traditional systems were always designed from the ground up to work with only structured data. With the passage of time, the enterprises needed to store more and more detailed information for longer periods of time, mostly in areas such as Health and Finance. Managing the volume and cost of this data growth within these traditional systems is usually a stress point for IT organizations.



SOURCE: IBM CORPORATION

Traditional approach: Move data to program

ACID (Atomicity, Consistency, Isolation, Durability) based systems and the ideology around them are still undeniable for running the business. These systems took a long time to build and support business decisions that run some of the enterprises today. Traditional systems can store around petabytes (PB) of data. However, these systems were not designed to solve a number of today's data challenges. The speed, complexity, and cost of using these traditional systems to address these new data challenges would be extremely high. Unstructured data isn't organized but contains tags, markers, or some method for organizing the data. Examples of unstructured data include social media data structures (Twitter, Facebook), RFID, GPS coordinates, machine sensors, and so on. The volume of this unstructured data, generated on a daily basis, is enormous compared to the structured data. Moreover, this unstructured data is just as critical as the structured internal data being stored in relational databases.

Storing large volumes of data through traditional systems is very expensive. So for most of the critical data , organizations, relying on traditional systems, have not had the capability to save it, organize it, and utilize it or analyze its benefits because of the storage costs. The ever increasing volume of data, the unstoppable velocity of the data that is being generated , and the complexity of working with unstructured data as well as the costs have kept these organizations from leveraging the details of the data and that pulled them back from making good business decisions in an ever-changing competitive environment. During that time, Google was evolving and it had the challenge to be able to rank the Internet. For this, it had to design a new way of solving the problem. It designed a platform for itself that offered inexpensive storage, was easily scalable, could access data very fast and store structured, semi-structured, and unstructured data to make it easy to analyze the data and find a relationship among the data.

MONITOR FOOTBALL PLAYER ACTIONS

The National Football League (NFL) installed RFID data sensors in the shoulder pads of each player to collect location data which helped to keep track and analyze player acceleration and speed, which was tested during 2015 Pro Bowl event .

Introduction to Hadoop

Apache™ Hadoop® is an open source software based on Big Data technology, written in Java, that enables processing of a very large amount of data by distributing these sets of data across different cluster(thousands of nodes) of servers. It is designed to scale up from a single machine to thousands of machines, and each of these machines offer local computation and storage. Hadoop was created by Doug Cutting and Mike Cafarella in 2005.

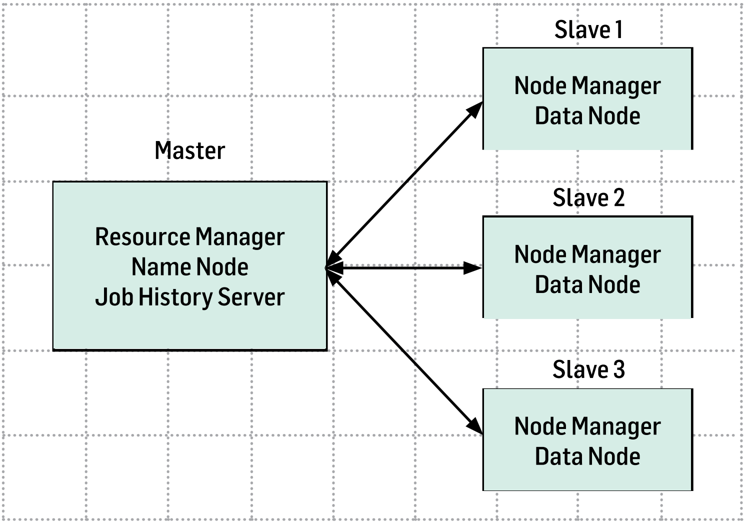
The Apache Hadoop framework is composed of the following four modules:

- **Hadoop Common:** It is also known as the Hadoop Core. It refers to the collection of common libraries and utilities that support other Hadoop modules. It is an essential module of the Hadoop Framework and is designed in such a manner that hardware failures are automatically handled in software by the Hadoop Framework. It is considered as the base/core of the Hadoop framework because it provides basic processes and essential services such as abstraction of the operating system and its file system. It also contains the necessary JAR files and scripts required to start Hadoop. It also provides source code and documentation from the Hadoop Community.

- **Hadoop YARN:** YARN is an acronym for Yet Another Resource Negotiator. YARN is a cluster management technology. YARN makes the Hadoop environment more suitable for priority based operational applications that can't wait for batch jobs to finish. It provides a central platform and resource management to deliver consistent operations, data governance tools and security across Hadoop clusters. YARN extends the power of Hadoop to adapt new technologies found within the data center so that they can take advantage of linear-scale storage, processing and cost effectiveness.
- **Hadoop Distributed File System (HDFS™):** HDFS is a Java-based file system that provides reliable and scalable storage of data, and it was designed to span large groups of commodity servers. Hadoop Distributed File System has demonstrated production scalability of up to 200 Petabytes of storage and a single cluster of 4500 servers, supporting more than a billion files and blocks. When that quantity and quality of enterprise data is available in Hadoop Distributed File System, and YARN enables multiple data access applications to process it, Hadoop users can answer the questions that jeopardized previous data platforms.

The Hadoop Distributed File System (HDFS) is based on the GFS (Google File System) and provides a distributed file system to run on large clusters of small computer machines in a reliable, fault-tolerant manner.

HDFS is a distributed storage system , fault-tolerant and scalable and works closely with a large number of concurrent data access applications, coordinated by YARN. HDFS works under a large number of physical and systemic circumstances. It distributes storage and computation across many servers and through this the combined storage resource can grow in a linear fashion with demand while remaining economical at every amount of storage.



SOURCE: THE CLOUD AVENUE

The Master-Slave node structure in Hadoop

HDFS is based on the master/slave architecture where master consists of a single NameNode that manages the file system metadata and single/ multiple slave(s) DataNodes that store the actual data.

A file in an HDFS namespace gets split into many blocks and these blocks are stored in a set of DataNodes. The NameNode determines the mapping of these blocks to the different DataNodes. The DataNodes takes care of read and write operation with the file system. They are also responsible for taking care of block replication, creation anddeletion based on instruction given by NameNode.

- **Hadoop MapReduce:** Hadoop MapReduce is a software based framework for writing applications that process large amounts of structured, semi-structured and unstructured data stored in the HDFS.

  The term MapReduce can be splitted into two different categories of tasks :

(a) **The Map Task:** This is the first task, which takes input data and converts it into a set of data and then individual elements are broken down into a finite ordered list of elements or tuples (key/value pairs).

(b) **The Reduce Task:** This task takes the output from a map task as input and combines those data arranged as a finite ordered list of elements into a smaller set of list of elements. The reduce task is always performed after the map task.

The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The JobTracker is responsible for tracking resource consumption/availability , scheduling the jobs component tasks on the slaves, resource management, and monitoring them and re-executing the failed tasks. The slave TaskTrackers execute the tasks assigned by the master and provide task-status information to the master periodically.

Economics of Hadoop

Big Data comes at Big Costs. Hadoop is not the only Big Data platform in the industry, but the software has created a buzz around in a short span of time. For Hadoop, the biggest motivator in the market is simple: Before Hadoop, data storage was expensive. Hadoop lets you store any amount of data, structured or unstructured, simply by adding more servers to a Hadoop cluster. These servers are x86 based machines which come at relatively low cost and add more processing power and more storage to the overall cluster. This makes data storage with Hadoop far less costly than its competitors.

"The cost of a Hadoop data management system, including hardware, software, and other expenses, comes to about $1,000 a terabyte--about one-fifth to one-twentieth the cost of other data management technologies." says Charles Zedlewski, VP of product at Cloudera. If we were to store the same terabyte of data at a network storage, it wouldn't be unreasonable to think of a price of $5,000 per terabyte or even more. Moreover, if we were to also include the hardware and other physical resources involved with the network storage, the cost would sky rocket to not less than $10,000-$15,000 per terabyte. On the other hand, if we thought of still sticking to the legacy data management technologies like RDBMS, the total cost might be more like $30,000 to $40,000 per terabyte.

Forrester Research coined the term "Hadooponomics". The main reason for Hadoop being huge in business for 2015 is its economical viability. According to Forrester "Hadooponomics" will definitely make adoption mandatory because Hadoop is simply a cheaper way to form large-scale storage repository of data and conduct analytics queries. Hadoop can scale storage and processing and leverage the cloud. The costs for large Hadoop distributions amount to 2,000 -3,000 dollars per node per year. In contrast, a SAP HANA node costs approximately 750,000 dollars per year. A renowned UK company compared the costs of conventional data storage with the estimated costs for a Hadoop replacement. One TB of data in an Oracle database generates costs in the amount of 48,000 Euro per year whereas it costed around 1,540 Euro per year for storage of the same data volume in Hadoop.

Hadoop Environment Setup on Linux

Hadoop is supported by GNU/Linux. Therefore, we have to install a Linux operating system for setting up Hadoop environment. If you are using a different OS, you always have the option to install Virtualbox.

Step 1 : Create a separate user for Hadoop

Fire up the Linux Terminal and type in the following code :

```
$ su
   password: <Your Root Password>
# useradd <username>
# passwd <password>
   New passwd: <Your Password for New User>
   Retype new passwd: <Your Password for New User>
```

Step 2 : SSH Setup and Key Generation

Before installing Hadoop into the Linux environment, we need to set up Linux using SSH (Secure Shell). SSH setup is required to do perform operations on a cluster. To authenticate users of Hadoop, it is mandatory to provide public/private key pair for a Hadoop user and share it with different users.

```
$ ssh-keygen -t rsa
$ cat ~/.ssh/id _ rsa.pub >> ~/.ssh/authorized _ keys
$ chmod 0600 ~/.ssh/authorized _ keys
```

Step 3 : Configure Java

Java is the main prerequisite for Hadoop. Please make sure that you have Java installed on your system and is made available to all the users. You can verify the existence of java in your system using the following command:

```
$ java -version
```

Step 4 : Download Hadoop

To download Hadoop in your system, type in the following command in the terminal :

```
$ su
password:
# cd /usr/local
# wget http://apache.claz.org/hadoop/common/hadoop-2.4.1/
hadoop-2.4.1.tar.gz
# tar xzf hadoop-2.4.1.tar.gz
# mv hadoop-2.4.1/* to hadoop/
# exit
```

Step 5 : Installing Hadoop

After downloading Hadoop, Hadoop cluster can be operated in the following three modes:

- Local/Standalone Mode (Default)
- Pseudo Distributed Mode
- Fully Distributed Mode

By default, Hadoop is configured in a Standalone Mode and the other two modes are beyond the scope of our discussion.

Fire up your terminal and type in the following series of commands :

Step 5.1 : Set up Hadoop variables

```
export HADOOP _ HOME=/usr/local/hadoop
```

Step 5.2 : Check Hadoop Installation

```
$ hadoop version
```

If everything is fine with your setup, then you should see the following result

```
Hadoop 2.4.1
Subversion https://svn.apache.org/repos/asf/hadoop/common
-r 1529768
Compiled by hortonmu on 2013-10-07T06:28Z
Compiled with protoc 2.5.0
From source with checksum 79e53ce7994d1628b240f09af91e1af4
```

It means your Hadoop's standalone mode setup is working fine. Hadoop is configured to run in a single machine as a non-distributed mode, by default.

Hadoop on Cloud aka Hadoop-as-a-Service(HaaS)

Regardless of the immense popularity of Hadoop, not all organizations are capable of implementing and maintaining a successful Hadoop environment due to lack of expertise in managing the systems. This has resulted in a rise of a large number of Hadoop-as-a-Service (HaaS) providers. HaaS providers present an outstanding opportunity to businesses that need to incorporate Hadoop but don't possess the internal resources or expertise to do so. HaaS providers offer a range of support and features, from basic access to Hadoop software and virtual machines, from preconfigured software in a Do It Yourself environment to full service support options that include job monitoring and tuning support.

The importance and the reason of fame for HaaS is a no-brainer. The technology giants like Google, Amazon, Microsoft, IBM, Rackspace and Hewlett-Packard are all offering HaaS based solutions, in some way or the other. However, all the offerings are still very command line and MapReduce-focused, with the highest level of abstraction generally being a powerful programming language like Python. There might be chances of GUI-based Hadoop services, trying to take some of the responsibility of creating out the complex Hadoop jobs, coming up soon.

HaaS makes sense for similar reasons as running any other software offering on the cloud- quick, one time use cases involving big data computation. For instance, in 2007, the New York Times harnessed the power of Amazon HaaS for a single day to do a one time conversion of TIFF documents to PDF. This could have taken days, if done with the aid of traditional approach or a massive amount of fortune, if they were to setup the entire Hadoop systems and data centre on their own. The cloud, with its promise to access the hardware resources instantly, is very appealing to businesses who need to have a platform that scales fast to meet their growing needs. For instance, it would take several weeks to get a hundred more machines into a data centre whereas it would be available with a HaaS provider in minutes.

SPEND YOUR FREE TIME BY SOLVING PUZZLES

Crossword Labs features several interesting crossword puzzles on Big Data with 6 questions each across and down to test your real knowledge on the evolving topic. If you are an expert in Big Data, you can create your own crossword for others to solve, thus gaining popularity.

The businesses generally collect relevant data from various platforms and pass it to an analytics application running on Hadoop to extract insights from them. The load on the computational resources of a Hadoop cluster vary based on the rate of these incoming data or scheduled runs. A fixed capacity Hadoop cluster built on physical machines is always on whether it is used or not – incurring unnecessary cost. The cloud overcomes this issue with its Pay-as-You-Use model. The businesses can schedule clusters of Hadoop to be working only for that period of time during the day when the data needs to be analyzed and pay only for what is being used.

Future of Hadoop

It is probably not the right time to comment on the future of Hadoop. The opportunities and the possibilities with Hadoop are endless, but there are still certain points that need to be addressed on a priority basis. For several years now, much of the hype surrounding big data has been connected with Hadoop. Companies have simply seen Hadoop as entwined together with the big data movement; where one goes, the other goes, and as a result, the future seemed to be bright for Hadoop and the vendors offering it. Many have pointed to the continued success of numerous Hadoop vendors like MapR , Cloudera and Hortonworks as evidence that Hadoop remains a force to be reckoned with. For Example, Cloudera has a valuation valued at around $5 billion and has almost doubled its annual revenue in 2015. Hortonworks has also seen a growth in revenue by more than 150 percent from year to year.

FAST FORWARDING TO 2025

According to a latest report by IDC, the total capacity of digital data is expected to reach 180 zettabytes in 2025 because of rapid growing number of devices such as smartphones and tablets.

However, a recent report from Gartner seemed to put a damper on Hadoop's aesthetics. According to the report, Hadoop was not growing as expected and that demand for Hadoop was low despite the growth in demand for big data solutions. Moreover, almost 54 percent of surveyed enterprises had no plans to invest in Hadoop in their to-do list. The report revealed that the biggest challenge businesses found was the lack of people with the necessary Hadoop skills to make the best use of it.

On the other hand, the report from Forrester Research contradicts the report from Gartner. As per Forrester, Hadoop have given the enterprises a value for their money and in the coming years CIOs will be making Hadoop a priority. Looking at the growth of Hadoop vendors like Cloudera and Hortonworks, it seems like investors are more than willing to spend their dime on this space and thus the future for Hadoop looks promising.

Regardless of the findings from Gartner, there is still very much scope for Hadoop to thrive and dominate the Big Data space. Gartner Report addresses some of the challenges and obstacles faced by companies, particularly Hadoop's operational complexity. This can be overcome by making Hadoop easier to use and lowering the barriers to adoption and thus more companies may be willing to take the risk. Moving Hadoop to the cloud can be one good way to solve this particular problem. Hadoop vendors should also look at integrating other big data technologies that can complement Hadoop as a way to diversify their offerings. As a technology, Hadoop has a lot to offer and there's a lot of potential for continued growth well into the future.

APPLICATIONS OF BIG DATA: MANUFACTURING AND GOVERNANCE

From better streamlining of processes in a manufacturing unit to helping create a better police force, big data has spurred a wave of changes that's not abating.

Big data in the manufacturing sector

Big data analysis works in favour of the manufacturing industries on multiple fronts. From reducing process flaws to improving production quality while raising overall efficiency, not to mention saving a lot of time and money, the advantages are being garnered by many businesses around the world.

Here, we take a closer look at some of the major advantages and how they are brought about.

1. Improving the process of manufacture

To get a clear picture of how big data helps improve the manufacturing process, McKinsey and Company cites a case where big data is used to glorious results by a biopharmaceutical company for the manufacture of pharmaceuticals. The company in question had to use live and genetically engineered cells while having to track about 200 variables so that they can ensure the purity of the manufacturing process. The resultant products were vaccines and blood components.

Mass scale production is affected by too many parameters to be tracked traditionally

Now, the fascinating thing was that two separate batches of the same substance manufactured using the same process resulted in a yield variation from a whopping 50 to 100 percent!Such variations could result in attracting regulatory attention.

To tackle the problem the project team divided the manufacturing process into different activity clusters. They assessed the process interdependencies using big data analytics and were able to identify nine parameters which had a direct impact on the vaccine yield. Using the analysis the company successfully increased the production of vaccine by 50 percent. Something that gave the company annual savings between $5 million and $10 million.

2. Custom design of products

Consumerism is at an age when everything from advertising to the product design itself is getting more and more customised, catering to the individual tastes keeping in tune with the ethos of 'It's my life' that modern brands seem to be promoting. And companies are increasingly turning to big data to help customize product designs for their consumer.

Tata Consultancy Services cites an interesting case in this regard-about a $2 billion company for which the major chunk of the revenue comes from manufacturing products to specific orders.The company used big data to analyse the behavioural patterns of repeat customers which led to a better understanding of delivering products in a more timely and profitable way.

For its crux , the analysis itself-the major part of it, at least had methods to ensure strong contracts. It also enabled the company to adopt a more lean manufacturing so that they can determine the products that were indeed viable and the ones that ought to be scrapped.

A customized product, delivered on time, always leads to higher customer satisfaction

3. Higher assurance of quality

As far as quality parameters go, Intel is one company that has been scoring high among consumers. Particularly impressive given that the chip maker's products are rarely seen by the consumer. Such high levels of quality come with meticulous attention to details, and with the help of big data analysis.

Intel tests each and every microchip that comes from its production line and this usually entails about 19,000 different tests- for each chip. Savvily enough, The global giant made use of big data for predictive analysis. This way, they successfully brought down the number of tests to ensure quality- a significant drop at that. This resulted in savings of $3 million in cost of manufacture for one line of Intel Core processors. And by further expanding the use of big data for the manufacture of microchips, Intel aims to clock an additional saving of $30 million.

Big Data makes the large number of quality paramters readable

4. Managing Supply Chain Risk

One of the key risk areas that companies face is with Supply Chain. From weather conditions to infrastructure breakdown, there are many things that could go wrong with a supply chain.

Big data analytics helps companies overcome many of these challenges. For instance, predictive analytics enable companies to figure out the probabilities of delay. In addition to analysing weather statistics for such natural

Analyzing historic data helps compnies predict risks to the Supply Chain calamities like tornadoes, earthquakes and hurricanes, analytics also help firms identity backup suppliers and also evolve contingency plans so that the production won't be interrupted in the event of a natural disaster.

5. Increasing the pace of IT integration for a better industry

It has become more than obvious that effective integration of Information Technology leads for better functioning industries. The German government's vision for Industrie 4.0 is an example of a political body putting this understanding to real world application.

Big Data analysis can reveal potential ways to gain production speed

The idea behind Industrie 4.0 was to develop Smart Factories. By using big data to optimize production schedules based on such parameters as availability of machines and suppliers, production schedules are being optimized. Using these optimization methods, the manufacturing chains are making headway in industries that are heavily regularized and have to rely a lot on German suppliers and manufacturers. By forecasting the process flow across multiple departments, IT integration is happening faster than ever before.

6. Measuring compliance and traceability

Thanks to the way relevant and measurable data is made available, measuring compliance and traceability to the level of individual machines is now possible. By using sensors on all the machines in a production unit, operations managers have a better hang on how every machine is functioning. Also, with advanced analytics parameters such as quality, performance and training variances for each machine and operators can be gauged. This helps in streamlining the workflows in a production facility.

7. Quantifying how daily production impacts financial performance

Analyzing every day production values alongside financial impact is highly useful

The connection between the daily production to the financial performance has been hazy for a long time for manufacturers. But thanks to big data and advanced analytics, this scenario is now changing. By having live information that gives a clear picture of how efficiently a factory floor is functioning, production planners can devise methods to scale their operations better. Needless to say, by linking daily production to the financial impact, the chances of improving profitability are also higher.

8. Bringing down relative cost to prototype

Companies can reduce the relative cost to prototype by taking note of the product testings that come before full-scale production. By measuring

Data analysis can reveal potentially beneficial ventures before prototyping starts

the relative time and cost that is required to develop a working prototype and making that data a benchmark, the company will have a better understanding of what can be a profitable project even before the bidding process is initiated.

Big data in the government sector

Perhaps no other arena of human endeavour stands to gain more from big data and analytics than political governance. After all, it doesn't require a wide stretch of imagination to see how improving traffic congestions, evaluating and predicting crime and keeping track of public resources are all distinct possibilities with big data.

And indeed, these are facets of governance in which many political powers are using big data to improve currently but the fact still remains that compared to the private sector, the government sector has been slow in the uptake of big data.

The major problem that seems to be plaguing government sectors across the globe

MANAGING DAILY ORDERS WITH HUGE DATABASES

Amazon.com has implemented Linux based technology to manage millions of orders every day and has the world's largest databases with 7.8TB, 18.5TB, and 24.7TB storage capacities.

as regard to big data implementation is skepticism. For instance, in the U.S alone it is estimated that a savings of 14 percent can be had by federal agencies using analytics programs-that's almost $500 billion. However, skepticism runs high as it was also found that of those who initiated an analytics project, only 31 percent actually believed that data strategies would yield positive results.

A study conducted by IBM called 'Realizing the promise of big data: Implementing big data projects' which is based on interviews conducted across 28 federal, state and city CIOs shows that the majority of those interviewed thought that big data is more a passing fad than anything else.

But that's not to say that governments are completely turning their heads away from big data. The governments of many countries have gradually warmed up to the idea. And there are quite a few ways in which they are putting big data to good use:

1. Improving traffic conditions

The whole idea of modernization, rightly or wrongly is linked to urbanization. And while a urbane life brings its own share of advantages, one of the universally lamented issue with urbanization is the traffic coils that are created in the process-a slow procession of metallic containers on wheels that perhaps exemplifies best one of the worst maladies of city life- of being so close to each other and yet being alone.

Live traffic data analysis is what gives you those accurate transit estimates

By using big data, governments are tackling the traffic issues with an heretofore impossible level of efficiency. By assessing information as to which routes are affected the most by traffic and at what times, it is now possible for the concerned departments to devise better methods to control traffic. Instead of blindly making all the affected routes as one way, for instance it's now possible to pinpoint the exact location on the route from where the traffic issue arises and offer alternative routes for commuters to bypass the lane.

2. Water management

Just like traffic management, one of the highly visible challenges that governments face is with water management. But this is one issue that is not necessarily limited to the cities but also goes far beyond the borders of the urbanist's domain.

A leak that's sprung in the water supply line is one of the most common issues faced in this respect. Big data analysis helps the authorities get a better picture of the areas such issues are most likely to happen-based on past data. By assessing the situation the authorities can then take preventive measures to thwart such incidents from happening in the future. Also significant is the ability to accurately plot the regions to which water supply etc. has to be extended in the future which big data brings. Based on information garnered from related departments, the concerned authorities can devise a plan for effectively extending the services, by causing minimal disruption to the daily lives of ordinary people. This follows the simple principle of being better able to serve by being better informed.

The massive scale of water supply systems in cities could use Big Data streamlining

Analysis of Medical Big Data can help detect disease patterns with any factor

3. Improving medical services in the public sector

Governments can now analyse data regarding the drugs that are prescribed to patients by doctors in the public sector. This helps them gain a clear picture of the types of drugs that are being prescribed, thereby helping to understand whether the patients are receiving the most-up-to-date medicines.

4. Improved energy efficiency

One of the key areas where governments are finding big data useful is the energy sector. By having access to data regarding energy consumption by government factories-down to the last machine and also other public utilities like street lamps etc., it's now easy to find which are the utilities that are wasting energy. To cite a simple example, if street lights along a highway remain turned up during the day, that constitutes a waste of time. Data as to whether a lamp is on or not can be made available live to the concerned authority, thanks to modern technology and this data feed leads to immediate and efficient actions.

Data regarding device usage can be used to track energy wasted

In an emergency situation, correct and fast information is crucial

5. Serving people better during unforeseen emergencies

One of the best uses of big data comes to the government in times when unforeseen events arise. For instance in the case of bad weather, local authorities can use data such as the roads that are damaged in the heavy downpour etc. to bring help bring help to the affected people using routes that are more accessible.

6. Improving adult social care

In states where social care is the norm rather than the exception, services can be further improved by analysing data that would help social workers make better decisions as to when and how to intervene. To cite an example, if someone living alone has a medical condition that warrants periodic checkups, the social service personnel can intervene on their behalf to ensure that the dependent undergoes such medical assessments on time-something that would never have been possible without access to the person's medical data.

Big Data technology can help senior citizens too

7. Higher efficiency of the police force

Big data is already proving to be a great boon for the police forces. By assimilating data that can be procured through social media, the police departments are becoming more responsive to threats and emergencies. The matter of monitoring social media data for identifying possibly nefarious activities-controversial though it is, is still something that's practised by security forces. With unprecedented access to personal and social information, it is also hoped that police forces around the world would become better at predicting and preventing crimes before they occur.

OBAMA KNOWS PROBLEM SOLVING

Barack Obama announced the Big Data Research and Development Initiative in 2012 spread across 6 departments to analyze how the technology can be deployed to solve problems related to government related issues.

These are but the first years of both the manufacturing and government sectors making use of big data, true more for the latter than the former. While it's a fact that meaningful data is more easily available today than ever before, analysing the plethora of data to glean meaning out of it is a skill that still requires a bit of polishing. For the time being at least, such skills are possessed by only a very limited number of individuals in both private and public sectors. And it will only be when the larger population-as both employees and citizens gain a semblance of the skill that all the information out there that goes by the name of big data become a game changer in a huge way. But with the internet becoming a great democratizer of knowledge, such days are not far away or so we can hope.

Meanwhile, governments and companies will continue to benefit from the big data revolution that's underway.

APPLICATIONS OF BIG DATA: MEDIA AND EDUCATION

Learning as well as creation and consumption of media content will never be the same again, thanks to big data

On the face of it, it may look as though they are two disparate things: education and the media. But there are more commonalities among them than one may think. Both are instrumental in informing the people about the world they live in, and both are powerful enough to affect-and change behaviours, sometimes spurring social changes on massive scales.

So, perhaps it's fitting that one of the latest things that's driving industries to the next level in big data is lending its power to both these sectors. But the applications that big data is being put to in each sector are by no means similar or the same. Let's start with a look at the field of education .

Applications in Education

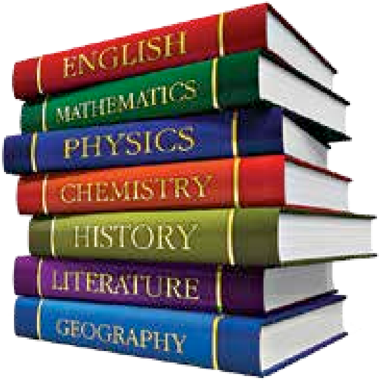
There is no need for an argument to establish that if there's one aspect of human life that always stands to gain from innovation, it's education. Being the foundation for actions related to just about anything, making positive changes in education can have beneficial effects in the immediate society and the world at large-to a much higher level than you may imagine.

Improving education has a direct impact on the society

And big data-being the catch phrase for many a company these days when it comes to improving performance -has a lot of potential in reforming the educational sector as well. Some companies and nations are already harnessing the power of big data to change the way people are learning-by using apps, online learning tools and analysis for shaping a better tomorrow. Here then are some of the most fascinating applications of big data in education.

Better student acquisition

Schools and colleges can make use of past performance and demographics data-pertaining to both current and previous students so that they can create profiles of applicants who are most likely to enroll. This can be augmented with social media data to evaluate how much those students value the establishment. Also, it's possible to leverage the social networks of current as well as prospective students to to identify potential new students from among their first-level friends.



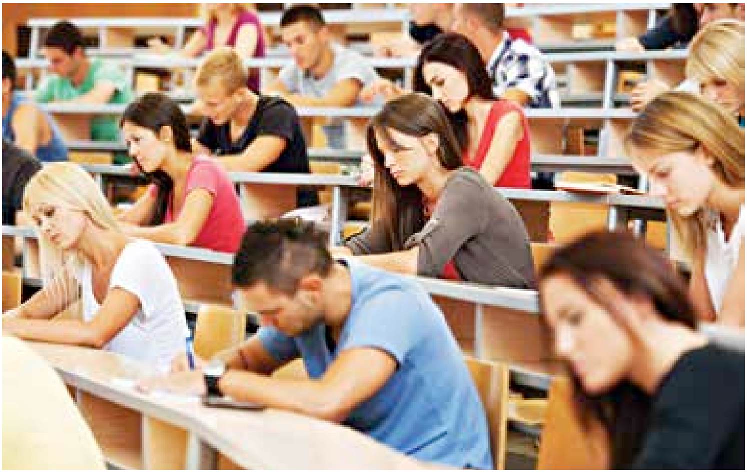
Comparing job data for curriculum choice

Selecting the right course/ curriculum for a student

By creating profiles of students based on their earlier performance, fields of interest from their social media pages and aptitude test results etc., educators can compare the profiles to that of courses to identify a good match. Also, various external data like skill demands and salaries to be expected when they join the workforce could help the student make a more informed decision.

Improve student's performance

The current performance level of the student can be measured by monitoring test performance etc. which can be compared with earlier test results or with results for similar students. Data from social media and teacher's notes could help create a broader profile of the student's behaviours and inclinations. Based on these data, student or class specific recommendations can be made, like individual or group tutoring, extra learning materials for subject areas that the students find hard or even a change of course taken.



Extensive data comparison can be done for analyzing a student's performance

Improve student retention

For an educational institution, very few things can be as bad as losing good students. By clubbing earlier analytics and marks along with such data as demographic, financial and social information one can evaluate the likelihood of attrition and also make recommendations for the institution to decide whether or not the student should be retained or not. This way, the authorities can also successfully intervene to prevent losing the students with higher potential but who are from a poor financial background.

RETAILING HITS ALL TIME HIGH

According to estimates, Walmart handles more than 1 million transactions per hour and are stored in databases having more than 2.5 petabytes (2560 terabytes) of data.

Increase the effectiveness of teachers

While it's all too common to see people creating a hoopla about increasing the performance of students, it's equally important to improve the effectiveness of the teachers if that result is to be brought about. The performance

A teacher can harness Big Data to get a better grasp of their strengths and weaknesses of a teacher can be measured using data pertaining to the subject matter, number of students, the demographics of the students, their aspirations and behavioural categories among others. This can help the teacher get a good idea of the areas s/he has to improve on. Not only that, this sort of analysis also helps in matching the right teachers to the right classes.

Effective problem management

Unfortunately enough, it's almost inevitable that there would be problem-makers in any student body. The application of big data could also make it easy to curb problems. For instance, checking previous data could give teachers information as to whether a student has indulged in malpractices before. Also, when it comes to such things as assignments, teachers always expect the submitted material to be original and not plagiarised. There are companies like iParadigms that use big data to compare a student's written piece of work with online resources and public databases. This way, it can be verified that the submitted material is indeed original.

Trends can often reveal potential problems

Applications in the media industry

In a way, it's funny-the fact that everyone is so hooked on entertainment even when the modern man and woman repeat the phrase, "Life is to be taken seriously" like a mantra. And the thirst for entertainment is so huge that even conventionally "serious" content like news must be catered in the mode of entertainment for the consumption of an audience. Perhaps it's the fact that life in the 21st century is too logical and serious that brings about this situation. Whatever be the case, the media industry doesn't show any sign of slowing down, coming out with ever more ways to keep the population entertained than before. And Big Data, it seems is helping them in multiple ways.

Predicting what the audience wants

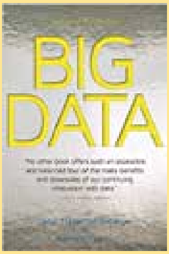
That's practically the holy grail for media companies. And it is revolutionising the way in which shows, movies etc are being conceived-not

Content prediction is the holy grail for the media industry

necessarily as a flash of inspiration but as an amalgamation of carefully selected data points. Knowing what your audience wants gives a media company an obvious advantage over their rivals. And one fine example of a company making use of the "predictive knowledge" that big data brings is the bidding war between HBO,AMC and Netflix for the Kevin Spacey starrer show, House of Cards.

All the three companies did realise that the show would bring in a lot of money. However, it was Netflix that used better data for their bidding. Their bidding was based on their fine-tuned analysis of viewing habits that took into consideration millions of showings. The analysis gave them insights into the types of shows that engaged the audience and realising that House of Cards was in that category, they gained enough confidence to make a bolder bid for the show, which led to them winning it.

LEARN BIG DATA WITH THE BESTSELLING BOOK



Big Data: A Revolution That Will Transform How We Live, Work, and Think is the first major book on the topic. Authored by Viktor Mayer-Schönberger and Kenneth Cukier, the book explains the hard topic in 272 pages and was selected as the finalist in Financial Times Business Book of the Year.

Creating new products

Media is traditionally considered as a game of gambling and going by that view one can say that more science is being pumped into the game, thanks to big data-making it less of a gambling in the process. This is true when it comes to creating new products based on big data driven insights than on hunches. For instance, The Weather Company escalated its capabilities so that they could sell as news services weather data and insights. And one of the new products that they came up with was WeatherFX-essentially a marketplace service with which advertisers could match display of their ads with various weather events, using the insight that some products would be more likely to sell under particular weather conditions than others.

Time based preferences can optimize program scheduling

For creating an optimized scheduling

Analyzing big data gives media companies the understanding about the types of contents that customers are most likely to watch at different times. They could also glean insights about the type of device that would be using to watch something. This data can be evaluated(sometimes the data can be available on a very local level-based on zip codes etc.) and a highly optimized scheduling can be created.

To increase customer acquisition and retention

Big data goes a long way in helping companies understand why consumers subscribe or unsubscribe their services. This helps them develop promotional and/or product strategies accordingly. In a media landscape where the competition is so tight that every single customer matters, this is the kind of insight that would give one company the edge over the other. Interestingly enough, it's not just the availability of data that gives one the advantage, it's also the way you analyse the data, and even which part of the myriad data you are looking at. For instance often overlooked data like sentiments expressed in social media or through email-that might have been considered as peripheral to the current context at that time could lead to great insights about the reasons why the customer sticks with a subscription or not.

Improved ad targeting

There has never been a better way to understand how people consume entertainment and their related behaviour than the present time, thanks to the transparency and the power to provide instant feedbacks that the digital revolution has brought in. And these insights taken along with demographic data would help advertisers create highly personalized advertising that could be pushed in the right context at the ideal time and place. One example-since people may consume entertainment content on multiple devices, big data can be used so that advertisers can understand when a particular consumer may use a second screen. This way, advertising campaigns can be better optimized across devices. Also, media firms will be able to raise their digital conversion rates if they offer micro-level segmentation of their customers to their ad networks.

Ads can be directed based on analyzed data

Some of the best educators consider creativity to be the basis of education. At least, they think that it ought to be made the basis of education to make the learning process more organic. And in the media world, there are many who lament that there's a general lack of creativity in the way the industry functions. That the shows and movies which are being produced are rarely fresh in terms of content.

Perhaps, big data with all the insights that it brings to the table would help remedy the scenarios for both the industries. By identifying the strong points of a student most accurately, one could provide him or her with the creative impetus to further strengthen the abilities. And by understanding that the audience has interests far wider than the things that are portrayed in the entertainment that's being given him/her, better shows and movies etc could be created. In other words, as with many things in the world, the potential is huge, it's how you use it that will determine how it will shape the world. d

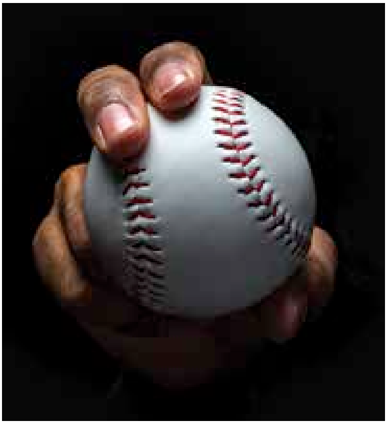
APPLICATIONS OF BIG DATA: SPORTS AND HEALTHCARE

We take a look at how processing massive amounts of data has helped these two industries – each involving high stakes and low tolerance for errors – expand and thrive in today's competitive world.

Part 1: Big Data In Sport

Analytics – a brave new world

Watch any sports broadcast today, and you are bound to be inundated by a dizzying stream of relevant numbers. Football channels will pounce on any break in play to show you ball possession percentages, pass completion percentages, chances created, shots on target, and the number of interceptions, tackles, blocks and dribbles. Cricket, between deliveries, brings you strike rates, batting/bowling averages, run rates and projected scores.



The Moneyball theory made data analytics in sports mainstream.

Fuelled by an exponentially growing semiconductor industry, commercially available computing power has dramatically increased over the past decade. With the growing commercialisation of sport, people's obsessive need to quantitatively deconstruct everything in sight resulted in increasingly obscure statistics being recorded and scrutinised. As a result, the statistical analytics domain has extended its purview to the world of sport, and for the most part, it has revolutionised sport as we know it.

At the end of the 2002 season, American Major League Baseball team Oakland Athletics were in trouble. They were on a limited budget compared to the bigger teams in the league, and several important players were about to leave at the end of the season. Desperate to stay competitive, general manager Billy Beane teamed up with Paul DePodesta – a Harvard economics graduate and baseball scout who was heavily reliant on statistical analysis. They found that most teams paid huge salaries to players scoring high on "old-school" baseball stats such as batting averages and total runs, but that those weren't good indicators of success. However, the "on-base percentage" (OBP) of a player was extremely underrated as a metric, and it turned out that it predicted the success of a player far better. Recruiting players based on OBP numbers turned out to be a stroke of genius – it gave them players who performed well, but whose market value was very low – as they reached the playoffs the next season.

That fateful season inspired the 2003 Michael Lewis book 'Moneyball' and the 2011 movie of the same name. More importantly, it changed the way the American baseball league – and eventually the world – analysed sports. There are a few lessons to be learnt from the Moneyball episode. First, identifying statistically significant factors for a team's success can give you a clear advantage. But second, and more importantly, this advantage does not last long. Sports is commercialised, and that brings with it the fundamentals of capitalism. Rivals soon picked up on the improved scouting methods, and the Oakland Athletics' competitive advantage was reduced within the next two seasons.

Conclusion: statistical analysis in sports is tricky business – there's definitely an edge to be gained if you're onto something, but the ubiquitous nature of statistics and the huge number of people working on them will ensure that the advantage is quickly evened out between rivals. What starts out as a competitive advantage one day soon becomes the bare minimum to be matched the next day. Another point is that success in sport always has – and always will – include the element of blind, stupid luck. No amount of mathematics can completely resolve the fine margins of uncertainty or account for human error (or genius) that so often almost singlehandedly decides the fates of teams. However, statisticians continue to drill deeper into all numbers available – and, indeed, try to obtain more of them – in search of even the smallest of payoffs, however short-lived those may be.

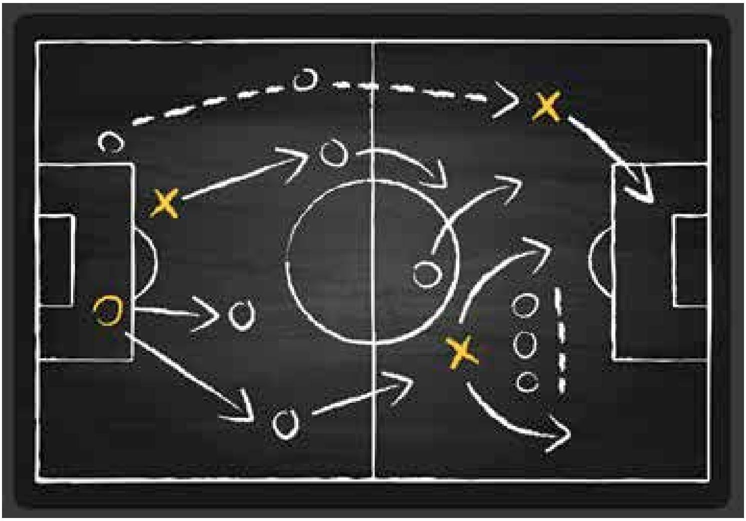
RELATIONSHIPS ARE NOT SIMPLE ON THE BACKEND

Matrimony.com, which adds over 12000 new subscribers on a daily basis makes use of IBMs latest technology solution not only to measure insights but also for the accurate prediction of matches from several petabytes of data.

Order from chaos – 'Insight' and 'Expected Goals (xG)'

We take a look at two of the most prominent applications of big data analytics to sports today: ESPNcricinfo's number-crunching tool, 'Insight', and an emerging way of analysing football matches that attempts complex analysis while retaining footballing sense amidst the confusion – the Expected Goals (xG) method.

ESPNcricinfo has established itself as the de-facto standard for official match statistics. It has lived a remarkable life. Beginning life in 1993 as a bot on the IRC (Internet Relay Chat) cricket channel that provided score updates on request, it had its own website and had catalogued all the test matches ever played by 1995. Over the next few years, its live coverage of matches made it extremely popular all over the world, which led to its acquisition by the Wisden group in 2003, and within 2 years Cricinfo was capable of



Football's obsession with statistics is growing in tandem with the application of big data technologies to all domains.

graphical representation of the match actions and had made its live coverage more comprehensive. It was rebranded as 'ESPNcricinfo' after it was bought by ESPN in 2007. If any doubts of its popularity remained, they were laid to rest when their servers crashed due to the traffic when Sachin Tendulkar scored the first ever double-century in an ODI on 24th February 2010.

ESPNcricinfo released their stats tool on Saturday, 14th February 2015, the inaugural day of the Cricket World Cup 2015. Driven by over 20 years' worth of detailed cricket data, it offers an easy interface for the casual consumer of the sport to pull up a neat data visualisation of almost any player statistic. For example, looking at Ravindra Jadeja's stats from all formats of the game, you can filter out his performances against Australia and notice that he is markedly stingier than average when it comes to conceding runs (an economy of 3.32 compared to 4.00 on average), and also has a much better bowling average than usual (25.60 per wicket as opposed to 31.04). You may want to drill down further and look only at his tests against Australia, which leads to even better numbers. In his 4 tests against them, he has bagged 24 wickets at an average of 17.45 at 2.16 per over and 48.33 balls/wicket. The corresponding stats for his overall test performances are 68, 16, 23.76, 2.27 and 62.77 respectively.

Bringing this kind of detailed drill-down ability to the last mile and enabling the casual consumer of the game see the sport with such extreme detail and convenience would have been impossible without big data. Every delivery ever bowled, every stroke ever played and every catch ever dropped – only because of a mammoth effort at cataloguing all possible detail has the world been rewarded with this magnificent view of the game that is arguably better than the real thing itself.

ESPNcricinfo is not, however, merely a data visualisation web app. Teams will naturally want to extract a lot more than a few colourful bar charts for measuring up their opponents. In exchange for a subscription fee, teams can obtain even deeper analyses and can fine tune the data even more to suit their purposes.

Another attempt at bringing clarity to an otherwise chaotic sport is the recently developed Expected Goals (xG) method of predicting football matches. Several people have developed their own xG models for predicting the final score-line based on the number of shots a team takes in the game, and the model has come under a fair amount of criticism as well. However, it comes with a lot of positives.

First, its aim is simple, and makes footballing sense – predicting the number of goals a team "should have" scored. Shots lead to goals – some shots are more likely to lead to goals than others. The xG method attempts to account for this in several ways – shots from farther out are less likely to be converted than nearer ones, shots taken right in front of goal are better bets than those taken from an angle, and shots taken after a rebound are more likely to catch the defence off-guard than those taken against a prepared defence.

Second, interpreting the results is easy. Comparing expected goals against the actual result of the match, we can tell whether the team underperforms or exceeds expectations. It's conceptually easy to grasp the idea of a ratio of empirical value divided by a theoretical value as a metric for performance of a team, and the accuracy of the model itself.

Third is the fact that it refers to goals as the objective. It's common in football for statistics such as ball possession or total passes completed to be emphasised to an unhealthy extent, often without context or consideration towards the nature of those passes or the zones in which the possession was kept. Goals are an objective metric using which football games are decided, and therefore the xG model is pertinent in football analysis and prediction.

As mentioned, there are several xG models made by several different people. However, the underlying premise in each remains the same – that of arriving at an "expected" number of goals for a team given details about the shots it took. The following discussion focuses on the xG model devised by Michael Caley, Analytics Editor for Howler magazine – a US publication about football (soccer, for them). It is by no means perfect, but then again, no predictive model ever has been.

Caley's xG method considers the following factors in order to calculate the probability of scoring a goal from each shot.

Shot location: it's found that as you go farther away from goal, the likelihood of scoring decreases exponentially with distance. Of course, the angle from which the shot is taken is accounted for as well, with an "adjusted distance" metric calculated as the actual distance from goal divided by the angle raise to the power 1.32.

KEEPING TRACK OF WATER LOSS

The Kerala Water Authority (KWA) has deployed IBM's Analytics and Mobility solution to analyze, track and manage water distribution in Thiruvananthapuram. The system regularly monitor and issue alerts using sensors and intelligent meters and the overall revenue collection improved to around 20%.

Shot type and assist type: For various combinations of these two, the exponential decay rate is empirically adjusted. For example, a header from a cross will become difficult to score very quickly as you move away from goal.

Speed of attack: a linear relationship was found between the speed of attack and the likelihood of a goal – since a fast break usually happens when the opposition defence is out of position.

Other factors like whether the shot was from a direct freekick or a rebound – which greatly decrease and increase the likelihood of a goal respectively – are accounted for similarly in the exponentially decreasing function.

Finally, all these values corresponding to each shot taken are added up, yielding a single value – the expected goals.

However, this system has some important shortcomings that yield inaccurate results for what would be considered boundary conditions or edge cases.

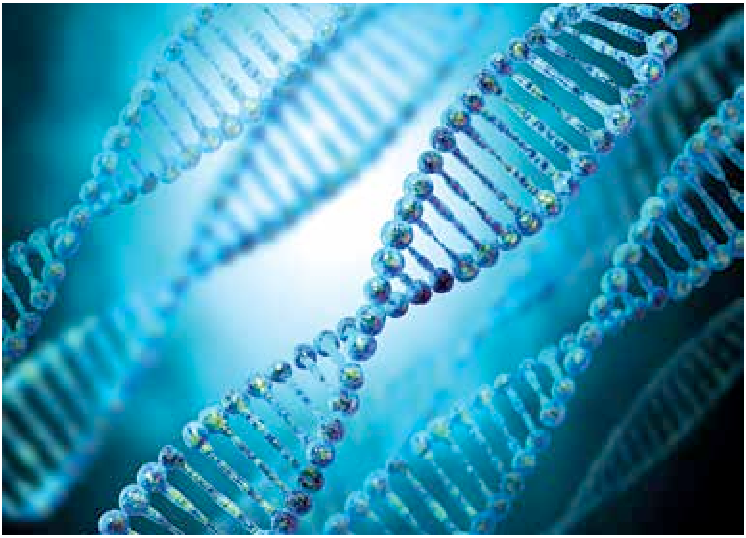
First, the position of the defenders at the time of the shot is not accounted for. Sure, metrics like the speed of the play indicate the defensive pressure to an extent, but the relative positions of the defenders themselves are almost always crucial in determining how likely the attacking team is to score.

Second, it doesn't take into account the skill of the player. There will obviously be a far greater chance of Lionel Messi scoring a goal from 20 yards out rather than Andy Carroll. Furthermore, forwards are significantly more likely to score than defenders, and so on.

Third, its results are skewed when the model considers the very best clubs in Europe, such as Barcelona, Bayern Munich and Real Madrid. Their actual goals scored are far higher than their expected values. This may be explained by their playing style, which tends to yield better chances than usual, and also simply by the high skill level of their players. This sort of qualitative description intrinsic to the game is difficult to factor in immediately into the equation.

Part 2: Data Therapy – Big Data In Healthcare

Another industry that has benefitted from the computing boom has been healthcare. As reception desks started including computers, patient records



Sequencing the human genome may hold the key to the holy grail of healthcare – personalised medicine.

began to be digitised. Eventually, the digitised records reached a critical mass and have now become the primary way of storing an individual's healthcare information. Pharmaceutical companies now involve data scientists to process those immense volumes of patient data for various objectives – patterns related to treatments, side-effects of drugs, and cost-benefit analyses are all emerging from the rich minefield of big data.

The nature of healthcare data is inherently extremely complex. It includes data of a scientific, demographic, personal data and, to an extent, qualitative nature as well. This complexity and richness of the data has prompted several software companies to develop analytical tools built specifically for the healthcare industry. Such companies and their tools have proliferated in the past half a decade or so.

Over the past few years, there have been several factors that have collectively given big data the requisite push to go mainstream in the healthcare industry as well.

First, increasing healthcare costs have brought in a need for a change in the approach to healthcare. Also, with the incomes of pharmaceutical companies as well as physicians being increasingly linked to the success of the treatment and not just the treatment numbers alone, there is now added incentive to find effective treatments. Second, the advent of dedicated healthcare analytics tools, along with the availability of large scale data, has made it feasible to apply big data in finding treatments. Finally, seeing how big data has brought positive results in other fields, it would appear to be a good bet for the healthcare domain as well.

Indeed, remarkable improvements have been made in healthcare once the medical records were available and the data could be mined for patterns. It has set up virtuous cycles that sustain themselves.

First, the trends in medical data have led to people being able to make cleaner lifestyle choices on their own, allowing people to take responsibility for their own well-being. Prevention is better than cure. The advent of wearable tech has allowed for fitness tracking devices such as the Fitbit and Samsung Gear Fit to become commonplace. The data from your body can be compared in real time with millions of people around you to predict an illness accurately, before it even occurs.

Second, treatments are now dispensed with the added support of empirical evidence. Treatments are now described not only in terms of biological impact, but also include a probabilistic description of their effectiveness. This kind of mathematical confidence has given a great boost to the credibility of modern medicine. A treatment can be trusted more when it has been proven to work successfully with other patients with a similar condition and lifestyle. This enables the physicians to dispense the treatment with more confidence, as well as the patients to accept the same with greater trust.

Third, such an approach makes it possible to reduce costs while maintaining the quality of treatments, since increasingly accurate predictions and diagnoses can now be made. Big data will make it easier to streamline medicine testing, allowing them to pick test subjects more intelligently and test treatments in a more controlled manner.

And finally, as predictive models are continuously improved, the innovation drives productivity in the industry and establishes an overall safer environment in the society. The holy grail of the healthcare industry is personalised medicine. Every individual receiving treatment specific to his/her precise genetic configuration would be the ideal scenario for a utopian healthcare industry. Such personalisation can only be achieved after understanding the human gene sequences in their entirety, and genome sequencing is an entire branch of bio-computation that involves gigantic amounts of data to be processed.

What does the future hold for big data? Like with any branch of computer science, big data – like any kind of data – will always have the threat of privacy and security breaches looming over it. The recent struggles of Apple against the FBI, though overtly unrelated, provide a fairly accurate precursor to a debate that will inevitable intensify in the near future. The value of such voluminous and personalised data is immeasurable, and everyone is only just beginning to grasp that.

KEEPING TAB OF RECORDS IN ONLINE GAMING

Image - http://bit.ly/1QJMDOv
Reliance Games examines huge data bank of nearly 38 million records on a daily basis. With over 50GB of data is being added daily, 400 events are captured from each device for each game session.

APPLICATIONS: BANKING

With millions of transactions happening every second, how exactly is the banking sector dealing with all that data?

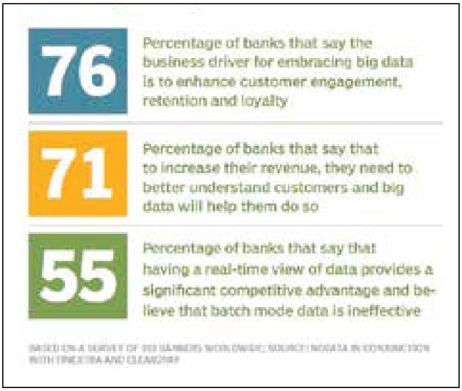
Overview

Banking is one of the many sectors in the economy that has to deal with massive volumes of data every day. Till now, the sector has been traditionally dominated by paperwork and bespoke non-standardised systems, which involve gigantic amounts of manual labour. However, the sheer quantity of data flowing into the banking sector makes it a perfect applicant for employing big data techniques to glean useful insights, which would help increase efficiency for many banks and generate additional streams of revenue.

Big Data benefits

Many key areas can benefit from data science. To cite a more general and clear example, the 2008 financial meltdown, which affected economies around the globe and led to huge debts and bankruptcies, could have been avoided by the use of big data systems to monitor data related to mortgage backed securities (MBSs) and their underlying assets. Such a system would have been able to detect anomalies in the provided data and help prevent the onslaught that followed, which still haunts the world economy today.

For a bank to properly determine which methods are to be employed, data science techniques such as hypothesis testing, crowdsourcing, machine learning, natural language processing and visualization can be applied to increase throughput and discover insights that were never thought to have previously existed.



76	Percentage of banks that say the business driver for embracing big data is to enhance customer engagement, retention and loyalty
71	Percentage of banks that say that to increase their revenue, they need to better understand customers and big data will help them do so
55	Percentage of banks that say that having a real-time view of data provides a significant competitive advantage and believe that batch mode data is ineffective

BASED ON A SURVEY OF [illegible] BANKERS WORLDWIDE; SOURCE: [illegible] IN CONJUNCTION WITH FINEXTRA AND [illegible]

Banking on Big Data

Fraud detection

In the current context, the areas within the banking industry that present the strongest near-term opportunities for tangible performance improvement include risk management/fraud detection and improved customer communication and loyalty.

One of the main priorities of the banking sector, espeically after the financial crisis of 2008, has been timely fraud detection, with one the earliest examples being the FICO Falcon fraud assessment system, which operated under a neural network shell. Most premium software available in the market have evolved and measure up to current standards, by leveraging the power of state-of-the-art big data techniques to capture information available from the copious amounts of data. Such systems usually look for anomalies buried in the spending patterns of a customer. These patterns are usually captured in from their transaction data, which contains the location, amount and time of said transaction, to detect whether any sort of credit card fraud or identity theft has been committed. Such methods can also minimise cases of false alarms being raised (false positive), and help smoothen the customer experience.

Customer Satisfaction

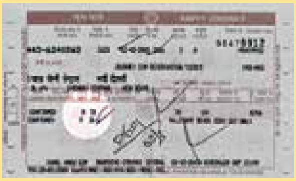
A different, but popular use is in risk and credit profiling. When a customer applies for a loan in a bank, banks have traditionally been known to take days to process the credit worthiness of a customer before replying in the affirmative or negative for the loan application. In order to better the cus-

Data Science in Banking

tomer experience and use more accurate data to come to a decision, banks have started making using of big data techniques to profile a candidate's credit risk. This is done by combing through previous transactions and bill payment timings and delays of the customer, and even making use of spending activity data, in terms of the transaction amounts and the frequency of said transaction, and then further correlating the information with the applicant's income level and financial health. This turns out to be a goldmine of insights, and makes for wholesome data that can be used to mark a candidate's credit risk, and further reduce the occurrence of picking up a non-performing asset, a problem that plagues a large number of Indian banks in today's context.

The idea of using big data architectures to better understand consumers and then seamlessly match appropriate offers to a customer's needs, allows financial institutions to optimize the management of profit-

KNOW THE STATUS OF YOUR TRAIN TICKET



If your train ticket is under waitlisted status then Indian Rail Train PNR status mobile app will be helpful. The app, by traversing over terabytes of data on the backend, predicts whether your train ticket will be confirmed on the Journey date by scanning the PNR number with 90% accuracy rate.

able, long-term customer relationships, and is one that is slowly catching up. Again, utilising the customer's social media presence and the vast amount of data that comes with it, has the potential to improve both effectiveness and efficiency of marketing efforts. This will not only reduce customer frustration because of unwanted promotional offers and events, but also help improve customer retention for banks and better customer relations.

Tracking transactions

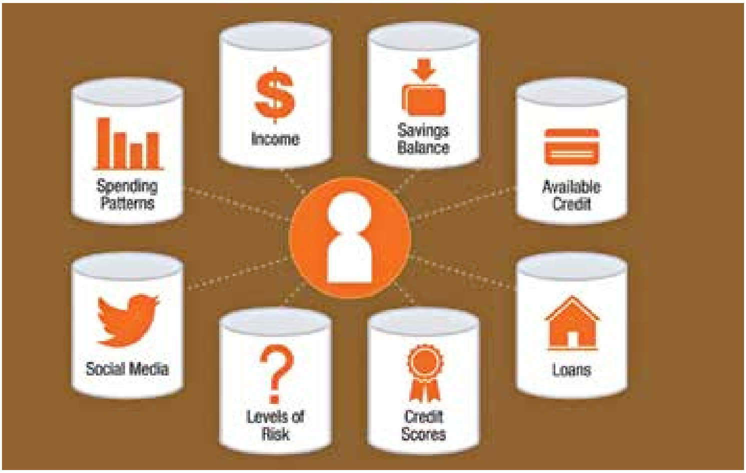
Taking the usage of transaction data a step further, banks can have unique insight into how, where, with whom and when customers are spending money, particularly by pairing this information with the social media activity of the customer. By analysing such information, banks can build an insight into customer intelligence, their spending patterns and behaviour. Unfortunately, bigger banks have been slow on the uptake, and their current usage of the treasure trove of data that is available out in the open has been less than satisfactory. On the other hand, big e-commerce players such as Amazon have successfully harnessed the power of social media and use the information to provide highly personalised services, such as product recommendations.

Another major application, one that was thrust into the spotlight after 2008, is the compliance and regulatory aspect of banking, that can be handled by engaging with data science techniques to maintain records of transactions and document everything that goes into each swap trade by implementing a deal monitoring system. Moreover, regulatory authorities also have the option of harnessing big data to fish out offenders and control the working of a sector that spawns petabytes of data a day.

Trading and portfolio management

But the most widely used function utilising the power of big data, is that of brokers and traders dealing on the stock market. The sheer efficiency and speed of a big data architecture makes it the perfect candidate for the foundation of a system, which provides the future prices of shares, securities and commodities and also predicts future market conditions. In fact, many automated systems that are currently being utilised, are capable of analysing the available data to buy and sell various securities on the market. This particular application now has its own field of study called "quantitative trading", and has taken world economies by storm. In today's context, when trading takes place by the second rather than minute, these systems have proved their worthiness and form the backbone of a huge chunk of trading that is taking place in the world economy.

Better yet, many investment banks have begun to leverage the power of this gigantic system to develop internal portfolios, while making use of market trends analysed by big data systems. The portfolio manager is informed of the variations in the portfolio value, and he/she takes appropriate steps to correct the portfolio, thereby increase their productivity and reducing the time needed to manually calculate benchmarks and trends that are essential to the formation and subsequent maintenance of the portfolio. Further extending the risk management algorithms underlying the banker's system can help extend support at any given time and quickly address any change in the volume of data being captured.



Big Data and Analytics in Banking

A possibility that has not been considered by banks, but is sure to makes wave when widely applied, is the segmentation of customers into different categories, in order to support sales, promotion, and marketing campaigns. This is possible by collecting and analyzing all available data and using big data technology to mine for intelligence from underlying data. Furthermore, on a foundational level, big data could provide the insights to develop segmentation strategies based on the huge volumes of data available on their social profiles. This would allow the organization to provide a highly personalized, consistent experience regardless of the channel selected by the customer.

Thus, the marketing department of a bank can determine which customers to target using the right offers, such as a pre-approved home loan for a customer in search for one.

Looking ahead – How and what banks need to do

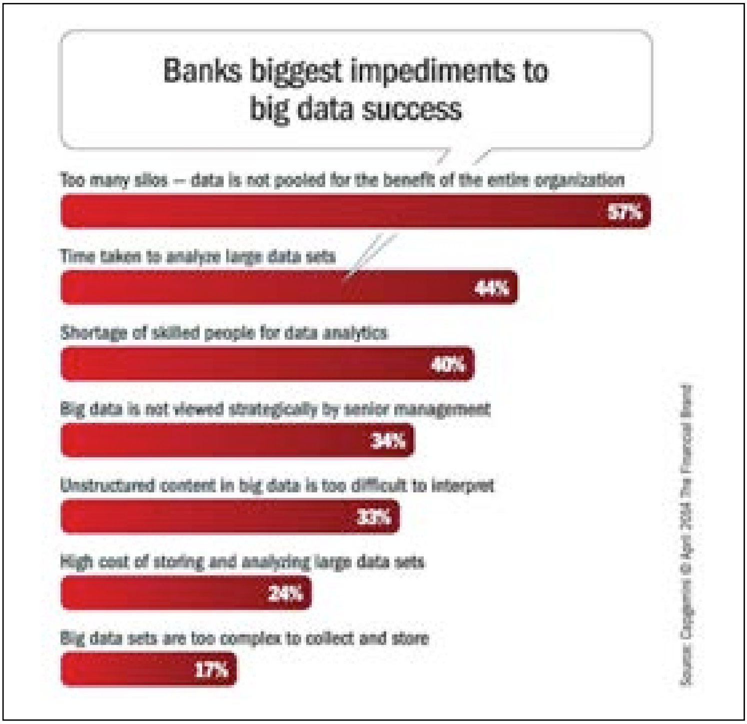
While we've discussed more conventional methods that address the current needs of banks, some more scientific applications can help reduce the risk exposure of a bank and avert any further financial crises. Using data to create operational risk modelling simulations, and using those insights to venture into newer streams of income, is a popular idea among those involved. Unfortunately, such simulations need much larger and concentrated data than is currently available. In order to prepare such a large database, data must be consistently and meticulously stored, in such a way that the relevant data can be extracted at a much later date. This brings us to a slightly inconvenient truth; one that has been hindering the adoption of big data in the banking sector at the moment.

VIEW THE PAST TRACK RECORD OF A CRICKET PLAYER

Insights, developed by ESPN-CricInfo dissect two decades of historical cricket statistics of the relevant player and are displayed in the form of charts and graphs in five different formats. This will help you to judge the best performing player for the upcoming Cricket seasons.

The current problem with the banking sector is most organisations have old and outdated legacy systems in place, which are incompatible with the newer and advanced database architectures. Extracting the older data, and creating a new integrated system with all of the organisations data pooled in under one umbrella, makes this a tedious and potentially costly task. Moreover, senior management in the banking sector are not completely aware of the potential benefits of employing big data systems, hence reducing the number of banks that are moving in the direction of intelligent systems.

However, this needs concentrated efforts from the side of a bank's management to slowly start the application of big data systems. The initial return-on-investment does not seem like it would yield high dividends. But on closer look, the potential for generating newer streams of revenue in the



Source: Capgemini © April 2014 The Financial Brand

Challenges faced by banks towards implementing Big Data systems

future should be good enough incentive for banks to start the adoption of such an architecture. It would begin with the clean and organised storage of data available to the bank, and end with the integrated identification of untapped sources of revenue. After being more aware of the required action, banks would have to allocate a fixed amount for the storage and organisation for newer data, while undertaking smaller pilot projects that would test the effectiveness of newer models. Any project that yielded accurate results should then be implemented on a larger scale, moving up gradually, till the architecture is seamlessly blended in with current workflow of the bank.

Even though technology and banking have never seemed to go hand by hand, management in banks must become more aware of key technological advancements that would benefit them in the long run. Big data initiatives must be perceived differently from traditional IT programs. Only then will banks be able to make the best use of their vast and growing repositories of data. d

APPLICATIONS OF BIG DATA: SCIENCE

Demanding a shift in mindset, big data is posing a challenge for scientists, something that many in the global scientific community rise to meet with bravado.

This isn't the first time in history that scientists are faced with a plethora of data to find logical connections that lead to discoveries. In fact, one can easily make the argument that that's always been the case-after all scientists down the ages have been looking at the greatest data warehouse for inspiration-nature. But what differentiate the age of big data from the preceding ones is the availability of a multitude of data for a very large number of people in the scientific community.

Whereas earlier the finding of data itself would be a strenuous task warranting physical explorations that are demanding as well as time consuming, such data is readily available these days-at least to a large extent. The demanding part then comes with the finding of meaning from the data that's available.

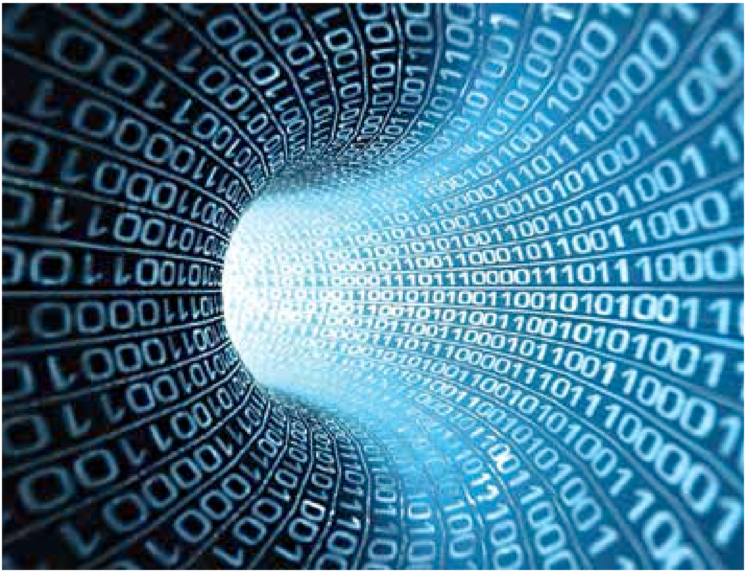
While it is just to assume that scientists would be overjoyed with this explosion of data, application of big data in science is not without its challenges. In fact, before taking a closer look at the applications themselves, it'll be beneficial for the reader to have at least a cursory knowledge of the challenges so that s/he can better appreciate the practicalities of analysing big data for practical scientific applications.

The challenges

1. Understanding huge chunks of data

The past decade has seen an overwhelming advancement in terms of data storage facilities-something that makes storing of huge volumes of data not just feasible but also easy. Due to massive storage facilities computer scientists are amassing data like never before. But collecting and storing information is quite distinct from gleaning understandings from it. The question of how to interpret this large database becomes significant, and a significant brain teaser at that.

The traditional statistical tests as well as computing models for making scientific inferences were created to analyze very small data samples garnered from relatively large populations. These models by their nature



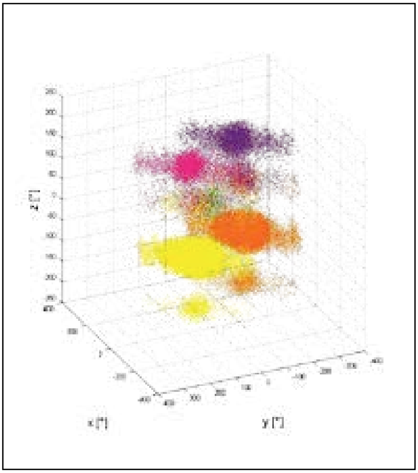
Scientific experiments can be of a massive scale, generating a lot of data becomes unsuitable for processing big data which sometimes entails huge samples that could even include a population in its entirety.

To tackle this issue, new statistical procedures are designed. Many of these procedures have either unknown run times or long run times that make the procedure impractical if the data is seriously huge. When faced with such situations, scientists may have to resort to ad hoc methodologies many of which could lead to drawing the wrong conclusions.

2. The issue of "high dimensionality"

High dimensionality refers to the property by which big data typically contains a host of multiple information about individual parameters that are sampled. And the more "dimensions" that the samples have the chances of finding spurious links between different data-correlations that are nothing but flukes are also higher.

For instance, a medical study could find a link between the success of a medicine with the patient's height. The reality might be that when the big data includes information on parameters including height, body weight, eye colour, shoe size and more some connections may appear to be important just by chance.



Big Data can handle multidimensional data

3. The challenge of overcoming biases

What comes to be known as big data, in fact is often garnered by combining information from multiple sources-something that happens at various times and by using different technologies or methods. This heterogeneity forces the data scientists to design more adaptive procedures to analyse the data. In other words, novel statistical thinking and methods for computation are called for. This in turn demands new ways of looking at the very idea of how science should work or even what science is-a paradigm shift in mindset which is perhaps the biggest challenge of all. After all, resistance to change isn't something the scientific community is immune to, as is evident from history.

Key applications of big data

Formidable though the challenges of using big data in science may be, there are various fields of science in which big data is already finding application-much to the delight of the beneficiaries: the common people. Here are a few significant arenas of science where big data analysis is bringing in a sea of changes.

1. Medical research

Being one of the scientific disciplines that have a major impact on people's lives, medicine is rightly being benefited by big data. And one of the main facets of medicine in which this is seen is in the shift from broader treatment methodologies to specifically targeted pharmaceutical testing. In other words, instead of giving the same treatment to a large segment of population, medical researchers can now devise treatment methods for people with highly specific genetic markers, thanks to the fact that genetic information is now more available than ever before for the medical community.

BUYING A HOUSE MADE EASY

CrediFi makes use of credit risk algorithms to analyze huge amounts of data from public and licensed sources to produce risk scores for a transparent real estate deal involving brokers, lenders, investors and owners.

Also, genomic researchers can use data regarding gene samples to identify the types of genes that are responsible for some particular disease. For instance, the Palo Alto based company, CardioDX made use of over 100 million gene samples which they analyzed to isolate the

Specific genetic data can be analyzed in medical sciences

23 primary genes that caused coronary artery disease. This helped predicting the disease possible.

2. Climate science

Ironically enough, the 21st century which is seeing many parts of the world surging ahead in the modernity wagon also faces one of the earliest challenges of humans- drastic climatic scenarios that range from abrupt droughts to other unexpected natural calamities like cyclones.

Climate data from all sources can be compared to generate accurate predictions

And what with the climate change that's being experienced across the world thanks to global warming, climatic conditions are becoming harder to predict. If this situation remains unchecked, both livelihoods and lives of millions around the world would remain in a perilous situation.

Thankfully enough, big data is proving to be a boon in this regard too. Climate science researchers now posses a huge volume of observational data from sensors that would help them create better models to predict the effects that climate change has on different parts of the world. Whether it be alerting the authorities about a possible calamity in the near future, thereby initiating an evacuation process that would save the lives of many or advising people about the type of crops that would give them a better yield in the changing climatic conditions of their regions, climate scientists now can help the people better, thanks to big data.

3.Military science

If there's one arena that's dependant on gathering intelligence for its success, it's military science. With the communication technology being widely accessible to a large population, eavesdropping on such channels has become something of a routine for military and security forces to identity possible malicious agents.

But that's not the only way in which the armed forces are harnessing big data for their functioning.

Defense departments are often early adopters of technology and robotics is one aspect of technology in which such departments are currently being the forerunners in adopting. Research projects that analyze texts in multiple languages are undertaken by defense departments so that they can create better autonomous systems like robotics. Mechanised systems like robots are often deployed in field operations and assessing the verbal data that come from fields that have people talking a native language become important for such systems to function efficiently.

SECURE A BRIGHT FUTURE WITH BIG DATA

As per Analytics and Big Data Salary Report 2016, a candidate with proficiency in both Big Data and Data Science earns Rs. 13.10 Lakhs yearly than a person with only Big Data who earns Rs. 9.80 Lakhs.

The military often employs cutting edge technology that involves data crunching

4. Nature conservation

One of the major challenges faced by mankind in the post-modern world is that of preserving nature-enough to make sustainable living more than just a fancy phrase. The myriad of ecological data that are available today help scientists devise meaningful plans to conserve nature. The data procured could be as diverse as temperature, measurement of soil and water parameters, data regarding the lifeforms like birds and mammals that are found in the region and aerial imaging. By intelligently analysing these data the scientists can construct conservation mechanisms that are suitable for the particular region. Instead of relying on hunches, the availability of hard data gives the scientists the right pointers to the right conservation measures.

Multiple parameters like temperature, soil, water can be analyzed in nature sciences

As far as application of big data in science goes, the limitation is set only by a lack of imagination. As mentioned before, there's a necessity to let go of biases and think afresh about the various possibilities of efficiently using big data in science. Scientific advancement has always been reliant on meaningful data and on that regard, this seems to be a great time. And what with some of the smartest minds in the world striving to convert huge chunks of data to insights that help make life better for humans, there's every reason to believe that science will amaze us even more in the years to come, thanks in no small way to big data. d

THE NETFLIX RECOMMENDATION ENGINE

Much has been said about The Netflix Prize. Let's see how 10% changed the game.

Introduction – The legacy of Netflix

Right from Frank Underwood's stone cold approach to politics to having one of the most complicated and technologically advanced recommendation systems, the legacy of Netflix has been quite an impactful one. The online streaming website and production house recently started services in India in the month of January prompting a simultaneous sigh of relief from the millions of Indians who no longer have to rely on dodgy websites to download their favourite TV shows and movies. As Indians are slowly discovering, the charm of Netflix goes beyond just having a common portal to access your TV shows. Netflix is your best friend, saving you the time it takes between finishing one TV show to finding another to binge watch, helping you get over the depression of finishing entire seasons in a day by providing you access to an endless stream of entertainment all of which you somehow miraculously enjoy. All of this is thanks to the incredible science that goes behind Netflix's recommendation system; a system so detailed that it even takes into account the variation of a viewer's moods and

The arrival of Netflix in India promises to revolutionise the way we watch TV and Movies

is smart enough to know that Mondays are more depressing than Fridays. In this article, we take an amateur's look at the big data analysis and latent machine learning that goes on behind the scenes of the world's favourite couch potato website.

The Netflix prize – Tiny details lead to big results

Much of the hullabaloo around the recommendation system used by Netflix started when the company announced a $1M prize to any team of data scientists who could predict user ratings for films purely from the history of previous ratings by the user with an accuracy of at least 10% more than Netflix's existing predictor system in 2006, Cinematch. When it was announced, the prize was not viewed as something that would be taken too seriously. In fact, the CEO of Netflix at the time, Reed Hastings admitted he wasn't expecting much from it and he was so confident in Cinematch that he once admitted that they 'thought we built the best darn thing ever.' However, in a matter of just six days, Cinematch was beaten by an algorithm by team WXYZ Consulting. After that the competition became intense for a period of three years, and, in June 2009, the team that called itself 'Bellkor's Pragmatic Chaos', reported a 10.10% improvement over Cinematch and were crowned winners of the contest.

The incredible part of the entries for the contest was the manner in which they quantified what may seem like very arbitrary and unquantifiable metrics. The Bellkor algorithm took into consideration two key ideas. Firstly, they divided the data set into slices that they called 'frequencies'. These frequencies were then grouped together and sorted based on the quantity of

Bellkor's Pragmatic Chaos with the winners' check for the Netflix Prize

movies rated based on the age of the movies. The key idea behind doing this is that the time at which a movie is released and the impression and rating it would have gotten immediately after its release is vastly different from the kind of rating it will receive later. To take an example, a movie like the original Die Hard (1988), is more likely to receive higher ratings now than it would have at the time of its release because of its growth as a cult classic. Some movies simply age better than others. By noticing the fact that users use different criteria to rate movies from different time periods, the algorithms took a huge step forwards in improving Cinematch. Another important factor they took into consideration was the time at which the rating was given. While this may seem pedantic and inconsequential, the cumulative effect of a million ratings all done on Monday, when viewers are less happy and are therefore less likely to rate a movie highly, is quite huge. The winning algorithms are a linear blend of two algorithms: Restricted Boltzmann Machines, a form of neural networking and the so called SVD++, a form of matrix factorization which was developed by the winning team to account for implicit information in the data. These small steps combined together pushed the Bellkor algorithm over the 10% mark. It would be unfair to mention the Netflix prize without lauding the efforts of the Runner-up team, 'The Ensemble', who achieved the same score as the Bellkor team, but were late by a mere 24 minutes. When 24 minutes was

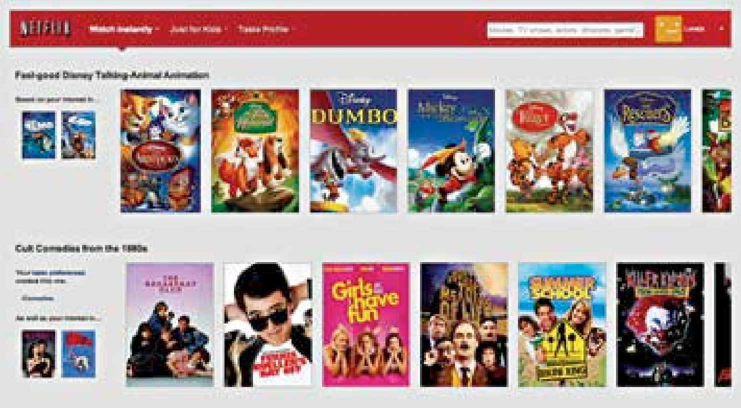
HOST YOUR OWN BIG DATA APPLICATION

Rackspace offers a enterprise-grade cloud environment based on Hadoop on demand and Spark powered by the Hortonworks Data Platform (HDP) with up to 10TB of storage including a separate package of On Metal 3.2TB.

all that separated one team from another and $1 Million from zilch, it gives a magnificent indicator of the kind of competition that existed between the various groups at the time.

The new algorithm

However, the current Netflix recommendation system no longer uses the same algorithm. This is mostly because of Netflix's diversification into an online streaming service rather than the DVD rental service of the 2000's. Overall ratings are no longer the chief concern and the priority has shifted to personalised rankings and page optimisation. So, how does Netflix do it?


NETFLIX
Watch Instantly
Just for Kids
Taste Profile
Feel-good Disney Talking-Animal Animation
Based on your interest in...
DUMBO
Cult Comedies from the 1980s
As well as your interest in...
Girls have fun

A few extremely specific Netflix categories

The current ranking system focuses on ensuring similarity of movies through an extremely selective categorization of genres while still trying to integrate diversity in an active manner, with the users' awareness. Therefore, the system actively 'learns' from the users choices and offers a customized experience. This explains why, apart from generic genres like Comedy and Drama, one is also likely to find categories like 'Teenage fantasy movies from the 2000's' instead of just 'Fantasy', if the user has a history of rating movies like Harry Potter highly. After such customization is made and specific categories are identified, the Netflix algorithm uses a variation of the Belkor algorithm to offer a list of top ranked movies. However, these top ranked movies are no longer based on a generic global system; they are top ranked movies as rated by users with similar tastes and preferences.

Understanding the ranking system

This all important list of rankings are based on many different factors. To understand how these different factors play a role in determining the overall ranking, we will take a simplified model in which the ranking of a movie depends on two factors: The predicted ranking from the Belkor algorithm and the overall popularity of the movie (which has a huge role especially in a socially integrated world where popular movies obtain higher ratings). We will call these factors R and P. For a particular movie, we obtain an overall score (S), using the formula

S=W(R)R + W(P)P

where W(R) and W(P) are the weights. The higher the weight W(R), the higher the priority given to predicted ranking when compared to the overall popularity and vice-versa. The movies and TV shows with the highest overall score place higher on the ranking system. The factors W(R) and W(P) play a pivotal role in this model and they influence the overall effectiveness of the model. There are multiple ways to fix these constants but the most effective way is through a simple Machine Learning algorithm. This algorithm picks up archived data from the servers and, using a large set of data points, varies over a long range of values for W(R) and W(P) and determines the constants which yield the most accurate values

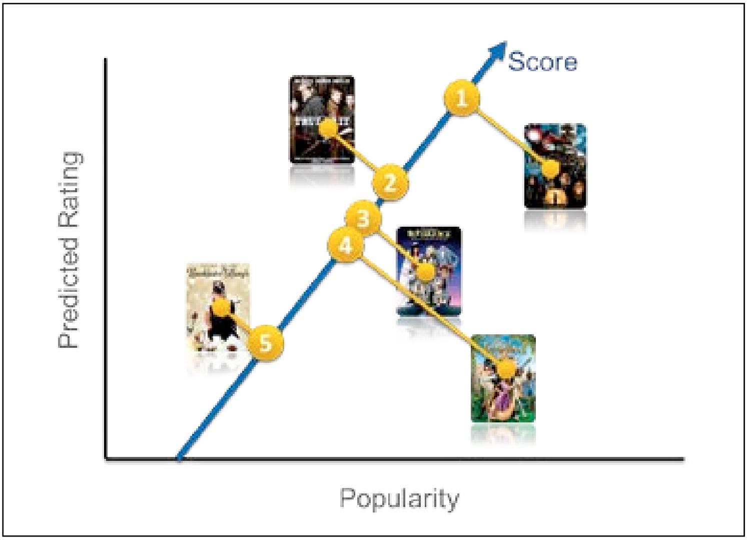
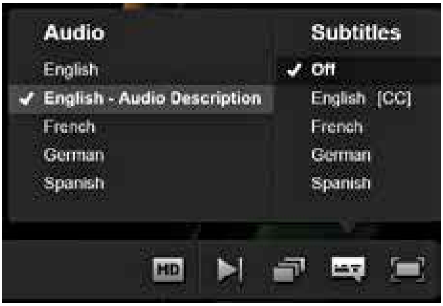


IMAGE SOURCE: WWW.TECHBLOG.NETFLIX.COM

An example of how the ranking system works of the score. The good thing about this method is that it is dynamic; it can be changed depending on the category of movies, the country or even with the demographic of users.

Striving for continuous improvement

These aren't the only forms of data that Netflix uses though. Through a continually evolving process the data scientists at Netflix keep varying the criteria for ranking and experiment with different variations. Their methodology of testing revolves around the same principles: Offline testing of a hypothesis using archived data, selective testing of the data on online subscribers over months and, if the hypothesis proves to be a more successful scheme than the existing methodology, it is rolled out to all subscribers. Multiple metrics are tracked to measure the performance of any particular model, including metrics specifically designed for this purpose such as the Root Mean Square Error (RMSE) which was popularised in the Netflix prize and is a measure of the variation of scores with diversity. The company also innovates with in-house competitions such as the Top10 Marathon, which was a focused 10 week period in which different algorithms were tested quickly, many of which are now part of the Netflix system.



Some of the multi-audio, multi-caption language options on Netflix streaming

As the number of countries in which Netflix operates grows, the demands of the data mining and machine learning algorithm increase exponentially. There is active research currently going on to develop algorithms that are diverse enough to integrate cultural awareness, enabling Netflix to truly establish itself as an international organisation.

Language is also a huge issue as the expansion of Netflix services into 21 different languages brings with it a host of challenges. At the same time, the availability of a service like Netflix ultimately leads to an expansion of tastes and an overall improvement in the internationalism of entertainment. The time is not too far away when one can expect to see Salman Khan dancing around computer screens in Kampala, a thought that somehow fails to be encouraging despite the magnificence of the technology that goes behind it.

The future

Other services are also expanding and offering users the best possible experience by tailoring themselves to suit the user's needs. Music streaming service such as Spotify and even news services are adopting clever algorithms and sorting through Terabytes of data on a daily basis. These personalised applications and services have one simple goal: Offering the best experience for each individual user. It will be fascinating to see how these companies enhance user experience next. d

HASSLE FREE RAILWAY TICKET BOOKING

With the help of Pivotal Big Data Suite, nearly 200,000 simultaneous purchases were made without compromising performance through the official online ticket portal of Indian railways with 150,000 tickets sold per hour, thereby boosting revenue to Rs. 600 million daily.

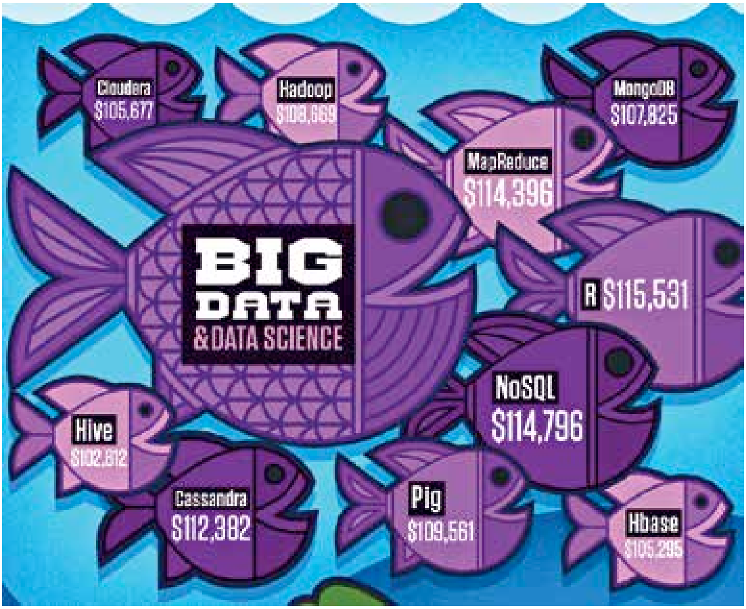
CAREERS IN BIG DATA

With the demand and usage of Big Data increasing everyday, we show you how to pursue a career in this field

Big Data is an evolving area which has ample job opportunities not only in India but also aboard. With the rapid growth of e-commerce companies, the need for storing a huge amount of critical content such as order details, payment information including product catalogs in a secured manner assumes great importance. Moreover, with the Aadhaar project attaining legality, there is a possibility of more service sectors making use of the system. Internationally, IoT is a growing technology that will massively increase the generation of data everyday. All these factors only indicate the growing need of skilled Big Data professionals in India and abroad.

Required Qualifications

In order to apply for a Big Data job, you need to possess a minimum Bachelors, Masters or PhD in Computer Science, Statistics, Mathematics, Physics, and Statistics from reputed colleges which include IITs and IISc. To apply for the data scientist position, you need to have a doctorate in computer science with an aggressive mind to tackle various core issues affecting businesses.



Nine skills that can land you in Big Data

In addition to a basic college degree, you need to have a good knowledge in any one or two of these skills - Apache Hadoop, Cassandra, Spark, Pig, MapReduce, Microsoft Azure, Oracle, NoSQL, MySQL, MongoDB, Microsoft SQL Server, Machine Learning, Data Mining, Statistical Analysis, Data Visualization and Warehousing, Operation Research, Semantic Web, Data Science and Artificial Intelligence in addition to proficiency in programming languages like Java, C, Ruby, Python, or Scala.

The companies will also look how good you are at tackling business problems. You should be able to communicate effectively with other team members in addition to educating them through frequent seminars, presentations and creation of ebooks. You should be ready to work in a fast paced challenging work environment and should be capable of meeting tight deadlines in addition to development, testing and implementation of program logic effectively.

A job in Big Data field requires working with both the back and front-end on a daily basis. You should be quick enough to easily translate business needs into end user applications in addition to an ability to quickly solve problems. Moreover, you should have a confident, passionate and enthusiastic attitude towards your work and other team members.

Industry overview in India

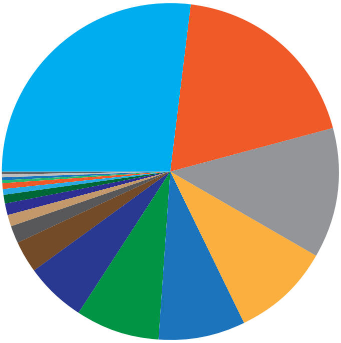
The concept of Big Data revolves around Volume, Velocity, Variety and Veracity. As a Big Data expert, you should have the capability to work on all these spheres. While you handle large volume of data of customers, you should also make sure that the speed doesn't get reduced when it is accessed. Moreover, you should handle a wide range of data from various database formats. When large amount of data gets automatically generated, companies are dependent on Big Data and analytics experts who are not only familiar with programming but also various back-end technologies.

According to estimates, analytics market in India will be doubled to $2.3 billion from the current $1 billion by the end of 2017-18. Moreover, Biometric data of about 99 crore people have been already collected under Aadhaar scheme and stored across two data centres based in India, which needs huge manpower for regular maintenance.

Big Data can also be used in security segment as well, especially with the fact that nearly 15,000 CCTV cameras were installed during the recent India visit of US President Barack Obama. Instead of employing 15,000 personnel to monitor, the technology captures the unstructured data of video streams and run pattern matching algorithms to identify potential problems based on certain pre-defined parameters.

TOP 20 INDUSTRIES HIRING BIG DATA EXPERTISE

Source: Wanted Analytics, 2014



27.14	Professional, Scientific, and Technical Services
18.89	Information
12.35	Manufacturing
9.62	Retail Trade
8.2	Sustainability, Waste Management & Remedial Services
8.13	Finance and Insurance
5.7	Wholesale Trade
3.04	Educational Services
1.71	Other Services (except Public Administration)
1.22	Accommodation and Food Services
1.05	Health Care and Social Assistance
0.76	Real Estate and Rental and Leasing
0.62	Construction
0.46	Transportation and Warehousing
0.42	Public Administration
0.28	Management of Companies and Enterprises
0.18	Arts, Entertainment, and Recreation
0.11	Mining, Quarrying, and Oil and Gas Extraction
0.1	Utilities
0.02	Agriculture, Forestry, Fishing and Hunting
100%	

We expect Big Data industry to grow at a rapid pace with the launch of online stores for all sectors such as electronics, groceries, food, jewellery, books and much more including sectors like libraries, railways, and airports. The analytics can also be used in Space field to capture, store and manage images rendered by satellites. Hospitals also require management of huge volume of records on a continuous basis and a system to measure analytics in a wide range of parameters on a regular basis is a compulsory requirement.

In short, the demand for professionals who are familiar with Big Data technologies is expected to grow in each and every sector which demands management of huge volume of data on a daily basis.

AIRPORT MANAGEMENT MADE SIMPLE

Dubai Airports has implemented Big Data analytics to automatically assign gates in a dynamic manner depending upon the physical location of the aircraft. Moreover, retail stores in the airport scan boarding passes and deliver customized notifications about the departure time and other details

Learning Resources

As you attempt to reach towards the goal of acquiring the required qualifications to secure your future in the Big Data analytics field, you need to master the relevant concepts with the help of learning resources in the form of articles, tutorials, videos, books and courses.

Beginner

- Simple and Easy Learning for Big Data & Analytics - *http://dgit.in/simplebigdata*
- Big Data basic concepts and benefits explained - *http://dgit.in/bigdatabasics*
- Big Data Basics - Part 1 - Introduction to Big Data - *http://dgit.in/bigdataintro*
- Introduction of Big Data - *http://dgit.in/bigdatacs*
- Learning Basics of Big Data in 21 Days - *http://dgit.in/bigdata21days*
- What is Big Data and how does it work? (Video) - *http://dgit.in/bdatawork*
- Big Data Fundamentals (Video) - *http://dgit.in/bigdatafunda*

Intermediate

- Learn Hadoop Free - *http://dgit.in/hadoopfree*
- Install Hadoop on Windows - *http://dgit.in/hadoopwin*
- Big Data Tutorial V4 - *http://dgit.in/bigdatav4*
- Free Hadoop Tutorial: Master BigData - *http://dgit.in/masterbdata*
- DZone Hadoop and Big Data Tutorials - *http://dgit.in/hadoopdzone*
- Big Data and Hadoop Quick Introduction (Video) - *http://dgit.in/bigdataquick*
- Big Data Explained (Video) - *http://dgit.in/bdataexplained*
- Top Six Business Intelligence Podcasts - *http://dgit.in/top6bi*
- Big Data Survey Report - *http://dgit.in/quickbigdata*

Advanced

- Big Data: What it is and Why it matters - *http://dgit.in/bigdatawhatwhy*
- Hadoop - Big Data Tutorial - *http://dgit.in/hadoopjava*
- Developing Big Data Applications with Apache Hadoop - *http://dgit.in/bdatahadoop*
- Amazon Web Services Tutorials for Big Data - *http://dgit.in/awsbdata*
- Oracle Big Data Learning Library - *http://dgit.in/oraclebdata*
- Implementing Big Data Analysis (Video) - *http://dgit.in/bdatanalysis*
- Build Big Data Solutions in Azure (Video) - *http://dgit.in/bigdata-azure*
- Top 9 Big Data Podcasts To Sharpen Your Business Skills - *http://dgit.in/top9bigdata*

Books

- Big Data For Dummies - *http://dgit.in/bdatadummies*
- Big Data, Black Book - *http://dgit.in/bdatablackbook*
- A Revolution That Will Transform How We Live, Work and Think - *http://dgit.in/bdatarevol*
- Principles and Best Practices of Scalable Real-Time Data Systems - *http://dgit.in/bdatascalable*
- Hadoop Explained (Free Kindle ebook) - *http://dgit.in/hadoopkindle*
- Easy Learning: Learn Hadoop MapReduce and Big Data - *http://dgit.in/bdataeasy*
- Data Science, Data Analysis and Predictive Analytics for Business - *http://dgit.in/analyticsdata*
- Hadoop: The Definitive Guide - *http://dgit.in/hadoopguide*
- Big Data Fundamentals: Concepts, Drivers & Techniques - *http://dgit.in/bigdataconcepts*

Courses

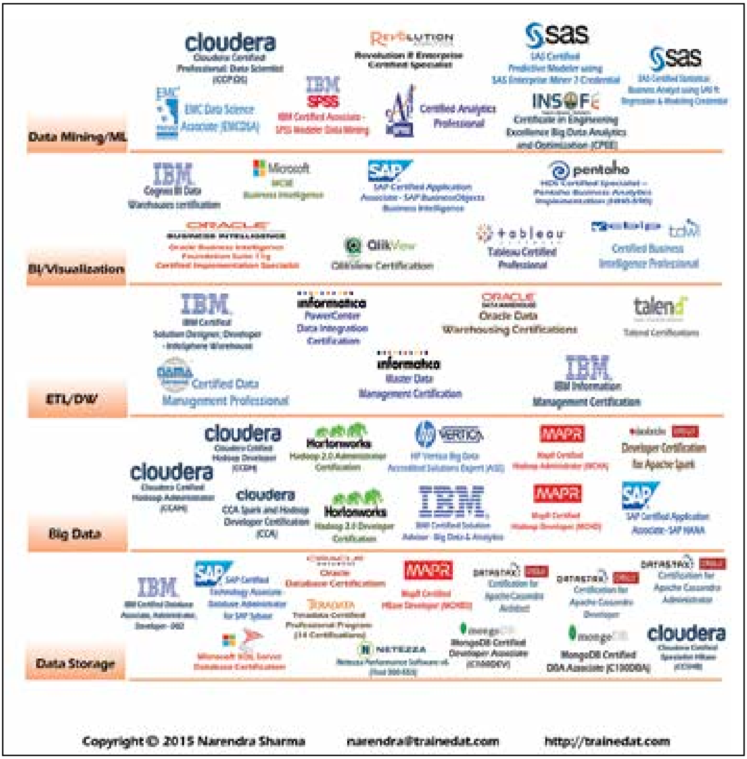
- Simplilearn offers 170 instructor led online training on the various concepts related to Big Data with, on-demand support and industry projects.
  http://dgit.in/simplilearn-bigdata
- Beginners guide to Hadoop and MapReduce is an online course developed by ChalkStreet and authored by Abhishek Roy, an Experienced Big Data trainer. Priced at an affordable Rs. 299, the course spans over 8 sections covering various aspects of Big Data and Hadoop.
  http://dgit.in/chalkstreet
- Pluralsight offers web based courses in a wide range of topics related both to Big Data analytics, Azure, HDFS, Tableau, Apache Spark, SQL, AWS, MongoDB,and NoSQL.
  http://dgit.in/pluralsight-bigdata
- Lynda's Techniques and Concepts of Big Data course by Barton Poulson explores the various aspects of Big Data spread over 8 chapters. Spanning over 2 hours, the course enables you to learn how big data creates an impact on consumers and businesses.
  http://dgit.in/lynda-bigdata
- NIIT Analytics conducts the Professional Certificate in Data Analytics course in select centers across India and covers three emerging topics with a refresher course in Maths. To quality for this course, you need to pass an eligibility test.
  http://dgit.in/niitanalytics

EXTENSIVE TECHNOLOGY BEHIND MEDICAL RECORDS

Nicklaus Children's Hospital based in Miami, USA is in the process of building a comprehensive and sophisticated data warehouse technology that will store, fetch and encrypt data from medical records and devices by using all the possible vital details of every child.

Certifications

Microsoft certifications are widely recognized by employers all over the globe. An MCSE certification on Business Intelligence consists of five exams with special reference to SQL Server but they have a fixed validity period.


cloudera
Cloudera Certified Professional: Data Scientist (CCP:DS)
Revolution R Enterprise Certified Specialist
sas
SAS Certified Predictive Modeler using SAS Enterprise Miner 7 Credential
sas
SAS Certified Statistical Business Analyst using SAS 9: Regression & Modeling Credential
EMC
EMC Data Science Associate (EMCDSA)
IBM SPSS
IBM Certified Associate - SPSS Modeler Data Mining
Certified Analytics Professional
INSOFE
Certificate in Engineering Excellence Big Data Analytics and Optimization (CPEE)
Data Mining/ML
IBM
Cognos BI Data Warehouses certification
Microsoft
MCSE
Business Intelligence
SAP
SAP Certified Application Associate - SAP BusinessObjects Business Intelligence
pentaho
HDS Certified Specialist – Pentaho Business Analytics Implementation
ORACLE
BUSINESS INTELLIGENCE
Oracle Business Intelligence Foundation Suite 11g Certified Implementation Specialist
QlikView
QlikView Certification
tableau
Tableau Certified Professional
tdwi
Certified Business Intelligence Professional
BI/Visualization
IBM
IBM Certified Solution Designer, Developer - InfoSphere Warehouse
informatica
PowerCenter Data Integration Certification
ORACLE
Oracle Data Warehousing Certifications
talend
Talend Certifications
DAMA
Certified Data Management Professional
informatica
Master Data Management Certification
IBM
IBM Information Management Certification
ETL/DW
cloudera
Cloudera Certified Hadoop Developer (CCDH)
Hortonworks
Hadoop 2.0 Administrator Certification
VERTICA
HP Vertica Big Data Accredited Solutions Expert (ASE)
MAPR
MapR Certified Hadoop Administrator (MCHA)
databricks
Developer Certification for Apache Spark
cloudera
Cloudera Certified Hadoop Administrator (CCAH)
cloudera
CCA Spark and Hadoop Developer Certification (CCA)
Hortonworks
Hadoop 2.0 Developer Certification
IBM
IBM Certified Solution Advisor - Big Data & Analytics
MAPR
MapR Certified Hadoop Developer (MCHD)
SAP
SAP Certified Application Associate - SAP HANA
Big Data
IBM
IBM Certified Database Associate, Administrator, Developer - DB2
SAP
SAP Certified Technology Associate - Database Administrator for SAP Sybase
ORACLE
Oracle Database Certification
TERADATA
Teradata Certified Professional Program (14 Certifications)
MAPR
MapR Certified HBase Developer (MCHBD)
DATASTAX
Certification for Apache Cassandra Architect
DATASTAX
Certification for Apache Cassandra Developer
DATASTAX
Certification for Apache Cassandra Administrator
Microsoft SQL Server Database Certification
NETEZZA
Netezza Performance Software v6 (Test 000-553)
mongoDB
MongoDB Certified Developer Associate (C100DEV)
mongoDB
MongoDB Certified DBA Associate (C100DBA)
cloudera
Cloudera Certified Specialist HBase (CCSHB)
Data Storage
Copyright © 2015 Narendra Sharma
narendra@trainedat.com
http://trainedat.com

View the various ways by which you can get certified

Cloudera offers four big data-related certifications aimed at data scientists, which requires you to pass three exams like Descriptive and Inferential Statistics on Big Data, Advanced Analytical Techniques on Big Data and Machine Learning at Scale in eight hours. There are separate certification exams available for Hadoop and HBase as well. The cost per exam is $600, which is higher than Microsoft exams.

The EMC Data Science Associate certification requires you to pass any one of the exams - Big Data Analytics or Backup Recovery Systems. Each exam includes 60 questions with 90 minutes time to complete them. Oracle Business Intelligence Foundation Suite 11g Certified Implementation Specialist certification requires you to pass only one exam with 75 questions in 120 minutes. A minimum of 63% passing score is required to earn the certificate.

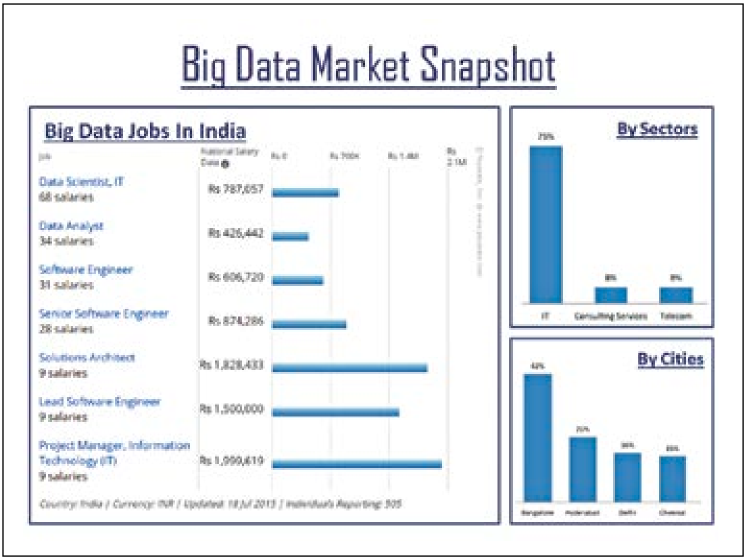
Among all the above mentioned certifications, Cloudera and Oracle are highly demanded by employers as per the data gathered from online job portals.

Career Path

With companies dealing with huge amount of data on a daily basis, the requirement for engineers with knowledge in Analytics has increased with the aim of effectively analyzing the assimilated data. You should try to learn tools such as SAS, SPSS, R, and SQL including the tricks employed by the tools to perform various activities.

As a beginner, you will likely land up in a System Engineer position from where you can slowly climb the ladder to Data Analyst, Product Manager, Research Associate, Lead Modelling Scientist and Data Scientist. You can join a small company to gain real world experience in Big Data Analytics and then move on to a big company, not only for good opportunities but also to earn more.

While the data analyst job fetches an average salary of Rs. 433,672 per year, a data scientist earns an average of Rs. 700,000 per year depending upon the qualification and experience, as per Big Data Analytics salary



See how big you can earn in Big Data field survey conducted by PayScale. According to a salary guide published by Robert Half Technology (RHT), the starting salary for a data architect in US ranges from $111,750 to $153,750 per year, which is a gain of 7.2% compared to previous year.

Conclusion

In addition to the knowledge you gained through courses and certifications, you should continuously make an attempt to update yourself in the emerging field of Big Data via blogs, magazines, whitepapers, books and other resources. Moreover, you should be equipped to learn concepts independently with little assistance to climb the ladder in a corporate environment.

NEWS AND REVIEWS

Comprehensive news, unbiased reviews and ratings to help you make the right purchase

DOWNLOADS

Check out the latest software downloads, games for your Windows, Linux, Mac and PDA/mobiles

All this and more in the world of Technology

VISIT NOW digit.in

RNI REGN. NO. DELENG18986

www.facebook.com

Join 700K+ members of the digit community

http://www.facebook.com/thinkdigit

facebook

digit.in

Digit

Your favourite magazine on your social network. Interact with thousands of fellow Digit readers.

http://www.facebook.com/IThinkGadgets

facebook

IThinkGadgets

I Think Gadgets

An active community for those of you who love mobiles, laptops, cameras and other gadgets. Learn and share more on technology.

http://www.facebook.com/GladtobeaProgrammer

facebook

Glad to be a Programmer

If you enjoy writing code, this community is for you. Be a part and find your way through development.

http://www.facebook.com/devworx.in

facebook

d>

Devworx

devworx, a niche community for software developers in India, is supported by 9.9 Media, publishers of Digit