Data Analytics— Overview and Analysis

Data Analytics is fast emerging as a 'must have' skill for finance professionals, particularly Chartered Accountants. Data when applied to some purpose and adds value for the recipient becomes information. Data Warehouse is an architecture and Big Data is a technology to handle huge data. Big data analytics is based on Data Mining, Statistical Algorithms, Machine Learning and Text Analytics and Natural Language Processing. Ronald van Loon, data scientist, speaker and author says "With more data driving operations in a business than ever before, leaders need to cultivate a culture that is data-driven, instead of believing in their gut instincts." In practical terms, various types of data analytics include Descriptive Analytics, Diagnostic Analytics, Prescriptive Analytics, Exploratory Analytics, Predictive Analytics, Mechanistic Analytics, Causal Analytics and Inferential Analytics. Read on for complete overview and analysis of various aspects of Data Analytics.

CA. Tushankur Saroha

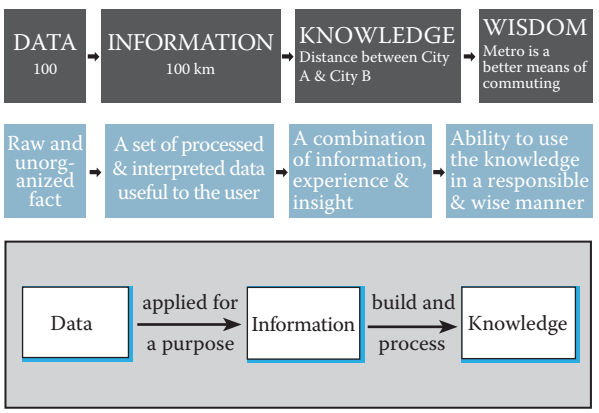
(The author is a member of the ICAI. He can be reached at tushankurs@hotmail.com)

Introduction to Data

How do we know that we KNOW?

- Because majority does that way in our industry
- Because that's what we used to do under the same circumstances
- Because it had delivered good results earlier
- Because our rivals do that way

Usually our arguments are based upon our personal experience and/or gut feelings. What is the scientific and structured way of drawing a conclusion? First, we need to understand few basic building blocks.



The boundaries between the three terms are not always in black & white. What is **data** to one person is **information** to someone else.

A. DATA:

Data when applied to some purpose and adds value for the recipient becomes information. For example, a set of raw cost figures is data. For the Cost Accountant tasked with solving a problem of high costs in one region or deciding the future focus of a cost saving drive, the raw data needs to be processed into a region-wise, head-wise detailed Costs report. It is the Costs report that provides information.

To be useful, data must be:

1. RELEVANT - to the specific purpose
2. COMPLETE – incomplete info is pure poison
3. ACCURATE – garbage-in would lead to garbage-out
4. TIMELY - data that arrives after a decision is made is of no value
5. IN THE RIGHT FORMAT – cleaned and evenly formatted like in excel sheet
6. AVAILABLE AT A SUITABLE PRICE – cost-benefit analysis; the benefits of the data must merit the cost of collecting or buying it.

B. INFORMATION:

It is all about

- GETTING THE RIGHT INFORMATION

A manager investigating poor punctuality of production staff on an assembly line needs information showing all the staff arrival data on that line. Data on other lines is irrelevant, unless late connections elsewhere are causing the problem.

- GETTING THE INFORMATION RIGHT

Just as important, the manager must use the data correctly. One day of shut-down will have a major impact on a week's results. Wrongly interpreting the results could identify a problem where no problem exists.

C. KNOWLEDGE:

- AVAILABLE KNOWLEDGE

For example, knowledge that has been captured and used to develop policies and standard operating procedures (SOP)

- INSTINCTIVE OR HIDDEN KNOWLEDGE

There are certain people in every business set-up who hold specific knowledge or have the 'know how' -- "I did something very similar to that last year and this happened…."

D. WISDOM:

Using the knowledge in a wise and responsible manner is Wisdom – a black box of personally encrypted data.

Basic Types of Data

1. **Sensor** – data from devices having sensors like cameras, GPS, thermometers, etc
2. **User Input** – using keyboard, tap, etc
3. **Interactions** – accessing net banking website through mobile phone
4. **Calculated** – data that is calculated using other data like calculating tax liabilities using revenue data of the company
5. **Metadata** – data about data. For example, an image may include metadata that describes how large the picture is, the colour depth, the image resolution, image created date, etc
6. **Information** – any processed data that has a meaning for its user
7. **Knowledge** – Information created by human thought process and critical analysis

8. **Artificial Knowledge** – Information created by technological systems such as predictions and interpretations

What is Big Data?

The more data in hand, the more chances to get useful insights from it. It means that data processing must manage:

- **High volume** (lots of data)
- **High velocity** (data arriving at high speed)
- **High variety** (many different data sources and formats)

It could be structured like historical bank payment transaction data or it could be semi structured like in XML or it could be totally unstructured like free flow text in social network.

Big data analytics is based on:

- **Data mining:** to sift through data to find patterns and relationships
- **Statistical algorithms:** to build models and predict outcomes
- **Machine learning:** The science of getting a computer to act without programming to handle new data and changes thereon to adapt and enrich models
- **Text analytics and natural language processing (NPL):** to analyse free form text and speech

Common Terminologies Used In Data Analytics

Data Warehouse – is electronic storage of large amount of data collected from varied sources to provide meaningful business insights. It is separate from Transactional databases. It is also known as Decision Support Database or Executive Information System. It has three components:

- Data sources from operational systems such as ERP, CRM, SCM, Excel
- Data Staging Area when data is cleaned and ordered
- Data Access Area where data is warehoused & presented

Example – Airlines use it to analyse route profitability, Retail chains use it for tracking customer buying patterns, Banking uses it to analyse the performance of its product.

Difference between Data Warehousing & Big Data:

Data Warehouse is an architecture and Big Data is a technology to handle huge data.

If an organization wants to know what is going on in its operations or next year planning based on current year performance data etc – it is preferable to choose data warehousing as it needs reliable data.

If organization needs to compare with a lot of big data, which contain valuable information and help them to take a better decision like how to lead more revenue or more profitability or more customers etc, they obviously preferred Big Data approach.

DATA MARTS – are the subsets of Data Warehouse used by specific business groups like HR, Finance, Sales, Inventory, Procurement & Resourcing. They are much smaller than Data Warehouses and usually controlled by a specific department.

BUSINESS INTELLIGENCE (BI) – encompasses a variety of data analysis tools & applications that access the data within Data Warehouse and creates reports & dashboards used in decision making

DATABASE - is generally used to capture and store data from a single source, such as an invoice transactional system. Databases aren't designed to run across very large data sets.

DATA LAKE – is a central storage for all kinds of structured, semi structured or unstructured raw data collected from multiple sources even outside of company's operational systems. Therefore, it is not a good fit for average business analytics but used as a playground by Data Scientists & other data experts as it allows more types of data analytics. It can be used for text searches, machine learning & real-time analytics.

DEEP LEARNING – A subset of machine learning that is automated form of predictive analytics.

ARTIFICIAL INTELLIGENCE (AI) - is the simulation of human intelligence processes by machines. These processes include learning (the acquisition of information and rules for

using the information), reasoning (using rules to reach approximate or definite conclusions) and self-correction. Machine vision and speech recognition uses AI.

DATA DREDGING (DATA FISHING) - is the misuse of data analysis to find patterns in data that appears to be accurate and real when in fact there is no real underlying effect. It is like seeking more information from a data set than it really contains.

DATA SCIENCE – is a combination of three skills: Statistical/Mathematical, Coding & Domain/Business knowledge.

Types of Analytics

Now-a-days, collection of large volume of data is easily possible for any organisation or a person with moderate budget. However, the process of extracting meaning out of collected raw data is known as 'Analytics'. Examples of analytics are:

a) Descriptive Analytics

After analysing the raw data, it helps to provide answers to the questions like –

What happened?

What is happening?

It uses conventional business intelligence and representations like Bar charts, Line graphs, Pie Charts, etc.

Example – assessing a client's credit risk by a bank using past financial performance or getting an insight into the Sales cycle by categorising customers based on their preferences

b) Diagnostic Analytics

It answers – **Why Something Happens?**

Example - A banking website maintains error and access logs that can be used to troubleshoot incidents

c) Prescriptive Analytics

It answers - **What to Do, To Obtain a Given Result?**

It is dedicated to finding the best course of action for a given situation. It provides a "laser-like" focus to answer precise questions.

Example – Google's self-driving car – how to navigate & when to stop

d) Exploratory Analytics

It focusses on identifying general patterns in the raw data to identify outliers and features that might not have been anticipated using other analytical types.

Example – a set of customers purchasing more than 50 products regularly from company's website identified as small retailers using the company's website as a wholesale platform.

e) Predictive Analytics

It answers - **What Will Likely Happen Next?**

It uses data, machine learning (ML) techniques, and statistical algorithms to determine the likelihood of future results based on historical data.

Example - detecting fraud cases, measuring the levels of credit risks & retaining the valuable clients in the banking system.

f) Mechanistic Analytics

Outside of engineering, mechanistic data analysis is extremely challenging and rarely undertaken.

Example - how wing design changes air flow over a wing, leading to decreased drag

g) Causal Analytics

It allows big data scientists to figure out what is likely to happen if one component of the variable is changed.

Example – outcome of antiseptic handwash on infection rate in a village community

h) Inferential Analytics

It is used to draw inferences beyond the immediate data available.

Example – inferring average salary of CA professionals in India based on the salary data available for Delhi based CAs only.

Data Analytics Tools

BASIC TOOLS:

1. MATHEMATICS - numbers
2. EXCEL – vlookup, count, pivot, formulae

3. BASIC SQL – Relational Database Management Systems (RDBMS)
4. WEB DEVELOPMENT – HTML, JavaScript, PHP

ADVANCE TOOLS:

1. HADOOP – open source cloud computing platform allows storage & processing of massive amount of data
2. R PROGRAMMING - open source programming language software that provides data scientists with a variety of features for analysing data
3. PYTHON PROGRAMMING - very powerful, open source and flexible programming language that is easy to learn, use and has powerful libraries for data manipulation, management, and analysis.
4. MATLAB - Its simple syntax is easy to learn and resembles C or C++
5. PERL - Originally developed as a scripting language for UNIX
6. JAVA – is preferred when it comes to the excellent performance of systems on a large scale though may not be appropriate for statistical modelling
7. JULIA - is a new programming language that can fill the gaps with respect to improving visualizations and libraries for data analysis

How Does Basic Data Analytics Work?

a) Defining the Problem

"The single most critical principal I apply when analysing data is a rule my high school math professor taught me at age 14: 'Don't write the first line of code until you can describe in plain English the problem you are attempting to solve!' Simply put, if you can't explain in plain English the business problem you are setting out to address, no amount of data analytics is ever going to solve it." - ***Dez Blanchfield, investor and data scientist***

Validate the problem you have identified:

1. Could it be a symptom of a bigger problem?
2. Or is it a freak one-off instance such as simple reporting error?

This preliminary assessment answers two questions:

3. **Is this actually a problem?** And if yes
4. **What is the core problem here?**

b) Identifying Potential Causes

1. LOOK FOR QUICK WINS – like 'turn your device off and back on'
   - Look for any obvious cause(s)
   - Double check the source showing the problem
   - Any abnormal causes that immediately come to mind
   - NO SUCCESS, then go to next step
2. ASK AROUND
   - Does this problem impact other teams too? Do they have any insight into the possible causes?
   - Gather any insight and move to next step
3. CREATE HYPOTHESES (POSSIBLE CAUSES)

"A fact is a simple statement that everyone believes. It is innocent, unless found guilty. A hypothesis is a novel suggestion that no one wants to believe. It is guilty, until found effective." - ***Edward Teller, Hungarian-American theoretical physicist***

- It is an educated guess that has not been confirmed yet.
- Think of several assumptions about the cause of the problem and then jot down how you might prove/disprove each one before analysing the data
- This helps prevent common mistakes like 'data dredging' or 'cherry picking'
- Sometimes looking at the data gives rise to a new hypothesis that needs testing

4. TECHNICALLY SPEAKING
   - It is a formal approach to formulate the hypotheses
   - Finding the cause (independent variable that can be changed or controlled)
   - Finding the effect (dependent variable as outcome)
   - Hypothesis can be proven wrong

"There are two possible outcomes: if the result confirms the hypothesis, then you've made a discovery. If the result is contrary to the hypothesis, then you've made a discovery." - ***Enrico Fermi, Italian-American physicist and the creator of the world's first nuclear reactor***

c) How to Analyse Data (What Does the Data Say?)

1. Determine And Segment Relevant Data

- Based on the hypotheses what data should be looked at?
- What metrics will help to prove or disprove the possible cause?
- Isolate different pieces of data that may be causing the problem to easily spot trends or anomalies

2. Eyeball The Data

- Based on the knowledge and common sense, take a note is there any aspect of the data that appears abnormal?
- If new to the role or company – use historical data to baseline the 'normal' and then check for any abnormality in the data

3. Assess The Impact

"One finds the truth by making a hypothesis and comparing observations with the hypothesis." - **David Douglass, American Physicist**

- Are the results statistically significant?
- If there is any abnormality or anomaly spotted in Step 2 above, is it significant enough to explain the problem
- Statistical significance does not mean 'important' but 'accurate' & 'verifiable'
- Are the anomalies or trends spotted 'practically significant'?

"We often worry about whether our sample size is large enough to provide reliable results when we find statistical significance, but we also need to take into consideration whether these differences are meaningful in a real way." *- Jennifer Shin, Founder at 8 Path Solutions and Faculty at UC Berkeley*

d) Avoiding Data Fallacies

"A common fallacy is assuming a dataset is trustworthy - until it's discovered later in analysis that it's not. Flip that. Make sure your data is trustworthy before you begin analysis." - *Tamara Dull, Director of Emerging Tech at SAS Institute*

1. Keep Your Data In Context

- Avoid cherry picking. Data can sometimes play tricks on us. Cherry picking is selecting only the bits of data that support your claim while discarding the parts that don't
- Ignore personal bias and motives while analysing the data

  "It is absolutely essential that one should be neutral and not fall in love with the hypothesis." - *David Douglass, American Physicist*

2. Start With A Hypothesis

- Avoid data dredging
- While looking for a cause of a problem, it might be tempting to dig through the data until a pattern emerges
- This pattern could be a 'false positive'
- Always begin with a hypothesis before analysing the data, check related metrics and test to see if the trend continues

ePass
2003 Series Tokens

www.charteredinfo.com

Thank You Once again...

ePass User-base has now crossed 90,00,000 Users,

THANK YOU

everyone for all support and trust. Ask ePass..!!

Get your existing ePass tokens updates to New CCA Guideline for more details log on to **update.epasstokens.com**

www.signer.digital

signer digital

Simplify Digital Document Signing

ISO 27001

For more details & DEMO call us:

Delhi	Mumbai	Bangalore	Ahmedabad	Kolkata	Nagpur
011-45037177	022-65228288	080-40921639	079-40083529	033-40078356	0712-6638888

"If you torture the data long enough, it will confess to anything." - *Ronald Coase, Nobel prize winning economist*

3. **Correlation Or Causation**
   - Avoid False causality
   - It is easy to assume that because two events happen at the same time, one has caused the other
   - But sometimes patterns that seem correlated may be correlated to a third factor and not to each other

e) Solve Problems, Make Smart Decisions

"Shallow men believe in luck or in circumstance. Strong men believe in cause and effect." - *Ralph Waldo Emerson, American essayist, lecturer, and poet*

- Once data inquiry is over and cause of the problem is ascertained, it is a good idea to update other teams that were involved or impacted
- Write a summary for your future reference and for others
  i. WHAT WAS THE PROBLEM?
  ii. WHAT THE DATA SHOWED?
  iii. WHAT WAS THE RESULTING ACTION?

"Whether one is fighting the war against cancer, or the battle to stop infectious diseases, or the cybersecurity war, or the battle to win the hearts of your customers, or pricing wars within competitive marketplaces, or engaged in other battles, the winners will most certainly be those who use data effectively to make smart business decisions – those who truly appreciate that knowledge is power." - *Kirk D. Borne, PhD, Astrophysicist, Principal Data Scientist at Booz Allen*

Data Analytics Work Flow in Action

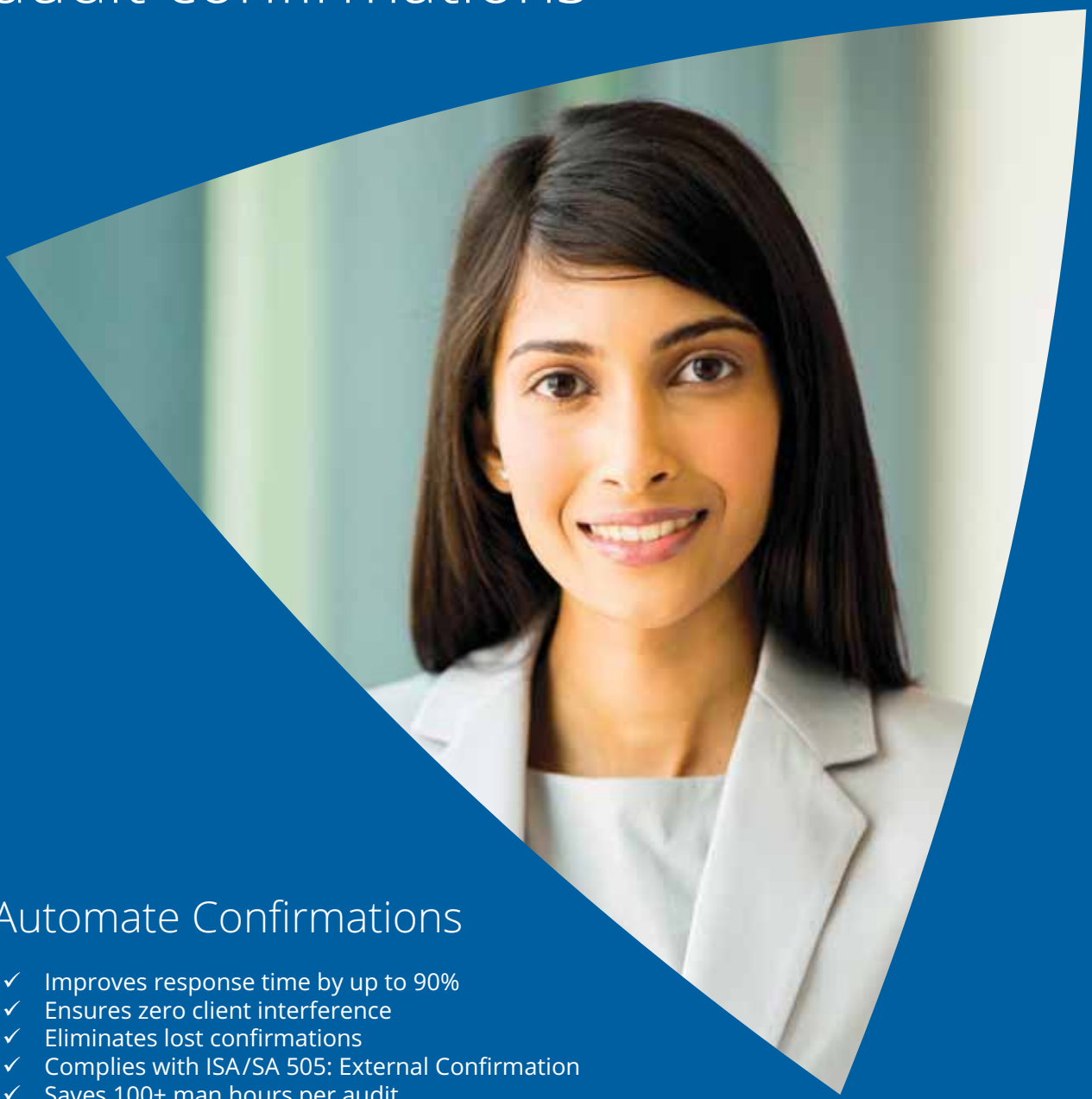
"With more data driving operations in a business than ever before, leaders need to cultivate a culture that is data-driven, instead of believing in their gut instincts." - *Ronald van Loon, data scientist, speaker, author, and founder.*

E-Commerce

1. **Defining The Problem:** On an on-line shopping website, more potential customers are abandoning their shopping carts. How average abandonment rate can be decreased?
2. **Identifying Potential Causes (Hypotheses):** The average cart abandonment rate has increased because of:
   i. an increase in the absolute number of people beginning a cart (have started placing items in the cart).
   ii. a recent change on a section of the checkout process.
   iii. seasonality (i.e. holidays, school breaks, etc.).
   iv. an end to a promotion which results in more people abandoning their carts.
   v. a specific product.
3. **Analyse Data:**
   i. Check if absolute numbers of customers visiting the site have changed?
   ii. Has the abandonment rate gone up because more people are starting a cart?
   iii. Are the absolute numbers of customers visiting the site unchanged?
   iv. Ask around if there have been any promotions? Any new product launched? Any seasonal impact? Anything change in the checkout process? Any price changes?
4. **Avoid data fallacies (comparing observations with hypotheses):**
   i. Number of people beginning a cart has gone up slightly but number of people completing the purchase remain the same. It shows there is something that causes less conversion rate in additional shoppers.
   ii. There has been a small change in the checkout flow. Now item image is also shown in the cart. This change coincides with the time when the rate of cart abandonment increased
   iii. Splitting the checkout process into different steps reveals that cart dropping is not on the page that has recently changed, and more customers are moving to the next page

Confirmation.com

One platform for all your audit confirmations



Automate Confirmations

- ✓ Improves response time by up to 90%
- ✓ Ensures zero client interference
- ✓ Eliminates lost confirmations
- ✓ Complies with ISA/SA 505: External Confirmation
- ✓ Saves 100+ man hours per audit

Over 100,000 professionals worldwide use Confirmation.com to perform their audit confirmations online rather than dealing with the hassle of sending them by mail. Work smarter with Confirmation.com.

World's largest network of online audit confirmations.

700,000 companies	100,000 professionals	3,000 law firms	3,500 banks	160+ countries	all use Confirmation.com

To learn more, contact Harsh Jogani at **harsh.jogani@confirmation.com** or call **+91 97735 85311**.

iv. To check the seasonal impact, current week's abandonment rate is compared to the same week in previous years. After a quick look at the calendar, sessions and email open rates it is concluded the seasonality is not the cause

v. It is noticed a promotion has recently ended. Usually, people abandon the cart once promotion is over. Proportion of mid-way checkout drops using the promotion code is one fourth of the rate of cart abandonment, so this is not a major contributing factor.

vi. Could it be due to inventory? Abandonment rate is broken down by product categories, but all products have similar abandonment rate.

vii. Another look at the checkout process shows that biggest drop off are on the page where Shipping charges are shown first time. It is observed that few cosmetic changes were done on the page around the same time when drop rate also increased.

viii. So new hypothesis is emerged – that potential customers are abandoning their carts because they're frustrated by the price. Their expectations are set at a lower price based on the product pages. As soon as they see the full price (including shipping), they're more likely to drop off.

5. **Solve Problem & Make Smart Decisions**

   i. It is time to validate the results. Product page can be reverted to old design or shipping price is added to the product price and impact on the abandonment rate can be tracked and validated.

Customer Support (CRM)

CRM Manager noticed that response time to the service requests/support ticket has increased and are too long.

1. **Defining The Problem:** Customer support ticket response time is too long. How can we reduce response time?
2. **Identifying Potential Causes (Hypotheses):** Ticket response time has increased because of
   i. an influx of service-related tickets which have longer response times compared to product-related tickets.
   ii. challenges specific to one call centre.
   iii. insufficient customer support team members which results in a backlog of tickets.
   iv. a recently launched product feature.
3. **Analyse Data:**
   i. Is there any shift in service related ticket compared to product related tickets?
   ii. Did it happen at the same time when response time has increased?
   iii. Do Service tickets take longer to resolve?
   iv. Does the magnitude of the impact match the core problem?
4. **Avoid data fallacies (comparing observations with hypotheses):**
   i. It is observed that service tickets take 15 minutes more than the product tickets.
   ii. Ratio between support & product related tickets remains the same as in previous month at 60:40.
   iii. Exploring the speculative possible cause

MPROFIT

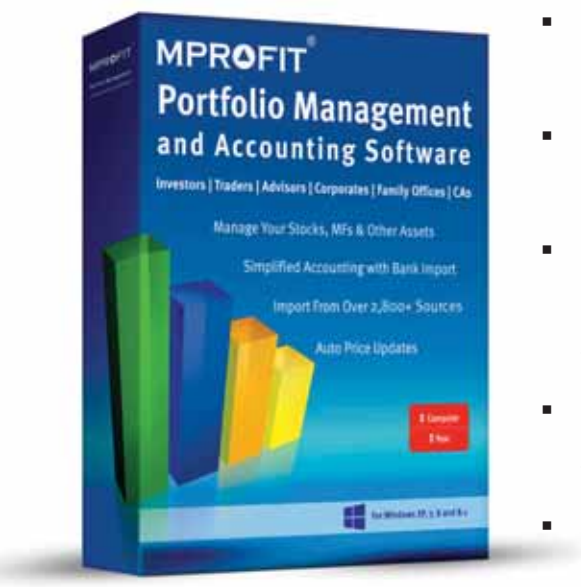
Portfolio Management and Accounting Software

Capital Gains with Grandfathering Simplified

Featured in Outlook MONEY

Featured on ET NOW

Over 150,000+ Downloads



- Manage Stocks, MFs, Bonds, FDs, Insurance, ULIPs, F&O and Other Assets
- Import contract notes, mutual fund and bank statements from over 3,000+ sources
- Capital gains reports for Stocks, Bonds, Equity & Debt MFs (with or without indexation)
- Detailed long-term capital gains reports for stocks & equity mutual funds with PP, FMV and CA as per grandfathering clause
- Simplified Accounting with Bank Import and auto generated vouchers from Portfolio Management
- Other reports such as Annualised Returns (XIRR), Income Reports, Trial Balance, P&L and Balance Sheet
- Auto price updates and historical prices

Available Online @ Paytm amazon.in flipkart.com

MProfit Software Pvt. Ltd. | Nariman Point, Mumbai
022-4002-4149 | www.mprofit.in

– is the long response time related to a specific geographic zone?

iv. By splitting the data into different regions, it is observed that average response time in the zones are: East zone - 60 mins, South zone – 20 mins, West zone – 15 mins & North zone – 25 mins.

v. By eyeballing the data, it is observed that East zone is taking 4x times more than other zones

vi. Talking to the East zone Manager revealed that team is struggling to connect their phones with new software application that is causing the backlog.

5. **Solve Problem & Make Smart Decisions**

i. It is time to validate the results. After resolving the technical snag, data is interpreted again, and it is found that response time now within the defined SLA.

Case Study

Data Analytics Use Case in Banking and Financial Services

1. Fraud Detection

Banks and financial services firms use analytics to differentiate fraudulent interactions from legitimate business transactions. By applying analytics and machine learning, they can define normal activity based on a customer's history and distinguish it from unusual behaviour indicating fraud. The analysis systems suggest immediate actions, such as blocking irregular transactions, which stops fraud before it occurs and improves profitability.

2. Compliance and Regulatory Requirements

Financial services firms operate under a heavy regulatory framework, which requires significant levels of monitoring and reporting and requires deal monitoring and documentation of the details of every trade. This data is used for trade surveillance that recognizes abnormal trading patterns.

3. Customer Segmentation

Banks have been under pressure to change from product-centric to customer-centric businesses. One way to achieve that transformation is to better understand their customers through segmentation. Big data analytics enable them to group customers into distinct segments, which are defined by data sets that may include customer demographics, daily transactions, interactions with online and telephone customer service systems, and external data, such as the value of their homes. Promotions and marketing campaigns are then targeted to customers according to their segments.

4. Personalised Marketing

One step beyond segment-based marketing is personalized marketing, which targets customers based on understanding of their individual buying habits. While it's supported by big data analysis of merchant records, financial services firms can also incorporate unstructured data from their customers' social media profiles to create a fuller picture of the customers' needs through customer sentiment analysis. Once those needs are understood, big data analysis can create a credit risk assessment to decide whether to go ahead with a transaction.

5. Risk Management

While every business needs to engage in risk management, the need may be largest in the financial industry. Regulatory schemes such as Basel III require firms to manage their market liquidity risk through stress testing. Financial firms also manage their customer risk through analysis of complete customer portfolios. The risks of algorithmic trading are managed through back testing strategies against historical data. Big data analysis can also support real-time alerting if a risk threshold is surpassed.

More than 25% of financial firms have already implemented big data analytics projects and are already obtaining a competitive advantage. Due to both regulatory requirements and the perceived value of big data analytics, financial firms will continue to implement big data analytics projects. This will require increased investments in data centre technology as well as increased hiring of staff with big data skills. ■