# Technology Digest

Bulletin of telecom technology

**Issue, March 2018**

**Telecom Regulatory Authority of India**


## Data Analytics

In this issue



Data analytics is the process of inspecting, cleansing, transforming, and modeling data with the objectives to discover useful information, suggest conclusions, and support decision-making. Data Analytics involves applying an algorithmic or mechanical process to derive insights. For example, running through several data sets to look for meaningful correlations between each other.

Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make informed business decisions and by researchers and academicians to verify or disprove scientific models, theories, and hypotheses. In advanced analytics projects, much of the required work takes place in collecting, integrating and preparing data and then developing, testing and revising analytical models to produce reliable results. Evolving data may facilitate thorough decision-making. For example, data collected by social networking websites related to user preferences, community interests, and segments can be analysed according to specified criteria such as demographics, age or gender, to reveal key user and customer trends and facilitate the social network's alignment of content, layout and overall strategy. A similar sort of thing happened in the case of Cambridge Analytica's Facebook data analysis and political advertising. Similarly, Google Analytics provides a tool to analyze data from all touchpoints in one place, for a deeper understanding of the customer experience of websites and mobile applications.

**Background**

Before computers, the 1880 Census in the US took over 7 years to process the collected data and finalize the report. With the invention of Von Neumann architecture, the data had been regarded and processed as data to be processed for data analysis. The appearance of RDB (Relational Database), in the 1980s proved to be a turning point, which allowed users to write Structured Query Language (SQL) to retrieve data from databases. Due to ever decreasing the cost of hard disk drives, there's significant increase in the amount of data, that's when William H. Inmon proposed a “data warehouse”, which is a system optimized for reporting and data analysis. Howard Dresner at Gartner, in 1989, proposed the term “BI (Business Intelligence)”. BI supports better business decision making through searching, collecting and analyzing accumulated data in business. Around the 1990s, another term “data mining” was coined which is the computational process to discover patterns in large datasets. The next substantial change was the internet. For the demand of searching a website on the web, Larry Page and Sergey Brin developed the Google search engine which processes and analyses big data in distributed computers. In the early 2010s, Amazon Redshift, which is a cloud-based data warehouse, and Google BigQuery, which processes a query in thousands of Google servers, were released.

For 10 years the prevailing trend in business intelligence (BI) and data analytics has been the move toward self-service. In 2018 and beyond, we'll see a growing list of what many call "smart" capabilities powered by machine learning (ML) and artificial intelligence (AI).

**Types of Data Analytics**

There are 4 types of analytics:

- **Descriptive Analytics**

  Descriptive analytics organizes raw data from multiple data sources to give valuable insights into the past events which simply signifies that something is wrong or right, without explaining why. Descriptive analytics is leveraged when the need is to understand the overall performance at an aggregate level and describe the various aspects. These are based on aggregate functions and summarize certain groupings based on simple counts of some events. An example of descriptive analytics is the analytics derived from organizations from the web server through Google Analytics tools.

- **Diagnostic Analytics**

  Organizations use diagnostic analytics using historical data that can be measured against other data to answer the question of why something happened. This type of analytics provides a possibility to drill down to find out dependencies and identify patterns. It basically provides a very good understanding of a limited piece of the problem to be solved and gives actionable insights. For example, diagnostic analytics can be used to assess the effectiveness of a social media marketing campaign to see what worked in the past campaigns what didn't.


Figure 1: Types of Analytics

- **Predictive Analytics**

  Predictive analytics provide estimates about the likelihood of a future outcome which is fundamentally based on probabilistic models. Prediction is just an estimate, the accuracy of which highly depends on data quality and stability of the situation, so careful treatment and continuous optimizations are required for this type of analytics. Supervised / unsupervised learning algorithms can be used for to train on a data set to make predictions or take actions. A set of such algorithms are classified as machine learning. Deep learning is a subset of Machine learning which is based on neural networks, a conceptual model of the brain. Different techniques involved in such analysis are Naive Bayes, SVM (Support Vector Machine), Neural networks, association rules, decision trees, logistic regression, etc.

- **Prescriptive Analytics**

  Prescriptive analytics attempts to quantify the effect of future decisions to advise on possible outcomes before the decisions are taken and recommend one or more possible courses of action. Prescriptive analytics uses a combination of techniques and tools such as business rules, algorithms, machine learning and computational modeling procedures. Highly sophisticated algorithms such as neural networks are also typically in the realm of prescriptive analytics as they are focused on making a specific prediction. These techniques are applied against input from many different data sets including historical and transactional data, real-time data feeds, and big data.

**Data Analytics – Process**

Data analytics process consist of following iterative phases:

- **Data requirements specification**

  The very first step, before data acquisition and data analytics, is the identification of goals, metrics, and lever which helps to set clear measurement priorities of the project and avoids meaningless data analysis. For example, the goal can be improving customer retention, one of the metrics can be the percent of renewed subscriptions, and the business levers can be the design of the renewal page, timing, and content of reminder emails and special promotions.

- **Data collection**

  After the requirement specification and setting measurement priorities, data are collected from diverse sources which enable better correlations, building better models and identify more actionable insights. There may be different methods that can be used to collect, e.g., web-based surveys, hand-held devices such as Personal Digital Assistants (PDAs), social networking sites, etc.

- **Data processing**

  This step involves pre-processing and organizing data in proper format on which analytics can be performed. For example, data collected can be organized in relational model in a database or other statistical software, on which further analytics can be performed.

- **Data cleaning**

  After pre-processing and organizing data, one of the most critical steps in the data value chain is to improve the quality of data because even with the best analysis, junk data may generate incorrect results and mislead decision making of the organizations. It involves spelling corrections, handling of missing data, weed out garbage or irrelevant data.

- **Data analysis**

  Once data is cleaned, it can be analyzed in several ways depending on the objective and type of analytics needed. Depending on the problem at hand, it may be required to make certain calculations using statistical formulae like mean, variance, standard deviation, correlation, ANOVA (ANalysis Of VAriance), ANCOVA (ANalysis of COVAriance) etc.

- **Communication**

  After the data analysis, analysts have to communicate the results to the users of the analysis to support their requirement. Results may be communicated effectively and efficiently using data visualization techniques. Data visualization displays information to communicate key messages contained in the data in the form of tables and charts.

**Statistical methods and coefficients for analytics:**

Some popular summary statistics for quantitative variables are as follows:

**Time-series analysis:** Time series analysis is best suited if a single variable is captured over a period, e.g., GDP growth rate over a period of 5 years.

**Ranking:** Different categorical subdivisions of a variable are ranked in ascending or descending order, e.g., ranking of countries based on per capita income.

**Part-to-whole:** Categorical subdivisions are measured as a ratio to the whole, e.g., market share of different service providers in the telecom market.

**Deviation:** Categorical subdivisions are compared against a reference or fixed value, e.g., the standard deviation of a typical data point about mean.

**Frequency distribution:** Number of observations of a variable in a given interval, e.g., census focused on the demography of a country may use frequency distribution of a number of people in different age groups (0-10, 10-20, 20-30, and so on).

**Correlation:** Comparison between observations represented by two variables (X, Y) to determine how strong relationship is between them. Correlation coefficients usually range from -1 (total negative relationship) to +1 (total positive relationship). 0 means no relationship between variables.

**Nominal comparison:** Arbitrary comparison of categorical subdivisions, e.g., comparison of sales volume by product code.

**Geospatial:** Comparison of variables across a geographic area or map layout, e.g., representation of population density over a geographic map of countries.

**Challenges to Effective Analytics**

1. **Dealing with data growth**

   Dealing with the volume of data being produced and the speed at which it is being produced is a challenge. Additionally, it is also a challenge to manage the enormous number of sources that are producing this data. A significant amount of data is unstructured which doesn't reside in a database, e.g., documents, photos, audio, videos, etc., which can be difficult to search and analyze. To collect and store a large set of unstructured data in native format, options like data warehouses are used.

2. **Real-time integration**

   The addition of new data to the existing dataset is also an issue. To drive decision-making process by analytics, it must be real-time or near real-time, especially in banking, healthcare, and other automated systems. To achieve that speed, organizations are looking to a new generation of ETL (Extract, Transform, Load) and analytics tools that dramatically reduce the time it takes to generate reports, preferentially, with real-time analytics capabilities to respond, immediately, to developments in the marketplace.

3. **Integrating disparate data sources**

   The variables associated with data leads to challenges in data integration. Data may come from diverse sources — enterprise applications, social media streams, email systems, employee-created documents, etc. The problem arises when a data lakes/ warehouse tries to combine inconsistent data from disparate sources, it encounters errors. Inconsistency, redundancy, logic conflicts, and missing data may all result in data quality challenges and, hence, impact the quality of analysis.

4. **Security and privacy of data**

   The tools used for data analytics, stores, manages, analyses, and utilizes the data from a different variety of sources which may lead to a risk of data exposure. Security and privacy concerns may increase with the production of more and more data, making it highly vulnerable. Hence, it is essential for analysts and data scientists to consider such issues and data should be processed in a manner to respect the privacy of concerned users.

**Data Analytics Tools**

There are many analytics tools for data analysis that are easy to use and have powerful capabilities which help to manage and interpret data in a better and effective manner. Here are some of the analytics tools:

1. **Tableau**

   It's a simple and intuitive tool which is exceptionally powerful in data visualization. It is very helpful for creating interactive data visualization with little/no programming skills required. There are five ways to access their products: Desktop (both professional and personal editions), Server, online (which scales to support thousands of users), Reader, and Public, with the last two free to use.

2. **R-Programming**

   It's an open-source programming language and development environment for statistical computation and graphical visualization. In addition to statistical and graphical techniques, it also provides time-series analysis, classification, clustering among others. It may also be used to develop statistical software applications.

3. **Pentaho**

   It simplifies the preparation and blend of data and provides a range of tools for data analysis, visualization, exploration, reporting, and prediction.

4. **Python**

   It's an open source scripting language, known for its simplicity, that supports libraries (numpy, scipy, and matplotlib) and function for distinct types of statistical operations. With the introduction of packages like TensorFlow and Keras, python also provides support for deep learning.

**Data Analytics in TRAI**

TRAI's mission is to create and nurture conditions for growth of telecommunications in the country in a manner and at a pace which will enable India to play a leading role in emerging global information society. One of the main objectives of TRAI is to provide a fair and transparent policy environment which promotes a level playing field and facilitates fair competition.

TRAI collects data from various sources related to the state of the telecom in India, e.g., through Performance Monitoring Reports (PMRs), through IDTs, through crowdsourced applications, etc. These data need to be analyzed properly to retrieve meaningful insights and formulation of regulations and other policies, thereafter. At present, for analysis of data, mostly, Tableau and MS Excel are being used in TRAI.

Data collected from Telecom Service Providers (TSPs) for various Key Performance Indicators (KPIs) related to Quality of Service (QoS) of the telecom operators, for example, have been analysed extensively, to obtain adequate metrics and their threshold for measurement of Drop Call Rate (DCR), before the formulation of the "The Standards of Quality of Service of Basic Telephone Service (Wireline) and Cellular Mobile Telephone Service (Fifth Amendment) Regulations, 2017".

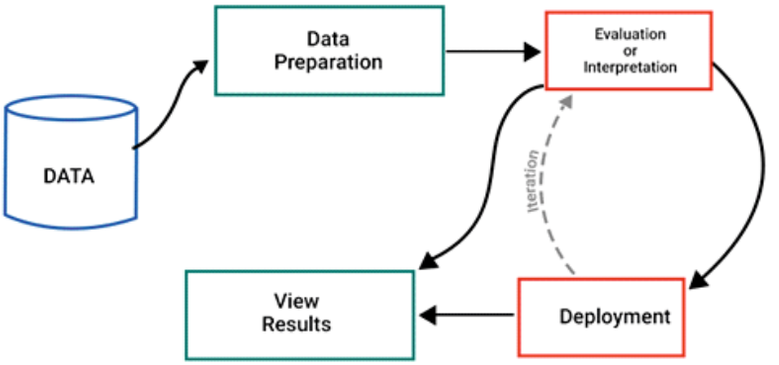


Figure 2: Workflow for Data Analytics in TRAI

Similarly, crowdsourced data collected through TRAI MySpeed and TRAI MyCall mobile applications are processed, filtered and analyzed to get insight into the state of the telecom in India vis-à-vis data speeds experienced by subscribers and customer perceived the quality of calls, respectively. Usually, bottom-up (data-driven) approach is used for the analytics in TRAI and the kind of analytics done is, mostly, descriptive and diagnostic in nature.

## Conclusion

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have provided a great boon to various companies and organizations, as it is helping them make better decisions, thus profiting the company. The convergence of these trends provides capabilities required to analyze astonishing data sets quickly and cost-effectively. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability.

## References


[1] Claus Weihs, Katja Ickstadt, Data Science: the impact of statistics, International Journal of Data Science and Analytics, January 2018.

[2] Anton Wirsch, Analysis of a Top-Down Bottom-Up Data Analysis Framework and Software Architecture Design, Working Paper, Composite Information Systems Laboratory (CISL), MIT, May 2014.

[3] The Four Realms of Analytics. (2015, June 04). Retrieved from http://www.vlamis.com/blog/2015/6/4/the-four-realms-of-analytics.html

[4] Types of Analytics: descriptive, predictive, prescriptive analytics. (2016, February 08). Retrieved from https://www.dezyre.com/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209

[5] Data Analysis (2018, March 27). Retrieved from https://en.wikipedia.org/wiki/Data_analysis

[6] Haitham Badi, Mohammad Fadhel, Sana Sabry, Mohamed Jasem, A survey of human-computer interaction technologies and techniques, Journal of Data Science and Analytics, August 2016.

[7] George Steinbuss, Klemens Bohm, Hiding outliers in high-dimensional data spaces, Journal of Data Science and Analytics, September 2017.

[8] Nicoletta Di Blas, Mirjana Mazuran, Paolo Paolini, Elisa Quintarelli, Letizia Tanca, Exploratory Computing: a comprehensive approach to data sensemaking, Journal of Data Science and Analytics, December 2016.

[9] Hendrik Blockeel, Declarative data analysis, Journal of Data Science and Analytics, part of Springer Nature 2017, November 2017.

[10] Christophe Ley, Stephane P. A. Bordas, What makes Data Science different? A discussion involving Statistics 2.0 and Computational Sciences, International Journal of Data Science and Analytics, December 2017.

[11] Claus Welhs, Katja Ickstadt, Data Science: the impact of statistics, International Journal of Data Science and Analytics, January 2018.

[12] Arno Slebes, Data science as a language: challenges for computer science – a position paper, International Journal of Data Science and Analytics, January 2018.

[13] Sujing Wang, Christophe F. Eick, A data mining framework for environmental and geo-spatial data analysis, International Journal of Data Science and Analytics, September 2017.

[14] Ravindra Khattree, Manoj Bahuguna, An alternative data analytic approach to measure the univariate and multivariate skewness, International Journal of Data Science and Analytics, February 2018.


---


Published by: Telecom Regulatory Authority of India
Editorial responsibility: TD Division, TRAI
Contributions, comments and suggestion: sroit@trai.gov.in