# The Basics of Data Analytics

For you to become a professional data scientist, working in data mining and business intelligence firms you have to understand the fundamentals of data analytics. The goal of this article is to help you firm up all the key concepts in data analytics.

By the end of the article, you should be in a position to describe different types of analytics, common terminologies used in analytics, tools and basic pre-requisites for analytics and the workflow of data analytics. Without further ado, let’s dive in to explore the basics of data analytics.

**Types of analytics**

Raw data is not any different from crude oil. These days, any person or institution with a moderate budget can collect large volumes of raw data. But the collection in itself shouldn’t be the end goal. Organizations that can extract meanings from the collected raw data are the ones that can compete in today’s complex and unpredictable environment.

At the core of any data refinement process sits what is commonly referred to as “analytics”. But different people use the word “analytics” to imply different things. If you’re in marketing and would like to understand data analytics, you should understand the different types of analytics. Below are examples of analytics:

· Descriptive

· Diagnostic

· Prescriptive

· Exploratory

· Predictive

· Mechanistic

· Casual

· Inferential

Let’s dive in to explore more about these analytics.

#1: Descriptive analytics

The main focus of descriptive analytics is to summarize what happened in an organization. Descriptive Analytics examines the raw data or content — that is manually performed — to answer questions such as:

· What happened?

· What is happening?

Descriptive analytics is characterized by conventional business intelligence and visualizations such as the bar charts, pie charts, line graphs, or the generated narratives. A simple illustration of descriptive analytics can be assessing credit risk in a bank. In such a case, past financial performance can be done to predict client’s likely financial performance. Descriptive analytics is useful in providing insights into sales cycle such as categorizing customers based on their preferences.

#2: Diagnostic analytics

As the name suggests, diagnostic analytics is used to unearth or to determine why something happened. For example, if you’re conducting a social media marketing campaign, you may be interested in assessing the number of likes, reviews, mentions, followers or fans. Diagnostic analytics can help you distill thousands of mentions into a single view so that you can make progress with your campaign.

#3: Prescriptive analytics

While most data analytics provides general insights on the subject, prescriptive analytics gives you with a “laser-like” focus to answer precise questions. For instance, in the healthcare industry, you can use prescriptive analytics to manage the patient population by measuring the number of patients who are clinically obese.

Prescriptive analytics can allow you to add filters in obesity such as obesity with diabetes and cholesterol levels to find out areas where treatment should be focused.

#4: Exploratory analytics

Exploratory analytics is an analytical approach that primarily focuses on identifying general patterns in the raw data to identify outliers and features that might not have been anticipated using other analytical types. For you to use this approach, you have to understand where the outliers are occurring and how other environmental variables are related to making informed decisions.

For example, in biological monitoring of data, sites can be affected by several stressors, therefore, stressor correlations are vital before you attempt to relate the stressor variables and biological response variables. The scatterplots and correlation coefficients can provide you with insightful information on the relationships between the variables.

However, when analysing different variables, the basic methods of multivariate visualization are necessary to provide greater insights.

#5: Predictive analytics

Predictive analytics is the use of data, machine learning techniques, and statistical algorithms to determine the likelihood of future results based on historical data. The primary goal of predictive analytics is to help you go beyond just what has happened and provide the best possible assessment of what is likely to happen in future.

Predictive models use recognizable results to create a model that can predict values for different type of data or even new data. Modeling of the results is significant because it provides predictions that represent the likelihood of the target variable — such as revenue — based on the estimated significance from a set of input variables. Classification and regression models are the most popular models used in predictive analytics.

Predictive analytics can be used in banking systems to detect fraud cases, measure the levels of credit risks, and maximize the cross-sell and up-sell opportunities in an organization. This helps to retain valuable clients to your business.

#6: Mechanistic analytics

As the name suggests, mechanistic analytics allow big data scientists to understand clear alterations in procedures or even variables which can result in changing of variables. The results of mechanistic analytics are determined by equations in engineering and physical sciences. Also, they allow data scientists to determine the parameters if they know the equation.

#7: Causal analytics

Causal analytics allow big data scientists to figure out what is likely to happen if one component of the variable is changed. When you use this approach, you should rely on a number of random variables to determine what’s likely to happen next even though you can use non-random studies to infer from causations. This approach to analytics is appropriate if you’re dealing with large volumes of data.

#8: Inferential analytics

This approach to analytics takes different theories on the world into account to determine the certain aspects of the large population. When you use inferential analytics, you’ll be required to take a smaller sample of information from the population and use that as a basis to infer parameters about the larger population.

Common terminologies used in data analytics

As you plan to begin using data analytics for the achievement of your bottom line, there are terminologies that you must learn. Below is a list of common terminologies and their meanings:

The severing of links between people in a database and their records to prevent the discovery of the source of the records.

· Business Intelligence (BI). Developing intelligent applications that are capable of extracting data from both the internal and external environment to help executives make strategic decisions in an organization.

· Automatic identification and capture (AIDC). It is any method that can automatically identify and collect data on items, and store them in a computer system.

· Avro. It is a data serialization system that facilitates encoding of a database schema in Hadoop.

· Behavioral analytics. It involves using data about people’s behavior to infer their intent and predict their future actions.

· Big Data Scientist. A professional who can develop algorithms that make sense from big data.

· Cascading. It is used in Hadoop to explain the concept of providing a higher level of abstraction. This allows developers to create complex jobs using different programming languages in the JVM.

· Cassandra. Cassandra is an open source and distributed database system developed by Facebook that is designed to deal with large volumes of data.

· Classification analysis. It is a systematic process of obtaining crucial and relevant information about raw data and its metadata.

· Database. A digital collection of logically related and shared data.

· Database administrator (DBA). A professional, often certified that is responsible for developing and maintaining the integrity of the database.

· Database management system (DBMS). A software that creates and manipulates database systems in a structured format.

· Data cleansing. The process of reviewing and revising data to eliminate duplicate entries, correct spelling mistakes and add missing data.

· Data collection. Any process that leads to the acquisition of data.

· Data-directed decision making. Using database as the basis to support making crucial decisions.

· Data exhaust. The by-product that is created by a person who uses database system.

· Data feed. A means for any person to receive a stream of data such as RSS.

· Data governance. A set of processes that promotes the integrity of the data stored in a database system.

· Data integration. The act of combining data from diverse and disparate sources and presenting it in a single coherent and unified view.

· Data integrity. The validity or correctness of data stored in a database. It ensures accuracy, timeliness, and completeness of data.

· Data migration. The process of moving data from one storage location or server to another while maintaining its format.

· Data mining. The process of obtaining patterns or knowledge from large sets of databases.

· Data science. A discipline that incorporates the use of statistics, data visualization, machine learning, computer programming and data mining database to solve complex problems in organizations.

· Data scientist. A professional who is knowledgeable in data science.

· Machine learning. Using algorithms to allow computers to analyze data for the purpose of extracting information to take specific actions based on specific events or patterns.

· MongoDB. It is a NoSQL database system that is oriented to documents and developed under the open source concept. It uses JSON to save data structures in documents with a dynamic scheme.

· Qualitative analysis. The process of analyzing qualitative data by interpreting words and text

· Quantitative analysis. The process of analyzing quantitative data by interpreting numerical data.

· Quartiles. The lower (Q1) quartile is the value below for which the bottom 25% of any sampled data lies, and the upper (Q3) quartile is the value above which the upper 25% of sampled data lies.

· R. It is an open source programming language for performing data analysis.

· Random sample. Every member of a given population has an equal chance of being selected in a random sample. The random sample is the representative of the population that is being studied.

· Representative. The extent to which the sampled data reflect accurately the characteristics of the selected population in an experiment.

· Research process. The process that is undertaken by researchers or data scientists to answer research questions and hypotheses.

· Research question. A specific question that is supposed to guide the research process.

· Sample. A subset (n) that is selected from entire population (N). Scatter

· Significance level. Setting the p-value.

· Standard deviation. It is a descriptive statistic — which is a measure of dispersion, or spread — of sampled data around the mean.

· The standard error of the mean. It is a measure of the accuracy of the sampled mean as an estimate of the entire population mean.

· Hypothesis. A precise statement which is a proposition that relates to the research question to be tested.

· Independent variable. The variable that determines the values of the other dependent (response) variable. For instance, blood pressure can be deemed to in response to changes in age.

Tools and basic prerequisites for a beginner in data analytics

By now, you are wondering, “Where should I start to become a professional data analyst?”

Well to become a professional data scientist, here is what you should learn:

· Mathematics

· Excel

· Basic SQL

· Web development

Let’s see how these fields are important in data analytics.

#1: Mathematics

Data analytics is all about numbers. If you relish working with numbers and algebraic functions, then you’ll love data analytics. However, if you don’t like numbers, you should begin to cultivate a positive attitude. Also, be willing to learn new ideas. Truth be told — the world of data analytics is fast-paced and unpredictable. Therefore, you can’t be contented. You should be ready to learn new technologies that are springing up to deal with changes in data management.

#2: Excel

Excel is the most all-around and common business application for data analytics. While many data scientists graduate with functional specific skill — such as data mining, visualization, and statistical applications — almost all these skills can be learned in Excel. You can start by learning the basic concepts of Excel such as the workbooks, the worksheets, the formula bar and the ribbon.

Once you’ve familiarized with concepts of Excel, you can proceed to learn the basic formulas such as sum, average, if, count, vlookup, date, max and min, getpivotdata. As you begin to become more comfortable with basic formulas, you can try out the complex formulas for regression and chi-square distributions.

#3: Basic SQL

Excel provides you with tools to slice and dice your data. However, it assumes you already have the data stored in your computer system. What about data collection and storage. As you’ll learn about seasoned data scientists, the best approach to deal with data is getting it or pulling it directly from its source. Excel doesn’t provide you with these functionalities.

Relational database management systems (RDBMS) — such as SQL Server, Ms. Access, and MySQL — support procedures for data collection. To master relational database management systems, you should be good in SQL (Structured Query Language) — the language that underpins all the RDBMS.

To fast-track the mastery of SQL, you should understand how the following statements/commands are used:

· Select

· From

· Where

· Group By

· Having

· Order By

Besides mastering the basic SQL commands, you should also understand the reason behind use of primary keys, foreign keys and candidate keys in these DBMSs’

#4: Basic web development

I know you’re thinking that web development is an odd-ball with regard to data analytics. But trust me, mastery of web development will be an added bonus to your data scientist career. If you want to work for consumer internet companies or work for IoT companies such as IBM, AWS, and Microsoft Azure, you have to be good in internet programming tools such as HTML, JavaScript and PHP.

**Advanced tools and prerequisites for data analytics**

If you wish to take your professional career to the next level, then basic pre-requisites for data analytics may be insufficient. Below are advanced tools and pre-requisites for data analytics:

#1: Hadoop

Hadoop is a cloud computing platform that you can use to perform highly parallelized operations on your big data. It is an open-source software framework that stored big data and allows applications to run on it in form of clusters. One advantage of Hadoop is that at allows users to store and process massive storage of data of any type. Because of enormous processing power, Hadoop is suited for analysis of big data with virtually limitless simultaneous tasks.

#2: R programming

Every person that start the journey of data science usually faces the common problem of selecting the best programming language. Today, there are a couple of programming languages that can perform data analytics. Each of these programming languages has their own fair share of pros and cons. However, R programming language is a tested programming language that you can try out.

R is very useful for data analytics due to its versatility nature especially in the field of statistics. It is an open source software that provides data scientists with a variety of features for analyzing data.

Below are reasons that make R popular in data analytics:

· It is simple, well developed and one of the efficient programming languages that support loops, recursive functions, conditionals, and input/output facilities.

· It provides programming operators that can perform calculations on vectors, arrays, matrices and lists.

· It has storage facilities therefore, data analysts can effective handle their data.

· It has graphical facilities that data analysts can use to display processed data.

#3: Python programming

Python is a very powerful, open source and flexible programming language that is easy to learn, use and has powerful libraries for data manipulation, management, and analysis. If you have basic skills in these programming languages, you’ll not have a problem with Python language.

In addition, Python combines the features of general-purpose programming language and those of analytical and quantitative computing. In the recent past, Python has been applied in scientific computing with highly quantitative domains. For instance, Python has found its applications in finance, physics, oil, gas and signal processing.

Similarly, Python has been used to develop popular scalable web applications such as YouTube. Because of its popularity, Python can help you with tools for big data and business analytics in science, engineering and other areas of scalable computing. You can use Python’s inbuilt libraries such as Panda and NumPy to help you with data analytics.

This is because it integrates well with existing independent IT infrastructure systems. Among modern programming languages, the agility productivity of Python-based applications is legendary.

#4: Database proficiency tools

Database proficiency tools — such as SQL Server, Ms Access, MongoDB, and MySQL — support procedures for data collection, storage and processing. To master these systems, you should be good in SQL (Structured Query Language) — the language that underpins all these systems.

To fast-track the mastery of SQL, you should understand how the following statements/commands are used:

· Select

· From

· Where

· Group By

· Having

· Order By

Besides mastering the basic SQL commands, you should also understand the reason behind the use of primary keys, foreign keys and candidate keys in these systems.

#5: MatLab

MatLab is a very powerful, flexible and open source programming language that is easy to learn, use and has powerful libraries for data manipulation, management, and analysis. Its simple syntax is easy to learn and resembles C or C++. If you have basic skills in these programming languages, you’ll not have a problem with MatLab language.

In addition, MatLab combines the features of general-purpose programming language and those of analytical and quantitative computing. In the recent past, MatLab has been applied in scientific computing with highly quantitative domains.

Similarly, MatLab has been used to develop popular scalable web applications such as YouTube. Because of its popularity, MatLab can help you with tools for big data and business analytics in science, engineering and other areas of scalable computing. This is because it integrates well with existing independent IT infrastructure systems.

#6: Perl

Perl is a dynamic and high-level programming language that you can use for data analytics. Originally developed as a scripting language for UNIX by Larry Wall, Perl has provided its UNIX-like features and flexibility of any programming language to develop robust and scalable systems.

With the advent of the internet in 1990’s, Perl usage exploded. Besides providing dominant features of CGI programming, Perl has also become a key language for data analysis because of its rich set of analysis libraries.

#7: Java

Java and its Java-based frameworks are found deep in the skeletons of virtually all the biggest Silicon Valley tech companies. When you look at Twitter, LinkedIn, or Facebook, you’ll find that Java is the backbone programming language for all the data engineering infrastructures. While Java doesn’t provide the same features of data analytics as Python and R, I can bet for Java when it comes to the excellent performance of systems on a large scale.

Java’s speed makes it one of the best languages for developing large-scale systems. While Python is significantly faster than R, the Java language provides an even greater performance compared to Python. It is because of this reason that Twitter, Facebook, and LinkedIn have picked Java as the backbone of their systems. However, Java language may not be appropriate for statistical modeling.

#8: Julia

Today, the vast majority of data analytics use R, Java, MatLab and Python for data analysis. However, there’s still some gap that requires being filled since there’s no language that’s one-stop-shop for data analysis needs. Julia is a new programming language that can fill the gaps with respect to improving visualizations and libraries for data analysis.

Even though the Julia programming community is in an infancy, more and more programmers will soon realize its potentials in data analysis and adopt it.

Data analytics workflow

Data analytics workflow can be explained in the following steps:

· Preparation phase

· Analysis phase

· Reflection phase

· Dissemination phase

Let’s dive in to explore these phases.

#1: Preparation stage

Before you analyze your data, you must acquire the data and reformat it into in a manner that is suitable for computation. You can acquire data from the following sources:

· Data files from online repositories such as the public websites. For instance, the U.S. Census data sets.

· Data files streamed on-demand through APIs. For instance, the Bloomberg financial data stream.

· Physical apparatus such as scientific lab equipment that has been attached to computers.

· Data from computer software such as log files from the web server.

· Manually entering the data in spreadsheet files.

#2: Analysis phase

At the core of any data analytics activity is analysis. This involves writing computer programs or scripts that analyze the data to derive helpful insights from it. You can use programming languages such as Python, Perl, MatLab, R or Hadoop.

#3: Reflection phase

At this stage, you’ll frequently be alternating between the analysis and the reflection stages as you work on your data to obtain the necessary information. While the analysis phase is a purely programming process, the reflection phase requires critical thinking and communication with your clients about the outputs obtained.

After inspecting your set of output files, you can take notes if you’re dealing with an experiment that that’s in either physical or digital format.

#4: Dissemination phase

Dissemination is the final phase of data analysis workflow. You can present your results using written reports such as the internal memos, PowerPoint presentations or business white papers. If you’re in the academic field, you can publish the academic paper.

Statistical process

The process of data analysis begins with identifying the population from which you’ll obtain data. Because it’s practically impossible to get data on every subject in the population, you should use an appropriate sampling technique to get a sample size that’s representative. The statistical process is a four-step phase activity that includes the following:

· Estimate the expected proportion of the population that you want to study. The proportion of that population must of interest to the study. If you have the agreed benchmark from literature review or prior studies, you can use it as the basis for your expected proportion. If in doubt, consult experts in that field to get the correct estimate.

· Determine the confidence interval for use in your analysis. Think of confidence level as the “margin of error” in your sample size. Now, all the empirical estimates are based on a sample that must have a certain degree of uncertainty. It’s a must for you to specify the desired total spectrum of the confidence interval.

· Set the value of the confidence level. This provides the precision or level of uncertainty in the analysis. Typically a 95% confidence level is widely used. However, a narrow confidence interval that has a high confidence level such as 99% is likely to be as representative of the population as possible.

· Use the statistical table to estimate your sample size. If the number that is required is too large, you can recalculate it with lower confidence levels or use wider intervals to choose a smaller sample size.

Descriptive and Inferential statistics

Statistics is broadly divided into two fields: descriptive and inferential. Descriptive statistics provides information about the distribution, variations, and the shape of the data. Think of descriptive statistics as that statistics which analyzes a big chunk of data to provide summary charts, bar graphs, pie charts using descriptive measures such as:

· Measures of central tendency such as mean, mode, and median.

· Measures of dispersion such as range, variance, and standard deviation.

· Measures of a shape such as skewness.

However, descriptive statistics doesn’t draw conclusions about the population from where the sample was obtained. If you’re interested in knowing the relationships, differences within your data or whether statistical significance exists, you have to use inferential statistics.

Inferential statistics provides these determinations and allow you to generalize your results obtained from the sample size to the larger population. Some of the models you’re likely to use for inferential statistics include:

· Chi-square distributions

· Correlation and regression models

· ANOVA

· ANCOVA