unionbank

TURBOCHARGED
FOR SUCCESS

Introduction to Data Analytics

October 2021

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

294bn

billion emails are sent

Radicati Group

320bn emails to be sent each day by 2021

306bn emails to be sent each day by 2020

3.9bn

people use emails

4TB

of data produced by a connected car

Intel

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

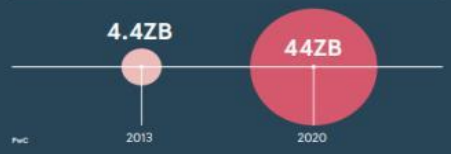
DEMYSTIFIYING DATA UNITS

From the more familiar "bit" or "megabyte", larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

ACCUMULATED DIGITAL UNIVERSE OF DATA



PwC

Searches made a day 5bn

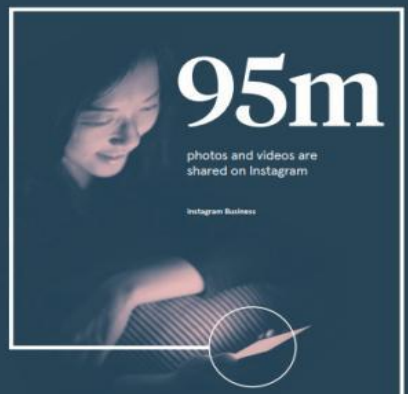
Searches made a day from Google 3.5bn

Smart Insights

463EB

of data will be created every day by 2025

IDC



95m

photos and videos are shared on Instagram

Instagram Business

28PB

to be generated from wearable devices by 2020

Statista

RACONTEUR

Table of Contents

Introduction to Data Analytics

Data analytics is the practice of using data to drive business strategy and performance.

It includes a range of approaches and solutions, from looking backward to evaluate what happened in the past to looking forward to do scenario planning and predictive modelling.

Introduction to Data Analytics

Brief Overview of Data Analytics

Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Descriptive Analytics describes the happenings over time, such as whether the number of views increased or decreased and whether the current month's sales are better than the last one.	It focuses on the reason for the occurrence of any event. It requires hypothesizing and involves a much diverse dataset. It examines data to answer questions, such as "Did the new ad strategy affect sales?"	This focuses on the events that are expected to occur in the immediate future.Predictive analytics tries to find answers to questions like, How many weather forecasts is expected this year's hot summer?	It indicates a plan of action. If the chance of a hot summer calculated as the average of the five weather models is above 58%, an evening shift can be added to the brewery, and an additional tank can be rented to maximize the production.

Introduction to Data Analytics

Our Perspective

The process of delivering analytics business results is one of continuous improvement. Starting anywhere in the analytics cycle, an organization can promptly begin to address its specific needs.

Applied Analytics

Issues
- What business problem are you solving

Information Management
- Govern
- Architect
- Integrate

Facts
- What data can be leveraged to understand the business and improve performance

Performance Optimization
- Monitor
- Report
- Recommend

Understanding
- What is currently happening or has happened related to our business and why?
- What should we do about it?

Analytics Insights
- Predict
- Embed
- Discover

Actions
- How do we look to the future and build analytics insights directly into business processes?

Introduction to Data Analytics

What is Data?

- Data are letters, numbers or symbols representing information, suitable for processing, communicating or interpreting.

- What does the string of data below represent?

04092003

- This string could be interpreted as

Introduction to Data Analytics

What is Data?

04092003

The string could be interpreted as any of the following;

- An amount (NUMERIC) $4,092,003 or $40,920.03
- An Invoice Date (DATE) April 9, 2003 or September 4, 2003
- An Account Number (CHARACTER) 04092003

Introduction to Data Analytics

- Data is a collection of text, numbers, images, audios and symbols with no meaning
- Data therefore must be processed, or provided with a context, before it can have meaning
- However, data is, the second most important thing in data analytics after the question you're trying to answer
- Often the data will limit, or enable the questions you're trying to ask, as such, data analytics starts with the question
- You might not have the data to be able to answer the question and may have to result to modification of the question or create a sub-question or a related question to attempt to give an answer with analytics.
- But having data in general, can't save you if you don't have a question that you're asking

Why Data Analytics (1/2)

Faster and Better Decision Making: Analytics is important for your business to the extent that making good decisions is. The practice of analytics is all about supporting decision making by providing the relevant facts that will allow you to make a better decision.

Seeing new opportunities: Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers.

Cost reduction. Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.

Why Data Analytics (2/2)

Application to Business Areas

Successful uses of Analytics are always rooted in real business problems

	Applying Analytics…	For example:
Customer & Growth	…to enhance the customer lifecycle, sales and pricing processes, and overall customer experience	• Detailed segmentation to better target cross-sell and up-sell activity • Understanding lifestyle factors to improve pricing & risk calculations for products • Predicting the impact of different compliance actions (e.g. court cases) on [music] licensing revenues • Identifying and managing the most profitable customers (customer lifetime value) across a portfolio of products and services
Operations & Supply-chain	…to provide insights across the organisation's value chain	• Analysing spend to identify efficiencies across the value chain • Identifying candidate locations for new sites (e.g. depots, or retail outlets) based on a range of geospatial factors • Monitoring traffic flows using mobile-phone location data
Finance	…to measure, control, and optimise financial management processes	• Consolidating financial reporting with other data to provide multi-dimensional views and more accurate financial forecasts • Simulating the impact of changes in the financial markets (e.g. stress testing of the banking system) • Analysing aggregated financial returns to suggest tax efficient structures
Workforce	…to enhance and optimise workforce processes and intelligence	• Reducing overtime by optimising scheduling • Forecasting demand to improve workforce planning • Identifying early indicators of attrition to improve retention • Analysing employee data to identify those most at risk of workplace accidents
Risk & Regulatory	…to measure, monitor and mitigate enterprise risk	• Identifying and investigating instances of fraud and error in payment systems • Monitoring compliance with financial regulations (e.g. sanctions, anti-bribery & corruption laws, etc) • Identifying cyber-security breaches from patterns of user behaviour

Key concepts of Data and Analytics

Understanding Data

Data is everywhere. Every time you use a credit card, make a telephone call, or log on to a website, you leave a digital footprint.

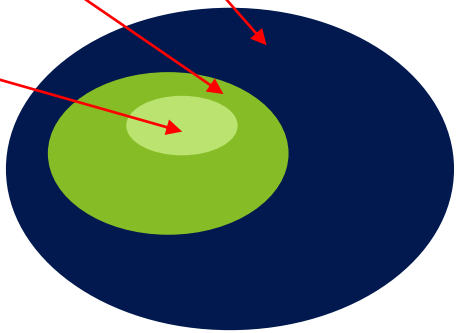
To work with raw data, One must know how to interpret it and what to do with it. In the same way, a computer must first be told how to interpret raw data and then how to manage it.

Key concepts of data and analytics

Files, Records and Fields

Modern data structures are built from files

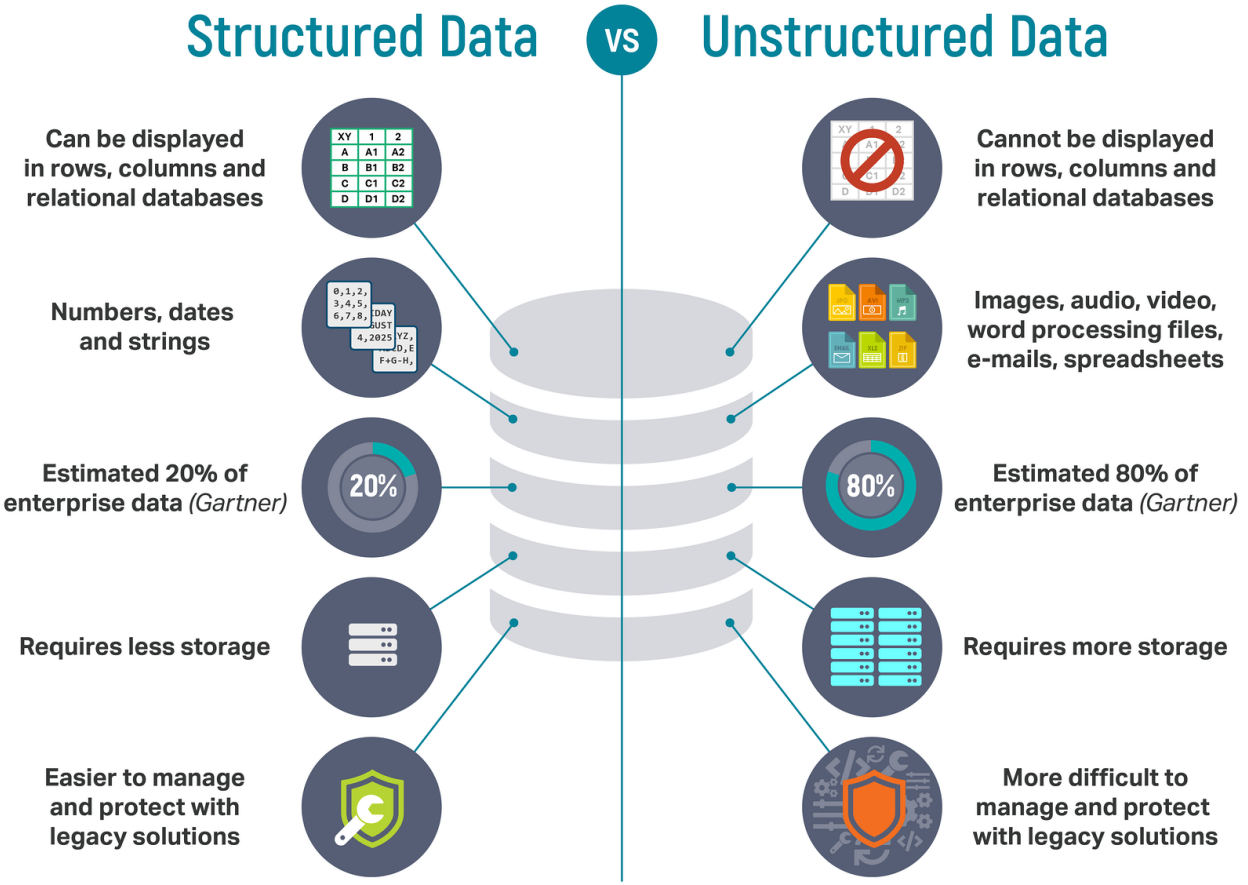
- **Files**
  - Named collection of records stored or processed as an individual entity
- **Records**
  - Subset of a file
  - Collection of related fields containing data items grouped for processing
- **Fields**
  - Subset of a record
  - A specified area of a record used for storing a particular class of data



Categories of Data

More Examples

- Customer information
- Payment or expense details
- Payroll details
- Transaction records
- Financial reports
- Often found in databases, spreadsheets and data collection software



More examples

- Email and instant messages
- Videos, audio files, pictures etc.
- Social media activities
- Corporate document repositories
- News feeds

What is Big Data

Definition

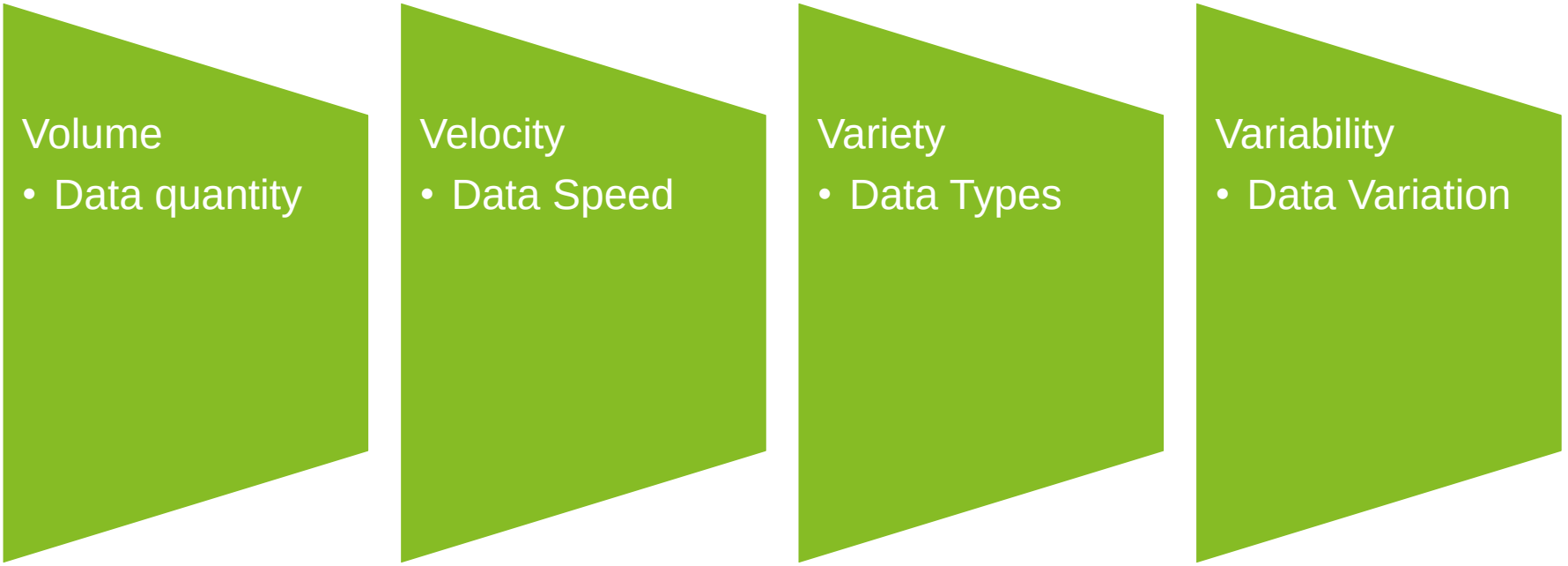
Big Data' is like 'small data', but bigger in size

Big Data is an extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

Characteristics of Big Data (1/5)

‘There are Four ‘4’ Vs of Big Data



Characteristics of Big Data (2/5)

Volume

The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. Also, whether a particular data can be considered as a Big Data or not, is dependent upon volume of data.

Hence, **'Volume'** is one characteristic which needs to be considered while dealing with 'Big Data'.

Characteristics of Big Data (3/5)

Velocity

The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, mobile devices, etc. The flow of data is massive and continuous.

Characteristics of Big Data (4/5)

Variety

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications.

Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. is also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

Characteristics of Big Data (5/5)

Variability

This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Benefit of Big Data Analytics (1/2)

Businesses can utilize outside intelligence while taking decisions

Access to social data from search engines and sites like Facebook, twitter are enabling organizations to fine tune their business strategies.

Improved Customer Service

Traditional customer feedback systems are getting replaced by new systems designed with 'Big Data' technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

Benefit of Big Data Analytics (2/2)

Better Operational Efficiency

'Big Data' technologies can be used for creating staging area or landing zone for new data before identifying what data should be moved to the data warehouse.

In addition, such integration of 'Big Data' technologies and data warehouse helps organization to offload infrequently accessed data.

Risk associated with Big Data (1/3)

As with any business initiative, a big data project involves an element of risk. Any project can fail for any number of reasons: bad management, under-budgeting, or a lack of relevant skills. However, big data projects bring their own specific risks. Highlighted below are some of the biggest risks of big data project

- **Data Security**
- **Data Privacy**
- **Cost of Implementation**

Risk in Big Data (2/3)

Data Security

Data theft is a rampant and growing area of crime – and attacks are getting bigger and more damaging. The bigger your data, the bigger the target it presents to criminals with the tools to steal and sell it.

Data Privacy

Closely related to the issue of security is privacy. But in addition to ensuring that people's personal data are safe from criminals, you need to be sure that the sensitive information you are storing isn't going to be divulged through less malevolent but equally damaging misuse by yourself or by people to whom you have delegated responsibility for analyzing and reporting on it.

Risk in Big Data (3/3)

Cost of Implementation

Data collection, aggregation, storage, analysis, and reporting all cost money. These costs can be mitigated by careful budgeting during the planning stages but getting it wrong at that point can lead to spiraling costs, potentially negating any value added to your bottom line by your data-driven initiative.

Computer Assisted Auditing Techniques (CAAT)

Definition

CAATs is the practice of using computers to automate the IT audit processes. CAATs normally includes using basic office productivity software such as spreadsheet, word processors and text editing programs and more advanced software packages involving use statistical analysis and business intelligence tools.

Common CAAT tools used for Data Analytics (1/2)

The analytics industry has seen a paradigm shift in how and which analytics tools are used. There are some popularly used data analytics tools in today's dispensation.

Excel - Excel is the most widely used analytics tool in the world.

SQL – SQL is a standard language for storing, **manipulating** and retrieving data in databases

ACL - ACL is the only software provider to seamlessly integrate the industry's standard data analytics capabilities into a comprehensive risk assessment, audit management, work-papers, issue tracking and remediation workflow with powerful visualization and dash-boarding.

ACL data analytics is a data extraction and analysis software used for fraud detection & prevention, and risk management. By sampling large data sets, ACL data analysis software is used to find irregularities or patterns in transactions that could indicate control weaknesses or fraud.

Common CAAT tools used for Data Analytics (2/2)

IDEA - IDEA is a powerful and user-friendly data analysis tool designed to help auditors, accountants and other finance professionals perform data analysis quickly to help improve audits and identify control breakdowns.

R – R is a statistical programming language and environment for statistical computing and graphics. R easily handles large data sets, and it has also become a lot more versatile.

Python – Python has been a favorite of programmers for long. This is mainly because it's an easy to learn language that is also quite fast.

It has developed into a powerful analytics tool with the development of analytical and statistical libraries and offers a comprehensive coverage of statistical and mathematical functions.

SAS – SAS is a robust, versatile and easy to learn tool with tons of modules.

Benefit of Computer assisted audit techniques

(1) CAATs enable the auditors to test huge volume of data, or the operation of the controls in a system, precisely and rapidly and are therefore, very cost efficient when operated appropriately.

(2) CAATs lessen the level of human error in testing and facilitate a very high level of audit evidence to be derived.

(3) The application of CAATs free up expensive human capital enabling auditors to focus more attention on key judgmental areas

How CAATs impact our audit (1/2)

- The audit program need to be run on the live database (i.e., actual files) of the client because the auditor is testing the actual system of the client. Some clients may be unwilling to let auditors run the audit program on the live/actual files as this need to be fully tested that it won't corrupt the actual database files.

Alternative solution is to run the audit software on backup copies of the live database. But the problem is ensuring that the copied files are true replicate of the live files.

How CAATs impact our audit (2/2)

- CAATs are too costly to setup and necessitate the support of the client. It is usually necessary for a continuing audit relationship to be present before it is worth committing the audit resources.

- Major changes in client systems often necessitate major changes in CAATs, which is costly. If the audit fee is based on the assumption that the prior year's CAATs can be used, and a change is made without warning, the client may have unlikely expectations about the level of service that can be provided for the fee.

Future of Computer assisted audit techniques

- Technologies that support the audit process, for example, scheduling, work paper management, etc;

- Continuous Control Monitoring (CCM), or functionality embedded in a business process that looks for situations that may indicate a control has been or may be breached;

- Data Analytics, where substantive analysis of data is used to support financial and operational auditing or conclusions made about the operation of controls; and,

- Specialist analysis technology, which use tools that will check user access rights for incompatible functions, security and configuration weaknesses.

Data Analysis Cycle (1/6)

All data analytics project have a defined life cycle that generally fits into the five (5) phases below:

- Planning
- Data access
- Data integrity verification
- Data analysis
- Reporting results

Data Analysis Cycle (2/6)

- Planning
- Data access
- Data integrity verification
- Data analysis
- Reporting results

Planning

Plan your work before you start.

Consider the steps required to achieve the objectives of the remaining 4 phases.

Formulate clear objectives, develop concise strategies, and budget the right amount of time.

Data Analysis Cycle (3/6)

- Planning
- Data access
- Data integrity verification
- Data analysis
- Reporting results

Data access

Access the data outlined in your strategic plan.

Includes locating, requesting, and transferring the data.

Data Analysis Cycle (4/6)

- Planning
- Data access
- Data integrity verification
- Data analysis
- Reporting results

Data integrity verification

Test the integrity of the data you receive. If you don't, results may be incomplete or incorrect.

Data Analysis Cycle (5/6)

- Planning
- Data access
- Data integrity verification
- Data analysis
- Reporting results

Data analysis

Perform the tests necessary to achieve your objectives.

Data Analysis Cycle (6/6)

- Planning
- Data access
- Data integrity verification
- Data analysis
- Reporting results

Reporting results

Create reports from your results and document your work.

Data Source Testing (1/2)

To ascertain the reliability of data provided for use, one must perform procedures to address concerns around the (1) Completeness and (2) Accuracy of the data by considering;

- The source data
- The report logic, which includes how the data is extracted and any calculations
- Any relevant parameters that are user entered

For example, consider the process to generate a typical terminated user listing. We need to determine for example

(1) whether the appropriate parameters were applied,

(2) whether the source data in the HR database is accurate and complete, and

(3) whether the report logic that generates the terminated user listing is appropriate.

See illustration on the next slide.

Data Source Testing (2/2)