Data

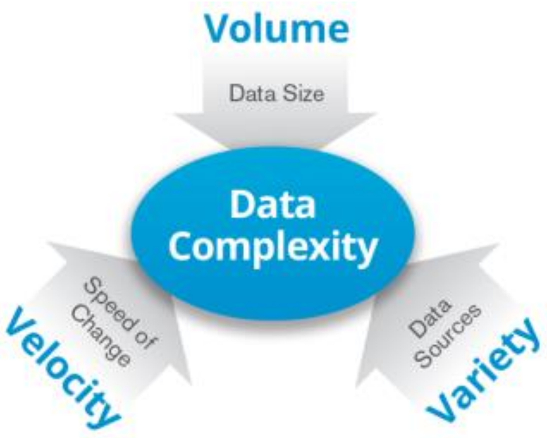
- **Data is a set of qualitative or quantitative variables – it can be structured or unstructured, machine readable or not, digital or analogue, personal or not. Ultimately it is a specific set or sets of individual data points, which can be used to generate insights, be combined and abstracted to create information, knowledge and wisdom.**
- **Traditional analysis tools and software can be used to analyse and “crunch” data.**

Big Data

- Big Data **refers to datasets whose size is beyond the ability of typical database software** tools to **capture, store, manage and analyze**.” *(McKinsey Global Institute)*
- “**Big Data is the term** for a **collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications**.” *(Wikipedia)*
- **“Big data” is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, decision making and process automation.”**

What is Big Data?

- **It refers to a massive amount of data that keeps on growing exponentially with time.**
- **It is so big that it cannot be processed or analyzed using conventional data processing techniques.**
- **The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.**
- **It includes data mining, data storage, data analysis, data sharing, and data visualization.**

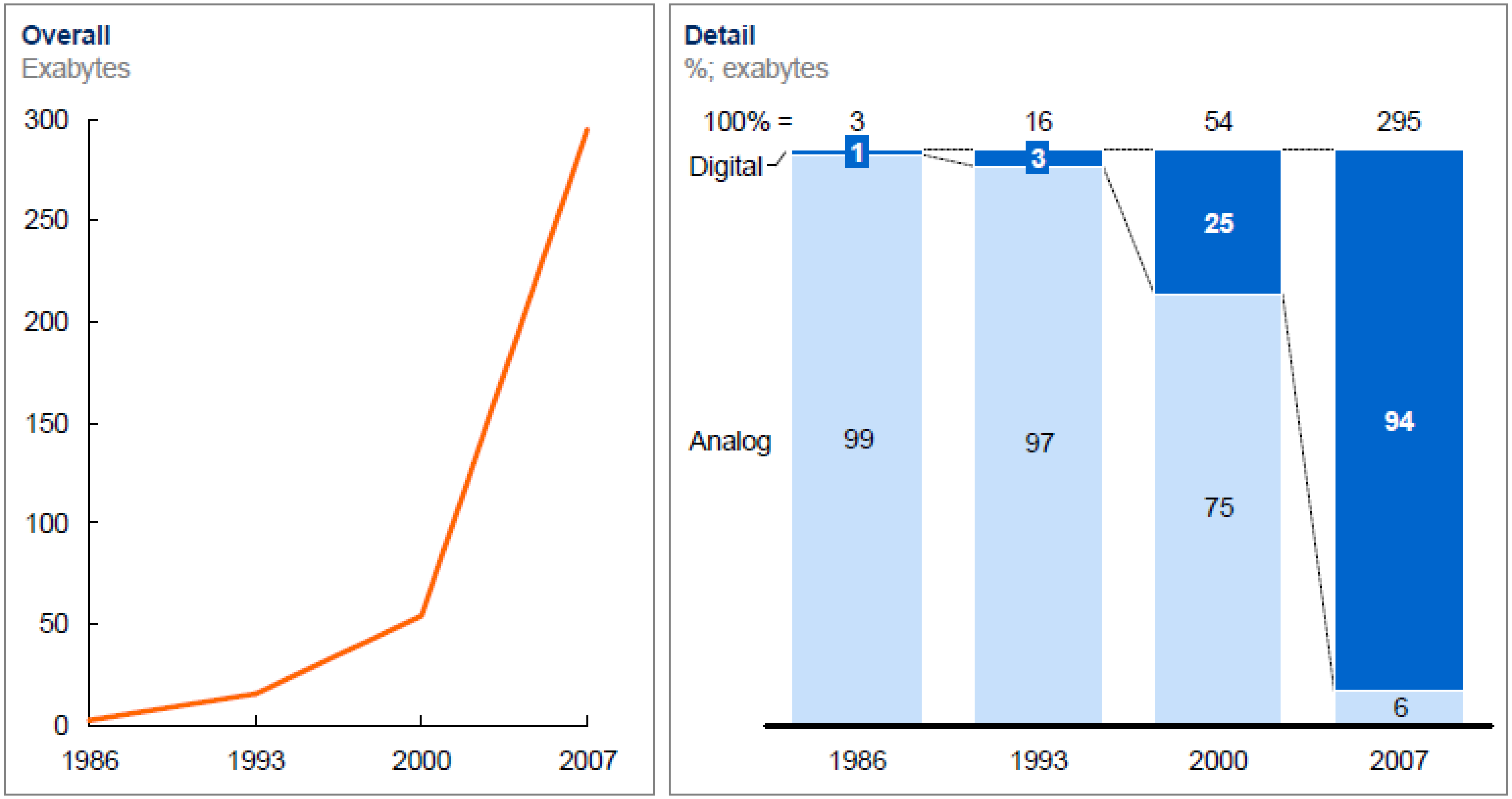


Key enablers for the growth of “Big Data” are:

- **Increase of storage capacities**

Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage



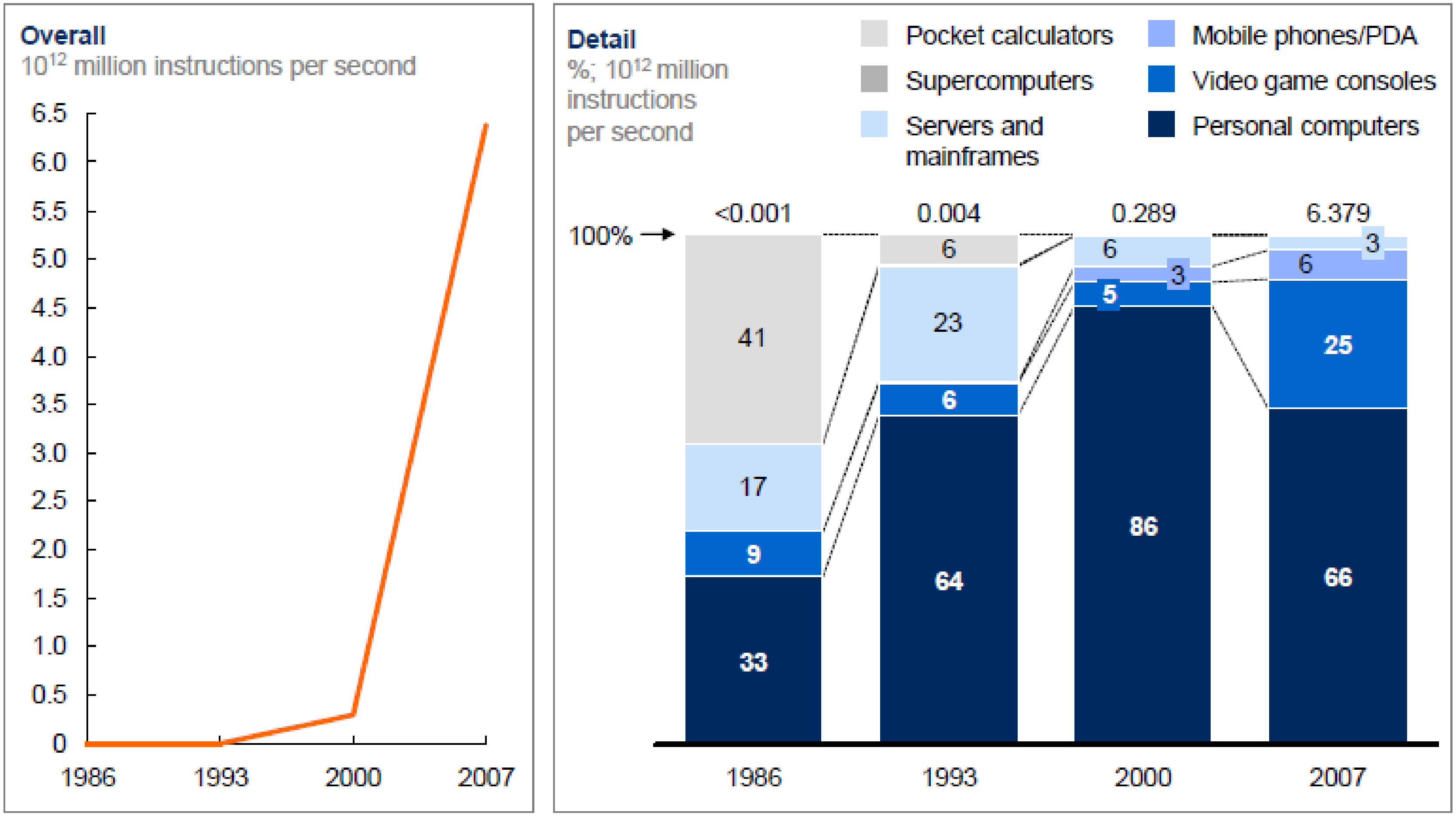
NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, “The world’s technological capacity to store, communicate, and compute information,” *Science*, 2011

- **Increase of processing power**

Computation capacity has also risen sharply

Global installed computation to handle information

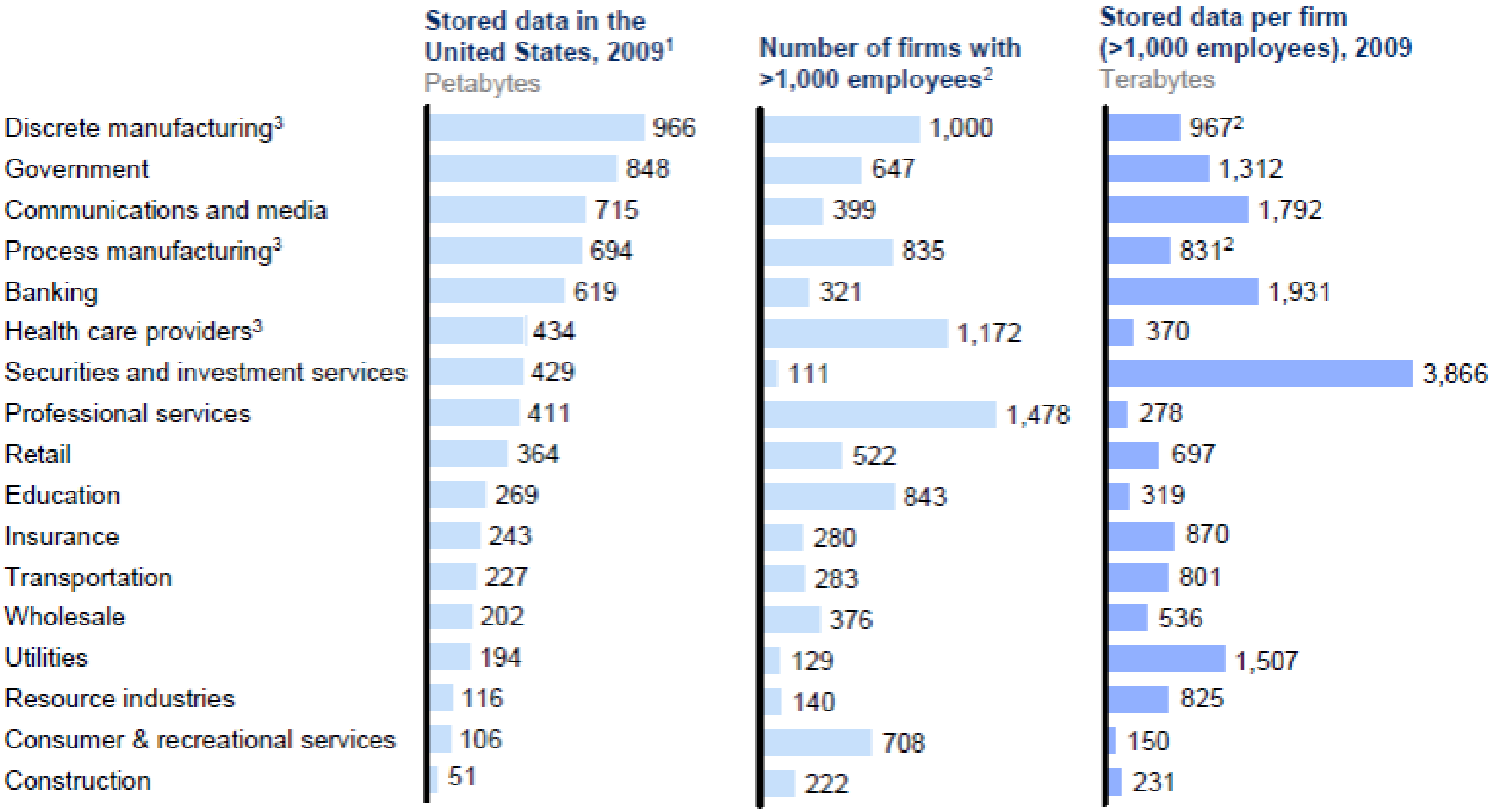


NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

- **Availability of data**

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte



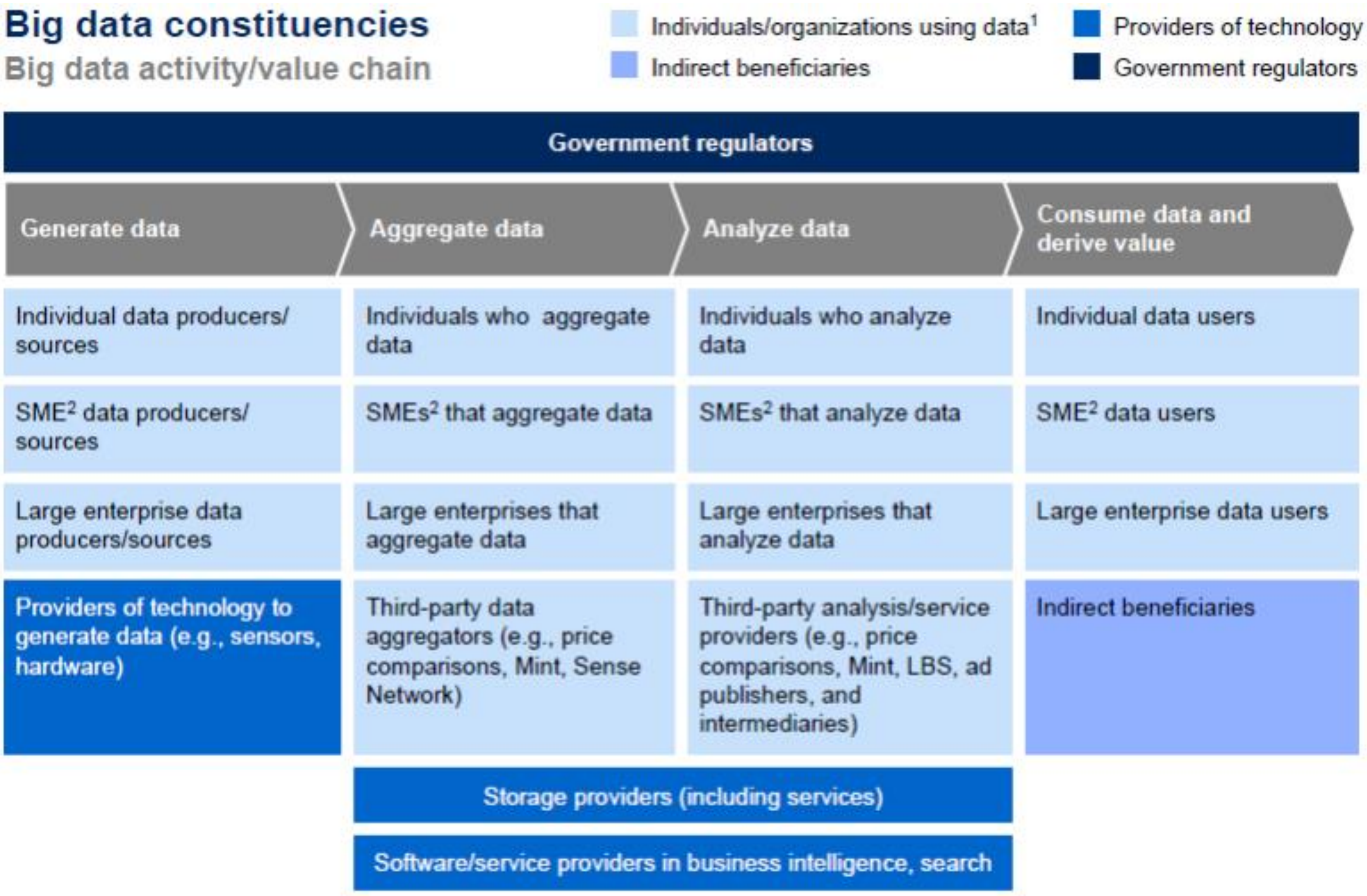
[1] Storage data by sector derived from IDC.

[2] Firm data split into sectors, when needed, using employment

[3] The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

Big data Value chains



Big data constituencies

Big data activity/value chain

□ Individuals/organizations using data[1]

▨ Indirect beneficiaries

🟦 Providers of technology

■ Government regulators

Government regulators

Generate data	Aggregate data	Analyze data	Consume data and derive value
Individual data producers/ sources	Individuals who aggregate data	Individuals who analyze data	Individual data users
SME[2] data producers/ sources	SMEs[2] that aggregate data	SMEs[2] that analyze data	SME[2] data users
Large enterprise data producers/sources	Large enterprises that aggregate data	Large enterprises that analyze data	Large enterprise data users
Providers of technology to generate data (e.g., sensors, hardware)	Third-party data aggregators (e.g., price comparisons, Mint, Sense Network)	Third-party analysis/service providers (e.g., price comparisons, Mint, LBS, ad publishers, and intermediaries)	Indirect beneficiaries

Storage providers (including services)

Software/service providers in business intelligence, search

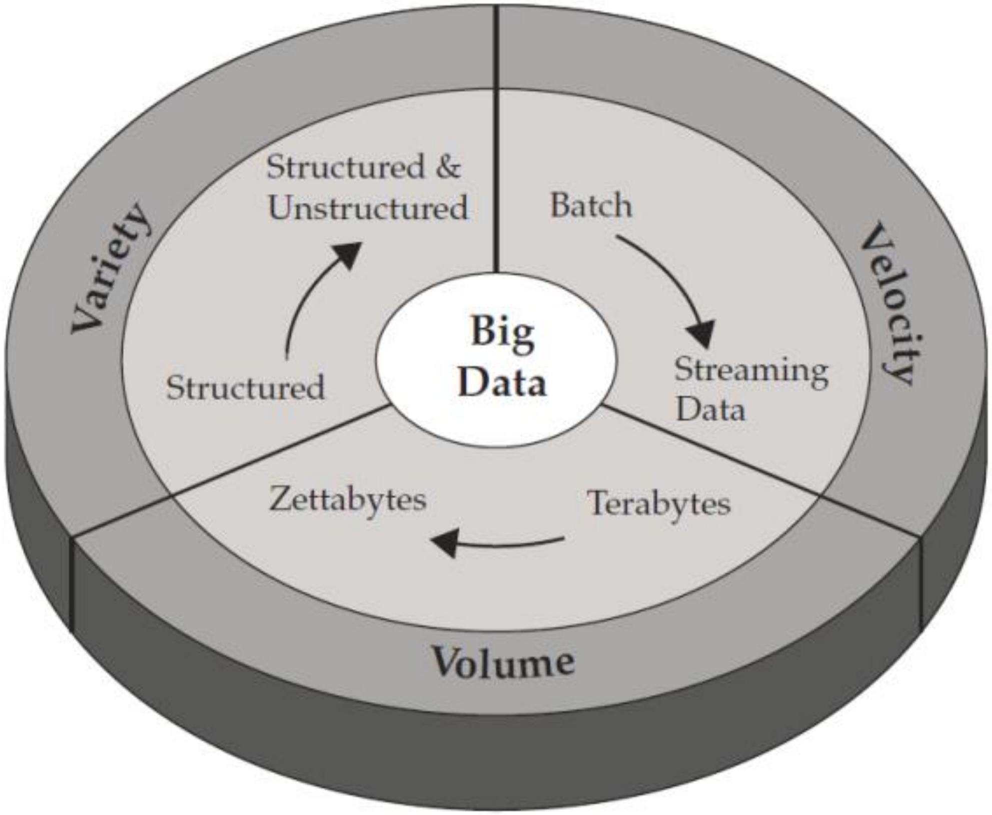
[1] Individuals/organizations generating, aggregating, analyzing, or consuming data.

[2] Small and medium-sized enterprises.

SOURCE: McKinsey Global Institute analysis

Big Data Characteristics

- **Big data is information too large or complex for traditional data processing methods to analyze**. To better understand this **definition, consider the three V's of big data**:



- **Velocity**: Unstructured data received in large amounts, such as Twitter data feeds, can sometimes comprise terabytes or petabytes of storage space.
- **Volume**: As internet use grows, businesses receive more data at once, thus requiring more processing capacity.
- **Variety**: Think of the diversity of extensions among the files in your database – MP4, DOC, HTML and more.

How is big data different from traditional data sources?

There are some important ways that big data is different from traditional data sources.

- First, ***big data can be an entirely new source of data***. For example, most of us have experience with online shopping. The transactions we execute are not fundamentally different transactions from what we would have done traditionally. An organization may capture web transactions, but they are really just more of the same transactions that have been captured for years (e.g. purchasing records). However, actually capturing browsing behaviour (how do you navigate on the site, for instance) as customers execute a transaction creates fundamentally new data.

- Second, ***sometimes one can argue that the speed of data feed has increase to such an extent that it qualifies as a new data source***. For example, your power meter has probably been read manually each month for years. Now we have a smart meter that automatically read it every 10 minutes. One are argue that it is the same data. It can also be argued that the frequency is so high now that it enables a very different, more in-depth level of analytics that such data is really a new data source.

- Third, ***increasingly more semi-structured and unstructured data are coming in***. Most traditional data sources are in the structured realm. Structure data are the ones like the receipts from your grocery store, the data on your salary slip, accounting information on the spreadsheet, and pretty much everything that can fit nicely in a relational database. Every piece of information included is known ahead of time, comes in a specified format and occurs in a specified order. This makes it easy to work with.

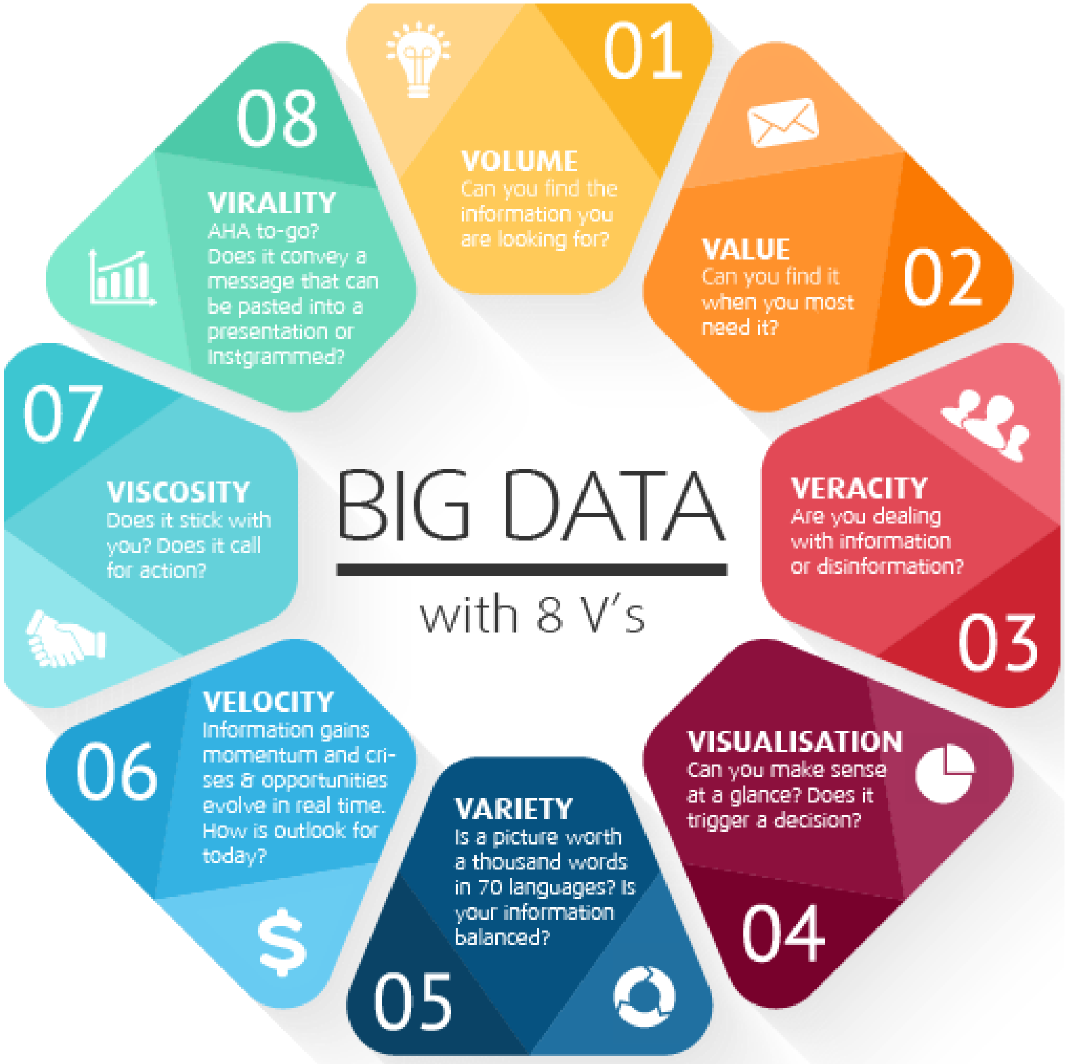
- ***Unstructured data sources are those that you have little or no control over its format***. Text data, video data and audio data all fall into this category. Unstructured data is messy to work with because the meaning of the bites and bits are not predefined. In between structured and unstructured data is semi-structured data. Semi-structured data is data that may be irregular or incomplete and have a structure that may change rapidly or unpredictably.

Big Data three V's: Volume, Velocity and Variety

- **Volume:** **Large amounts of data, from datasets with sizes of terabytes to zettabyte.**
- **Velocity:** **Large amounts of data from transactions with high refresh rate resulting in data streams coming at great speed and the time to act on the basis of these data streams will often be very short . There is a shift from batch processing to real time streaming.**
- **Variety:** **Data come from different data sources. For the first, data can come from both internal and external data source. More importantly, data can come in various format such as transaction and log data from various applications, structured data as database table , semi-structured data such as XML data, unstructured data such as text, images, video streams, audio statement, and more.**

Big Data 8 V's

- **Volume:** Can you find the information you are hunting for?
- **Value:** Can you find it when you most need it?
- **Veracity:** Are you dealing with information or disinformation?
- **Visualization:** can you make sense and trigger a decision
- **Variety:** is it worth?
- **Velocity:** Information gains momentum and opportunity evolves in real time?
- **Viscosity:** does the information stick?
- **Virality:** Does it convey message and can be placed in presentations?


BIG DATA
with 8 V's
01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
$
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go?
Does it convey a message that can be pasted into a presentation or Instgrammed?

The value of big data for business

Big data benefits businesses because it helps to:

- Quickly uncover the causes **behind issues**, **defects** and **failures**.
- Helps in **determining patterns.**
- Rapidly **recalculate entire risk portfolio**.
- Identify **fraudulent or malicious cyber activity** before its **worst consequences can occur**.
- Inform **your full data** and **analytics strategy**.

Types of Big Data

Structured

- **Structured data mean data that can be processed, stored, and retrieved in a fixed format.**
- **It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms.**
- **For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.**

Unstructured

- **Unstructured data refers to the data that lacks any specific form or structure whatsoever.**
- **This makes it very difficult and time-consuming to process and analyze unstructured data.**
- **Email is an example of unstructured data.**

Semi-structured

- **Semi structured is the third type of big data.**
- **Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data.**
- **To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.**

Advantages of Big Data (Features)

- **Predictive analysis:** Big Data analytics tools can predict outcomes accurately, thereby, allowing businesses and organizations to make better decisions, while simultaneously optimizing their operational efficiencies and reducing risks.
- **Forecasting and formulation of quality decisions:** By harnessing data from social media platforms using Big Data analytics tools, businesses around the world are streamlining their digital marketing strategies to enhance the overall consumer experience.
- **Improvement and guarantee on quality:** Big Data provides insights into the customer pain points and allows companies to improve upon their products and services.
- **Consensus:** Being accurate, Big Data combines relevant data from multiple sources to produce highly actionable insights. Big Data tools can help reduce this, saving you both time and money by proper management if of irrelevant data.

- **Increase Profitability: Big Data analytics could help companies generate more sales leads which would naturally mean a boost in revenue. Businesses are using Big Data analytics tools to understand how well their products/services are doing in the market and how the customers are responding to them.**
- **With Big Data insights, you can always stay a step ahead of your competitors. Big Data insights allow you to learn customer behavior to understand the customer trends and provide a highly 'personalized' experience to them.**

Who is using Big Data?

- **Healthcare: With the help of predictive analytics, medical professionals are now able to provide personalized healthcare services to individual patients. Apart from that, fitness wearables, telemedicine, remote monitoring – all powered by Big Data and AI – are helping change lives for the better.**
- **Academia: Big Data is also helping enhance education today. Education is no more limited to the physical bounds of the classroom – there are numerous online educational courses to learn from.**
- **Banking: The banking sector relies on Big Data for fraud detection. Big Data tools can efficiently detect fraudulent acts in real-time such as misuse of credit/debit cards, archival of inspection tracks, faulty alteration in customer stats, etc.**
- **Manufacturing: According to TCS Global Trend Study, the most significant benefit of Big Data in manufacturing is improving the supply strategies and product quality.**

- **IT: IT companies around the world are using Big Data to optimize their functioning, enhance employee productivity, and minimize risks in business operations. By combining Big Data technologies with ML and AI, the IT sector is continually powering innovation to find solutions even for the most complex of problems.**
- **Retail: Big Data has changed the way of working in traditional brick and mortar retail stores. Over the years, retailers have collected vast amounts of data from local demographic surveys, POS scanners, RFID, customer loyalty cards, store inventory, and so on. Now, they've started to leverage this data to create personalized customer experiences, boost sales, increase revenue, and deliver outstanding customer service.**
- **Transportation: Transportation companies use Big Data technologies to optimize route planning, control traffic, manage road congestion, and improve services. Additionally, transportation services even use Big Data to revenue management, drive technological innovation, enhance logistics, and of course, to gain the upper hand in the market.**