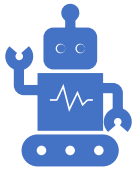
WHAT IS BIG DATA

UNIT 1



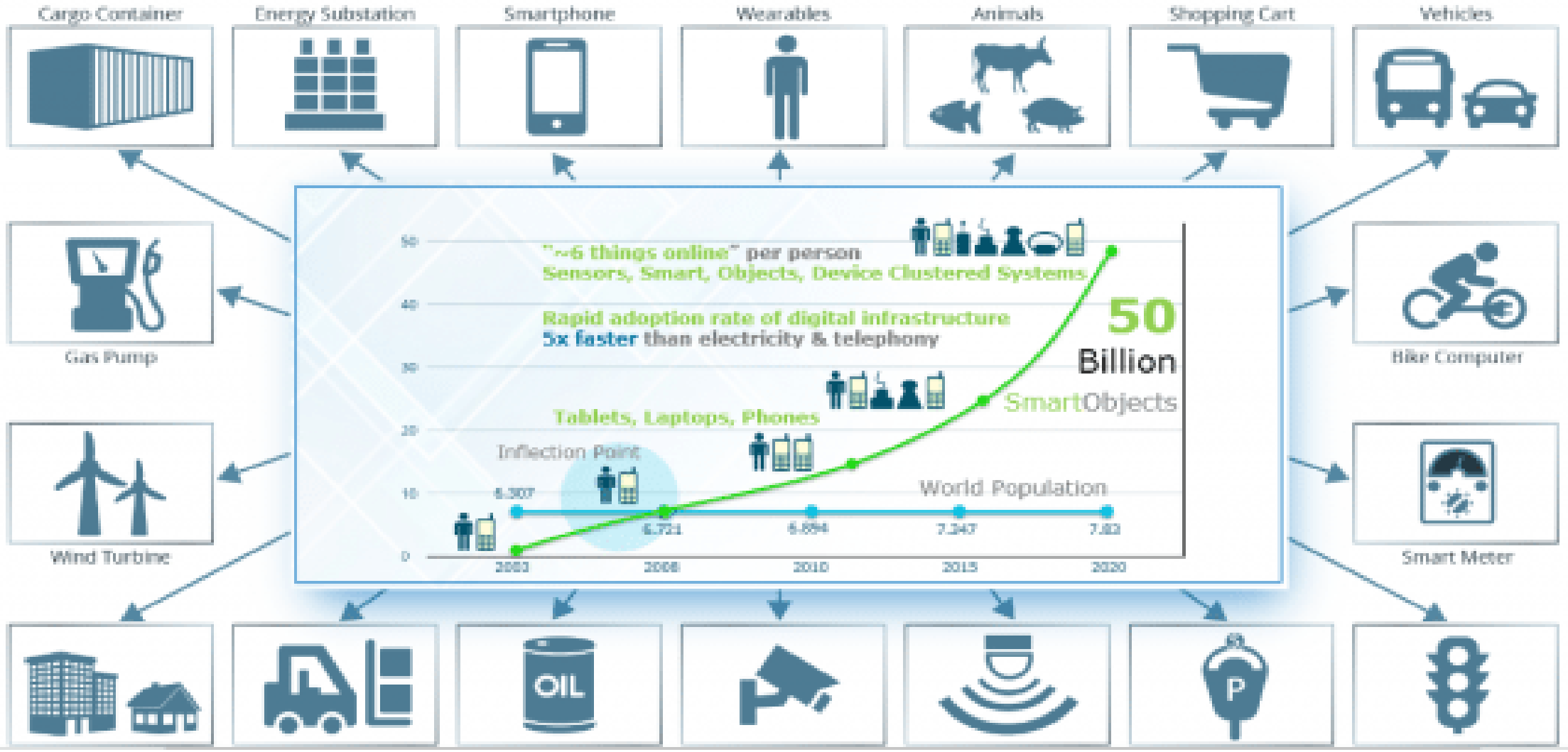
Why Learn Big Data?

To get an answer to Why You should learn Big Data? Let's see what industry leaders have to say:

- **GARTNER** – Big Data is the ***new Oil***.
- **IDC** – Its market will be growing ***7 times faster*** than the overall IT market.
- **IBM** – It is not just a technology – it's a ***Business Strategy*** for capitalizing on information resources.
- **IBM** – Big Data is the ***biggest buzz word*** because technology makes it possible to analyze all the available data.
- **McKinsey** – There will be a ***shortage of 2000000*** Big Data professionals by the end of 2020.

Industries today are searching new & better ways to maintain their position and be prepared for the future. According to experts, **Big Data analytics** provides leaders a path to capture insights and ideas to stay ahead in the tough competition.

Big Data Driving Factors



Big Data Driving Factors

The quantity of data on planet earth is growing ***EXPONENTIALLY*** for many reasons.

VARIOUS SOURCES & our day to day activities generates lots of data.

With the invent of the web, the whole world has gone online, every single thing we do leaves a ***DIGITAL TRACE***.

Major sources of Big Data are social media sites, sensor networks, digital images/videos, cell phones, purchase transaction records, web logs, medical records, archives, military surveillance, eCommerce, complex scientific research & so on.

All these information amounts to around some ***QUINTILLION*** bytes of data.

By 2022, the data volumes will be around ***45 Zettabytes*** which is equivalent to adding every single grain of sand on the planet multiplied by 80.

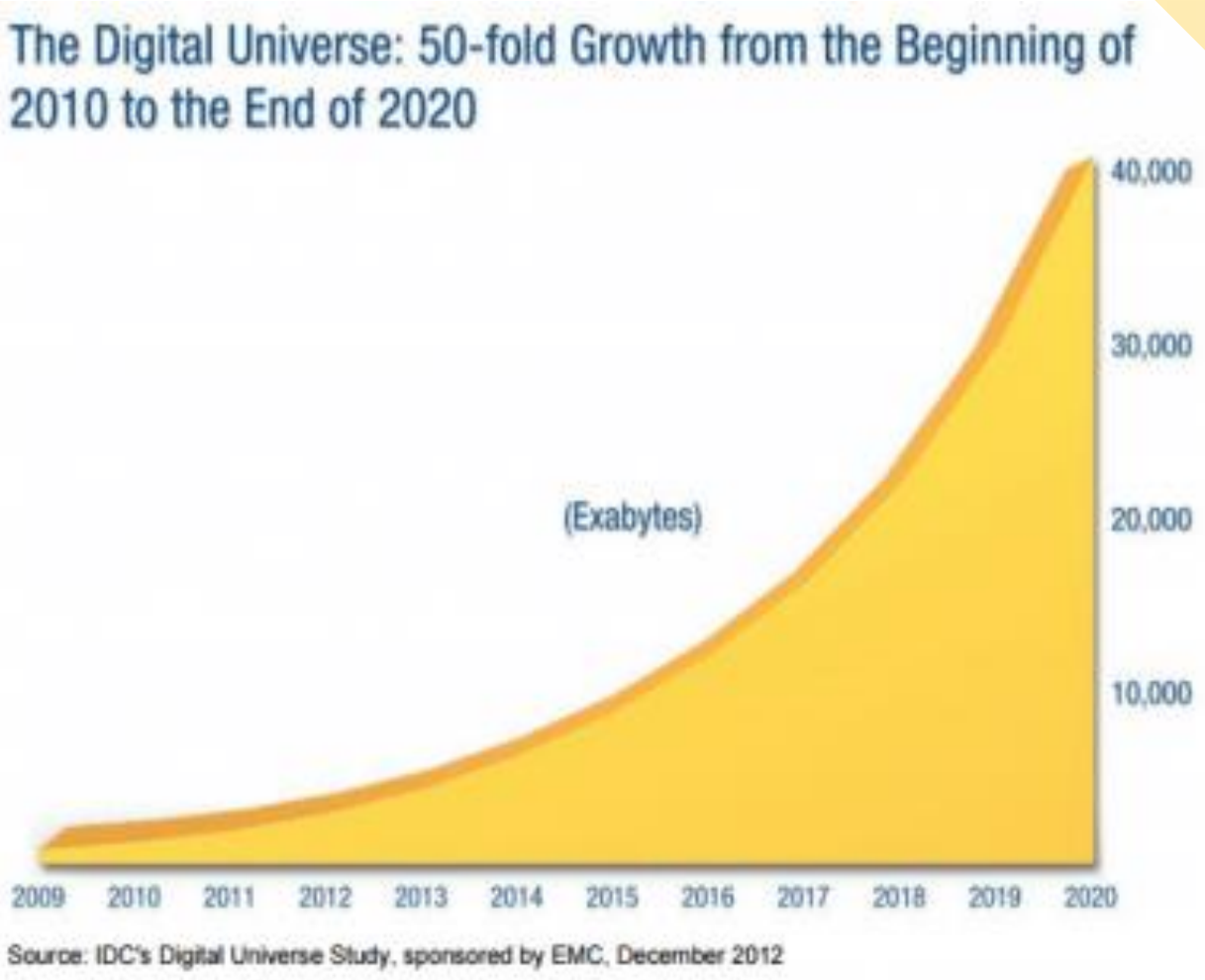
What is Big Data?

- Big Data is a term used for a ***collection*** of data sets that are large & complex, which is difficult to store & process using available database management tools or traditional data processing applications.
- The ***challenge*** includes capturing, curating, storing, searching, sharing, transferring, analyzing & visualization of this data.
- **80%** of the data getting generated today is unstructured and cannot be handled by our traditional technologies.
- It is saying that- *"**An image is a worth of thousand words**"*.

Big Data Characteristics

1. VOLUME

- Volume refers to the '***amount of data***', which is growing day by day at a very fast pace.
- The size of data generated by humans, machines & their interactions on social media itself is massive.
- Researchers have predicted that 45 Zettabytes (45,000 Exabytes) will be generated by 2020, which is an increase of ***300 times*** from 2005.



Big Data Characteristics

2. VELOCITY

- Velocity is defined as the ***PACE*** at which different sources generate the data every day.
- This flow of data is massive & continuous.
- There are ***1.03 billion*** Daily Active Users (Facebook DAU) on Mobile as of now, which is an increase of **22%** year-over-year.
- This shows ***how fast*** the number of ***users are growing*** on social media and how fast the data is getting generated daily.
- If you are able to handle the velocity, you will be able to generate insights and take **DECISIONS** based on real-time data.

Big Data Characteristics

3. VARIETY

- As there are many sources which are contributing to Big Data, the **TYPE OF DATA** they are generating is different.
- It can be ***structured, semi-structured or unstructured***.
- Earlier, we used to get the data from excel & databases, now the data are coming in the form of ***images, audios, videos, sensor data etc.*** as shown in the image.
- Hence, this variety of unstructured data creates ***problems*** in capturing, storage, mining & analyzing the data.

Big Data Characteristics

4. VERACITY

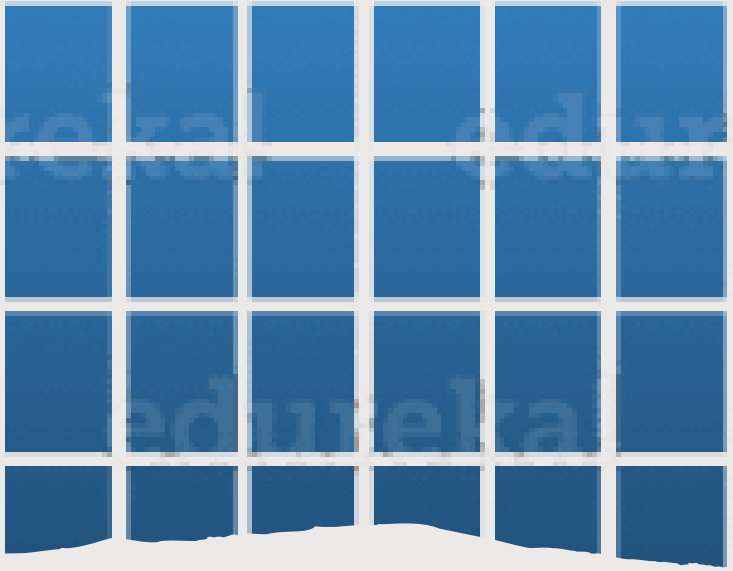
- Veracity refers to the data in ***doubt /uncertainty*** of data available due to data inconsistency & incompleteness.
- In the image shown, you can see that few values are ***missing*** in the table.
- Also, a few values are ***hard to accept***, for example – 15000 minimum value in the 3rd row, it is not possible.
- This ***inconsistency*** & ***incompleteness*** is Veracity.
- Hence, maybe ***difficult to trust***.
- With many forms of big data, ***quality & accuracy*** are difficult to control like Twitter posts with hashtags, abbreviations, typos and colloquial speech.
- Due to uncertainty of data, ***1 in 3*** business leaders don't trust the information they use to make decisions.
- It was found in a survey that ***27%*** of respondents were unsure of how much of their data was inaccurate.
- Poor data quality costs the US economy around ***$3.1 trillion*** a year.

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

Big Data Characteristics

5. VALUE

- After discussing Volume, Velocity, Variety & Veracity, there is another V that should be taken into account when looking at Big Data i.e. VALUE.
- It is all well & good to have access to big data but unless we can turn it into ***value*** it is useless.
- ***By turning it into value*** I mean, Is it ***adding*** to the benefits of the organizations who are analyzing big data?
- Is the organization working on Big Data achieving ***high ROI*** (Return On Investment)?
- Unless, it ***adds to their profits*** by working on Big Data, it is useless.



Types of Big Data

Big Data could be of **3** types:

- Structured
- Semi-Structured
- Unstructured

Types of Big Data

1.STRUCTURED

The data that can be stored & processed in a ***fixed format*** is called as Structured Data.

Data stored in a relational database management system (***RDBMS***) is one example of ‘structured’ data.

It is ***easy to process*** structured data as it has a fixed schema.

Structured Query Language (SQL) is often used to manage such kind of Data.

Types of Big Data

2. SEMI-STRUCTURED

It is a type of data which ***does not*** have a formal structure of a data model.

For example, a table definition in a relational DBMS, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that makes it easier to analyze.

XML files/ JSON documents are examples.

Types of Big Data

3. UNSTRUCTURED

The data which have ***unknown*** form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format.

Text Files & multimedia contents like images, audios, videos are examples of unstructured data.

The unstructured data is growing quicker than others, experts say that ***80%*** of the data in an organization are unstructured.

Examples of Big Data

- Daily we upload millions of bytes of data. **90%** of the world's data has been created in last 2 years.
- Walmart handles more than **1 million** customer transactions every hour.
- Facebook stores, accesses & analyzes **30+ Petabytes** of user generated data.
- **230+ millions** of tweets are created every day.
- More than **5 billion** people are calling, texting, tweeting & browsing on mobile phones worldwide.
- YouTube users upload **48 hours** of new video every minute of the day.
- Amazon handles **15 million** customer click stream user data per day to recommend products.
- **294 billion** emails are sent every day. Services analyses this data to find the spams.
- Modern cars have close to **100 sensors** which monitors fuel level, tire pressure etc. , each vehicle generates a lot of sensor data.

Applications of Big Data

SMARTER HEALTHCARE: Making use of the petabytes of ***patient's data***, the organization can extract meaningful information & then build applications that can ***predict*** the patient's deteriorating condition in advance.

TELECOM: Telecom sectors collects information, analyzes it & provide solutions to different problems. By using Big Data, companies have been able to significantly ***reduce*** data packet loss, which occurs when networks are overloaded, and thus, providing a seamless connection to their customers.

RETAIL: Retail has some of the ***tightest margins***, and is one of the greatest beneficiaries of big data. The beauty of using big data in retail is to understand consumer behavior. Amazon's recommendation engine provides suggestion based on the browsing history of the consumer.

Applications of Big Data

TRAFFIC CONTROL: Traffic congestion is a major challenge for many cities globally. Effective use of data & sensors will be key to ***managing*** traffic better as cities become increasingly densely populated.

MANUFACTURING: Analyzing big data can reduce component defects, improve product quality, increase efficiency, and save time & money.

SEARCH QUALITY: Every time we are extracting information from google, we are ***simultaneously generating*** data for it. Google stores this data & uses it to improve its search quality.

Applications of Big Data

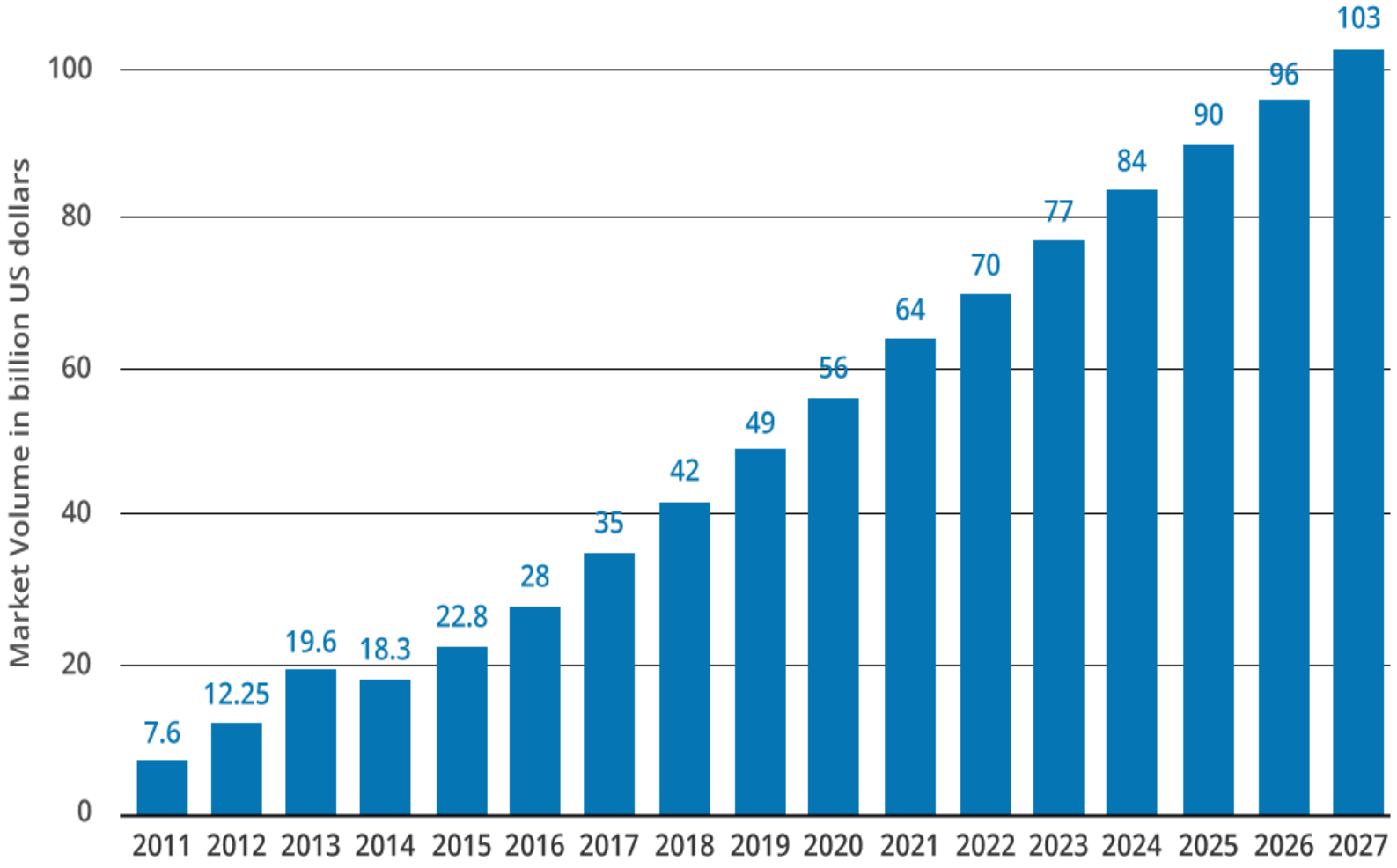
EDUCATION: Here, Big Data Analytics is being absorbed slowly & gradually. Opting for big data powered technology as a learning tool instead of traditional lecture methods, enhanced the learning of students as well as aided the teachers to track their performance better.

INSURANCE: These companies use big data analytics for risk assessment, fraud detection, marketing, customer insights, customer experience & more.

FINANCE: Banks & financial services firms use analytics to differentiate ***fraudulent*** interactions from legitimate business transactions. The analytics systems suggest immediate actions, such as blocking irregular transactions, which stops fraud before it occurs & improves profitability.

Trends in Big Data Analytics

The image depicts the **market revenue of Big Data** in billion U.S. dollars from the year 2011 to 2027.



Facts & Statistics by Forbes

50% respondents report fundamental change in their business and marketing functions

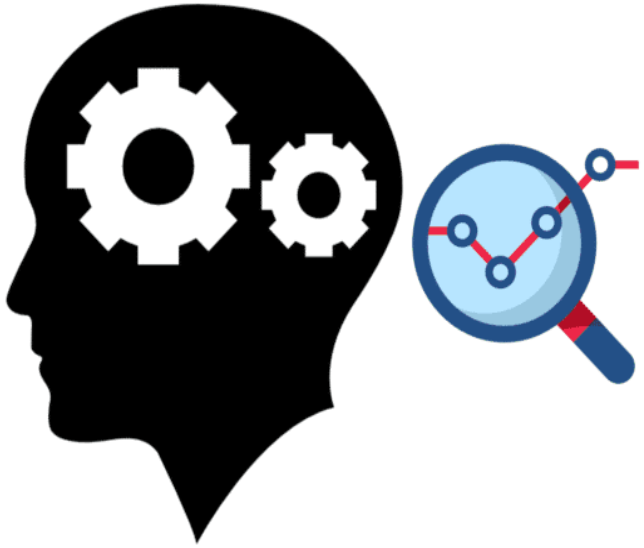
Accenture reveals that 79% of enterprise executives claim that companies that do not embrace Big Data will face extinction

$5.3 billion in 2018

$19.4B in 2026,

Decreasing expenses by 49.2% and creating new avenues for innovation by 44.3%

Starbucks uses behavioural analytics to cater to its customers



Starbucks gather a lot of info about their customers' coffee-buying habits from their preferred drinks to what time of day they're usually ordering

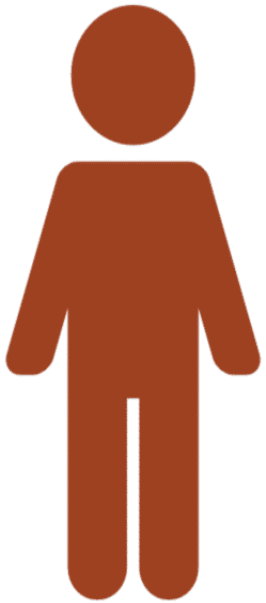
The company directs exciting offers and coupons to their customers and ensures to maintain their interest

Use case: STARBUCKS

Use case: PROCTER&GAMBLE

P&G uses Market Basket Analysis and price optimization to optimize their products

Market Basket Analysis, analyses customer buying habits by finding associations between the different items that customers place in their “shopping baskets”



The company uses simulation models and predictive analysis in order to create the best design for its products.