

INTRODUCTION TO BIG DATA



Table of Contents

- ▶ **Introduction**
- ▶ **How Big is Big Data**
- ▶ **Evolution of Big Data**
- ▶ **What is Big Data**
- ▶ **Characteristics of Big Data (5 V's)**
- ▶ **Types of Big Data**
- ▶ **Example of Big Data**
- ▶ **Challenges with Big Data**
- ▶ **Big Data Analytics & Web Analytics**
- ▶ **Industrial Example of Big Data**
- ▶ **Big Data and Marketing**
- ▶ **Big data and finance service**
- ▶ **Big data healthcare**

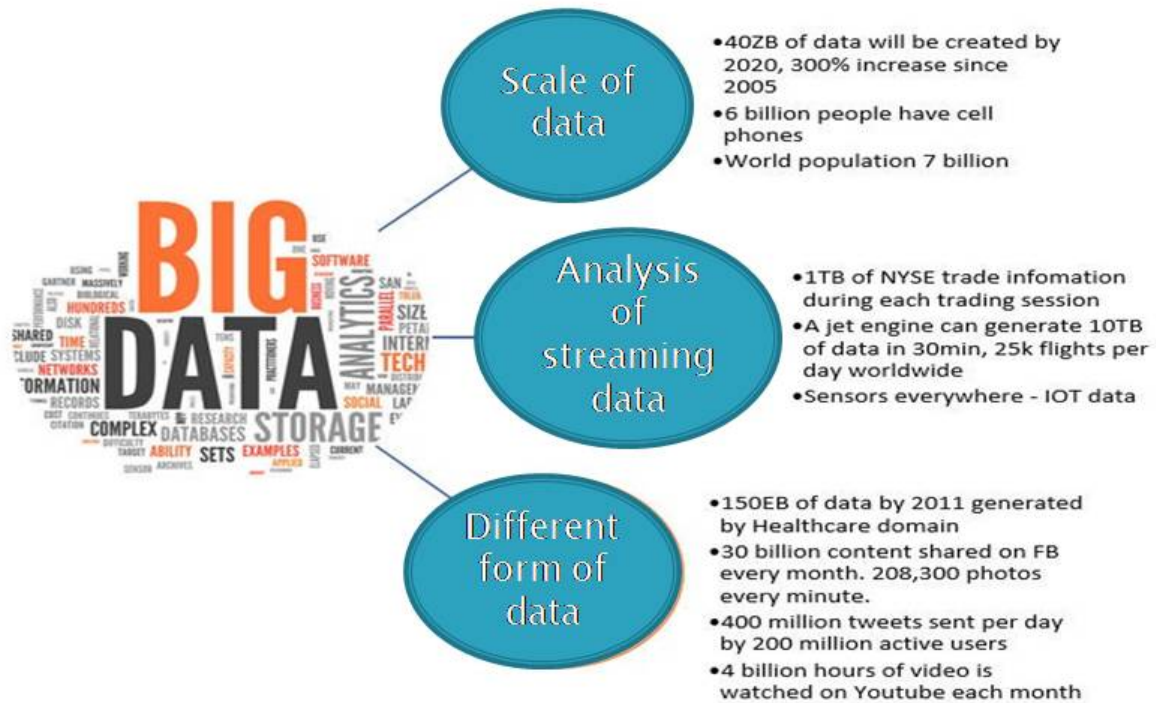


Introduction

- 'Big Data' is similar to 'normal Data', but bigger in size.
- 'Big Data' is also a data but with a huge size. 'Big Data' is a term used to describe collection of data that is huge in size and yet growing exponentially with time.
- This data could be either structured or unstructured



How Big is Big Data



Where is this “Big Data” coming from ?

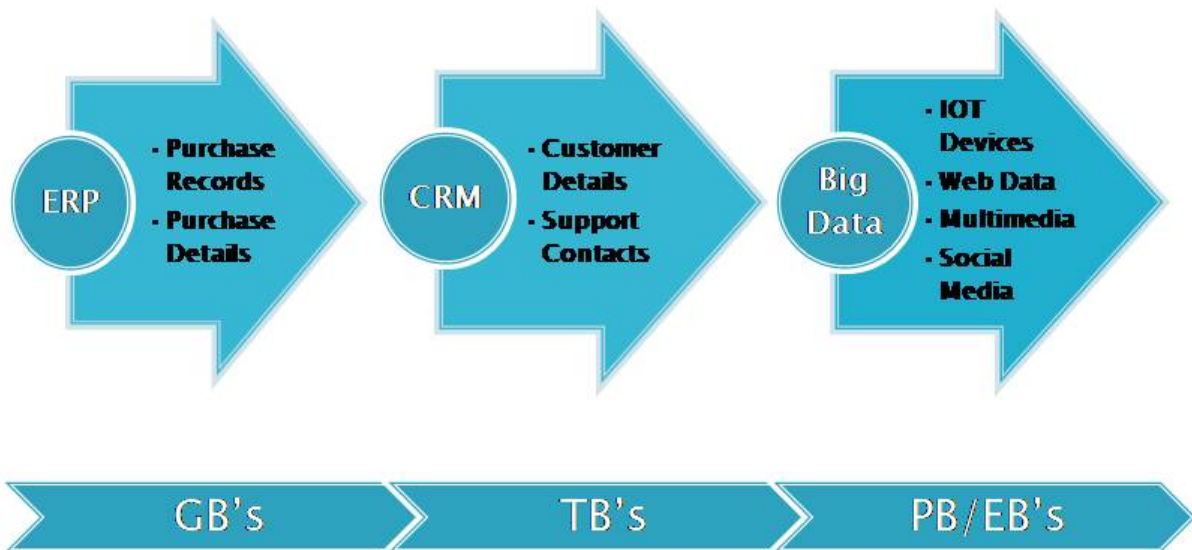


- Walmart handles more than 1–million customer transactions every hour.
- Facebook handles 40 billion photos from its user base.
- Decoding the human genome originally took 5 years to process; now it can be achieved in one week.

Big Data Market Forecast



Evolution of Big Data

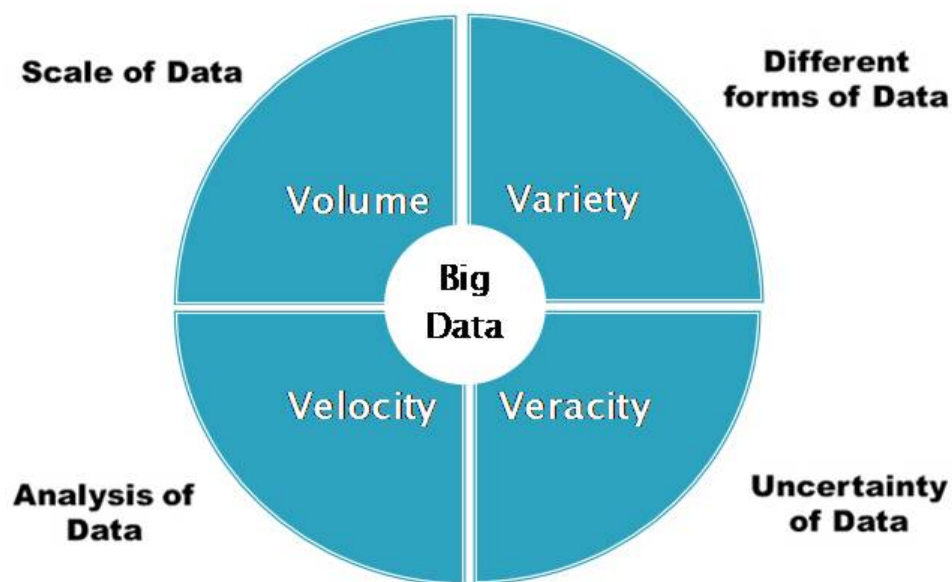


What is Big Data?

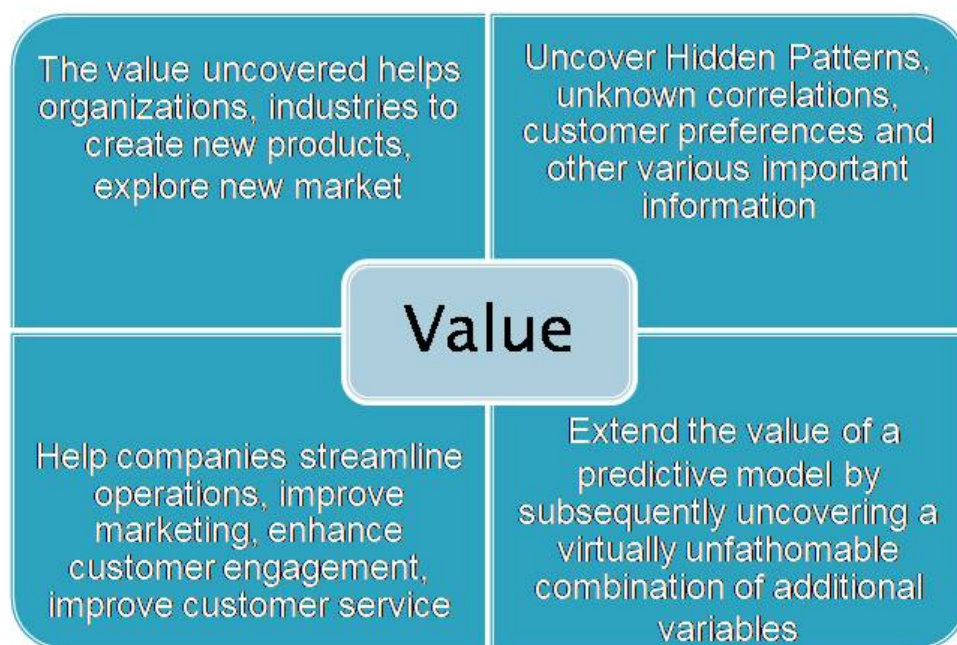
- ▶ Big data is a term used for a **collection of data sets** that are **large and complex**, which is difficult to store and process using available database management tools or traditional data processing applications.
- ▶ The challenge includes capturing, curating, storing, searching, sharing, transferring, analyzing and visualization of this data.



Characteristics of Big Data (The 4 V's)

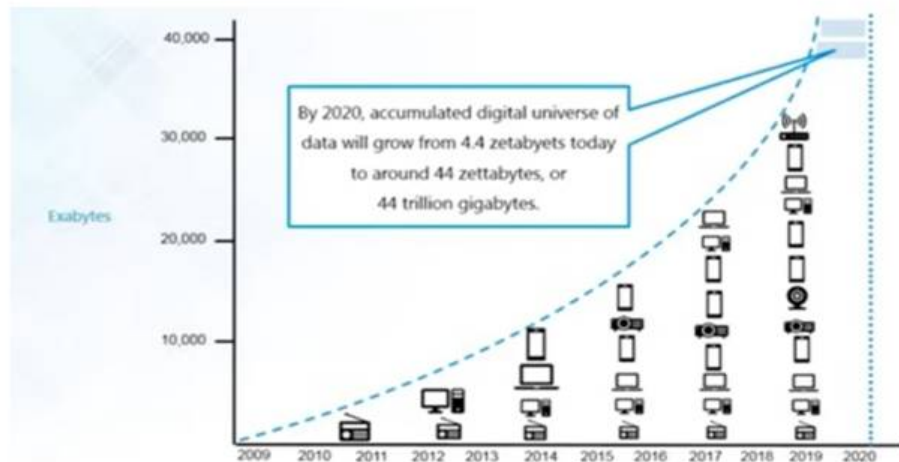


Characteristics of Big Data (The 5 V's)



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 and their interactions on social media itself is massive.



Velocity

Velocity is defined as the pace at which different sources generate the data every day. This flow of data is massive and 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 numbers 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.



Veracity

Veracity refers to the data in doubt or uncertainty of data available due to data inconsistency and incompleteness. In the image below, you can see that few values are missing in the table. Also, a few values are hard to accept, for example - 15000 minimum values in the 3rd row, it is not possible. This inconsistency and incompleteness is Veracity.



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. Hence, there is a variety of data which is getting generated every day. Earlier, we used to get the data from excel and databases, now the data are coming in the form of images, audios, videos, sensor data etc. as shown in below image. Hence, this variety of unstructured data creates problems in capturing, storage, mining and analyzing the data.



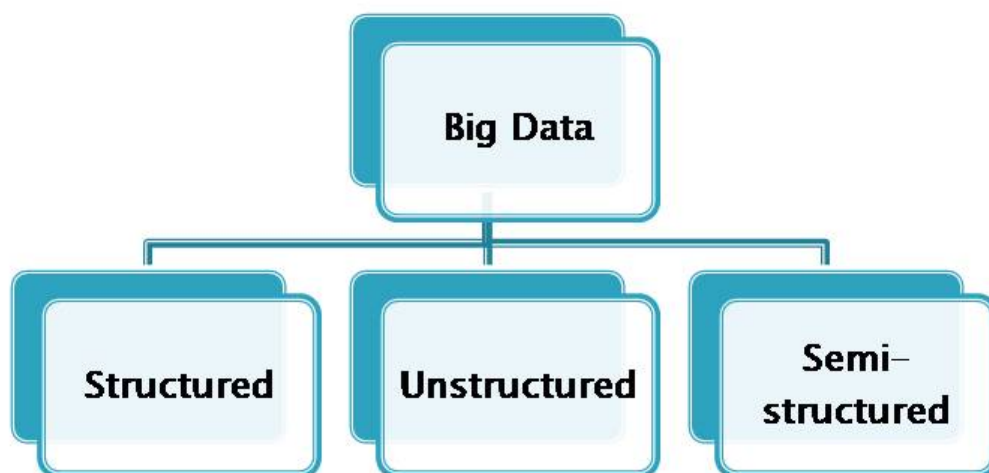
Value

Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights.



Types of data under Big Data

- ▶ Big Data could be of three types:



Structured Data

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

Example of Structured Big Data

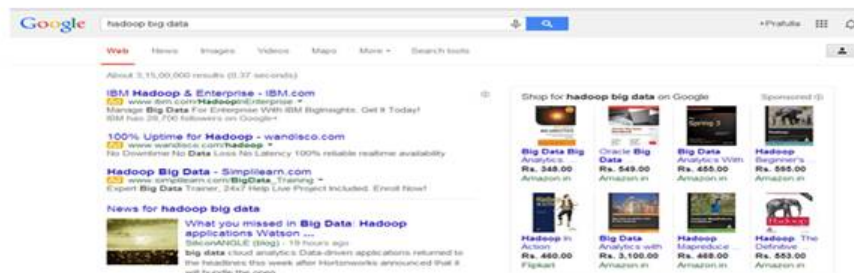
Emp_ID	Name of EMP	Gender	Department	Salary
2345	Rajesh Kulkarni	Male	Finance	40000/-
5642	Kuldeep Singh	Male	Admin	50000/-
8732	Pratibha Singh	Female	Finance	55000/-



Unstructured Data

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

Example of Unstructured Big Data



Semi-structured Data

Semi-structured data can contain both the forms of data. Example of semi-structured data is a data represented in XML file

Example of semi-structured Big Data

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>  
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>  
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>  
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>  
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

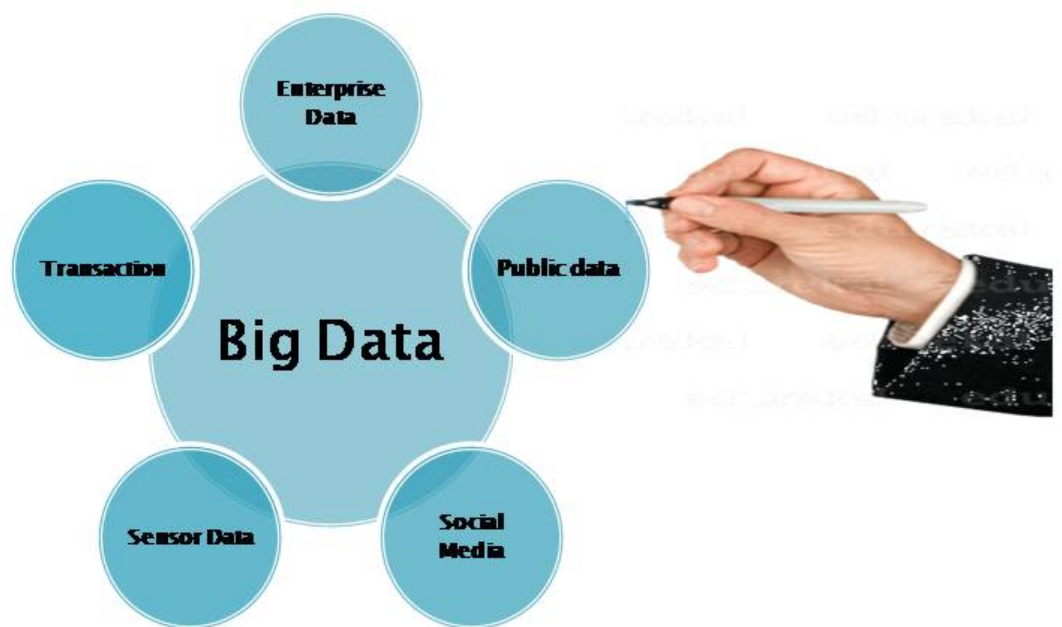


Why Big Data is important?

- 1. Understanding and Targeting Customers**
- 2. Understanding and Optimizing Business Processes**
- 3. Personal Quantification and Performance Optimization**
- 4. Improving Healthcare and Public Health**
- 5. Improving Sports Performance**
- 6. Improving Science and Research**
- 7. Optimizing Machine and Device Performance**
- 8. Improving Security and Law Enforcement**
- 9. Improving and Optimizing Cities and Countries**
- 10. Financial Trading**



Examples of Big Data



A Single View to the Customer



Challenges with Big Data

- 1. Data Quality**
- 2. Discovery**
- 3. Storage**
- 4. Analytics**
- 5. Security**
- 6. Lacks of Talent**



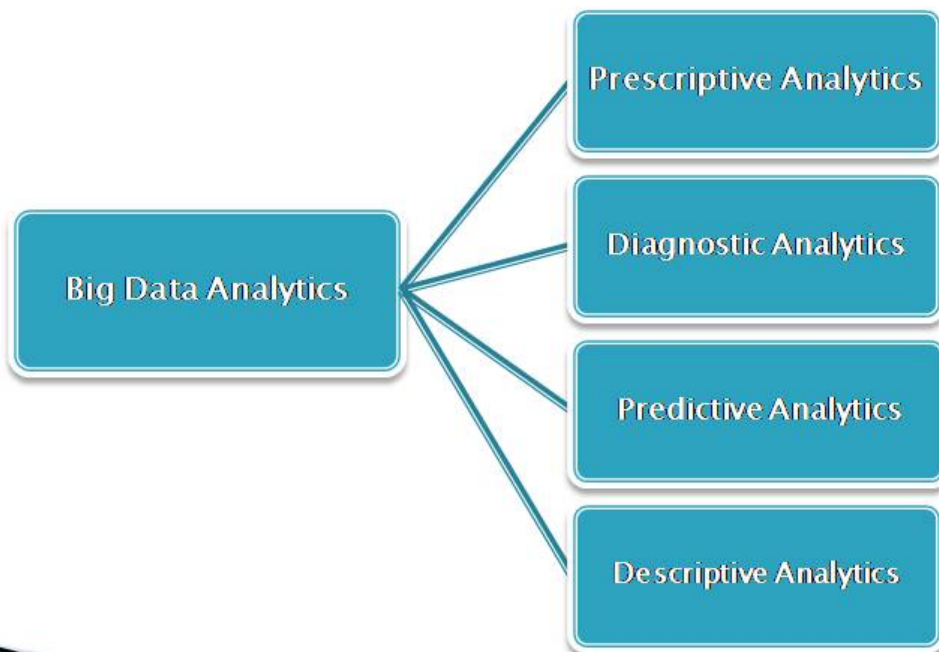
Big Data Analytics

- ▶ Big Data analytics is the process of collecting, organizing and analyzing large sets of data (called Big Data) to discover patterns and other useful information.
- ▶ Big Data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions



Big Data Analytics

- ▶ Big Data Analytics could be of four types:



Benefits of Big Data Analytics

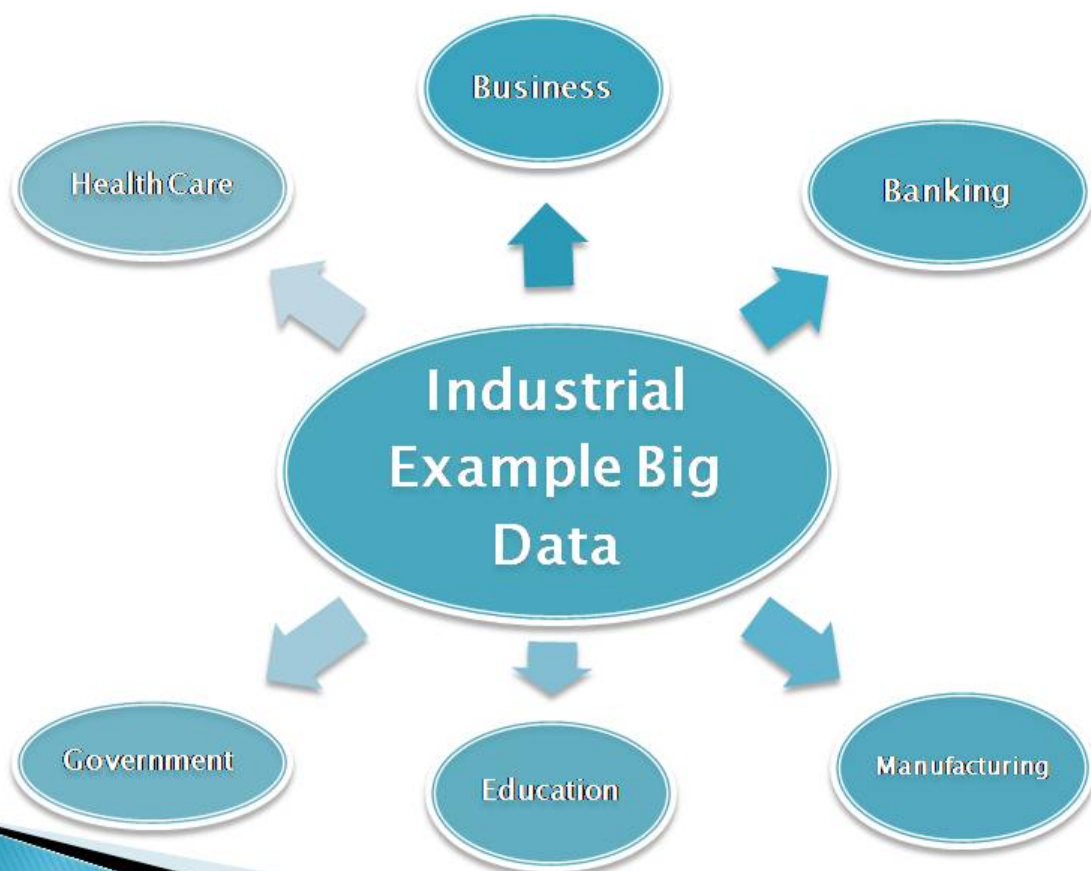


Web Analytics


1. It is the collection, reporting and analysis of websites data.
2. Provides the information about the number of visitors the number of page views.
3. However, Web analytics is not just a process for measuring **web traffic** but can be used as a tool for business and market research, and to assess and improve the effectiveness of a website



Industrial Example of Big Data



Big Data And Marketing

- **Big data is the biggest game-changing opportunity and paradigm shift for marketing since the invention of the phone or the Internet going mainstream.**
 - **Big data refers to the ever-increasing volume, velocity, variety, variability and complexity of information. For marketing organizations, big data is the fundamental consequence of the new marketing landscape, born from the digital world we now live in.**
 - **The term “big data” doesn’t just refer to the data itself; it also refers to the challenges, capabilities and competencies associated with storing and analyzing such huge data sets to support a level of decision-making that is more accurate and timely than anything previously attempted: big data-driven decision-making.**
- 

Three types of big data are key for marketing:

1. **Customer:** The big data category most familiar to marketing may include behavioral, attitudinal and transactional metrics from such sources as marketing campaigns, points of sale, websites, customer surveys, social media, online communities and loyalty programs.
2. **Operational:** This big data category typically includes objective metrics that measure the quality of marketing processes relating to marketing operations, resource allocation, asset management, budgetary controls, etc.
3. **Financial:** Typically housed in an organization's financial systems, this big data category may include sales, revenue, profits and other objective data types that measure the financial health of the organization.



Having big data doesn't automatically lead to better marketing

Organizations that want to succeed in marketing should do the following things well:

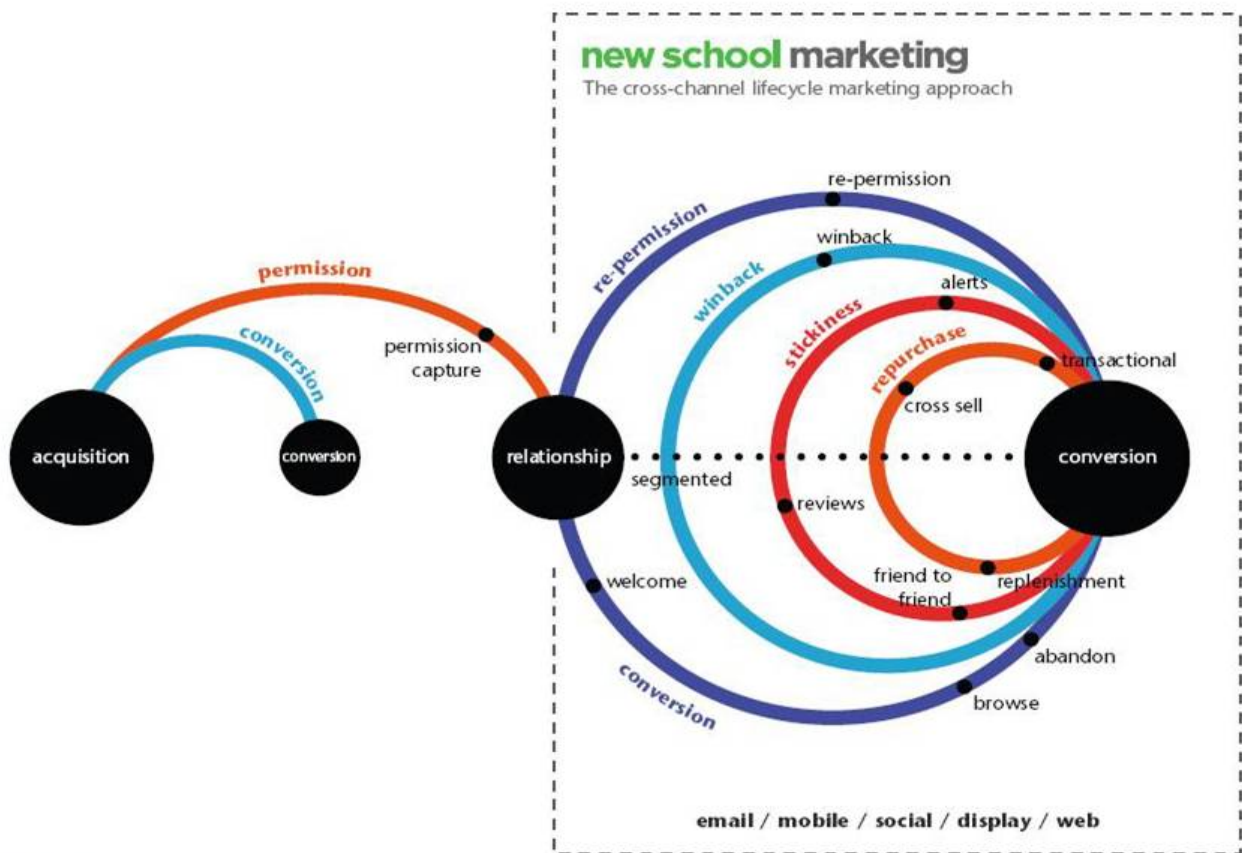
- 1. Successful discovery of new opportunities**
- 2. Understand consumer decision journey.**
- 3. Monitor Google Trends to inform your global/local strategy**
- 4. Create real-time personalization to buyers**
- 5. Identify the specific content that moves buyers down the sales funnel**
- 6. Make it quick and simple**



Big Data and the New School of Marketing

Dan Springer, CEO of Responsys, defines the new school of marketing: "Today's consumers have changed. They've put down the newspaper, they fast forward through TV commercials, and they junk unsolicited email. Why? They have new options that better fit their digital lifestyle. They can choose which marketing messages they receive, when, where, and from whom. They prefer marketers who talk with them, not at them. New School marketers deliver what today's consumers want: relevant interactive communication across the digital power channels: email, mobile, social, display and the web."





Finance Service

Fraud and Compliance

- Cyber attack prevention
- Regulatory compliance
- Criminal behaviour
- Credit Card fraud detection

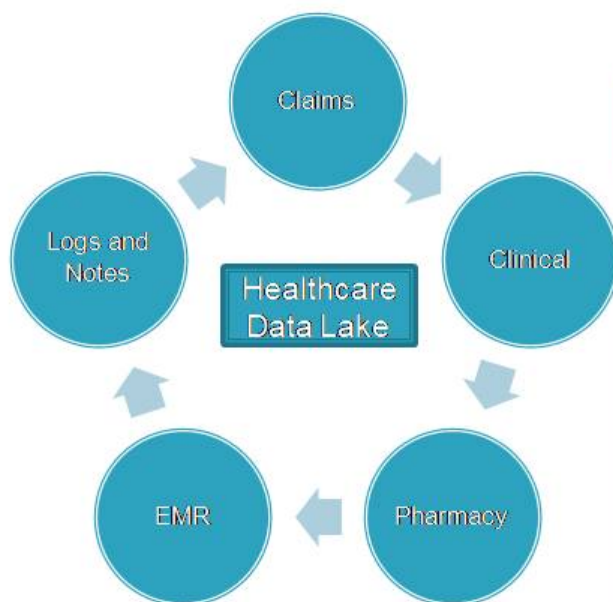
EDW Optimization

- Offload expensive analytics
- Offload expensive data preparation at lower cost
- Data discovery
- Deal with various data types

Risk Management

- Real time risk alerting system
- Analyse credit risk, counter- or third party risk
- Utilizing simulations that use huge volumes of data and require massive parallel computing power

Big Data and Healthcare



Healthcare IOT

- Most of the data is of unstructured variety created by Healthcare IOT's
- Devices monitoring everything of patient from blood sugar level, heart rate, etc.
- Smart devices already in place can detect if medicines are being taken regularly at home
- Lower costs and improve patient care

Reducing Fraud, Waste and Abuse

- Prevent healthcare fraud by using Predictive analytics. Centre for Medicare and Medicaid services prevented \$210.7 million using the same
- Identifying fraud by analyzing large historical unstructured data of historical claims and by using ML algorithms to detect anomalies and patterns

Electronic Health Records (EHRs)

- Every patient has his own digital record which includes demographics, medical history, allergies, laboratory test results etc.
- Records are shared via secure information systems. Every record is comprised of one modifiable file, which means that doctors can implement changes over time with no paperwork and no danger of data replication