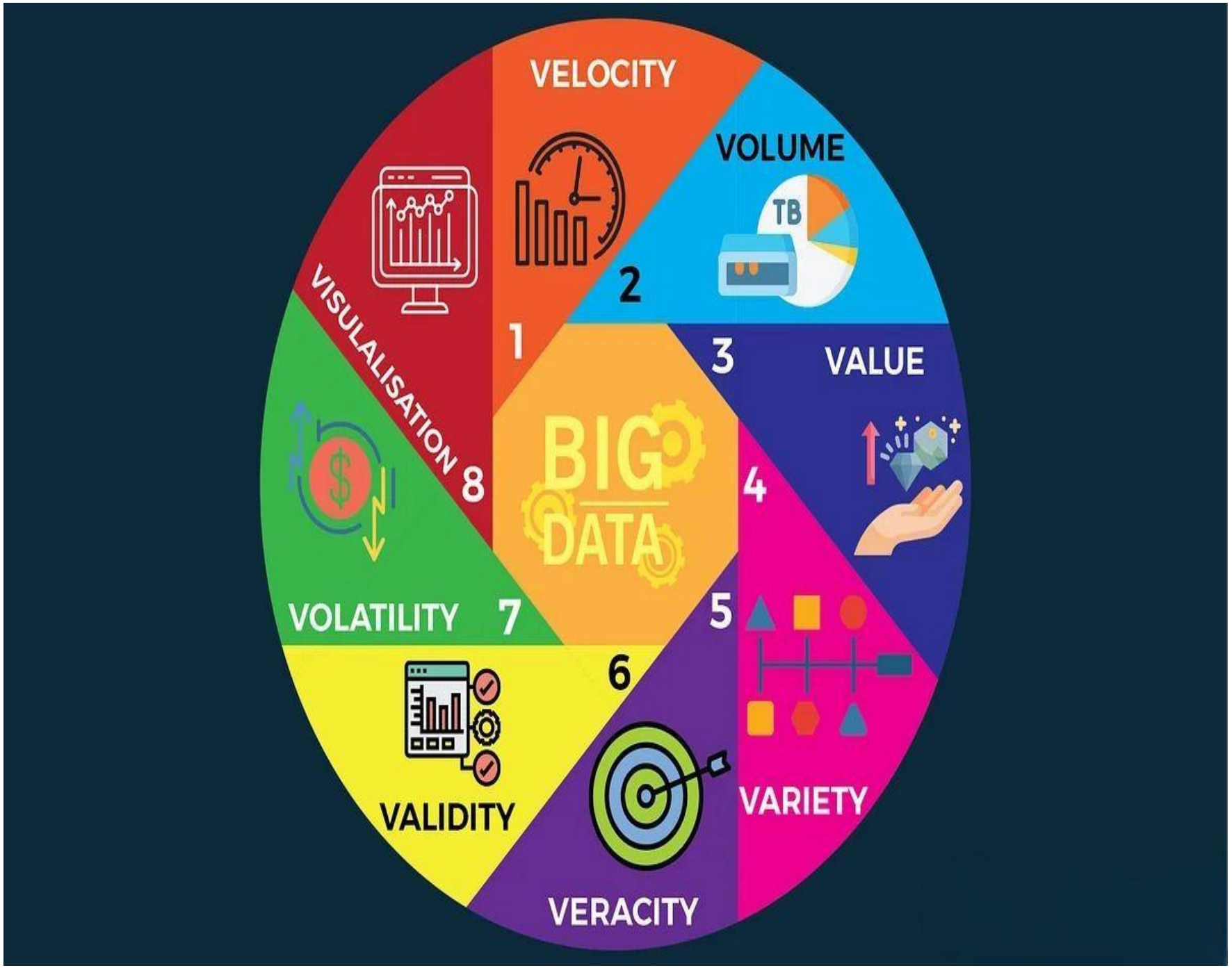
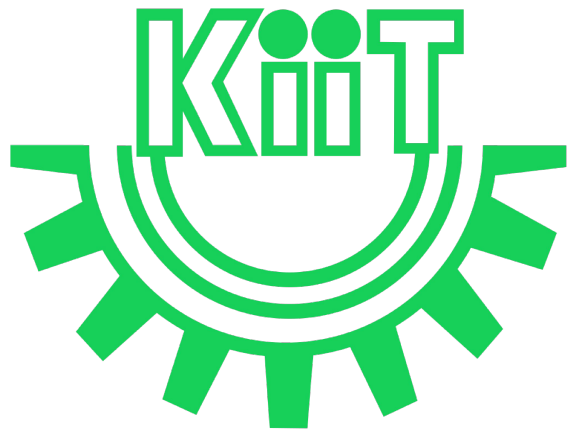


CS 3032: Big Data

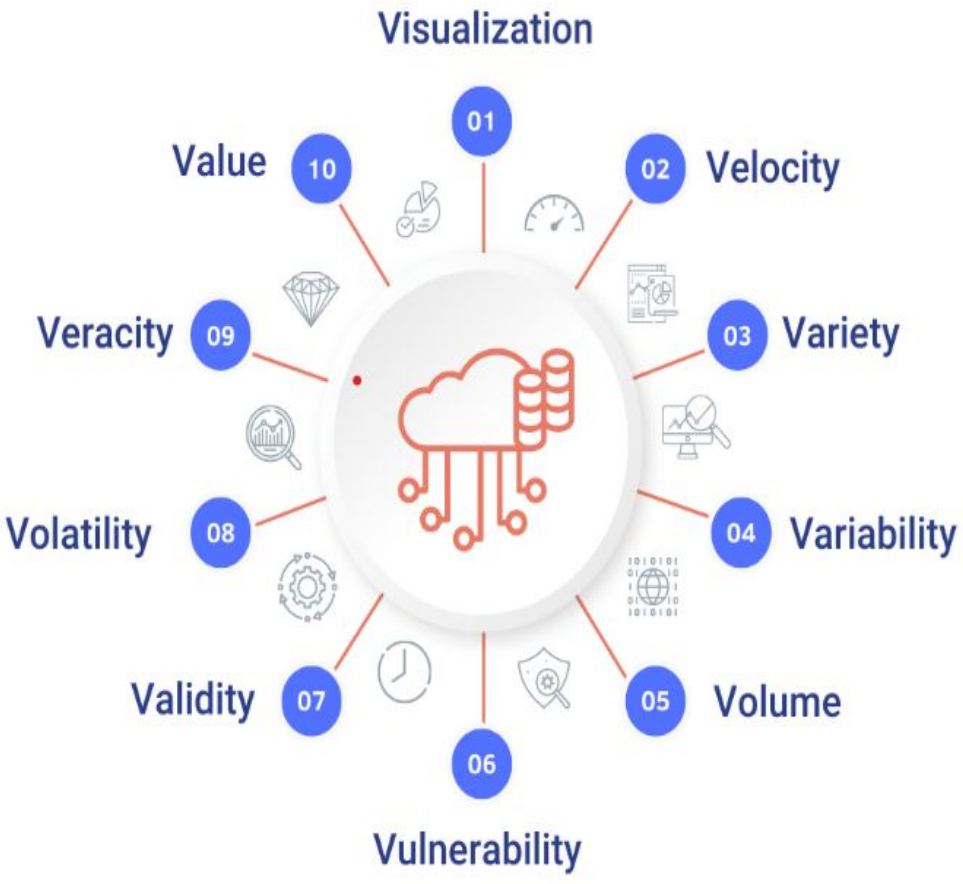
Lec-2

In this Discussion . . .

- Elements of Big Data
- Data Analytics
- Evolution of analytics scalability
- Big Data Analytics



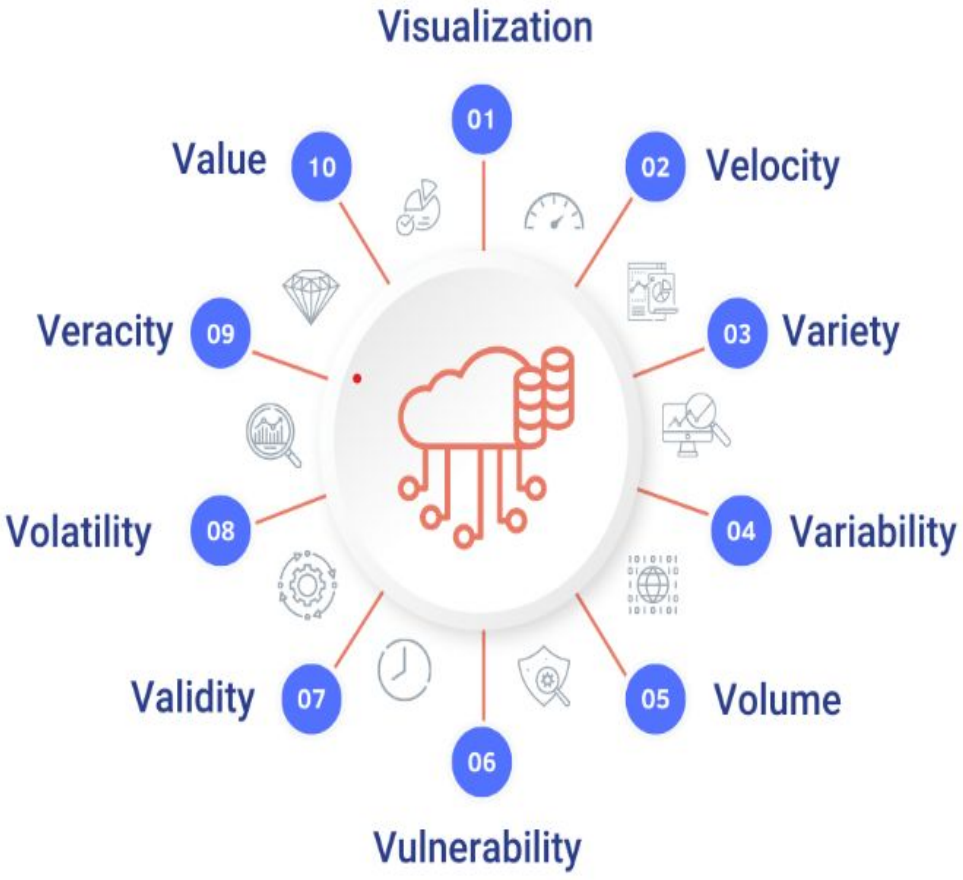
Elements of Big Data



Volume:

- refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, M2M sensors, photographs, video, etc.
- The vast amounts of data have become so large that in fact it can no longer store and perform data analysis using traditional database technology.
- So using distributed systems, where parts of the data is stored in different locations and brought together by software.

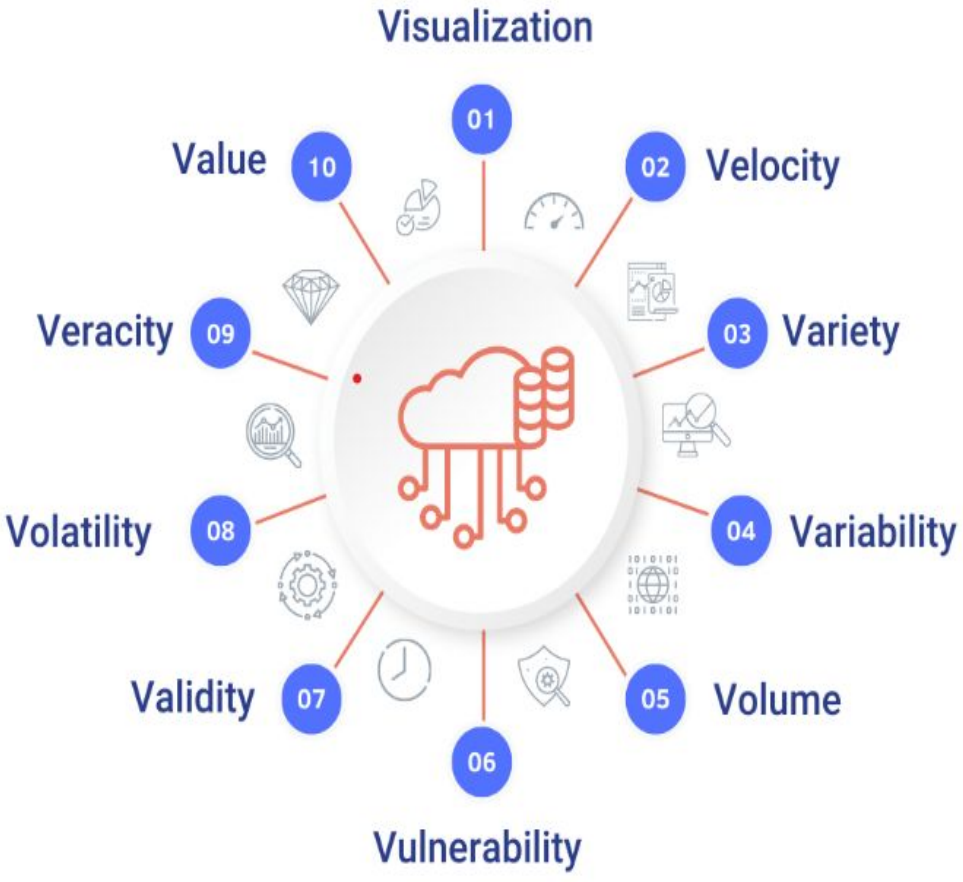
Elements of Big Data



Velocity:

- refers to the speed at which vast amounts of data are being generated, collected and analyzed
- Velocity is important for businesses that need their data to be quickly available for making informed decisions.
- Velocity adds to volume, allowing us to grapple with data as a dynamic quantity.

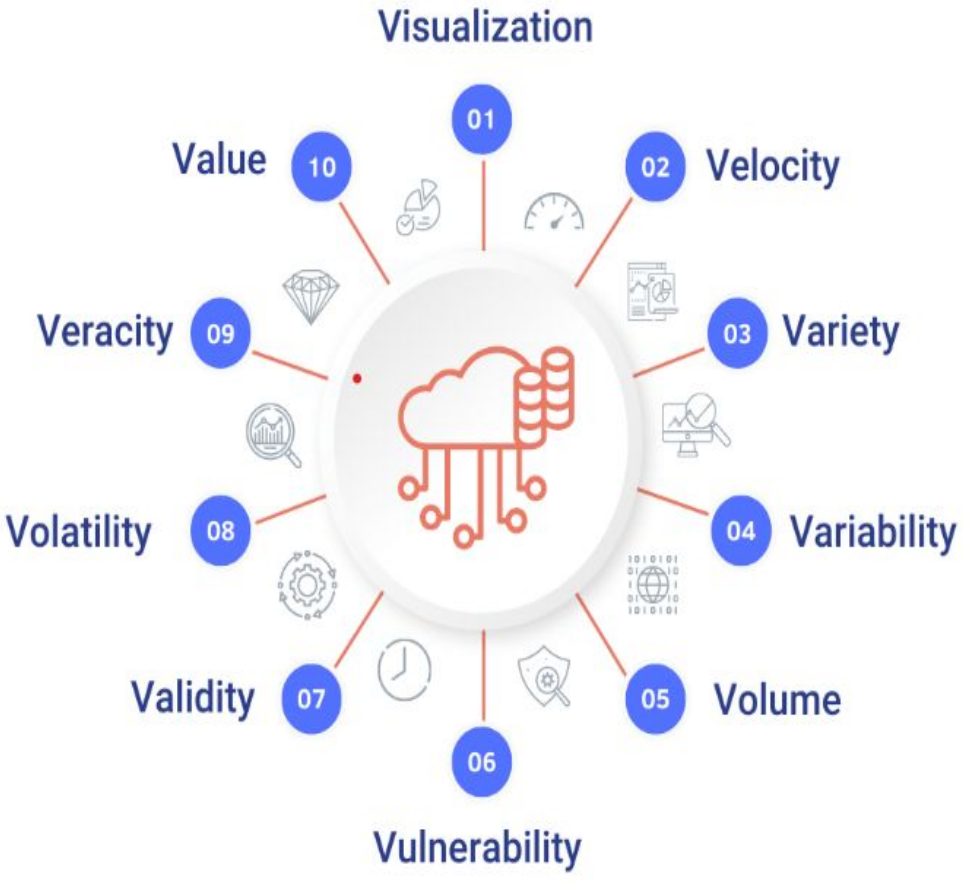
Elements of Big Data



Variety:

- defined as the different types of data the digital system now use.
- In fact, the availability of clean data is among the top challenges facing data scientists.
- According to Forbes, most data scientists spend 60% of their time cleaning data.

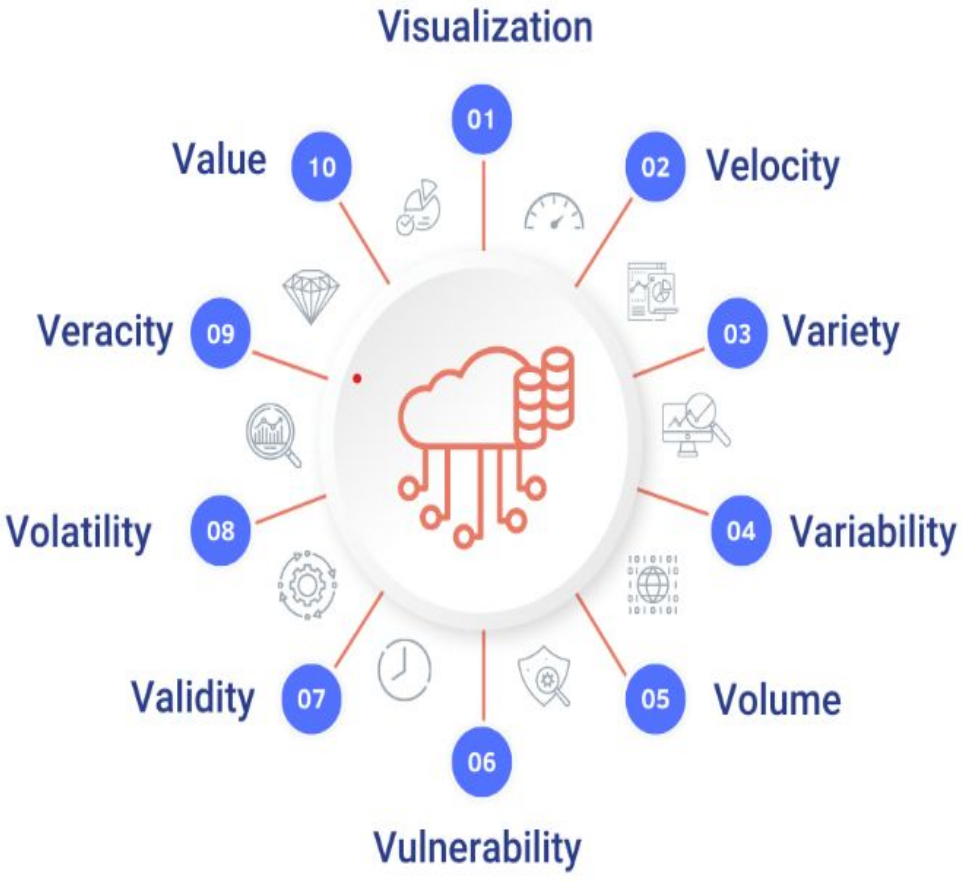
Elements of Big Data



Veracity:

- refers to the quality or trustworthiness of the data or the data source.
- Numerous factors can contribute to the reliability of the input they provide at a particular time in a particular situation.
- Veracity is particularly important for making data-driven decisions for businesses as reproducibility of patterns relies heavily on the credibility of initial data inputs.

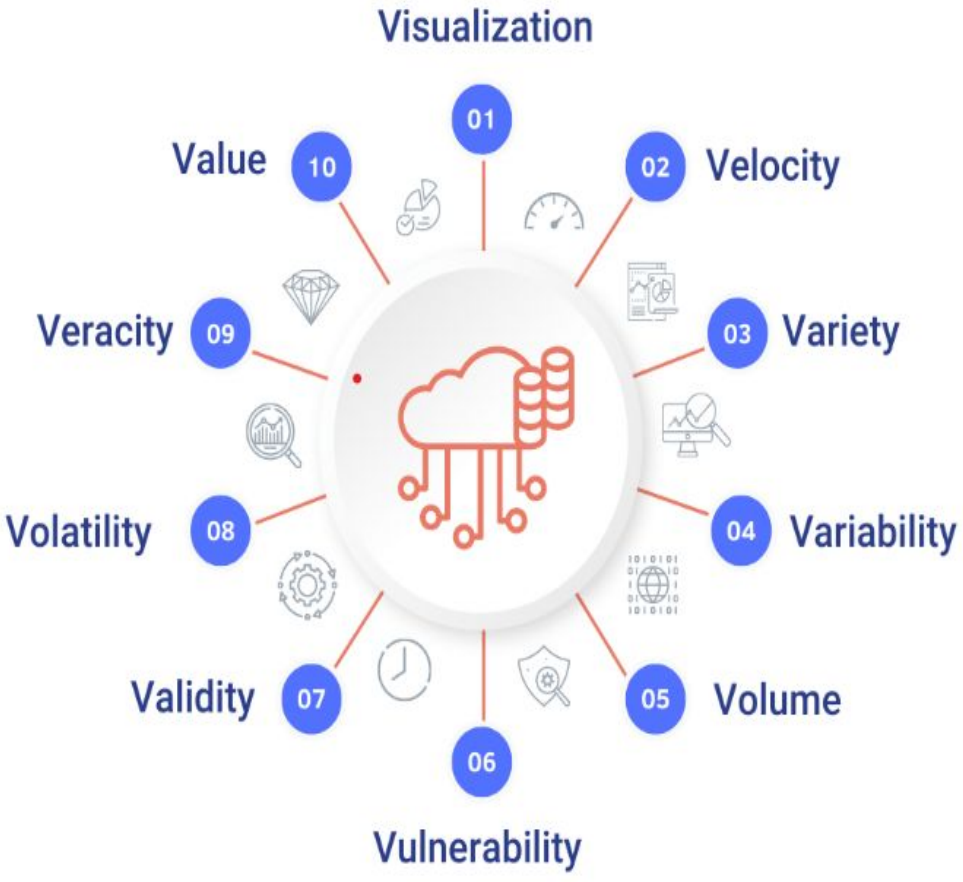
Elements of Big Data



Variability:

- Is a measure of the inconsistencies in data and is often confused with variety.
- This kind of inconsistency in data is an important feature as it places limits on the reproducibility of data.
- Variability also accounts for the inconsistent speed at which data is downloaded and stored across various systems, creating a unique experience for customers consuming the same data.

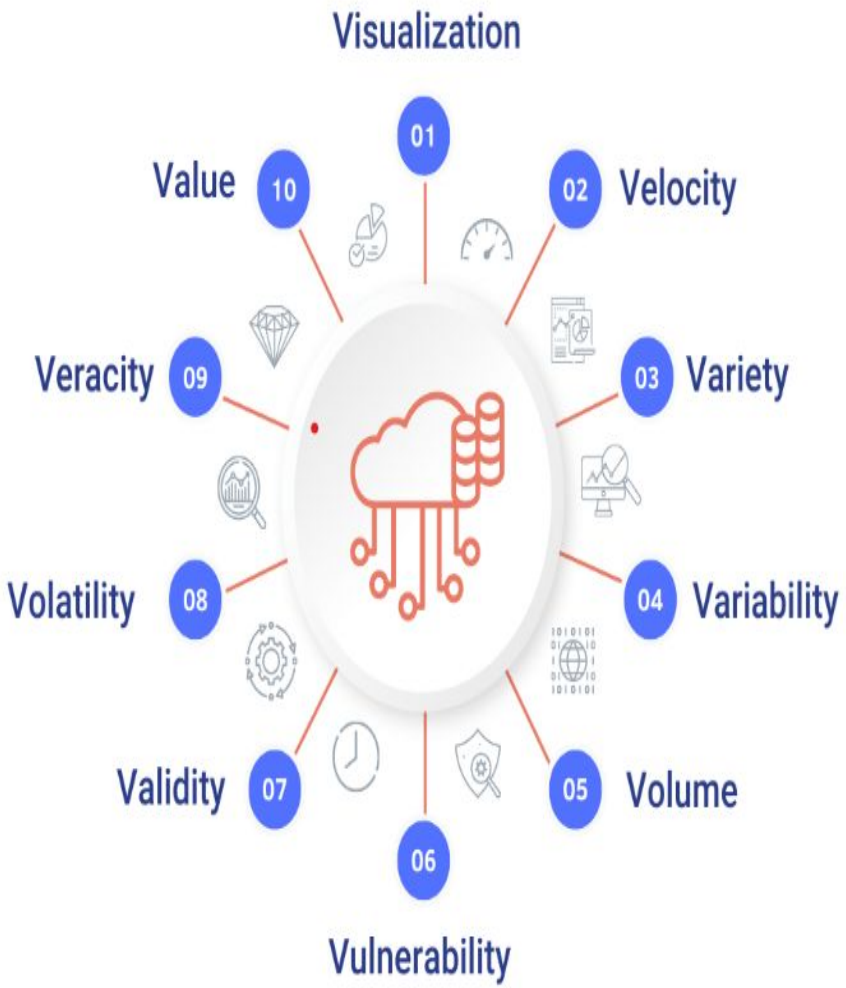
Elements of Big Data



Validity:

- Validity pertains to the accuracy of data for its intended use.
- For example- We may acquire a dataset pertaining to data related to your subject of inquiry, increasing the task of forming a meaningful relationship and inquiry. Registered charity data contact lists

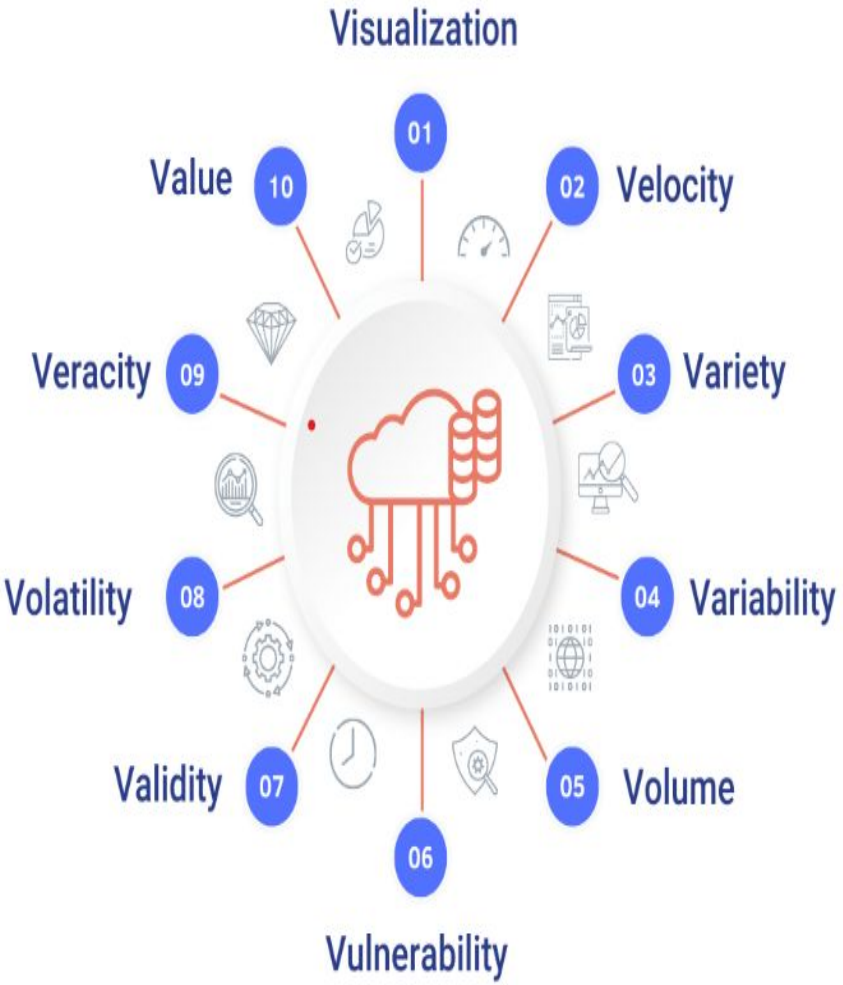
Elements of Big Data



Volatility:

- Refers to the time considerations placed on a particular data set.
- It involves considering if data acquired a year ago would be relevant for analysis for predictive modeling today.
- This is specific to the analyses being performed. Similarly, volatility also means gauging whether a particular data set is historic or not.
- Usually, data volatility comes under data governance and is assessed by data engineers.

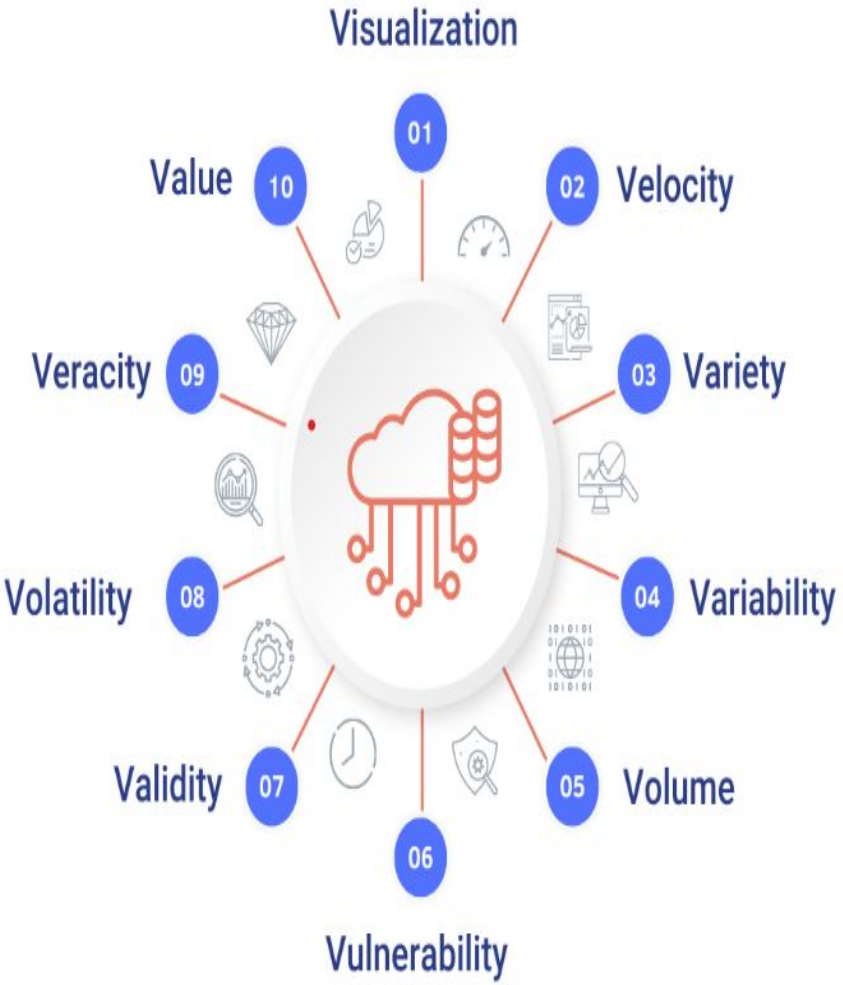
Elements of Big Data



Vulnerability:

- Big data is often about consumers.
- We often overlook the potential harm in sharing our shopping data, but the reality is that it can be used to uncover confidential information about an individual.
- For instance, Target accurately predicted a an innocent person's guilt for a murder before it even happened.
- To avoid such consequences, it's important to be mindful of the information we share online.

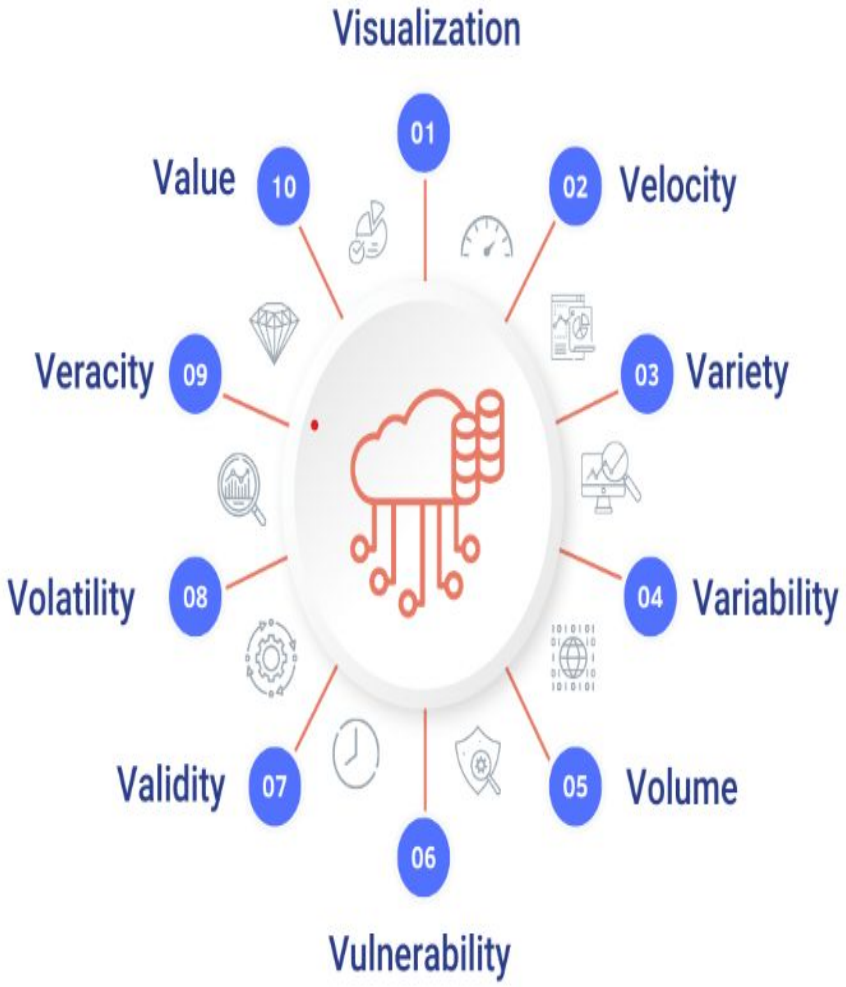
Elements of Big Data



Visualization:

- With a new data visualization tool being released every month or so, visualizing data is key to insightful results.
- The traditional x-y plot no longer suffices for the kind of complex detailing that goes into categorizations and patterns across various parameters obtained via big data analytics.

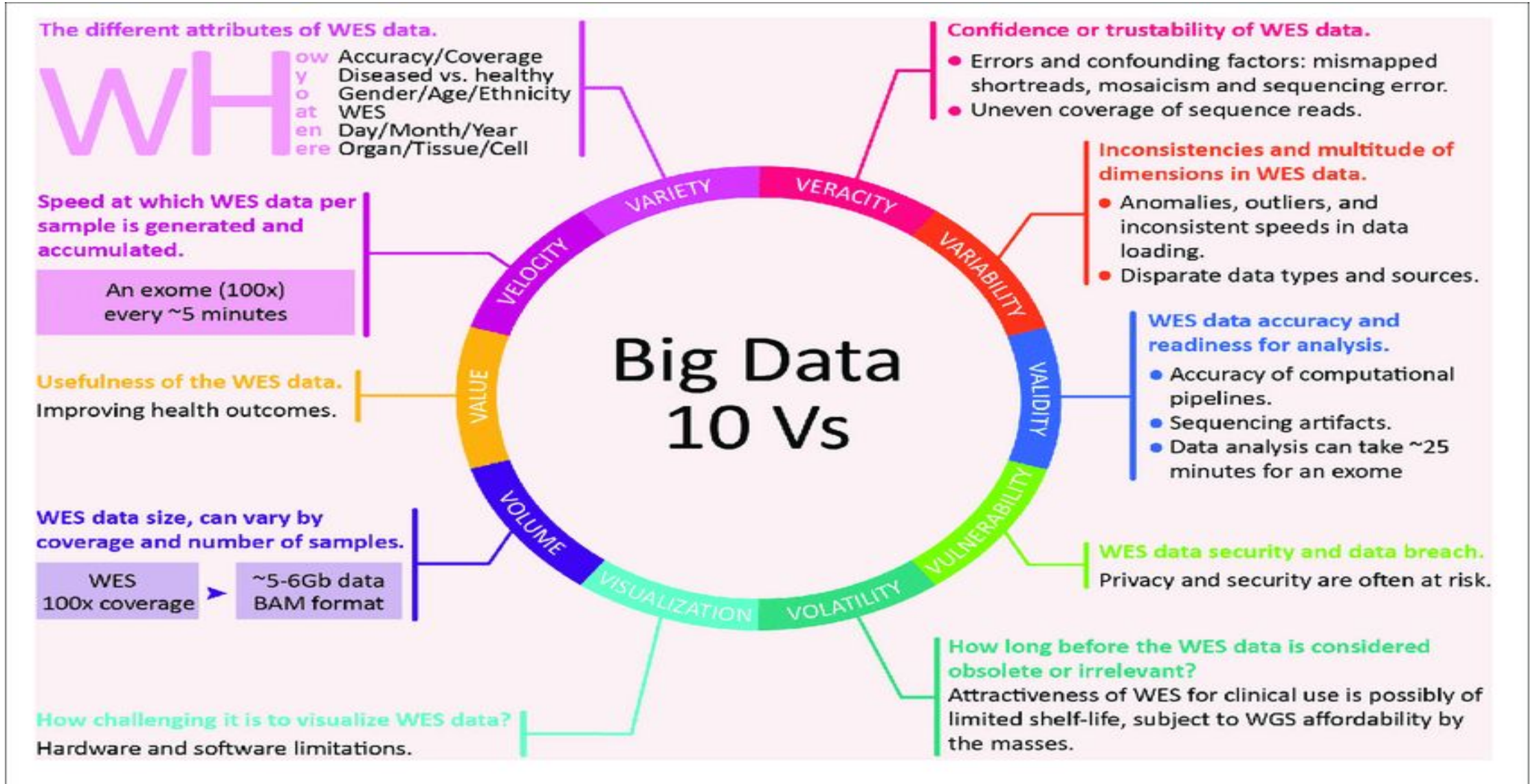
Elements of Big Data



Value:

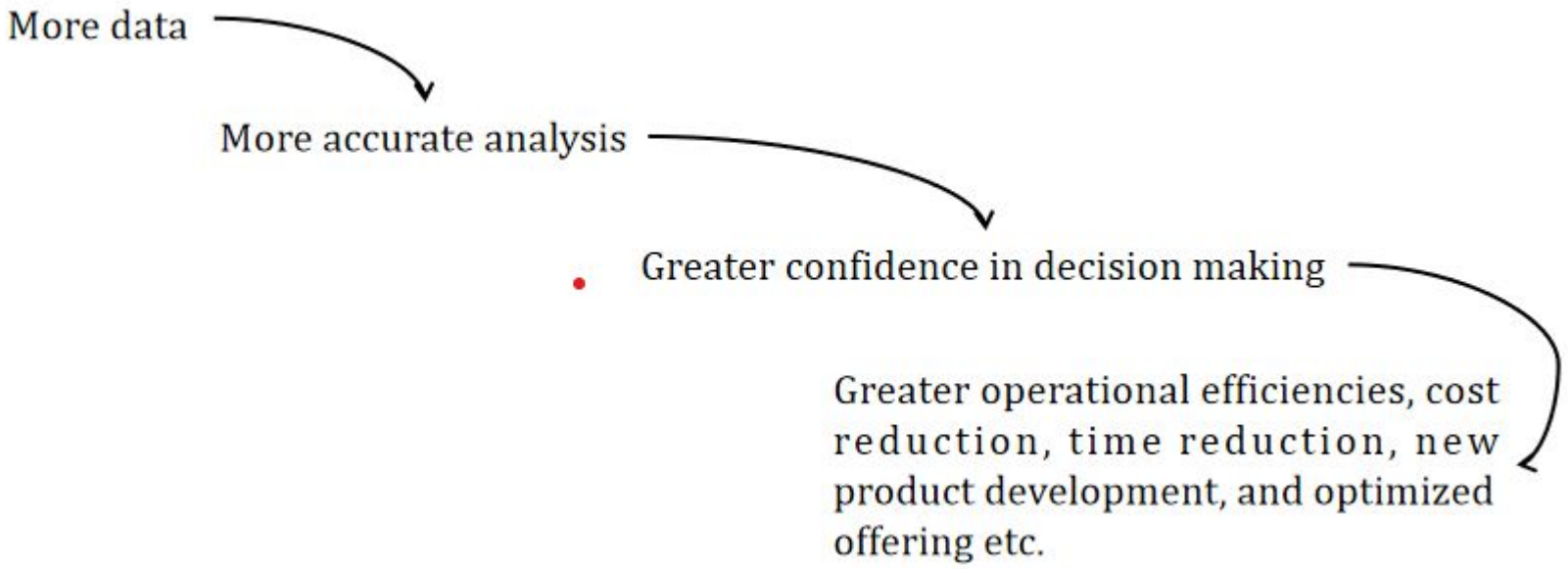
- BIG data is nothing if it cannot produce meaningful value.
- Consider, again, the example of Target using a 16-year-old's shopping habits to predict his theft.
- While in this case, it violates privacy, in most other cases, it can generate incredible customer value by bombarding them with the specific product advertisement they require.

Elements of Big Data (An Example with reference to exome sequencing)



Why Big Data ?

- More data for analysis will result into greater analytical accuracy and greater confidence in the decisions based on the analytical findings.
- This would entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services and optimizing existing services.



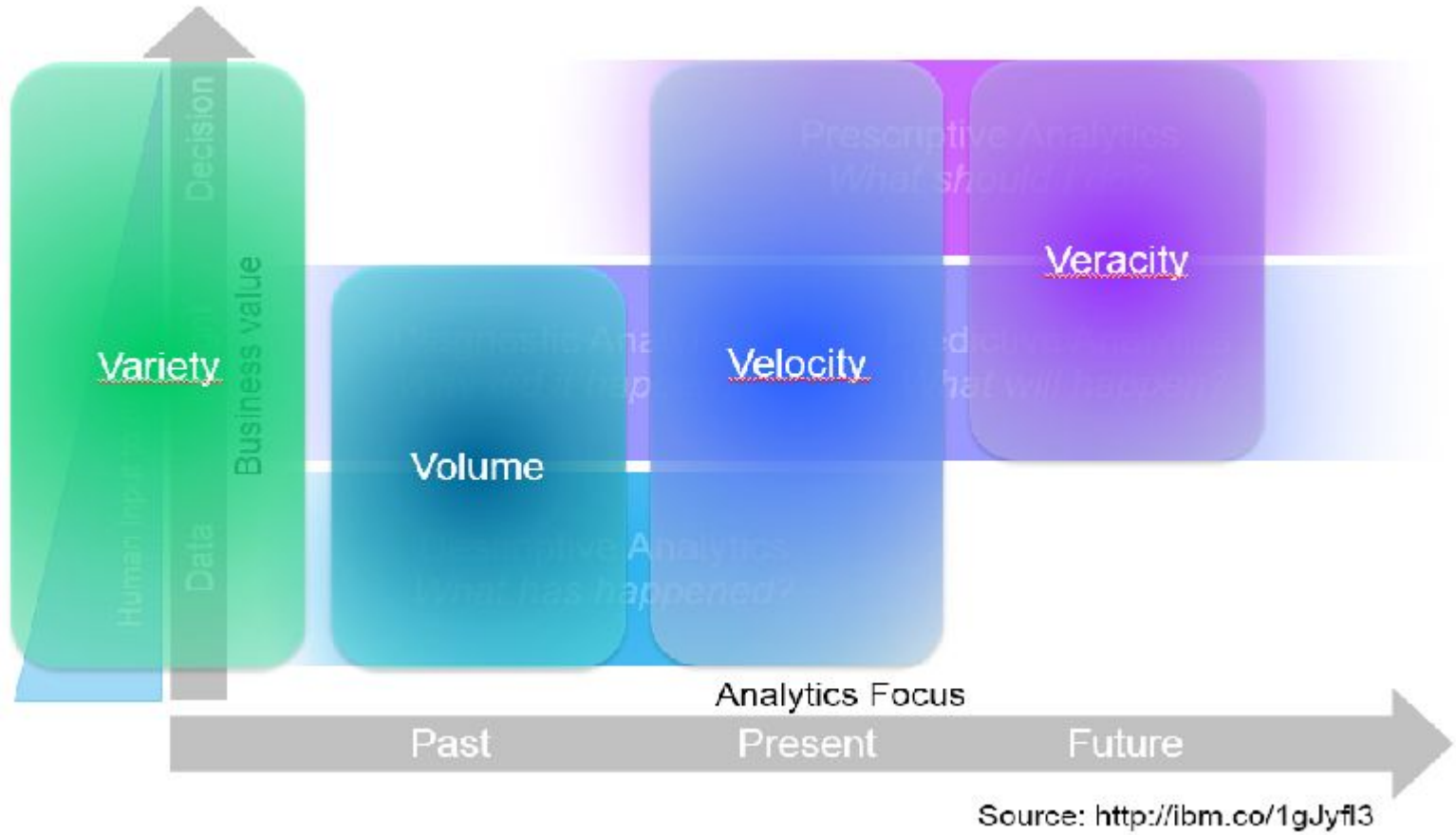
Data Analytics

- **Data analytics** describe an advanced scientific field wherein analysts collect raw data from the past and draw inferences meaningfully for proper action about the information contained.
- They use various statistical tools, machine learning, and other technical tools.
- Companies further use the inferences to perform smart business decisions.

Data Analytics Types : 4 Types

Approach	Explanation
Descriptive	What's happening in my business? • Comprehensive, accurate and historical data • Effective Visualisation
Diagnostic	Why is it happening? • Ability to drill-down to the root-cause • Ability to isolate all confounding information
Predictive	What's likely to happen? • Decisions are automated using algorithms and technology • Historical patterns are being used to predict specific outcomes using algorithms
Prescriptive	What do I need to do? • Recommended actions and strategies based on champion/challenger strategy outcomes • Applying advanced analytical algorithm to make specific recommendations

Mapping of Big Data's Vs to Analytics Focus



Source: http://ibm.co/1gJyfl3

Mapping of Big Data's V's to Analytics Focus

Historical data can be quite large. There might be a need to process huge amount of data many times a day as it gets updated continuously. Therefore ***volume is mapped to history***.

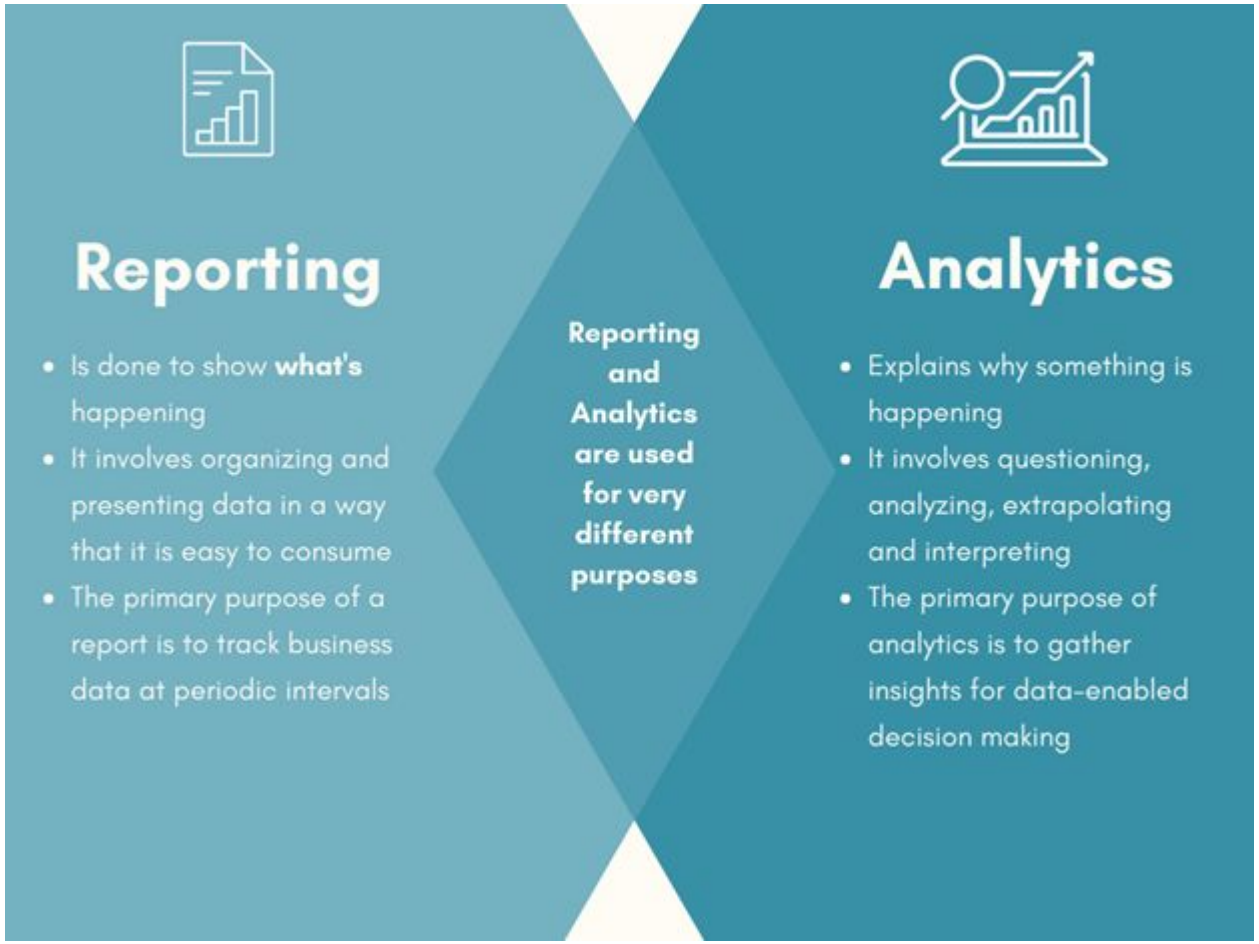
Variety is pervasive. Input data, insights, and decisions can span a variety of forms, ***hence it is mapped to all three***.

Mapping of Big Data's V's to Analytics Focus

High velocity data might have to be processed to help real time decision making and plays across *descriptive, predictive, and prescriptive analytics when they deal with present data*. *Predictive and prescriptive analytics create data about the future.*

Data is uncertain, by nature and its veracity is in doubt. *Therefore veracity is mapped to prescriptive and predictive analytics when it deal with future.*

Analytics Vs. Reporting



Reporting: The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

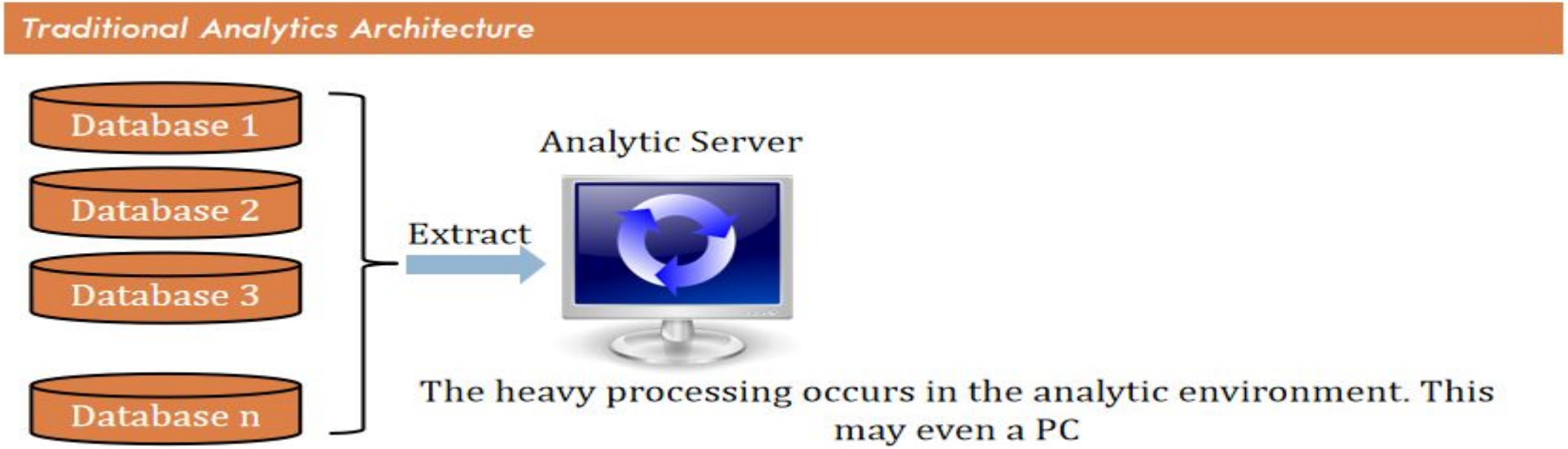
Analytics: The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.

Analytics Vs. Reporting

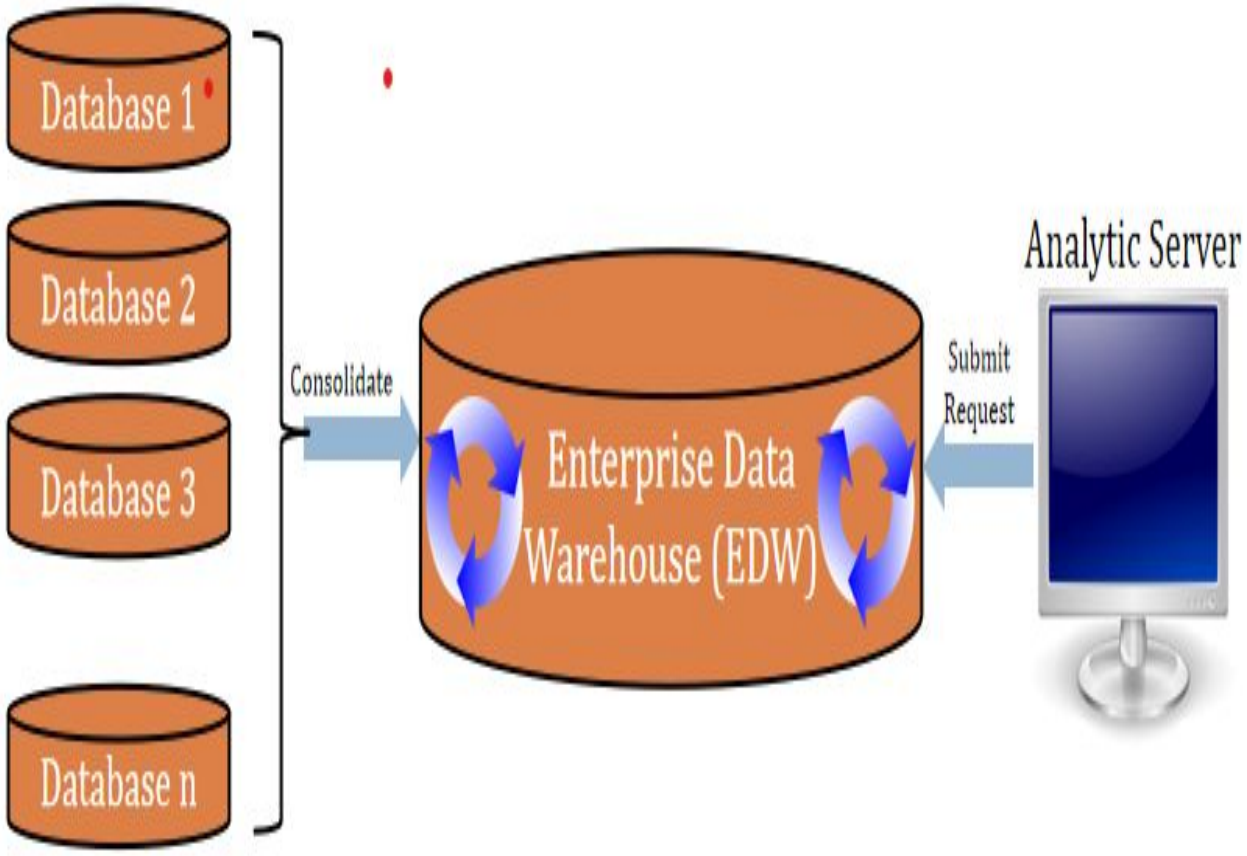
- Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected ranges.
  - Good reporting should raise questions about the business from its end users.
- The goal of analytics is to answer questions by interpreting the data at a deeper level and providing actionable recommendations.

Evolution of Analytics Scalability

- It goes without saying that the world of big data requires new levels of scalability.
- As the amount of data that organizations process continues to increase, the same old methods for handling data just won't work anymore.
- Organizations that don't update their technologies to provide a higher level of scalability will quite simply choke on big data.
- Luckily, there are multiple technologies available that address different aspects of the process of taming big data and making use of it in analytic processes.



Evolution of Analytics Scalability (Contd.)



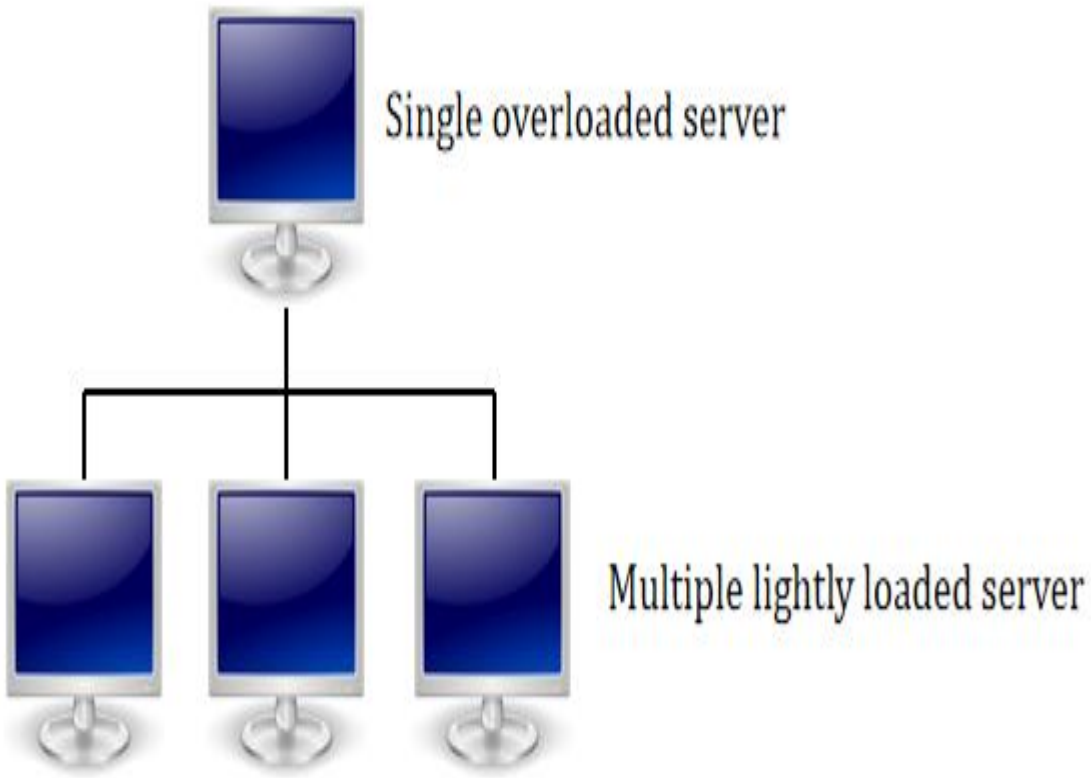
In an in-database environment, the processing stays in the database where the data has been consolidated. The user's machine just submits the request; it doesn't do heavy lifting.

Evolution of Analytics Scalability (Contd.)

- **Massively Parallel Processing Database Analytics Architecture:**
  - **Massively parallel processing** (MPP) database systems is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data.
  - **An MPP database spreads data out into independent pieces managed by independent storage and central processing unit (CPU) resources.**
  - Conceptually, it is like having pieces of data loaded onto multiple network connected personal computers around a house.
  - **The data in an MPP system gets split across a variety of disks managed by a variety of CPUs spread across a number of servers.**

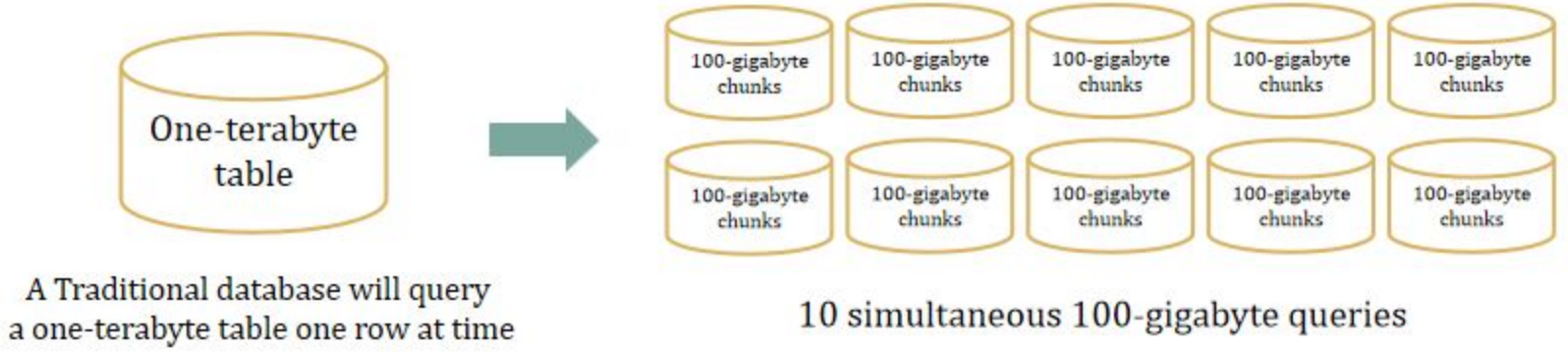
Evolution of Analytics Scalability (Contd.)

- **Massively Parallel Processing Database Analytics Architecture:**



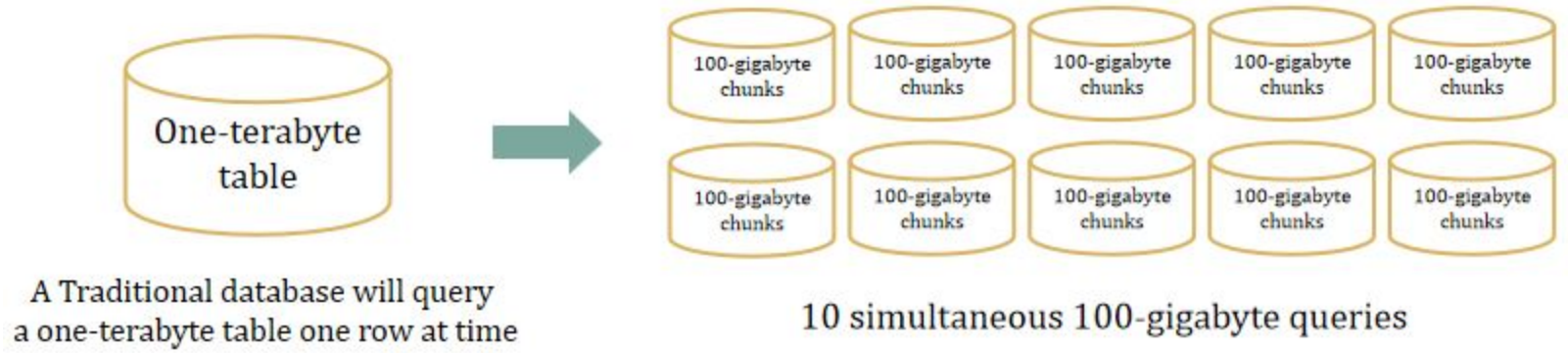
Instead of single overloaded database, an MPP database breaks the data into independent chunks with independent disk and CPU

Example of MPP Database



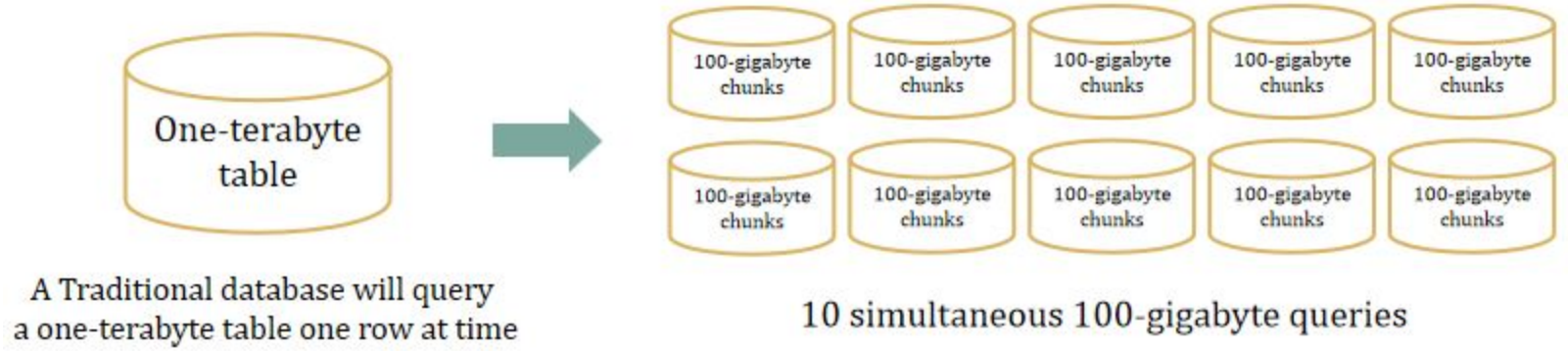
MPP database is based on the principle of **SHARE THE WORK!**

Example of MPP Database



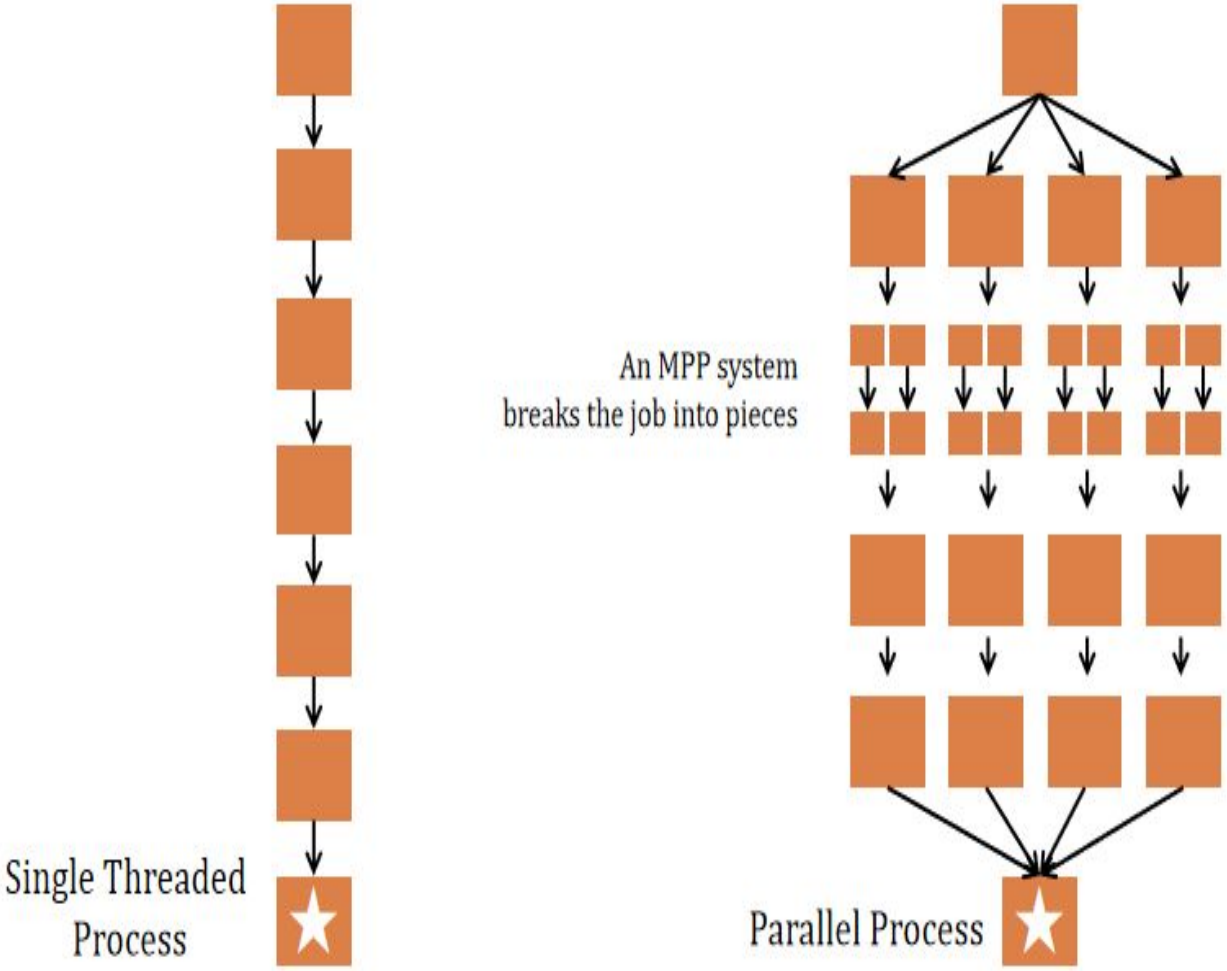
- **A MPP database spreads data out across multiple sets of CPU and disk space.**
- Think logically about dozens or hundreds of personal computers each holding a small piece of a large set of data.
- This allows much faster query execution, since many independent smaller queries are running simultaneously instead of just one big query

Example of MPP Database



- **If more processing power and more speed are required, just bolt on additional capacity in the form of additional processing units.**
- **MPP systems build in redundancy to make recovery easy and have resource management tools to manage the CPU and disk space.**

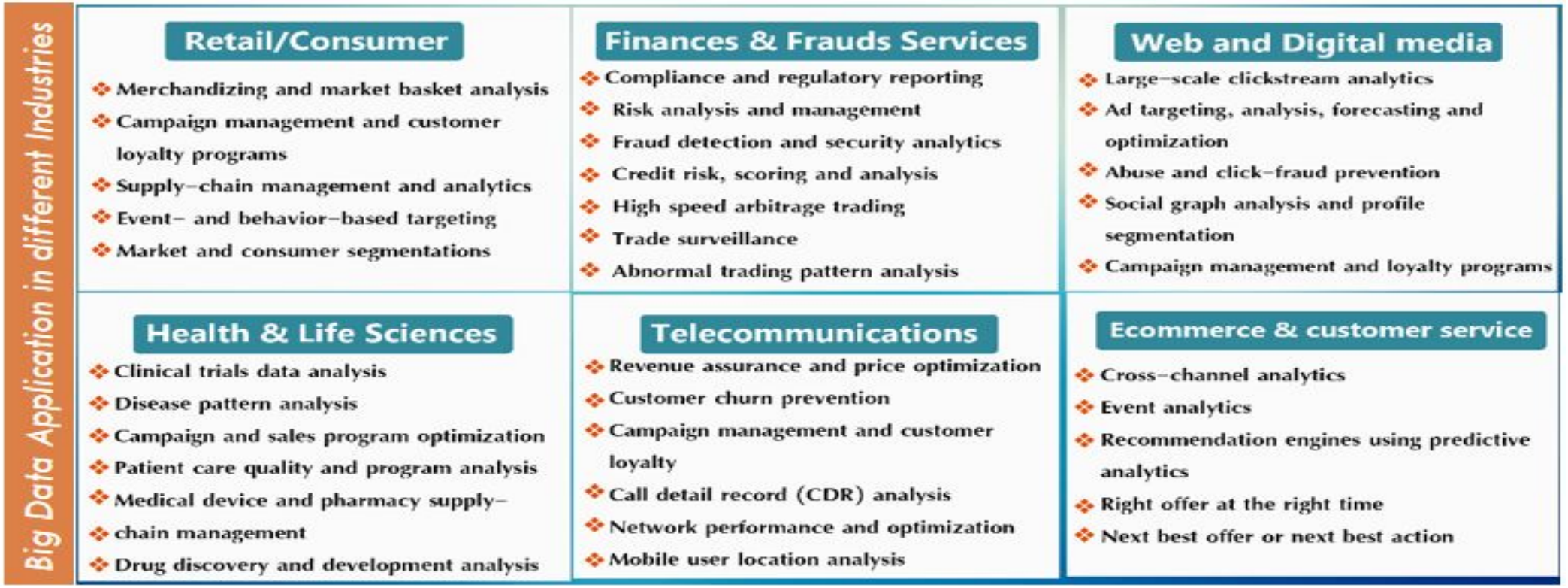
Example of MPP Database



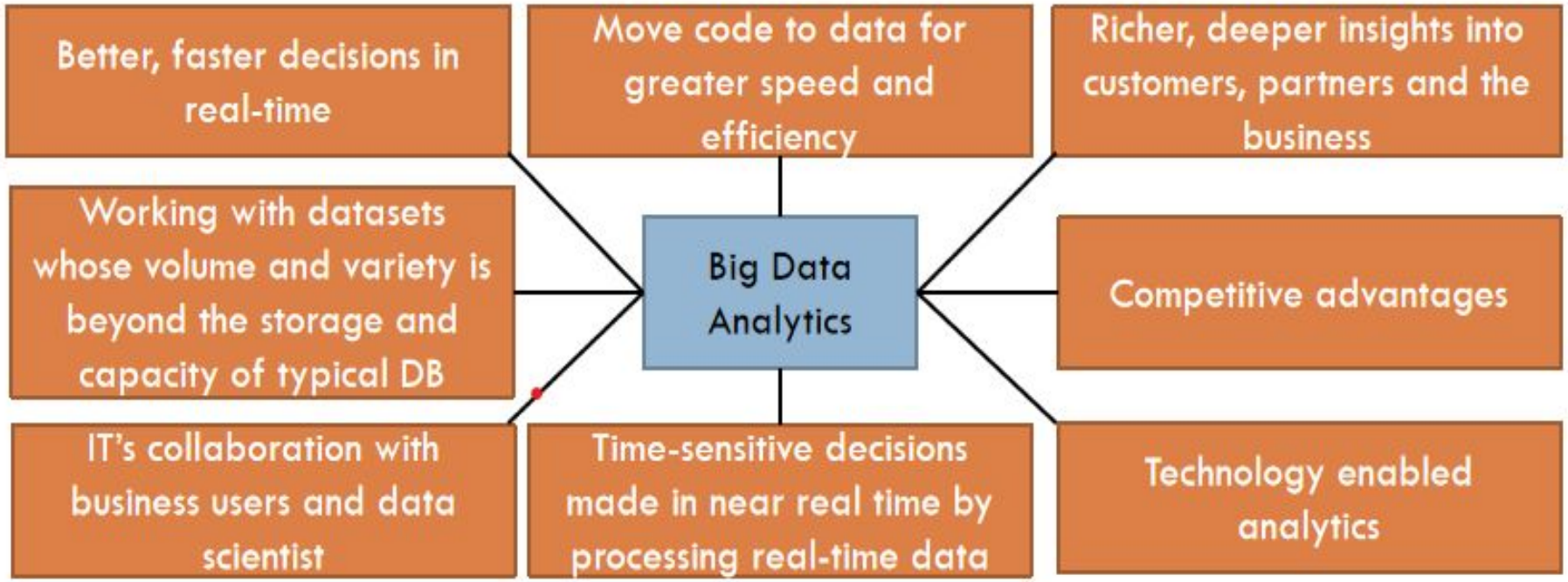
An MPP system allows the different sets of CPU and disk to run the process concurrently

Big Data Analytics

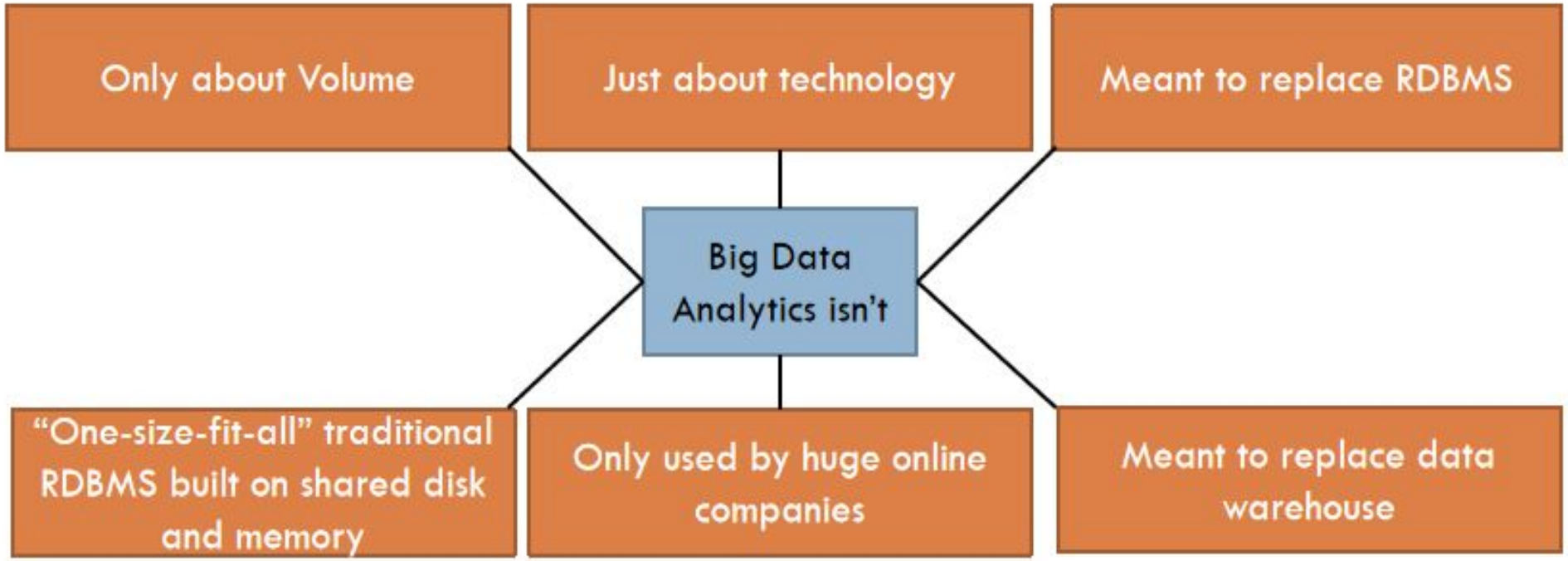
- Big data analytics is the process of extracting useful information by analysing different types of big data sets. It is used to discover hidden patterns, outliers, unearth trends, unknown co-relationship and other useful info for the benefit of faster decision making.



What is Big Data Analytics?



What isn’t Big Data Analytics?



References

1. https://www.researchgate.net/publication/331042286_Advancing_Personalized_Medicine_Through_the_Application_of_Whole_Exome_Sequencing_and_Big_Data_Analytics
2. https://datasciencedojo.com/blog/10-vs-of-big-data/
3. https://www.orbitanalytics.com/understanding-the-difference-between-reporting-and-analytics/
4.