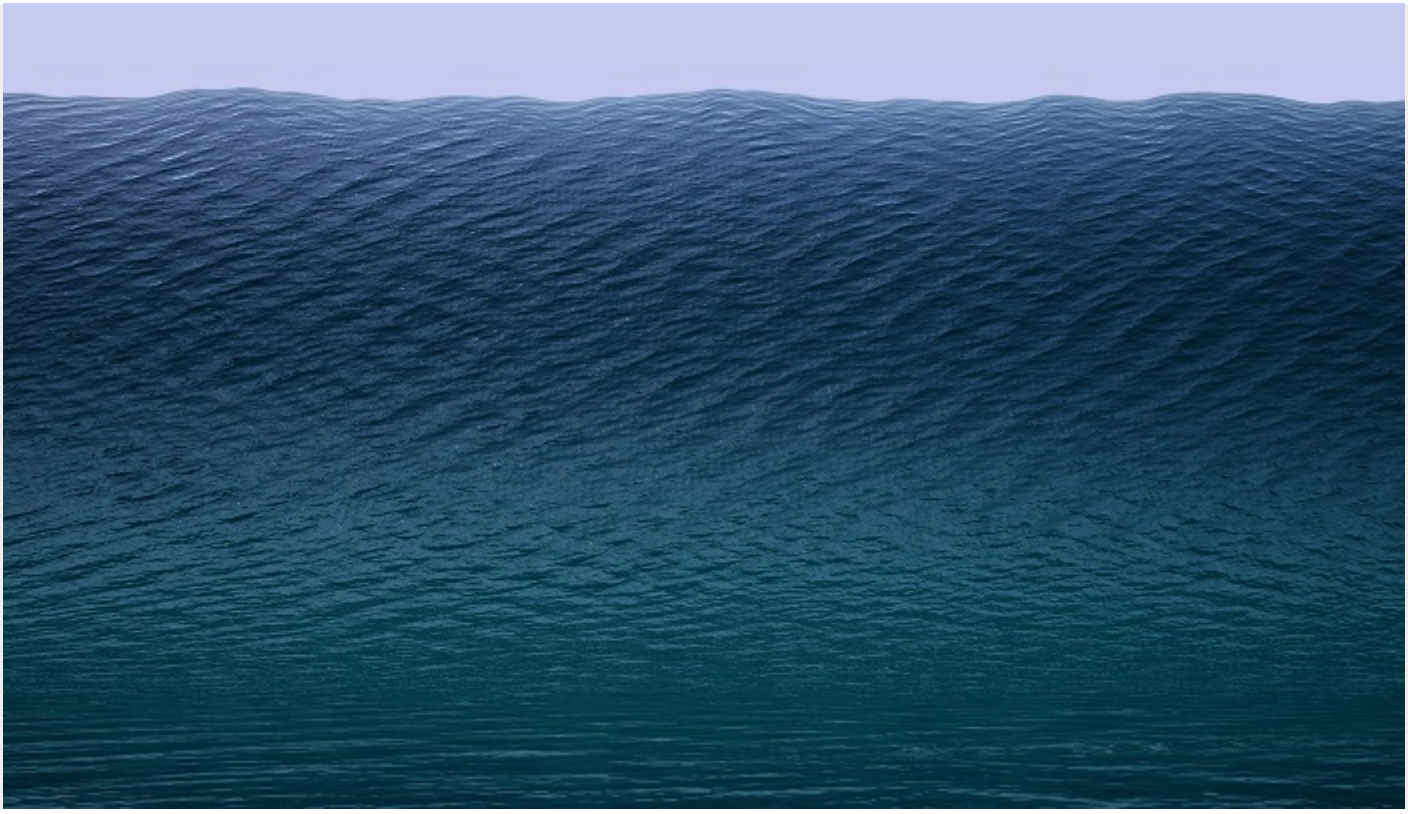
BIG DATA: A BIG INTRODUCTION



The digital universe is continuously expanding—just like the physical universe, except that the digital world alone has generated more data than the number of stars in the entire observable physical universe.

44 zettabytes! That's 44 with trailing zeros ($44x10^{21}$). That's 40 times more bytes than the number of stars in the observable universe.

 By 2025, there will be 175 zettabytes of data in the global datasphere. The growth in data volume is exponential.

All of this data is aptly called Big Data. In this article, we will:

- Introduce Big Data
- Explain core concepts
- Compare small and thick data
- Highlight the latest Big Data trends for business
- Point you to plenty of resources

What is Big Data?

Big data is the term for information assets (data) that are characterized by high volume, velocity, and variety that are systematically extracted, analyzed, and processed for decision making or control actions.

The characteristics of Big Data make it virtually impossible to analyze using traditional data analysis methods.

The importance of big data lies in the patterns and insights, hidden in large information assets, that can drive business decisions. When extracted using advanced analytics technologies, these insights help organizations understand how their users, markets, society, and the world behaves.

3 Vs of Big Data

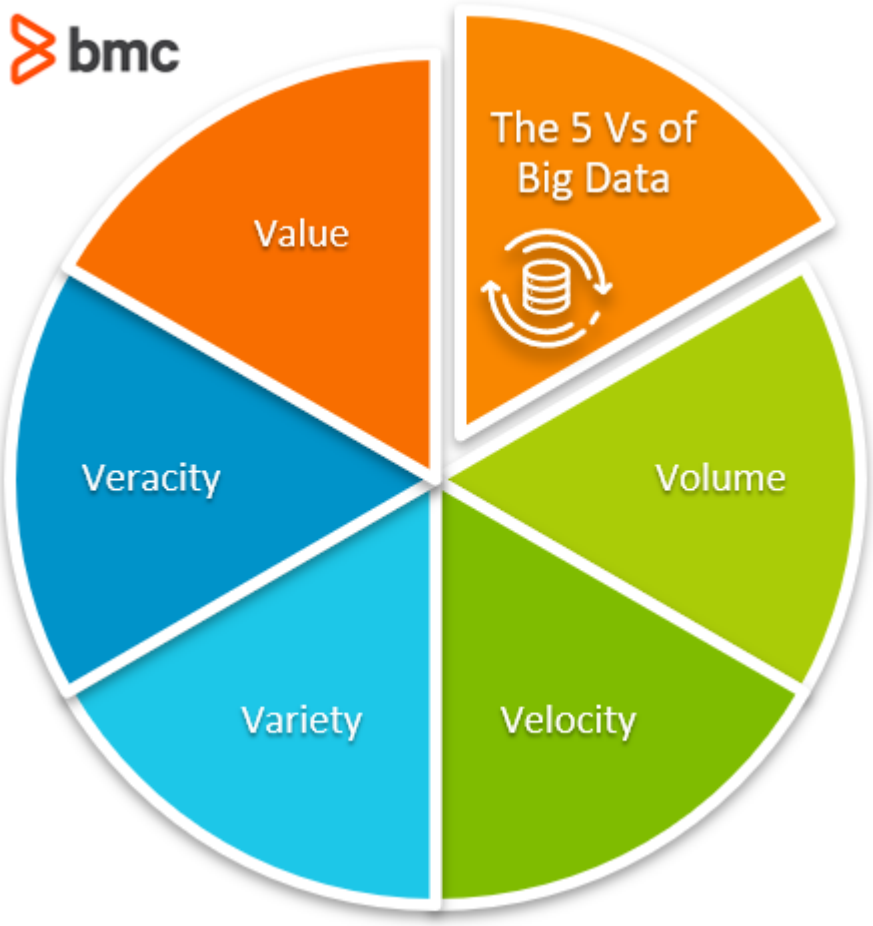
For an information asset to be considered as Big Data, it must meet the 3-V criteria:

- **Volume**. The size of data. High volume data is likely to contain useful insights. A minimum threshold for data to be considered big usually starts at terabytes and petabytes. The large volume of Big Data requires hyperscale computing environments with large storage and fast IOPS (Input/Output Operations per Second) for fast analytics processing.

- **Velocity**. The speed at which data is produced and processed. Big Data is typically produced in streams and is available in real-time. The continuous nature of data generation makes it relevant for real-time decision-making.
- **Variety**. The type and nature of information assets. Raw big data is often unstructured or multi-structured, generated with a variety of attributes, standards, and file formats. For example, datasets collected from sensors, log files, and social media networks are unstructured. So, they must be processed into structured databases for data analytics and decision-making.

More recently, two additional Vs help characterize Big Data:

- **Veracity**. The reliability or truthfulness of data. The extent to which the output of big data analysis is pertinent to the associated business goals is determined by the quality of data, the processing technology, and the mechanism used to analyze the information assets.
- **Value**. The usefulness of Big Data assets. The worthiness of the output of big data analysis can be subjective and is evaluated based on unique business objectives.



Big data vs small data vs thick data

In contrast to these characteristics, there are two other forms of data: small data and thick data.

Small Data

Small Data refers to manageable data assets, usually in numerical or structured form, that can be analyzed using simple technologies such as Microsoft Excel or an open source alternative.

Thick Data

Thick Data refers to text or qualitative data that can be analyzed using manageable manual processes. Examples include:

- Interview questions
- Surveys
- Video transcripts

When you use qualitative data in conjunction with quantitative big data, you can better understand the sentiment and behavioral aspects that can be easily communicated by individuals. Thick Data is particularly useful in the domains of medicine and scientific research where responses from individual humans hold sufficient value and insights—versus large big data streams.

Big Data trends in 2021-2022

Big Data technologies are continuously improving. Indeed, data itself is fast becoming the most important asset for a business organization.

Prevalence of the Internet of Things (IoT), cloud computing, and Artificial Intelligence (AI) is making it easier for organizations to transform raw data into actionable knowledge.

Here are three of the most popular big data technology trends to look out for in 2021:

- **Augmented Analytics.** The Big Data industry will be worth nearly $274 billion by the end of 2021. Technologies such as Augmented Analytics, which help organizations with the data management process, are projected to grow rapidly and reach $18.4 billion by the year 2023.
- **Continuous Intelligence.** Integrating real-time analytics to business operations is helping organizations leapfrog the competition with proactive and actionable insights delivered in real-time.
- **Blockchain.** Stringent legislations such as the GDPR and HIPAA are encouraging organizations to make data secure, accessible, and reliable. Blockchain and similar technologies are making their way into the financial industry as a data governance and security instrument that is highly resilient and robust against privacy risks. This EU resource discusses how blockchain complements some key GDPR objectives.

Big Data best practices for businesses

Certainly the world of data is growing exponentially. Are your data and data processes up to the tasks that you're asking of it?

BMC Blogs has many resources for understanding and working with Big Data. Browse the BMC Machine Learning & Big Data Blog or dive deeper into these areas:

Data basics

- Data Quality: Top Concepts & Best Practices for Enterprise IT
- Structured vs Unstructured Data: A Shift in Privacy
- Data Architecture: Components, Standards & Changing Architectures
- Data Streaming Explained: Pros, Cons & How It Works

- What Is a Data Pipeline?

Data storage

- Data Storage Explained: Data Lake vs Warehouse vs Database
- Cold vs Hot Data Storage: What's The Difference?
- CAP Theorem for Databases: Consistency, Availability & Partition Tolerance
- Database-as-a-Service (DBaaS) Explained

Data management

- Data Management vs Data Governance: Main differences
- What Is Data Governance? Why Do I Need It?
- 3 Simple Data Monetization Strategies for Companies
- What Is Test Data Management (TDM)?

Data security

- Introduction To Data Security
- Data Loss Prevention & DLP Solutions
- Big Data Security Issues in the Enterprise

Data analysis & analytics

- Data Analytics vs Data Analysis: What's The Difference?
- Big Data vs Analytics vs Data Science: What's The Difference?
- Data Visualization: Getting Started with Examples, part of our Data Visualization Guide

Machine learning, data science & AI

- Predictive Analytics vs Machine Learning: What's The Difference?
- Data Annotation & Its Role in Machine Learning
- 3 Keys to Building Resilient Data Pipelines

Data theory & thought leadership

- Data Ethics for Companies
- What Is Data Gravity?
- Mindful AI: 5 Concepts for Mindful Artificial Intelligence
- What Is Goodhart's Law? Balancing Authenticity & Measurement

Learning Big Data

    - Data Engineer vs Data Scientist: What's the Difference?
    - Data Science Certifications: An Introduction
    - Enabling the Citizen Data Scientists

Big data tutorial series

These tutorial series are part of BMC Guides.

Amazon Redshift
Apache Cassandra
Apache Spark
AWS
Data Visualization
Docker
DynamoDB
ElasticSearch
Hadoop
Kubernetes
Microsoft Power BI
MongoDB
Pandas
Microsoft Power BI
scikit-learn
Snowflake
Tableau Online