big data analytics

What is big data analytics?

Big data analytics is the often complex process of examining [big data](#) to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

On a broad scale, [data analytics](#) technologies and techniques give organizations a way to analyze data sets and gather new information. Business intelligence (BI) queries answer basic questions about business operations and performance.

Big data analytics is a [form of advanced analytics](#), which involve complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by analytics systems.

Why is big data analytics important?

Organizations can use [big data analytics systems and software to make data-driven decisions](#) that can improve business-related outcomes. The benefits may include more effective marketing, new revenue opportunities, customer personalization and improved operational efficiency. With an

effective strategy, [these benefits can provide competitive advantages over rivals](#).

How does big data analytics work?

Data analysts, [data scientists](#), predictive modelers, statisticians and other analytics professionals collect, process, clean and analyze growing volumes of structured transaction data as well as other forms of data not used by conventional BI and analytics programs.

Here is an overview of the four steps of the big data analytics process:

1. Data professionals **collect** data from a variety of different sources. Often, it is a mix of [semistructured](#) and unstructured data. While each organization will use different data streams, some common sources include:

-   internet [clickstream](#) data;
-   web server logs;
-   cloud applications;
-   mobile applications;
-   social media content;
-   text from customer emails and survey responses;
-   mobile phone records; and
-   machine data captured by [sensors](#) connected to the internet of things (IoT).

2. Data is prepared and **processed**. After data is collected and stored in a [data warehouse or data lake](#), data professionals must organize, configure and partition the data properly for analytical queries. Thorough [data preparation](#) and processing makes for higher performance from analytical queries.
3. Data is **cleansed** to improve its quality. Data professionals scrub the data using scripting tools or data quality software. They look for any errors or inconsistencies, such as duplications or formatting mistakes, and organize and tidy up the data.
4. The [collected, processed and cleaned data](#) is **analyzed** with analytics software. This includes tools for:

- [data mining](#), which sifts through data sets in search of patterns and relationships
- predictive analytics, which builds models to forecast customer behavior and other future actions, scenarios and trends
- machine learning, which taps various algorithms to analyze large data sets
- [deep learning](#), which is a more advanced offshoot of machine learning
- [text mining](#) and statistical analysis software
- artificial intelligence (AI)
- mainstream business intelligence software
- data visualization tools

Key big data analytics technologies and tools

Many different types of tools and technologies are used to support big data analytics processes. Common technologies and tools used to enable big data analytics processes include:

- **Hadoop**, which is an open source framework for storing and processing big data sets. Hadoop can handle large amounts of structured and unstructured data.
- **Predictive analytics** hardware and software, which process large amounts of complex data, and use machine learning and statistical algorithms to make predictions about future event outcomes. Organizations use predictive analytics tools for fraud detection, marketing, risk assessment and operations.
- **Stream analytics** tools, which are used to filter, aggregate and analyze big data that may be stored in many different formats or platforms.
- **Distributed storage** data, which is replicated, generally on a non-relational database. This can be as a measure against independent node failures, lost or corrupted big data, or to provide low-latency access.
- **NoSQL** databases, which are non-relational data management systems that are useful when working with large sets of distributed data. They do not require a fixed schema, which makes them ideal for raw and

unstructured data.

- **A data lake** is a large storage repository that holds native-format raw data until it is needed. Data lakes use a flat architecture.
- **A [data warehouse](#)**, which is a repository that stores large amounts of data collected by different sources. Data warehouses typically store data using predefined schemas.
- **Knowledge discovery/big data mining** tools, which enable businesses to mine large amounts of structured and unstructured big data.
- **In-memory data fabric**, which distributes large amounts of data across system memory resources. This helps provide low latency for data access and processing.
- **Data virtualization**, which enables data access without technical restrictions.
- **Data integration software,** which enables big data to be streamlined across different platforms, including Apache, Hadoop, MongoDB and Amazon EMR.
- **Data quality software**, which cleanses and enriches large data sets.
- **[Data preprocessing](#) software,** which prepares data for further analysis. Data is formatted and unstructured data is cleansed.
- **Spark,** which is an open source cluster computing framework used for batch and stream data processing.

Big data analytics applications often include data from both internal systems and external sources, such as weather data or demographic data on consumers compiled by third-party information services providers. In addition, streaming analytics applications are becoming common in big data environments as users look to perform [real-time analytics](#) on data fed into Hadoop systems through stream processing engines, such as Spark, Flink and Storm.

Early big data systems were mostly deployed on premises, particularly in large organizations that collected, organized and analyzed massive amounts of data. But cloud platform vendors, such as Amazon Web Services (AWS), Google and Microsoft, have made it easier to set up and manage Hadoop clusters in the cloud. The same goes for Hadoop suppliers such as Cloudera, which supports the distribution of the big data framework on the AWS, Google and [Microsoft Azure](#) clouds. Users can now spin up clusters in the cloud, run them for as long as they need and then take them offline with usage-based pricing that doesn't require ongoing software licenses.

Big data has become increasingly beneficial in [supply chain analytics](#). Big supply chain analytics utilizes big data and quantitative methods to enhance decision-making processes across the supply chain. Specifically, big supply chain analytics expands data sets for increased analysis that

goes beyond the traditional internal data found on enterprise resource planning ([ERP](#)) and supply chain management (SCM) systems. Also, big supply chain analytics implements highly effective statistical methods on new and existing data sources.

Business intelligence vs. advanced analytics

	BUSINESS INTELLIGENCE	ADVANCED ANALYTICS
Answers the questions:	▪ What happened? ▪ When? ▪ Who? ▪ How many?	▪ Why did it happen? ▪ Will it happen again? ▪ What will happen if we change x? ▪ What else does the data tell us that we never thought to ask?
Includes:	▪ Reporting (KPIs, metrics) ▪ Automated monitoring and alerting (thresholds) ▪ Dashboards ▪ Scorecards ▪ OLAP (cubes, slice and dice, drilling) ▪ Ad hoc query ▪ Operational and real-time BI	▪ Statistical or quantitative analysis ▪ Data mining ▪ Predictive modeling ▪ Multivariate testing ▪ Big data analytics ▪ Text analytics

TechTarget

Big data analytics is a form of advanced analytics, which has marked differences compared to traditional BI.

Big data analytics uses and examples

Here are some examples of [how big data analytics can be used to help organizations](#):

- **Customer acquisition and retention.** Consumer data can help the marketing efforts of companies, which can act on trends to increase customer satisfaction. For example, [personalization engines](#) for Amazon, Netflix

and Spotify can provide improved customer experiences and create customer loyalty.

- **Targeted ads.** Personalization data from sources such as past purchases, interaction patterns and product page viewing histories can help generate compelling targeted ad campaigns for users on the individual level and on a larger scale.
- **Product development.** Big data analytics can provide insights to inform about product viability, development decisions, progress measurement and steer improvements in the direction of what fits a business' customers.
- **Price optimization.** Retailers may opt for pricing models that use and model data from a variety of data sources to maximize revenues.
- **Supply chain and channel analytics.** [Predictive analytical models](#) can help with preemptive replenishment, B2B supplier networks, inventory management, route optimizations and the notification of potential delays to deliveries.
- **Risk management.** Big data analytics can identify new risks from data patterns for effective risk management strategies.
- **Improved decision-making.** Insights business users extract from relevant data can help organizations make quicker and better decisions.

Big data analytics benefits

The benefits of using big data analytics include:

- Quickly analyzing large amounts of data from different sources, in many different formats and types.
- Rapidly making better-informed decisions for effective strategizing, which can benefit and improve the supply chain, operations and other areas of strategic decision-making.
- Cost savings, which can result from new business process efficiencies and optimizations.
- A better understanding of customer needs, behavior and sentiment, which can lead to better marketing insights, as well as provide information for product development.
- Improved, better informed [risk management](#) strategies that draw from large sample sizes of data.

Unstructured data types

Text files and documents	Server, website and application logs	Sensor data	Images
Video files	Audio files	Emails	Social media data

ICONS: VIEWLPIXEL/ADOBE STOCK, ORKHAN/ADOBE STOCK

©2018 TECHTARGET. ALL RIGHTS RESERVED TechTarget

TechTarget

Big data analytics involves analyzing structured and unstructured data.

Big data analytics challenges

Despite the wide-reaching benefits that come with using big data analytics, its use also comes with challenges:

- **Accessibility of data.** With larger amounts of data, storage and processing become more complicated. Big data should be stored and maintained properly to ensure it can be used by less experienced data scientists and analysts.
- **Data quality maintenance.** With high volumes of data

coming in from a variety of sources and in different formats, [data quality management](#) for big data requires significant time, effort and resources to properly maintain it.

- **Data security.** The complexity of big data systems presents unique security challenges. Properly addressing security concerns within such a complicated big data ecosystem can be a complex undertaking.
- **Choosing the right tools.** Selecting from the vast array of big data analytics tools and platforms available on the market can be confusing, so organizations must know how to pick the best tool that aligns with users' needs and infrastructure.
- With a potential lack of internal analytics skills and the high cost of hiring experienced data scientists and engineers, some organizations are finding it hard to fill the gaps.

History and growth of big data analytics

The term *big data* was first used to refer to increasing data volumes in the mid-1990s. In 2001, Doug Laney, then an analyst at consultancy Meta Group Inc., expanded the definition of big data. This expansion described the increasing:

- Volume of data being stored and used by organizations;

- Variety of data being generated by organizations; and
- Velocity, or speed, in which that data was being created and updated.

Those three factors became known as the [3Vs](#) of big data. Gartner popularized this concept after acquiring Meta Group and hiring Laney in 2005.

Another [significant development in the history of big data](#) was the launch of the Hadoop distributed processing framework. Hadoop was launched as an [Apache](#) open source project in 2006. This planted the seeds for a clustered platform built on top of commodity hardware and that could run big data applications. The Hadoop framework of software tools is widely used for managing big data.

By 2011, big data analytics began to take a firm hold in organizations and the public eye, along with Hadoop and various related big data technologies.

Initially, as the Hadoop ecosystem took shape and started to mature, big data applications were primarily used by large internet and e-commerce companies such as Yahoo, Google and Facebook, as well as analytics and marketing services providers.

More recently, a broader variety of users have embraced big data analytics as a key technology driving [digital](#)

[transformation](#). Users include retailers, financial services firms, insurers, healthcare organizations, manufacturers, energy companies and other enterprises.

REFERENCE:

Chai, Wesley, et al. "What Is Big Data Analytics and Why Is It Important?" Business Analytics, TechTarget, 14 Dec. 2021, https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics