Big Data Analytics (https://www.selecthub.com/category/big-data-analytics/)

What Are The Types Of Big Data?

By Richard Allen (https://www.selecthub.com/author/richard-allen/)

Big Data Analytics (https://www.selecthub.com/category/big-data-analytics/)

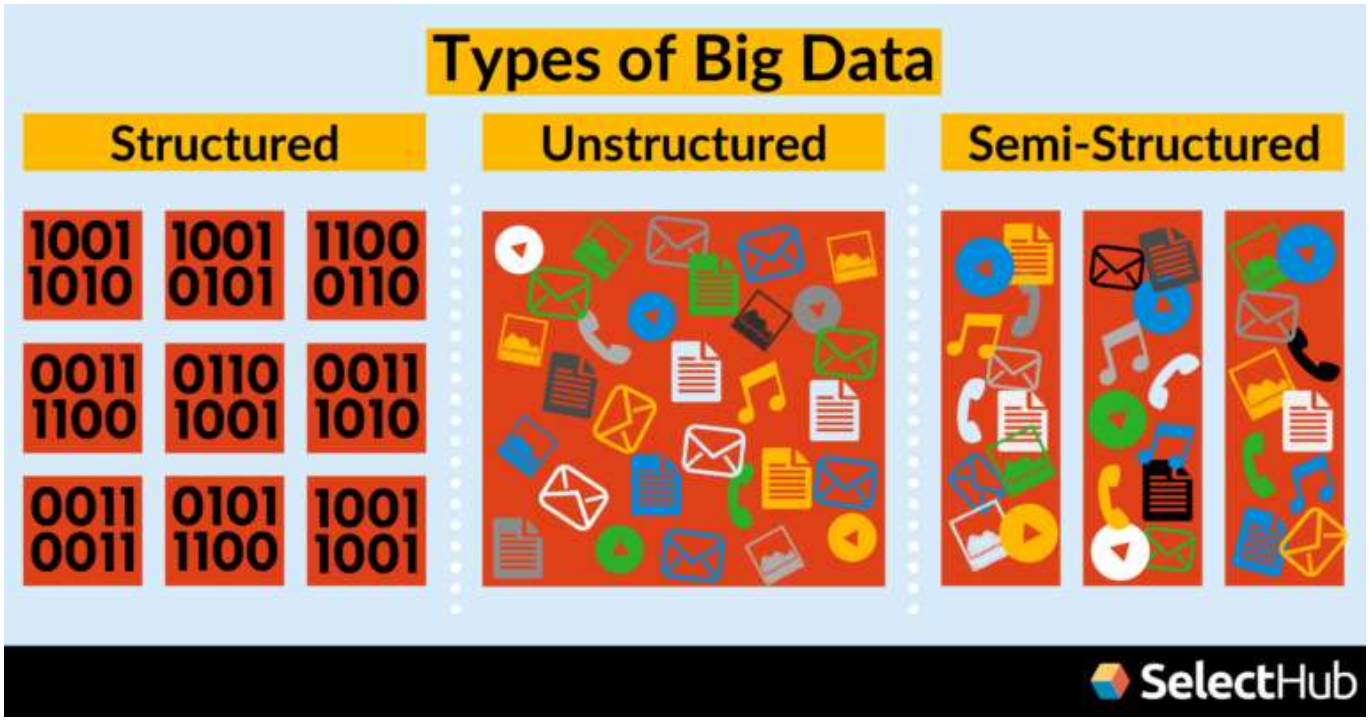
2 comments (https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/?noamp=mobile#comments)

Last Reviewed: September 20, 2024

As the Internet age surges on, we create an unfathomable amount of data every second. So much so that we've denoted it simply as big data. Naturally, businesses and analysts want to crack open all the different types of big data for the juicy information inside. But it's not so simple. The different types leverage varying big data tools (/big-data-analytics-tools/) and have different complications that accompany working with each individual data point plucked out of the vast ether.

Free

Compare Top Big Data Software Leaders (https://pmo.selecthub.com/request-custom-scorecard/?category=Big%20Data%20Analytics%20Tools)



To quantify it, more than 33 zettabytes of data (https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf) floated around the internet, in servers and on computers a couple of years back.

No, that's not a word we made up. Nor did we pull it from "Star Wars", "Star Trek" or "The Hitchhiker's Guide to the Galaxy." It's 33 trillion gigabytes. In granular bytes, it's a three, then another three…followed by 21 zeros.

That's a lot of tweets, statuses, selfies, bank accounts, flight paths, street maps, product prices and any other piece of digital information you can think of.

If you can harness, process and present it all, data can become an invaluable tool for your business. You can understand why your business stands where it does in comparison to competitors, generate projections for future business or develop deep insights on an entire market.

But with so much quantity comes an equal amount of variety. Not all of that data is readily usable in analytics and has to undergo a transformation known as data cleansing to make it understandable. Some of it carries some clues to help the user tap into its well of knowledge.

Big data is classified in three ways:

- Structured Data
- Unstructured Data
- Semi-Structured Data

These three terms, while technically applicable at all levels of analytics, are paramount in big data. Understanding where the raw data comes from and how it has to be treated before analyzing it only becomes more important when working with the volume of big data. Because there's so much of it, information extraction needs to be efficient to make the endeavor worthwhile.

The structure of the data is the key to not only how to go about working with it, but also what insights it can produce. All data goes through a process called extract, transform, load (https://www.sas.com/en_us/insights/data-management/what-is-etl.html) (ETL) before it can be analyzed. It's a very literal term: data is harvested, formatted to be readable by an application, and then stored for use. The ETL process for each structure of data varies.

Let's dive into each, explaining what it means and how it relates to big data analytics.

Structured Data

Structured data (https://www.dummies.com/programming/big-data/engineering/structured-data-in-a-big-data-environment/) is the easiest to work with. It is highly organized with dimensions defined by set parameters.

Think spreadsheets; every piece of information is grouped into rows and columns. Specific elements defined by certain variables are easily discoverable.

It's all your quantitative data:

- Age
- Billing
- Contact
- Address
- Expenses
- Debit/credit card numbers

Because structured data is already tangible numbers, it's much easier for a program to sort through and collect data.

Structured data follows schemas (https://www.tutorialspoint.com/dbms/dbms_data_schemas.htm): essentially road maps to specific data points. These schemas outline where each datum is and what it means.

A payroll database will lay out employee identification information, pay rate, hours worked, how compensation is delivered, etc. The schema will define each one of these dimensions for whatever application is using it. The program won't have to dig into data to discover what it actually means, it can go straight to work collecting and processing it.

Working With It

Structured data is the easiest type of data to analyze because it requires little to no preparation before processing. A user might need to cleanse data (https://www.sisense.com/glossary/data-cleaning/) and pare it down to only relevant points, but it won't need to be interpreted or converted too deeply before a true inquiry can be performed.

One of the major perks of using structured data is the streamlined process of merging enterprise data with relational. Because pertinent data dimensions are usually defined and specific elements are in a uniform format, very little preparation needs to be done to make all sources compatible.

The ETL process for structured data stores the finished product in what is called a data warehouse (/business-intelligence/business-intelligence-and-data-warehousing/). These databases are highly structured and filtered for the specific analytics purpose the initial data was harvested for.

Relational databases (https://searchdatamanagement.techtarget.com/definition/relational-database) are easily-queried datasets. They allow users to find external information and either study it standalone or integrate it with their internal data for more context. Relational database management systems use SQL, or Structured Query Language, to access data, providing a uniform language across a network of data platforms and sources.

This standardization enables scalability in data processing. Time spent on defining data sources and making them cooperate with each other is reduced, expediting the delivery of actionable insight.

The qualitative nature and readability of this classification also grant compatibility with almost any relevant source of information. The amount of data used is limited only by what the user can get their hands on.

Unfortunately, there's only so much structured data available, and it denotes a slim minority of all data in existence.

Free

[Get our Big Data Requirements Template](https://pmo.selecthub.com/big-data-requirements-onsite/)

Unstructured Data

Not all data is as neatly packed and sorted with instructions on how to use as structured data is. The consensus is no more than 20% of all data is structured (https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/).

So what's the remaining four-fifths of all the information out there? Since it isn't structured, we naturally call this unstructured data.

Unstructured data is all your unorganized data:

You might be able to figure out why it constitutes so much of the modern data library. Almost everything you do with a computer generates unstructured data. No one is transcribing their phone calls or assigning semantic tags to every tweet they send.

While structured data saves time in an analytical process, taking the time and effort to give unstructured data some level of readability is cumbersome.

For structured data, the ETL process is very simple. It is simply cleansed and validated in the transform stage before loading into a database. But for unstructured data, that second step is much more complicated.

To gain anything resembling useful information, the dataset needs to be interpretable. But the effort can be much more rewarding than processing unstructured data's simpler counterpart. As they say in sports, you get out what you put in.

Working With It

The hardest part of analyzing unstructured data is teaching an application to understand the information it's extracting. More often than not, this means translating it into some form of structured data.

This isn't easy and the specifics of how it is done vary from format to format and with the end goal of the analytics. Methods like text parsing, natural language processing and developing content hierarchies via taxonomy are common.

Almost universally, it involves a complex algorithm blending the processes of scanning, interpreting and contextualizing functions.

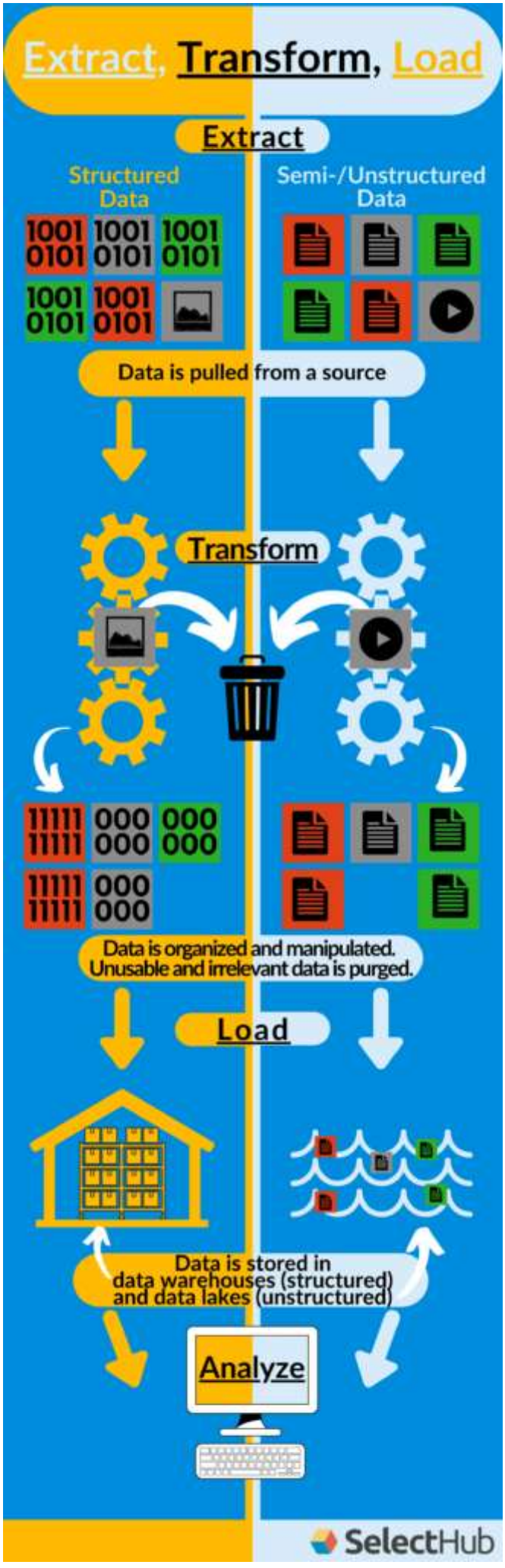
This brings us to an important point: context is almost, if not as, important as the information wrung out of the data. Alissa Lorentz, then the vice president of creative, marketing and design at Augify, explained: a query on an unstructured data set might yield the number 31, but without context it's meaningless. It could be "the number of days in a month, the amount of dollars a stock increased…, or the number of items sold today."

The contextual aspect is what makes unstructured data ubiquitous in big data: merging internal data with external context makes it more meaningful. The more context (and data in general), the more accurate any sort of model or analysis is.

This context can be created from unstructured datasets, like NoSQL databases, or human dictation. We can tell applications and AI what data means. In fact, you've probably been doing it for years every time Google asks you to prove you're not a robot (https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it). The world of machine learning, or AI teaching itself how to improve and discover patterns, is becoming instrumental in the world of big data because of its ability to autonomously improve on models.

In contrast to structured data, which is stored in data warehouses (/business-intelligence/data-warehouse-requirements-gathering/), unstructured is placed in data lakes (https://en.wikipedia.org/wiki/Data_lake), which preserve the raw format of the data and all of the information it holds. In warehouses, the data is limited to its defined schema. This is not true of lakes which make the data more malleable.

Products like Hadoop (/big-data-analytics-tools/hadoop/) are built with extensive networks of data clusters and servers, allowing all of the data to be stored and analyzed on a big data scale.


Extract, Transform, Load
Extract
Structured Data
Semi-/Unstructured Data
Data is pulled from a source
Transform
Data is organized and manipulated.
Unusable and irrelevant data is purged.
Load
Data is stored in
data warehouses (structured)
and data lakes (unstructured)
Analyze
SelectHub

Semi-Structured Data

Semi-structured data toes the line between structured and unstructured. Most of the time, this translates to unstructured data with metadata attached to it. This can be inherent data collected, such as time, location, device ID stamp or email address, or it can be a semantic tag attached to the data later.

Let's say you take a picture of your cat from your phone. It automatically logs (https://hackernoon.com/your-digital-photos-can-reveal-information-about-you-732783dd857b) the time the picture was taken, the GPS data at the time of the capture and your device ID. If you're using any kind of web service for storage, like iCloud, your account info becomes attached to the file.

If you send an email, the time sent, email addresses to and from, the IP address from the device sent from, and other pieces of information are linked to the actual content of the email.

In both scenarios, the actual content (i.e. the pixels that compose the photo and the characters that make up the email) is not structured, but there are components that allow the data to be grouped based on certain characteristics.

Free

[Compare Top Big Data Software Leaders](https://pmo.selecthub.com/request-custom-scorecard/?category=Big%20Data%20Analytics%20Tools)

Working With It

Semi-structured splits the gap between structured and unstructured data, which, using the right datasets, can make it a huge asset. It can inform AI training and machine learning by associating patterns with metadata.

Semi-structured data has no set schema. This can be both a benefit and a challenge. It can be more difficult to work with because effort must be put in to tell the application what each data point means. But this also means that the limits in structured data ETL in terms of definition don't exist.

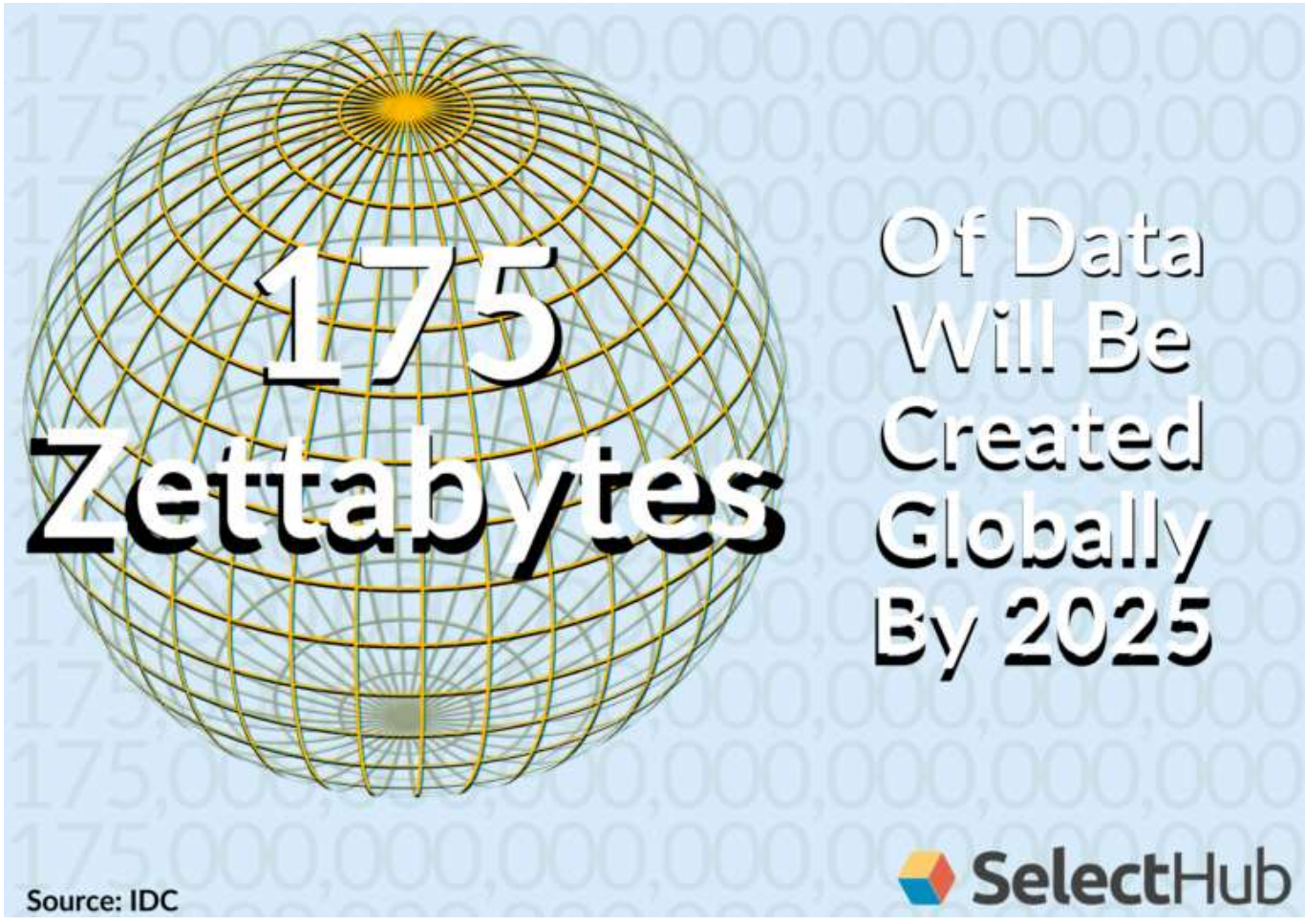
Queries on semi-structured datasets can be organized by schema creation through the metadata, but they are not bound by them. Information extracted from the actual content, as it would be for all unstructured data, can be further contextualized with the metadata for deeper insights that can provide demographic information.

Markup languages like XML (https://queue.acm.org/detail.cfm?id=1103834) allow text data to be defined by its own contents, rather than conform to a schema. The relational model is built out of the data, rather than filling data into a pre-configured form. It gives semantics to the content, rather than use its prescribed meaning.

XML specifically allows data to be organized into a tree structure, stemming attributes and decorations from individual nodes, potentially metadata and semantic tags. This allows layered analysis and deeper intelligence gathered from semi-structured intelligence.

Subtypes of Data

Though not formally considered big data, there are subtypes of data that hold some level of pertinence to the field of analytics. Often, these refer to the origin of the data, such as geospatial (locational), machine (operational logging), social media or event-triggered. These subtypes can also refer to access levels: open (i.e. open source), dark/lost (siloed within systems that make them inaccessible to outsiders, like CCTV systems) or linked (web data transmitted via APIs and other connection methods).



Interacting with Data Through Programming

Different programming languages will accomplish different things when working with the data. There are three major players on the market:

- **Python:** Python (https://www.python.org/about/) is an open-source language, and is regarded as one of the simplest to learn. It utilizes concise syntax and abstraction. Because of its popularity and open-source nature, it has an extensive community of support with near-endless libraries that enable scalability and connections with online applications. It is compatible with Hadoop.
- **R:** For more sophisticated analytics and specific building, R (https://www.r-project.org/about.html) is the language of choice. It is one of the top coding languages available for data manipulation and can be used at every step of an analytics process all the way through to visualization. It provides users with a community-developed network of archived packages, called CRAN, enabling more than 15,000 functions to be implemented with little coding. One of its drawbacks is the fact that it does all of its processing in-memory, meaning the user will likely need to distribute analytics over several devices to handle big data.
- **Scala:** On the come up in popularity is Scala (https://www.scala-lang.org/), a Java based-language. It was used to develop several Apache products, including Spark, a major player in the big data platforms market. It utilizes both object-oriented and functional processing, meaning it can handle both structured and unstructured data alike.

Other languages like Java, SQL, SAS, Go and C++ are used commonly in the market and can be utilized to accomplish big data analytics.

Free

[Compare Top Big Data Software Leaders](https://pmo.selecthub.com/request-custom-scorecard/?category=Big%20Data%20Analytics%20Tools)

Next Steps

Big data paves the way for virtually any kind of insight an enterprise could be looking for, be the analytics prescriptive, descriptive, diagnostic or predictive (/business-intelligence/predictive-descriptive-prescriptive-analytics/). The realm of big data analytics is built on the shoulders of giants: the potential of data harvesting and analyzing has been known for decades, if not centuries.

If you're selecting a solution, the types of big data analytics you're working with is something you need to consider. Don't know where to begin in your adventure? Our crash course on big data analytics (/business-analytics/crash-course-big-data/) can start to point you in the right direction. Not sure what else to look for in a big data product? Our features and requirements article (/big-data-analytics/big-data-analytics-requirements/) provides insight on what to look for and our customizable tool scorecard (https://pmo.selecthub.com/request-custom-scorecard/?category=Big%20Data%20Analytics%20Tools) ranks products in areas like text, content, statistical, social media and spatial analytics.

What more do you want to know about structures in big data? Which type of big data has been the most beneficial to your business? What questions do you have that we didn't answer here? Sound off in the comments section.

2 comments

Join the conversation

‹ ›

Pramod Kumar Sharma - November 26, 2024
reply (https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/?replytocom=195142#respond)

Thanks, this is a good learning course.

shyam singh - August 10, 2023
reply (https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/?replytocom=179247#respond)

This was a good learning course. Thank you.

Leave a Reply

Your email address will not be published. Required fields are marked *

Your message

Your name *

Your email *

Save my name, email, and website in this browser for the next time I comment.

Post Comment

Search

Compare the Top Big Data Analytics Tools

Pricing, Ratings, and Reviews for each Vendor. PLUS… Access to our online selection platform for free.

(https://pmo.selecthub.com/request-custom-leaderboard/?category_slug=big-data-analytics-tools&product_slug&slug=big-data-analytics-tools&category=Big%20Data%20Analytics%20Tools&scorecard_id=266)

Requirements Template for Big Data Analytics Tools

Jump-start your selection project with a free, pre-built, customizable Big Data Analytics Tools requirements template.

(https://pmo.selecthub.com/big-data-requirements-onsite/)

Why SelectHub (/why/)
Browse Products (/categories/)
Research Methodology (/research-methodology/)
Editorial Guidelines (/editorial-guidelines/)
Create a Project (https://app.selecthub.com/projects/new)
Dashboard (https://app.selecthub.com/dashboard)
Managed Selection Services (https://www.selecthub.com/managed-selection-services/)
Claim Your Product Listing (https://pmo.selecthub.com/claim-your-product/)
For Vendors (/seller/)
Thought Leader Program (/thought-leaders/)
Awards Program (/awards/)
Careers (https://www.selecthub.com/careers/)
About Us (https://www.selecthub.com/about/)
Privacy Policy (/policies/)

611 S. Congress Avenue, Suite 130
Austin, TX, 78704. 855.850.3850

(https://www.facebook.com/selecthub) (https://twitter.com/selecthub)
(https://www.linkedin.com/company/selecthub/) (https://www.youtube.com/@SelectHub)

© 2024, SelectHub. All rights reserved. Various trademarks held by their respective owners.
All original content is copyrighted by SelectHub and any copying or reproduction (without references to SelectHub) is strictly prohibited.