**Top 15 Big Data Tools (Big Data Analytics Tools) In 2022**

**List and Comparison of the top open source Big Data Tools and Techniques for Data Analysis:**

As we all know, data is everything in today’s IT world. Moreover, this data keeps multiplying by manifolds each day.

Earlier, we used to talk about kilobytes and megabytes. But nowadays, we are talking about terabytes.

Data is meaningless until it turns into useful information and knowledge which can aid the management in decision making. For this purpose, we have several top big data software available in the market. This software help in storing, analyzing, reporting and doing a lot more with data.



**What You Will Learn:**[[hide](https://www.softwaretestinghelp.com/big-data-tools/)]

* [Top 15 Big Data Tools for Data Analysis](https://www.softwaretestinghelp.com/big-data-tools/#Top_15_Big_Data_Tools_for_Data_Analysis)
  + [#1) Integrate.io](https://www.softwaretestinghelp.com/big-data-tools/#1_Integrateio)
  + [#2) Adverity](https://www.softwaretestinghelp.com/big-data-tools/#2_Adverity)
  + [#3) Dataddo](https://www.softwaretestinghelp.com/big-data-tools/#3_Dataddo)
  + [#4) Apache Hadoop](https://www.softwaretestinghelp.com/big-data-tools/#4_Apache_Hadoop)
  + [#5) CDH (Cloudera Distribution for Hadoop)](https://www.softwaretestinghelp.com/big-data-tools/#5_CDH_Cloudera_Distribution_for_Hadoop)
  + [#6) Cassandra](https://www.softwaretestinghelp.com/big-data-tools/#6_Cassandra)
  + [#7) Knime](https://www.softwaretestinghelp.com/big-data-tools/#7_Knime)
  + [#8) Datawrapper](https://www.softwaretestinghelp.com/big-data-tools/#8_Datawrapper)
  + [#9) MongoDB](https://www.softwaretestinghelp.com/big-data-tools/#9_MongoDB)
  + [#10) Lumify](https://www.softwaretestinghelp.com/big-data-tools/#10_Lumify)
  + [#11) HPCC](https://www.softwaretestinghelp.com/big-data-tools/#11_HPCC)
  + [#12) Storm](https://www.softwaretestinghelp.com/big-data-tools/#12_Storm)
  + [#13) Apache SAMOA](https://www.softwaretestinghelp.com/big-data-tools/#13_Apache_SAMOA)
  + [#14) Talend](https://www.softwaretestinghelp.com/big-data-tools/#14_Talend)
  + [#15) Rapidminer](https://www.softwaretestinghelp.com/big-data-tools/#15_Rapidminer)
  + [#16) Qubole](https://www.softwaretestinghelp.com/big-data-tools/#16_Qubole)
  + [#17) Tableau](https://www.softwaretestinghelp.com/big-data-tools/#17_Tableau)
  + [#18) R](https://www.softwaretestinghelp.com/big-data-tools/#18_R)

…………………………………………………………….

### #1) Integrate.io

[Integrate io_logo](https://www.integrate.io/?utm_source=software-testing-help&utm_medium=referral+&utm_campaign=ecommerce)

[**Integrate.io**](https://www.integrate.io/?utm_source=software-testing-help&utm_medium=referral+&utm_campaign=ecommerce) is a platform to integrate, process, and prepare data for analytics on the cloud. It will bring all your data sources together. Its intuitive graphic interface will help you with implementing ETL, ELT, or a replication solution.

Integrate.io is a complete toolkit for building data pipelines with low-code and no-code capabilities. It has solutions for marketing, sales, support, and developers.

Integrate.io will help you make the most out of your data without investing in hardware, software, or related personnel. Integrate.io provides support through email, chats, phone, and an online meetings.

**Pros:**

* Integrate.io is an elastic and scalable cloud platform.
* You will get immediate connectivity to a variety of data stores and a rich set of out-of-the-box data transformation components.
* You will be able to implement complex data preparation functions by using Integrate.io’s rich expression language.
* It offers an API component for advanced customization and flexibility.

**Cons:**

* Only the annual billing option is available. It doesn’t allow you for the monthly subscription.

**Pricing:** You can get a quote for pricing details. It has a subscription-based pricing model. You can try the platform for free for 7-days.

### #2) Adverity

[Adverity Logo](https://hubs.ly/H0RjyDs0)

[**Adverity**](https://hubs.ly/H0RjyDs0) is a flexible end-to-end marketing analytics platform that enables marketers to track marketing performance in a single view and effortlessly uncover new insights in real-time.

Thanks to automated data integration from over 600 sources, powerful data visualizations, and AI-powered predictive analytics, Adverity enables marketers to track marketing performance in a single view and effortlessly uncovers new insights in real-time.

This results in data-backed business decisions, higher growth, and measurable ROI.

**Pros**

* Fully automated data integration from over 600 data sources.
* Fast data handling and transformations at once.
* Personalized and out-of-the-box reporting.
* Customer-driven approach
* High scalability and flexibility
* Excellent customer support
* High security and governance
* Strong built-in predictive analytics
* Easily analyze cross-channel performance with ROI Advisor.

**Pricing:** The subscription-based pricing model is available upon request.

### #3) Dataddo

[Logo Dataddo](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2021/09/Logo-Dataddo.png)

[**Dataddo**](https://www.dataddo.com/integration/?utm_source=softwaretestinghelp&utm_medium=paid_media&utm_campaign=paid-listings&utm_content=Big-Data-Tools) is a no-coding, cloud-based ETL platform that puts flexibility first – with a wide range of connectors and the ability to choose your own metrics and attributes, Dataddo makes creating stable data pipelines simple and fast.

Dataddo seamlessly plugs into your existing data stack, so you don’t need to add elements to your architecture that you weren’t already using, or change your basic workflows. Dataddo’s intuitive interface and quick set-up lets you focus on integrating your data, rather than wasting time learning how to use yet another platform.

**Pros:**

* Friendly for non-technical users with a simple user interface.
* Can deploy data pipelines within minutes of account creation.
* Flexibly plugs into users’ existing data stack.
* No-maintenance: API changes managed by the Dataddo team.
* New connectors can be added within 10 days from request.
* Security: GDPR, SOC2, and ISO 27001 compliant.
* Customizable attributes and metrics when creating sources.
* Central management system to track the status of all data pipelines simultaneously.

### #4) Apache Hadoop



Apache Hadoop is a software framework employed for clustered file system and handling of big data. It processes datasets of big data by means of the MapReduce programming model.

Hadoop is an open-source framework that is written in Java and it provides cross-platform support.

No doubt, this is the topmost big data tool. In fact, over half of the Fortune 50 companies use Hadoop. Some of the Big names include Amazon Web services, Hortonworks, IBM, Intel, Microsoft, Facebook, etc.

**Pros**:

* The core strength of Hadoop is its HDFS (Hadoop Distributed File System) which has the ability to hold all type of data – video, images, JSON, XML, and plain text over the same file system.
* Highly useful for R&D purposes.
* Provides quick access to data.
* Highly-available service resting on a cluster of computers

**Cons**:

* Sometimes disk space issues can be faced due to its 3x data redundancy.
* I/O operations could have been optimized for better performance.

**Pricing:** This software is free to use under the Apache License.

### #5) CDH**(Cloudera Distribution for Hadoop)**

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/CDH.png)

CDH aims at enterprise-class deployments of that technology. It is totally open source and has a free platform distribution that encompasses Apache Hadoop, Apache Spark, Apache Impala, and many more.

It allows you to collect, process, administer, manage, discover, model, and distribute unlimited data.

**Pros**:

* Comprehensive distribution
* Cloudera Manager administers the Hadoop cluster very well.
* Easy implementation.
* Less complex administration.
* High security and governance

**Cons**:

* Few complicating UI features like charts on the CM service.
* Multiple recommended approaches for installation sounds confusing.

However, the Licensing price on a per-node basis is pretty expensive.

**Pricing:** CDH is a free software version by Cloudera. However, if you are interested to know the cost of the Hadoop cluster then the per-node cost is around $1000 to $2000 per terabyte.

### #6) Cassandra

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Cassandra.png)

Apache Cassandra is free of cost and open-source distributed NoSQL DBMS constructed to manage huge volumes of data spread across numerous commodity servers, delivering high availability. It employs CQL (Cassandra Structure Language) to interact with the database.

Some of the high-profile companies using Cassandra include Accenture, American Express, Facebook, General Electric, Honeywell, Yahoo, etc.

**Pros**:

* No single point of failure.
* Handles massive data very quickly.
* Log-structured storage
* Automated replication
* Linear scalability
* Simple Ring architecture

**Cons**:

* Requires some extra efforts in troubleshooting and maintenance.
* Clustering could have been improved.
* Row-level locking feature is not there.

**Pricing:** This tool is free.

### #7) Knime

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Knime.png)

KNIME stands for Konstanz Information Miner which is an open source tool that is used for Enterprise reporting, integration, research, CRM, data mining, data analytics, text mining, and business intelligence. It supports Linux, OS X, and Windows operating systems.

It can be considered as a good alternative to SAS. Some of the top companies using Knime include Comcast, Johnson & Johnson, Canadian Tire, etc.

**Pros:**

* Simple ETL operations
* Integrates very well with other technologies and languages.
* Rich algorithm set.
* Highly usable and organized workflows.
* Automates a lot of manual work.
* No stability issues.
* Easy to set up.

**Cons:**

* Data handling capacity can be improved.
* Occupies almost the entire RAM.
* Could have allowed integration with graph databases.

**Pricing:** Knime platform is free. However, they offer other commercial products which extend the capabilities of the Knime analytics platform.

### #8) Datawrapper

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Datawrapper.png)

Datawrapper is an open source platform for data visualization that aids its users to generate simple, precise and embeddable charts very quickly.

Its major customers are newsrooms that are spread all over the world. Some of the names include The Times, Fortune, Mother Jones, Bloomberg, Twitter etc.

**Pros:**

* Device friendly. Works very well on all type of devices – mobile, tablet or desktop.
* Fully responsive
* Fast
* Interactive
* Brings all the charts in one place.
* Great customization and export options.
* Requires zero coding.

**Cons:**Limited color palettes

**Pricing:** It offers free service as well as customizable paid options as mentioned below.

* Single user, occasional use: 10K
* Single user, daily use: 29 €/month
* For a professional Team: 129€/month
* Customized version: 279€/month
* Enterprise version: 879€+

### #9) MongoDB

[Mongo DB](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Mongo-DB.png)

MongoDB is a NoSQL, document-oriented database written in C, C++, and JavaScript. It is free to use and is an open source tool that supports multiple operating systems including Windows Vista ( and later versions), OS X (10.7 and later versions), Linux, Solaris, and FreeBSD.

Its main features include Aggregation, Adhoc-queries, Uses BSON format, Sharding, Indexing, Replication, Server-side execution of javascript, Schemaless, Capped collection, MongoDB management service (MMS), load balancing and file storage.

Some of the major customers using MongoDB include Facebook, eBay, MetLife, Google, etc.

**Pros:**

* Easy to learn.
* Provides support for multiple technologies and platforms.
* No hiccups in installation and maintenance.
* Reliable and low cost.

**Cons:**

* Limited analytics.
* Slow for certain use cases.

**Pricing:** MongoDB’s SMB and enterprise versions are paid and its pricing is available on request.

### #10) Lumify

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/lumify.png)

Lumify is a free and open source tool for big data fusion/integration, analytics, and visualization.

Its primary features include full-text search, 2D and 3D graph visualizations, automatic layouts, link analysis between graph entities, integration with mapping systems, geospatial analysis, multimedia analysis, real-time collaboration through a set of projects or workspaces.

**Pros:**

* Scalable
* Secure
* Supported by a dedicated full-time development team.
* Supports the cloud-based environment. Works well with Amazon’s AWS.

**Pricing:** This tool is free.

### #11) HPCC

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/HPCC.png)

HPCC stands for **H**igh-**P**erformance **C**omputing **C**luster. This is a complete big data solution over a highly scalable supercomputing platform. HPCC is also referred to as DAS (**Data** **A**nalytics **S**upercomputer). This tool was developed by LexisNexis Risk Solutions.

This tool is written in C++ and a data-centric programming language knowns as ECL(Enterprise Control Language). It is based on a Thor architecture that supports data parallelism, pipeline parallelism, and system parallelism. It is an open-source tool and is a good substitute for Hadoop and some other Big data platforms.

**Pros:**

* The architecture is based on commodity computing clusters which provide high performance.
* Parallel data processing.
* Fast, powerful and highly scalable.
* Supports high-performance online query applications.
* Cost-effective and comprehensive.

**Pricing:** This tool is free.

### #12) Storm

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Storm.png)

Apache Storm is a cross-platform, distributed stream processing, and fault-tolerant real-time computational framework. It is free and open-source. The developers of the storm include Backtype and Twitter. It is written in Clojure and Java.

Its architecture is based on customized spouts and bolts to describe sources of information and manipulations in order to permit batch, distributed processing of unbounded streams of data.

Among many, Groupon, Yahoo, Alibaba, and The Weather Channel are some of the famous organizations that use Apache Storm.

**Pros:**

* Reliable at scale.
* Very fast and fault-tolerant.
* Guarantees the processing of data.
* It has multiple use cases – real-time analytics, log processing, ETL (Extract-Transform-Load), continuous computation, distributed RPC, machine learning.

**Cons:**

* Difficult to learn and use.
* Difficulties with debugging.
* Use of Native Scheduler and Nimbus become bottlenecks.

**Pricing:** This tool is free.

### #13) Apache SAMOA

SAMOA stands for Scalable Advanced Massive Online Analysis. It is an open-source platform for big data stream mining and machine learning.

It allows you to create distributed streaming machine learning (ML) algorithms and run them on multiple DSPEs (distributed stream processing engines). Apache SAMOA’s closest alternative is BigML tool.

**Pros:**

* Simple and fun to use.
* Fast and scalable.
* True real-time streaming.
* Write Once Run Anywhere (WORA) architecture.

**Pricing:** This tool is free.

### #14) Talend

**Talend Big data integration products include:**

* Open studio for Big data: It comes under free and open source license. Its components and connectors are Hadoop and NoSQL. It provides community support only.
* Big data platform: It comes with a user-based subscription license. Its components and connectors are MapReduce and Spark. It provides Web, email, and phone support.
* Real-time big data platform: It comes under a user-based subscription license. Its components and connectors include Spark streaming, Machine learning, and IoT. It provides Web, email, and phone support.

**Pros:**

* Streamlines ETL and ELT for Big data.
* Accomplish the speed and scale of spark.
* Accelerates your move to real-time.
* Handles multiple data sources.
* Provides numerous connectors under one roof, which in turn will allow you to customize the solution as per your need.

**Cons:**

* Community support could have been better.
* Could have an improved and easy to use interface
* Difficult to add a custom component to the palette.

**Pricing:** Open studio for big data is free. For the rest of the products, it offers subscription-based flexible costs. On average, it may cost you an average of $50K for 5 users per year. However, the final cost will be subject to the number of users and edition.

Each product is having a free trial available.

### #15) Rapidminer

[Rapidminer](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Rapidminer.png)

Rapidminer is a cross-platform tool which offers an integrated environment for data science, machine learning and predictive analytics. It comes under various licenses that offer small, medium and large proprietary editions as well as a free edition that allows for 1 logical processor and up to 10,000 data rows.

Organizations like Hitachi, BMW, Samsung, Airbus, etc have been using RapidMiner.

**Pros:**

* Open-source Java core.
* The convenience of front-line data science tools and algorithms.
* Facility of code-optional GUI.
* Integrates well with APIs and cloud.
* Superb customer service and technical support.

**Cons:**Online data services should be improved.

**Pricing:** The commercial price of Rapidminer starts at $2.500.

The small enterprise edition will cost you $2,500 User/Year. The medium enterprise edition will cost you $5,000 User/Year. The Large enterprise edition will cost you $10,000 User/Year. Check the website for the complete pricing information.

### #16) Qubole

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Qubole.png)

Qubole data service is an independent and all-inclusive Big data platform that manages, learns and optimizes on its own from your usage. This lets the data team concentrate on business outcomes instead of managing the platform.

Out of the many, few famous names that use Qubole include Warner music group, Adobe, and Gannett. The closest competitor to Qubole is Revulytics.

**Pros:**

* Faster time to value.
* Increased flexibility and scale.
* Optimized spending
* Enhanced adoption of Big data analytics.
* Easy to use.
* Eliminates vendor and technology lock-in.
* Available across all regions of the AWS worldwide.

**Pricing:** Qubole comes under a proprietary license which offers business and enterprise edition. The business edition is **free of cost** and supports up to **5 users**.

The **enterprise edition** is subscription-based and paid. It is suitable for big organizations with multiple users and uses cases. Its pricing starts from **$199/mo**. You need to contact the Qubole team to know more about the Enterprise edition pricing.

### #17) Tableau

[tableau](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Tableau.png)

Tableau is a software solution for business intelligence and analytics which present a variety of integrated products that aid the world’s largest organizations in visualizing and understanding their data.

The software contains three main products i.e.Tableau Desktop (for the analyst), Tableau Server (for the enterprise) and Tableau Online (to the cloud). Also, Tableau Reader and Tableau Public are the two more products that have been recently added.

Tableau is capable of handling all data sizes and is easy to get to for technical and non-technical customer base and it gives you real-time customized dashboards. It is a great tool for data visualization and exploration.

Out of the many, few famous names that use Tableau includes Verizon Communications, ZS Associates, and Grant Thornton. The closest alternative tool of Tableau is the looker.

**Pros:**

* Great flexibility to create the type of visualizations you want (as compared with its competitor products).
* Data blending capabilities of this tool are just awesome.
* Offers a bouquet of smart features and is razor sharp in terms of its speed.
* Out of the box support for connection with most of the databases.
* No-code data queries.
* Mobile-ready, interactive and shareable dashboards.

**Cons:**

* Formatting controls could be improved.
* Could have a built-in tool for deployment and migration amongst the various tableau servers and environments.

**Pricing:** Tableau offers different editions for desktop, server and online. Its pricing **starts from $35/month**. Each edition has a free trial available.

**Let us take a look at the cost of each edition:**

* Tableau Desktop personal edition: $35 USD/user/month (billed annually).
* Tableau Desktop Professional edition: $70 USD/user/month (billed annually).
* Tableau Server On-Premises or public cloud: $35 USD/user/month (billed annually).
* Tableau Online Fully Hosted: $42 USD/user/month (billed annually).

### #18) R

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/R.png)

R is one of the most comprehensive statistical analysis packages. It is open-source, free, multi-paradigm and dynamic software environment. It is written in C, Fortran and R programming languages.

It is broadly used by statisticians and data miners. Its use cases include data analysis, data manipulation, calculation, and graphical display.

**Pros:**

* R’s biggest advantage is the vastness of the package ecosystem.
* Unmatched Graphics and charting benefits.

**Cons:**Its shortcomings include memory management, speed, and security.

**Pricing:** The R studio IDE and shiny server are free.

**In addition to this, R studio offers some enterprise-ready professional products:**

* RStudio commercial desktop license: $995 per user per year.
* RStudio server pro commercial license: $9,995 per year per server (supports unlimited users).
* RStudio connect price varies from $6.25 per user/month to $62 per user/month.
* RStudio Shiny Server Pro will cost $9,995 per year.

### Additional Tools

**#19) Elasticsearch**

Elastic search is a cross-platform, open-source, distributed, RESTful search engine based on Lucene.

It is one of the most popular enterprise search engines. It comes as an integrated solution in conjunction with Logstash (data collection and log parsing engine) and Kibana (analytics and visualization platform) and the three products together are called as an Elastic stack

**#20) OpenRefine**

OpenRefine is a free, open source data management and data visualization tool for operating with messy data, cleaning, transforming, extending and improving it. It supports Windows, Linux, and macOD platforms.

**#21) Stata wing**

Statwing is a friendly to use statistical tool that has analytics, time series, forecasting and visualization features. Its starting price is $50.00/month/user. A free trial is also available.

**#22) CouchDB**

Apache CouchDB is an open source, cross-platform, document-oriented NoSQL database that aims at ease of use and holding a scalable architecture. It is written in concurrency-oriented language Erlang.

**.**

**#23) Pentaho**

Pentaho is a cohesive platform for data integration and analytics. It offers real-time data processing to boost digital insights. The software comes in enterprise and community editions. A free trial is also available.

**.**

**#24) Flink**

Apache Flink is an open-source, cross-platform distributed stream processing framework for data analytics and machine learning. This is written in Java and Scala. It is fault tolerant, scalable and high-performing.

**#25) DataCleaner**

Quadient DataCleaner is a Python-based data quality solution that programmatically cleans data sets and prepares them for analysis and transformation.

**#26) Kaggle**

Kaggle is a data science platform for predictive modeling competitions and hosted public datasets. It works on the crowdsourcing approach to come up with the best models.

**#27) Hive**

Apache Hive is a java based cross-platform data warehouse tool that facilitates data summarization, query, and analysis.

**.**

**#28) Spark**

[Apache Spark](https://www.softwaretestinghelp.com/apache-spark-tutorial/) is an open source framework for data analytics, machine learning algorithms, and fast cluster computing. This is written in Scala, Java, Python, and R.

**#29) IBM SPSS Modeler**

SPSS is a proprietary software for data mining and predictive analytics. This tool provides a drag and drag interface to do everything from data exploration to machine learning. It is a very powerful, versatile, scalable and flexible tool.

**#30) OpenText**

OpenText Big data analytics is a high performing comprehensive solution designed for business users and analysts which allows them to access, blend, explore and analyze data easily and quickly.

OpenText Big data analytics is a high performing comprehensive solution designed for business users and analysts which allows them to access, blend, explore and analyze data easily and quickly.

**#31) Oracle Data Mining**

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/ODM.png)

ODM is a proprietary tool for data mining and specialized analytics that allows you to create, manage, deploy and leverage Oracle data and investment

**#32) Teradata**

[Teradata](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Teradata.png)

Teradata company provides data warehousing products and services. Teradata analytics platform integrates analytic functions and engines, preferred analytic tools, AI technologies and languages, and multiple data types in a single workflow.

**#33) BigML**

[BigML](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/BigML.png)

Using BigML, you can build superfast, real-time predictive apps. It gives you a managed platform through which you create and share the dataset and models.

**#34) Silk**

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Silk.png)

Silk is a linked data paradigm based, open source framework that mainly aims at integrating heterogeneous data sources.

**#35) CartoDB**

[](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Carto-DB.png)

CartoDB is a freemium SaaS cloud computing framework that acts as a location intelligence and data visualization tool.

**#36) Charito**

[Charito](https://www.softwaretestinghelp.com/wp-content/qa/uploads/2018/06/Charito.png)

Charito is a simple and powerful data exploration tool that connects to the majority of popular data sources. It is built on SQL and offers very easy & quick cloud-based deployments.

**.**

**#37) Plot.ly**

**Plot.ly** holds a GUI aimed at bringing in and analyzing data into a grid and utilizing stats tools. Graphs can be embedded or downloaded. It creates the graphs very quickly and efficiently.