**Chapter VIII Data Analysis Tools and Technique**

IG DATA is a term used for a collection of data sets so large and complex that it is difficult to process using traditional applications/tools. It is the data exceeding Terabytes in size. Because of the variety of data that it encompasses, big data always brings a number of challenges relating to its volume and complexity. A recent survey says that 80% of the data created in the world are unstructured. One challenge is how these unstructured data can be structured, before we attempt to understand and capture the most important data. Another challenge is how we can store it. Here are the top tools used to store and analyse Big Data. We can categorise them into two (storage and Querying/Analysis).

**1. Apache Hadoop**

Apache Hadoop is a java based free software framework that can effectively store large amount of data in a cluster. This framework runs in parallel on a cluster and has an ability to allow us to process data across all nodes. Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster. This also replicates data in a cluster thus providing high availability.

**2. Microsoft HDInsight**

It is a Big Data solution from Microsoft powered by Apache Hadoop which is available as a service in the cloud. HDInsight uses Windows Azure Blob storage as the default file system. This also provides high availability with low cost.

**3. NoSQL**

While the traditional SQL can be effectively used to handle large amount of structured data, we need NoSQL (Not Only SQL) to handle unstructured data. NoSQL databases store unstructured data with no particular schema. Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data. There are many open-source NoSQL DBs available to analyse big Data.

**4. Hive**

This is a distributed data management for Hadoop. This supports SQL-like query option HiveSQL (HSQL) to access big data. This can be primarily used for Data mining purpose. This runs on top of Hadoop.

**5. Sqoop**

This is a tool that connects Hadoop with various relational databases to transfer data. This can be effectively used to transfer structured data to Hadoop or Hive.

**6. PolyBase**

This works on top of SQL Server 2012 Parallel Data Warehouse (PDW) and is used to access data stored in PDW. PDW is a data ware housing appliance built for processing any volume of relational data and provides an integration with Hadoop allowing us to access non-relational data as well.

**7. Big data in EXCEL**

As many people are comfortable in doing analysis in EXCEL, a popular tool from Microsoft, you can also connect data stored in Hadoop using EXCEL 2013. Hortonworks, which is primarily working in providing Enterprise Apache Hadoop, provides an option to access big data stored in their Hadoop platform using EXCEL 2013. You can use Power View feature of EXCEL 2013 to easily summarise the data. (More information).

Similarly, Microsoft’s HDInsight allows us to connect to big data stored in Azure cloud using a power query option. (More information).

**8. Presto**

Facebook has developed and recently open-sourced its Query engine (SQL-on-Hadoop) named Presto which is built to handle petabytes of data. Unlike Hive, Presto does not depend on MapReduce technique and can quickly retrieve data.

Here are thousands of Big Data tools out there for data analysis today. Data analysis is the process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. To save your time, in this post, I will list 30 top big data tools for data analysis in the areas of**open source data tools, data visualization tools, sentiment tools, data extraction tools and databases.**

**Open Source Data Tools**

**1. Knime**

KNIME Analytics Platform is the leading open solution for data-driven innovation, helping you discover the potential hidden in your data, mine for fresh insights, or predict new futures.

With more than 1000 modules, hundreds of ready-to-run examples, a comprehensive range of integrated tools, and the widest choice of advanced algorithms available, KNIME Analytics Platform is the perfect toolbox for any data scientist.

**2. OpenRefine**

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it, transforming it from one format into another, and extending it with web services and external data. OpenRefine can help you explore large data sets with ease.

**3. R-Programming**

What if I tell you that Project R, a GNU project, is written in R itself? It’s primarily written in C and FORTRAN. And a lot of its modules are written in R itself. It’s a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R’s popularity substantially in recent years.

Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others.

**4. Orange**

Orange is open source data visualization and data analysis for novice and expert, and provides interactive workflows with a large toolbox to create interactive workflows to analyse and visualize data. Orange is packed with different visualizations, from scatter plots, bar charts, trees, to dendrograms, networks and heat maps.

**5. RapidMiner**

Much like KNIME, RapidMiner operates through visual programming and is capable of manipulating, analysing and modelling data. RapidMiner makes data science teams more productive through an open source platform for data prep, machine learning, and model deployment. Its unified data science platform accelerates the building of complete analytical workflows – from data prep to machine learning to model validation to deployment – in a single environment, dramatically improving efficiency and shortening the time to value for data science projects.

**6. Pentaho**

Pentaho addresses the barriers that block your organization’s ability to get value from all your data. The platform simplifies preparing and blending any data and includes a spectrum of tools to easily analyse, visualize, explore, report and predict. Open, embeddable and extensible, Pentaho is architected to ensure that each member of your team — from developers to business users — can easily translate data into value.

**7. Talend**

Talend is the leading open source integration software provider to data-driven enterprises. Our customers connect anywhere, at any speed. From ground to cloud and batch to streaming, data or application integration, Talend connects at big data scale, 5x faster and at 1/5th the cost.

**8. Weka**

Weka, an open source software, is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a data set or called from your own JAVA code. It is also well suited for developing new machine learning schemes, since it was fully implemented in the JAVA programming language, plus supporting several standard data mining tasks.

For someone who hasn’t coded for a while, Weka with its GUI provides easiest transition into the world of Data Science. Being written in Java, those with Java experience can call the library into their code as well.

**9. NodeXL**

NodeXL is a data visualization and analysis software of relationships and networks. NodeXL provides exact calculations. It is a free (not the pro one) and open-source network analysis and visualization software. It is one of the best statistical tools for data analysis which includes advanced network metrics, access to social media network data importers, and automation.

**10. Gephi**

Gephi is also an open-source network analysis and visualization software package written in Java on the NetBeans platform. Think of the giant friendship maps you see that represent LinkedIn or Facebook connections. Gelphi takes that a step further by providing exact calculations.

**Data Visualization Tools**

**11. Datawrapper**

Datawrapper is an online data-visualization tool for making interactive charts. Once you upload the data from CSV/PDF/Excel file or paste it directly into the field, Datawrapper will generate a bar, line, map or any other related visualization. Datawrapper graphs can be embedded into any website or CMS with ready-to-use embed codes. So many reporters and news organizations use Datawrapper to embed live charts into their articles. It is very easy to use and produces effective graphics.

**12. Solver**

Solver specializes in providing world-class financial reporting, budgeting and analysis with push-button access to all data sources that drive company-wide profitability. Solver provides BI360, which is available for cloud and on-premise deployment, focusing on four key analytics areas.

**13. Qlik**

Qlik lets you create visualizations, dashboards, and apps that answer your company’s most important questions. Now you can see the whole story that lives within your data.

**14. Tableau Public**

Tableau democratizes visualization in an elegantly simple and intuitive tool. It is exceptionally powerful in business because it communicates insights through data visualization. In the analytics process, Tableau’s visuals allow you to quickly investigate a hypothesis, sanity check your gut, and just go explore the data before embarking on a treacherous statistical journey.

**15. Google Fusion Tables**

Fusion TablesMeet Google Spreadsheets cooler, larger, and much nerdier cousin. Google Fusion tables is an incredible tool for data analysis, large data-set visualization, and mapping. Not surprisingly, Google’s incredible mapping software plays a big role in pushing this tool onto the list. Take for instance this map, which I made to look at oil production platforms in the Gulf of Mexico.

**16. Infogram**

Infogram offers over 35 interactive charts and more than 500 maps to help you visualize your data beautifully. Create a variety of charts including column, bar, pie, or word cloud. You can even add a map to your infographic or report to really impress your audience.

**Sentiment Tools**

**17. Opentext**

The OpenText Sentiment Analysis module is a specialized classification engine used to identify and evaluate subjective patterns and expressions of sentiment within textual content. The analysis is performed at the topic, sentence, and document level and is configured to recognize whether portions of text are factual or subjective and, in the latter case, if the opinion expressed within these pieces of content are positive, negative, mixed, or neutral.

**18. Semantria**

Semantria is a tool that offers a unique service approach by gathering texts, tweets, and other comments from clients and analysing them meticulously to derive actionable and highly valuable insights. Semantria offers text analysis via API and Excel plugin. It differs from Lexalytics in that it is offered via API and Excel plugin, and in that it incorporates a bigger knowledge base and uses deep learning.

**19. Trackur**

Trackur’s automated sentiment analysis looks at the specific keyword you are monitoring and then determines if the sentiment towards that keyword is positive, negative or neutral with the document. That’s weighted the most in Trackur algorithm. It could use to monitor all social media and mainstream news, to gain executive insights through trends, keyword discovery, automated sentiment analysis and influence scoring.

**20. SAS Sentiment Analysis**

SAS sentiment analysis automatically extracts sentiments in real time or over a period of time with a unique combination of statistical modelling and rule-based natural language processing techniques. Built-in reports show patterns and detailed reactions. So you can hone in on the sentiments that are expressed.

With ongoing evaluations, you can refine models and adjust classifications to reflect emerging topics and new terms relevant to your customers, organization or industry.

**21. Opinion Crawl**

Opinion Crawl is an online sentiment analysis for current events, companies, products, and people. Opinion Crawl allows visitors to assess Web sentiment on a topic – a person, an event, a company or a product. You can enter a topic and get an ad-hoc sentiment assessment of it. For each topic you get a pie chart showing current real-time sentiment, a list of the latest news headlines, a few thumbnail images, and a tag cloud of key semantic concepts that the public associates with the subject. The concepts allow you to see what issues or events drive the sentiment in a positive or negative way. For more in-depth assessment, the web crawlers would find the latest published content on many popular subjects and current public issues, and calculate sentiment for them on ongoing basis. Then the blog posts would show the trend of sentiment over time, as well as the Positive-to-Negative ratio.

**Data Extraction Tools**

**22. Octoparse**

Octoparse is a free and powerful website crawler used for extracting almost all kind of data you need from the website. You can use Octoparse to rip a website with its extensive functionalities and capabilities. Its point-and-click UI helps non-programmers to quickly get used to Octoparse. It allows you to grab all the text from the website with AJAX, Javaxript and thus you can download almost all the website content and save it as a structured format like EXCEL, TXT, HTML or your databases.

More advanced, it has provided Scheduled Cloud Extraction which enables you to refresh the website and get the latest information from the website.

**23. Content Grabber**

Content Graber is a web crawling software targeted at enterprises. It can extract content from almost any website and save it as structured data in a format of your choice, including Excel reports, XML, CSV and most databases.

It is more suitable for people with advanced programming skills, since it offers many powerful scripting editing, debugging interfaces for people in need. Users are allowed to use C# or VB.NET to debug or write script to control the crawling process programming.

**24. Import.io**

Import.io is a paid web-based data extraction tool to pull information off of websites used to be something reserved for the nerds. Simply highlight what you want and Import.io walks you through and “learns” what you are looking for. From there, Import.io will dig, scrape, and pull data for you to analyse or export.

**25. Parsehub**

Parsehub is a great web crawler that supports collecting data from websites that use AJAX technologies, JavaScript, cookies and etc. Its machine learning technology can read, analyse and then transform web documents into relevant data. As a freeware, you can set up no more than five publice projects in Parsehub. The paid subscription plans allows you to create at least 20 private projects for scraping websites.

**26. Mozenda**

Mozenda is a cloud based web scraping service. It provides many useful utility features for data extraction. Users will be allowed to upload extracted data to cloud storage.

**27. Scraper**

Scraper is a Chrome extension with limited data extraction features but it’s helpful for making online research, and exporting data to Google Spreadsheets. This tool is intended for beginners as well as experts who can easily copy data to the clipboard or store to the spreadsheets using OAuth. Scraper is a free web crawler tool, which works right in your browser and auto-generates smaller XPaths for defining URLs to crawl. It may not offer all-inclusive crawling services, but novices also needn’t tackle messy configurations.

**Databases**

**28. Data.gov**

The US Government pledged last year to make all government data available freely online. This site is the first stage and acts as a portal to all sorts of amazing information on everything from climate to crime.

**29. US Census Bureau**

US Census Bureau is a wealth of information on the lives of US citizens covering population data, geographic data and education.

**30. The CIA World Factbook**

The World Factbook provides information on the history, people, government, economy, geography, communications, transportation, military, and transnational issues for 267 world entities.

**31. PubMed**

PubMed, developed by the National Library of Medicine (NLM), provides free access to MEDLINE, a database of more than 11 million bibliographic citations and abstracts from nearly 4,500 journals in the fields of medicine, nursing, dentistry, veterinary medicine, pharmacy, allied health, health care systems, and pre-clinical sciences. PubMed also contains links to the full-text versions of articles at participating publishers’ Web sites. In addition, PubMed provides access and links to the integrated molecular biology databases maintained by the National Center for Biotechnology Information (NCBI). These databases contain DNA and protein sequences, 3-D protein structure data, population study data sets, and assemblies of complete genomes in an integrated system. Additional NLM bibliographic databases, such as AIDSLINE, are being added to PubMed. PubMed includes “Old Medline.” Old Medline covers 1950-1965. (Updated daily)

Or some years now, [data storage vendors](http://www.enterprisestorageforum.com/storage-management/slideshows/top-10-storage-vendors-with-staying-power.html) have been incorporating big data and analytics tools such as Hadoop into their products. However, the performance requirements of these applications have sometimes proven hard to align with a prodigious need for raw capacity. The typical approach has been to provide an architecture that spans several storage products in order to provide both the performance needed as well as the capacity that these workloads require at a reasonable cost.

“This usually results in a lot of complexity — multiple consoles and systems, for example — as well as high operating costs,” said Laz Vekiarides, chief technology officer of [ClearSky Data](https://www.clearskydata.com/).

But that is changing. A new generation of storage tools seek to add more analytics and deal better with big data. Here are some of the top trends.

1. Beyond Hadoop

The initial wave of [big data storage tools](http://www.enterprisestorageforum.com/storage-technology/top-20-big-data-analytics-solutions-for-major-storage-environments.html) focused on Hadoop and MapReduce. They then began to add analytics via the likes of SAS, Splunk and SAP HANA. But now they are heading into new pastures.

“Storage solutions are expanding their big data and analytics feature sets into fields such as video, image, click steam analytics, telemetry, transcoding and other forms of data transition,” said Greg Schulz, an analyst with [StorageIO Group](https://storageio.com/).

2. Edge Computing

Vekiarides said that the amount of big data that needs to be retained for analytics is growing fast. Therefore, elastic tools that allow users to scale up or down are essential. Edge computing is required to support analytics when billions of tiny computers are generating data everywhere.

“The analytics data sets and their processing can’t reside solely in the cloud because the latency is too great,” said Vekiarides. “There is a need to store and process in nearby locales.”

3. Storage Remains Essential

There has been some talk in the media about data storage becoming irrelevant — nothing more than a commodity or a utility. But not matter how fancy the analytics or how huge the big data application, they must be underpinned with high-performance and high-capacity storage.

“Today’s analytics-based initiatives can't exist without robust storage to support them,” said Vekiarides. “As enterprises prioritize machine data analytics and management, they need to ensure IT infrastructure can support such initiatives and produce return on investment.”

4. Multi-Cloud

Multi-cloud storage is emerging as a key area for storage and big data. The main public cloud storage providers (AWS, Microsoft and Google) continue to enhance their services with integrated big data capabilities, including search (AWS EMR and Athena, for example). In addition, their compute bursting capabilities offer advantages in many forms of compute-intensive, "on-demand" analytics workloads.

“A class of multi-cloud data management solutions is emerging that provides the ability to store and manage data across these key cloud providers, as well as in on premises private clouds,” said Speciale.

5. Embedded Intelligence

No longer is it enough to offer an add-on for intelligence, analytics or big data. Everything is being integrated.

“Look for storage solutions that embed intelligence for metadata, search and policy-based data workflows,” said Speciale. “This can help support the analytics applications used by data scientists when storing, archiving and most importantly, mining big data for its true value.”

6. IoT and Machine Learning

Internet of Things (IoT) devices are generating millions of small files at an accelerated rate. Machine learning is adding to the problem with its ability to chew through vast amounts of data in short order. Storage platforms now must up their game to store more and process it far faster than ever to be able to cope with the latest IoT and machine data applications.

“IoT data must be stored, protected and analyzed using cloud-scale technology, because users cannot rack and stack on-prem NAS devices and backup software fast enough to keep up,”

7. Zero Tolerance

In an always-on, always -connected digital economy, where instant gratification is an expectation, there is intense pressure to move toward zero recovery point objectives (RPOs) and zero recovery time objective (RTOs) for data recovery, from today’s commonplace 24-hour or longer RPOs and RTOs.

“This includes recovery from the thousands of malware and ransomware attacks that penetrate organizations every day, such as recent attacks on the National Institutes of Health in the UK, Maersk in Denmark, Toyota in Japan and FedEX in the U.S.”

8. Hybrid Storage

Mike McNamara, senior manager of product and solution marketing at NetApp Storage, noted that storage infrastructure is constantly evolving to meet scalability demands brought on by big data. As a result, many enterprises, he said, are converting to a hybrid storage model, scaling out to the cloud entirely to utilize capacity or weighing their options to keep data on premise or in a private cloud.

“Modern enterprises have been wary to move to the public cloud for loss of control of their data or concerns over security,” he said. “As data management tools evolve and support cloud migration and integrate strong security measures, enterprises have become more comfortable moving to a public cloud while keeping control of their data.”

9. Vertical Focus

Point tools that sat on top of storage and cloud platforms characterized the early big data and analytics tools. But the latest crop includes several that seek to integrate all these functions into one stack. With that accomplished, the next step is to tailor these tools to certain industries.

“We are beginning to see vertical-focused big data solutions, targeting specific industries like healthcare and manufacturing,”

10. Storage Intelligence

There is a lot going on in the analytics, big data and IoT side that harnesses the underlying storage. But you can sometimes forget to add intelligence *about* storage. McNamara said enterprises are looking for smart management tools that provide integrated analytics within storage. This enables them to conduct resource monitoring and flag when an aspect of the infrastructure is being underutilized, for example.

“From there, enterprises can re-architect their environment to invest in areas where they need the highest performance,” said McNamara.

Oday’s market is flooded with an array of Big Data tools. They bring cost efficiency, better time management into the data analytical tasks. Here is the list of best big data tools with their key features and download links.

### 1) Hadoop:

[](https://www.guru99.com/bigdata-tutorials.html)

The Apache Hadoop software library is a big data framework. It allows distributed processing of large data sets across clusters of computers. It is designed to scale up from single servers to thousands of machines.

**Features:**

* Authentication improvements when using HTTP proxy server
* Specification for Hadoop Compatible Filesystem effort
* Support for POSIX-style filesystem extended attributes
* It offers robust ecosystem that is well suited to meet the analytical needs of developer
* It brings Flexibility In Data Processing
* It allows for faster data Processing

**Download link:** <https://hadoop.apache.org/releases.html>

### 2) HPCC:

[https://www.guru99.com/images/2-2017/092917_0347_Top15BigDat2.png](https://hpccsystems.com/)

[HPCC](https://hpccsystems.com/) is a big data tool developed by LexisNexis Risk Solution. It delivers on a single platform, a single architecture and a single programming language for data processing.

**Features:**

* Highly efficient accomplish big data tasks with far less code.
* Offers high redundancy and availability
* It can be used both for complex data processing on a Thor cluster
* Graphical IDE for simplifies development, testing and debugging
* It automatically optimizes code for parallel processing
* Provide enhance scalability and performance
* ECL code compiles into optimized C++, and it can also extend using C++ libraries

**Download link:** <https://hpccsystems.com/try-now>

### 3) Storm:

[](http://storm.apache.org/)

[Storm](http://storm.apache.org/) is a free and open source big data computation system. It offers distributed real-time, fault-tolerant processing system. With real-time computation capabilities.

**Features:**

* It benchmarked as processing one million 100 byte messages per second per node
* It uses parallel calculations that run across a cluster of machines
* It will automatically restart in case a node dies. The worker will be restarted on another node
* Storm guarantees that each unit of data will be processed at least once or exactly once
* Once deployed Storm is surely easiest tool for Big data analysis

**Download link:** <http://storm.apache.org/downloads.html>

### 4) Qubole:

[](https://www.qubole.com/)

[Qubole](https://www.qubole.com/) Data is Autonomous Big data management platform. It is self-managed, self-optimizing tool which allows the data team to focus on business outcomes.

**Features:**

* Single Platform for every use case
* Open-source Engines, optimized for the Cloud
* Comprehensive Security, Governance, and Compliance
* Provides actionable Alerts, Insights, and Recommendations to optimize reliability, performance, and costs
* Automatically enacts policies to avoid performing repetitive manual actions

**Download link:**<https://www.qubole.com/>

### 5) Cassandra:

[](https://www.guru99.com/cassandra-tutorial.html)

The [Apache Cassandra](https://www.guru99.com/cassandra-tutorial.html) database is widely used today to provide an effective management of large amounts of data.

**Features:**

* Support for replicating across multiple data centres by providing lower latency for users
* Data is automatically replicated to multiple nodes for fault-tolerance
* It is most suitable for applications that can't afford to lose data, even when an entire data centre is down
* Cassandra offers support contracts and services are available from third parties

**Download link:** <http://cassandra.apache.org/download/>

### 6) Statwing:

[https://www.guru99.com/images/2-2017/092917_0347_Top15BigDat6.png](https://www.statwing.com/)

[Statwing](https://www.statwing.com/) is an easy-to-use statistical tool. It was built by and for big data analysts. Its modern interface chooses statistical tests automatically.

**Features:**

* Explore any data in seconds
* Statwing helps to clean data, explore relationships, and create charts in minutes
* It allows creating histograms, scatterplots, heat maps, and bar charts that export to Excel or PowerPoint
* It also translates results into plain English, so analysts unfamiliar with statistical analysis

**Download link:** <https://www.statwing.com/>

### 7) CouchDB:

[](http://couchdb.apache.org/)

[CouchDB](http://couchdb.apache.org/) stores data in JSON documents that can be accessed web or query using JavaScript. It offers distributed scaling with fault-tolerant storage. It allows accessing data by defining the Couch Replication Protocol.

**Features:**

* CouchDB is a single-node database that works like any other database
* It allows running a single logical database server on any number of servers
* It makes use of the ubiquitous HTTP protocol and JSON data format
* Easy replication of a database across multiple server instances
* Easy interface for document insertion, updates, retrieval and deletion
* JSON-based document format can be translatable across different languages

**Download link:** <http://couchdb.apache.org/>

### 8) Pentaho:

[https://www.guru99.com/images/2-2017/092917_0347_Top15BigDat8.png](http://www.pentaho.com/)

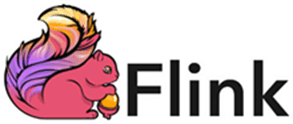
[Pentaho](http://www.pentaho.com/) provides big data tools to extract, prepare and blend data. It offers visualizations and analytics that change the way to run any business. This Big data tool allows turning big data into big insights.

**Features:**

* Data access and integration for effective data visualization
* It empowers users to architect big data at the source and stream them for accurate analytics
* Seamlessly switch or combine data processing with in-cluster execution to get maximum processing
* Allow checking data with easy access to analytics, including charts, visualizations, and reporting
* Supports wide spectrum of big data sources by offering unique capabilities

**Download link:** <http://www.pentaho.com/download>

### 9) Flink:

[](https://flink.apache.org/)

Apache [Flink](https://flink.apache.org/) is an open-source stream processing big data tool. It is distributed, high-performing, always-available, and accurate data streaming applications.

**Features:**

* Provides results that are accurate, even for out-of-order or late-arriving data
* It is stateful and fault-tolerant and can recover from failures
* It can perform at a large scale, running on thousands of nodes
* Has good throughput and latency characteristics
* This big data tool supports stream processing and windowing with event time semantics
* It supports flexible windowing based on time, count, or sessions to data-driven windows
* It supports a wide range of connectors to third-party systems for data sources and sinks

**Download link:** <https://flink.apache.org/>

### 10) Cloudera:

[https://www.guru99.com/images/2-2017/092917_0347_Top15BigDat12.png](https://www.cloudera.com/)

[Cloudera](https://www.cloudera.com/) is the fastest, easiest and highly secure modern big data platform. It allows anyone to get any data across any environment within single, scalable platform.

**Features:**

* High-performance analytics
* It offers provision for multi-cloud
* Deploy and manage Cloudera Enterprise across AWS, Microsoft Azure and Google Cloud Platform
* Spin up and terminate clusters, and only pay for what is needed when need it
* Developing and training data models
* Reporting, exploring, and self-servicing business intelligence
* Delivering real-time insights for monitoring and detection
* Conducting accurate model scoring and serving

**Download link:** <https://www.cloudera.com/>

### 11) Openrefine:

[https://www.guru99.com/images/2-2017/092917_0347_Top15BigDat13.png](http://openrefine.org/)

[Open Refine](http://openrefine.org/) is a powerful big data tool. It helps to work with messy data, cleaning it and transforming it from one format into another. It also allows extending it with web services and external data.

**Features:**

* OpenRefine tool help you explore large data sets with ease
* It can be used to link and extend your dataset with various web services
* Import data in various formats
* Explore datasets in a matter of seconds
* Apply basic and advanced cell transformations
* Allows to deal with cells that contain multiple values
* Create instantaneous links between datasets
* Use named-entity extraction on text fields to automatically identify topics
* Perform advanced data operations with the help of Refine Expression Language

**Download link:** <http://openrefine.org/download.html>

### 12) Rapidminer:

[](https://rapidminer.com/)

[RapidMiner](https://rapidminer.com/) is an open source big data tool. It is used for data prep, machine learning, and model deployment. It offers a suite of products to build new data mining processes and setup predictive analysis.

**Features:**

* Allow multiple data management methods
* GUI or batch processing
* Integrates with in-house databases
* Interactive, shareable dashboards
* Big Data predictive analytics
* Remote analysis processing
* Data filtering, merging, joining and aggregating
* Build, train and validate predictive models
* Store streaming data to numerous databases
* Reports and triggered notifications

**Download link:** [https://my.rapidminer.com/nexus/account/index.html#downloads](https://my.rapidminer.com/nexus/account/index.html)

### 13) DataCleaner:

[](https://datacleaner.org/)

[DataCleaner](https://datacleaner.org/) is a data quality analysis application and a solution platform. It has strong data profiling engine. It is extensible and thereby adds data cleansing, transformations, matching, and merging.

**Feature:**

* Interactive and explorative data profiling
* Fuzzy duplicate record detection
* Data transformation and standardization
* Data validation and reporting
* Use of reference data to cleanse data
* Master the data ingestion pipeline in Hadoop data lake
* Ensure that rules about the data are correct before user spends their time on the processing
* Find the outliers and other devilish details to either exclude or fix the incorrect data

**Download link:** <http://datacleaner.org/>

### 14) Kaggle:

[](https://www.kaggle.com/)

[Kaggle](https://www.kaggle.com/) is the world's largest big data community. It helps organizations and researchers to post their data & statistics. It is the best place to analyse data seamlessly.

**Features:**

* The best place to discover and seamlessly analyse open data
* Search box to find open datasets
* Contribute to the open data movement and connect with other data enthusiasts

**Download link:**<https://www.kaggle.com/>

### 15) Hive:

[](https://www.guru99.com/hive-tutorials.html)

[Hive](https://www.guru99.com/hive-tutorials.html) is an open source-software big data too. It allows programmers analyse large data sets on Hadoop. It helps with querying and managing large datasets real fast.

**Features:**

* It Supports SQL like query language for interaction and Data modelling
* It compiles language with two main tasks map, and reducer
* It allows defining these tasks using Java or Python
* Hive designed for managing and querying only structured data
* Hive's SQL-inspired language separates the user from the complexity of Map Reduce programming
* It offers Java Database Connectivity (JDBC) interface

**Download link:** <https://hive.apache.org/downloads.html>

Big Data Analytics software is widely used in providing meaningful analysis of a large set of data. This software helps in finding current market trends, customer preferences, and other information.

Here are the 11 Top Big Data Analytics Tools with key feature and download links.

### 1) Microsoft HDInsight:

[](https://azure.microsoft.com/en-us/services/hdinsight/)

[Azure HDInsight](https://azure.microsoft.com/en-us/services/hdinsight/) is a Spark and Hadoop service in the cloud. It provides big data cloud offerings in two categories, Standard and Premium. It provides an enterprise-scale cluster for the organization to run their big data workloads.

**Features:**

* Reliable analytics with an industry-leading SLA
* It offers enterprise-grade security and monitoring
* Protect data assets and extend on-premises security and governance controls to the cloud
* High-productivity platform for developers and scientists
* Integration with leading productivity applications
* Deploy Hadoop in the cloud without purchasing new hardware or paying other up-front costs

**Download link:** <https://azure.microsoft.com/en-in/free/>

### 2) Skytree:

[https://www.guru99.com/images/2-2017/092817_0946_Top11BigDat2.png](http://www.skytree.net/)

[Skytree](http://www.skytree.net/) is a big data analytics tool that empowers data scientists to build more accurate models faster. It offers accurate predictive machine learning models that are easy to use.

**Features:**

* Highly Scalable Algorithms
* Artificial Intelligence for Data Scientists
* It allows data scientists to visualize and understand the logic behind ML decisions
* Skytree via the easy-to-adopt GUI or programmatically in Java
* Model Interpretability
* It is designed to solve robust predictive problems with data preparation capabilities
* Programmatic and GUI Access

**Download link:** <http://www.skytree.net/>

### 3) Talend:

[https://www.guru99.com/images/2-2017/092817_0946_Top11BigDat3.png](https://www.talend.com/)

[Talend](https://www.talend.com/) is a big data tool that simplifies and automates big data integration. Its graphical wizard generates native code. It also allows big data integration, master data management and checks data quality.

**Features:**

* Accelerate time to value for big data projects
* Simplify ETL & ELT for big data
* Talend Big Data Platform simplifies using MapReduce and Spark by generating native code
* Smarter data quality with machine learning and natural language processing
* Agile DevOps to speed up big data projects
* Streamline all the DevOps processes

**Download Link:** <https://www.talend.com/download/>

### 4) Splice Machine:

[](https://www.splicemachine.com/)

[Splice Machine](https://www.splicemachine.com/) is a big data analytic tool. Their architecture is portable across public clouds such as AWS, Azure, and Google.

**Features:**

* It can dynamically scale from a few to thousands of nodes to enable applications at every scale
* The Splice Machine optimizer automatically evaluates every query to the distributed HBase regions
* Reduce management, deploy faster, and reduce risk
* Consume fast streaming data, develop, test and deploy machine learning models

**Download link:** [https://www.splicemachine.com](https://www.splicemachine.com/)

### 5) Spark:

[](https://spark.apache.org/)

[Apache Spark](https://spark.apache.org/) is a powerful open source big data analytics tool. It offers over 80 high-level operators that make it easy to build parallel apps. It is used at a wide range of organizations to process large datasets.

**Features:**

* It helps to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk
* It offers lighting Fast Processing
* Support for Sophisticated Analytics
* Ability to Integrate with Hadoop and Existing Hadoop Data
* It provides built-in APIs in Java, Scala, or Python

**Download link:** <https://spark.apache.org/downloads.html>

### 6) Plotly:

[](https://plot.ly/)

[Plotly](https://plot.ly/) is an analytics tool that lets users create charts and dashboards to share online.

**Features:**

* Easily turn any data into eye-catching and informative graphics
* It provides audited industries with fine-grained information on data provenance
* Plotly offers unlimited public file hosting through its free community plan

**Download link:** <https://plot.ly/>

### 7) Apache SAMOA:

[Apache SAMOA](https://samoa.incubator.apache.org/) is a big data analytics tool. It enables development of new ML algorithms. It provides a collection of distributed algorithms for common data mining and machine learning tasks.

**Download link:** <https://samoa.incubator.apache.org/>

### 8) Lumify:

[](http://www.altamiracorp.com/index.php/lumify/)

[Lumify](http://www.altamiracorp.com/index.php/lumify/) is a big data fusion, analysis, and visualization platform. It helps users to discover connections and explore relationships in their data via a suite of analytic options.

**Features:**

* It provides both 2D and 3D graph visualizations with a variety of automatic layouts
* It provides a variety of options for analysing the links between entities on the graph
* It comes with specific ingest processing and interface elements for textual content, images, and videos
* It spaces feature allows you to organize work into a set of projects, or workspaces
* It is built on proven, scalable big data technologies

**Download link:** <http://www.altamiracorp.com/index.php/lumify/>

### 9) Elastic search:

[](http://www.elastic.co/)

[Elastic search](http://www.elastic.co/) is a JSON-based big data search and analytics engine. It is a distributed, RESTful search and analytics engine for solving numbers of use cases. It offers horizontal scalability, maximum reliability, and easy management.

**Features:**

* It allows combine many types of searches such as structured, unstructured, geo, metric, etc.
* Intuitive APIs for monitoring and management give complete visibility and control
* It uses standard RESTful APIs and JSON. It also builds and maintains clients in many languages like Java, Python, NET, and Groovy
* Real-time search and analytics features to work big data by using the Elasticsearch-Hadoop
* It gives an enhanced experience with security, monitoring, reporting, and machine learning features

**Download link:** <https://www.elastic.co/downloads/elasticsearch>

### 10) R-Programming:

[](https://www.r-project.org/)

[R](https://www.r-project.org/) is a language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests.

**Features:**

* Effective data handling and storage facility,
* It provides a suite of operators for calculations on arrays, in particular, matrices,
* It provides coherent, integrated collection of big data tools for data analysis
* It provides graphical facilities for data analysis which display either on-screen or on hardcopy

**Download link:** <https://www.r-project.org/>

### 11) IBM SPSS Modeler:

[IBM SPSS Modeler](https://www.ibm.com/us-en/marketplace/spss-modeler) is a predictive big data analytics platform. It offers predictive models and delivers to individuals, groups, systems and the enterprise. It has a range of advanced algorithms and analysis techniques.

**Features:**

* Discover insights and solve problems faster by analysing structured and unstructured data
* Use an intuitive interface for everyone to learn
* You can select from on-premises, cloud and hybrid deployment options
* Quickly choose the best performing algorithm based on model performance

**Download link:** [https://www.ibm.com/us-en/marketplace/spss-modeler/purchase#product-header-top](https://www.ibm.com/us-en/marketplace/spss-modeler/purchase)

## Why Data Analysis?

Companies that are not leveraging data analytic tools and techniques are falling apart. Since data analytics tools capture in products that automatically glean and analyse data, deliver information and predictions, you can improve prediction accuracy and refine the models. Here we discuss the top 10 Data Analytics Tools for success.



### Goals of Performing Data Analysis

You can analyse data. Extract actionable and commercially relevant information to boost performance. Several extraordinary analytical tools are available, that is the free and open source so that you can leverage it to enhance your business and develop skills.

## Top Data Analytics Tools

Here is the list of top Analytics tools for data analysis that are available for free (for personal use), easy to use (no coding required), well-documented (you can Google your way through if you get stuck), and have powerful capabilities (more than excel). These data analysis tools will help you manage and interpret data in a better and more effective way. Here, we have explored the top 10 Data Analytics tools in Big Data.

### #1 Tableau Public

#### What is Tableau Public

Tableau, one of the top 10 Data Analytics tools, is a simple and intuitive and tool which offers intriguing insights through data visualization. Tableau Public’s million row limit, which is easy to use fares better than most of the other players in the data analytics market.

With Tableau’s visuals, you can investigate a hypothesis, explore the data, and cross-check your insights.

#### Uses of Tableau Public

* You can publish interactive data visualizations to the web for free.
* No programming skills required.
* Visualizations published to Tableau Public can be embedded into blogs and web pages and be shared through email or social media. The shared content can be made available s for downloads.

#### Limitations of Tableau Public

* All data is public and offers very little scope for restricted access
* Data size limitation
* Cannot be connected to
* The only way to read is via OData sources, is Excel or txt.

### #2 OpenRefine

#### What is OpenRefine

Formerly known as GoogleRefine, the data cleaning software that helps you clean up data for analysis. It operates on a row of data which have cells under columns, quite similar to relational database tables.

#### Uses of OpenRefine

* Cleaning messy data
* Transformation of data
* Parsing data from websites
* Adding data to the dataset by fetching it from web services. For instance, OpenRefine could be used for geocoding addresses to geographic coordinates.

#### Limitations of OpenRefine

* Open Refine is unsuitable for large datasets.
* Refine does not work very well with big data.

### #3 KNIME

#### What is KNIME?

KNIME, ranked among the top Data Analytics tools helps you to manipulate, analyse, and model data through visual programming. It is used to integrate various components for data mining and machine learning via its modular data pipelining concept.

#### Uses of KNIME

* Rather than writing blocks of code, you just have to drop and drag connection points between activities.
* This data analysis tool supports programming languages.
* In fact, analysis tools like these can be extended to run chemistry data, text mining, python, and R.

#### Limitation of KNIME

Poor data visualization

### #4 RapidMiner

Same as KNIME. Poor data visualization.

#### What is RapidMiner?

RapidMiner provides machine learning procedures and data mining including data visualization, processing, statistical modelling, deployment, evaluation, and predictive analytics.

RapidMiner, counted among the top 10 Data Analytics tools, is written in the Java and fast gaining acceptance.

#### Uses of RapidMiner

It provides an integrated environment for business analytics, predictive analysis, text mining, data mining, and machine learning.

Along with commercial and business applications,**RapidMiner** is also used for application development, rapid prototyping, training, education, and research.

#### Limitations of RapidMiner

* RapidMiner has size constraints with respect to the number of rows.
* For RapidMiner, you need more hardware resources than ODM and SAS.

### #5 Google Fusion Tables

#### What is Google Fusion Tables?

When talking about Data Analytics tools for free, here comes a much cooler, larger, and nerdier version of Google Spreadsheets. An incredible tool for data analysis, mapping, and large dataset visualization, Google Fusion Tables can be added to business analytics tools list. Ranked among the top 10 Data Analytics tools, Google Fusion Tables is fast gaining popularity.

#### Uses of Google Fusion Tables

* Visualize bigger table data online:
* Filter and summarize across hundreds of thousands of rows.
* Combine tables with other data on the web:

You can merge two or three tables to generate a single visualization that includes sets of data. With Google Fusion Tables, you can combine public data with your own for a better visualization.

You can create a map in minutes!

#### Limitations of Google Fusion Tables

* Only the first 100,000 rows of data in a table are included in query results or mapped.
* The total size of the data sent in one API call cannot be more than 1MB.

[](https://www.digitalvidya.com/blog/data-analytics-tools/#ctcpopup-da)

### #6 NodeXL

#### What is NodeXL?

NodeXL is a free and open-source network analysis and visualization software. Ranked among the top 10 Data Analytics tools, it is one of the best statistical tools for data analysis which includes advanced network metrics, access to social media network data importers, and automation.

#### Uses of NodeXL

**This is one of the best data analysis tools in Excel that helps in:**

1. Data Import
2. Graph Visualization
3. Graph Analysis
4. Data Representation

**NodeXL** integrates into Microsoft Excel 2007, 2010, 2013, and 2016. It opens as a workbook with a variety of worksheets containing the elements of a graph structure like nodes and edges. It can import various graph formats like adjacency matrices, Pajek .net, UCINet .dl, GraphML, and edge lists.

#### Limitations of NodeXL

* Multiple seeding terms are required for a particular problem.
* Need to run the data extractions at slightly different times.

### #7 Wolfram Alpha

#### What is Wolfram Alpha?

Wolfram Alpha, one of the top 10 Data Analytics tools is a computational knowledge engine or answering engine founded by Stephen Wolfram. With Wolfram Alpha, you get answers to factual queries directly by computing the answer from externally sourced ‘curated data’ instead of providing a list of documents or web pages.

#### Uses of Wolfram Alpha

* Is an add-on for Apple’s Siri
* Provides detailed responses to technical searches and solves calculus problems.
* Helps business users with information charts and graphs, and helps in creating topic overviews, commodity information, and high-level pricing history.

#### Limitations of Wolfram Alpha

* Wolfram Alpha can only deal with the publicly known number and facts, not with viewpoints.
* It limits the computation time for each query.

### #8 Google Search Operators

#### What is Google Search Operators?

It is a powerful resource which helps you filter Google results instantly to get most relevant and useful information.

#### Uses of Google Search Operators

* Fast filtering of Google results.
* Google’s powerful data analysis tool can help discover new information or market research.

### #9 Solver

#### What is Excel Solver?

The Solver Add-in is a Microsoft Office Excel add-in program that is available when you install Microsoft Excel or Office. Ranked among the best-known Data Analytic tools is a linear programming and optimization tool in excel? This allows you to set constraints. It is an advanced optimization tool that helps in quick problem-solving.

#### Uses of Solver

The final values found by Solver are a solution to interrelation and decision. It uses a variety of methods, from nonlinear optimization and linear programming to evolutionary and genetic algorithms, to find solutions. It is one of the top 10 Data Analytic tools in use.

#### Limitations of Solver

* Poor scaling is one of the areas where Excel Solver lacks.
* It can affect solution time and quality.
* Solver affects the intrinsic solvability of your model.

### #10 Dataiku DSS

#### What is Dataiku DSS?

Ranked among the top 10 Data Analytic tools, Dataikuis a collaborative data science software platform that helps the team build, prototype, explore, and deliver their own data products more efficiently.

#### Uses of Dataiku DSS

It provides an interactive visual interface where they can build, click, and point or use languages like SQL. This data analytics tool lets you draft data preparation and modulation in seconds. Helps you coordinate development and operations by handling workflow automation, creating predictive web services, model health on a daily basis, and monitoring data.

#### Limitation of Dataiku DSS

* Limited visualization capabilities
* UI hurdles: Reloading of code/datasets
* Inability to easily compile entire code into a single document/notebook
* Still, need to integrate with SPARK

## Big Data Cloud

## Top 5 Data Analytics Tools and Techniques

Here are some of the top Data Analytic tools and techniques that can be used for better performance:

### Visual Analytics

There are different ways to analyse the data. One of the simplest ways to do is to create a graph or visual and look at it to spot patterns. This is an integrated method that combines data analysis with human interaction and data visualization.

### Business Experiments

Experimental design, AB testing, and business experiments are all techniques for testing the validity of something. It is trying out something in one part of the organization and comparing it with another.

### Regression Analysis

It is a statistical tool for investigating the relationship between variables. For instance, the cause and effect relationship between product demand and price.

### Correlation Analysis

A statistical technique that allows you to determine whether there is a relationship between two separate variables and how strong that relationship may be. It is best to use when you know or suspect that there is a relationship between two variables and wish to test the assumption.

### Time Series Analysis

It is the data that is collected at uniformly spaced time intervals. You can use it when you want to assess changes over time or predict future events on the basis of what happened in the past.

**Bonus**: Want to Transform your Career in Data Analytics? Attend Live Data Analytics Orientation Session. Alternatively, you may also go for an advanced certification in Data Analytics.

Application of Storage Technology with Database

NoSQL Database:

NoSQL encompasses a wide variety of different database technologies that were developed in response to the demands presented in building modern applications:

* Developers are working with applications that create massive volumes of new, rapidly changing data types — structured, semi-structured, unstructured and polymorphic data.
* Long gone is the twelve-to-eighteen month waterfall development cycle. Now small teams work in agile sprints, iterating quickly and pushing code every week or two, some even multiple times every day.
* Applications that once served a finite audience are now delivered as services that must be always-on, accessible from many different devices and scaled globally to millions of users.
* Organizations are now turning to scale-out architectures using open source software, commodity servers and cloud computing instead of large monolithic servers and storage infrastructure.

Relational databases were not designed to cope with the scale and agility challenges that face modern applications, nor were they built to take advantage of the commodity storage and processing power available today.

## NoSQL Database Types

* **Document databases** pair each key with a complex data structure known as a document. Documents can contain many different key-value pairs, or key-array pairs, or even nested documents.
* **Graph stores** are used to store information about networks of data, such as social connections. Graph stores include Neo4J and Giraph.
* **Key-value stores** are the simplest NoSQL databases. Every single item in the database is stored as an attribute name (or 'key'), together with its value. Examples of key-value stores are Riak and Berkeley DB. Some key-value stores, such as Redis, allow each value to have a type, such as 'integer', which adds functionality.
* **Wide-column stores** such as Cassandra and HBase are optimized for queries over large datasets, and store columns of data together, instead of rows.

## The Benefits of NoSQL

When compared to relational databases, NoSQL databases are [more scalable and provide superior performance,](https://www.mongodb.com/mongodb-scale) and their data model addresses several issues that the relational model is not designed to address:

* Large volumes of rapidly changing structured, semi-structured, and unstructured data
* Agile sprints, quick schema iteration, and frequent code pushes
* Object-oriented programming that is easy to use and flexible
* Geographically distributed scale-out architecture instead of expensive, monolithic architecture

## Dynamic Schemas

Relational databases require that schemas be defined before you can add data. For example, you might want to store data about your customers such as phone numbers, first and last name, address, city and state – a SQL database needs to know what you are storing in advance.

This fits poorly with [agile development](https://www.mongodb.com/agile-development) approaches, because each time you complete new features, the schema of your database often needs to change. So if you decide, a few iterations into development, that you'd like to store customers' favourite items in addition to their addresses and phone numbers, you'll need to add that column to the database, and then migrate the entire database to the new schema.

If the database is large, this is a very slow process that involves significant downtime. If you are frequently changing the data your application stores – because you are iterating rapidly – this downtime may also be frequent. There's also no way, using a relational database, to effectively address data that's completely unstructured or unknown in advance.

NoSQL databases are built to allow the insertion of data without a predefined schema. That makes it easy to make significant application changes in real-time, without worrying about service interruptions – which means development is faster, code integration is more reliable, and less database administrator time is needed. Developers have typically had to add application-side code to enforce data quality controls, such as mandating the presence of specific fields, data types or permissible values. More sophisticated NoSQL databases allow validation rules to be applied within the database, allowing users to enforce governance across data, while maintaining the agility benefits of a dynamic schema.

## Auto-sharding

Because of the way they are structured, relational databases usually scale vertically – a single server has to host the entire database to ensure acceptable performance for cross- table joins and transactions. This gets expensive quickly, places limits on scale, and creates a relatively small number of failure points for database infrastructure. The solution to support rapidly growing applications is to scale horizontally, by adding servers instead of concentrating more capacity in a single server.

'Sharding' a database across many server instances can be achieved with SQL databases, but usually is accomplished through SANs and other complex arrangements for making hardware act as a single server. Because the database does not provide this ability natively, development teams take on the work of deploying multiple relational databases across a number of machines. Data is stored in each database instance autonomously. Application code is developed to distribute the data, distribute queries, and aggregate the results of data across all of the database instances. Additional code must be developed to handle resource failures, to perform joins across the different databases, for data rebalancing, replication, and other requirements. Furthermore, many benefits of the relational database, such as transactional integrity, are compromised or eliminated when employing manual sharding.

NoSQL databases, on the other hand, usually support auto-sharding, meaning that they natively and automatically spread data across an arbitrary number of servers, without requiring the application to even be aware of the composition of the server pool. Data and query load are automatically balanced across servers, and when a server goes down, it can be quickly and transparently replaced with no application disruption.

[Cloud computing](https://www.mongodb.com/big-data-explained) makes this significantly easier, with providers such as Amazon Web Services providing virtually unlimited capacity on demand, and taking care of all the necessary infrastructure administration tasks. Developers no longer need to construct complex, expensive platforms to support their applications, and can concentrate on writing application code. Commodity servers can provide the same processing and storage capabilities as a single high-end server for a fraction of the price.

## Replication

Most NoSQL databases also support automatic database replication to maintain availability in the event of outages or planned maintenance events. More sophisticated NoSQL databases are fully self-healing, offering automated failover and recovery, as well as the ability to distribute the database across multiple geographic regions to withstand regional failures and enable data localization. Unlike relational databases, NoSQL databases generally have no requirement for separate applications or expensive add-ons to implement replication.

## Integrated Caching

A number of products provide a caching tier for SQL database systems. These systems can improve read performance substantially, but they do not improve write performance, and they add operational complexity to system deployments. If your application is dominated by reads then a distributed cache could be considered, but if your application has just a modest write volume, then a distributed cache may not improve the overall experience of your end users, and will add complexity in managing cache invalidation.

Many NoSQL database technologies have excellent integrated caching capabilities, keeping frequently-used data in system memory as much as possible and removing the need for a separate caching layer. Some NoSQL databases also offer fully managed, integrated in-memory database management layer for workloads demanding the highest throughput and lowest latency.

## NoSQL vs. SQL Summary

|  |  |  |
| --- | --- | --- |
|  | SQL Databases | NoSQL Databases |
| **Types** | One type (SQL database) with minor variations | Many different types including key-value stores, [document databases](https://www.mongodb.com/document-databases), wide-column stores, and graph databases |
| **Development History** | Developed in 1970s to deal with first wave of data storage applications | Developed in late 2000s to deal with limitations of SQL databases, especially scalability, multi-structured data, geo-distribution and agile development sprints |
| **Examples** | MySQL, Postgres, Microsoft SQL Server, Oracle Database | MongoDB, Cassandra, HBase, Neo4j |
| **Data Storage Model** | Individual records (e.g., 'employees') are stored as rows in tables, with each column storing a specific piece of data about that record (e.g., 'manager,' 'date hired,' etc.), much like a spreadsheet. Related data is stored in separate tables, and then joined together when more complex queries are executed. For example, 'offices' might be stored in one table, and 'employees' in another. When a user wants to find the work address of an employee, the database engine joins the 'employee' and 'office' tables together to get all the information necessary. | Varies based on database type. For example, key-value stores function similarly to SQL databases, but have only two columns ('key' and 'value'), with more complex information sometimes stored as BLOBs within the 'value' columns. Document databases do away with the table-and-row model altogether, storing all relevant data together in single 'document' in JSON, XML, or another format, which can nest values hierarchically. |
| **Schemas** | Structure and data types are fixed in advance. To store information about a new data item, the entire database must be altered, during which time the database must be taken offline. | Typically dynamic, with some enforcing data validation rules. Applications can add new fields on the fly, and unlike SQL table rows, dissimilar data can be stored together as necessary. For some databases (e.g., wide-column stores), it is somewhat more challenging to add new fields dynamically. |
| **Scaling** | Vertically, meaning a single server must be made increasingly powerful in order to deal with increased demand. It is possible to spread SQL databases over many servers, but significant additional engineering is generally required, and core relational features such as JOINs, referential integrity and transactions are typically lost. | Horizontally, meaning that to add capacity, a database administrator can simply add more commodity servers or cloud instances. The database automatically spreads data across servers as necessary. |
| **Development Model** | Mix of open-source (e.g., Postgres, MySQL) and closed source (e.g., Oracle Database) | Open-source |
| **Supports multi-record ACID transactions** | Yes | Mostly no. MongoDB 4.0 and beyond support multi-document ACID transactions. [Learn more](https://www.mongodb.com/transactions) |
| **Data Manipulation** | Specific language using Select, Insert, and Update statements, e.g. SELECT fields FROM table WHERE… | Through object-oriented APIs |
| **Consistency** | Can be configured for strong consistency | Depends on product. Some provide strong consistency (e.g., MongoDB, with tuneable consistency for reads) whereas others offer eventual consistency (e.g., Cassandra). |

## Implementing a NoSQL Database

Often, organizations will begin with a small-scale trial of a NoSQL database in their organization, which makes it possible to develop an understanding of the technology in a low-stakes way. Most NoSQL databases are also open-source, meaning that they can be downloaded, implemented and scaled at little cost. Because development cycles are faster, organizations can also innovate more quickly and deliver superior customer experience at a lower cost.

As you consider alternatives to legacy infrastructures, you may have several motivations: to scale or perform beyond the capabilities of your existing system, identify viable alternatives to expensive proprietary software, or increase the speed and agility of development. When selecting the right database for your business and application, there are five important dimensions to consider.