**DATA STORAGE**

What is Big Data Storage?

Big Data Storage is a new technology poised to revolutionize how we store data. The technology was first developed in the early 2000s when companies were faced with storing massive amounts of data that they could not keep on their servers.

The problem was that traditional storage methods couldn't handle storing all this data, so companies had to look for new ways to keep it. That's when Big Data Storage came into being. It's a way for companies to store large amounts of data without worrying about running out of space.

Big Data Storage Challenges

Big data is a hot topic in IT. Every month, more companies are adopting it to help them improve their businesses. But with any new technology comes challenges and questions, and big data is no exception.

The first challenge is how much storage you'll need for your extensive data system. If you're going to store large amounts of information about your customers and their behavior, you'll need a lot of space for that data to live.

It's not uncommon for large companies like Google or Facebook to have petabytes (1 million gigabytes) of storage explicitly dedicated to their big data needs, and that's only one company!

Another challenge with big data is how quickly it grows. Companies are constantly gathering new types of information about their customer's habits and preferences, and they're looking at ways they can use this information to improve their products or services.

 As a result, big data systems will continue growing exponentially until something stops them. It means it's essential for companies who want to use this technology effectively to plan how they'll deal with it later on down the road when it becomes too much for them alone!

**Big Data Storage Key Considerations**

Big data storage is a complicated problem. There are many things to consider when building the infrastructure for your big data project, but there are three key considerations you must consider before you move forward.

* Data velocity: Your data must be able to move quickly between processing centers and databases for it to be helpful in real-time applications.
* Scalability: The system should be able to expand as your business does and accommodate new projects as needed without disrupting existing workflows or causing any downtime.
* Cost efficiency: Because big data projects can be so expensive, choosing a system that reduces costs without sacrificing the quality of service or functionality is essential.

Finally, consider how long you want your stored data to remain accessible. If you're planning on keeping it for years (or even decades), you may need more than one storage solution.

**Key Insights for Big Data Storage**

Big data storage is a critical part of any business. The sheer volume of data being created and stored by companies is staggering and growing daily. But without a proper strategy for storing and protecting this data, your business could be vulnerable to hackers—and your bottom line could suffer.

**Here are some critical insights for big data storage:**

* Have a plan for how you'll organize your data before you start collecting it. It will ensure you can find what you need when you need it. Here are some critical insights for big data storage:
* Ensure your team understands security's essential when dealing with sensitive information. Everyone in the company needs to be trained on best practices for protecting data and preventing hacks.
* Remember backup plans! You never want to get stuck and unable to access your information because something went wrong with the server or hardware it's stored.

**Data Storage Methods**

Warehouse and cloud storage are two of the most popular options for storing big data. Warehouse storage is typically done on-site, while cloud storage involves storing your data offsite in a secure location.

### Warehouse Storage

Warehouse storage is one of the more common ways to store large amounts of data, but it has drawbacks. For example, if you need immediate access to your data and want to avoid delays or problems accessing it over the internet, there might be better options than this. Also, warehouse storage can be expensive if you're looking for long-term contracts or need extra personnel to manage your warehouse space.

### Cloud Storage

Cloud storage is an increasingly popular option since it's easier than ever to use this method, thanks to advancements in technology such as Amazon Web Services (AWS). With AWS, you can store unlimited data without worrying about how much space each file takes up on their servers. They'll automatically compress them before sending them over, so they take up less space overall!

## Data Storage Technologies

Apache Hadoop, Apache HBase, and Snowflake are three big data storage technologies often used in the data lake analytics paradigm.

### Hadoop

[Hadoop](https://www.simplilearn.com/tutorials/hadoop-tutorial/what-is-hadoop) has gained considerable attention as it is one of the most common frameworks to support big data analytics. A distributed processing framework based on open-source software, Hadoop enables large data sets to be processed across clusters of computers. Large data sets were initially intended to be processed and stored across clusters of commodity hardware.

### HBase

With HBase, you can use a [NoSQL](https://www.simplilearn.com/rise-of-nosql-and-why-it-should-matter-to-you-article" \o "NoSQL" \t "_blank) database or complement Hadoop with a column-oriented store. This database is designed to efficiently manage large tables with billions of rows and millions of columns. The performance can be tuned by adjusting memory usage, the number of servers, block size, and other settings.

### Snowflake

Snowflake for Data Lake Analytics is an enterprise-grade cloud platform for advanced analytics applications built on top of Apache Hadoop. It offers real-time access to historical and streaming data from any source and format at any scale without requiring changes to existing applications or workflows. It also enables users to quickly scale up their processing power as needed without having to worry about infrastructure management tasks such as provisioning and

**DATA QUALITY**

**What’s the Definition of Data Quality?**

In simple terms, data quality tells us how reliable a particular set of data is and whether or not it will be good enough for a user to employ in decision-making. This quality is often measured by degrees.

**But What Is Data Quality, in Practical Terms?**

Data quality measures the condition of data, relying on factors such as how useful it is to the specific purpose, completeness, accuracy, timeliness (e.g., is it up to date?), consistency, validity, and uniqueness.

Data quality analysts are responsible for conducting data quality assessments, which involve assessing and interpreting every quality data metric. Then, the analyst creates an aggregate score reflecting the data’s overall quality and gives the organization a percentage rating that shows how accurate the data is.

To put the definition in more direct terms, data quality indicates how good the data is and how useful it is for the task at hand. But the term also refers to planning, implementing, and controlling the activities that apply the needed quality management practices and techniques required to ensure the data is actionable and valuable to the data consumers.

Now, let us look at data quality dimensions after you better understand what is data quality.

**Data Quality Dimensions**

There are six primary, or core, dimensions to data quality. These are the metrics analysts use to determine the data’s viability and its usefulness to the people who need it.

### Accuracy

The data must conform to actual, real-world scenarios and reflect real-world objects and events. Analysts should use verifiable sources to confirm the measure of accuracy, determined by how close the values jibe with the verified correct information sources.

### Completeness

Completeness measures the data's ability to deliver all the mandatory values that are available successfully.

### Consistency

Data consistency describes the data’s uniformity as it moves across applications and networks and when it comes from multiple sources. Consistency also means that the same datasets stored in different locations should be the same and not conflict. Note that consistent data can still be wrong.

### Timeliness

Timely data is information that is readily available whenever it’s needed. This dimension also covers keeping the data current; data should undergo real-time updates to ensure that it is always available and accessible.

### Uniqueness

Uniqueness means that no duplications or redundant information are overlapping across all the datasets. No record in the dataset exists multiple times. Analysts use data cleansing and deduplication to help address a low uniqueness score.

### Validity

Data must be collected according to the organization’s defined business rules and parameters. The information should also conform to the correct, accepted formats, and all dataset values should fall within the proper range.

**How Do You Improve Data Quality?**

People looking for ideas on how to improve data quality turn to[data quality management](https://www.simplilearn.com/what-is-data-quality-management-article) for answers. Data quality management aims to leverage a balanced set of solutions to prevent future data quality issues and clean (and ideally eventually remove) data that fails to meet data quality KPIs (Key Performance Indicators). These actions help businesses meet their current and future objectives.

There is more to data quality than just[data cleaning](https://www.simplilearn.com/data-cleaning-why-and-how-to-get-started-article). With that in mind, here are the eight mandatory disciplines used to prevent data quality problems and improve data quality by cleansing the information of all bad data:

### Data Governance

[Data governance](https://www.simplilearn.com/what-is-data-governance-article) spells out the data policies and standards that determine the required data quality KPIs and which data elements should be focused on. These standards also include what business rules must be followed to ensure data quality.

### Data Profiling

[Data profiling](https://www.simplilearn.com/data-profiling-tools-and-best-practices-article) is a methodology employed to understand all data assets that are part of data quality management. Data profiling is crucial because many of the assets in question have been populated by many different people over the years, adhering to different standards.

### Data Matching

Data matching technology is based on match codes used to determine if two or more bits of data describe the same real-world thing. For instance, say there’s a man named Michael Jones. A customer dataset may have separate entries for Mike Jones, Mickey Jones, Jonesy, Big Mike Jones, and Michael Jones, but they’re all describing one individual.

### Data Quality Reporting

Information gathered from data profiling, and data matching can be used to measure data quality KPIs. Reporting also involves operating a quality issue log, which documents known data issues and any follow-up data cleansing and prevention efforts.

### Master Data Management (MDM)

Master Data Management frameworks are great resources for preventing data quality issues. MDM frameworks deal with product master data, location master data, and party master data.

### Customer Data Integration (CDI)

CDI involves compiling customer master data gathered via CRM applications, self-service registration sites. This information must be compiled into one source of truth.

### Product Information Management (PIM)

Manufacturers and sellers of goods need to align their data quality KPIs with each other so that when customers order a product, it will be the same item at all stages of the supply chain. Thus, much of PIM involves creating a standardized way to receive and present product data.

### Digital Asset Management (DAM)

Digital assets cover items like videos, text documents, images, and similar files, used alongside product data. This discipline involves ensuring that all tags are relevant and the quality of the digital assets.

**Data Quality Best Practices**

Data analysts who strive to improve data quality need to follow best practices to meet their objectives. Here are ten critical[best practices](https://www.simplilearn.com/data-profiling-tools-and-best-practices-article) to follow:

* Make sure that top-level management is involved. Data analysts can resolve many data quality issues through cross-departmental participation.
* Include data quality activity management as part of your data governance framework. The framework sets data policies and data standards, the required roles and offers a business glossary.
* Each data quality issue raised must begin with a root cause analysis. If you don’t address the root cause of a data issue, the problem will inevitably appear again. Don’t just address the symptoms of the disease; you need to cure the disease itself.
* Maintain a data quality issue log. Each issue needs an entry, complete with information regarding the assigned data owner, the involved data steward, the issue’s impact, the final resolution, and the timing of any necessary proceedings.
* Fill data owner and data steward roles from your company’s business side and fill data custodian roles from either business or IT whenever possible and makes the most sense.
* Use examples of data quality disasters to raise awareness about the importance of data quality. However, while anecdotes are great for illustrative purposes, you should rely on fact-based impact and risk analysis to justify your solutions and their required funding.
* Your organization’s business glossary must serve as the foundation for metadata management.
* Avoid typing in data where possible. Instead, explore cost-effective solutions for data onboarding that employ third-party data sources that provide publicly available data. This data includes items such as names, locations in general, company addresses and IDs, and in some cases, individual people. When dealing with product data, use second-party data from trading partners whenever you can.
* When resolving data issues, make every effort to implement relevant processes and technology that stops the problems from arising as close as possible to the data onboarding point instead of depending on downstream data cleansing.
* Establish data quality KPIs that work in tandem with the general KPIs for business performance. Data quality KPIs, sometimes called Data Quality Indicators (DQIs), can often be associated with data quality dimensions like uniqueness, completeness, and consistency.

# What is data operations? How can it benefit your organization?

Data operations enables your organization to maximize the business value of your data and its underlying infrastructure. It’s the overall approach to designing, building, moving and using your data, both on-premises and in the cloud. It’s the key to digital transformation initiatives such as cloud migration, DevOps, open-source database adoption and data governance.

Does it feel as though your database technology gets in the way of using your data to solve business problems? Maybe you’re aware of how much data your organization is accumulating, but you’re having trouble getting good value from your data assets and infrastructure**.**

Then it’s time to start thinking about data operations**.**

## What kind of business problems can data operations solve?

The more data you accumulate — and the more you rely on it — the more problems arise.

* **Poorly-planned cloud migrations** – Have you moved workloads to the cloud in a hurry, without the chance to analyze and monitor them first? When performance problems arose, how did you know where to start looking? Maybe the cloud environment introduced them, maybe not. Data operations ensures that the root cause is not so easily misdiagnosed or hidden.
* **Increased agility of database changes** – Organizations are trying to apply DevOps practices like continuous integration and continuous deployment to react more quickly to changes in the business. But DBAs tend to not like quick changes to data structures as it can create risks to data, so bottlenecks crop up when IT slows the pace of change in their effort to mitigate risk.
* **The balance between “always on” and costs** – You can have it all — [high availability](https://blog.quest.com/high-availability-architecture-considerations-and-techniques-to-achieve-five-9s/) of mission-critical applications, cloud environment, Oracle databases and offsite [data replication](https://www.quest.com/community/blogs/b/database-management/posts/data-replication-what-is-it-and-what-are-the-advantages-of-using-it) — but it costs money. Data operations helps you strike a balance.
* **Skills gaps** – New databases and infrastructure options tend to de-centralize your IT landscape, yet for support and troubleshooting, everybody always comes back to IT. Can your team keep up with the constant innovation? [Research from ESG](https://www.quest.com/docs/empowering-everyone-to-bring-the-right-data-to-every-decision-white-paper-148808.pdf) shows that 36 percent of organizations have glaring gaps in cloud architecture/planning, 34 percent in IT architecture/planning and 32 percent in IT orchestration and automation.
* **Disruption of the data pipeline** – When your business depends on data and analytics, all eyes are on the data pipeline. The responsibility for keeping data flowing places new urgency on internal systems and data ingestion points.
* **Self-service data consumption** – If you’re trying to empower line-of-business (LOB) users, too much data can be as big a problem as too little data. Data operations addresses the problem of locating the right data from so many sources and understanding how to connect, access and interpret it.
* **Reactive mindset** – Being constantly reactive means that database or infrastructure performance problems can overtake you and affect user experience in business-critical applications. You’re better off anticipating those problems than reacting to them.
* **Transformation of operations teams** –Autonomous databases, artificial intelligence (AI) and machine learning (ML) are changing the traditional role of DBAs, who are looking for new ways to add value to the business. Data operations enables DBAs to embrace change and grow from being experts in the database to being experts in the data.

## Data operations is different from DataOps

DataOps provides greater collaboration and delivery of data and insights at real-time speeds to decision makers or decision-making applications. The [essence of DataOps is the automation of processes](https://thenewstack.io/overcoming-the-obstacles-to-data-democratization-and-dataops/), similar to those used in DevOps, that help [democratize data](https://blog.quest.com/why-data-democratization-why-now-what-does-it-look-like/). DataOps doesn’t refer to supporting infrastructure.

Data operations, on the other hand, takes a broader view. It includes the data and the data pipeline: the hybrid infrastructure where data resides and the operational needs of data availability, integrity and performance. Both the data and the pipeline have business value and the goal of data operations is to maximize that value. Within the pipeline is the infrastructure that needs testing, performance monitoring, cost analysis, tuning, securing and so forth.

## Four basic steps for implementation

### 1. Assemble the data.

To put the right data in front of decision makers, IT first designs the data in a way that maintains the organization’s data management framework. So, the first goal of data operations is to capture business requirements that can be translated into an accurate, usable, logical data model.

As obvious as that may sound, consider the downside of not capturing those requirements. If you skip the logical modeling and dive into creating physical data structures and relationships, it’s far more work to incorporate business rules later. Get it right the first time by creating a full, logical model because the friction you’ll face is not worth the little bit of time you’ll have saved.

[Data modeling tools](https://www.erwin.com/products/erwin-data-modeler/) introduce two elements at this point: rigor and documentation. Rigor is important because there is more than one way to model data for say, a customer entity. But in your organization, there is only one right way to model it, and rigor ensures that you find that right way and build the model accordingly. As in constructing a building, form (the physical design) follows function (the logical data model).

Documentation is important because of the need to support discovery and documentation of data from anywhere. The goals are consistency, clarity and artifact reuse across initiatives like data integration, master data management, metadata management, big data, business intelligence and analytics. Once you’ve documented the elements of the data model and its physical translation, you have a way to find, visualize, define, deploy and standardize enterprise data assets. Whoever (usually DBAs) creates the data structures to match the logical model’s requirements needs a guide to make that translation. Documentation helps them generate the data definition language (DDL) scripts for those structures accurately, adjust if needed and assemble the data engine.

Next, come the [database tools](https://www.quest.com/products/toad-for-oracle/) that drive productivity, improve manageability and increase quality in a heterogeneous database environment. The more efficiently developers can create high-quality code that’s free from defects and DBAs can handle or automate routine tasks, the more they can focus on innovation.

### 2. Move the data to where it needs to be.

There’s a simple truth that drives much of cloud spending: The greater the resource load on the platform, the higher the cost once it’s in the cloud. In this step, data operations includes [performance monitoring](https://www.quest.com/solutions/database-performance-monitoring/) to ensure there is no problem when a workload goes to the cloud. That means right-sizing your cloud computing resources — CPU, storage, memory, network — to handle the load without breaking your budget.

Here is a simple breakdown of right-sizing tasks:

1. Estimate on-premises cost of the entire workload.
2. Before migrating, perform a cost study of what looks like the best-fit cloud service tier.
3. Optimize the workloads.
4. Do a pre-migration load test on the database.
5. Document everything.

Your goal is to study workloads for expected cloud resource usage and costs, based on performance on-premises. If you shrink your virtual machines and trim database resource requirements early, you’ll have slimmed-down baselines on hand to compare to workload performance in the cloud after migration.

### 3. Manage the data across the enterprise.

In a completely on-premises environment, you [monitor and optimize databases](https://www.quest.com/solutions/database-performance-monitoring/) to maintain performance levels; in the cloud, you monitor and optimize to control cost, also.

Right after cloud migration, the most pressing question centers on performance: How does application performance in the cloud compare to what it was on-premises? According to the [2020 PASS Database Management Survey](https://www.quest.com/whitepaper/database-professionals-look-to-the-future-2020-trends-in-database-admi8141411/) report, in the cloud, 24 percent of DBAs worry about performance, compared to 41 percent for on-premises databases. The trend points to less time spent on performance management, but the need to monitor performance still grows as more databases go to the cloud.

Within a month or so, concerns switch to cost:

* Are cloud costs acting as expected?
* If costs are higher than estimated, does anyone in the organization understand why?
* How do you know what to adjust?
* Have you selected the right service tier for your workloads?

The goal of data operations in this step is to provide meaningful information to management and the CIO for tuning the current cloud migration strategy. The most useful tools are the ones that show where the organization is overpaying for and under-benefitting from cloud computing. Typical culprits include migrated virtual machines whose costs are increasing because their workloads don’t fit the service tier.

### 4. Make the data useful.

We’ve seen how the first three steps help in the democratization of data.

To get their hands on the right data, decision makers need applications. Customer-dependent applications are changing all the time because business needs call for it. The goal of data operations is to keep pace with the changes and keep the right data flowing through the organization.

[Database DevOps](https://www.quest.com/products/toad-devops-toolkit/) has evolved as a way of keeping quick-turnaround builds reliably safe for production. Like application developers, database developers unit test database code to reduce defects, perform code reviews to reduce technical debt and automatically create change scripts to accurately deploy their changes to the production system.

Finally, companies can fulfill the primary goal of making their data work for them by extracting the important insights necessary to help inform the decisions that drive strategy and enable the business to grow.

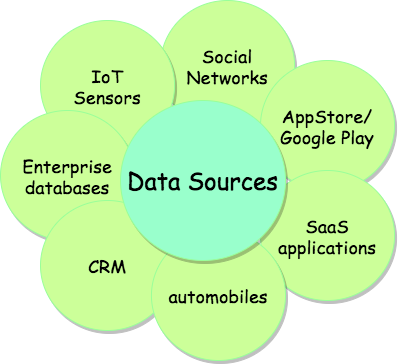
## Conclusion

Data operations seeks to streamline the delivery of data through increased data democratization into downstream processes like analytics and AI/ML. At the business level, that means extracting insights at scale and maximizing the full value of both data and infrastructure.

The drivers for change that are implemented and managed by IT need to come from the business so that they align with business strategy and drive growth. Companies can attain those goals through [data empowerment](https://www.quest.com/solutions/data-empowerment/): a strategy that combines the three pillars of data operations, [data protection](https://www.quest.com/community/blogs/b/database-management/posts/data-protection-and-data-privacy-how-do-you-really-protect-your-users-information) and [data governance](https://blog.erwin.com/blog/the-what-why-of-data-governance/).

**DATA INGESTION**

Data Ingestion is the first layer in the **Big Data Architecture —**this is the layer that is responsible for collecting data from various data sources—IoT devices, data lakes, databases, and SaaS applications—into a target data warehouse. This is a critical point in the process — because at this stage the size and complexity of the data can be understood, which will affect the architecture or every decision we make down the road.



**Why Do we need Data ingestion layer?**

1. Availability — The data is available to all users: BI analysts, developers, sales and anyone else in the company can access the data.
2. Uniformity — a quality data ingestion process can turn different types of data into a unified data that is easy to read and perform statistics and manipulations on.
3. Save money and time — a data ingestion process saves engineers time in trying to collect the data they need and develop efficiently instead.

**What are the challenges with data ingestion?**

1. Complexity — Writing data ingestion processes can be complex due to data velocity and variety, and some times so development times can be costly in time and resources.
2. Data security — When transferring data from one place to another there is a risk of security to sensitive data.
3. Unreliability — During the process, the reliability of the data may be compromised and thus cause the data to be of no value or in the worst case to make incorrect decisions based on untrue data.

**Data ingestion types**

There are three common ways to do data ingestion and the use will be according to the product needs — is it important to collect data in real time or can it be done once in a while in a timed manner?

**Real Time data ingestion**

This is the process to collect and process data from the various data sources in real time — also known as streaming, we will use this approach when the data collected is time-sensitive.

For example — the data coming from oil tanks’ sensors will be critical in cases of leakage.

In real time cases, the velocity of the data will be high, so the solution will contain a queue to avoid losing events. The data will be extracted, processed and saved at fast as we can.

**Batching data ingestion**

Batching data ingestion meaning that data moved from the data sources to the data target on scheduled intervals. Batch ingestion is useful when companies need to collect data on a daily basis.

**SCALABILITY IN BIG DATA**

Ability of a process, network, software or organization to grow and manage increased demands in a capable manner is Scalability. Scalable infrastructure is one that has the capability to add more computers or servers in a network in order to handle the increased workload. Scaling can be difficult, but absolutely necessary in the growth of a successful data-driven company. There are a few signs that it’s time to implement a scaling platform. When users begin complaining about slow performance, or service outages, it’s time to scale. Don’t wait for the problem to turn into major source of contention in the [minds of your customers](https://ngdata.com/resources/experience-your-customers-your-brand/). This can have a massively negative impact on retaining those customers. If possible, try to anticipate the problem before it becomes severe. In addition to this, increased application latency, slow read queries rises and database writes are also important indicators that a scale is needed.

**SECURITY IN BIG DATA**

Ability of a process, network, software or organization to grow and manage increased demands in a capable manner is Scalability. Scalable infrastructure is one that has the capability to add more computers or servers in a network in order to handle the increased workload.

These threats include the theft of information stored online, ransomware, or DDoS attacks that could crash a server. The issue can be even worse when companies store information that is sensitive or confidential, such as customer information, credit card numbers, or even simply contact details. Additionally, attacks on an organization’s big data storage could cause serious financial repercussions such as losses, litigation costs, and fines or sanctions.

**How Can You Implement Big Data Security?**

There are several ways organizations can implement security measures to protect their [big data analytics tools](https://www.sisense.com/product/big-data-analytics/). One of the most common security tools is encryption, a relatively simple tool that can go a long way. Encrypted data is useless to external actors such as hackers if they don’t have the key to unlock it. Moreover, encrypting data means that both at input and output, information is completely protected.

Building a strong firewall is another useful big data security tool. Firewalls are effective at filtering traffic that both enters and leaves servers. Organizations can prevent attacks before they happen by creating strong filters that avoid any third parties or unknown data sources.

Finally, controlling who has root access to [BI tools](https://www.sisense.com/bi-tools/) and analytics platforms is another key to protecting your data. By developing a tiered access system, you can reduce the opportunities for an attack.

[**Traditional data**](https://www.geeksforgeeks.org/traditional-data-mining-life-cycle-crisp-methodology/)**:** Traditional data is the structured data that is being majorly maintained by all types of businesses starting from very small to big organizations. In a traditional database system, a centralized database architecture used to store and maintain the data in a fixed format or fields in a file. For managing and accessing the data Structured Query Language (SQL) is used.

**2. Big data:** We can consider big data an upper version of traditional data. Big data deal with too large or complex data sets which is difficult to manage in traditional data-processing application software. It deals with large volume of both structured, semi structured and unstructured data. Volume, Velocity and Variety, Veracity and Value refer to the 5’V characteristics of big data. Big data not only refers to large amount of data it refers to extracting meaningful data by analyzing the huge amount of complex data sets. semi-structured

The difference between Traditional data and Big data are as follows:

| **Traditional Data** | **Big Data** |
| --- | --- |
| Traditional data is generated in enterprise level. | Big data is generated outside the enterprise level. |
| Its volume ranges from Gigabytes to Terabytes. | Its volume ranges from Petabytes to Zettabytes or Exabytes. |
| Traditional database system deals with structured data. | Big data system deals with structured, semi-structured,database, and unstructured data. |
| Traditional data is generated per hour or per day or more. | But big data is generated more frequently mainly per seconds. |
| Traditional data source is centralized and it is managed in centralized form. | Big data source is distributed and it is managed in distributed form. |
| Data integration is very easy. | Data integration is very difficult. |
| Normal system configuration is capable to process traditional data. | High system configuration is required to process big data. |
| The size of the data is very small. | The size is more than the traditional data size. |
| Traditional data base tools are required to perform any data base operation. | Special kind of data base tools are required to perform any databaseschema-based operation. |
| Normal functions can manipulate data. | Special kind of functions can manipulate data. |
| Its data model is strict schema based and it is static. | Its data model is a flat schema based and it is dynamic. |
| Traditional data is stable and inter relationship. | Big data is not stable and unknown relationship. |
| Traditional data is in manageable volume. | Big data is in huge volume which becomes unmanageable. |
| It is easy to manage and manipulate the data. | It is difficult to manage and manipulate the data. |
| Its data sources includes ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data etc. | Its data sources includes social media, device data, sensor data, video, images, audio etc. |

**REAL LIFE APPLICATIONS**

1. Transportation

Big Data powers the GPS smartphone applications most of us depend on to get from place to place in the least amount of time. GPS data sources include satellite images and government agencies.

Airplanes generate enormous volumes of data, on the order of 1,000 gigabytes for transatlantic flights. Aviation analytics systems ingest all of this to analyze fuel efficiency, passenger and cargo weights, and weather conditions, with a view toward optimizing safety and energy consumption.

Big Data simplifies and streamlines transportation through:

* Congestion management and traffic control  
  Thanks to Big Data analytics, Google Maps can now tell you the least traffic-prone route to any destination.
* Route planning  
  Different itineraries can be compared in terms of user needs, fuel consumption, and other factors to plan for maximize efficiency.
* Traffic safety  
  Real-time processing and predictive analytics are used to pinpoint accident-prone areas.

* + Advertising and Marketing

Ads have always been targeted towards specific consumer segments. In the past, marketers have employed TV and radio preferences, survey responses, and focus groups to try to ascertain people’s likely responses to campaigns. At best, these methods amounted to educated guesswork.

Today, advertisers buy or gather huge quantities of data to identify what consumers actually click on, search for, and “like.” Marketing campaigns are also monitored for effectiveness using click-through rates, views, and other precise metrics.

For example, Amazon accumulates massive data stories on the purchases, delivery methods, and payment preferences of its millions of customers. The company then sells ad placements that can be highly targeted to very specific segments and subgroups.

1. Banking and Financial Services

The financial industry puts Big Data and analytics to highly productive use, for:

* Fraud detection  
  Banks monitor credit cardholders’ purchasing patterns and other activity to flag atypical movements and anomalies that may signal fraudulent transactions.
* Risk management  
  Big Data analytics enable banks to monitor and report on operational processes, KPIs, and employee activities.
* Customer relationship optimization  
  Financial institutions analyze data from website usage and transactions to better understand how to convert prospects to customers and incentivize greater use of various financial products.
* Personalized marketing  
  Banks use Big Data to construct rich profiles of individual customer lifestyles, preferences, and goals, which are then utilized for micro-targeted marketing initiatives.

1. Government

Government agencies collect voluminous quantities of data, but many, especially at the local level, don’t employ modern data mining and analytics techniques to extract real value from it.

Examples of agencies that do include the IRS and the Social Security Administration, which use data analysis to identify tax fraud and fraudulent disability claims. The FBI and SEC apply Big Data strategies to monitor markets in their quest to detect criminal business activities. For years now, the Federal Housing Authority has been using Big Data analytics to forecast mortgage default and repayment rates.

The Centers for Disease Control tracks the spread of infectious illnesses using data from social media, and the FDA deploys Big Data techniques across testing labs to investigate patterns of foodborne illness. The U.S. Department of Agriculture supports agribusiness and ranching by developing Big Data-driven technologies.

Military agencies, with expert assistance from a sizable ecosystem of defense contractors, make sophisticated and extensive use of data-driven insights for domestic intelligence, foreign surveillance, and cybersecurity.

1. Media and Entertainment

The entertainment industry harnesses Big Data to glean insights from customer reviews, predict audience interests and preferences, optimize programming schedules, and target marketing campaigns.

Two conspicuous examples are Amazon Prime, which uses Big Data analytics to recommend programming for individual users, and Spotify, which does the same to offer personalized music suggestions.

1. Meteorology

Weather satellites and sensors all over the world collect large amounts of data for tracking environmental conditions. Meteorologists use Big Data to:

* Study natural disaster patterns
* Prepare weather forecasts
* Understand the impact of global warming
* Predict the availability of drinking water in various world regions
* Provide early warning of impending crises such as hurricanes and tsunamis

1. Healthcare

Big Data is slowly but surely making a major impact on the huge healthcare industry. Wearable devices and sensors collect patient data which is then fed in real-time to individuals’ electronic health records. Providers and practice organizations are now using Big Data for a number of purposes, including these:

* Prediction of epidemic outbreaks
* Early symptom detection to avoid preventable diseases
* Electronic health records
* Real-time alerting
* Enhancing patient engagement
* Prediction and prevention of serious medical conditions
* Strategic planning
* Research acceleration
* Telemedicine
* Enhanced analysis of medical images

1. Cybersecurity

While Big Data can expose businesses to a greater risk of cyberattacks, the same datastores can be used to prevent and counteract online crime through the power of machine learning and analytics. Historical data analysis can yield intelligence to create more effective threat controls. And machine learning can warn businesses when deviations from normal patterns and sequences occur, so that effective countermeasures can be taken against threats such as ransomware attacks, malicious insider programs, and attempts at unauthorized access.

After a company has suffered an intrusion or data theft, post-attack analysis can uncover the methods used, and machine learning can then be deployed to devise safeguards that will foil similar attempts in the future.

1. Education

Administrators, faculty, and stakeholders are embracing Big Data to help improve their curricula, attract the best talent, and optimize the student experience. Examples include:

* Customizing curricula  
  Big Data enables academic programs to be tailored to the needs of individual students, often drawing on a combination of online learning, traditional on-site classes, and independent study.
* Reducing dropout rates  
  Predictive analytics give educational institutions insights on student results, responses to proposed programs of study, and input on how students fare in the job market after graduation.
* Improving student outcomes  
  Analyzing students’ personal “data trails” can provide a better understanding of their learning styles and behaviors, and be used to create an optimal learning environment.
* Targeted international recruiting  
  Big Data analysis helps institutions more accurately predict applicants’ likely success. Conversely, it aids international students in pinpointing the schools best matched to their academic goals and most likely to admit them.

**DATA MODEL CONSTRAINTS**

**Integrity Constraints**

The basic structure consists of:

* Attribute
* Value
* Relation

**Attributes** can be either single-valued or multi-valued, while **values**correspond to specific objects like strings or integers. **Relationships**between different entities can be represented as links where each link has its own unique identifier number called a primary key which enables retrieval from other

One example from this relational model would be an email address table where each row contains One example from this relational model would be an email address table where each row contains contact information such as name, company address etc., just like you might find on your computer desktop’s contacts list applet (the ones usually found at top right). You can also think about any spreadsheet document which has multiple cells separated by column headers: there are many ways these could represent different types of individuals who

The relational model, or RM for short, represents the database as a collection of related tables. The columns in these tables denote entities and their relationships to one another. For example: House has-a Street will be represented by two rows – one representing the “House” entity with three values (name=house), streetName=”Street”, address=”123 Main St.”) and other row defining an instance of its relationship to Street(“has-a”)with value (“street”).

**Operations performed in Relational Model**

In the relational model, four types of operations are performed.

1. **Insert Operation:** Insert Operation is used to save the values in a new tuple.
2. **Modify/Update Operation:** Modify operation updates (changes) the existing values in a tuple.
3. **Search Operation:** We retrieve information.
4. **Delete Operation:** Used to delete the tuples.

**Why is it called a relational database?**

The data is stored in a structured format i.e. in columns and rows. In this type of storage, we can easily access the values. It is called a relational database because the values within the same table are closely related to each other. Tables within the database can also be related to each other. We can run different queries to fetch data or some specific value from multiple tables.

**TYPES OF Big Data Model**

**Structured Data**

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

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

It’s all your quantitative data:

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

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

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

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

**Working With It**

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

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

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

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

**Unstructured Data**

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

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

Unstructured data is all your unorganized data:

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

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

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

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

**Working With It**

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

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

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

This brings us to an important point: context is almost, if not as, important as the information wrung out of the data. Alissa Lorentz, then the vice president of creative, marketing and design at Augify, explained in a [guest article for Wired](https://www.wired.com/insights/2013/04/with-big-data-context-is-a-big-issue/): a query on an unstructured data set might yield the number 31, but without context it’s meaningless. It could be “the number of days in a month, the amount of dollars a stock increased…, or the number of items sold today.”

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

This context can be created from unstructured datasets, like NoSQL databases, or human dictation. The world of machine learning, or AI teaching itself how to improve and discover patterns, is becoming instrumental in the world of big data because of its ability to autonomously improve on models.

**Semi-Structured Data**

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

Let’s say you take a picture of your cat from your phone. It automatically logs the time the picture was taken, the GPS data at the time of the capture and your device ID. If you’re using any kind of web service for storage, like iCloud, your account info becomes attached to the file.

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

**Big data processing tools**

**Hadoop**

The [Apache Hadoop](https://www.guru99.com/bigdata-tutorials.html) software library is a big data framework. It allows distributed processing of large data sets across clusters of computers. It is one of the best big data tools 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 has big data technologies and tools that 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

### [Zoho Analytics](https://guru99.link/recommends-big-data-analytics-tool" \t "_blank)

[Zoho Analytics](https://guru99.link/recommends-big-data-analytics-tool) is a self-service business intelligence and analytics platform. It allows users to create insightful dashboards and visually analyze any data in minutes. It features an AI powered assistant that enables users to ask questions and get intelligent answers in the form of meaningful reports.

**Features:**

* 100+ readymade connectors for popular business apps, cloud drives and databases.
* Wide variety of visualization options–charts, pivot tables, summary views, KPI widgets and custom themed dashboards.
* Unified business analytics for analyzing data from across business apps.
* Augmented analytics using AI, ML and NLP.
* White label BI portals and embedded analytics solutions.

### [Atlas.ti](https://prf.hn/l/6lo4Nnx)

[Atlas.ti](https://prf.hn/l/6lo4Nnx) is all-in-one research software. This big data analytic tool gives you all-in-one access to the entire range of platforms. You can use it for qualitative data analysis and mixed methods research in academic, market, and user experience research.

**Features:**

* You can export information on each source of data.
* It offers an integrated way of working with your data.
* Allows you to rename a Code in the Margin Area
* Helps you to handle projects that contain thousands of documents and coded data segments.

### HPCC

[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:**

* It is one of the Highly efficient big data tools that accomplish big data tasks with far less code.
* It is one of the big data processing tools which 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

### Storm

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

**Features:**

* It is one of the best tool from big data tools list which is benchmarked as processing one million 100 byte messages per second per node
* It has big data technologies and tools that 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 Bigdata analysis

### Cassandra

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 centers by providing lower latency for users
* Data is automatically replicated to multiple nodes for fault-tolerance
* It one of the best big data tools which is most suitable for applications that can’t afford to lose data, even when an entire data center is down
* Cassandra offers support contracts and services are available from third part

### Stats iQ

[Stats iQ](https://www.qualtrics.com/au/iq/stats-iq/) by Qualtrics is an easy-to-use statistical tool. It was built by and for big data analysts. Its modern interface chooses statistical tests automatically.

**Features:**

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

**CouchDB**

[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 is one of the big data processing tools that 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