algoscale

Building a Scalable Data Warehouse

10 Essential Factors to Consider

algoscale

Abstract

The total amount of data created and consumed worldwide is predicted to grow and reach 180 zettabytes by 2025. With this exponential growth in the amount of available data, every organization is braved by two significant challenges: where to store all this data and how to make sense of it. And the best solution to both these challenges is using data warehouses.

This whitepaper explores the critical factors that must be considered while building a robust and scalable data warehouse. It aims at helping decision-makers overcome the most common challenges and adopt the right strategy to ensure seamless data warehouse implementation.

algoscale

Table of Contents

Problem

According to TrustRadius, in the past decade, the total amount of data created in the world has grown by 5,000%.

Today, almost every organization is struggling to manage enormous volumes of data that are being generated from multiple sources. This produced data coming from varied business verticals is exerting tremendous pressure on existing organizational resources.

Most large-scale organizations deploy solutions such as Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) that generate a vast amount of unstructured data. This unstructured data is not pre-organized in any specific format and typically contains text-rich information such as names and addresses. Statistics reveal that unstructured data is anticipated to grow at over 10% CAGR from 2019 to 2025. And as per Forbes, 95% of organizations cite that management of unstructured data is the biggest challenge for their business.

Let's take a scenario:

Say, your business enterprise needs to combine data from varied internal tools in order to facilitate more informed business decisions.

For instance, you may want to track the weekly activities of your most valuable customers. And this requires you to assimilate information from varied sources, i.e. payment data from your credit card processor, fiscal data from your accounting software, and activity data that your customers generate within your system.

Now, combining all this data isn't easy, especially if it is not located in one central location. Moreover, you need to separate the analytical data from the transactional data so your analysts can work on the former without any disruption.

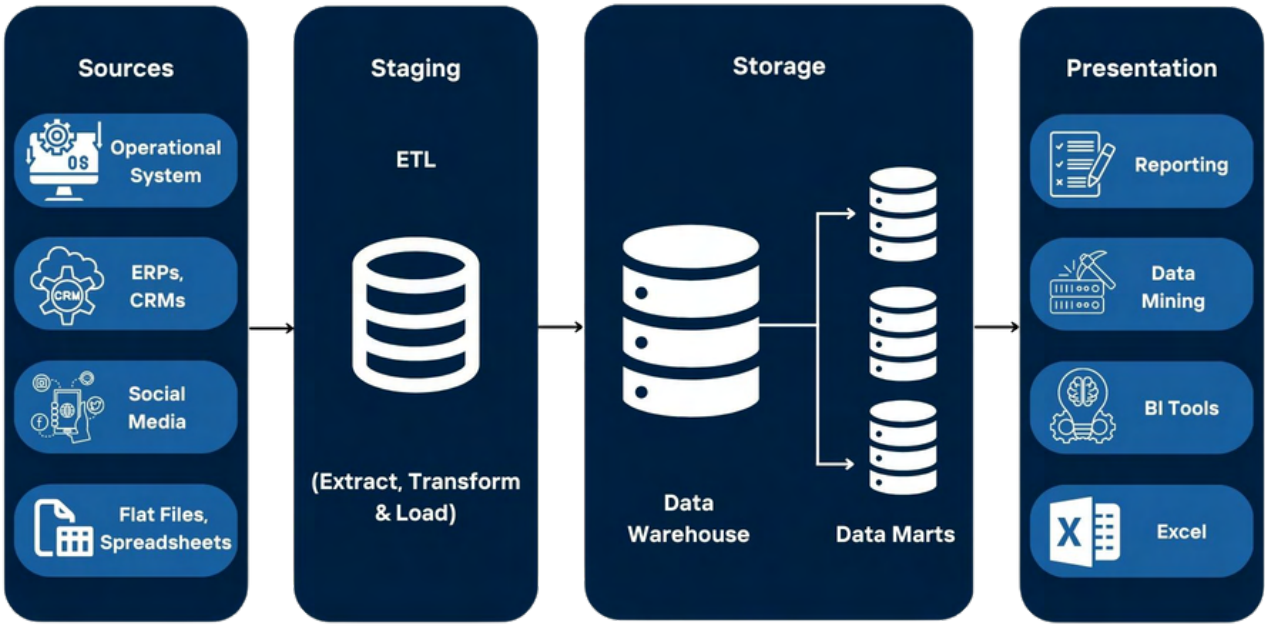
But how can you do that?

algoscale

Solution

The short answer? By implementing a data warehouse.

A data warehouse is a system that collects and consolidates enterprise information from multiple sources and converts it into a format that is suitable for reporting and analytical querying to support informed business decisions.



Data Lifecycle

In other words, a data warehouse serves as a single source of truth. And because the data is aggregated from multiple systems into one location, the data is most precise, offering decision-makers the best information they need to make the right business choices. In practice, this might mean gathering data from your ERP, customer portals, accounting system, and marketing database into one single repository.

Here's how Algoscale helped a marketing agency build a scalable data warehouse to accumulate data from multiple sources and generate weekly assessment reports. Check our Case Study here.

Cloud vs. On-Premises Data Warehouse

Now that you have understood what a data warehouse is, it's time to make an important choice: whether to opt for an on-premises data warehouse or a cloud data warehouse. Both approaches have their pros and cons and are made to suit different requirements. Let's look at each of them in detail to gain a better understanding.

On-Premise Data Warehouse

An on-premise data warehouse means that an organization deploys either an open-source or a paid data warehouse system on its own infrastructure.
Hosting a data warehouse on your organization's own infrastructure has certain pros and cons

On-Premise Data Warehouse	
Pros	**Cons**
The organization exercises complete control over its data. This solution is ideal for enterprises that have stringent security policies in place.	Building and implementing an on-premise system requires a lot of effort on the development front.
The data is always located close to the source so it eases concerns over network latency and makes data governance much easier.	Scaling an on-premise setup is not easy as it involves a hefty cost of getting new hardware.
Most on-premise data warehouse systems provide simpler interfaces to data sources as they use very little third-party cloud data.	Scaling down at zero is not an option available with an on-premise setup.

Cloud Data Warehouse

As the name suggests, a cloud-based data warehouse is located on the cloud. In this, the organization doesn't have to deploy or maintain a data warehouse at all. The warehouse is built and maintained by the cloud provider and all the features and functionalities needed to operate the data are offered as web APIs.

Some of the most common examples of cloud data warehouse systems are AWS Redshift, Google BigQuery, Microsoft Azure SQL Data warehouse, etc.

Using a cloud-based data warehouse has the following pros and cons.

Cloud Data Warehouse	
Pros	**Cons**
Scaling operations with a cloud data warehouse is very convenient. The organization only has to pay for the storage and processing capacity it uses.	The organization's data is not located in the internal system which may raise significant security concerns for some industries.
With cloud data warehouses, scaling down is also easy, thereby offering great flexibility to organizations with budget restrictions.	There might be some network latency issues since no data is present in the internal network of the organization.
The organization does not require any special teams for building, deploying, and maintaining a reliable data warehouse.	

Finally, the decision between on-premise and cloud data warehouses depends on your precise business needs. For instance, for organizations looking for high-processing volumes throughout the day, investing in an on-premise set makes more sense.

algoscale

10 Essential Factors to Consider While Choosing a Data Warehouse

A key aspect of using data warehouses is the ability to integrate them with your organization's existing systems and implement efficient data visualization. An excellent use case for data warehousing is to incorporate it with tools such as Power BI, enabling business owners to develop custom reports and gather insights more quickly.

However, to ensure a successful implementation of a data warehouse, there are certain factors that you need to keep in mind.

1. Your precise business requirements

Most organizations fail to implement a data warehouse because they do not have a defined business use case for it. Conversely, organizations that kick-start the process by recognizing a business problem for their data can remain focused on finding a solution.

Here are some of the top reasons why your business might require a data warehouse:

- **Enhancing decision-making:** Typically, organizations make decisions without evaluating the gathered data as opposed to flourishing businesses that devise smart data-driven strategies. **In fact, as per a survey by Harvard Business Review, 53% of organizations report not using data as a critical asset.** Data warehousing boosts the speed and efficiency of data access, enabling business leaders to create a data-driven organization and have a clear edge over competitors.

- **Standardizing your data:** Data warehouses gather data from multiple sources and store it in a unified format, making it easier for analysts to obtain actionable insights from it. This reduces the overall likelihood of errors.

- **Reducing costs**: Data warehouses allow business managers to dig deep into historical data and gauge the success of past endeavors. This is an excellent way to refine your approach, drive business growth, and minimize costs.

2. Type of data

The next critical consideration is **determining the type of data** you plan to store in the data warehouse. For instance, will it be **structured or unstructured**? Based on this, you will choose between a **relational database and a non-relational database.**

Essentially, a relational database is used for structured data that can easily fit into neat rows and columns of a spreadsheet. Alternatively, a non-relational database is suitable for semi-structured or unstructured data that comes from various sources such as images, emails, social media posts, audio, videos, etc.

Additionally, for unstructured data, building a data lake may also be a feasible choice. A relatively new concept, a data lake helps you to maintain a 360-degree view of all digital and in-store touchpoints, thus engaging your customers at every stage.

Algoscale's professional data lake consultants can help you grow your business by leveraging a data lake for eCommerce, retail, & CPG. Click here to read more.

3. ETL vs. ELT Consideration

The seamless movement of data from multiple sources to the data warehouse and the associated transformation takes place through an extract-transform-load (ETL) or extract-load-transform (ELT) workflow.

Now **deciding between ETL and ELT is a crucial step** that must be undertaken before building a data warehouse. In an ETL workflow, the accumulated data is transformed before it is loaded into the warehouse. Thus, there is no need for further transformation to create analytical reports. Until recently, the ETL functionality had been the standard that was deployed in almost all data warehouses. However, now, several data warehouses also use ELT which does not require complete data transformation before it's loaded. It can be transformed as and when the need arises.

It is seen that ELT workflow offers organizations the following benefits:

- With an increasing number of businesses moving to the cloud, ELT is a **much more feasible option.** It is agile, requires less maintenance, and enables businesses of all sizes to make the most of current technologies to discover insights.

- With ELT, organizations **don't need to necessarily know the transformation logic** at the time of designing the data flow structure.

- ELT allows for **better handling of unstructured data** as, in most cases, organizations don't know beforehand what they need to do with the data.

It is important to decide between ETL and ELT before building a data warehouse. An ELT workflow requires a warehouse that has an extremely high processing ability. Our experts at Algoscale can help you understand the advantages and challenges of both workflows to make sure you make the best decision for your business.

CHARACTERISTICS	ETL	ELT
The order of the process	Data is pulled, moved and transformed on the Staging layer, and then transferred to the target server	The data is pulled and transferred directly to the target server where the transformations will be performed.
Maintenance	It requires more maintenance and more knowledge	Virtually maintenance-free as we move raw data
Processing time	Processing time increases as the data volume increases because all transformations must take place	Processing time is significantly less dependent on the amount of data, because we migrate raw data
Infrastructure	An on -premises environment that is expensive and difficult to scale is essential	It uses cloud services such as Saas or PaaS, which do not need to be installed. They enable dynamic scalability.
Costs	High initial and running costs	Low start-up costs, downstream costs depending on data volume

4. Architecture Consideration

Designing a good data warehouse architecture is a complex task as it deals with historical data from multiple sources. There are three different approaches for **building the layers of a data warehouse:** single-tier, two-tier, and three-tier.

Single and two-tier architectures are not usually recommended due to their limitations. On the other hand, the three-tier architecture is widely used and comprises three layers:

- **Bottom tier:** This is the database of the warehouse. All the gathered data is cleansed, transformed, and loaded into this layer.
- **Middle tier:** This layer includes the OLAP server and serves as a mediator between the database and the end users.
- **Top tier:** This layer includes all tools and APIs that you use to get data out from the warehouse.

At Algoscale, we recommend you opt for architectures that are based on massively parallel processing. Even if your current use case necessitates only minimal processing abilities, we still do not advise you to get a single instance-based data warehouse as it is immensely difficult to scale. Get in touch to know more.

5. Choosing ETL Tools

Another significant consideration when building a scalable data warehouse is selecting the right ETL tools. Essentially, an ETL tool handles the execution and scheduling of all data mapping tasks.

You can choose from the following options while selecting ETL tools for your data warehouse:

- **Custom-built tools:** As the name suggests, custom-built ETL tools are built from scratch within the organization. It leverages a myriad of languages and open-source frameworks to build a custom ETL framework that performs tasks by following the business logic and configuration provided. Although an expensive option, custom-built ETL tools feature superior interfacing ability with the internal sources of data.
- **Third-party managed ETL tools:** Many data warehouse providers such as Microsoft or AWS offer ETL tools as a service. The most popular ones include AWS Glue, Apache Nifi, and Airflow. By opting for these, you can avoid the hassles of designing, developing, and maintaining ETL tools and only focus on proving business logic. However, these third-party managed tools may not offer seamless interfacing ability with your data sources.

At Algoscale, we can help you move your data from hundreds of different sources into your data warehouse without any hassle. We make the entire process of building and implementing a data warehouse as simple as possible for our users.

Read to find out the difference between Apache Nifi and Airflow and how the two can benefit your data warehouse capabilities.

6. Metadata Consideration

Next on the list is metadata consideration, which is a crucial component of the data warehouse architecture. Metadata provides a foundation for data and elucidates the database in a data warehouse architecture. It facilitates the process of creating, processing, maintaining, and utilizing the data warehouse.

Now, two types of metadata are found in a data warehouse. These include

- **Technical metadata:** Technical metadata includes data about the data warehouse that is used by designers and warehouse administrators. This data enables them to seamlessly conduct warehouse development and management tasks.

- **Business metadata:** This data is very important as it provides context for information that is stored in the data warehouse. In other words, it helps end-users to get a perspective on the data present in the warehouse.

7. Using an Agile Approach

Before building a data warehouse, you must consider the approach you want to take. Often, modern data warehouses take several months, sometimes even years, to build. During this time, businesses cannot realize any value from their investment. Additionally, the requirements of every business evolve with time, and many times, differ drastically from the initial set of requirements.

Therefore, it is best to opt for an agile approach instead of a Big Bang approach. With the latter, the projects are often put on hold because of the long duration. However, with an agile approach, you can ensure that your data warehouse evolves as per your changing business requirements.

At Algoscale, our data warehouse experts follow an iterative process. This means that the warehouse is developed in multiple sprints while accommodating all of the business's changing requirements. Click here to know more about our approach.

8. Support of BI Tools

Business Intelligence (BI) tools allow for the easy accumulation, integration, examination, and presentation of business data. These tools enable businesses to make premeditated business decisions that are aimed at growth.

Now, combining BI tools with your data warehouse is an excellent practice as it helps business enterprises drive smart and strategic decision-making. Here are some tasks that you can perform by uniting the power of BI tools and data warehouses.

- Conduct **data mining**, i.e. extract usable data from an extensive set of raw data. This will help to discover trends, patterns, and themes.

- Analysts can **query data** to ascertain its accuracy. They can have all the data in the right place and ask questions about their data to obtain reliable and quantitative information.

- Conduct **data visualization**, i.e. represent data in the form of charts, infographics, diagrams, etc. to boost understanding and make better and informed decisions.

- Translate **data analysis** into layman's terms via data storytelling. With the right BI tools, these stories will be most persuasive to drive strategic decisions.

While many organizations use data warehouses without implementing BI tools, the process will only put a strain on transactional databases, increase load time, and reduce performance. At Algoscale, we make sure that your data warehouse supports good BI tools such as PowerBI and Tableau. This helps to maintain a strong relationship with your data and use it to power the most efficient business decisions.

9. Providing a CDC Policy for Real-time Data

Another thing to consider while building a data warehouse is defining a CDC policy. CDC refers to Change Data Capture. It ensures that you capture any changes that are made in a database in real-time. For instance, a lot of historical data gets changed or even lost when loading new data into the warehouse. But with a CDC policy, every change made to the data is well-captured and stored in relational tables called change tables.

When evaluated carefully, these tables provide an overview of historical data and how it has changed over time. Additionally, CDC can also be used to manage real-time analytics dashboards and easily optimize data migrations.

10. Enduring Storage Costs and Maintenance

Now that you have considered all of the above factors and built a data warehouse, you also need to think about its ongoing costs and maintenance. In most cases, the cost of maintaining a warehouse is much higher than the cost of building one.

To make sure your business will be able to seamlessly operate, here are the different ongoing costs you must consider:

- **Storage and compute:** As your monthly data grows, so will your storage bill costs.
- **People:** Devote time to prepare a good team to keep your data warehouse system running smoothly. Delegate tasks, such as performance tuning or vacuuming, to ensure effective functioning.
- **Opportunity costs:** Maintaining your own data warehouse is a highly time-consuming job. And often, this translates to lost opportunities. So consider your opportunity costs and everything that you are missing out on before making any decision.

Investing in building and maintaining a data warehouse is a big decision that will impact you in the long run. So give some thought to where you are now and where you will be in the next six months to make a choice that is feasible for your business.

How Can Algoscale Help?

Building a data warehouse is an excellent solution to centralizing and quickly analyzing your business' data. It improves data availability, increases analytical capabilities, enhances the quality of information for reporting and analysis, and makes working with data a secure process.

The experts at Algoscale can help you build and implement a customized data warehouse that is meant to fulfill your precise business requirements. **We will help you build a data warehouse that scales as your data needs grow, ensuring you never have to lose critical data points.**
Also, our customer support provides first-rate support for all your queries and issues.
Schedule a personalized demo today to find out more about our solutions.

algoscale

Contact us

Get all your questions answered by our team.

India:

+91-120-416-5801

US:

+1-862-234-9997

askus@algoscale.com

Noida, India

Algoscale Technologies Private Limited

D-76, Sector 63
Noida, UP 201301

NJ, USA

Algoscale Technologies, Inc.

One Gateway Center Suite 2600
Newark, NJ 07102