The True Cost of

BUILDING A DATA WAREHOUSE

Discover Hidden Costs and Avoid Unpleasant Surprises

Cooladata

Cooladata

TABLE OF CONTENTS

Cooladata

In today's fast-paced digital marketplace, business intelligence generates a lot of information. Most of this **raw data** comes from real-time analytics tools. Yet to be truly useful to a business intelligence team, all information needs to be collected and **unified**.

More often than not, a **data warehouse** solution is the perfect fit. It puts all of your heterogenous data into one place, automatically archives all of your data, and makes your information easy to access and analyze by your team.

And yet, a "data warehouse" often raises more questions than it answers. What's the difference between a data warehouse and a database? What resources do we need to build one in-house? What about if we build it off-site? How much will all of this cost?

Luckily, we're here to help.

In this article, we will cover several basic **types** of data warehouses and the **components** required to build them. We will begin by defining the essential components of data warehouses:

- Required hardware
- Essential software
- Human resources to maintain a data warehouse

Then, we will evaluate different **costs** associated with maintaining and expanding a data warehouse. At the end of this article, the long road ahead should be a bit more traversable.

WHAT IS A DATA WAREHOUSE?

The thought of consolidating so much data into a **single stream of data** may seem like a nightmare. Yet it really isn't so scary if you know the requirements and costs associated with building or buying this solution.

Understanding data warehouses doesn't have to be intimidating

Before discussing the costs and investments of a project this large, we should take a moment to **define** some terms.

There's an important distinction exists between a data warehouse and a database. Regular databases usually store information in a way that facilitates their retrieval for one type of service or a certain kind of data.

In other words, **databases** are great for a specific business process, but are under-optimized for analytical queries.

"Databases are great for a specific business process, but are under-optimized for analytical queries"

Enter in the **data warehouse**, which combines **many different sources of information** (possibly from many databases) into a format that is suitable for **analytical use.** Ideally, a data warehouse should automatically refresh its contents in order to keep up with the intelligence and live data sources that feed it information.

If you're interested in building a data warehouse from scratch, you should know that there are three major components:

1. **Storage**: your data warehouse will need servers that are either cloud-based or on-premise.
2. **Software**: your data warehouse will need software that pulls data from your live-streaming services and unifies them on the servers.
3. **Human Resources:** your data warehouse will need routine upkeep as well as data analysts who will use the data to garner actionable insights.

First, let's take an in-depth look at the storage costs and requirements of a full-fledged data warehouse.

GETTING STARTED

So what **requirements** should your data warehouse have?

It has to be **fast**: data warehouses need to be able to manage multiple users and lots of server requests simultaneously, so the faster your write and read speeds, the better.

Data warehouses also have to be **scalable**. A good big data platform needs to be able to handle

over a billion events easily, ideally without losing access to individual events.

Finally, a good data warehouse needs to **store** months of your analytics at a time. A permanent history option would be a perfect fit for most business intelligence solutions, though that option may be too cost-prohibitive for smaller businesses.

"A solid data warehouse must be fast, scalable, and have ample storage."

1. Storage

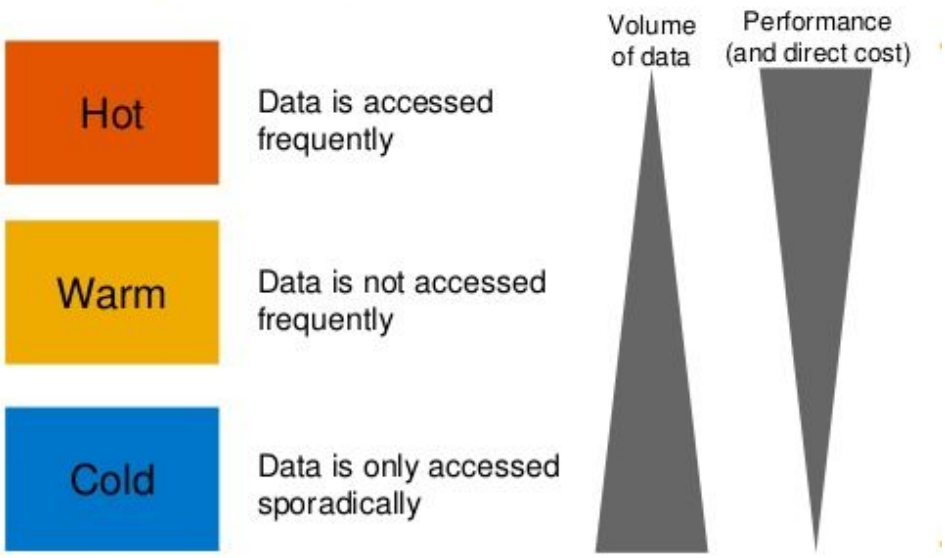
The first decision you will have to make is whether you want to store your data warehouse on in-house servers or in the cloud. If you already have a scalable server solution built for your business, this part is more manageable because you can expand some of your storage to fit a data warehouse.

Let's assume, however, you are starting completely from scratch.

Cloud Storage

Storing your data warehouse using cloud-based storage may be a viable option for businesses seeking a faster and more-easily scalable data warehouse storage solution that doesn't require any up-front hardware expenditures.

There are three types of cloud-based storage you should be aware of: cold, warm, and **hot** storage. The basic idea is that the "**temperature**" of your data determines how often that data is accessed.



Hot data is accessed frequently, requiring the best storage available

Since your data warehouse will be feeding into your analytics, you will need "**hot**" storage -- the fastest and the most accessible cloud-based storage available -- in order to have a functional data warehouse.

Most analytical storage solutions also charge an extra fee for accessing the data stored on their servers. Each query and event will cost you: BigQuery charges an extra five cents per gigabyte, while Azure charges five cents for every 10,000 rows of data that it has to process. While BigQuery or Azure are cheaper for storage than Redshift, using your data will cost extensively will most likely end up costing extra in the long run.

Below is a price breakdown for three of the top "hot" storage solutions currently available on the market.

1. **Amazon Redshift.** Amazon's hot storage solution costs **0.08$ per gigabyte (GB)**. If that sounds low, it isn't: Amazon's solution is the priciest of the bunch. Luckily, Redshift offers this solution at a flat cost, meaning that the prices won't go up or down depending on your data use. Still, expect to be paying over **$1000 per terabyte (TB) per year** if you're looking to use Amazon.

2. **Google BigQuery.** Google's storage solution costs **0.02$ per GB**. Don't let that low price fool you: this is a variable cost. Google also charges an additional price for accessing this data -- you will be paying an extra $0.05 for every 100GB that the service accesses for you. If you're running a data warehouse that is accessed often by your analytics team, then the prices can really start to add up. Expect to be spending around **$720+ per TB per year** if you're using BigQuery.

3. **Microsoft Azure**. Azure does pricing in blobs, which is a fun way of saying that prices go up the more data that you store on Azure's servers. So for the first 50 TB per month, you will be paying **0.0184$ per GB, after that, the prices go up**. Again, if that seems low, it isn't -- Azure charges "access prices": in their case, this means that every 10,000 write operations will cost an extra 0.05$. If we think in terabytes, then Azure will likely cost **$700+ per TB per year**.

Features	Amazon Redshift	Google BigQuery	Microsoft Azure
Per TB/month	$81.92	$20.48	$18.43
Per Event	N/A	$0.05/100GB	$0.05/10k Queries

Total cost for storage: in today's marketplace, every additional terabyte of data will cost you upwards of $1000 per year in storage and access costs. If we had to make a conservative estimate of total costs, then a mid-sized data warehouse would use about **$12,000 per year** in storage alone.

"Storage costs for an in-house data warehouse can be $12K per month."

Cloud Storage Challenges

The trouble with cloud storage is the sheer unpredictability of the pricing. Since most solutions use access pricing, certain months of heavy analytics can make the bill significantly heftier than usual. For instance, loading data, allocating instances, and creating clusters will all incur access prices. Streaming and querying the data for visualization events will also pad your bill at the end of each month, so keep that in mind when selecting the right solution for your company.

Another issue is the constantly inflating prices. Most data warehouse solutions can easily burn through one terabyte of data per month, which they will have to archive alongside last month's data.

While using cloud service may eschew initial build costs, the unpredictability of the pricing models and their constant fluctuations means that you will never be sure what your per annum data warehouse costs will be, especially if your data warehouse storage grows heavily during that time.

2. Software

Most software with this functionality is called **ETL software**, ETL being short for **Extract**, **Transform**, and **Load**.

While ETL can be done through certain open-source solutions like Apache Hadoop and Talend, let's take a look at software options that use ETL to incorporate popular database types and SaaS platforms into your data warehouse.

Data Centralization

Here are a few options that centralize your data on the cloud storage services outlined above. Most of the current offerings integrate fairly easily either into Amazon Redshift or into Google BigQuery:

1. **FlyData.** FlyData is a fully-featured solution that syncs your databases with your Amazon Redshift storage. The service supports VPN tunneling and error handling, as well as multiple database formats: MySQL, CSV, TSV, JSON, and PostGRESQL, among others. FlyData does not provide any visualization solutions, so it may have to be used in parallel with services that do provide that functionality, such as Chartio or Tableau. Pricing is variable, and ranges anywhere **between $200 and $2000** per month.

2. **RJMetrics.** RJMetrics is similar in functionality to FlyData, but offers more robust support for SaaS products such as Salesforce, Shopify, and Google Ecommerce. Otherwise, it supports many of the same databases as FlyData, with the exception of also supporting MongoDB, Microsoft SQL Server, and Heroku. Pricing is virtually the same as Flydata: between $500 and $2000 per month.

3. **Fivetran.** Fivetran, much like FlyData and RJMetrics, funnels all of the data from your databases and SaaS services into your Amazon Redshift or Google BigQuery database. It's important to note the service's pricing aren't readily available online, though we'd expect their pricing model is similar to FlyData and RJMetrics.

Features	FlyData	RJMetrics	Fivetran
Less than 10 people	$200	$500	$200-500
More than 10 people	$2000	$2000	N/A

After you have your ETL software installed and ready to go, you're pretty much finished building the software-side of your data warehouse -- all that's left is visualization.

Data Visualization

While some enterprise business analysts work exclusively with open-source visualization such as R and D3.js, you may want to consider investing into some data visualization services that can automate your team's workflow. Some popular options include:

1. **Chartio.** A business intelligence tool that generates visualizations in real-time. The service has a drag-and-drop interface that writes SQL queries for you, although its learning curve has been criticized in the past. Again, the costs of this service are dependant on the size of your solution, and can range anywhere **between $500 and $2000** per month.

2. **Tableau.** Tableau offers many of the same services as Chartio, but offers a user-based pricing model: the software costs **$70 per user, per month**. So if you have a team of ten people using Tableau, you will be spending around $700 per month.

3. **QlikView.** Qlikview also offers pricing per user, starting at $25 per user, per month. This means that you'll be paying upwards of **$250 per month** for a ten-person team.

Total cost for software: taking a conservative median price for both ETL and visualization software, your projected costs will most likely be around **$2000** per month, or **$24,000 per year.**

Features	Chartio	Tableau	QlikView
Per user/month	~$40	$70	$25
Enterprise-level (10+ users)	$2000	$2000	$250

"Data warehouse software costs can be $2K per month, or $24K per year."

Keep in mind this is a ballpark estimate. When starting to build your own in-house data warehouse budget, consider the following:

- Your software prices are bound to go up as time passes. Increased storage brings increased ETL usage and higher storage access prices, inflating the cost of your data warehouse per month.

- Pricing models are wildly unpredictable, depending entirely on the size of your storage, the amount of data you have going into your warehouse, and the size of your team.

We've put together the below chart that organizes the storage and software options we've covered, along with some others to consider:

Component	Cloud Service/Open Source	On Premises/Private Cloud
Collectors	Cloudfront, Amazon Kinesis	Storm, Kafka
Process	Amazon EMR , Google data pipeline	Hadoop distributions, Talend, Informatica
Storage	Amazon S3, Google storage	EMC, IBM, HP
Analytical DB	Google Big Query, Amazon Redshift, Impala, Spark	Vertica, Exasol, Infrobright
Real-time DB	MongoIO, Redis Labs	MongoDB, Couchbase, Cassandra
Visualization Analytics	ChartIO, D3.JS, Google Spreadsheet	Looker, Tableau, QlikView, MicroStrategy

Data warehouse software and storage service options

Now that you have your software and hardware ready to go, you will need a support and visualization team ready to run your new business intelligence solution.

3. Human Resources

For your database support team, you will need a dedicated Systems Manager, Backend Developer, and a Software Engineer. These people will make sure that your data warehouse is running smoothly and that all data is manageable and secure.

For the visualization team, you will need at least one Data Analyst, (although a Data Visualization specialist would be a plus).

Below we've outlined job descriptions and conservative cost estimates associated with staffing a fully-operational data warehouse. The salary ranges were assessed using the median income of each respective position on payscale.com.

1. **Information Systems (IS) Manager**. An IS Manager is required to oversee the team running your warehouse and keep all the systems in check. Expect to pay around **$7000-10,000 per month** for the services of a qualified IS Manager.

2. **Backend Developer.** Backend developers are responsible for installing and maintaining all of your ETL software and making sure that it is working in tandem with your storage service. Backend developers often offer their services for around **$6000-8000 per month.**

3. **Database Architect (DBA):** A DBA will determine the structural requirements of your data warehouse and propose the best solution for unifying all of your existing data sources into it. A qualified DBA will cost around **$10,000-12,000 per month.**

4. **Data Analyst:** Your data analyst will be responsible for analyzing and visualizing your business intelligence in a way that will produce actionable insights. A data analyst will also run you **about $5000-8000 per month**.

Total cost of human resources: in short, staffing can easily be the most expensive part of your data warehouse. Assuming you only get one person for each position outlined above, your costs can be as high as **$28,000-$38,000 a month**, or roughly **$432,000 a year**.

	1 TB per Q	1TB per Month	3TB per Month	Monthly Costs
BackEnd dev	1	1.5	2.5	$8,000
Infra/system MGMT	0.2	0.5	0.5	$10,000
DBA	0.3	0.5	1	$10,000
Analyst	1	2	2	$8,000
Total headcount	2.5	4.5	6	
Total monthly headcount cost	$21,000	$38,000	$51,000	

HR Costs for maintaining an in-house data warehouse

If your data warehouse grows in size (as it undoubtedly will), you will need more than one person for each role, easily doubling or tripling your costs in data warehouse maintenance.

ADDING UP YOUR COSTS

Now that we've covered all of the components necessary to build your data warehouse, let's put the costs together. Assuming you want to build a data warehouse that will use, on average, one terabyte of storage and 100,000 queries per month, your total yearly cost for storage, software, and staff will be around **$468,000.**

"Annual in-house data warehouse costs can be around $468K."

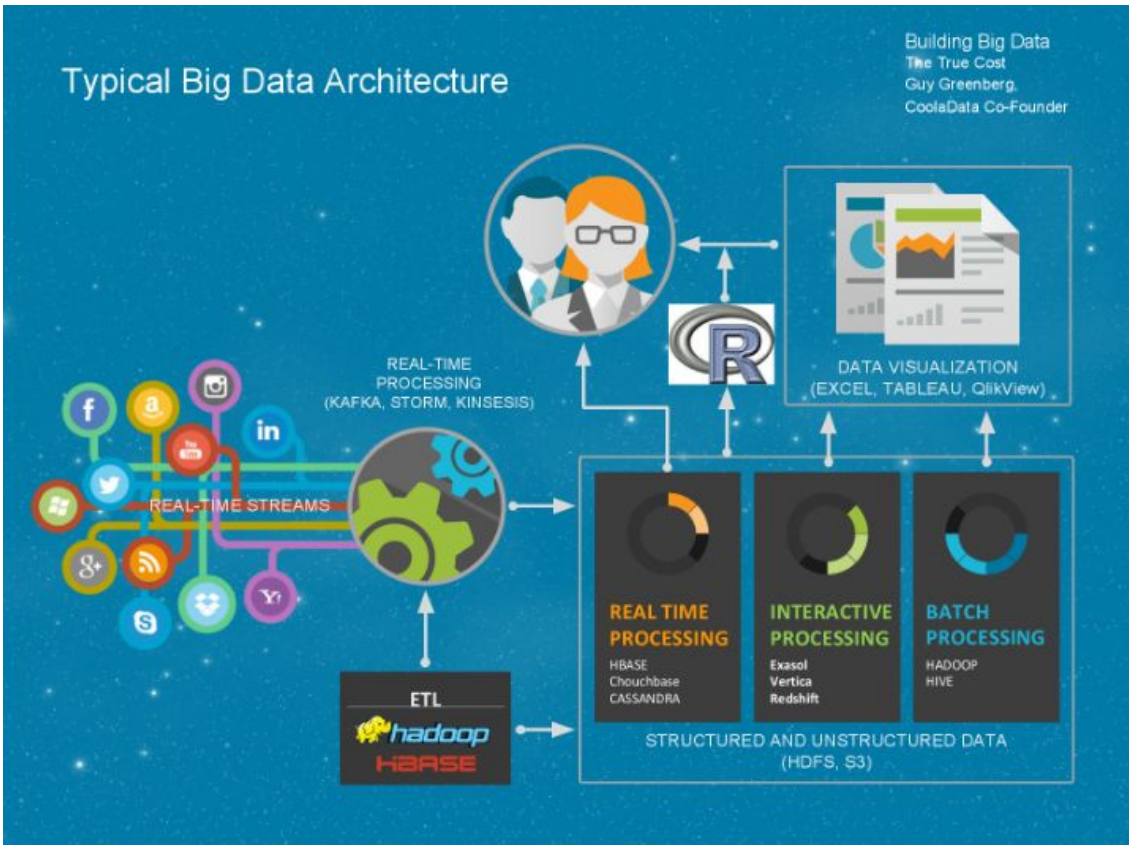
Keep in mind that these costs vary from business to business, from industry to industry, and sometimes from month to month. It is impossible to predict, with any certainty, what your costs will be at the end of the year.

Though the software and technology surrounding data storage and unification are changing all the time, the options listed above should be safe bets in terms of their reliability and price. This article should be used more as an approximation rather than as a concrete roadmap.

CONCLUSION

Building a data warehouse from scratch is no easy task. A large project such as this requires more than a year of setup, configuration, and optimization before it is ready for business intelligence purpose.

And remember, your database warehouse is only one aspect of your entire data architecture:



Typical Big Data Architecture

It is no surprise then, that [according to Gartner](), over 60 percent of all big data projects fail to go past the experimentation stages and are subsequently abandoned.

Sometimes messy problems have elegant solutions. If you're interested in buying instead of

building your own data warehouse, Cooladata's data unification and visualization services might help not only reduce the costs of building your own data warehouse, but also help you get your business intelligence off the ground instantly.

Let us know in the comments what your experiences have been with the services and storage providers listed above. Have you been successful in launching a large data warehouse in the past?

ABOUT COOLADATA

Cooladata is an end-to-end solution that lets companies capture, unify, analyze, visualize and share their data to empower every team to make smarter decisions, faster.

Cooladata provides a secure, fully-managed analytics and data warehouse solution optimized for behavioral and time-series analysis.

- Product teams use Cooladata to improve user engagement, retention and monetization.
- Marketing teams use Cooladata to unify all their data in one platform so they can truly understand the entire customer journey, increase conversions and ROI.
- Data teams use Cooladata to perform ad-hoc analysis of their data, and answer complex questions in seconds without writing long SQL queries.

For data teams considering building their own data warehouse for behavioral analysis, Cooladata gives you a solution as powerful and flexible, but at a fraction of the cost.

Cooladata is backed by 83 North / Greylock Israel, Carmel Ventures and Salesforce Ventures.

cooladata.com