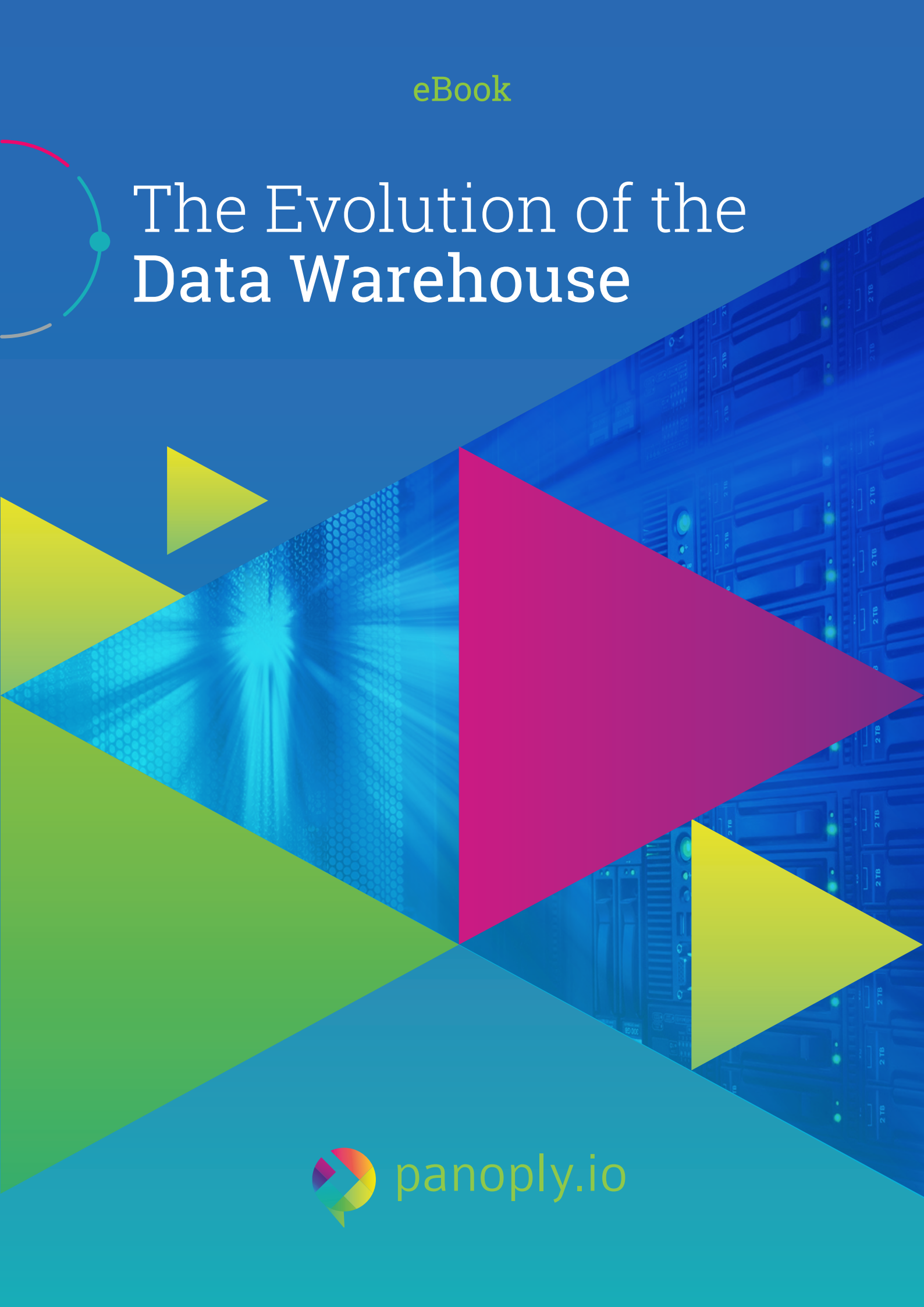
eBook

The Evolution of the **Data Warehouse**



panoply.io

Index

1

Complexity in Data Warehousing

The concept of the "business data warehouse" dates back to the late 1980s when several software companies developed a framework for building decision support systems. While system designers conceived such systems as an approach to reduce the processing impact reporting had on business-critical operational systems such as point of sale (POS) and supply chain management (SCM) systems, they evolved into real-time dashboards that are mission-critical to most organizations. These data warehouse systems evolved over time as overall computer performance increased. This evolution has allowed businesses to collect more data from more disparate sources and ultimately do more with that data.

The Internet of Things

Societal trends such as social media and smartphones, as well as connected devices in manufacturing and logistics, mean that there is an increasing amount of data to collect and connect with traditional business systems for deeper analysis. Moreover, the home device market continues to explode, with new models of just about every appliance in the home now capable of sending a plethora of data points per day to a database.

Data Warehouse Architecture

There are three primary processes and structures involved in the creation of a traditional data warehouse:

- **Extract-Transform-Load (ETL)**
- **The Data Warehouse**
- **Online Analytical Processing (OLAP) Cubes**

The following sections elaborate on these facets of data warehousing.

ETL in A Nutshell

At some point, data needs to move from source systems to an analysis target—typically a data warehouse. This process is known as the extract-transform-load (ETL) process, and it is often the most challenging part of any data warehouse project. The ETL process involves cleaning the data, which means taking data out of a variety of source formats and consolidating it into a format suitable for analysis. As such, ETL can be a time-consuming and tedious process. Moreover, the ETL process is not always a static one. It can change midstream as business requirements shift, leading to delays and additional work in the course of the analytics effort.

Many ETL processes run just once each day, which means business users often do not see their most recent data; the results can be up to 24 hours out of date, which can feel like a lifetime for some businesses. Although efforts have been made by many organizations to provide more real-time delivery of data into the data warehouse, this can be a chal-

lenging process, as it involves adding the ability to capture changed data from the source systems in real time. Capturing real-time data changes adds physical and administrative burdens to the source systems.

Tangible System Components

After loading the data into the target system via the ETL process, there are two more components in the big picture of an analytics strategy: the warehouse itself, which handles batches of queries, and optionally an online analytical processing (OLAP) cube, which supports the use of ad-hoc queries without overtaxing the warehouse itself.

Finally, sitting atop the data warehousing system there is the reporting and data visualization layer, which allows business users to create insight from data through visualizations such as dashboards and reports, which make analysis easier. All components of these decision support systems have evolved in recent years to keep pace with the explosion of data volume tracked, the new types of data developed, as well as new data source types that now exist.

Database Pace Losing Face

In many cases, traditional database vendors have not been able to innovate rapidly enough to continue to meet the ever-changing needs of their data analysis customers..

A rather recent trend in the space is the concept of a data lake. This term refers to a storage repository that holds data in its native format until an analysis process acts on it. A unique ID is applied to each piece of data upon ingestion. Data lakes are discussed in more detail in a later section.

From a cost perspective, data warehouses are some of the most expensive resources in an IT organization. Most of this comes down to pure infrastructure costs—on-site enterprise storage is still quite expensive, and data warehouses are colossal systems ranging from terabytes to petabytes.

Trends in Computing

Over the last decade, the amount of computing power that can be brought to bear to work on larger volumes of data has increased dramatically. However, even with significant raw computing power, traditional relational database systems may encounter major performance issues when trying to query large volumes of operational, transactional data, resulting in what has become known as a big data problem.

That's Some Big Data!

The term big data is widely overused and is associated with several types of systems, including Hadoop and massively parallel processing (MPP) data warehouses. Since both data warehouses and big data solutions follow the pattern of writing data once and reading it many times, they are prime candidates to be optimized for read performance.

What is Hadoop?

Hadoop is an open-source software project that provides a platform to store and process massive data sets. The Hadoop Distributed File System can handle very large files as well as store an astronomical quantity of files. The MapReduce framework processes data in parallel which results in substantially increased performance over the serial processing methods used historically.

These systems combine the processing power of multiple servers working together to analyze massive amounts of data quickly. While the implementation of each of these technologies is quite different, they both serve the same purpose—to quickly process lots of data. Other technologies in the space, such as columnar data storage, allow for massive amounts of scalability for those types of queries. Columnar data stores will be discussed further in a following section.

Abstraction, Consolidation, and Virtualization

Virtualization is another development in computing that laid the groundwork for many other technologies. Like client-server before it, virtualization allowed companies to increase density in their server rooms and harness the power of modern processors. Virtualization has become the de facto standard architectural choice for new applications and services in the data center, and it has become important for one key reason: Virtualization enabled infrastructure to become software-defined, which allows for high degrees of automation. This concept is called "software-defined infrastructure."

In essence, it means things like networks, storage, and server configuration can all be turned into software and can be automated.

What Does "Software-Defined" Really Mean?

In a software-defined data center, all infrastructure is abstracted in some way from the underlying hardware – generally through virtualization – pooled, and the services that operate in the environment are entirely managed in software.

The Great Democratizer

Cloud computing is the most impactful innovation in computing in a generation. What started with the outsourcing of basic computing tasks such as e-mail has evolved into organizations' having far more options regarding where and how to run workloads. In some cases, companies have outsourced all of their computing tasks to third parties, such as Microsoft, Amazon, or Google. Cloud has changed many paradigms, democratizing many parts of IT that only large enterprises could afford in the past.

A good example of this is massively parallel data warehouses. Before cloud computing came along, the only way to have an MPP data warehouse was to buy a very expensive appliance, which likely included a costly support and services contract. Most companies simply could not afford the investment, which often carried a starting price tag of a million US dollars and could even run into the tens of millions for some

systems. With cloud computing and the ability to "rent" hardware and licensing, companies can get up and running in a matter of minutes, and have data streaming into their cloud-based data warehouse shortly thereafter. Best of all, the up-front investment is practically zero dollars, since companies pay only for what they use.

Another factor in cloud computing's favor is the ability to scale computing resources up and down as workload demands increase and decrease. For a data warehouse, which users primarily leverage during business hours, a small domestic company can reduce resources allocated to the service during the overnight period, reducing their overall costs to rent the platform.

Although there are a number of different ways to consume cloud-based services, one common model is called platform as a service (PaaS), which is very cost effective and allows the service provider to rapidly make changes to meet the needs of end customers more quickly. For example, most database vendors have major releases of their software every 1–2 years; in a PaaS model, new code and features can be developed and deployed as often as monthly.

Security and Justified Paranoia

Data is under threat of attack from a myriad of sources, both internal and external. We see the fallout from the attacks on the news almost weekly. Cloud computing does create concerns around security for many. After all, as the data leaves an on-site environment, security may now be in the hands of a provider, rather than internal staff. Traditionally, data

warehouse security leverages roles, with some semblance of row-level security—either at the reporting layer or in the database itself. Encryption of data at rest is common in organizations that have specific regulatory requirements, as is encryption or obfuscation of data within the database.

One common attack vector facing many online data gathering systems revolves around SQL injection. SQL injection can occur when systems retrieve data via an Internet (or intranet) front end (for example, a registration form) with data submission URLs that pass SQL code to the database system. A nefarious user can carefully manipulate these URLs to grant themselves permission in the database and then retrieve or manipulate data in the target system to which they should not have access. Many systems are vulnerable to this type of hacking due to improper programming techniques.

Diversity of Data

Another trend impacting data warehousing projects is an increase in the overall diversity of data. Data no longer resides exclusively in relational databases in a perfect tabular format. Many social media sources, for example, use JavaScript object notation (JSON) for their data; many application programming interfaces (APIs) use eXtensible Markup Language (XML); and some organizations still have mainframe data that is fixed width. This medley of data types can be even more troublesome for the ETL process than the problems previously described. Typically, custom code is required to parse and manage these types of data.

In addition to the custom code required to parse this data, it

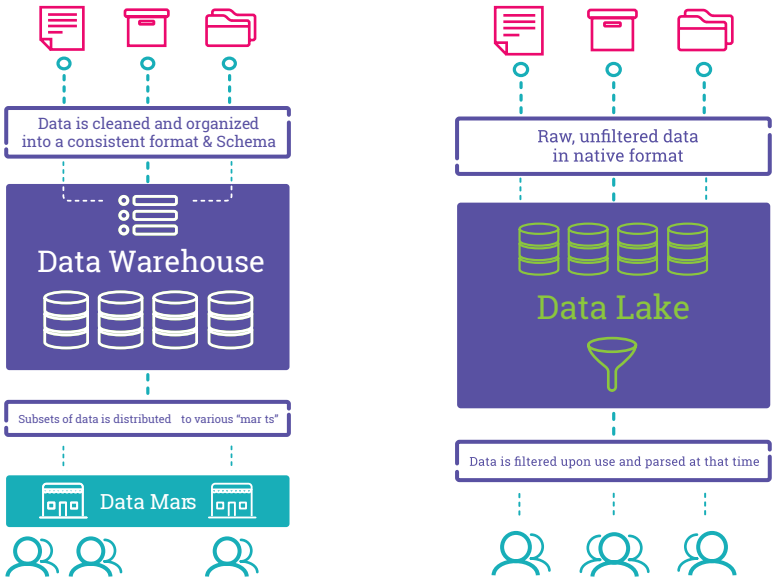
is very expensive to parse from a computational perspective. Moreover, the data tends to have dynamic formats that can break a rigid traditional ETL process.

Data Lakes

A recent trend that takes advantage of a number of advances in computing is the concept of a data lake. A data lake takes advantage of low-cost local storage to house a vast amount of raw data from source systems in its native format until the data is needed for analysis. Each element of data in the lake is assigned a unique identifier and tagged with metadata tags. This data is typically queried to filter to a smaller set of data, which can be deeply analyzed.

Data Warehouse vs. Data Lake

The primary difference between a data warehouse of old and a data lake is the position of data processing in the overall data pipeline. In a traditional data warehouse, the ETL process described earlier cleans and structures the data upon ingest. In contrast, a data lake stores raw data in its native format until it's needed.



Introducing Panoply.io

If you think that all of this sounds rather daunting, you're absolutely right. And that's where Panoply.io comes in. Panoply.io is an end-to-end platform for analytical data warehousing. Its purpose is to abstract away the complexity of the technologies, components, and configurations required to maintain a robust data warehouse, allowing companies to utilize their data with their favorite tools instantly.

Panoply.io sits at the very intersection of all of the aforementioned trends. The offering takes advantage of the facilities provided by Amazon Web Services (AWS), including leveraging the following services:

- **Amazon Redshift** - Scale-out Data Warehouse
- **Amazon Elasticsearch Service** – Hosted Elasticsearch cluster
- **Amazon S3** – Highly durable object storage

These services form the foundation of the Panoply.io solution and you will learn more about how each service works together with other custom components to deliver a robust end-to-end platform for data analytics.

Panoply.io was born in - and lives in - the cloud. As such, Panoply.io can be rapidly deployed, so you can get started with data analysis quickly and eliminate the costly lead time and the capital expense of getting hardware into place. Additionally, the cloud platform makes scaling a breeze as your computing needs grow. Since the infrastructure in the cloud is abstracted as software, scale-up and scale-down can be au-

tomated based on workload demands to synchronize costs with workload needs. Panoply.io takes workload sizing to the next level by gathering data about your workloads and auto-scaling the size of your infrastructure predictively based on observed usage patterns.

It gets better. Normally, when designing an application in the cloud or on-site, the architect has to choose what size virtual machines and what type of storage to use. Instead, Panoply.io uses the query information mentioned above and applies it to your underlying hardware configuration. Tuning hardware based on observed queries means that you always have the optimal hardware configuration (from both a cost and performance perspective) for your workload, with minimal effort on your part.

Data Security

Security is a major concern identified by a number of customers regarding cloud computing. They need to know: by shipping their business data off-site to a cloud provider, what risks are they incurring?

The answer is good news for many.

In many cases, cloud provider security protocols are far more rigorous than those found in enterprise data centers. Cloud providers have worked doggedly to provide security at every level of their environments. Their very business depends on not being breached. This focus on security, combined with their high levels of data protection, means your data is just as secure—if not more secure—at most cloud providers as it is on-site.

On top of the security inherent in AWS, the Panoply.io architecture has security at its very core. All of the data stored in Panoply.io is encrypted, both at the file system and application levels. Panoply.io also supports two-factor authentication to protect against social engineering attacks. All data is secured in transport using TLS encryption and hardware-accelerated AES 256-bit encryption. This means there is no performance penalty for protecting your data.

Additionally, Panoply.io has a fully developed permissions model, which means you can restrict access to specific data objects. A particularly compelling feature is the ability to have data access automatically expire after a defined period of time. Log-ins coming from unusual locations trip the alarm provided by an anomaly detection algorithm, and customers have the ability to block those connections if they choose. Finally, all queries are audited and can be reviewed by the customer.

Data Transformation

Panoply.io does not include an ETL tool. As mentioned earlier, ETL can be a painful, manual, and expensive process. In its place, there is a bit of a twist: Panoply.io takes advantage of a more modern process—extract, load, and transform (ELT).

ETL vs. ELT

As reflected in the acronym, the difference is simply one regarding the order of operations. In ETL, data is extracted from the source, transformed, and loaded into the data warehouse. In ELT, by contract, data is extracted from the source and immediately ingested. It is then transformed later on read. Consider how this process structure relates to data warehouses versus data lakes.

ELT is quickly becoming the norm for big data systems where the schema is applied upon read. This transformation model allows users to write their transformations in SQL or Python and represent the transformations by displaying the number of views. This process has the advantage of working retroactively on data through a simple code change to the view as opposed to making major changes to the ETL process.

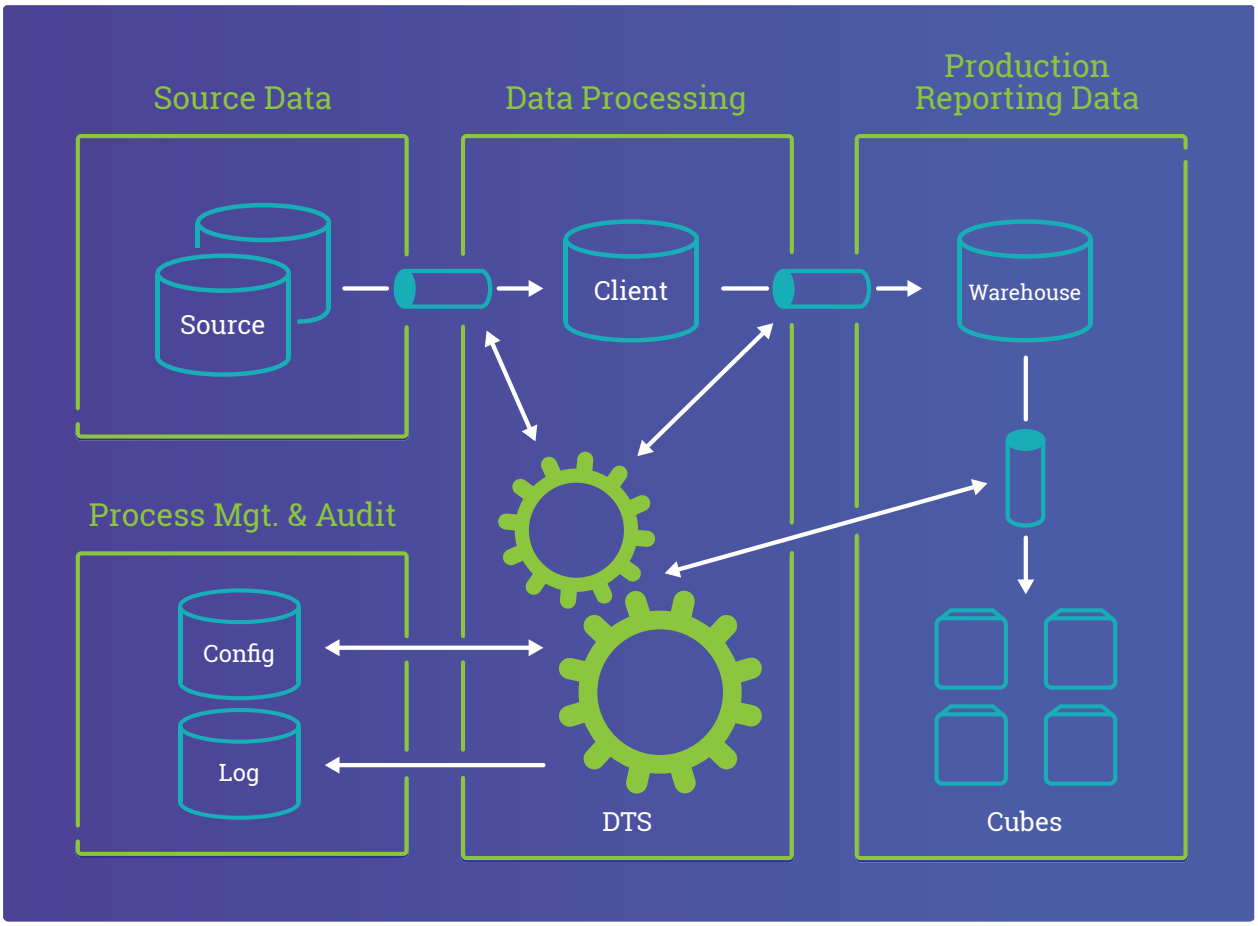


Figure 1 – An example data transformation process (designer needs to redraw picture)

As far as extracting and loading data goes, Panoply.io offers an array of default data source configurations such as enterprise database systems and various web services that offer a data extraction API. If Panoply.io hasn't already built an

integration that meets your needs, a framework is provided to allow users to create custom integrations with other data sources as well.

Administrators can schedule all of these data sources for periodic updates, or you can push data directly from your code into a Kafka cluster (a streaming data solution) using the provided SDKs. Pushing data via code is possible with a variety of languages, and the Panoply.io platform processes them in real time.

This automated process is a significant step forward in data warehousing. By eliminating the largest time sink in data warehousing projects, Panoply.io speeds up the time-to-value for a data analytics solution. Many common data types and formats (CSV, JSON, XML, and log files) are automatically identified by the system and parsed accordingly. Nested formats like lists and objects are flattened into tables with a reference structure that is built automatically. Data type discovery is performed on all ingested data, as is relationship detection between tables. Slowly changing tables are automatically generated for all of your data, which essentially provides a version history, allowing you to use SQL to query your data at any given point in time. Accessing this historical data without the auto-generated data versioning was a manual process that required a great deal of effort on legacy systems.

Having this kind of data access also allows tighter integration of external data sources into business intelligence eco-

systems. For example, you may wish to incorporate social media data from advertising campaigns to align sales data with marketing efforts. Since most of the resources in question probably export JSON, and Panoply.io supports API-based ingestion, you can include social data in your analysis with minimal effort and see deeper, richer detail. Additionally, you can easily integrate with other popular data sources like Salesforce and Google Analytics.

Integrated BI Tools

Integrated reporting tools have always been part of the business intelligence (BI) landscape, as far back as Crystal Reports and Brio. Such visualization tools are a vital part of the BI process. Recently, comprehensive external tools like Tableau and Microsoft Power BI have become very popular, as they support a variety of modern data sources and deliver powerful visualizations across large amounts of data. These tools are extremely popular with business users, as they can quickly develop data models and charts without the user needing to have comprehensive knowledge of SQL or any other programming language. In addition to Tableau and Power BI, Panoply.io supports tools such as R and Spark for statistical and streaming analysis of data.

2

Houston, We Have an Efficiency Problem

Especially in larger enterprises, the lead time that is required to procure storage hardware can delay – or even derail - the start of an analytics project. There tends to be a large degree of organizational friction caused by IT resource silos, and that friction has a tendency to impact the progress of projects. It certainly can complicate the initial launch of data projects unnecessarily. Additionally, the tension between the multiple teams involved in a project often lengthens the project lifecycle. For example, the following teams might need to be involved to do a data warehouse project in an enterprise with a large IT organization:

- **IT Infrastructure**
- **IT Storage**
- **Database Administration**
- **Development**
- **IT Business Intelligence**
- **IT Project Management**
- **Business Team Requesting Project**

Managing and aligning all of these resources to reach a common goal is quite challenging, and this is why in large organizations these projects can take years to get off the ground—not to mention get completed. Worse yet, sometimes the projects aren't even successful. There is often contention for these resources, especially the shared ones such as the infrastructure and database teams, which means the data warehouse project may not be the highest priority for the IT organization.

Deploying Analytics Systems Is Tough

Besides the political and organizational challenges of deploying and maintaining an analytics platform, there are significant technical challenges as well. The following are some of the most intimidating hurdles for any organization attempting a deployment or refresh of these types of systems today.

Development and Data Logic

When trying to build an analytics system in-house, many organizations reach out to third-party consulting firms to support the development of these systems. While this can be beneficial since the resources are fully dedicated to the project, it can also cause delays and missteps in the project because in-depth knowledge of your business rules and processes is required while building the solution. Hired guns aren't always properly brought up to speed, including the transfer of tribal knowledge and undocumented intricacies of the business. Miscommunication and misunderstanding are common due

to the lack of intimate familiarity with the company. Even when working with internal development staff, it can be challenging to convey the requirements that are needed.

Data warehousing projects tend to follow more of a waterfall model of development than an Agile one, which means rework tends to be more comprehensive and costly in terms of time and resources.

The other challenge that organizations must overcome is the cumbersome and fragile nature of the ETL process. Every change to some downstream processes also potentially impacts other processes all the way back up to the top of the ETL process. This cascade means reconfiguring and redeploying code in addition to going through new testing cycles. While some attempts have been made to better automate ETL processes, they are limited in their adoption. Many organizations still rely on manual techniques or primitive scripts that can be greatly impacted by changes.

Since development of the ETL process is customized and newly developed for each customer, there is limited use of frameworks and repeatability between industries. This means that almost every time an organization wants to develop an analytics system, they are starting from scratch with regard to their ETL process.

Systems Management

For smaller companies, one of the benefits of cloud computing is that they are effectively outsourcing the management

of their infrastructure. As mentioned earlier, large IT organizations have a significant percentage of their staff dedicated to keeping the lights on. The IT Operations team usually performs some of the following duties:

- Managing storage and free space
- Patching servers and databases
- Replacing failed hard drives
- Managing backups and restores

In a smaller organization, all of these tasks might be handled by just one or two people who also tend to have other responsibilities in the IT organization. They can quickly become overburdened, however, which makes proactively implementing new projects problematic. Just getting the hardware for a project can become challenging in a smaller organization, much less getting a large project planned and implemented.

Performance

In organizations of all sizes, there are often issues with the performance of large data systems. When you start working with terabytes and petabytes of data, systems must be optimized for ideal performance. In the traditional relational database world, this meant optimizing indexes across the data warehouse. That optimization used to be a time-consuming, tedious art, but it has evolved over time. In larger organizations, the solution is sometimes as primitive as purchasing more robust hardware to meet the needs of the system. While this can be an acceptable solution, it is expensive, sometimes

even wasteful, and does not solve the underlying inefficiency.

For organizations that have moved to big data systems like Hadoop, performance tuning is even trickier. While there are techniques for optimizing performance of these systems, the expertise in that space is rare and expensive. In some cases, adding more nodes to the cluster can solve the problem, but "throwing hardware at the problem" is not a resolution method that can scale forever.

In recent years, many relational database vendors have made inroads in large system performance by using columnar storage techniques. This structure, used in Amazon Redshift (upon which Panoply.io runs), involves taking tables of data and turning them on their side; the rows turn 90 degrees and become columns. (See Figure 2) Structuring the data in this way has the benefit of introducing a great deal of compression potential into the data. Since the data in columns tends to have a higher rate of duplication, compression of columns can be up to five times more effective than other reduction techniques. The other major benefit of this technique is that the columns not involved in a given query never get scanned; this greatly reduces the storage operations needed to return the query results.

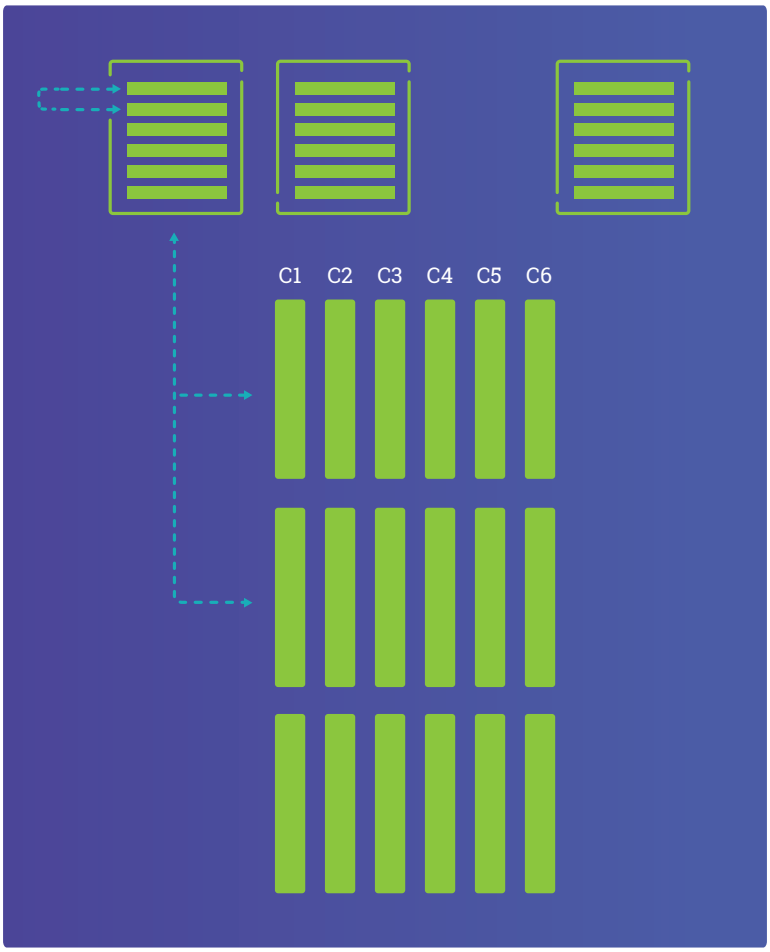


Figure 2 - Column store example (needs to be redrawn)

While this technique is good, many organizations are stuck on older versions of database software that do not support modern features such as columnar data storage. Friction concerning upgrades can happen for many reasons, such as: licensing, organizational standards, or dependence on third-party software that does not offer support. Being stuck on legacy database software means that a great deal of time is spent troubleshooting performance rather than writing code that delivers business value. Further, smaller organizations may lack the expertise necessary to perform appreciable performance tuning on these systems.

Tuning data warehouses involves a tedious process of capturing queries, evaluating execution plans, and gradually implementing improvements through a testing cycle. It can take several days, and often several weeks, to resolve a particularly problematic performance issue. It is important not to underestimate the resources necessary to meet these needs, and automation tools should be considered to speed things along. Automation is an inevitable part of the process for companies that want to deliver successful IT projects in the post-cloud era.

Another challenge for many organizations is a lack of knowledge around newer solutions. Their business may be in a position to take advantage of emerging techniques such as machine learning and predictive analytics, but they lack the organizational knowledge to deploy these types of systems. Or in a large IT organization, these systems may be "off menu" items that would require extra time to configure and deploy.

How Does Panoply.io Address These Problems?

The first thing to consider is the procurement and deployment problem. By delivering a solution based fully in the cloud, Panoply.io eliminates the costly lead time and delays associated with deploying hardware and lets you quickly get to the true value of your data. Instead of the months of lead time followed by months of development data, you can have your solution up in minutes.

By fully committing to this cloud-based model, Panoply.io eliminates some of the previously described problems, as well as creates some heretofore unknown advantages. Let's look at a few of the advantages working with Panoply.io has over data warehousing solutions of yesteryear.

Performance

As opposed to the manual query tuning process mentioned above, Panoply.io captures metrics on all of your query runs. This information is fed to a self-tuning process that automatically optimizes your data and index structures based on your query patterns and workloads. The tuning goes as far as implementing techniques like partitioning (splitting a table into several smaller sub-tables in a manner that is invisible to users and queries), the implementation of which required a great deal of manual effort to complete in the past.

By leveraging machine learning, Panoply.io analyzes all of your queries and will try to optimize them by transparently rewriting them using a more optimal query. Optimization can include changing join methods or reducing implicit conversions that may consume more compute resources than necessary. By performing these optimizations, Panoply.io does not just improve the performance of a single query but improves the overall throughput of the system by eliminating overhead and bottlenecks in the data pipeline. This feature eliminates the time-consuming, tedious process of troubleshooting query performance and allows you to focus on deriving business value instead. Panoply.io will notify you if there are any queries you should reconsider or that could be further optimized.

Panoply.io stores data in a columnar format, which is optimized for reading and loading of data in a multi-tiered fashion. This design allows for optimal performance without sacrificing on cost. Most organizations want to store more data than they can query at any given time. In practice, this means is that for economic reasons, the bulk of your data resides in a Hadoop/S3 store for archiving and backup/recovery purposes. Only the hot data (data frequently accessed by SQL queries) is stored in a fully managed Amazon Redshift data warehouse, which is optimized to deliver the best performance for frequent queries. The final tier in this solution is a small set of data that is stored in Elasticsearch to support small and fast queries. The Elasticsearch component acts as a results cache for those queries. While this architecture would be extremely complex to implement in an on-premises environment, Panoply.io abstracts it behind a single JDBC endpoint that you can use to query seamlessly.

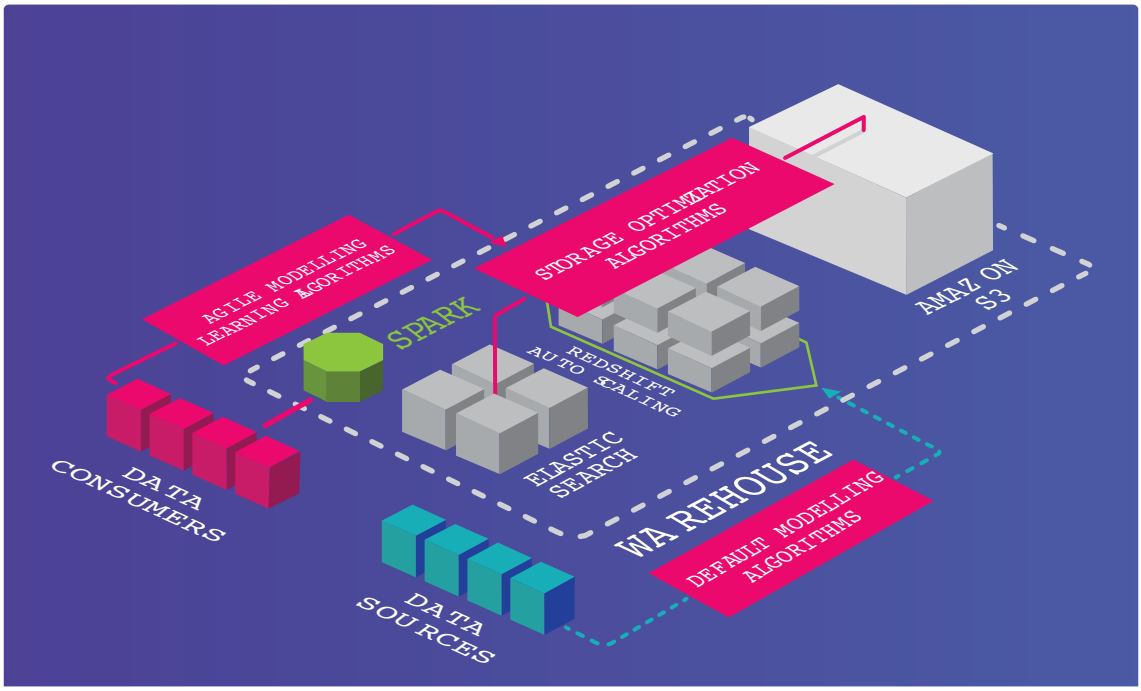


Figure 3 - Panoply.io data lake warehouse architecture
(needs to be redrawn and typos corrected)

Remodeling

To maximize the realization of the benefits from the aforementioned performance optimizations, Panoply rebuilds your indexes whenever it detects changes in your query patterns. These rebuild actions are kicked off by statistical analysis of your queries and data. Additionally, a separate task which redistributes data across nodes takes place asynchronously, to provide better data locality and therefore better performance. Since moving data is expensive and has a higher potential for negative impact, the redistribution algorithm is much more conservative than the reindexing algorithm. In a traditional system, these processes must both be arranged by the database administrator in conjunction with the business team. Panoply.io removes this burden entirely by automating the process altogether.

Self-Service

For many years, the Holy Grail of business intelligence solutions has been the concept of self-service BI. Tools like Power BI and Tableau have gone a long way toward making this possible. By offering a user-friendly interface from which to access the data, these tools allow business users to construct charts and dashboards using their own knowledge and vision of the data they intend to review. Panoply.io takes this to the next level by further abstracting all of the data from a myriad of source systems to provide a single data interface where users can connect all of their business intelligence tools.

Backups

Backup and recovery are an essential part of any data solu-

tion—you want to protect the investments you have made in your data against hardware failure or user error. Panoply.io leverages AWS's backup infrastructure to back up all of your data across two Availability Zones on different continents. The system takes incremental backups of data whenever changes are made, and full backups run periodically. These backups are not simple snapshots; you have the ability to restore to any point in time and debug any changes in data. You also have direct access to your backups, allowing you to write your own data analysis scripts that run against them or load them to any internal database.

Aggregations

Another part of legacy data warehouse projects is building an aggregation model. Typically, aggregation is accomplished using an OLAP cube, which allows users to query the data warehouse in a more ad-hoc fashion. Users can slice and dice data based on key values and filters. Building this OLAP model requires additional development time, and OLAP queries are batch-processed daily, meaning that the business may be looking at day-old data at times. Panoply.io automates this process by analyzing your metadata and data sources to identify logical entities and build key aggregations automatically.

You also have the ability to extend this functionality by building transformation views, which are instantiated views with programmable, user-defined functions. By building these, you can customize your warehouse to meet all of your business needs and eliminate the tedious process of getting to a starting point where you can use your data.

Machine Learning

Machine learning is a field of computer science that uses math to identify patterns and train computers to act without being explicitly programmed. This sort of training is used a number of ways internally within Panoply.io—optimization of your queries and hardware architecture happens by collecting data and self-training on it. These techniques are similar to the pattern identification of data types, which allows for a high degree of automation in the data warehouse stack. This transformative power automates a large portion of a process that used to require substantial manual effort.

Additionally, you may wish to leverage the power of machine learning with your own data sets. Panoply.io can serve as a back end for this in order to take advantage of machine learning tools developed in open source projects like Apache Spark and Mahout.

3

Using the Cloud to Maximize Return on Investment

In a traditional IT environment, corporations spent 70% of their resources on keeping the lights on. Operations consisted of a wide variety of expenses—pure hardware lifecycle issues like replacing servers every three years, adding storage (you'll learn more about the cost of storage later), network equipment, and data center power and cooling. For an organization that wants to protect its data, this means multiplying these costs by two or three times to provide redundancy across multiple data centers. In many cases, it also means having staff in each of those locations to support that hardware.

There's a significant financial outlay before even considering expensive support contracts on both hardware and software—storage arrays come with expensive support contracts, as does relational database management software. Many large organizations pay tens of millions of dollars in support costs annually.

Cloud computing changes some of these paradigms—in most cases, you can "rent" hardware and software, transferring what was traditionally a capital expense to a more palatable operating expense. In the next section, you will learn about how that can be beneficial to your business.

Capital Expense Versus Operating Expense

While this is a technical book, it helps to review some basic accounting to paint a clearer picture of this on-premises and cloud discussion. A capital expense, or CAPEX, is an accounting term that refers to physical assets that the business amortizes over several years. What this means is from a tax perspective, your business cannot deduct the full cost of these assets in the fiscal year they are acquired. For example, if you build a data center and fill it with storage, servers, and network gear, you can deduct only a percentage of that cost each year—this is called amortization. This is because your company is expected to derive value from the asset over several years.

CapEx spending means different things to different companies—in the case of small start-ups, it is just not practical to have the amount of capital on hand to make such a major investment. Before cloud computing, it was tough to stand up the required amount of infrastructure to host their solutions. For larger public companies, increasing capital expense means impacting budgets, return on investment, and potentially stock price. These are all things that are of primary im-

portance to the CFOs of both companies.

The other side of organizational expenses is operational expenses, or OpEx. These are a little bit easier to understand and qualify. Unlike CapEx, which a business amortizes over multiple years, operational businesses expense hit the books in the year of the cost and, for taxable entities, create a savings in the form of expense deduction. Your company derives the value from the expense immediately. Some simple examples of this are travel expenses, or as we are talking about here, cloud computing services.

There is no hard-and-fast rule about what should be OpEx versus what should be CapEx. However, if you think about options involving the two, you can form a better understanding. For a smaller firm, the notion of purchasing all computing services as OpEx provides a number of benefits. It drops the barrier to entry and democratizes a lot of services that were previously available only to much larger companies. For example, in the past, the concept of a massively parallel processing engine like Panoply.io required the purchase of a dedicated appliance from a vendor like Oracle, Teradata, or Microsoft. At a minimum, this required an investment of hundreds of thousands of dollars of hardware, with a multiplier for consulting fees for implementation, and software costs on top of that. Above that is the added cost of annual support on both the hardware and software, which on a system like that alone can add up to hundreds of thousands of dollars.

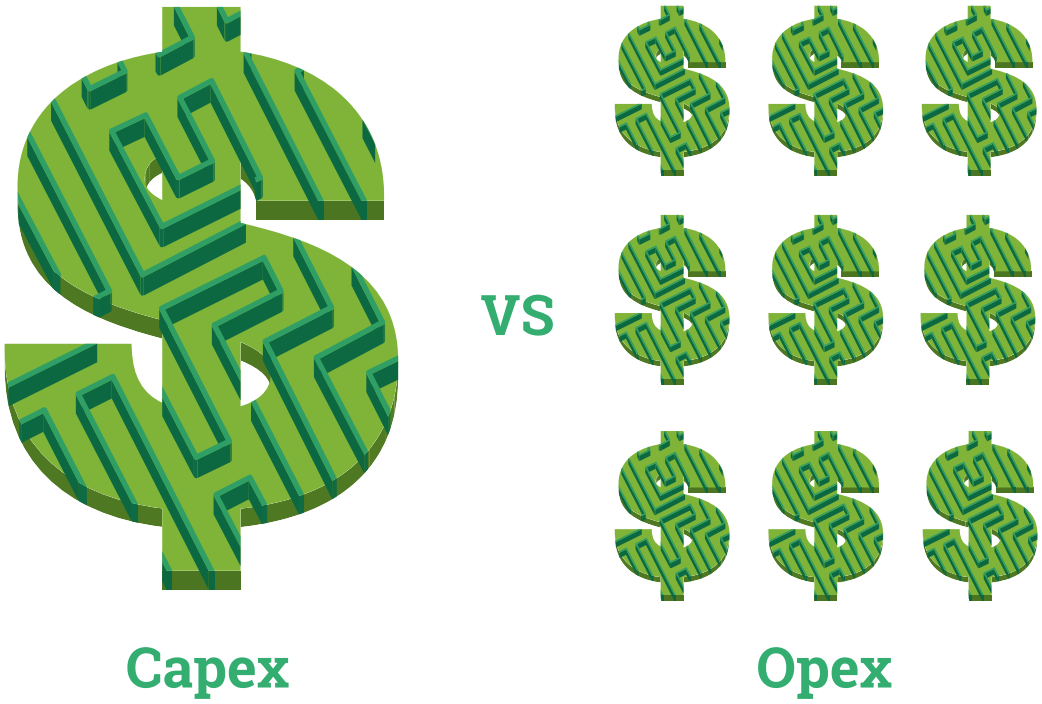


Figure 4. CapEx (smaller amount of big, one-time expenses) versus OpEx (many smaller, sometimes recurring expenses)

Another example of technology that was democratized by cloud computing is the ability to have a second data center for disaster recovery (DR). While some organizations can tolerate some degree of downtime without impacting their business, in our increasingly connected world, many applications need to be able to survive the loss of a primary data center while mostly maintaining uptime. Before cloud computing, this meant either renting expensive co-location space or building and managing a second data center in a second geography. Now a company can cheaply provide DR using cloud resources and scaling them as needed in the event of a disaster.

Another way to look at this CapEx and OpEx discussion is to think about a department within a large company. The department may wish to do a data analytics project; however,

the associated CapEx costs will require executive approval and may need to pass through several budget cycles. Additionally, as mentioned earlier, the lead time associated with a major IT project could double the time it takes to get the project off the ground. With a cloud solution, if the smaller amount of funding required for the pay-as-you-go solution is in the departmental budget, the department can fund the project internally with a much shorter approval cycle. Once the connectivity to the cloud provider comes online, and an administrator makes the initial data source connections, the department can start to receive value from the project in very short order.

Cloud Cost Savings

The other benefit of moving workloads to the cloud is reducing the IT staff needed to manage systems or reallocating those resources to higher value tasks. A smaller organization might be able to get by with having only a couple of IT resources to work with the cloud provider and manage the connectivity to the cloud and the resources consumed in the cloud. In a larger organization, IT staff can be moved from jobs where they are merely responsible for keeping servers up and running and into higher value tasks like data management and helping business units derive value from their use of IT systems.

A larger organization would need to move the majority of its resources into the cloud to capture the benefits mentioned above. Moving to the cloud is a non-trivial strategy transi-

tion that may take many years to accomplish. However, for a smaller company, this shift is revolutionary and game changing. For a small and fast-growing company, the ability to have all of its human resources focused on mission-centric work rather than maintenance activity such as managing the upkeep of systems can be a huge benefit.

The Hidden Cost of Data Inefficiency

In a data-driven world, one of a company's biggest assets is its data. The ability to gather metrics on customers, supply chain, and marketing campaigns provides the ability to make better decisions throughout the company, and support more informed strategic decisions. In larger firms, the silos created between departments may pose a challenge to bringing all of this data together in one place to perform an analysis. In a smaller company, the source systems may be a mix of software-as-a-service (SaaS) products and traditional software that the firm may lack the technical resources to bring together.

There are very real and costly ramifications of this sort of inefficiency. At some organizations, limitations such as those just described keep them from reporting their monthly sales until 30 days after month end. In that case, the organization would be unable to report its financial data promptly to companies that were interested in them for acquisition, for example. Other companies may have useful data in place but may take days or months to get it into their target systems, which can impact their agility and flexibility in decision-making. The sooner you have higher quality information, the faster you can make critical business decisions.

Storage Optimization

One of the largest costs in any large IT organizations is storage. This is confusing to many people who aren't deeply involved in IT infrastructure and may go into a retail electronics store and see a 4-terabyte hard drive for less than $200. However, enterprise-class storage, the kind that supports analytics systems, comes with a hefty price tag. In some cases, it can be up to $3000 per terabyte. Why the dramatic difference in pricing? There are a number of reasons. For one, enterprise storage needs several layers of redundancy for protection in the event of hard drive or disk controller failures. Additionally, the storage in most larger organizations is network connected and requires the use of dedicated switches and expensive fiber optic cable to connect it to the servers it supports. These specialized storage devices require dedicated staff to maintain and support the storage array.

The presence of cloud computing changes the paradigm of storage. While in some areas of cloud computing (particularly IaaS virtual machines) the cost savings of using the cloud are debatable, storage is not one of them. Storage is markedly cheaper in the cloud, due to both economies of scale and engineering improvements from the cloud providers.

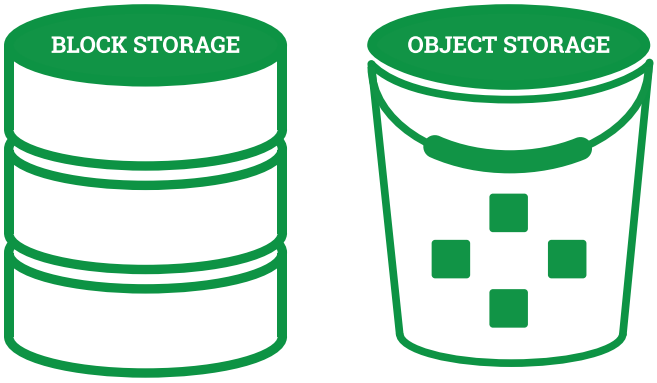


Figure 5. Block versus object-based storage

Cloud providers can offer impressive density at a low cost using a storage technique known as object-based storage. They reap additional efficiency because they are writing their own management software which is devoid of all the support and extra costs of legacy enterprise storage vendors. The object-based storage model also offers flexibility and allows cloud providers to simplify deployment of storage and build their own redundancy model without having to rely on a third party.

Object Storage versus File and Block

File and block storage are well understood by most IT practitioners, especially those who deal with data. However, object storage is a new subject for many. The primary differentiator is that object storage neatly bundles up the data and associated metadata and stores it with a unique identifier. As opposed to traditional enterprise storage which often consumes proprietary hardware and uses RAID to protect data, object stores often consume commodity software and use object-level replication or an erasure coding algorithm to protect data.

Why Use Object Storage?

The nature of object storage affords the following advantages over file and block:

- Low cost due to commodity hardware
- Enhanced durability due to erasure coding scheme
- Freedom from the limitations imposed by RAID

Automated Backup

Backups have come a long way since the days of the LTO tape being shipped to an off-site provider. Even though technologies have changed, the costs associated with storing backups have not changed that much. Backups are just additional data that needs to be stored, powered, and cooled. Administrators are also needed to ensure the success of backup jobs and perform storage capacity management.

One of the benefits of cloud-based storage is its nearly infinite capacity. A significant upside to using a platform as a service (PaaS) solution like Panoply.io that includes automated backups is the elimination of the need for the administrator to worry about backups, or the need for a third-party provider to provide off-site redundancy for the backup media. The redundancy is built into the cloud platform, and Panoply.io ensures that the data is backed up at regular intervals. There is no volume to run out of space; there's no tape library to fail—the backups just happen and are always there. Solid backups help business leaders sleep better at night while also reducing overall infrastructure costs.

Developer Efficiency

Given the very manual and waterfall-based nature of ETL processing, it accounts for the largest portion of most traditional data analytics projects. Gathering, cleaning, and transforming data can take up to 80% of developer effort on a project. Unfortunately, this part of the project isn't even something that adds any business value; it is merely a part of the work required to start building any analytics system.

Imagine if, from Day One of your analytics project, your developers were focused only on building the best analytics queries and algorithms, and they had free time to focus on determining how to answer the next questions your business will have rather than merely cleaning and shredding data and worrying about a character change breaking that day's data load. With the automated data transformation that Panoply.io offers, that dream becomes a reality.

Ease of Data Management

A final area where Panoply.io lowers your TCO is the ease of access to your data. Once your data is loaded into your warehouse, it is time to use it. Panoply exposes a standard SQL endpoint that supports JDBC and standardized ANSI-SQL. Your users and analysts can plug in any tool they want—Tableau, Spark, R, or any other standard data analysis tool. Panoply.io also delivers a set of extensible cloud-based analysis tools, which can lower your costs even more by allowing the consolidation of tools. This cloud analysis layer is completely open-source.

Other features included are the data management layer, which is a metadata editing tool for your data. Using this included tool from Panoply.io, you can identify your most popular queries and tune them based on your knowledge of the data. You can view and modify tables and data relationships such as primary and foreign keys. Additionally, if you want more control of your data and tuning, you can turn off Panoply.io's automatic data jobs or adjust them to align their timing to your business needs better.

Summary

Moving into a PaaS solution from a traditional on-premises solution can certainly be a leap of faith, as it requires giving up some measure of control. However, a powerful and robust platform like Panoply.io can provide higher velocity to insight, allowing your business to make better-informed decisions and improve growth opportunities. Additionally, your IT staff can be more productive because they will have a quicker time to results and less manual effort getting data into source systems.

Panoply.io provides end-to-end data management-as-a-service. Its unique self-optimizing architecture utilizes machine learning and natural language processing (NLP) to model and streamline the data journey from source to analysis, reducing the time from data to value as close as possible to none.

Panoply.io 1161 Mission St. • San Francisco, CA 94103 • hello@panoply.io