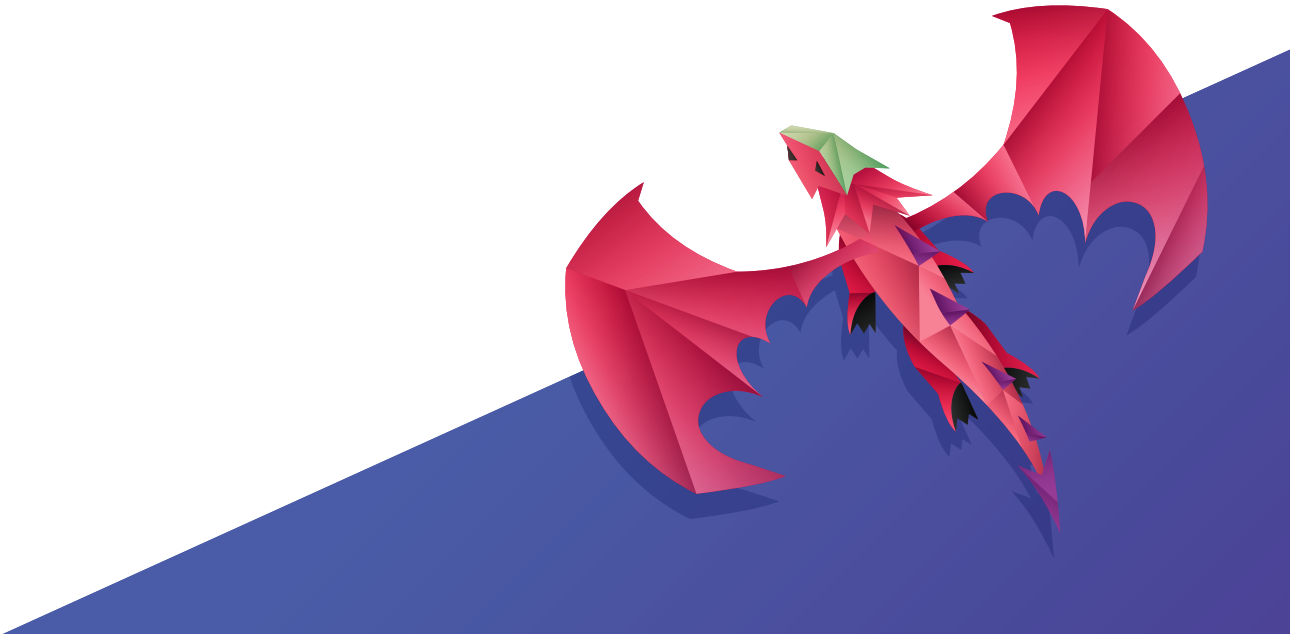
panoply.io

Leveraging AWS to Overcome Complexity in Data Warehousing

Table of Contents



The Data Warehouse

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. 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**

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 challenging 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 a few more components to be considered in the big picture of an analytics strategy:

• The warehouse itself, which handles batches of queries

• Optionally, an online analytical processing (OLAP) cube, which supports the use of ad-hoc queries without overtaxing the warehouse

• 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.

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.

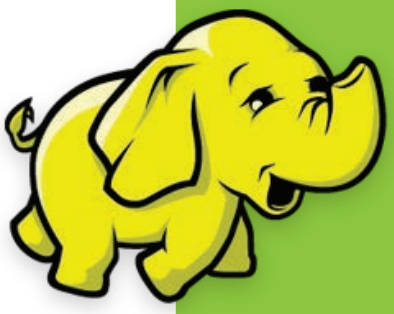
The pace of change in the data center today isn't making these systems any easier to manage. Let's look at some of the recent and current trends impacting computing, especially with respect to big data.

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. 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.



ABOUT 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 in- creased performance over the serial processing methods used historically. There are multiple commercial distributions of Hadoop such as those.

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 in- crease 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

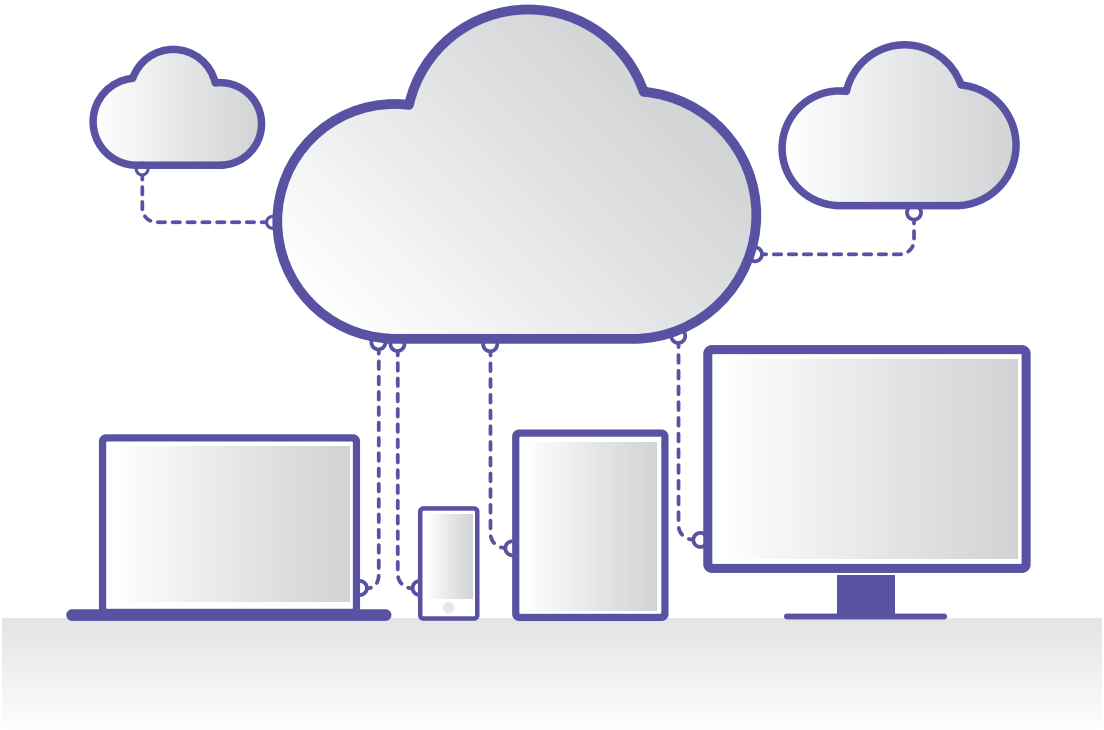
Cloud computing is the most impactful innovation in computing in a generation. What started with the outsourcing of basic computing tasks such as e-mail has evolved into organizations' having far more options regarding where and how to run workloads. In some cases, companies have outsourced all their 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 upfront 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 some 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 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

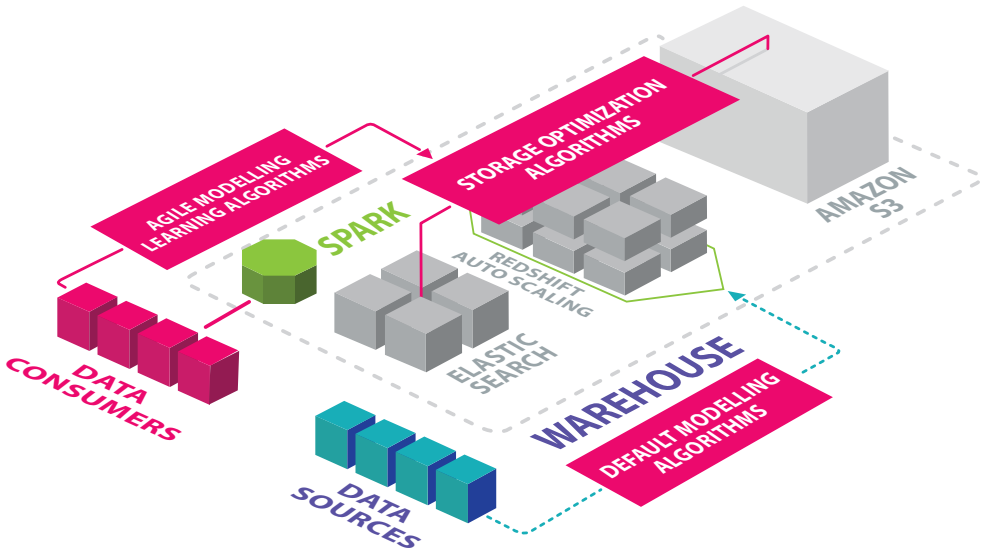
The primary difference bet- ween a data warehouse of old and a data lake is the position of data processing in the overall data pipe- line. In a traditional data warehouse, the ETL pro- cess described earlier cleans and structures the data upon ingest. In contrast, a data lake stores raw data in its native format until it's needed.

Introducing Panoply.io

If you think that all of this sounds rather daunting, you're 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 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 automated 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?

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).

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.

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 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 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 ecosystems. 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.

LEARN MORE ABOUT PANOPLY.IO

Panoply.io is a cloud-based data management platform for analytics that streamlines time and value for data engineers, scientists and analysts – automating the full data stack without the overhead of preparing and modeling data, or managing infrastructure – cutting down development time by 80%. To learn more about this data stack and about Panoply.io in general, visit **https://panoply.io**

Author: James Green, vExpert

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