aws | panoply

Modern Data Management

Next Generation Data Tools
Eliminate Data Maintenance

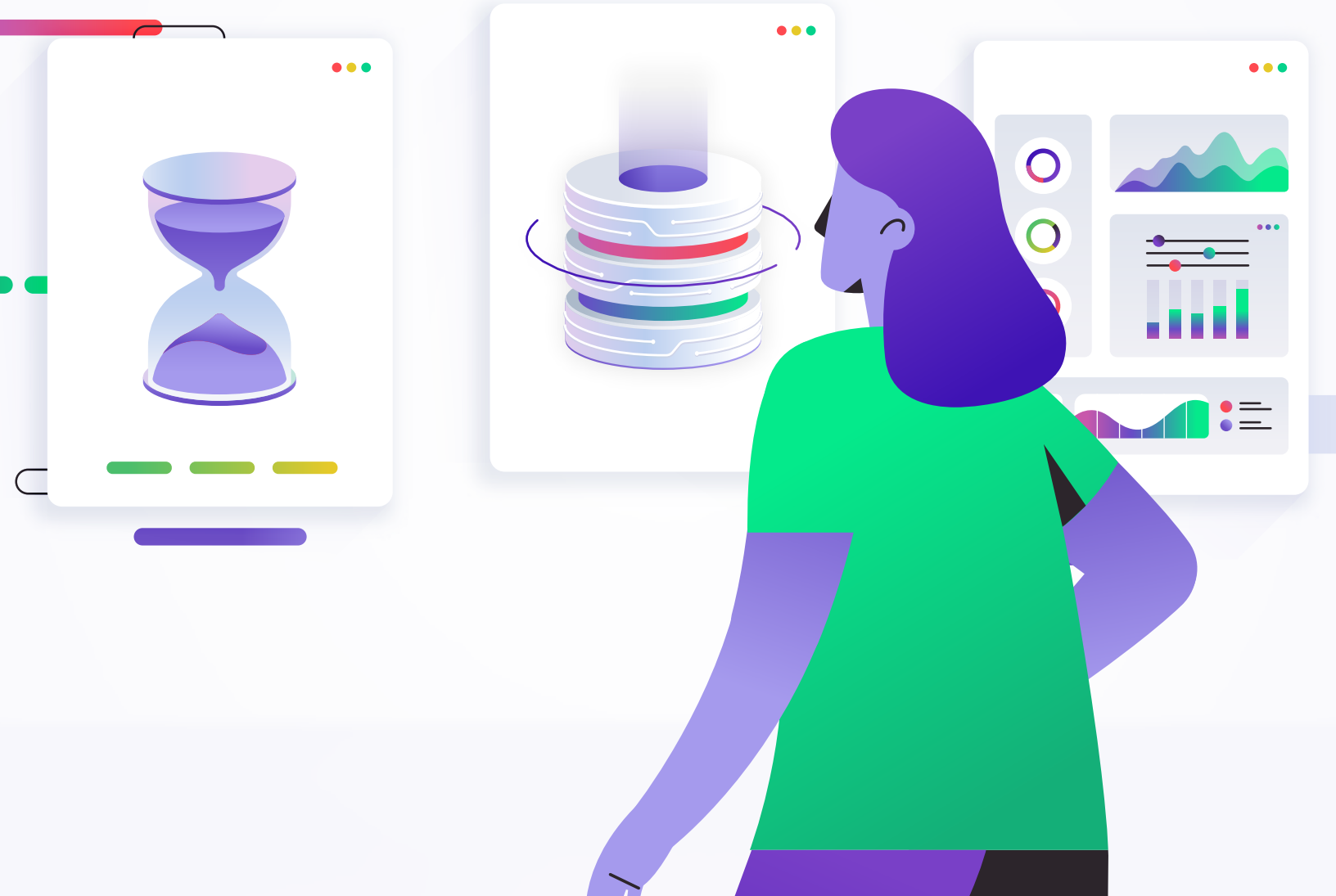


Table of Contents

Introduction

As more companies shift toward integrating data into every level of their business operations, the volume of data handled even by small companies has exploded. Data on customer behavior, market characteristics, product inventories and more can all be tracked in real time to provide critical business insights. Practically, this has the effect of making companies incredibly agile, able to identify signals in their data as they come in, and make quick decisions about how to respond.

But the benefits of having large amounts of data relating to every aspect of your business come with new challenges, chiefly in the form of data management and access.

As companies increase the size and complexity of their data operations, approaches that worked at a small scale introduce friction, slowing the time from data collection to actionable insights.

This sort of friction can be reduced with proper data management approaches that help organize, integrate and manage access to data for all the relevant stakeholders.

An effective data management approach will make it easy to scale your data operation with your business and share valuable insights across the organization, from IT to analysts to executives. For organizations trying to get a handle on their data, building an effective data management operation starts with setting up a "data stack", a collection of processes and applications that will automate the most cumbersome aspects of data management.

Today's Data Management Workflow Lacks "Flow"

Most businesses' data operations run on a collection of apps: Microsoft Excel or Google Sheets for collecting, organizing and visualizing data, Dropbox or Sharepoint for storing and sharing data, applications databases, and third-party data sources like Google Analytics, Salesforce or Shopify for tracking specific types of customer and commercial data. This is great for business, as it allows the end user to jump in with a familiar app and start working with their data right away. But spreadsheets (Excel-based or not) are an inefficient method for storage and manipulation of data, especially at scale or over the long term. They can be difficult to share and collaborate on, and provide few effective solutions for version control and master recordkeeping. Likewise, without fanatical attention to organizational practices, storage solutions like Dropbox or Sharepoint can quickly become tangles of outdated and disorganized folders that make it difficult to get the right data into the right hands at the time it's needed.

Time Spent in a Data Analytics Workflow is not Evenly Distributed Across Processes

Another common problem is that data management itself is far more time consuming than the value-add aspects of business intelligence and data science. Data from multiple surveys of analysts and data scientists indicates that data processing and cleanup makes up the majority of their day-to-day activities, followed by data visualization/ dashboard construction, basic statistics and advanced modeling.

Basic data handling and management was a task that required participation from one half to three quarters of the members of analytics teams, including senior data analysts and scientists.

Gartner

JET BRAINS

Team Members Performing Data Processing Prior to Analysis

In two surveys conducted by Gartner and the software company JetBrains, responses to survey questions about data-related activities showed that basic data handling and management was a task that required participation from one half to three quarters of the members of analytics teams, including senior data analysts and scientists. In practice, this means that a large portion of high-value team members who could be spending their time searching for useful signals in a company's data and passing those insights to other stakeholders are instead stuck dealing with simple data management tasks. With the right mix of strategy and automation, many of these data management tasks can be eliminated, freeing up time for uncovering actionable insights.

Other survey responses give further indication that data wrangling (collection, cleanup, loading, etc) remains the top activity of most data analysts and data scientists in terms of time spent. This shows that tools that make this process easier or more efficient should be extremely valuable to people working in the data analytics community, and for organizations seeking to make use of data generally.

Additionally, survey data indicate that, despite the hype around machine learning and AI, the majority of respondents are not using advanced analytics in their data practices, but are instead generally spending their time building overviews of their data in easily

digestible visualizations and dashboards. It's also telling that Excel remains in the top 5 tools used by respondents to KDNuggets's 2018 data analytics, data science and machine learning tool use survey, which showed that Excel use actually grew from 2017-2018, despite the continuing dominance of more advanced approaches like Python and SQL as data scientists' languages of choice.

It stands to reason, then, that developing effective data management infrastructure should be the top priority for any organization seeking to capitalize on the potentially vast amounts of data available from its customers and the wider market. Building a strong data management foundation will let the gains in productivity flow upward toward analysis and data-based decision-making and increase data democratization. This will make it possible to have key data available at all levels of the organization, from analysts to executives.

Poor Data Management Infrastructure is a Barrier to Success

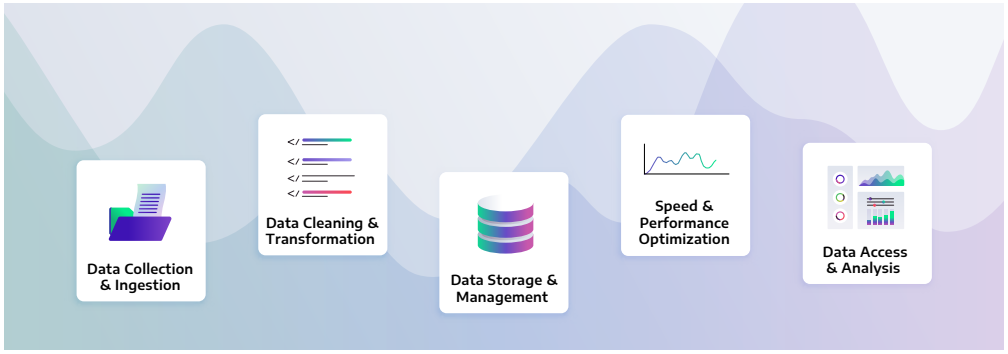
Traditional approaches to data management have involved on-premises deployments of expensive server hardware that is difficult to maintain and often requires extensive in-house IT operations to manage it. Moreover, on-premises data management and data warehousing operations, historically speaking, have seen high rates of failure—and even when they are successfully implemented, they are slow to prove their value to the wider organization.

More recently, the rise of cloud computing services like Amazon Web Services, Google Cloud Platform and others have taken on much of the burden of setup, deployment and server maintenance and freed up significant resources for their customer companies, but even with the increased success rates associated with the move from on-premises deployments to cloud-based offerings, there remains room for improvement. In surveys of data warehouse users conducted by Panoply, over 60% of respondents using the biggest cloud warehouse vendors (Amazon Redshift, Google BigQuery and MS Azure SQL Server) still rated their data warehouse solution as "difficult" or "very difficult" to use. Asked why these technologies were so difficult to use, on average, 47% of respondents pointed to the complexity of the interfaces and setups involved.

In essence, cloud computing platforms have made setting up a data management workflow much easier than it has ever been, but interacting with the most common cloud-based platforms still requires specialized knowledge and often remains the domain of the IT department. This has the effect of maintaining an inefficient separation of data operations from decision-making operations and slowing the time from data collection to the generation of actionable insights.

Fortunately, the next phase of development in the data management space is already underway. The rise of cloud computing has made it easier for companies with a high degree of technical skill to offer SaaS data management solutions that offload even more of the difficult work of server and database maintenance, producing true self-service data management and business intelligence tools that can be combined to create highly effective, agile and resilient data stacks.

Data Stacks: An Overview



In order to set up a functional data operation, an organization will usually need to combine several services into a data stack, as shown above. Fundamentally, an effective data stack will make it possible to perform five basic operations:

Data Collection & Ingestion: Many organizations will have to develop their own methods and approaches for collecting relevant data, but there are several types of data that can be ingested from existing vendors. Data from payment processors, ad platforms, CRMs, CMSs, ecommerce platforms, web and mobile analytics tools, and social media sources can all be gathered at this stage and combined with an organization's internally-collected data.

Data Cleaning & Transformation: Data often comes into a pipeline in need of editing, formatting, combining or rearrangement before it can be useful. This could be something as simple as combining a "first name" and "last name" field into one, or standardizing the date format in a collection of records. The end goal of these transformations is always to make the data ready to be passed to the next level of the data stack.

Data Storage & Management: Once ingested, data needs to be stored in a place that is accessible for analyses. A sound strategy at this level of the stack is critical. Without one, organizations can find themselves facing a whole new set of problems. Manual management of databases—or clusters of databases—requires a new suite of technical skills and introduces numerous decisions which heavily impact cost. Tools that can manage these processes automatically are key components of the modern data stack.

Speed & Performance Optimization: With the high availability of data on potentially every aspect of a company's business, it's not enough simply to collect, clean and store data. As datasets grow, further considerations need to be made—can the data be further formatted to optimize the space it takes up, or the time it takes to query it? Tools that automate these processes are what allow businesses to move at the speed of their data.

Data Access & Analysis: Data analytics is the top of the data pyramid, the operation that every other part of the data stack is designed to support. The term as used here encompasses a range of approaches at varying levels of complexity, from familiar business intelligence approaches such as dashboard construction and chart-making to more complex, machine learning-based discriminators, recommenders and predictors.

Any stack that can make these five basic operations possible has the potential to generate useful insights for an analytics-focused organization. This conceptual framework should be helpful as we move into a discussion of the specific types of tools available to cover the basic operations of a data stack. Below, we dig deeper into ingestion, transformation, storage, and data access/analytics.

Data Ingestion, Cleaning, and Transformation Tools

Traditionally, data transformation has been tightly coupled with data transfer due to technological constraints. Today, cloud computing has freed up the transformation process, allowing for greater flexibility and control in data processing. It is important to understand the nuances between traditional ETL and modern ELT.

ETL – Traditional Data Transformation: In general, the related operations of data ingestion, cleaning and transformation are handled by ETL tools, where "ETL" stands for Extract, Transform and Load. An effective ETL tool will manage the extraction of data from various sources, whether they're documents full of unstructured data, spreadsheets or staging databases, transform the data for later use, and load them into a final destination database. However, the traditional ETL approach was developed to function in an environment of more constrained data storage and data availability. Modern data analytics operations often have to process large amounts of data as quickly as possible, and the traditional ETL framework can slow this process down.

ELT – Shorter Time from Extraction to Insight: A faster approach for data ingestion and processing in data analytics is E-L-T, where the transformation process happens after the data is extracted and loaded into a centralized data repository. This eliminates potentially long wait times between extraction and transformation, and allows analysts to get their hands on raw data as soon as it's loaded into the data repository. Many modern data analytics operations have begun to shift over to an ELT framework in order to increase the agility and flexibility of their data pipelines.

Read more: ETL vs ELT and ETL Tools

Data Storage Tools

As the amount of analyzable data in most organizations has grown, the available options for data storage have proliferated. At the most basic level, all data storage will happen in a database of some sort that can manage the 4 basic "**CRUD**" operations: Create, Read, Update and Delete. Databases come in many flavors, but SQL-based relational databases have maintained a longstanding dominance in the field, and for good reason: the codebase is solid, and the resulting databases are robust and relatively easy to configure.

A common flow for a data analytics operation will involve collecting data from different sources, storing them in intermediary databases for further processing or local analysis, and then pulling data from multiple databases into a central repository for analyses that will integrate multiple data sources. These central repositories can take several

forms, but the most common types are referred to as **data warehouses** and **data lakes**. The differences between these two styles of data repository are outlined below.

Data Warehouse: Data warehouses are stores of structured, curated data pulled from separate intermediary databases that make it easy for analysts from different parts of an organization to access and analyze for their respective purposes. Data warehouses are most useful for analysts doing regular, directed projects for specific business purposes.

Data Lake: A data lake also pulls in data from multiple sources, but generally contains a much wider array of data types, including unstructured data and data from sources outside the organization. Where data warehouses are designed to be a central repository for data for known and specific purposes, data lakes are intended to contain data that might not be useful at the moment but could play into some potential future analysis. With the price of storage continuing to decrease over time, data lake-style approaches have become more economically feasible, but they are generally most useful for more exploratory analyses undertaken by data scientists and researchers.

Read more: Data Warehouse vs Data Lake

Data Analytics Tools

"Data analytics" covers a wide range of activities of varying degrees of complexity, but the landscape of data analytics tools can be subdivided into two main categories: Business intelligence tools and advanced analytics. Business intelligence-type activities will likely make up the majority of an analytics operation, while advanced analytics approaches tend to be deployed somewhat less frequently in order to answer experimental questions or identify trends in high-dimensional data that would be exceedingly difficult for human analysts on their own.

Business Intelligence: The types of questions answered by business intelligence-style analytics are often what's been referred to as the "dark matter" of analytics—the necessary, but relatively simple, questions that will need to be answered the majority of the time: current inventory, number of customers, incoming/outgoing payments, average number of purchases per customer, etc. The data that answers these types of questions can usually be presented in dashboards, simple data plots and reports.

Advanced Analytics: Advanced analytics uses more complex statistical techniques, machine learning and potentially huge datasets to generate predictions based on current data and identify key performance indicators. Models of future consumer behavior based on past data, fraud detection and recommender systems are all examples of advanced analytics applications.

Read more: BI Tools and Analytics Tools

Choosing the Right Tools

Choosing the right tools for the lower levels of a data stack is one of the most crucial decisions organizations face as they seek to set up a functional data analytics operation. Inefficiencies at the data ingestion, storage and retrieval stages will propagate upward in the stack, slowing the generation of actionable business insights at best, and dramatically reducing their value at worst. In short, the choice of data collection and storage tools can potentially make or break a data analytics operation. That being said, the decision process will likely still be confusing for those unfamiliar with the data and analytics landscape, but thinking in terms of key factors can help to structure the decision making process.

Key Factors to Consider in Building a Data Stack

Speed of Setup: The faster a potential tool can be evaluated for utility, the easier it will be to build an effective data stack. A major limiting factor in evaluating that utility will often be at the point of setup—the longer it takes to setup and install, the longer it will take to evaluate, and hence more time lost if it turns out not to be a useful addition to the data stack.

Ease-of-Use: For your organization to effectively democratize its data operations to every individual across departments or disciplines, your tools must be accessible. Note that this doesn't necessarily mean choosing graphical interfaces over code-based approaches, though that can help some organizations. More important is whether the tools are easy to learn and maintain.

Integrations: Whether a product integrates easily with other data services and products is another important consideration. A tool that needs special knowledge or bespoke code in order to work with another component of the data stack will limit the stack's overall flexibility. Worse still, a tool that can only work with components from the same vendor can force an organization into a single product ecosystem rather than allowing you to select the best fit for each layer of the stack.

Scalability: Your organization's data stack needs to be able to grow with the business and the amount of information it's bringing in. Being able to scale a data operation without burning engineering and IT hours or losing time on long migrations, or service interruptions is a major consideration for a growing business.

Features that Fit: A data management solution may claim to offer an advanced AI that can predict customer behavior 10 years into the future, but most day-to-day business analytics will be far less complicated. Choosing a tool that allows your organization to generate the insights it will need the majority of the time is a far better allocation of resources. Furthermore, because more layers can often be added on to a data stack, they can be swapped out as an organization's needs evolve toward something more complex.

Maintenance Automation: A data management solution that requires extensive, hands-on maintenance by your IT team is likely not a great fit for a modern data management stack. Tools that manage day-to-day maintenance tasks under the hood increase the stability and flexibility of an organization's data management operation and free up personnel resources for more important tasks in a growing organization.

Self-Service: Ultimately, an effectively organized, data-oriented organization should seek to make data and analytics available to every level of the company. What this means in practice is that the data-surfacing components of a company's stack should be just as accessible to business users as engineers. Those trying to build truly self-service data stacks should seek to include tools with intuitive, straightforward user interfaces.

Pricing: The costs of setting up a data stack can vary wildly, depending on the complexity of the operation, but inexpensive solutions are available at almost every level of the stack at this point. The ideal data management solution, especially for young and growing companies, will offer an easy, approachable, all-in-one setup at a low price point that scales predictably with the company's growth. That's not to say that modularity isn't desirable—an effective data stack should still be made up of components that can be swapped in and out to account for growth or changing needs—but bundling some features, especially at the level of data management, can help control costs and reduce feature redundancy across the stack.

When it comes to data management, Panoply makes an excellent addition to an effective, modern data stack. With a large—and always growing—set of pre-built data connectors and readymade integrations for a wide range of business intelligence and data analysis tools, Panoply slots easily into growing data operation.

The Panoply Advantage: End-to-End Data Management Platform

Easy-to-Use

- Get up and running with a cloud data warehouse in minutes, without complex technical configuration
- Collecting data sources is a simple, one-click process using your credentials
- No maintenance—we take care of all ongoing technical maintenance so you don't have to worry about it
- Integrations offer seamless connection to any BI visualization tool, and Panoply is also compatible with standard SQL, R, and Python

All-in-One

- **Data Ingestion –** No need for additional ETL tools or caching layer—Panoply automates data ingestion from 100+ data sources from APIs, files, and databases in just a few clicks
- **Storage Management –** Panoply automatically adapts server configurations to accommodate greater scalability
- **Query Performance Optimization –** Panoply optimizes the querying process by intelligently materializing your queries with machine learning

Self-Service Data Management

- Democratizes analytics for faster business decisions
- Upfront, simple, transparent pricing
- No extra infrastructure or headcount needed
- Allows business leaders and analysts to focus on the products they sell instead of spending time seeking backend connections and database storage

Next Steps

There are a lot of decisions involved in developing a data management program. Up-to-date information can be hard to find and even harder to understand. If you want to learn more about how businesses of all sizes are modernizing their data management programs, feel free to schedule a time to talk with one of our knowledgeable data architects. We can help you understand your current data management needs, what your options are, and how to move forward with confidence.

Panoply was rated the #1 on G2 Crowd's Data Warehouse Implementation Index which accounts for ease-of-setup, implementation time, user adoption, and other factors. Find out how easy it is to get started with a 21 day Free Trial. It only takes a few minutes to integrate all data from 100+ data integrations. You'll be analyzing and visualizing your data in no time!

aws | panoply

Start your free trial or schedule a personalized demo at panoply.io