panoply.io

Conquering Inefficiency in Data Analytics

Table of Contents

Introduction

Big data and analytics are coming to the forefront of enterprise IT requirements today and for a good reason. Insight and decisions based on analysis of the data at an organization's fingertips can be a significant contributor to the bottom line. However, in this new era where previously unthinkable quantities of data have become the norm, organizations certainly face some challenges with the requirements of modern data systems.

Particularly 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

In this paper, you'll learn more about some of the challenges that organizations face with data warehousing deployments and operations today. But there is hope for the future; you'll also learn about the ways Panoply.io shifts the data warehousing paradigm to overcome some of these challenges.

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. 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.

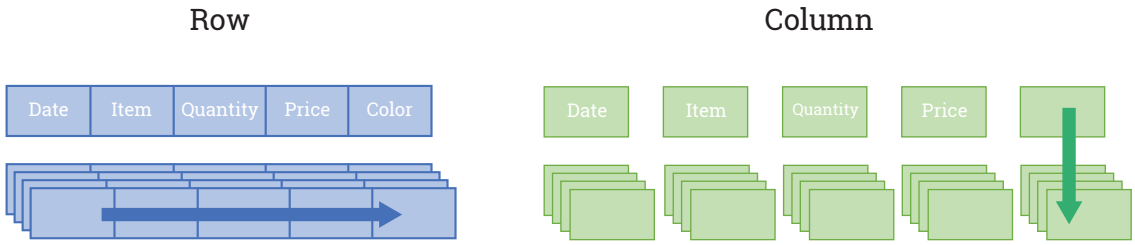


Figure 1 - Row-based storage on the left, columnar storage on the right

While the columnar storage 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.

Conquering Data Warehousing Inefficiency with Panoply.io

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.

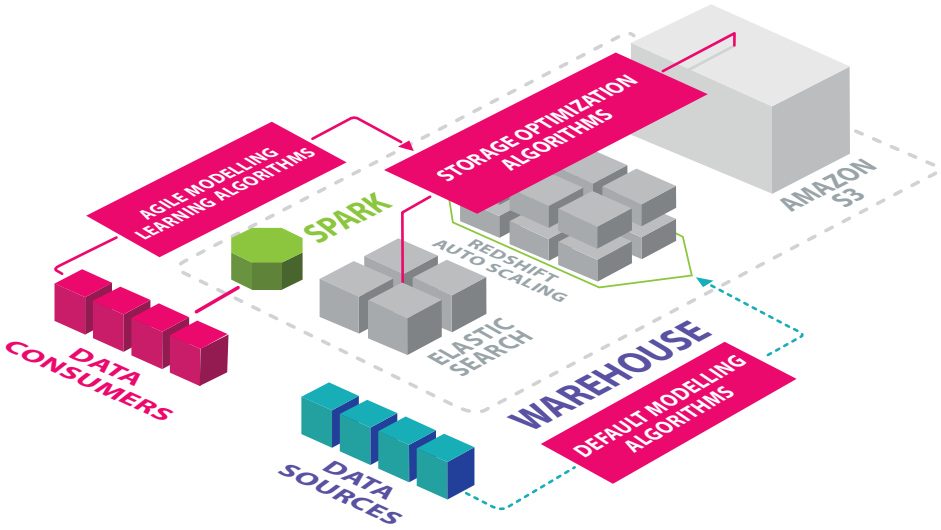
PERFORMANCE

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 (Java database connect) endpoint that you can use to query seamlessly.



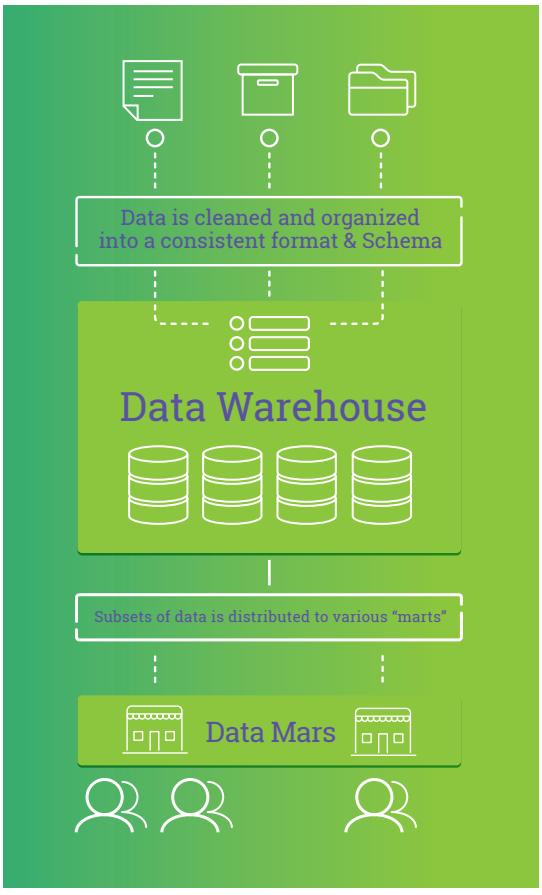
REMODELING

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

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.



Panoply.io data lake warehouse architecture

Backups

Backup and recovery are an essential part of any data solution—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 in 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.

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