panoply.io

Redshift vs BigQuery

Closing the Gaps in Data Warehousing

Table of Contents

Executive Summary

Database architects, analysts, and administrators are constantly seeking more cost-efficient, fast, and reliable ways to collect, store, and best use their organization's data. In doing so, enterprises are turning to cloud services solutions for their data warehousing. A high demand for virtualization and sophisticated analysis is propelling disruption in the market, with innovation and initiatives from cloud industry leaders such as Amazon, Microsoft and Google.

Recently, we have witnessed a great deal of discussion surrounding the best-performing data warehouse technologies in the cloud, with competing claims and rebuttals emerging in the blogosphere. At a private event in September, Google Cloud presented performance benchmarks comparing two data warehouse giants: Google BigQuery and AWS Redshift. Following the event, many questioned the methodology used in the testing and subsequent marketing of Google's claims; the biggest objection being that Google "cherry-picked" a single query out of 22 that generated favorable results for their data warehousing solution.

In response, Amazon Web Services (AWS) came out with a **rebuttal piece** [1] last October titled **"Fact or Fiction: Google BigQuery Outperforms Amazon Redshift as an Enterprise Data Warehouse?**" Amazon performed their own tests (initiated specifically to verify Google's claims), which they contend proved Google BigQuery to be, in fact, the lower-performing of the two solutions.

Parallel to this back and forth between the two giants, Periscope Data posted a **detailed piece** [2] entitled **"Interactive Analytics: Redshift vs Snowflake vs BigQuery,"** comparing performance benchmarks of the three top developers of modern cloud analytics solutions.

[1] https://aws.amazon.com/blogs/big-data/fact-or-fiction-google-big-query-outperforms-amazon-redshift-as-an-enterprise-data-warehouse/

[2] https://www.periscopedata.com/blog/interactive-analytics-redshift-bigquery-snowflake.html

Our Perspective

In July 2016, **we published a full comparison of Redshift vs. BigQuery** [3] detailing our test methodology, the results, and further considerations that led our team to choose AWS Redshift over BigQuery as the core database to power our Panoply.io stack.

As a cloud-based data management platform for analytics that streamlines time and value for data engineers, scientists, and analysts, this discussion and its implications are relevant to our mission and the promise we make to our customers.

In this white paper, we'll review these two data warehousing powerhouses in-depth, and discuss how the cloud brings added value to the market.

[3] http://panoply.io/blog/a-full-comparison-of-redshift-and-bigquery/

Two Cloud Data Warehousing Methodologies

Google BigQuery is an enterprise data warehouse that aims to address the time-consuming and expensive process of storing and querying massive datasets by enabling super-fast SQL queries. Amazon Redshift is a fast, fully-managed, petabyte-scale data warehouse focused on providing a quick, cost-effective way to analyze data using existing business intelligence tools. When comparing the two, it's important to note that BigQuery is a serverless service without the need to consider allocation of compute and storage resources. Amazon Redshift's users, on the other hand, must account for the amount and management of resources their data warehouse clusters require.

Based on their internal tests, Redshift allows for maximum performance through its high level of customizability, enabling users to get the maximum performance from their cluster. In addition, according to Amazon **TPC-H** [4] and **TPC-DS** [5] performance tests, **Redshift outperforms BigQuery by 2-6 times, respectively.**

Although we did not run these particular tests -- we based ours on more of a real-world comparison [**"A Comparison of Approaches to Large-Scale Data Analysis" by Pavlo et al (SIGMOD 2009)**] [6] - our results were similar. Up against BigQuery, Redshift is much better in terms of usability, performance, and cost-effectiveness for the majority of analytical use cases, especially at scale. However, as we highlight below, a major drawback of Redshift is the need for constant low-level tuning, including with VMs and database configurations.

[4] http://www.tpc.org/tpch/

[5] http://www.tpc.org/tpcds/

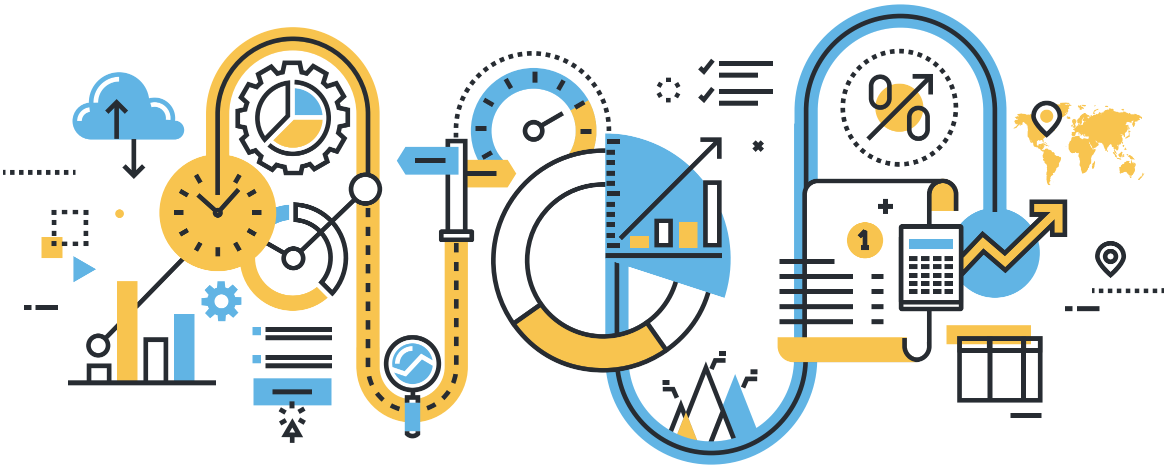
[6] http://database.cs.brown.edu/sigmod09/benchmarks-sigmod09.pdf

Cloud Data Warehousing Challenges

The evolution of the cloud brings with it changes in IT culture. As the cloud continues to enable IT professionals within an organization to offload tasks, companies can make use of the newly found time, and shift the focus of their IT team members from answering email requests and troubleshooting to more global business needs.

One notable Redshift challenge is that resizing a large cluster requires a significant amount of dedicated staff time when your cluster is read-only. Furthermore, analysts are not able to change the data, so depending on the size of your cluster, you need to take into account data specialists who can run concurrent queries.

BigQuery is faster at resizing the cluster than Redshift. Data analysts are therefore able to both employ a solution and administrate it - eliminating the need for data specialists. This offers an advantage over Redshift, where maintaining an optimal data warehouse infrastructure requires specialized skills and greater effort to continuously tune the clusters that meet your specific use case.



Closing the Gaps

Although Redshift and BigQuery each provide strong data warehousing capabilities, there are ways to optimize each solution, and make operations more efficient. Automating the data stack removes the overhead of preparing and modeling data or managing infrastructure, and streamlines data management.

For example, contrary to previous findings that didn't consider optimization, we found that when reasonably optimized, **Redshift outperforms BigQuery in 9 out of 11 use cases hands down** [7]. This is especially true when the rest of your infrastructure is already on AWS.

CHOOSING ELT OVER ETL

For the last couple of decades, Extract, Transform, Load (ETL) has been the traditional approach for data warehousing and analytics. With Redshift's scale and operational efficiency, however, ETL can actually be viewed as a rigid and outdated paradigm. The Extract, Load, Transform (ELT) approach changes this paradigm.

Leveraging Redshift scalability with a solution like Panoply.io allows you to move away from overnight ETL processes. Panoply.io follows the ELT process, whereby all of the raw data is instantly available in real time, and transformations happen asynchronously during query time. This makes Panoply.io both a data lake and data warehouse, allowing users to have continual, real-time access to their raw data. They can iterate their transformations in real time, with updates instantly applied on newly-inserted data, as well. Furthermore, customized, advanced transformations are also possible via the Panoply.io UI console, and take just minutes to set up and run.

IMMEDIATE ACCESS TO YOUR DATA

BigQuery automatically replicates data to ensure its availability and durability. However, complete loss of data due to disaster is less common than the need for fast, instant restore; for example, the retrieval of a specific table or even a specific record. For both purposes, Redshift automatically stores backups to S3 and enables you to revisit data at any point in time over the last 90 days. However, retrieval also includes a series of actions that can make instant recovery a cumbersome, lengthy operation.

[7] http://panoply.io/blog/a-full-comparison-of-redshift-and-bigquery/

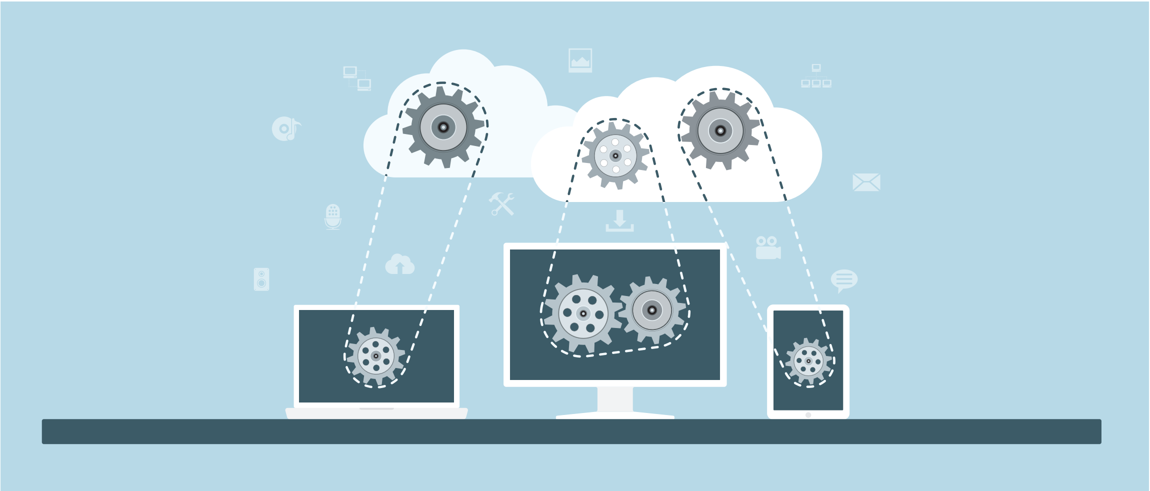
In order to facilitate data retrieval from S3, leveraging Panoply.io's Revision History tables, users can keep track of every change to any of their database rows, right within their data warehouse. This makes it immediately available to analysts with simple SQL queries, rendering file uploading to S3 and database extraction redundant.

EMBOLDENING THE CLOUD

AWS disrupts the world of infrastructure IT, and Amazon Redshift in particular disrupts the traditional data warehouse market.

Most research has confirmed its effectiveness as a well-balanced service, providing the performance, flexibility, efficiency, and robustness required from a modern, cloud-based data warehouse. Running and managing petabytes of data at scale, Redshift makes traditional infrastructure challenges obsolete.

The system leverages a great cloud infrastructure and empowers it by eliminating its management overhead, including both the underlying infrastructure and the database layers. This opens the door for more organizations to benefit from their data in ways they could not even imagine possible in the past, such as running a cross-department performance analysis. Panoply.io enables this not only for large enterprises, but also for small and medium organizations. The latter cannot afford or are typically sensitive to the costs attached to the DBAs and IT specialists required to establish and run a data warehouse.



SIMPLIFYING DATA MANAGEMENT

With Redshift, you need deep knowledge and a particular skillset in order to use and optimize it effectively. This is really where we add significant value. Panoply.io utilizes machine learning and natural language processing (NLP) to automate standard data management activities – saving thousands of code lines, and countless hours of debugging and research. The system also simplifies how you keep track of vast amounts of data – by identifying patterns, providing notification of anomalies, and generating alerts when the results of arbitrary SQL queries exceed predefined thresholds.

With a Redshift/Panoply.io combination, you don't need dedicated engineers to manage your organization's data management needs. What previously took in-house data specialists' time, and cost companies unnecessary overhead, now becomes automated and seamless.

The Winner of the Methodology Debate? You

Data warehouse consumerization by analysts starts with operational simplicity. This is the key when it comes to any product, but in particular for organizations that are constantly acquiring, combining, integrating, and analyzing data; and whose performance is measured by rapid response and reporting.

Both data warehousing cloud giants have their advantages and disadvantages when it comes to performance, usability and overhead. Panoply.io streamlines data management and the complexities of the cloud based data warehouse into a single click, without compromising on functionality.

We invite you to **learn more** about the Panoply.io platform and capabilities and start your **free trial here**.

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.