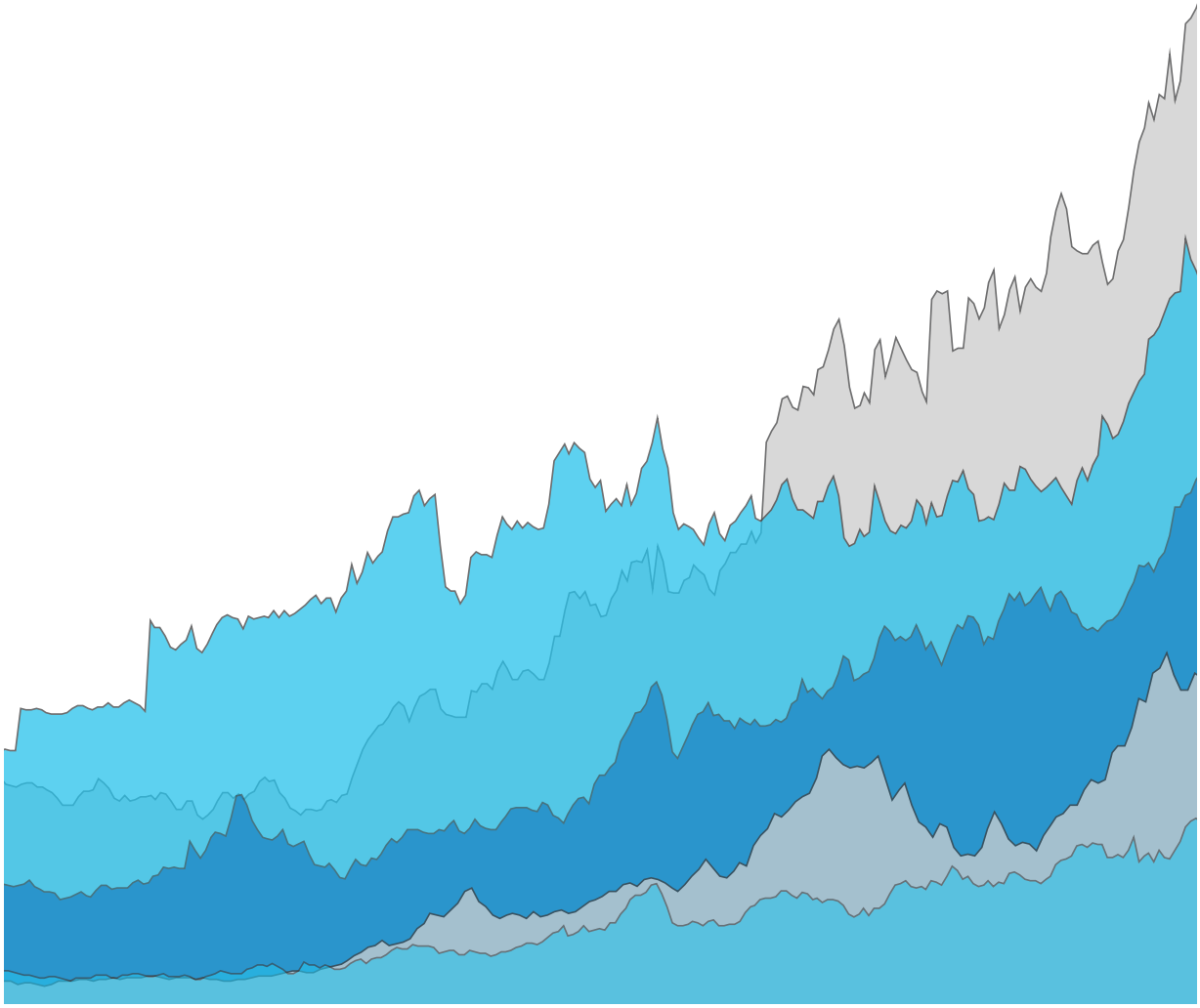


Delivering rapid-fire analytics with Snowflake and Tableau

Todd Beauchene, Snowflake Computing
Saqib Mustafa, Snowflake Computing
Jon Bock, Snowflake Computing
Ross Perez, Tableau Software

Contents

Abstract

Until recently, advancements in data warehousing and analytics were largely incremental. Small innovations in database design would herald a new data warehouse every 2-3 years, which would quickly become overwhelmed with rapidly increasing data volumes. Knowledge workers struggled to access those databases with development intensive business intelligence tools designed for reporting, rather than exploration and sharing. Both databases and business intelligence tools were strained in locally hosted environments that were inflexible to growth or change.

Snowflake and Tableau represent a fundamentally different approach. Snowflake's multi-cluster shared data architecture was designed for the cloud and to handle logarithmically larger data volumes at blazing speed. Tableau was made to foster an interactive approach to analytics, freeing knowledge workers to use the speed of Snowflake to their greatest advantage.

Both products are independently revolutionary, but in combination they can allow you to overcome many of the analytical challenges faced today. This paper will describe the methods and techniques you can use to fully utilize the power of Tableau and Snowflake together, along with best practices for optimizing your processes in both products.

Common analytics problems

Performance and Scalability

Analysts are the first victims of performance limitations. Analytics workloads are often pushed to off-peak times to reduce the effects of limited scalability on concurrency. Database credentials are closely held. Perhaps most frustratingly, there are often specific and complex rules for querying the database that can limit the ability of business users to find the data that they need. In many cases, because of the complexity of working with the database and the development intensive nature of legacy BI products, business users don't have any access to the data and information they need.

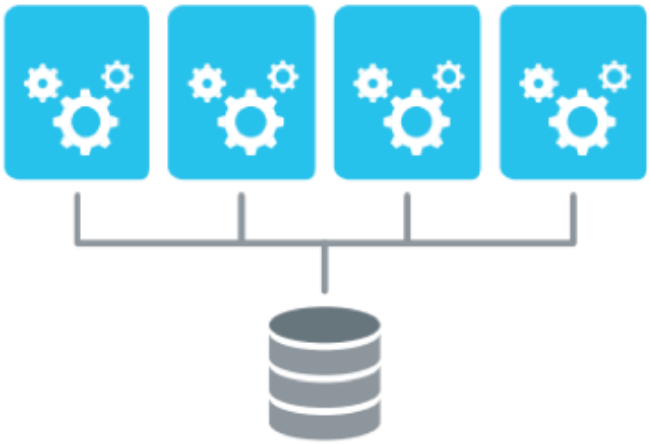


Fig 1. Shared disk architecture is limited by the performance of the disk

Fig 2. Shared nothing architecture is limited by the need to distribute and query data across nodes

As data volumes have skyrocketed, scalability has become the overarching concern for database and analytics experts alike. Traditional database architectures have been unable to address those concerns completely. Shared disk data warehouses (see figure 1) are hampered with concurrent queries bottlenecking at the disk. Shared nothing data warehouses (see figure 2) struggle to partition data efficiently for multiple needs, as well as to handle joins and queries that involve multiple partitions. The larger the data volumes, the more acute each of those limitations become.

Inflexibility

Limited scalability and performance expose another common problem: inflexibility. When facing performance and scalability problems, the knee-jerk reaction is to simply buy more database. Of course, due to the logarithmic nature of performance degradation, that rarely buys much time. It also exposes another problem: the inability to right size. People naturally purchase their data warehouse to match their needs at the point of highest demand, but rarely is that capacity used around the clock. When dealing with products that can cost millions of dollars, that unused capacity can be expensive.

Many data warehouses are also limited in the type of data they can store. The rise of the Internet of Things, and the prevalence of data formats like JSON in general, has led to a surge in the amount of semi-structured data organizations need to store and analyze. But, many traditional data warehouses are unable to house this data, and if they can will rarely be able to query it in conjunction with other types of data.

Traditional analytics tools suffer from inflexibility of a different nature. As businesses change and adapt, their dashboards, reports and analytics evolve as well. Traditional analytics tools are often so rigid that changes to existing reports can take months, involve multiple technical resources, and hamper the ability of anyone to actually find the information they need.

Complexity

Poor scalability and flexibility inevitably lead to a third problem: complexity. Many database administrators spend the better part of their days endlessly tweaking and tuning the knobs on their database to ensure that everything is optimally performing. It's a challenging job, changing distributions, sort keys, compression, and worrying about encryption. A tweak by a BI user in one area might lead to problems in another.

BI professionals have to deal with complexity brought about by their legacy tools. These legacy tools (and some of the new ones) have onerously complex calculation and visualization engines that force basic business users to ask for help with relatively straightforward analytics questions. This is time consuming for the whole team, hampers IT with distracting tasks, and prevents the line of business from being able to find the answers that they need.

A different approach with Snowflake and Tableau

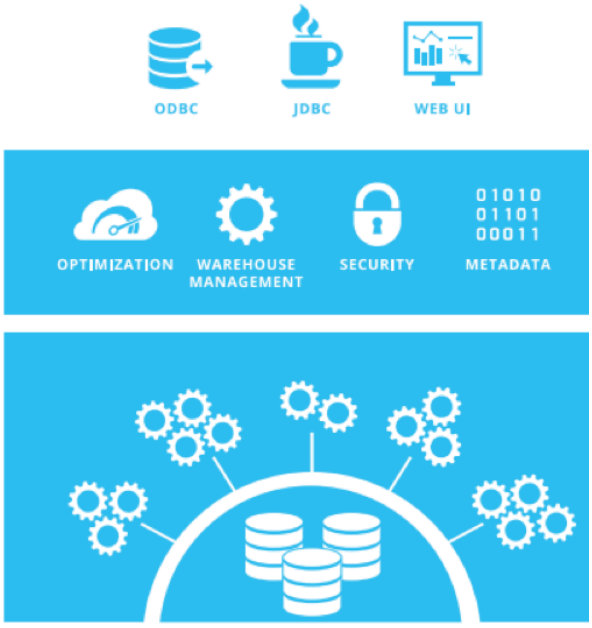
Snowflake and Tableau were built differently, and the effect on analytics can be dramatic.

Infinite Scale

Snowflake utilizes a new architecture built for the cloud: multi-cluster, shared data (see figure 3). From the end user perspective, it's like any other SQL database, but the architecture is fundamentally different. All of the underlying data is stored in the cloud on Amazon S3. Compute is handled with independent clusters (or groups of clusters) called virtual data warehouses. The operations of each virtual data warehouse are completely independent of one another, and have no effect on the integrity or referencability of the underlying data. This means that you can store an infinite amount of data, and scale your compute to match an

infinite workload.

Organizations that use Tableau are in an excellent position to take advantage of the scalability of Snowflake. Because Tableau was designed with simple connectivity and an easy-to-use drag and drop interface, anyone with a question (not just technical users) can connect to Snowflake and start finding answers in seconds. Specific workloads, like a reporting dashboard, can utilize their own virtual data warehouse, ensuring reliable performance independent of other



activity on the database.

Fig 3. Snowflake's multi-cluster, shared data architecture

Choice

Because each Snowflake virtual data warehouse is independent, and can be scaled up and down on demand, organizations are able to adjust their capabilities (and cost) to demand. In other words, you are able to choose and change your data warehouse to meet your needs at any time. It's simple, but revolutionary in a world with fixed cost and massive up-front investment.

Snowflake is also able to handle structured and semi-structured data at the same time. There's no specific set up requirements or preparation, and views can easily be created which will allow structured and semi-structured tables to be queried in Tableau at the same time.

Tableau matches Snowflake's flexibility with a range of options for analyzing and sharing data. Choose from thousands of different visualizations which can be combined, with a drag and drop, into custom interactive dashboards. When it's time to share data or dashboards with others, Tableau has fully managed, cloud hosted, and local options. Anyone with permissions can use a web editor to tweak or refine dashboards and visualizations, making updates and

changes seamless. Data in Google Sheets, local CSVs, and even legacy databases can be easily combined with Snowflake when needed.

Simplicity

Both Snowflake and Tableau were designed to be straightforward to use and manage. Since Snowflake is a data warehouse as a service, you would expect the infrastructure to be fully managed, but the service extends far beyond that. Unlike many databases, Snowflake has few "knobs" to turn: it adapts to usage patterns and dynamically responds. Encryption is automatic.

Tableau is similarly straightforward. There's no code or custom SQL needed to connect to your data in Snowflake. Inevitable changes to dashboards and systems can be handled with a right click or drag and drop.

How to get the most out of Snowflake and Tableau

What you don't need to do

As you've already seen, Snowflake and Tableau are an analytics system that requires very little optimization. We'll offer some general guidelines below on how to get the most from both tools, but it's important to note that there's a great deal that you won't need to do when using them together.

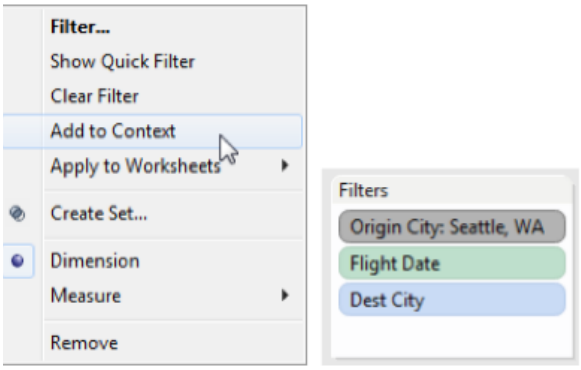
For instance, there's no need to create or manage indexes. You won't need to optimize your SQL, or even write SQL at all for much of your work within Tableau. There's no need to worry about data partitioning, or workload management either because those are handled automatically by Snowflake.

Once these traditional points of optimization are eliminated, there are smaller and more targeted groups of best practices that should be straightforward to follow. We'll focus first on proper initial set-up, and then dig in to the methods you can use to find and troubleshoot problems.

Set yourself up for success

Both Snowflake and Tableau provide multiple features that, if used properly, can help you to avoid performance problems altogether.

- **Workload Isolation in Snowflake** – By leveraging different virtual warehouses when defining your Snowflake connections in Tableau, you can ensure that separate query workloads do not impact each other. This can prevent data exploration from interfering with reporting. As a best practice, many organizations will have a virtual data warehouse defined for exploration, and then they will create a separate virtual data warehouse for reporting. You can easily switch the connection when publishing a Tableau dashboard by connecting to the reporting data warehouse, and then right clicking on the existing datasource and selecting "Replace Data Source".
- **Add filters in Tableau before you start** – Tableau's iterative nature means that every drag and drop will query the database. When working with large amounts of data, add filters as you connect to the datasource, or before creating the view, so that the underlying SQL that Tableau sends to Snowflake will be appropriately limited.
- **Use Tableau's context filters** – Most dimensional filters in Tableau can be added to context, effectively creating a temporary subset that significantly limits the queries Tableau sends to the database. These types of filters should only be applied to filters you do not intend to change. Refer to Filtering in the Tableau online help for how to create



context filters. For more information about performance improvement with context filters, see Speeding up Context Filters in the Tableau Online Help.

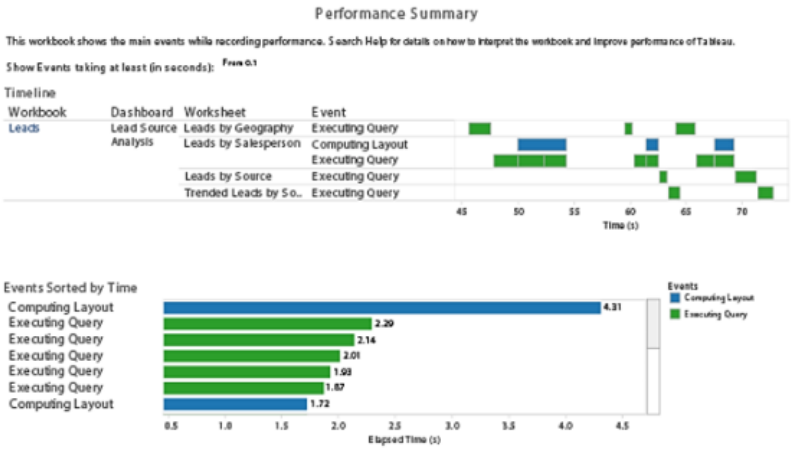
- **Bring your semi-structured data into Snowflake** – Snowflake has native support for JSON, AVRO and Parquet data. Often, these types of datasets are loaded into separate systems that are difficult to query with Tableau. However, since Snowflake supports these data types and makes them accessible for analysis in Tableau, it benefits you to bring them into Snowflake for analysis. This data can be ingested without predefining the schema, and

a basic view can then make that data available in Tableau. Additionally, tables containing semi-structured data can be joined to any other table including other tables that contain semi-structured data to provide flexible data models.

Finding performance problems

If you've already isolated your workload and have filtered as much as possible, but are still experiencing sub-optimal performance, it might be time to dig deeper. These tools will help you to more accurately pinpoint where you are running into problems. After identifying problems with the tools in this section, read on to find suggested fixes within Snowflake, and in the construction of your Tableau dashboards.

- **Tableau Performance recorder** - Performance Recorder is a powerful built-in tool that pinpoints slow queries and helps optimize workbooks for maximum performance. It does this by tracking the elapsed time for an individual workbook to execute a query and compute the layout. Hovering over one of the green bars below will show the user the query that's being generated against Snowflake. If a query is performing poorly, read on for more information on how to optimize your analytics tasks in Tableau for performance. For instructions on how to create or interpret a performance recording, please follow one of these links.
    - Performance Recorder on Tableau Desktop



    - Interpret a Performance Recording on Tableau Desktop
    - Performance Recorder on Tableau Server
    - Interpret a Performance Recording on Tableau Server

Status	Query ID	SQL Text	User	Warehouse	Size	Start Time	End Time	Total Duration
✔	b9bf4420-...	show parameters like ...	TEST	LOAD_WH		9:15:07 AM	9:15:07 AM	55ms
✔	25a4e43f-...	show parameters like ...	TEST	LOAD_WH		9:15:07 AM	9:15:07 AM	92ms

- **Snowflake History** – In a similar vein, it might make sense to view the same queries from within Snowflake. If you log in to your Snowflake instance, and click "History" from the top of the screen, you can see the queries that have been executed, and how long they took to execute.

Addressing frequent queries and concurrency

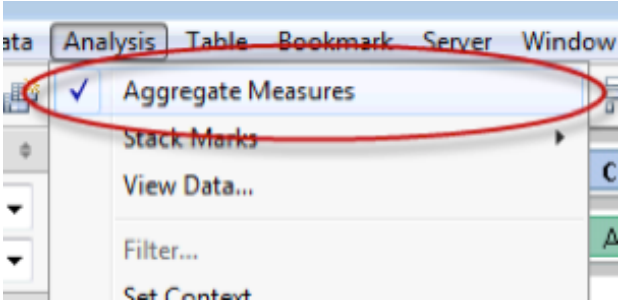
If the Tableau performance recorder shows a relatively performant workbook, but you are still seeing degraded performance, it's possible you have a concurrency problem. In other words, there may be too many queries going to the database at the same time. There are several ways to help address this.

- **Automatic scaling in Snowflake** – Snowflake's Multi-Cluster Warehouse future provides the ability to add compute resources automatically as additional Tableau users increase the concurrent load on the database. This feature also automatically scales down compute resources once demand subsides. Many organizations will enable this automatic scaling on their reporting data warehouse.
    - Learn how to enable automatic scaling
- **Query Result Caching in Snowflake** – Snowflake automatically caches all query results to provide extremely fast response times for queries that run multiple times throughout the day. This cache is intelligent enough to prevent users from ever seeing outdated data, but can significantly reduce the impact of queries that are run frequently. A best practice is to pre-populate the result cache after each data load for commonly run queries.

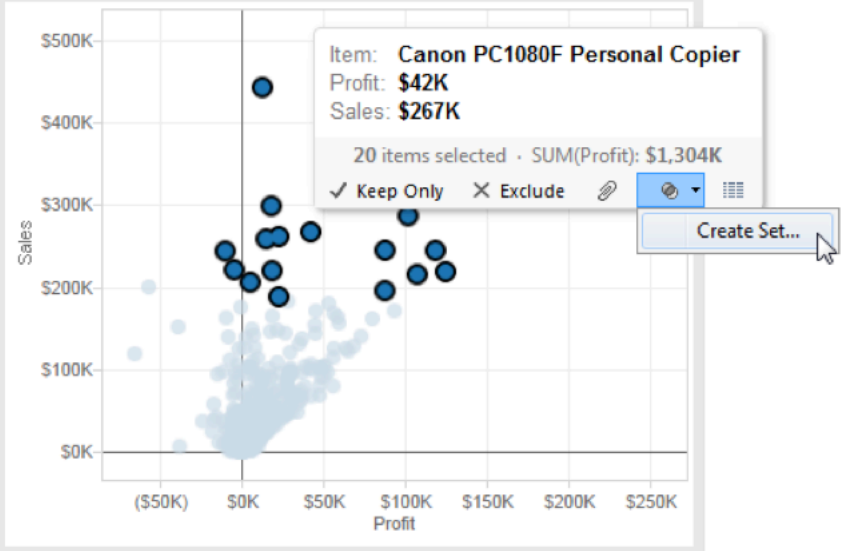
Optimizing your Tableau dashboards

If the Tableau performance recorder shows a workbook with long and complex queries, you can use these methods to optimize your dashboards and views.

- **Aggregate measures** – Aggregated measures are significantly less expensive than



disaggregated measures. Slow views are often hampered by the need to pull every value from every row to show disaggregated measures. This can be avoided by aggregating the data. To do this, make sure the Aggregate Measures option on the Analysis menu is selected. For more information, see Disaggregating and Aggregating Data in the Tableau



Knowledge Base.

- **Sets** – Sets can be used instead of quantitative filters to remove members based on a range of measure values. For instance, a set can be created that only returns the Top 50 items in a dimension, rather than all of the items in a dimension. For more information, see Creating and Using Sets in the Tableau online help.
- **Include only necessary columns** – When creating a group from a selection as described in Sorting, Grouping, and Sets in the Tableau online help, make sure to include only the columns of interest. Each additional column included in the set will result in decreased performance.

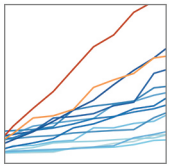
How to get started

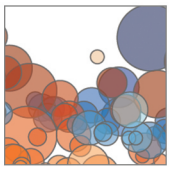
If deployed together, Tableau and Snowflake can help any organization to deliver a scalable, flexible and simple analytics platform. Free trials of both products are available on-demand at any time, from the links below.

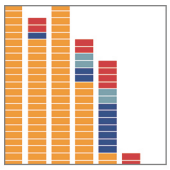
- Try Tableau Desktop
- Try Snowflake On-demand

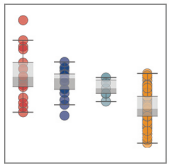
About Tableau

Tableau Software helps people see and understand data. Offering a revolutionary new approach to business intelligence, Tableau allows you to quickly connect, visualize, and share data with a seamless experience from the PC to the iPad. Create and publish dashboards and share them with colleagues, partners, or customers—no programming skills required. See how Tableau can help your organization by starting your free trial at tableau.com/trial.









Additional Resources

Download Free Trial

Try Snowflake On-demand

Related Whitepapers

Why Business Analytics in the Cloud?

5 Best Practices for Creating Effective Campaign Dashboards

See All Whitepapers

Explore Other Resources

· Product Demo

· Training & Tutorials

· Community & Support

· Customer Stories

· Solutions