The Bloor Group

ANALYTICS AT THE SPEED OF THOUGHT

Tableau's in-memory software architecture; what it is and what it means for BI

Robin Bloor, Ph.D.

WHITE PAPER

Introduction and Executive Summary

Tableau is a software company that specializes in highly visual and rapid-fire reporting and analytics, using breakthrough technology for both its visualizations and data management. In its current version, Tableau is taking business intelligence (BI) to a whole new level of fast. This white paper examines Tableau's visual capabilities and its new data engine, which together make for a formidable partnership for delivering BI solutions.

Here is a brief summary of the contents of this paper:

- Visual analysis for data discovery is a must in today's BI space. Knowing that the human brain processes information visually, Tableau's developers created an intuitive product that aligns with human thought. Much like our tendency to ponder questions and arrive at conclusions – often termed "stream of consciousness" – the Tableau user interface allows for the progression of analytics via familiar visual representations. In doing so it uses:
  - A visual drag and drop interface
  - Conventional graphics, such as tables, charts and maps
  - Interfaces for desktop, web browser, and iPad/mobile devices
- While Tableau has always made it possible for users to connect to live data sources, either local files (e.g., Excel) or databases (e.g., Oracle, SQL Server, Teradata), previous versions of Tableau could be constrained by the underlying data sources, both in the speed of analytics and the volume of data that could be processed. Tableau was configured either for local usage (e.g., files, spreadsheets) or to query data directly in data marts or other data stores. Both modes of usage could give rise to latency problems. Running natively on a PC using local data sources, Tableau lacked the ability to handle large amounts of data. When accessing data in data marts the processing speed depended primarily on the speed of the database being accessed. So while Tableau's analytics capabilities were up to the task, the user could experience sluggish performance if the back-end database was slow. With Version 6 of Tableau, this became a thing of the past.
- Tableau Version 6 introduced a new data engine that provides in-memory analytics optimized for speed. The improvement in performance it delivers is dramatic.
- It achieves speed through emulating some of the high-performance capabilities of column-store databases, such as Sybase IQ and Vertica, making effective use of data compression and storing data in columns. The approach is tailored for use in the single PC environment by the inclusion of a virtual memory paging capability.
- Because of this data engine Tableau is now capable of processing very large data sets locally – up to about a billion rows – and because the majority of that data is held in memory, it produces results rapidly. Additionally, it is still capable of accessing data in data marts. Organizations with fast, highly performant databases can still access those directly with Tableau.
- Tableau has been a BI platform since the introduction of Tableau Server in Tableau 4.0, and many customers use Tableau as their primary BI platform. Report authors can

build reports, dashboards, or even drill-down analytics and distribute them to less sophisticated users with existing data and user security enforced.

- It is clear from our interviews with two Tableau Version 6 customers (Kaleida Health and Nokia) that Tableau is a changed product. It alters the way in which BI can be carried out. Because Tableau can now do analytics so swiftly and gives people the choice to connect directly to fast databases or use Tableau's in-memory data engine, it has become much more powerful in respect of data exploration and data discovery. This leads to analytical insights that would most likely have been missed before.

Beyond Rapid-Fire BI

The name Tableau is synonymous with visualization and rapid-fire BI.

Born from years of academic research, Tableau stormed out of Stanford in 2003, carrying on its shoulders the highly innovative database visualization language, Visual Query Language (VizQL™). This language had the same analytic and query capabilities as its predecessors, SQL and MDX, but it took reporting one important step further. Its visual language facilitated an unlimited number of unique and customizable picture expressions of data. Recognizing the need for more effective BI, creators Pat Hanrahan and Chris Stolte developed VizQL™ as a means of solving an area of perpetual neglect: visual analysis.

Let's face it, we 21st century humans are visual creatures. Since the advent of television, the way we process knowledge has been shaped by movement, images, and color as opposed to the mundane rows of text and numbers that were once the only expression of BI. When personal computers became more about graphics than programming, it only enhanced the user's need for visual representation. To some purists, this was a travesty; however, it could not be undone.

A founding principle of Tableau is that data analysis should mimic the user's inherent ability to think visually. Instead of running a query against the typical SQL database, producing a report and creating a visualization to accompany the results, Tableau's visualization engine marries the two. In the Tableau world, analysis and the visual aids that represent and explain the analysis should not be separate.

Information Discovery

Tableau's user interface doesn't just help the user create visualizations. It serves as a data exploration tool that partners with the user and his train of thought, allowing for a visual cycle of analysis. Because of Tableau's careful, incremental, stream of consciousness visualizations, the user often finds unexpected results, which in turn call for deeper analysis. After all, the goal of analysis is not to find one answer to one question; it is to find a collection of answers and propagate more questions.

Tableau does not inundate the user with a vast array of unconventional graphics; it relies on the expectations of the end user. What is familiar? What will be easy to understand? Simple, elegant and richly colored bar charts, tables, maps, and time series are the most common types of business graphics. Although nearly any type of graphic can be generated, Tableau's practice demonstrates that design is second to the data and must act as an enhancement, not a distraction.

Versatile visual display is not the only compelling reason to use Tableau. All of Tableau's products – Desktop, Server, Digital and Public – are fast, user-driven applications with the option of being configured for desktop and/or the web. They promote user control.

Though it has traditionally been the case, it is not the job of IT to read the minds of business users concerning exactly what type of report or representation of data is required. Little satisfaction can be gained from either side when a report is requested and generated, only to be requested again for lack of a specific or forgotten detail. And while neither side is particularly at fault, it has led to the widening ravine that separates the back end from the front end.

Tableau Limited

The power to make good decisions depends on the data available. Tableau gives users the analytical power they need and it makes exporting results as easy as a few clicks of the mouse. Where IT is able to deliver the appropriate data service, business professionals can serve themselves the data and reports. However, there were limitations.

Prior to Version 6, it was not practical to extract large amounts of data in Tableau, due to the query performance. With Version 6, Tableau has made it practical to extract and work with large volumes of data at fast query speeds.

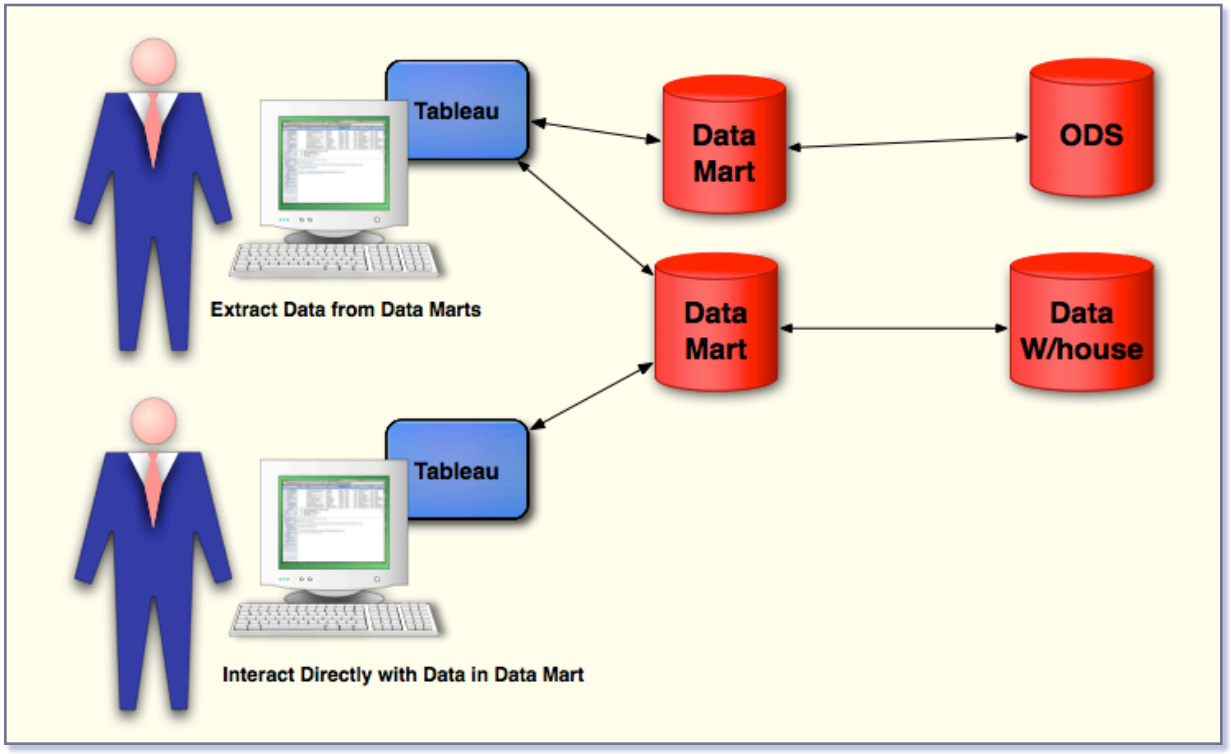

Figure 1. Tableau's Operation, Constrained by Data Limitations

This is illustrated in Figure 1. Tableau could be used in two ways. It could extract data from one or more data marts and then analyze that locally on the PC. As such, prior to Version 6, it would be practically constrained by the query speed, so it would not be possible to analyze high data volumes.

Alternatively, it could connect directly to a data mart and work directly with the data mart's database. This would enable Tableau to operate on a higher volume of data, but unless the data mart were a high-speed analytical database, performance would be compromised by query speed of the data mart and analytical work would be slower.

In practice this limitation had reduced Tableau's scope of operation. It meant that multiple BI products (dashboards, OLAP tools, analytics tools) might be deployed where Tableau itself would have been sufficient if it were able to deal with greater data volumes. This limitation was eliminated with the release of Tableau Version 6.

The Joy of Architecture

In Version 6, Tableau was given a wholly new data engine with an in-memory database at its core. This complements Tableau's ability to directly connect to data sources. The addition of an in-memory data engine makes the advantages of high-speed analytical databases available to users who are working with slow databases. This has had a major impact on the way that users can exploit Tableau, which in turn has altered the way businesses use and deploy Tableau. Of course, the data engine does not change the analytics and BI capabilities of Tableau itself, but it radically changes the size of the data sets that Tableau can operate on,and, because of that, it creates new possibilities for how Tableau can be deployed within the business. Describing how Tableau's in-memory architecture works will make this clearer.

A Marriage of Memories: Real and Virtual

It helps to think about Version 6 of Tableau, in the way that it is illustrated in Figure 2. Tableau embodies a Data Engine in addition to the components of the product that execute analytical routines, manage the user interface, and display information.

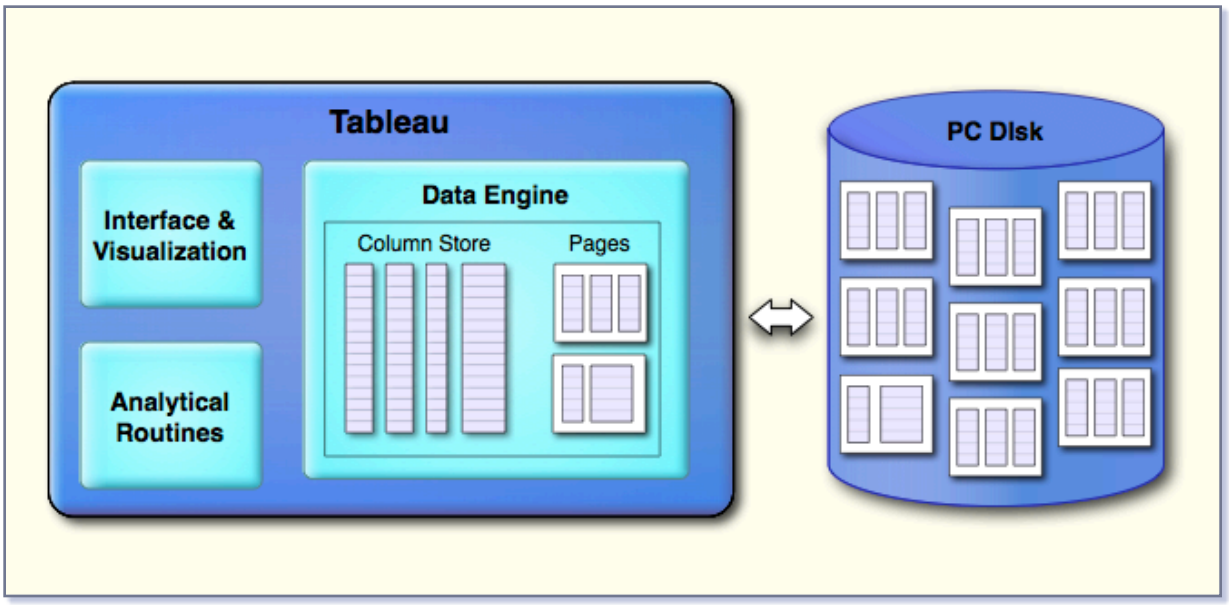

Figure 2. Tableau's In-Memory Architecture

Tableau's in-memory database works in almost exactly the same manner as many large column-store databases, such as Sybase IQ or Vertica. With such databases, which are focused on processing multi-terabytes of data, two specific performance techniques are employed. The database tables are partitioned into columns with each column distributed across multiple disks and multiple machines. Additionally, data in each column is compressed dramatically using various data compression techniques.

Tableau Version 6 works in a similar way, but in miniature. With these big database products the data is spread over many disks and many servers, all of which work in parallel to deliver the answer to a query rapidly. Tableau only has the resources of a single PC at its disposal, so

it cannot take such an approach. However, it does take advantage of the parallelism if it can. Nowadays PC CPUs come with multiple cores on each CPU and some even come with 2 CPUs. So Tableau will spread the workload across these processing resources, utilizing as much processor power as is available.

Databases make extensive use of memory because it is so much faster – roughly a million times faster – to read data from memory than from disk. With column-store databases, because the data is stored in a compressed form, much more of it can reside in memory at any given time than with traditional databases. For the sake of the speed it delivers, Tableau's data engine fully exploits the available memory on a PC in the same way.

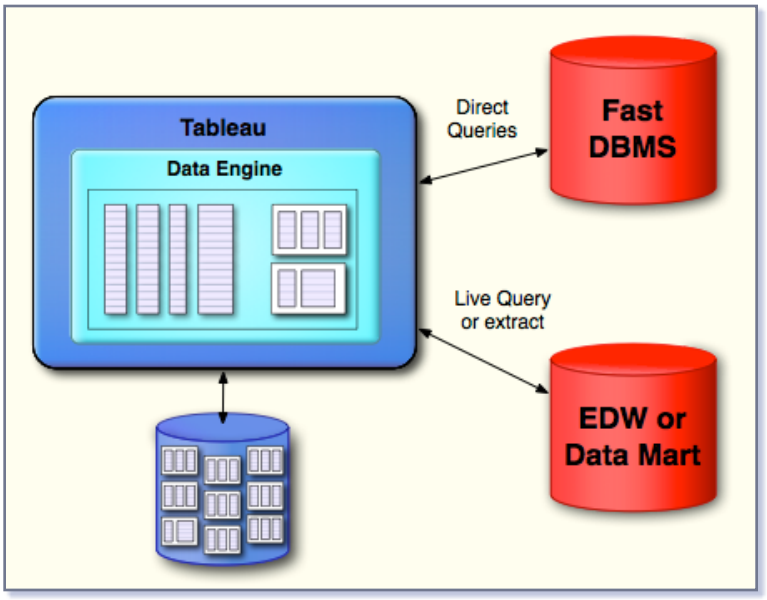

Figure 3. Tableau Direct Connect to Data

When it is processing more data than can be fit into memory, it writes it out to disk using a virtual memory capability. It divides the data into pages of equal size, writing out the pages to disk when memory is full and then reading them back in when needed. This activity is algorithmically controlled, so that the pages of data least likely to be required are pushed out to disk. When they are eventually needed, they are read in and other pages are pushed out.

In practice the efficiency of data compression is about 10 to 1. Naturally it varies according to the data, but 10 to 1 is a good "rule of thumb." With a PC that is configured with a large amount of memory (32 gigabytes), this means that Tableau can work on hundreds of gigabytes of data at a time. The limit is somewhere in the region of a billion rows of data.

Of course, not all PCs have such a generous amount of memory. But Tableau 6.0 will work happily with less – it just limits the data volumes that it can work on. There is no minimum requirement on the amount of memory; the memory on commodity laptops is sufficient. There is both a 32 bit and 64 bit implementation. In effect it can be implemented on almost any Windows device except a cheap Netbook.

Tableau Version 6 can also work via direct connection to a fast DBMS such as the column store databases we have described. This is illustrated in Figure 3. When connected to such data stores, it can query them directly and this may be the fastest way to analyze the data –and it is likely to be necessary when dealing with terabytes of data. However for lower volumes, it is a choice that can easily be made and simply tested by switching from extract mode to direct query mode.

Authors and Readers: The Tableau Server

Tableau is not a power-user-only capability. It's true that power users can serve their own needs with Tableau, if granted access to the data they wish to sift through. But in respect of corporate BI it is best to think in terms of report authors, who create targeted reports and data visualizations in Tableau, and report readers, who are the consumers of such Tableau reports. There is a server version of Tableau which caters specifically for report readers. It is almost identical to the PC version of Tableau, able to use the in-memory data engine or query data sources directly.

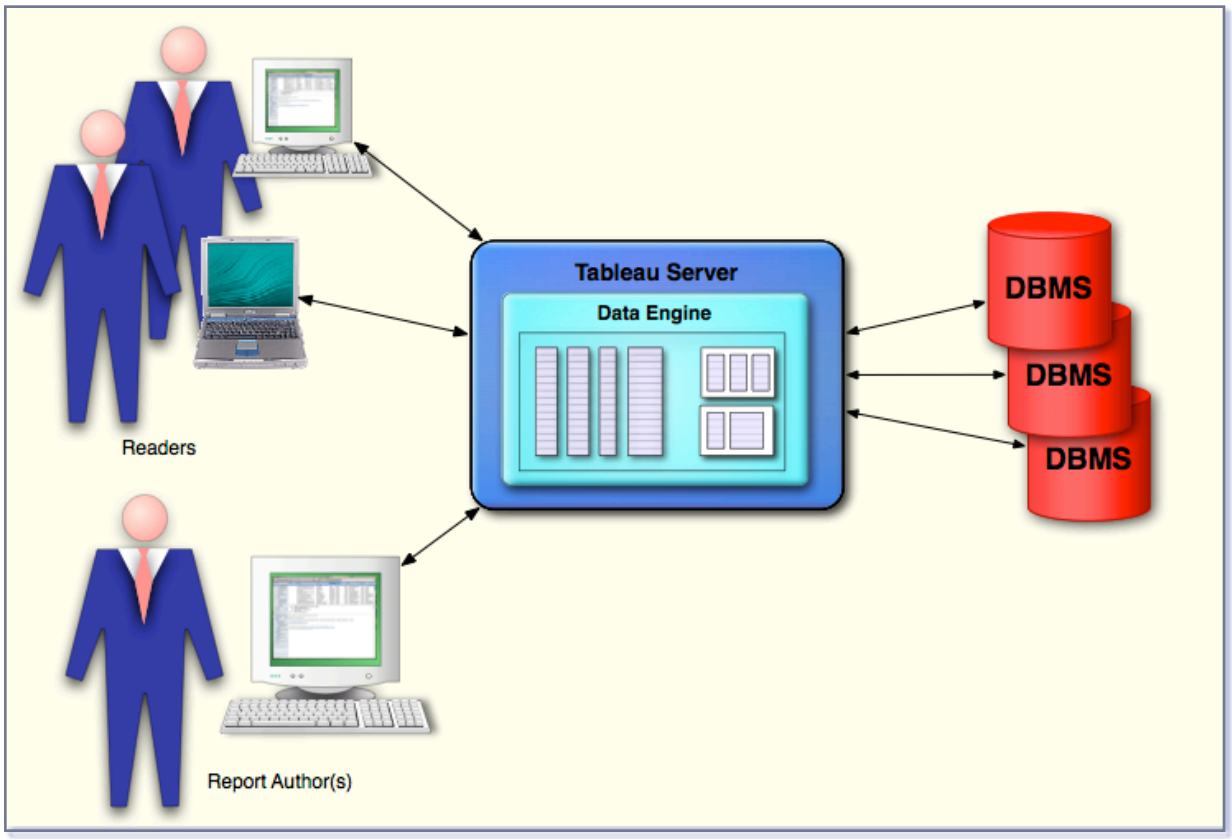

Figure 4. Tableau Server Implementation

This is illustrated in Figure 4. Report authors simply upload parameterized queries that they have created to the server and report readers can view them through a browser. If they have a local version of Tableau they can also download these reports and use them when offline.

The Tableau server scales up to work on clusters and thus is capable of serving a large number of report readers, including agents and supply chain partners. The data sources the server depends on will normally be refreshed on a regular basis.

Putting It All Together

The majority of companies do not have very large databases. It's true that some industries – telecoms, banking, e-commerce sites – do amass very sizable databases, but most companies do not. For many companies a one terabyte database is unusual or unheard of. And yet, until recently, it has been necessary for IT departments to provide a fair amount of infrastructure

in the form of data flows, data marts, and OLAP stores just to service the BI needs of the company.

With Tableau Version 6, the game changes. For such companies, Tableau can probably feed directly from the data warehouse or the operational data store. It is likely that Tableau will be able to replace some of the dashboards, reporting tools, and OLAP tools, considerably simplifying the situation for the BI user. Even in organizations where "big data" is a fact of life, Tableau has the power to simplify the BI environment to some degree.

The User Experience – At the Speed of Thought

The two dominant vectors of Tableau Version 6 are speed and scope. The in-memory data engine delivers results at lightning speed, far faster than any previous version of Tableau and far faster than most users have ever experienced, except on small amounts of data. This exceptional speed carries through even to the analysis of very large data sets that Tableau was previously unable to process.

We talked to some Tableau users to get a picture of the difference this makes.

Kaleida Health

Kaleida Health is the largest health care provider in Western New York. With 10,000 employees, five hospitals, a number or clinics and nursing homes, and a visiting nurse association – not to mention millions of patient records. Kaleida needed a BI tool that could handle large data sets quickly and painlessly.

Kaleida found traditional reporting tools inadequate to handle its requirements. Driven by the need for a deeper analytic capability, Kaleida created a corporate analysis department around Tableau. Instead of simply pushing out mass reports, Kaleida's BI department wanted a team that could give users what they wanted: reports they could understand.

Jennifer Keubler, Corporate Analyst with Kaleida, said although their data warehouse is robust, Oracle Reporting proved to be clunky and unfriendly. "We had nurses and non-financial people, and people that didn't really understand reporting and visualization, trying to interpret these reports." The result, she said, left much to be desired. Users often didn't notice the data that should have stood out.

"We created the corporate analysis department to work closely with the managers and senior executives on different projects they were doing, to not only pull the data for them, but also look at it, present it, and dig deep into the details, and find things that they should be looking at, and then highlight them in PDF files or PowerPoint presentations using Tableau," said Kuebler.

Kuebler said she spends 95 percent of her day – and runs all of her calculations – in Tableau, primarily pulling data from Oracle databases into Excel first. For monthly analytics, she extracts data directly into Tableau from the data warehouse and is able to save and re-run queries, which is a huge time-saver when dealing with millions of records, with 40+ fields per record.

Speaking of the previous version of Tableau she noted, "I was always able to use, to pull in, as many records as I wanted, but it just ran so incredibly slow that I'd do things in pieces rather than all at once," she said.

The new version eliminated that problem. "Instead of having to re-run everything, it's just a couple clicks. A lot of other people will just run high level reports or summaries and they can't really dig into the details, whereas I start at the detail level then create the summaries. So if somebody wants somethings different, I already have a file with all the details. It's just a matter of moving things around."

Besides speed, another feature of Tableau 6.0 is the ability to link different tables. Kuebler said she uses this function to merge files to an Excel table or link them to different extracts. One example is if she has an extract by zip code, but wants to look at data by county. "If I don't have county in my file, I have an Excel file that links all the ZIP codes to a county, so I can just link it to that, and the run information by county."

Kuebler's team was asked to look at the emergency room and the patients who visit it more than ten times a year. The data revealed that there was a problem among the Medicaid population: they were frequently using ER and ambulance services inappropriately for stomachaches, headaches, fever.

"It kind of spiraled and spiraled, and the next thing you know, it was a local news story, using the data that we pulled from Tableau," said Kuebler. "And when they did the news story, they made it sound like there was this great research project and the team digging for months. And it was really just a one-day project."

Kuebler also uses Tableau to manage resource utilizations: who's using what supplies and how much those supplies cost. This ultimately leads to efficiency and standardization across the system, something she said could previously not be done internally.

"It's fairly easy to do using Tableau and working with managers and doctors, and this is something that prior to us instituting our BI department, we'd have to contract out with health care consultants, to take a look at this stuff and pull our data."

Just like any competitive enterprise, a health care system also needs to gauge the marketplace. Kaleida uses Tableau to compare itself to other hospitals across the country, looking at the length of patient stay, hospital practices, market share and partnerships with doctors. "It's endless, the types of things we can look at," said Kuebler. "I could go on forever about this product. It's made my life a whole lot easier."

Nokia

Ravi Bandaru, Nokia's Product Manager for Advanced Data Visualizations and Data Analytics ,has been using Tableau since July, 2010. Within his organization, he said about 350 to 400 people use Tableau – either in desktop or interactive form – and that it has brought people together.

"Earlier, people were scared of using existing BI tools to discover data, so they were relying on intermediaries, the IT guys, to bring the data to them," he said.

He noted that Tableau's in-memory capabilities basically offer two benefits: they provide a do-it-yourself interface, and they increase the speed of query performance.

"It's letting the analyst do more analysis himself or herself without IT coming between them and their data," he said. "Using this kind of in-memory capability, I do see this being useful in exploring more complex and largish data sets, which were inaccessible before."

Nokia uses Tableau for marketing analysis, and they layer Tableau on top of their own database. Bandaru said that normally running live queries against their database does not give the kind of response times that they need, but running live queries against Tableau's data engine provides the type of instant responses they expect and can work with. He said that it is useful for ad hoc analysis and that most analysts refresh their data on a weekly basis.

Bandaru said he sometimes recommends Tableau for certain users, and other times they come to him wanting to use it. The typical end user he deals with is not versed in analytics, statistics or SQL.

Said Bandaru, "If the end user is currently data savvy or analytics savvy, he loves using Tableau because it's providing him access to lots of data sources now." But a less sophisticated user might just use it as another BI tool, and it's up to management to provide more education to explain that "you can actually leverage this to do these kinds of complex things which are impossible while using the rest of the tools."

Nokia is currently leveraging Tableau Version 6/6.1, and Bandaru said he expects some improvements in 7.0, such as sharing extracts across workbooks, allowing the extract to work as a data source, and having these extract be available to multiple users across the globe.

He said this type of streamlined collaboration would be particularly useful for a new employee who is not yet familiar with company operations. Today, he said, the problem is that they always have to go to the original data source, but he said a future benefit would be to have that employee able to extract and leverage data that has been created by any department, anywhere.

Another potential improvement Bandaru said he would like to see is a more seamless integration between Tableau and Hadoop. Currently, analysts have to move data from Hadoop through a SQL database before they can start using it in Tableau, but he said his ideal use case would be to "extract the data into the data engine of Tableau and start using it."

Despite Tableau's capabilities, he said he has a feeling that not many users are exploring the full potential of the tool. He said the primary motivation is to "bring some workbooks and publish it to wider audiences and let them actually use some of the data or some of the insights" instead of, for example, using Tableau in an offline capacity.

Bandaru said he tries to ensure product adoption by offering in-house collaboration tools, such as Wikis, training sessions, documents, and best practices.

The Difference that Makes a Difference

These user experiences of Tableau Version 6 illustrate the movement of Tableau from being a BI tool in the traditional sense, to becoming a BI platform that can take responsibility for a large amount of the BI needs of any organization. Technically, the difference is in the architecture. Direct-connect leverages existing highly performant data sources. And the in-memory data engine drives the possibilities. The efficient caching and processing capabilities mean that data sets do not need even need to be fully loaded into memory before analysis commences.

But the user is unlikely to know or even care about that. It's speed and scope that they notice. Analysis can be performed at the speed of thought, leveraging more data on less hardware. This is true ad hoc analysis where the user does not have to determine in advance which measures to aggregate or query. The user can explore the data in every one of its dimensions, digging down into detail or summarizing into categories. Almost every form of data visualization is there, available at speed of thought and capable of processing very large data sets.

About The Bloor Group

The Bloor Group is a consulting, research and technology analysis firm that focuses on open research and the use of modern media to gather knowledge and disseminate it to IT users. Visit both www.TheBloorGroup.com and www.TheVirtualCircle.com for more information.

❏ PO Box 200638❏ Austin, TX 78720 ❏ Tel: 512–524–3689 ❏

www.TheVirtualCircle.com

www.BloorGroup.com