SnT2017 T4.1-P03

Disclaimer
The views expressed on this poster are those of the author and do not necessarily reflect the view of the CTBTO

Business Intelligence Software as a Self-Service Data Analysis Platform for the CTBTO

E. Tomuta, Y. Kotselko, R. Le Bras, T. Edwald
CTBTO

CTBTO PREPARATORY COMMISSION

Introduction

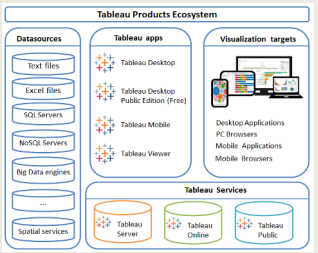
Business Intelligence (BI), as a collection of software, strategies, processes and services that support decision making in an enterprise has received a lot of attention in recent years. Some BI technologies have been around for more than a decade. They are often packaged with larger relational database management systems and include On-line Analytical Processing (OLAP) capabilities as well as data warehouse solutions and supporting tools, such as Extract-Transform-Load (ETL) capabilities. The realm of data science includes Complex Event Processing and advanced data analytics that make machine learning algorithms and tools (clustering, decision trees, neural networks, etc.) available for main-stream use. Advanced reporting capabilities, another BI component, are an essential ingredient in self-service capabilities, the ability for end-users to perform relatively sophisticated data analysis operations, using intuitive, interactive tools, without the involvement of IT specialists or data analysts.

This poster explores the applicability of Tableau Software, a commercial BI tool, to two specific use cases in the area of performance monitoring for station and network processing, focusing, in particular, on the creation and use of dashboards, self-service and interactive visual analysis capabilities. While we do not argue for the adoption of Tableau as the BI tool for the CTBTO, the use cases chosen reveal some required capabilities in the are of interactive visualization and data exploration that several existing BI tools provide, with some variations. Thus they can form the basis for a wider evaluation of such tools as part of identifying an organization-wide BI platform. We summarize the lessons we learned about Tableau Software by exploring the two use cases, identifying shortcomings of this software, as well as features that turned put to be particularly useful for the interactive data analysis tasks at hand. We conclude with a brief discussion of other commercial and open source BI toolsets that may be alternatives to Tableau or complement it in specific functionality areas, such as SiSense, PowerBI, Bokeh, Plotly or Grafana. The poster does not address the more general field of "data science", encompassing tools that support machine learning, classification, cluster analysis, data mining, etc. In addition to commercial tools like KNIME, RapidMiner, SAS, IBM SPSS, Alteryx, open source platforms based on Python and R are strong contenders in this field.

Tableau Suite of Products

The Tableau suite of products for the enterprise is centered around *Tableau Server*, the main platform through which interactive worksheets (charts) and dashboards are published. *Tableau Desktop* is the tool used to author such dashboards. This entails specifying data sources, defining their relationships, creating interactive charts and grouping charts into dashboards. Sophisticated filtering, highlighting, grouping functionality can be used in Tableau Desktop to synchronize data in separate interactive charts and to allow the end user to discover relationships within the data. This can largely be done interactively without the need to write code. A wide range of data sources is supported, including relational databases, various types of structured text files, excel spreadsheets, XML and JSON files, Hadoop and others. Interactive charts supported out of the box include: bar charts, line and area charts, pie charts, scatter plots, histograms, box plots, Gantt charts, tree maps, word clouds. Data values can also be visualized in spreadsheet format using Tableau's text tables/cross-charts features. Geographic data can be used to create various types of maps, including proportional symbol, choropleth, point distribution maps, flow maps, spider maps.
Charts produced using Tableau Desktop and published on Tableau Server can be explored using end-user tools such as Tableau Reader, Tableau Mobile (for mobile clients), a web browser interface, and, of course, Tableau Desktop.
More recently, Tableau also offers a cloud solution, via *Tableau Online*, which can be used to store interactive dashboards and charts as an alternative to Tableau Server. This however, also requires that the input data for such interactive charts be stored on Tableau Online. Charts stored on Tableau Online or Tableau Server can be embedded in web pages, thus providing analytics capabilities to existing web applications. Conversely, Tableau Desktop offers the possibility to refer through a URL to an information resource accessible via HTTP and to display this information when the user executes a specified action on a data point or data set.

The visualizations shown in this poster were created using Tableau Desktop Public Edition, a freely available, reduced-functionality version of Tableau Desktop whose output can only be stored on a public Tableau server.



Use Case 1: Visualizing Data on Waveform Quality Control Masks and Correlating it with Station Performance Measures

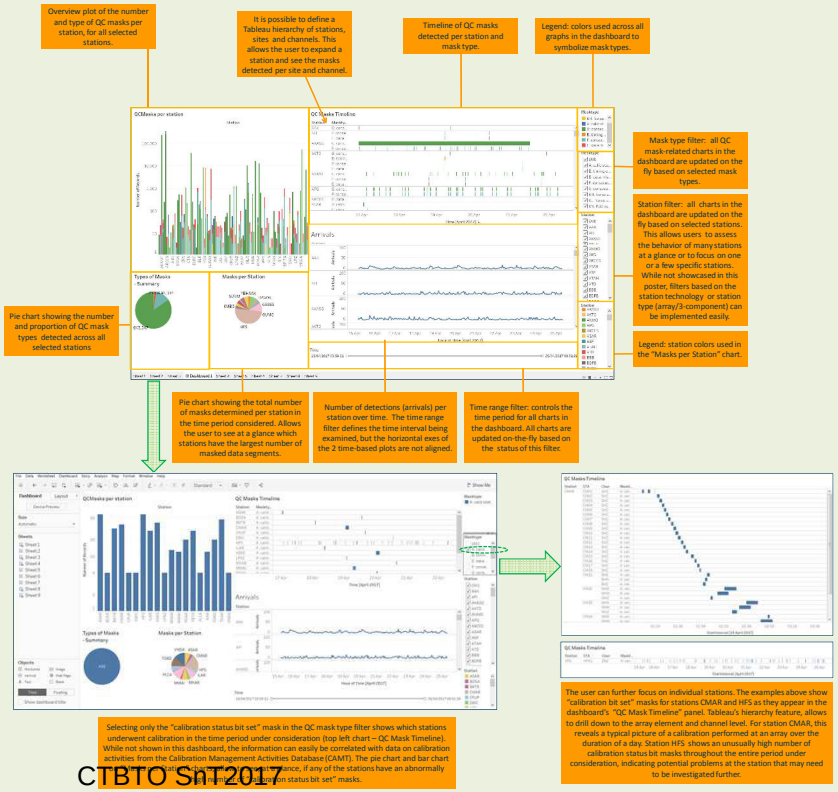
This use case explores how a BI tool like Tableau can be used to allow CTBTO performance officers and station maintenance officers to view data for waveform quality control (QC) masks at various stations over a specific period of time and to drill-down and analyse features of interest observed in the data. Determining waveform segments with quality problems is performed as a first step of IDC automatic processing of SHI data. As of November 2015, information about the types of quality problems detected and the waveform segments on which they were detected is stored in IDC databases in the form of QC masks. The types of QC masks being determined and stored include: calibration status bit set in the CD (Continuous Data) stream, continuous constant values (long and short segments), data with timing errors (clock differential determined to be too large during acquisition, and 3 types of masks related to station noise. The latter indicate that the noise is below or, respectively above the normal levels expected at the station based on Power Spectral Density (PSD) estimates using Welch's algorithm.

There is no software that allows users to visualize qc mask data at present.

The Tableau dashboard that supports this use case, brings together the following elements:

- Number and type of masks detected over a particular period of time at an SHI station.
- Total number of masks detected across all stations during the time period under consideration.
- Timeline of masks detected per station and channel.
- Number of detections at a station, that might correlate with information provided by QC masks.

The screenshots below illustrate the elements of the dashboard as well as various mechanisms to filter data, drill down to higher levels of detail and synchronize datasets used across several charts in the dashboard.



CTBTO SnT2017

Use Case 2: Determining Potentially Missed Associations in Network Performance Monitoring

The aim of this use case is to explore how Tableau can be used to visualize and characterize potentially missed associations in automatic processing and interactive review. This can be done by considering all unassociated arrivals at a given station within a specific time period and considering events to which they might have been associated based on time only (i.e. requiring that the time of the arrival is within 3 hours of a potential event). We then produce a scatter plot that shows the time difference and distance between each such unassociated arrival and a potential event to which it could be associated. In a typical case, such a plot will have no discernable structure, except for vertical lines (of constant distance) that correspond to different unassociated arrivals being matched to the same candidate event. However, the presence in the scatter plot of transversal linear features that match travel time curves indicate potentially missed associations.
The Tableau dashboard that supports this use case is shown below and consists in the following charts:

- The scatter plot mentioned above. We are displaying unassociated and associated arrivals in different colors. The associated arrivals help visualize the travel time curves we are looking for.
- A geographic map showing locations of candidate events for selected unassociated arrivals.
- A polar plot showing azimuth and slowness characteristics of arrivals selected in the main scatter plot.

We found that features such as zooming, sophisticated selection capabilities (circle, lasso, rectangle, adding to the selection), the ability to switch associated arrivals on and off in the scatter plot were particularly useful for exploring the dataset under consideration. The capability to filter the data shown in several charts on the dashboard by the selection in a source chart is a particular powerful feature. For this particular use case, it helped study characteristics of clusters of arrivals seen on the scatter plot in terms of their azimuth and slowness and the events to which they could potentially be associated.

Lessons learned from exploring the two use cases

- As showcased in the two use cases presented, Tableau provides interactive analysis and data exploration capabilities that may be very valuable for CTBTO performance monitoring purposes. While many tools exist that allow users to create plots of the types shown in this poster, BI software packages like Tableau bring additional *interactive visual analysis* capabilities that help correlate data in different views, similar to what DTK-DIVA offers.
- Tableau is being marketed as a self-service data analytics tool, with the implication that end users can create their own visualizations without software development skills. We found this to be partially true. Tableau's interface allows creating sophisticated dashboards such as the ones shown in this poster, without writing almost any code, however an in-depth understanding of relational database concepts and database schemas, scripting, user interface design (layout containers, etc.) are required in order to create effective dashboards. In particular for scenarios like the ones we explored, where the database schemas holding the desired information are relatively complex, it seems more realistic to view authorship of dashboards as an IT function, while end-users with minimal or no Tableau training would be the consumers of these visualization tools.
- Tableau visualizations are targeted at business users, and many chart types common in science and engineering (polar plots, surface or contour plots, geographic heat maps) are difficult or impossible to create in Tableau. The capability to associate and display external information referable via a URL to a data point in a Tableau chart helps to partially overcome this shortcoming. In the dashboard created for Use Case 1, it would, for instance, be possible to display spectral density plots for time periods during which abnormal noise characteristics were determined, or even to allow the user to visualize the masked waveform segments themselves.
- The capability to interact with Tableau dashboards through a web interface and to embed Tableau dashboards in other web applications seems particularly valuable, since it would enable us to add analytics capabilities to existing and new web applications such as the IDC Secure Web Portal, Performance Reporting Tool (PRTool), Incident Reporting System (IRS2) or MuTIP (Multi-technology Integration Portal).
- Much of the data the CTBTO is concerned with in the context of performance monitoring is time series data. While Tableau supports visualization of such data, other tools such as Grafana, provide more advanced features like aligning multiple time-line based plots and defining cursors across several such plots.
- For the two use cases shown on this poster, and for IDC performance monitoring in general, datasets of several hundreds thousands of records or even millions of records are common. Response times of the free version of Tableau Desktop used to create this poster were often quite slow and techniques to improve performance for such datasets need to be further investigated, especially given Tableau's claim to support big data analytics.

Other systems that could be considered as a BI platform for the CTBTO

The Business Intelligence and Data Analytics field is rapidly evolving. Among the commercial tools, Tableau, Microsoft's Power BI and Qlik's QlikView are named by Gartner as Leaders in its "Magic Quadrant report for Business Intelligence tools for 2017. A company that has seen a lot of momentum in the last few years is SiSense, which, in addition to front-ends for creating and consuming interactive dashboards, also provides a back-end that allows non-technical users to join and analyze large datasets from multiple sources.
An interesting alternative to leading commercial solutions, are open source tools like the Python-based Bokeh, a platform often used in conjunction with the Pandas Phyton Data Analysis Library, to create browser-based interactive visualizations, or Plotly, a collaborative browser-based analytics and visualization platform that integrates with Python, MATLAB or R. Both Bokeh and Plotly can produce interactive plots for Jupyter notebooks, a python-based technology, that provides browser-based rich documents, that combine text with embedded interpreted code and graphics.
Compared to commercial software, platforms such as Plotly and Bokeh offer a richer set of charts in particular for science and engineering (contour plots, polar plots, 3D surface and line plots). They are generally are aimed at end users familiar with a scripting language and data analysis library such as Python (numpy) or R. A somewhat more niche open-source analytics and visualization tool is Grafana, a tool that was at least originally aimed at system performance monitoring. While supported chart types are more restricted than, for instance, those available in Plotly or Bokeh, Grafana has advanced capabilities for log file analysis and for visualization and analysis of time-series and event data, that are quite relevant for the CTBTO.

Summary and Next Steps

We showcased how Tableau software can be used to support interactive data analysis for performance monitoring in the area of waveform processing (station and network processing). Despite being limited in scope, we believe the use cases chosen are fairly representative of the requirements BI software must fulfill in order to effectively support IDC performance monitoring in terms of interactive visualization capabilities. We consider this work as an important first step towards an organization-wide effort to explore existing BI tools with a view to adopt one or a small set of tools for organization-wide use. In this context, several other potential applications of BI need to be considered, e.g system/application monitoring, optimizing activities required to support IMS facilities (see also poster T4.1-P14), business process monitoring and optimization. Requirements on integration capabilities with existing systems and usage modes (self-service for non-IT end users vs. data analytics platform that can integrate with existing applications) must also be considered.