Eckerson Group

Integrated Data Preparation Tools for Visual Discovery

Chor-Ching Fan
Senior Research Consultant, Eckerson Group
December 2015

Research Sponsored by Rocket

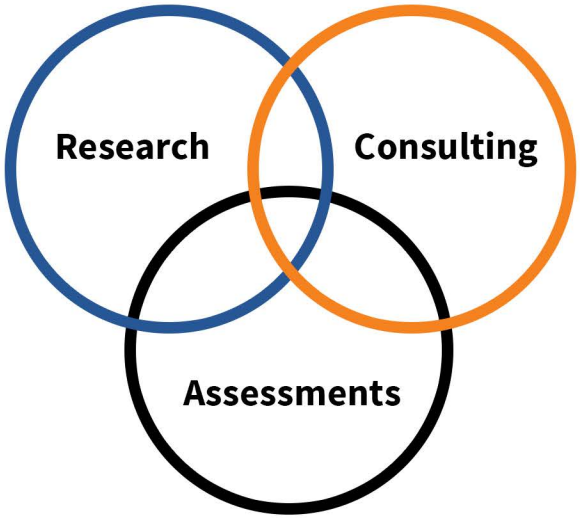
About the Author

Chor-Ching Fan is an IT and product management executive who has deep experience launching integration and analytics solutions with a focus on user and business impact. Chor-Ching is a principal consultant with Marketlocity and collaborates with the Eckerson Group to advance benefits from data management and analytics technology.

About Eckerson Group

Eckerson Group is a research and consulting firm of veteran practitioners who help business analytics leaders use data and technology to drive better insights and actions. Eckerson Group's researchers and consultants each have more than 20 years of experience in the field and are uniquely qualified to help business and technical leaders optimize their investments in business intelligence, analytics, big data management, and the Internet of things.

To explore our research and learn about our consulting services, visit **www.eckerson.com**.



Abstract

Visual discovery tools have played a key role in fueling the growth of self-service business intelligence (BI). But in order to generate insights without IT assistance, business users need to be able to access, integrate, and manipulate data from various data sources. BI vendors are now offering lightweight data integration functionality that replaces much of the data preparation work analysts formerly did in Microsoft Excel or Access. Some of these data preparation tools stand alone, but others are embedded in visual discovery products. This paper evaluates integrated data preparation tools, identifying their key features and characteristics for organizations seeking to empower business analysts and promote self-service BI.

Introduction

Visual discovery tools promise to empower business users with self-service access to data and insights, including the ability to analyze data, visualize the results, and publish their findings. Although these tools excel at visualizing data, most don't have rich features for preparing data for analysis. Instead, analysts normally need to source curated data from a data warehouse or data mart, then use Excel or some other tool to manipulate and integrate the data before analyzing it.

To address this gap in functionality, BI vendors are now offering a new class of products that enable business analysts to access, profile, clean, transform, combine, and format data for analysis. Data preparation tools come in two flavors: (1) Pure-play or standalone data preparation tools, and (2) integrated data preparation tools that are embedded in visual discovery or BI tools.

Standalone Tools. Standalone data preparation tools provide rich and comprehensive data preparation functionality for data-savvy users such as veteran business analysts, data scientists, BI developers, and data engineers. Some standalone data preparation tools run on Hadoop, making it easy for users to manipulate large volumes of both structured and unstructured data. Given their rich functionality, scalable platforms, and support for multi-structured data, standalone data preparation tools are often used as a substitute for traditional data integration tools or to support data science teams.

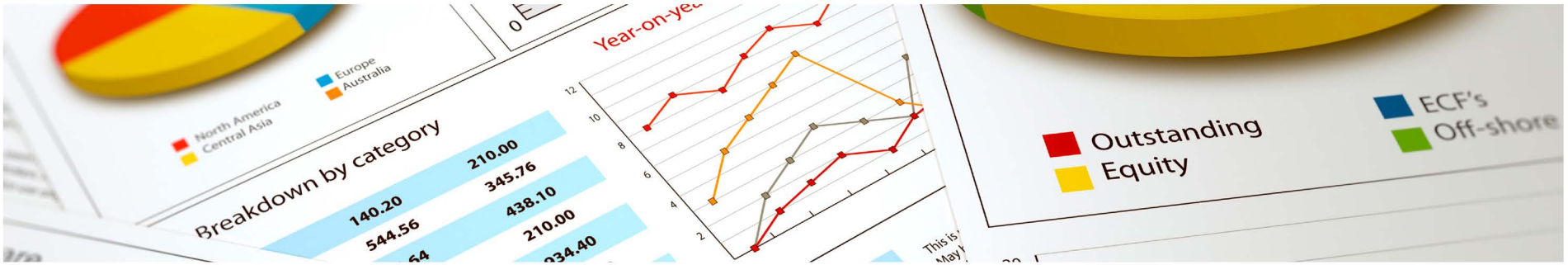
Integrated Tools. This paper focuses on integrated data preparation tools, which are bundled with visual discovery or business intelligence (BI) solutions. They are often presented seamlessly as a set of functional modules within the BI tool set. Compared with standalone offerings, integrated data preparation tools are geared to a less technical audience, including business managers, users, and newly minted analysts. These tools are easier to use than standalone tools; they have an intuitive graphical user interface, provide one-click access to host discovery tools, and make liberal use of embedded intelligence to automate many of the tasks involved in preparing data for analysis

Integrated Data Preparation

Benefits. Given their emphasis on ease of use, integrated data preparation promises to broaden the roles of users who can conduct self-service analyses and generate new insights. Integrated data preparation also safeguards a company's investment in visual discovery tools, ensuring that business users can effectively use them to generate new insights without IT assistance. Perhaps most important, it eliminates the need for an organization to purchase additional tools to support self-service business users.

Challenges. It's challenging to strike the right balance between ease of use and robust functionality. Integrated data preparation must deliver enough functionality to address common data integration tasks, but not so many features that average users are overwhelmed. Additional functions must be exposed on demand as users are ready and able to employ them; few business users want to hit a functionality wall when solving an urgent business problem. Finally, like any BI tool, integrated data preparation must provide governance features that prevent users from creating data silos that undermine enterprise information consistency.

When integrated data preparation achieves the right balance between ease of use, functionality, and governance, companies can make huge leaps toward achieving the goal of self-service BI.



Seven Characteristics

Integrated data preparation tools are just emerging, but a common set of attributes generally distinguishes them as tools designed for self-service business users. Certainly, the defining feature of integrated data preparation tools is that they are embedded in business intelligence and data discovery products. Although there is considerable overlap with stand-alone tools, integrated data preparation tools generally exhibit following seven characteristics. This paper uses Rocket Software's Discover product to illustrate many of these attributes.

1 Focus on Core Functions

Integrated data preparation adheres to the 80/20 rule. That is, its graphical user interface highlights the most commonly used data preparation functions and de-emphasizes (or hides) the others. The functions support a typical data preparation workflow in which analysts connect to a data source, profile the data set, combine it with another, change headings, parse strings, create hierarchies and aggregates, filter rows and columns, apply calculations, and publish the results. Other, less-used functions are available in drop-down menus or a built-in scripting language.

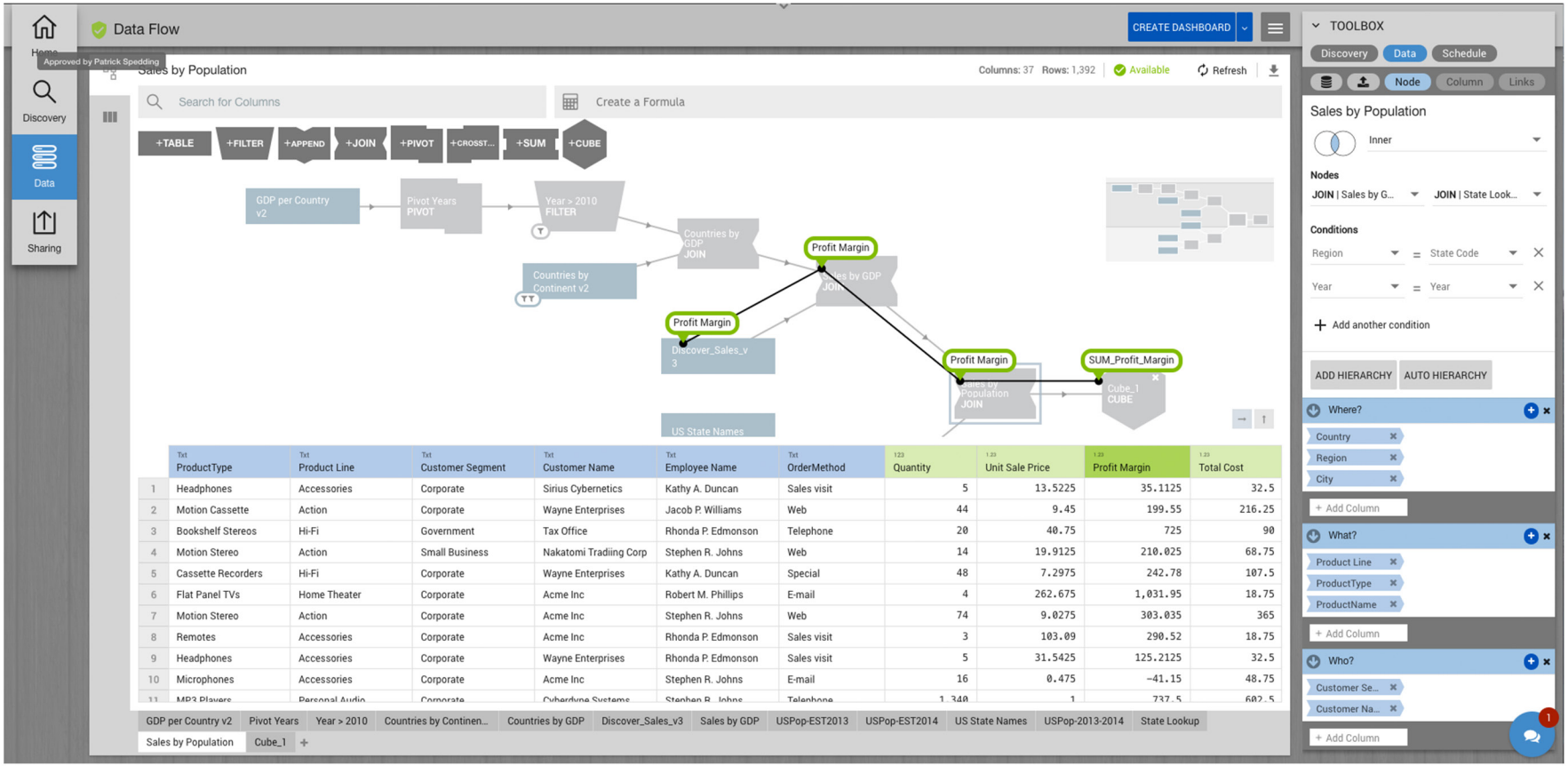
Business users commonly need to perform the following functions to prepare data for analysis in a visual discovery tool:

- **Connect** to a data source and import data into a prep tool.
- **Append** records to an existing data set, such as adding sales transaction data from Region X to Region Y.
- **Join** columns of two different data sources through a common column.
- **Filter** records based on the value in one or more columns.
- **Create** custom groups and segments.
- **Sum** the total value of a numeric field for a given set of records, such as calculate the total number of concert attendees for a given artist in 2015.
- **Export** results to a specific target, such as Excel, an OLAP cube, or an application programming interface.
- **Pivot** data values from rows to columns, or vice versa.
- **Parse** or clean data by removing repeated text or by concatenating fields.
- **Calculate** data using custom functions, such as applying a mathematical expression to a data set to create a new column.

2 Intuitive Graphical Interface

Integrated data preparation tools have an intuitive graphical user interface (GUI) that shows all steps in the assembly of a data set and displays the data lineage. (See figure 1 below.) The GUI also enables users to switch between data preparation and visualization screens, enabling them to visualize and then quickly modify data sets if necessary.

Figure 1. Graphical User Interface



The visual representation of steps in the assembly of a custom data set using data preparation functionality. On the right are configuration properties for a highlighted step. At the bottom of the screen is a preview of data produced by a highlighted step in the data flow. Source: Rocket Software.

The ideal interface for intergrated data preparation supports the following capabilities:

- **Visual data flows.** Users can create and view the steps in the assembly of a data output. They can click on any step in the visual flow to view or edit the details and view data outputs, simplifying navigation and increasing user productivity. Data flows become complex with conditional logic and multiple paths and sub-flows.
- **Point-and-click configuration.** Users can configure each step by pointing and clicking rather than coding. Steps might specify tasks that join, format, sort, or calculate data.
- **Preview pane.** A preview pane displays data generated by each step in the data flow. This helps users ascertain at a glance whether each step generates the correct data output.
- **Data lineage.** The data flow displays the lineage of a data set from source to target. In some tools, analysts can highlight a data element and trace its origins during the assembly process. (See figure 1.)
- **Zoom and pan.** A zoom and pan function helps users isolate a single part of a complex workflow so they can more easily analyze and edit it.

3 Embedded Intelligence

Embedded intelligence uses human-built rules and algorithms to make or suggest changes to data, simplifying usage and increasing productivity. For instance, a data preparation tool with embedded intelligence might automatically recommend join paths, hierarchies, parsing functions, or time-series aggregations. Business users can accept or modify the recommended actions.

Some of the common applications of embedded intelligence are:

- **Data typing.** Identifies date and text fields, geographical coordinates, dimensions, metrics, and table identifiers, and classifies them accordingly.
- **Auto join.** Joins two or more tables based on common data values in different columns.
- **Auto hierarchies.** Suggests data hierarchies based on the structure of the data.
- **Auto appends.** Adds new rows of similarly structured data to an existing data set.
- **Auto aggregation.** Aggregates data based on hierarchies or other groupings.

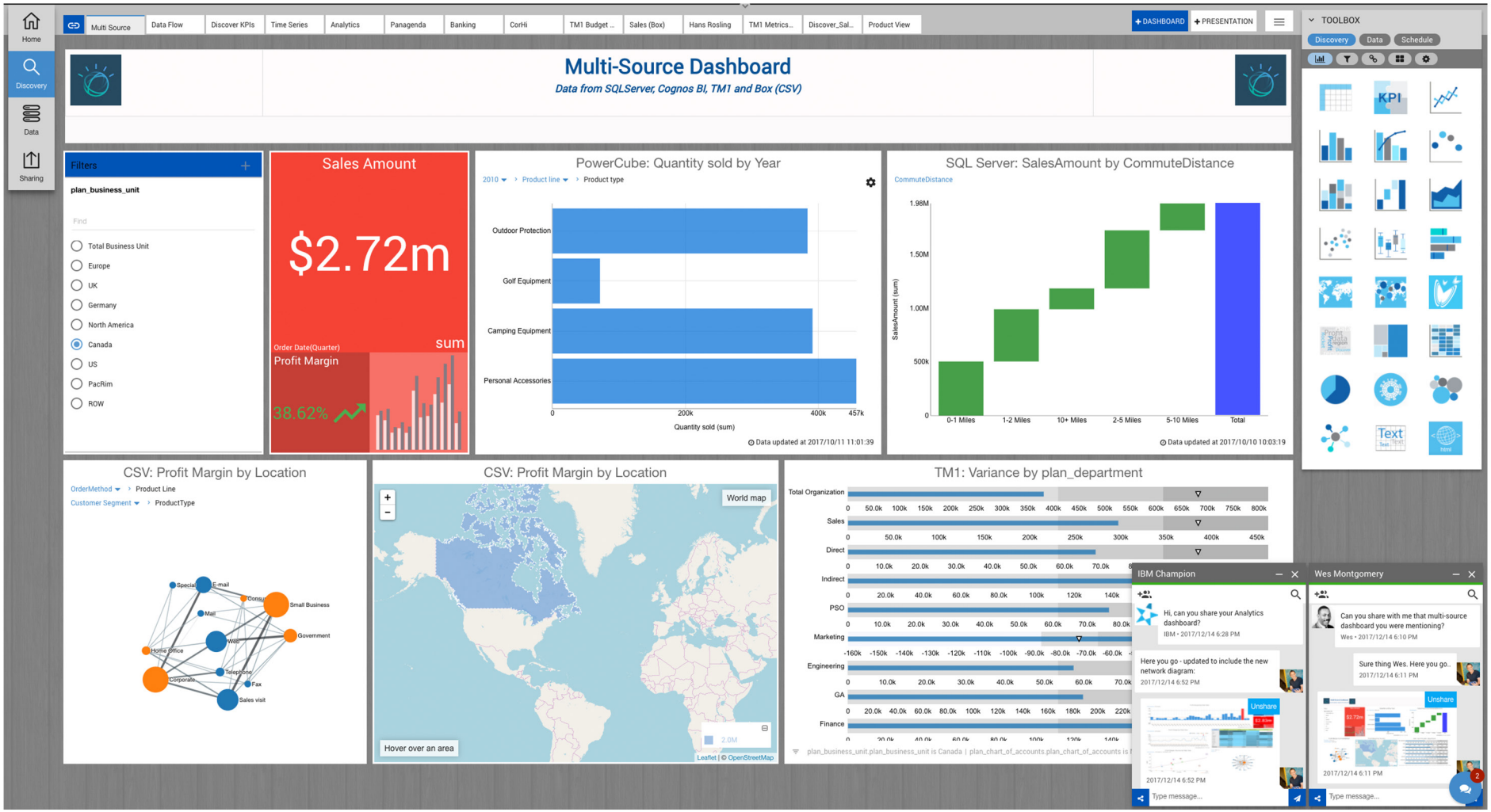
4 Tight Integration with Visual Discovery

By definition, integrated data preparation tools are embedded within visual discovery or BI products, enabling users to move seamlessly between the two so they can shape and visualize data in an iterative approach.

For example, an analyst evaluating the effectiveness of a marketing campaign might move back and forth between visual discovery and data preparation functions to smooth out price changes, achieve parity on promotion periods, or add third-party weather data. Behind the scenes, the integrated tool set orchestrates the movement of data and parameters between views, facilitating one-click access.

Some BI tools go a step further and provide side-by-side displays of data preparation and visualization functions. For example, while looking at charts in a dashboard, users may be able to reconfigure the visualized data set, changing join types, filters, and dimensions. Conversely, while in a data preparation module, users can select charts and change visualization filters. (See figures 2 and 3.)

Figure 2. Integrating Data Preparation and Visual Discovery



With one click, data preparation users can toggle to a discovery module (see above) where they can visualize data and collaborate with colleagues.

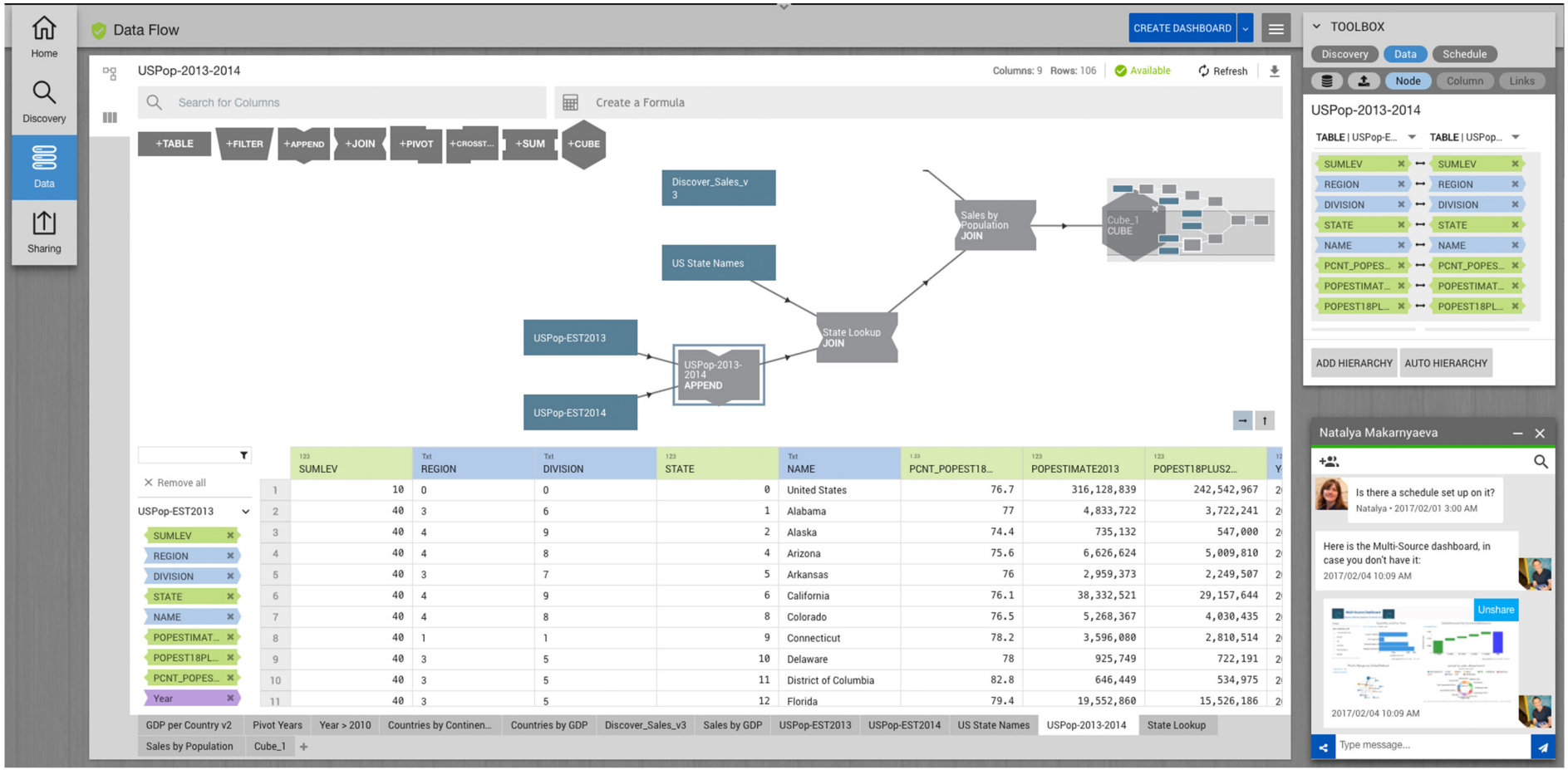
Key integration points between visual discovery and data preparation tools include:

- **One-click access.** During data preparation development, users can click once to visualize and analyze data, eliminating the need to export and import data. They can also click back to the integrated data preparation module to further modify the data.
- **Data profiling and lineage.** Visual discovery users can easily examine the attributes or lineage of a data set.
- **Visual prototyping.** Users can visualize data when creating new data flows to validate or refine the output.
- **Cross-functional visibility.** With some tools, users can view functionality from data preparation and visual discovery modules in side-by-side panels.

5 Collaboration

The best data preparation tools support collaboration and sharing among users. For example, threaded discussions and real-time chats make it easy for a team of users to discuss results and decide on a plan of action. And social-style collaboration—such as rating, tagging, commenting, and following—makes it easy for users to find the right data sets or visualizations and stay abreast of who is doing what types of analyses.

Figure 3. Collaboration Features



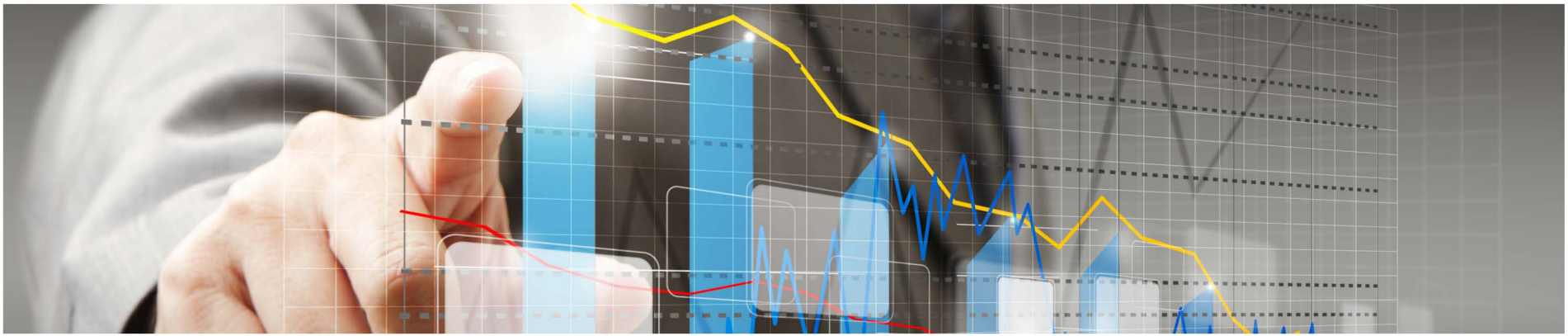
Threaded discussions and live chats assist users in assembling new data sets or to gain confidence in using an existing data set

Key collaboration attributes for integrated data preparation tools include:

- **Chats and discussion threads.** Threaded discussions and live chats are stored with the data and charts, keeping a history of conversations about data, insights, and actions. Besides reducing bureaucratic friction and aiding decision making, these features help new workers get up to speed quickly and understand the context for decisions.
- **Rate, rank, tag, and follow.** These social features enable business users to communicate their opinions about various data and visual assets as well as quickly find and evaluate the suitability of data or charts to answer business questions.
- **Publishing and centralized sharing.** Authors can share data sets and findings with colleagues, communicating insights broadly. Ideally, administrators set permissions up front that define who can publish what content to which types of users. (See "Governance," next.)

6 Governance

Integrated data preparation tools enable administrators to govern the publishing process and access control for data preparation and discovery output. Without adequate data governance, self-service data prepration can turn into BI chaos, creating a maelstrom of reports with contradictory metrics and ultimately undermining user confidence.

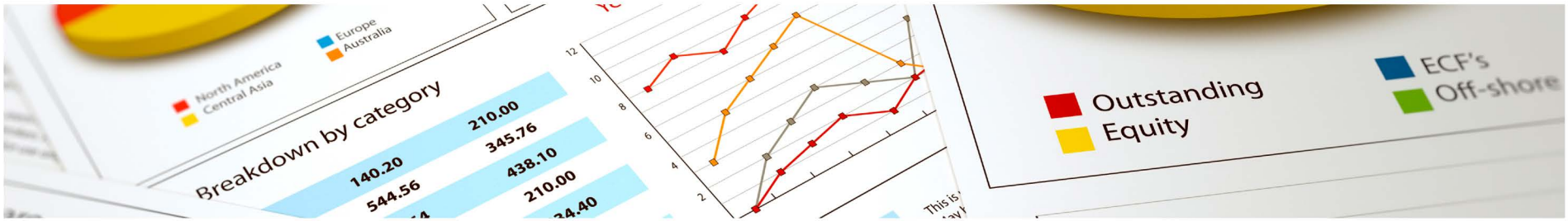


There are several important features for fostering a governed discovery environment:

- **Server-based administration.** Administrators can specify permissions for data access, publishing, and sharing, ensuring that the self-service output doesn't turn into data silos.
- **Directory services.** The tools should integrate with existing directory services, making it easy for administrators to establish permissions, set access rights, and define user groups.
- **Data governance policies.** Tools need to support multiple governance policies for controlling data output. Some organizations may want to vet all new output with a governance team, while others may define context-sensitive policies.
- **Auditing and logging.** Tools need to audit and log all user interactions with the data and tool to support troubleshooting, compliance, and auditing.
- **Usage monitoring.** Tools need to capture and display information about the use of data elements, queries, and reports. This enables administrators to predefine commonly generated data sets as well as remove duplicate efforts.
- **Lineage and impact analysis.** Data lineage helps business users gain confidence in the data by answering questions such as "Where did you get that data?" "What formula or statistical method did you use?" or "What filters did you apply?" (See figure 1.)

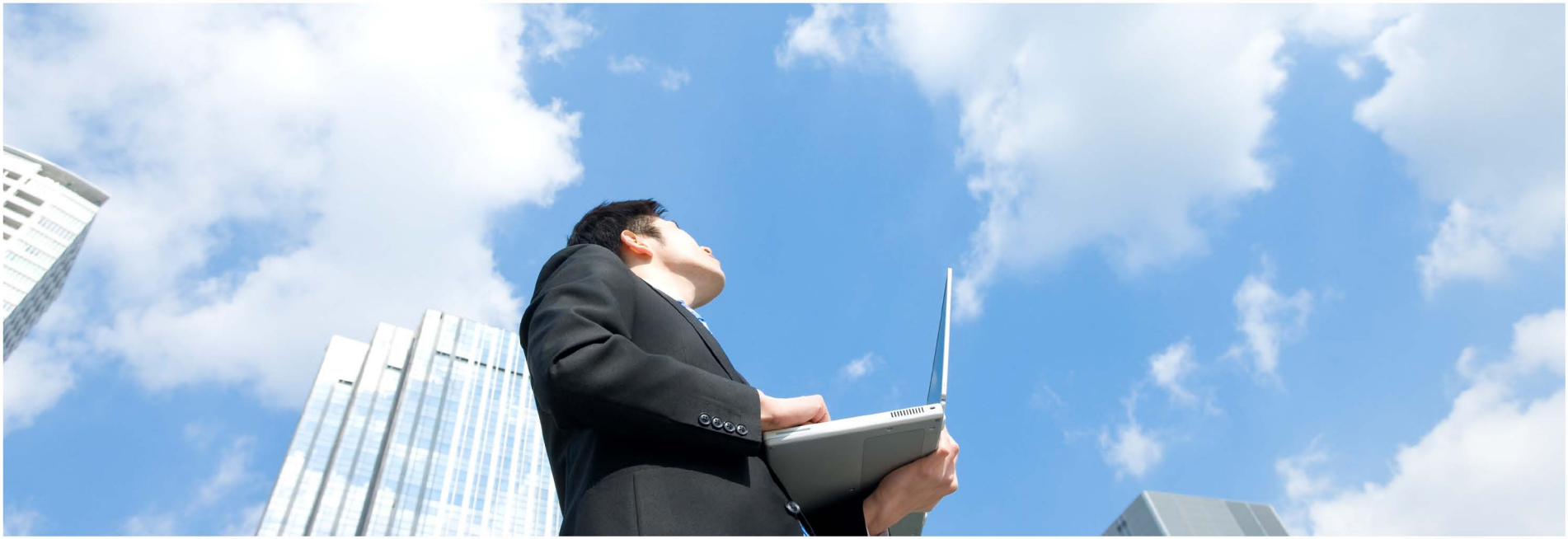
7 Integrated Support and Training

Training and support for integrated data preparation tools is geared to business users rather than professional data engineers or data scientists. The best training and support features are embedded within the tools, available on demand, and context sensitive. They should provide users with answers to questions as they arise without requiring them to switch to a separate screen or application.



Here are key attributes of effective training and support features for integrated data preperation tools:

- **Online content.** Tools need to provide one-click access to Web-based videos, tutorials, and other instructional material.
- **Context-sensitive help.** Rather than forcing users to search a generic knowledge base, context-sensitive help provides users with information about features or functions as they use them.
- **Access to experts.** Some tools take context-sensitive help to the next level and let users send instant messages to trainers, data owners, and designated experts.
- **Alerts.** Some tools support built-in messaging or ticker tapes that notify users about new or updated data sets or individuals who have subscribed to a user's data set, chart, report, or dashboard.
- **User community.** Users often learn best from each other, so many vendors now support online communities or forums where users can discuss tools, answer each other's questions, and share insights on best practices.



Summary

Data preparation functionality for visual discovery tools addresses a growing demand for self-service analysis among business users. Integrated data preparation tools target self-service business users (as compared to standalone solutions, which are typically geared toward data scientists and engineers). These tools focus on ease of use by providing an intuitive GUI, seamless integration with a host discovery tool, embedded intelligence to automate common data preparation tasks, and embedded collaboration and training features.

Many BI vendors are now shipping visual discovery tools with integrated data preparation functionality. This approach will broaden the population of business users who can create data sets without IT assistance and help make good on the promise of self-service BI. With adequate governance, integrated data preparation tools can transform the way organizations leverage data for decisions.