Google Fusion Tables: Web-Centered Data Management and Collaboration

Hector Gonzalez, Alon Y. Halevy, Christian S. Jensen*, Anno Langen,
Jayant Madhavan, Rebecca Shapley, Warren Shen, Jonathan Goldberg-Kidon†
Google Inc.

ABSTRACT

It has long been observed that database management systems focus on traditional business applications, and that few people use a database management system outside their workplace. Many have wondered what it will take to enable the use of data management technology by a broader class of users and for a much wider range of applications.

Google Fusion Tables represents an initial answer to the question of how data management functionality that focussed on enabling new users and applications would look in today's computing environment. This paper characterizes such users and applications and highlights the resulting principles, such as seamless Web integration, emphasis on ease of use, and incentives for data sharing, that underlie the design of Fusion Tables. We describe key novel features, such as the support for data acquisition, collaboration, visualization, and web-publishing.

Categories and Subject Descriptors: H.3.5 Online Information Services: [Data sharing, Web-based services]; H.2 Database Management: [Miscellaneous]

General Terms: Design

Keywords: Cloud Services, Visualization, Collaboration

1. INTRODUCTION

Given the sea-change in computing environments driven by the cloud, the Web, and the proliferation of connected, powerful personal computing devices, it is tempting to ask the following question: How would we design data management functionality for today's connected world? To put this question in context, recall that the foundations of database management systems were established several decades ago when the focus was on high-throughput business transactions, and the processing of complex SQL queries, and under the assumption that data typically belongs to a single enterprise.

While there will continue to be significant need for such systems, an increasing body of evidence points to drastically different requirements. First, data management needs to support collaboration among multiple users and multiple organizations at its very core. Second, data management systems need to appeal to a broader audience of users who are less technically skilled [5]. Third, data management and the Web need to be integrated seamlessly—data collection, presentation and visualization should be immediately compatible with the Web [7, 8].

This paper describes Google Fusion Tables, a cloud-based data management and integration service that aims to meet the aforementioned requirements. The Fusion Tables service was launched in June, 2009, and has since received considerable use (see *tables.googlelabs.com*). While we have witnessed a wide range of applications for Fusion Tables, the original audience, for which the service was designed, is organizations that are struggling with making their data available online (internally or externally), and communities of users that need to collaborate on data management across multiple enterprises and organizations.

Fusion Tables enables users to upload tabular data files (in Spreadsheet, CSV, and KML formats) of up to 100MB. The data can contain geographical objects such as points, lines and polygons. The system provides several ways of visualizing the data (charts, maps, timelines). The table can also be exported as KML so it can be viewed on Google Earth. The system supports filters and aggregates as a means of querying the data. Integration of data from multiple sources is supported by means of joins across tables that may belong to different users. Users can keep the data private, share it with a select set of collaborators, or make it public. The discussion feature of Fusion Tables allows collaborators to post or respond to comments at the level of individual rows, columns, or cells. Users can interact with data through a Web interface or through programs that access the data through an API. The features we currently support represent an initial subset of new and traditional data management functionality that were deemed to be of the highest priority.

This paper describes the design goals of Fusion Tables and the functionality we built in support of this design. A companion paper [4] provides details of the underlying architecture and our implementation.

We begin by outlining our design goals and the principles underlying the design of Fusion Tables. Section 3 describes

*On leave from Aalborg University.

†On leave from M.I.T.

SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA.

selected features that address our goals in the context of data acquisition, sharing, collaboration and visualization. Section 4 describes our API. Finally, Section 5 covers related work, and Section 6 concludes.

2. DESIGN FOUNDATIONS

Fusion Tables is designed with new applications in mind and according to a set of guiding principles that we consider important to enabling the intended applications.

2.1 New Applications

The goal of Fusion Tables is not to replace traditional database management systems and applications, and neither is the goal to simply move such applications into the cloud. In contrast, the objective is to offer data management functionality that exploits today's computing environment in order to effectively enable new users and uses of data management technology.

The following are example applications for which Fusion Tables was designed. Each application either mirrors or is inspired by actual, ongoing uses of Fusion Tables.

- Ecologists in the rain forests of Costa Rica collect specimens of animal and plant life. They want to maintain records of the specimens and also want to include the related genetic information that is being produced for them by a laboratory in Canada.
- A non-profit that wants to publish datasets about the availability and usage of water resources. They would like to use visualizations as a means of painting a compelling story about the dire state of water availability and quality around the planet (see *www.circleofblue.org* [2]).
- The Ministry of Health in an African country wants to obtain community input on the current status of health clinics dispersed across the country.
- The International Coffee Organization collects and distributes data about coffee exports and imports, and wants to make this data available to interested parties globally.
- An epidemiologist seeks to turn dry tables of numbers into a visual story, creating broader awareness more quickly, this way facilitating faster and more effective responses to disease trends.
- Congressional staffers want to visualize data to help a senator make an argument.
- The administrators of a web site managing a database of biking trails want to publish the contents of their database in such a way that they can programmatically manage their data while their users explore (browse and filter) the trails on a Map (see *www.mtbguru.com* [6]).
- A dairy farm in Brazil that is being managed by its owner in Thailand and his partner in California wants to enable data-based collaboration among the three sites.

2.2 Underlying Principles

Fusion Tables aims to adhere to a small set of guiding principles that we believe are important in enabling applications such as those just outlined. Subsequent sections describe how these principles are currently reflected in Fusion Tables. We note that in all cases, following these principles offers an agenda for a continuous process of improvements.

Provide Seamless Integration with the Web

In today's computing environment, access to the Internet can increasingly be taken for granted. It is therefore important that data management functionality is integrated seamlessly with the Web, and is able to leverage other properties and services on the Web. Other web properties may serve as entry points into data management as well as venues for publishing and visualizing data. Other services, e.g., geocoding of location names, can be used to add value to the data.

First, Fusion Tables allows users to publish their visualizations on the Web. We enable users to create bar charts, pie charts, timelines, motion charts, etc., and embed them on any web page. Especially popular are map visualizations that enable users to display geo-spatial datasets on Google Maps.

Second, public datasets on Fusion Tables can be crawled by search engines and hence have a chance of showing up as search results. The data is therefore automatically accessible through web search, the primary method for locating data on the Web.

Third, Google and others already have a powerful collaboration model for documents and spreadsheets, and Fusion Tables is designed to integrate seamlessly with such established models.

In a nutshell, we wanted the data management service to be an integral component of the eco-system of data on the Web.

Emphasize Ease of Use

A fundamental emphasis on ease of use is essential in reaching a much broader class of potential users with data management needs, but who are not IT professionals and who may not have any training in data management. In keeping with this principle, design decisions are made that prioritize ease of use over other requirements. One aspect of ensuring ease of use is to apply pay-as-you-go data management principles [3], the key idea being that a user should see an immediate benefit of investing time in using the data management functionality. As part of this, little initial investment should be required by the user.

For example, being a cloud-based service, Fusion Tables requires no initial installation. As another example, Fusion Tables does not require the user to declare a schema up front, but rather tries to automatically determine the data types of columns for common and useful data types.

Provide Incentives for Sharing Data

Users often desire to share data with others. However, they are faced with a number of disincentives. Data owners are afraid of loss of attribution, of misuse and corruption of their data, and of others not being able to find the data easily. Fusion Tables aims to address such concerns.

As already mentioned, when a user specifies that a certain dataset is public, we make that data crawlable by search engines, so it can appear in search results. Likewise, while all

datasets can be visualized in different ways on the Fusion Tables web site, only the public ones can have their visualizations embedded on web pages.

Facilitate Collaboration

Collaboration among multiple parties on the Web from different enterprises and organizations is a key to data management today. Valuable insight arises when data is combined from multiple sources and when data is scrutinized from the perspectives of multiple users.

Fusion Tables facilitates such joining of data and enables collaborators to discuss and comment on the data at several levels of granularity.

3. DATA MANAGEMENT WITH FUSION TABLES

We cover novel aspects of Fusion Tables, covering first the acquisition of data, then the support for collaboration and sharing, and finally the support for visualizations.

3.1 Data Acquisition

Fusion Tables enables users to upload files containing structured data. The currently supported formats include CSV (Comma Separated Values), different spreadsheet formats (Excel, Open Office, and Google Spreadsheets), and KML (Keyhole Markup Language).

To achieve ease of use, the number of steps a user needs to go through before the data is in the systems is reduced as much as possible. Rather than having the user declare a schema for the tabular data and then having the user also describe how the data to be imported matches the schema, the system tries to detect automatically which row in the imported file is the header row (i.e., specifying the names of the columns), and it simply asks the user to verify its guess. In addition, even though some processes (e.g., indexing, type guessing) are going on in the background, we try to maintain a responsive import process.

Further, the system does not ask the user to specify data types for the columns identified. Instead, as we describe shortly, it attempts to guess the types from the data (some of our users may not even understand the concept of schema versus instance or that of a data type). In keeping with pay-as-you-go principles, users can always specify data types if they so desire. The system also encourages the users to specify any other descriptive information about their data that may be useful to others.

3.2 Data Sharing and Collaboration

The data import step also addresses concerns that some users may have about sharing data. The typical concerns we hear from data owners involve (1) loss of control over their data once they upload and share it with others, (2) losing credit for creating the data, and (3) the possibility that others will use or interpret the data incorrectly.

Attribution and export: Fusion Tables provides several features to address these concerns. First, users can finely control who they share the data with. Second, users can specify an attribution for the data that is *always* carried around with the data, no matter what transformation gets applied to the data. Third, while the original owners of the data can always export it outside Fusion Tables (e.g., to save a local copy), they can restrict the ability of other users to do so.

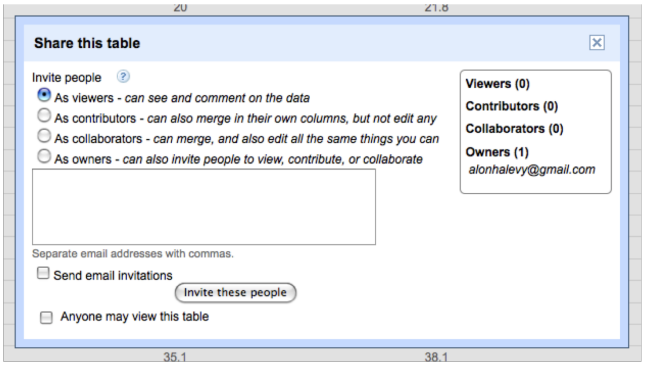

Figure 1: Collaborating with others. In addition to specifying collaborators with read or write permissions, Fusion Tables also allows collaborators who can add columns to an existing dataset, but still keeping distinct write permissions.

Search: Next, making data public in order to share it with others is useless unless the data can easily be discovered by interested users. Thus, rendering the data in Fusion Tables searchable is an important ongoing effort. Our main goal is to make public data discoverable by search engines, so they can direct users to the data when this is relevant to their queries. Whenever a user makes a dataset public in Fusion Tables, we create a corresponding HTML page that is crawlable by search engines. As such, some queries will obtain these tables as results.

It is also important to enable discovery of data inside Fusion Tables. Thus, we are pursuing efforts to enable an advanced search for tables from *within* Fusion Tables. This service will be known to a much narrower set of users (as is often the case for specialized search engines); it is meant to support those users who have a specific need to explore the collection of structured data. Searching over a repository of tables is neither a simple nor a solved problem. The typical signals that are used for document retrieval may not apply in this context [1]. Our search service is based on an extension of the techniques developed by Cafarella et al. [1].

Sharing and integration: As a step towards supporting collaboration, Fusion Tables enables users to explicitly and easily share data with one another, even if the users are not in the same organization; and it enables users to merge data from multiple owners.

Fusion Tables follows the typical model of document sharing in the cloud. A user can decide to invite a set of collaborators to either view a table or provide them update access as well. In addition, a user can decide to make a dataset public, which enables anyone to view and comment on it.

The basic collaboration model is extended with the ability to invite *contributors* (see Figure 1), who can contribute their own columns to a table, with the different owners of columns maintaining write conditions on their own columns.

For example, in the scenario of the ecologists in Costa Rica, the ecologists contribute the columns about the specimens, and the lab in Canada contributes the columns with the genetic information. This yields a shared, jointly owned dataset that they can explore together. This scenario illus-

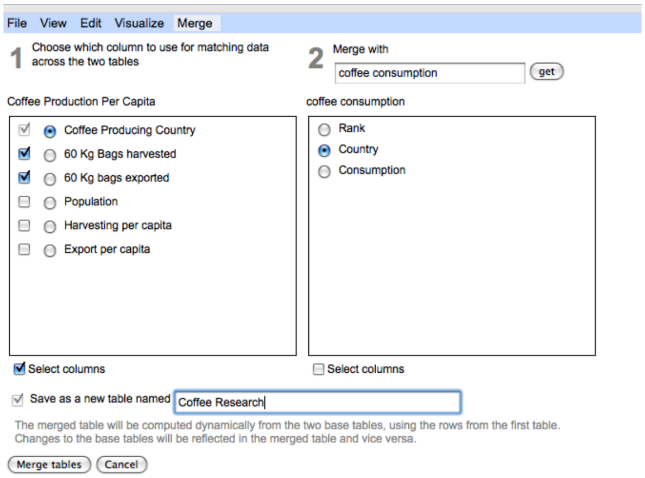

Figure 2: Merging data from multiple sources. Tables belonging to different owners can be merged by a join on a column containing values from the same domain.

Figure 3: Displaying an intensity map after importing data about coffee consumption. Fusion Tables detects a geo-location column in the data and offers map visualizations.

trates the ability to merge data from multiple tables with different owners. Despite the progress on data integration tools, such sharing of data across multiple enterprises is still very hard in practice. In Fusion Tables, we decided to begin by supporting the simplest kind of data integration possible. When multiple parties have data about the *same* entities, the first step in initiating a collaboration is to be able to see the data side by side. Hence, Fusion Tables supports a Merge operation that essentially performs a join on a key column that is shared by two (or more) sources. For example, Figure 2 shows how we can merge data about coffee production and coffee consumption on a common key (country). The tables may come from two different sources. In the case of the ecologists in Costa Rica and the lab in Canada, they would collaborate by merging their tables on the Specimen ID column.

Discussions: Sharing with others or integrating data from multiple sources typically just represents an intermediate step in the lifecycle of the data, where the different collaborators want to study the data.

Fusion Tables offers a discussion feature that supports in-depth collaboration. Discussions may point out outliers, may detect incorrect data, or may question the underlying assumptions and semantics of the data. Specifically, Fusion Tables lets users post and respond to comments at several levels of granularity: rows, columns, and individual cells. We find that enabling discussions at all these levels of granularity is crucial for large datasets because it is otherwise hard to keep track of the discussions and of their specific contexts.

Discussions are handled in an append-only fashion so that new comments are simply appended to the existing comment trails. An interesting aspect of the discussion mechanism is that if a change is made to the value of a cell then the change also becomes part of the discussion trail on the cell. This preserves the context of the comments and enables documentation of the reasons for changing the value. When a user views a table, a discussion panel shows the user which parts of the table are being discussed actively.

We note that discussions are associated with a particular view of the data. Specifically, if a user creates two views V_1 and V_2 from a table T, then the discussions on V_1 will not be visible in V_2 and vice-versa nor will they be visible in T itself. The reason for this design is that the views V_1 and V_2 may be visible to different sets of users, and therefore some discussions may need to be hidden. In addition, discussions may serve different purposes, so even if they are visible by the same users, we may want to keep them apart.

3.3 Data Manipulation and Visualization

Once the data is imported, Fusion Tables enables users to explore their data with a combination of data visualization and SQL-like querying.

Fusion Tables makes it easy to visualize the data in different ways. In particular, if the system finds that a certain column has values that are mostly geographic locations, then it will offer the user several map viewing options, as exemplified in Figures 3, 4 and 5. Likewise, the presence of date and time columns enable timelines and motion charts, while numeric columns enable bar charts and pie charts.

The applicability of a particular visualization depends on the data types of the columns in a dataset. Thus, considerable attention has been given to recognizing when a column in a dataset is a set of locations that can be plotted on a map, or when it is a time point that can be shown on a time line.

We chose to focus on the geo-location and time data types because they by far overwhelm any other data types in terms of the opportunities they present for useful visualizations, and because they occur frequently in practice. Once a user chooses a type of visualization, the system uses the column data type information available to guess how to specifically visualize the data.

When a visualization has been created, the user can ask the system for an HTML snippet that can be pasted into another web property such as a blog. The visualization then

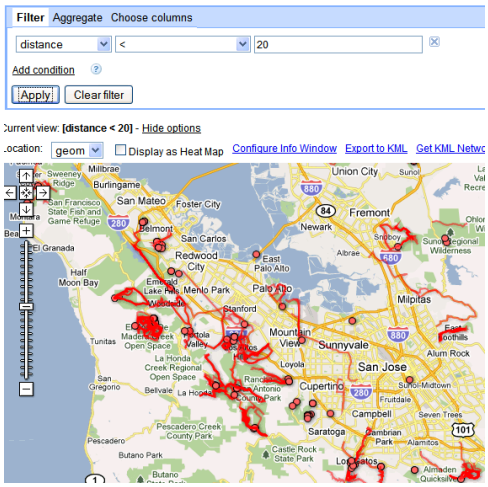

Figure 4: Visualization of all bike trails in the San Francisco Bay Area that are shorter than 20 miles.

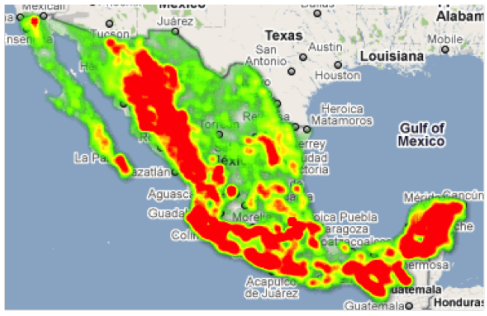

Figure 5: Heatmap for the forest cover in Mexico.

appears as a live gadget on that property, meaning that the visualization acts as a view on the underlying data and thus tracks the changes in the data. The user can also share a visualization with others by sending them a URL (called a *snap*) of the visualization.

The result is a very short path from data import to a useful visualization. The visualization feature with its simple task flow has been incredibly popular with our users.

A very popular component of Fusion Tables is the rendering of large geographic datasets. We allow users to upload tables with street addresses, points, lines, or polygons. We render these tables as map layers. The rendering is done on the server side, i.e., we send the client a collection of small images (tiles) that contain the rendered map. Figure 4 shows an example of rendering the bike trails in the San Francisco bay area that are shorter than 20 miles. Figure 5 shows an example of a heat map created by Fusion Tables, simply by selecting the option on the map menu.

We currently provide only the most common query facilities. We support common selection predicates, grouping and aggregation, and projecting out a subset of columns. We described briefly our join capability in Section 3.2. The query facilities will be expanded over time. Finally, we also have an API for inserting, deleting, and updating rows in a table.

4. FUSION TABLES API

An important aspect of being a platform for data management and collaboration is to provide developers with a way to extend the functionality of the site. We accomplish this through an API.

The API allows external developers to write applications that use Fusion Tables as a database. For example, the site *mtbguru.com*, has written an application that synchronizes their collection of bike routes with a table in Fusion Tables. The map visualization in Figure 4 was created over their dataset.

The API supports querying of data through select statements, update of the data through insert, delete, and update statements, and data definition through a create table statement. All access to data through the API is authenticated through pre-existing methods for all Google properties.

5. RELATED WORK

Fusion Tables is inspired in part by ManyEyes (manyeyes.com) that enables users to upload data and visualize it in several ways. We go further by providing data management capabilities and a sharing model that does not require always making your data public. We strive to preserve the ease of use provided by spreadsheets, but adapt it to larger datasets where the data and the presentation need not necessarily be one of the same. Several online database management tools exist (e.g., DabbleDB (dabbledb.com), Socrata (socrata.com), Factual (factual.com), but Fusion Tables focuses on the collaboration aspects of data management and handles larger datasets. In comparison to other products, Fusion Tables emphasizes the deep integration into a maps infrastructure that is proving immensely popular. Wolfram Alpha is a search engine for structured data, but our focus here is on enabling users to manager their own data.

There are also several project related to structured data at Google. Google Public Data is an effort to import public government data and provide high-quality and carefully-chosen visualizations of data in response search queries. For example, a query on "california unemployment rate" will yield a thumbnail of a graph with the data that the user can explore in more detail. The Google Squared Service lets users specify categories of objects (e.g., US Presidents, espresso machines) and explore attributes of these entity sets. In this case, the data populating the tables is automatically extracted from various sources on the Web, and may not always be accurate.

6. CONCLUSIONS

The goal of Fusion Tables is to enable a much larger class of users to manage their data and to do so in a way that is integrated with their other online activities. Fusion Tables is part of a bigger effort to encourage data owners to publish data on the Web and to make it easier for users to discover data that is relevant to their needs.

There are many obvious extensions that need to be made to to Fusion Tables, starting from providing more expressive data modeling and query capabilities and providing adequate performance on larger datasets. However, our strategy is to engage our users and prioritize their most acute needs. In particular, the API and the advanced management of geographical data were inspired by frequent user requests.

7. REFERENCES

[1] M. J. Cafarella, A. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. WebTables: Exploring the Power of Tables on the Web. In *VLDB*, 2008.

[2] Google Brings Water Data to Life. http://www.circleofblue.org/waternews/2009/world/google-brings-water-data-to-life/.

[3] M. Franklin, A. Halevy, and D. Maier. From Databases to Dataspaces: A New Abstraction for Information Management. *SIGMOD Record*, 34(4), 2005.

[4] H. Gonzalez, A. Halevy, C. Jensen, A. Langen, J. Madhavan, R. Shapley, and W. Shen. Google Fusion Tables: Data Management, Integration and Collaboration in the Cloud. In *SOCC*, 2010.

[5] H. V. Jagadish, A. Chapman, A. Elkiss, M. Jayapandian, Y. Li, A. Nandi, and C. Yu. Making Database Systems Usable. In *SIGMOD*, 2007.

[6] MTBGuru tracks as seen through Google Fusion Tables. http://blog.mtbguru.com/2010/02/24/mtbguru-tracks-as-seen-through-google-fusion-tables/.

[7] B. Shneiderman. Extreme Visualization: Squeezing a Billion Records into a Million Pixels. In *SIGMOD*, 2008.

[8] F. B. Viégas and M. Wattenberg. Transforming data access through public visualization. In *SIGMOD*, 2009.