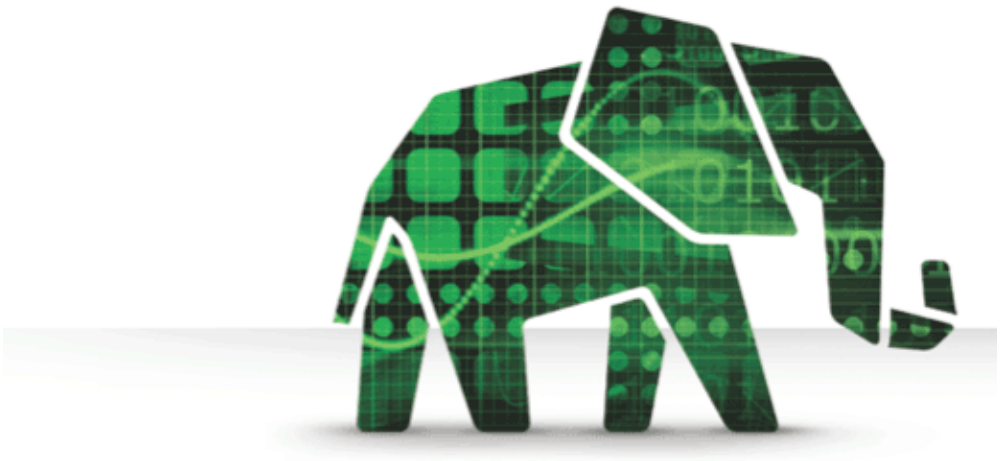
Hortonworks We do Hadoop.

Using Tableau Software with Hortonworks Data Platform

September 2013



http://www.hortonworks.com

Hortonworks We do Hadoop.

Modern businesses need to manage vast amounts of data, and in many cases they have accumulated this data for years. Many enterprises have built large-scale environments for transactional data with analytic databases, but are now inundated with new types of data, such as social media data, server logs, clickstream data, web logs, machine/sensor data, and geolocation data. These new data sources all share the common Big Data characteristics of volume (size), velocity (speed), and variety (type), and have sometimes been thought of as low value, or even as "exhaust data": too expensive to store and analyze.

It is these types of data that are turning the conversation from "data analytics" to "big data analytics." With Hortonworks, businesses are learning to see these types of data as inexpensive, accessible sources of insight and competitive advantage.

The Hortonworks Data Platform allows you to store, process, and manage data at scale. It is designed to integrate with and extend existing data applications. With Hortonworks, enterprises can retain and process more data, join new and existing data sets, and lower the cost of data analysis.

Tableau can be used with Hortonworks to explore this expanded data set. Tableau can access the data in the Hortonworks Data Platform, visualize that data, and provide valuable insights for your business. Tableau can also combine the data in the Hortonworks Data Platform with data in traditional analytics databases to create a blended view of multiple data sources.

The combined capabilities of Hortonworks and Tableau make Big Data less expensive, more accessible, and easier to understand and use for business advantage.

In the following sections, we will show you:

- The main features of the Hortonworks Data Platform and Tableau.
- Where Tableau fits in with the Hortonworks Data Platform as part of a modern data architecture.
- How you can use Tableau with Hortonworks for data exploration and visualization.

Hortonworks

3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks

Hortonworks We do Hadoop.

The Hortonworks Data Platform

The Hortonworks Data Platform (HDP) is an enterprise-grade, hardened Apache Hadoop distribution that enables you to store, process, and manage large data sets.

Apache Hadoop is an open-source software framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed for high-availability and fault-tolerance, and can scale from a single server up to thousands of machines.

The Hortonworks Data Platform combines the most useful and stable versions of Apache Hadoop and its related projects into a single tested and certified package. Hortonworks offers the latest innovations from the open source community, along with the testing and quality you expect from enterprise-quality software.

The Hortonworks Data Platform is designed to integrate with and extend the capabilities of your existing investments in data applications, tools, and processes. With Hortonworks, you can refine, analyze, and gain business insights from both structured and unstructured data – quickly, easily, and economically.

Hortonworks Data Platform: Key Features and Benefits

With the Hortonworks Data Platform, enterprises can retain and process more data, join new and existing data sets, and lower the cost of data analysis. Hortonworks enables enterprises to implement the following data management principles:

- **Retain as much data as possible.** Traditional data warehouses age, and over time will eventually store only summary data. Analyzing detailed records is often critical to uncovering useful business insights.

- **Join new and existing data sets.** Enterprises can build large-scale environments for transactional data with analytic databases, but these solutions are not always well suited to processing nontraditional data sets such as text, images, machine data, and online data. Hortonworks enables enterprises to incorporate both structured and unstructured data in one comprehensive data management system.

Hortonworks

3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks

- **Archive data at low cost.** It is not always clear what portion of stored data will be of value for future analysis. Therefore, it can be difficult to justify expensive processes to capture, cleanse, and store that data. Hadoop scales easily, so you can store years of data without much incremental cost, and find deeper patterns that your competitors may miss.

- **Access all data efficiently.** Data needs to be readily accessible. Apache Hadoop clusters can provide a low-cost solution for storing massive data sets while still making the information readily available. Hadoop is designed to efficiently scan all of the data, which is complimentary to databases that are efficient at finding subsets of data.

- **Apply basic data cleansing and data cataloging.** Categorize and label all data in Hadoop with enough descriptive information (metadata) to make sense of it later, and to enable integration with transactional databases and analytic tools. This greatly reduces the time and effort required to integrate with other data sets, and avoids a scenario in which valuable data is eventually rendered useless.

- **Integrate with existing platforms and applications.** Hortonworks connects seamlessly with many leading analytic, data integration, and database management tools.

Tableau

Tableau is a data analysis tool that can be used for data exploration and visualization. Tableau is designed to support people's natural tendency to think visually. Rather than typing data into forms or clicking through wizards, Tableau features an intuitive drag-and-drop interface. You can connect to data in a few clicks, then visualize and create interactive dashboards with a few more.

Traditional business intelligence (BI) platforms have required users to build elaborate "universes," "cubes," or "temporary tables" before any real work can be done. Tableau eliminates those steps completely. There's no requirement to pull data into a silo – you work directly from your database.

Hortonworks
3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks

Tableau includes a "Show Me" feature – a visualization best practices engine – that enables you to easily view your data using different visualizations, such as graphs, bar and pie charts, and map-based data representations. Tableau also enables you to share your visualizations on a secure server with colleagues, customers, and partners.

With Tableau you can connect directly to databases, cubes, data warehouses, files, spreadsheets, and Hadoop. Your connection is live, so you see up-to-the-minute data. It takes only a few clicks, and no programming is required. In minutes you'll be accessing data, consolidating numbers, and visualizing results without advance set-up. Tableau is true ad-hoc business analytics.

Reference Architecture

Traditional Enterprise Data Architecture

Today, nearly every enterprise already has some sort of database management system already in place. Generally, these environments are structured as follows:

- Data comes from a set of data sources – most typically from enterprise applications such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), and any custom applications used to gather data.

- That data is extracted, transformed, and loaded into a data system such as a Relational Database Management System (RDBMS), an Enterprise Data Warehouse (EDW), or even a Massively Parallel Processing (MPP) system.

- A set of analytical applications – either packaged (e.g. Tableau) or custom – then access the data in those systems to enable users to garner insights from the data.

Hortonworks
3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks

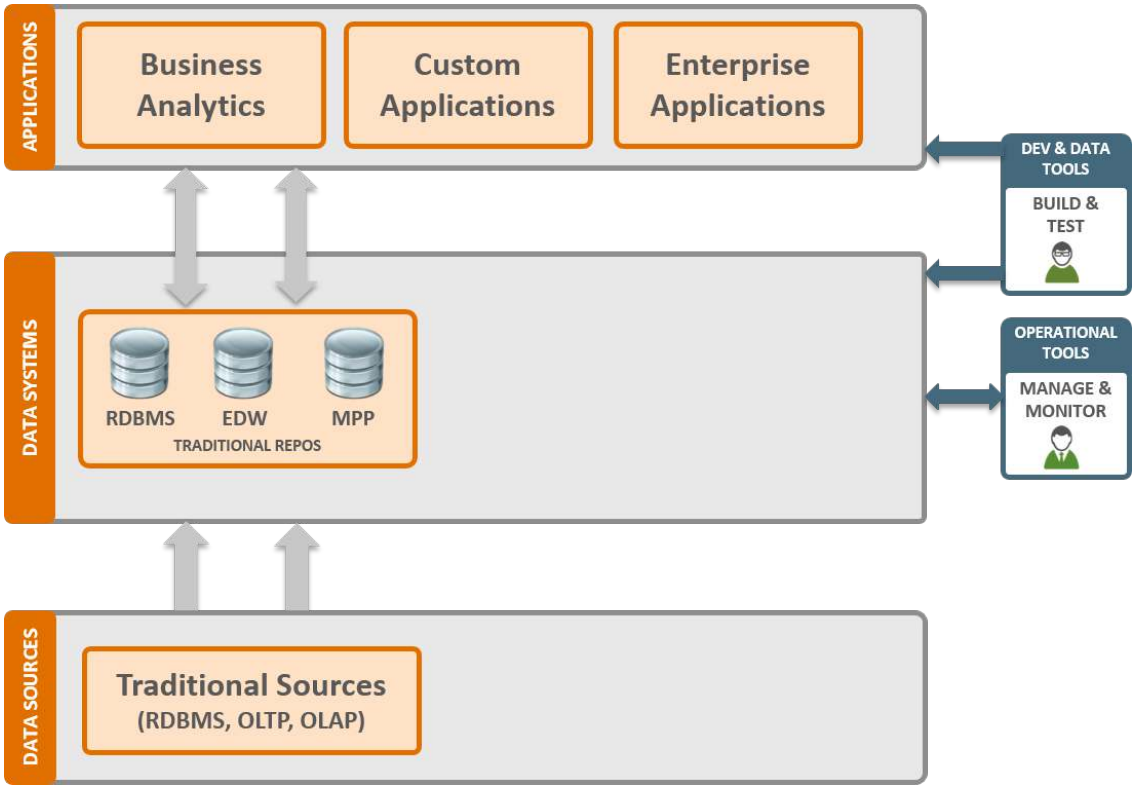


Figure 1: Traditional Database Architecture

Modern Data Architecture

In addition to traditional transactional data in analytic databases, enterprises now also need to gather, process, and analyze new unstructured data sets that are growing exponentially.

This new information can include text, images, machine-generated data, and online data from social media. It also includes data such as log files that was once thought of as having relatively little value; too expensive to store and analyze. These new types of data are turning the focus from "data analytics" to "big data analytics" because so much insight can be gleaned from these new data sources for business advantage.

The Hortonworks Data Platform is increasingly being introduced into enterprise environments to manage the massive amounts of these new types of data – as well as existing data – in an efficient and cost-effective manner.

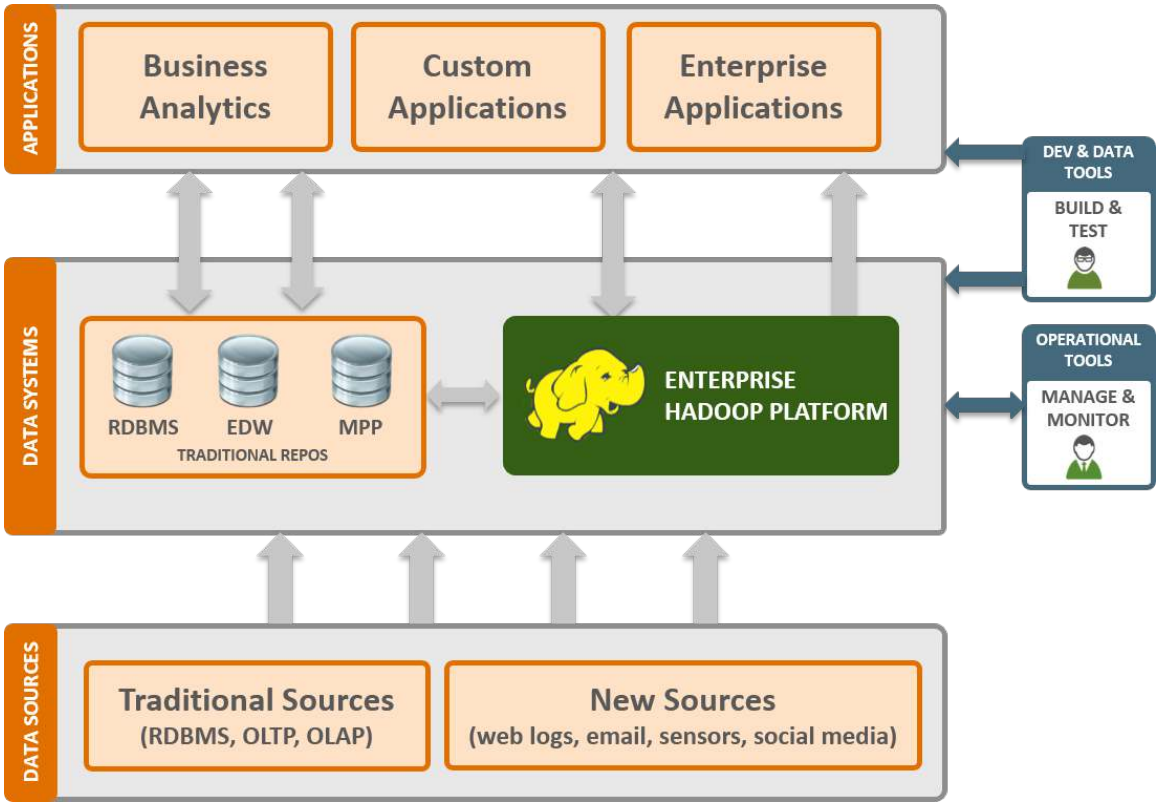


Figure 2: Modern Database Architecture

The Hortonworks Data Platform does not replace traditional data systems used for building analytic applications – the RDBMS, EDW and MPP systems – but is instead designed to integrate with and extend these systems.

By providing a framework to capture, store, and process vast quantities of both structured and unstructured data in a cost efficient and highly scalable manner, the Hortonworks platform is driving the creation of a new generation of enterprise database systems.

Hortonworks

3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks

Hortonworks We do Hadoop.

Tableau and the Hortonworks Data Platform

Tableau can be used with Hortonworks to explore your expanded data set. Tableau can directly access the data in the Hortonworks Data Platform, as well as the data in traditional analytic databases, and can combine them in a single view using a capability known as "data blending." Tableau can then explore and visualize the blended data, providing valuable business insights.

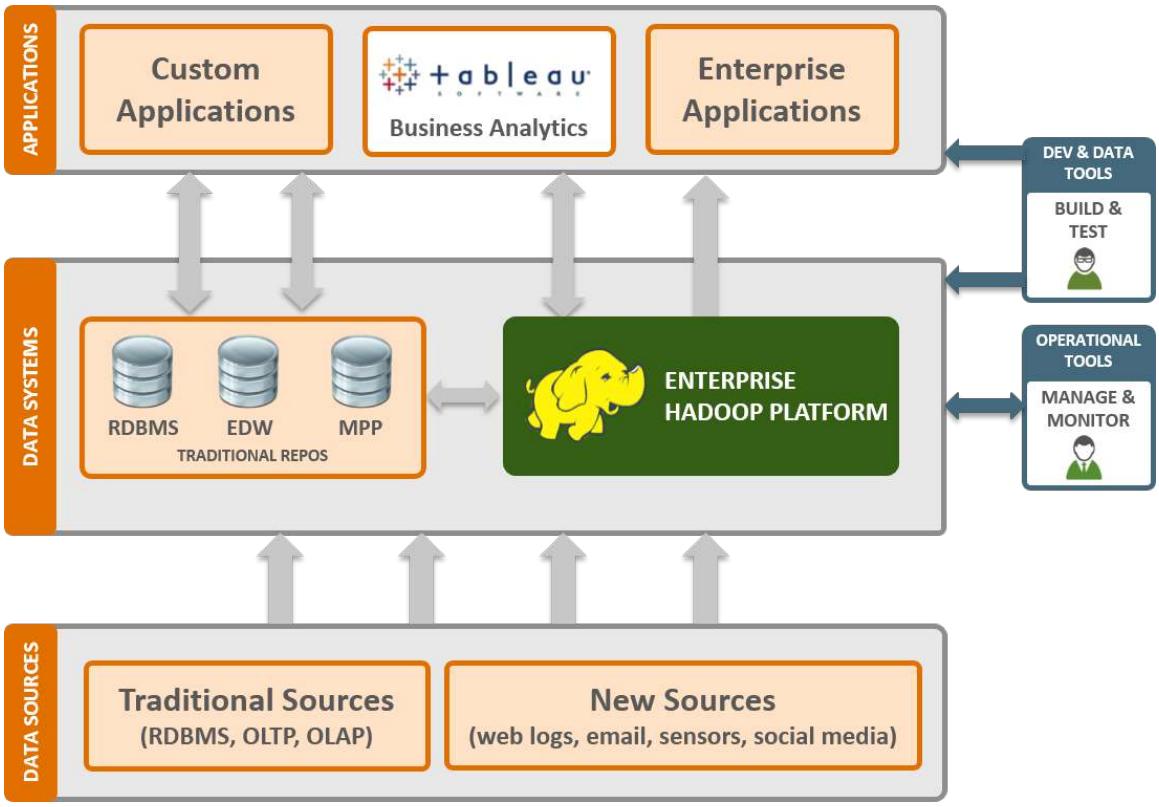


Figure 3: Tableau and Hortonworks

About Hortonworks

Hortonworks is a leading commercial vendor of Apache Hadoop, the preeminent open source platform for storing, managing and analyzing big data. Our distribution, Hortonworks Data Platform powered by Apache Hadoop, provides an open and stable foundation for enterprises and a growing ecosystem to build and deploy big data solutions. Hortonworks is the trusted source for information on Hadoop, and together with the Apache community, Hortonworks is making Hadoop more robust and easier to install, manage and use. Hortonworks provides unmatched technical support, training and certification programs for enterprises, systems integrators and technology vendors.

Hortonworks

3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks

Use Cases

Enterprises can combine Tableau with the Hortonworks Data Platform for the following use cases:

- Data Exploration
- Data Visualization

Data Exploration

In the Data Exploration use case, organizations are capturing and storing a large quantity of new data (sometimes referred to as a data lake) in Hadoop, and then exploring that data directly.

Data Exploration can be used to explore information that was previously ignored (social media data, server logs, clickstream data, web logs, machine/sensor data, and geolocation data), generate reports and visualizations from that data, and use new or existing analytic applications to leverage these new types of data.

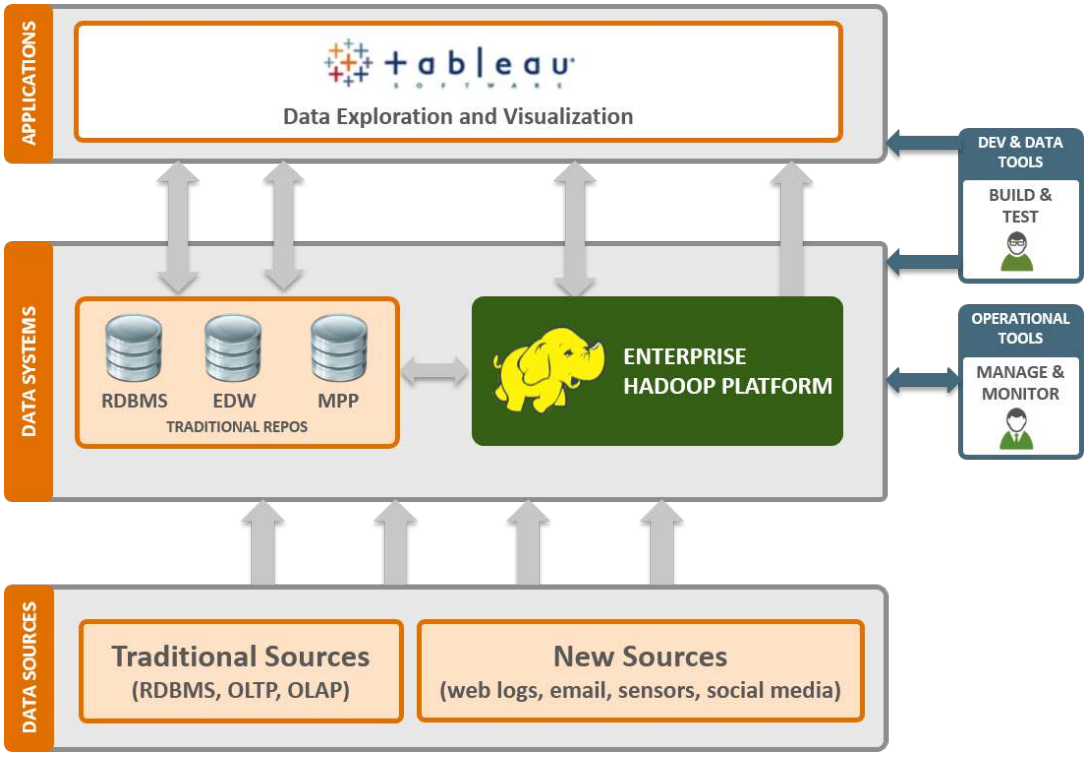


Figure 4: Data Exploration with Tableau and HDP

Data Visualization

Traditionally, gaining insights from a set of data has meant writing SQL queries to extract information from a database – often requiring the assistance of a programmer – and then working with spreadsheets to derive insights from tables of data.

Data visualization leverages people's natural tendency to think visually. It's much easier for people to understand data when they see it visually represented. It's much more difficult for people to try to extract meaning by looking at a table of data.

To illustrate this, let's use Tableau to visualize some sample clickstream data from an online retail store. Let's take a look at website visits by product category in the state of Florida.

With Tableau and Hortonworks, you can connect to the data directly and visualize the latest data. With just a few clicks in Tableau, you end up with the following visualization of the retail store data:

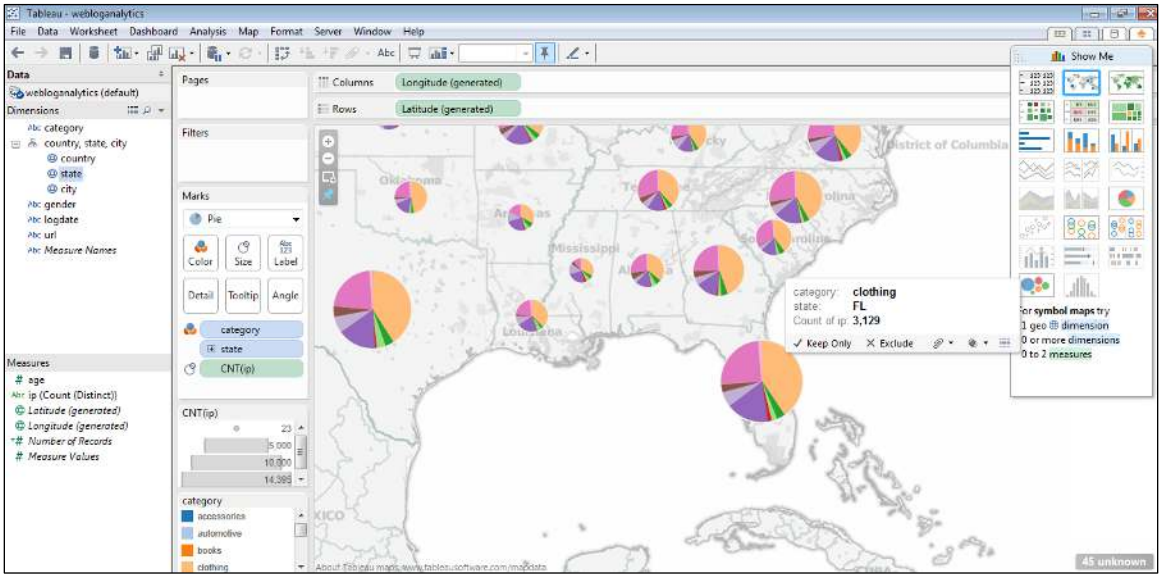


Figure 5: Sample Retail Store Data in Tableau

Here we can instantly see the product category details for each state by moving the pointer over the pie charts. At a glance we can see that clothing is the largest category in Florida, followed by shoes and handbags. With a few more clicks, we could visualize that same data by age or gender, or change the view to a bar chart or tree map.

This combination of ease-of-use and broader access means that a business or financial analyst no longer needs to wait for a database specialist in order to access data. Tableau also enables you to share your interactive visualizations on a secure server with colleagues, customers, and partners, providing them with the tools they need to answer their own questions. It's true democratization of data.

Getting Started with Hortonworks and Tableau

Here are a few links to help you get started with Hortonworks and Tableau:

- The Hortonworks Sandbox – This free download contains a stand-alone, single-node Hadoop environment, along with a set of hands-on, step-by-step tutorials.

- Tableau trial version – This page contains links to fully functional trial versions of Tableau Desktop, Tableau Server, and Tableau Online.

- Hortonworks Hive ODBC driver – The Hortonworks Add-Ons page contains links to the Hortonworks Hive ODBC driver. On Windows, Tableau requires the 32-bit version of the Hortonworks ODBC driver, even when running on 64-bit versions of Windows.

- Best Practices for Hadoop Data Analysis with Tableau

Hortonworks
3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks