Using Tableau with the Hortonworks Data Platform

David Hoyle
13 July 2013

Executive Summary

Modern businesses need to manage vast amounts of data, and in many cases they have accumulated this data for years.

Many companies have built large-scale environments for transactional data with analytic databases, but are now inundated with new types of data, such as social media activity, clickstream data, web logs, financial transactions, videos, and machine sensor data from equipment in the field.

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: 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, you 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.

Tableau is designed for ease-of-use, and to support people's natural tendency to think visually. Tableau also lets you share data visualizations with colleagues, customers, and partners.

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 complete data management solution.
- How you can use Tableau with Hortonworks for data exploration and visualization.

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 in a single tested and certified package. Hortonworks offers the latest innovations from the open source community, and the testing and quality you expect from enterprise-quality software.

The Hortonworks Data Platform provides a 100% open source, enterprise-grade Hadoop distribution. We are committed to working within the 100% open source Apache Software Foundation with no commercial holdbacks. Other Hadoop vendors take a proprietary approach that can lead to closed interfaces and vendor lock-in.

The Hortonworks Data Platform is designed to integrate with your existing 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: Key Features and Benefits

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

- **Retain as much data as possible.** Traditional data warehouses age, and over time will eventually store only summary data. Analyzing detailed records is 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 able to process nontraditional data such as social media data, server logs, clickstream data, machine sensor data, and geolocation data. Hortonworks enables you to incorporate both structured and unstructured data in one comprehensive data management system.

- **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 must be readily accessible. Apache Hadoop clusters can provide a low-cost solution for storing massive data sets – with the information still readily available.

- **Apply basic data cleansing and data cataloging.** You can categorize and label all data in Hadoop with descriptive information (metadata), and enable integration with transactional databases and analytic tools. This reduces the time and effort required to integrate with other data sets.

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

The Hortonworks Data Platform is the foundation for your next-generation enterprise data architecture – one that addresses both the volume and complexity of today's data.

Tableau

Tableau is a data analysis tool you can use 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 need to pull data into a silo – you work directly from your database.

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.

Connecting Tableau to the Hortonworks Data Platform

Tableau connects to the Hortonworks Data Platform via the Hortonworks ODBC driver. ODBC (Open Database Connectivity) is a standard protocol used to access database management systems (DBMS). You must install the Hortonworks ODBC driver in order to access the Hortonworks Data Platform with Tableau.

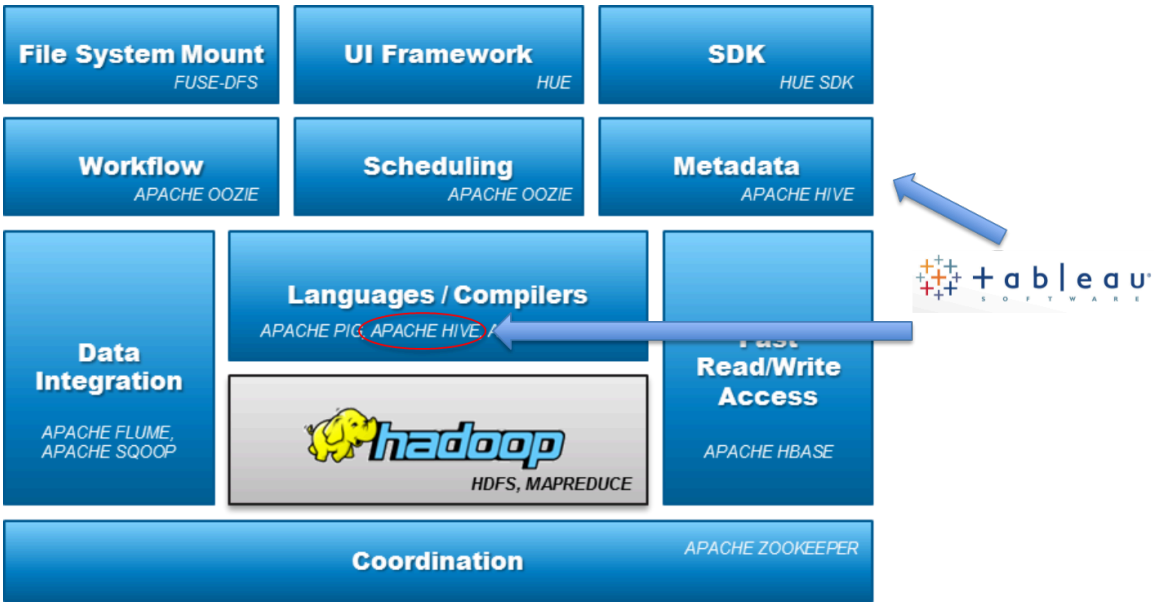


Figure 1: Tableau and HDP

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.

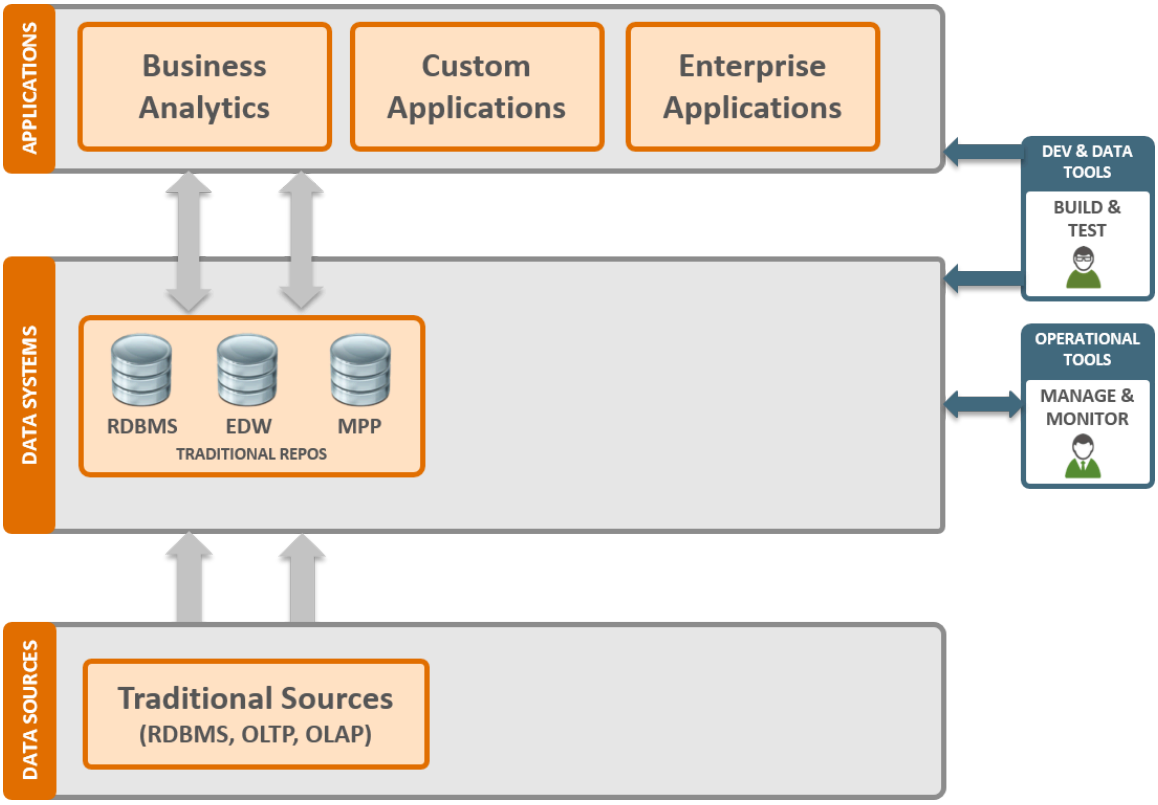


Figure 2: Traditional Database Architecture

Modern Big Data Architecture

In addition to traditional transactional data in analytic databases, companies now must also 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 were once thought to have 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.

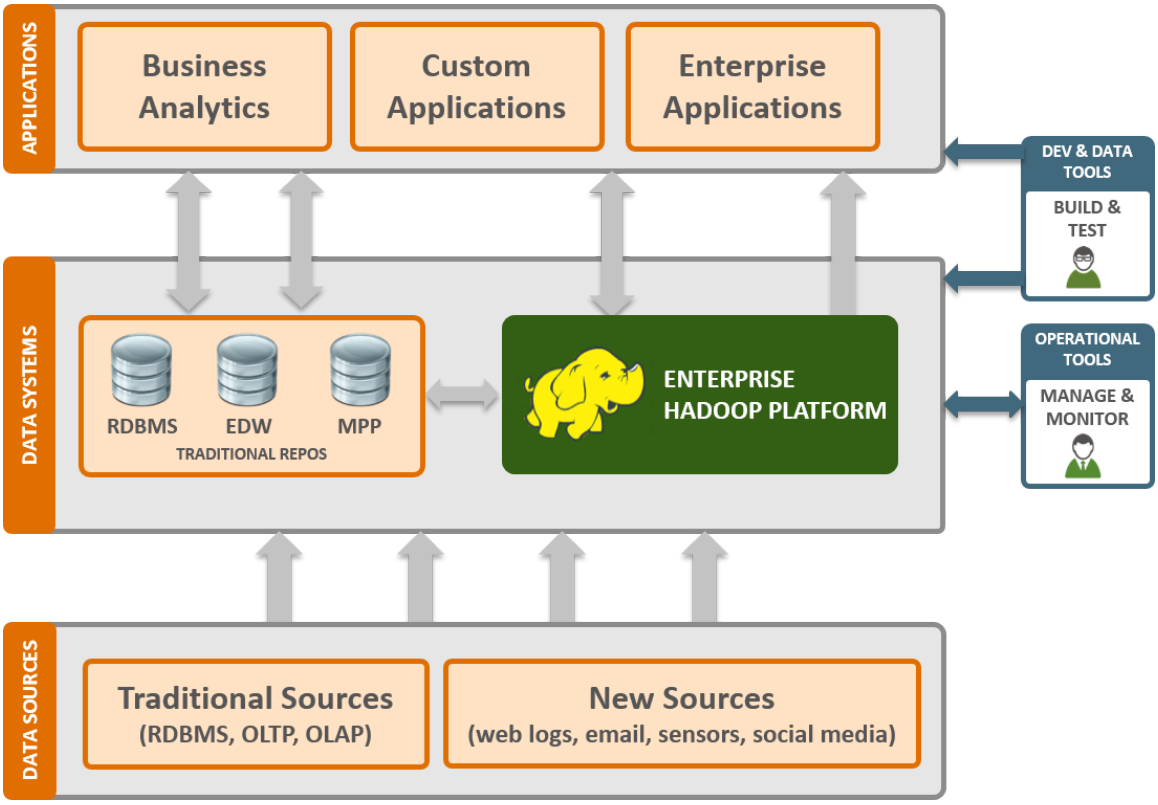


Figure 3: Modern Database Architecture

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.

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.

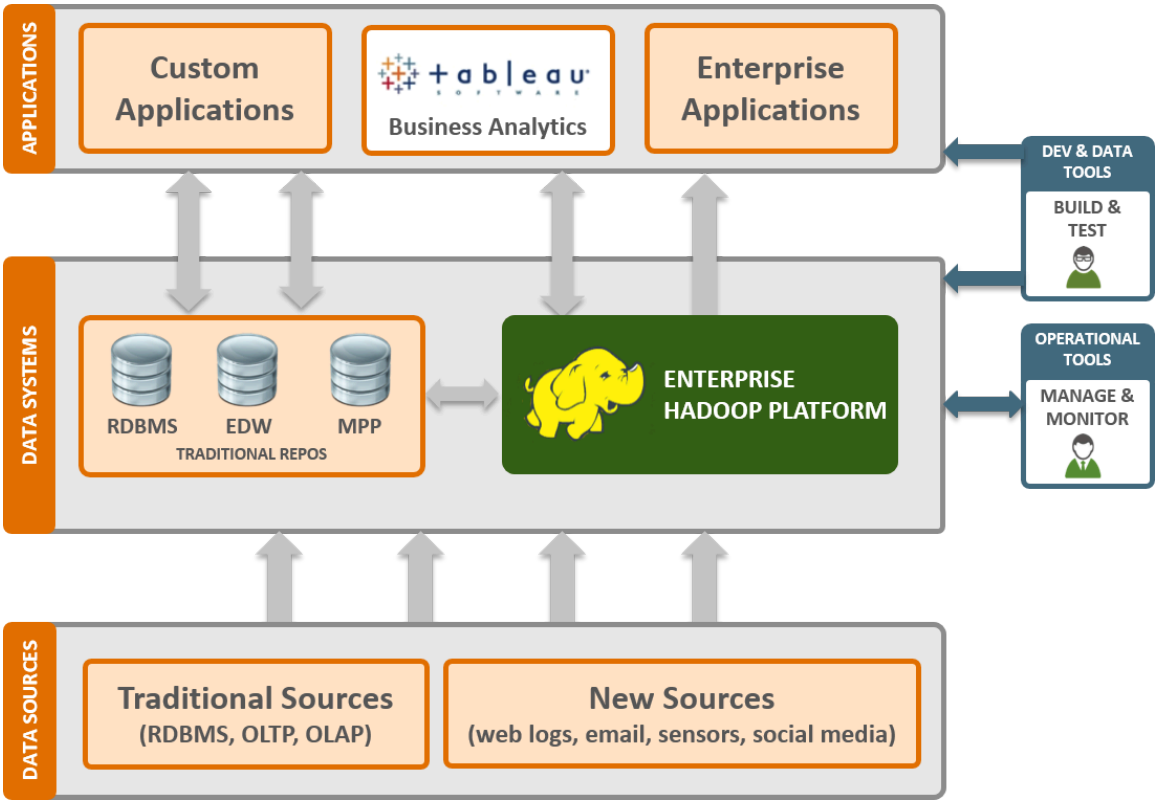


Figure 4: Tableau and Hortonworks

Use Cases

The following use case examples combine Tableau with the Hortonworks Data Platform:

- Data Exploration
- Data Visualization

Data Exploration

In the Data Exploration use case, companies 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 (text, machine data, social media data, and other online data), generate reports and visualizations from that data, and use new or existing analytic applications to leverage these new types of data.

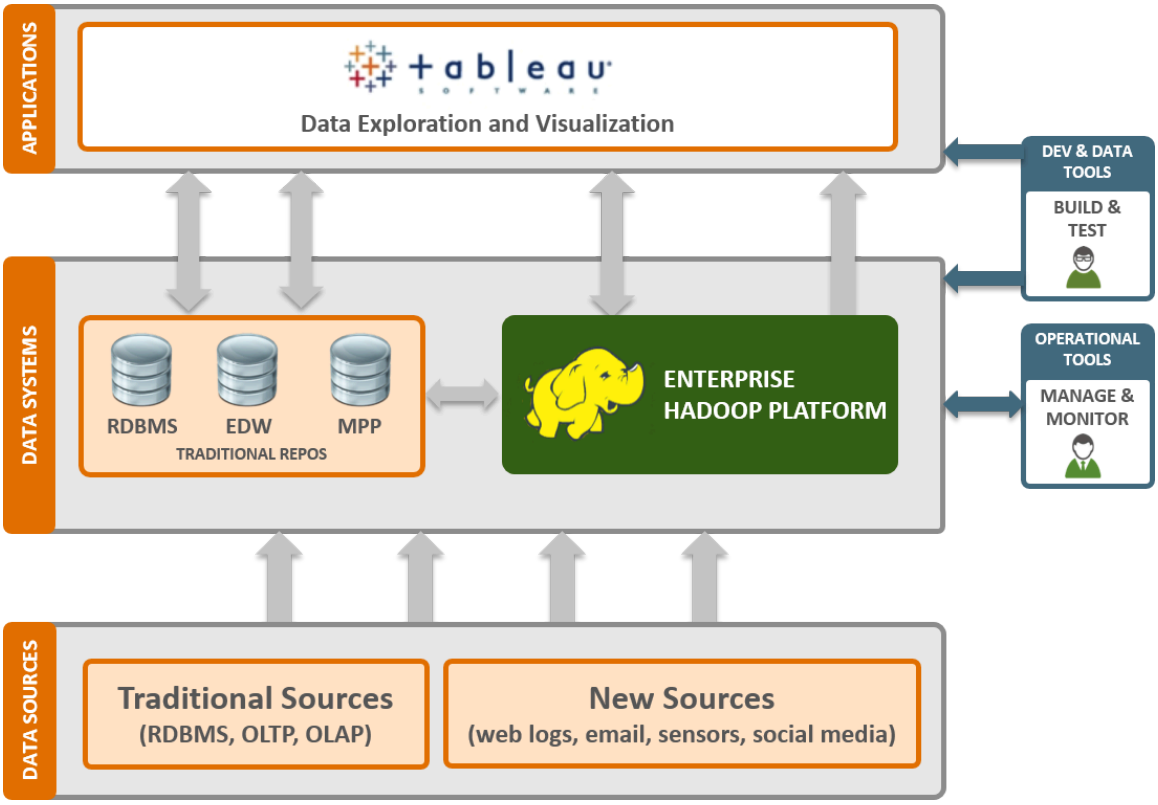


Figure 5: Data Exploration with Tableau and HDP

An example might be a telecommunications company capturing massive quantities of machine data, and then using data exploration to see if any patterns emerge that indicate when equipment is likely to fail. Given the huge quantities of data involved, they may not have previously been able to do this in a cost effective manner.

The ability to predict equipment failure (and respond proactively) is extraordinarily valuable, because it is far less expensive to do preventative maintenance than it is to pay for emergency repair or replacement. For example, if a restaurant's refrigerator fails, the franchise loses thousands of dollars in spoiled food, and a day's revenue.

Nearly every industry can take advantage of the Data Exploration use case. In financial services, organizations can use data exploration to perform forensics, or to identify fraud. A professional sports team may use data exploration to analyze trades or their annual draft, as depicted in the movie *Moneyball*.

Many companies are using Hadoop to explore and unify customer data, resulting in improved customer satisfaction and increased sales. Tableau can combine the data in the Hortonworks Data Platform with data from traditional analytic databases, creating a blended view of the combined data.

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 structured data in traditional analytic databases. Tableau can then explore and visualize that data, providing valuable business insights.

Data Visualization

Q: Why does data visualization matter?

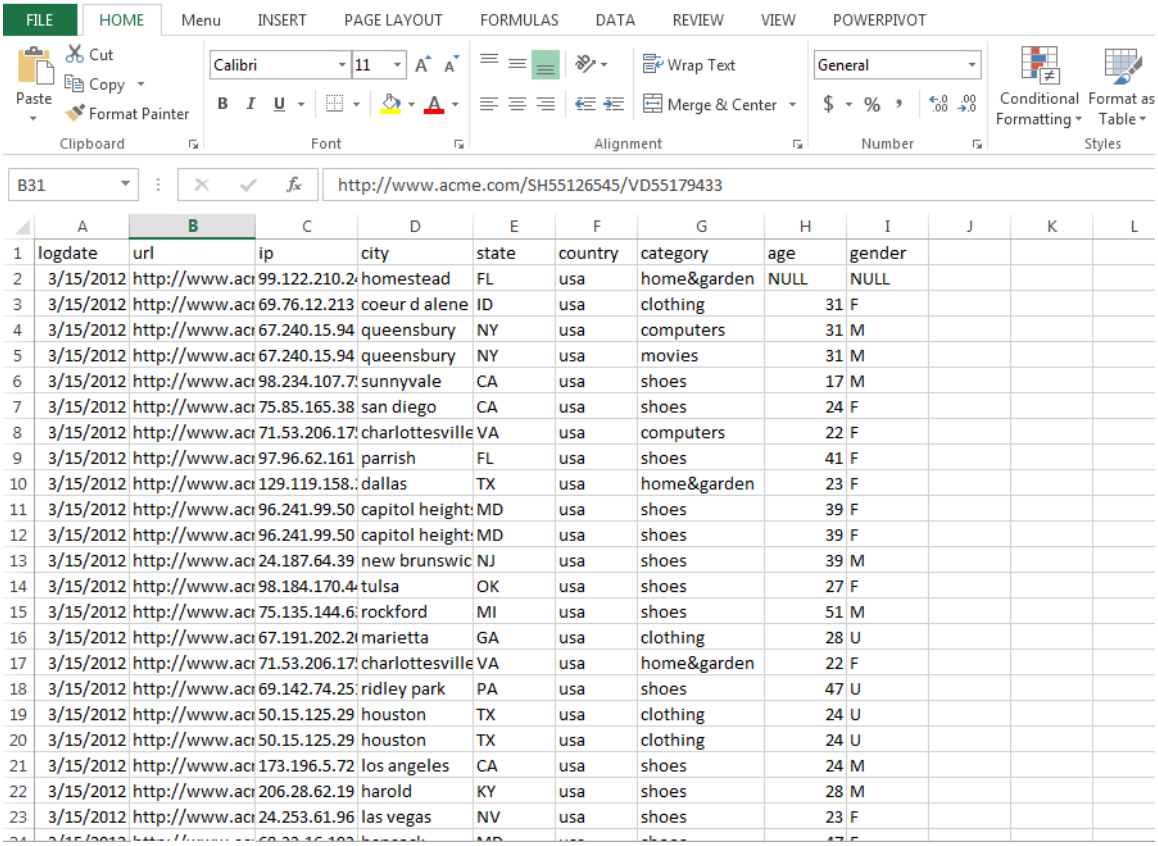
A: Visualization makes it much easier for people to understand data.

Traditionally, gaining insights from a set of data has meant writing SQL queries (or finding someone who knows how to write SQL queries) to extract information from a database, and then scrolling through spreadsheets trying to derive insights from row upon row 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 numbers.

"One well-crafted visualization is worth more than one-hundred thousand lines of data."

To illustrate this, let's look at some sample data from an online retail store using the traditional, non-visual approach. If you don't know SQL, first you have to find someone who can write the queries to generate a table from the database. But first you probably need to spend some time explaining to them exactly what you're looking for. Eventually, you get your table and open it up in Excel. Now let's say you're interested in looking at website visits by product category in the state of Florida. Here's the Excel spreadsheet:



	A	B	C	D	E	F	G	H	I
1	logdate	url	ip	city	state	country	category	age	gender
2	3/15/2012	http://www.acı	99.122.210.2	homestead	FL	usa	home&garden	NULL	NULL
3	3/15/2012	http://www.acı	69.76.12.213	coeur d alene	ID	usa	clothing	31	F
4	3/15/2012	http://www.acı	67.240.15.94	queensbury	NY	usa	computers	31	M
5	3/15/2012	http://www.acı	67.240.15.94	queensbury	NY	usa	movies	31	M
6	3/15/2012	http://www.acı	98.234.107.7	sunnyvale	CA	usa	shoes	17	M
7	3/15/2012	http://www.acı	75.85.165.38	san diego	CA	usa	shoes	24	F
8	3/15/2012	http://www.acı	71.53.206.17	charlottesville	VA	usa	computers	22	F
9	3/15/2012	http://www.acı	97.96.62.161	parrish	FL	usa	shoes	41	F
10	3/15/2012	http://www.acı	129.119.158.	dallas	TX	usa	home&garden	23	F
11	3/15/2012	http://www.acı	96.241.99.50	capitol height	MD	usa	shoes	39	F
12	3/15/2012	http://www.acı	96.241.99.50	capitol height	MD	usa	shoes	39	F
13	3/15/2012	http://www.acı	24.187.64.39	new brunswic	NJ	usa	shoes	39	M
14	3/15/2012	http://www.acı	98.184.170.4	tulsa	OK	usa	shoes	27	F
15	3/15/2012	http://www.acı	75.135.144.6	rockford	MI	usa	shoes	51	M
16	3/15/2012	http://www.acı	67.191.202.2	marietta	GA	usa	clothing	28	U
17	3/15/2012	http://www.acı	71.53.206.17	charlottesville	VA	usa	home&garden	22	F
18	3/15/2012	http://www.acı	69.142.74.25	ridley park	PA	usa	shoes	47	U
19	3/15/2012	http://www.acı	50.15.125.29	houston	TX	usa	clothing	24	U
20	3/15/2012	http://www.acı	50.15.125.29	houston	TX	usa	clothing	24	U
21	3/15/2012	http://www.acı	173.196.5.72	los angeles	CA	usa	shoes	24	M
22	3/15/2012	http://www.acı	206.28.62.19	harold	KY	usa	shoes	28	M
23	3/15/2012	http://www.acı	24.253.61.96	las vegas	NV	usa	shoes	23	F

Figure 6: Sample Retail Store Data in Excel

If you know Excel, you know you can sort by state or category, or maybe even write a formula or a macro to extract the product category data from Florida. But that may take some time. And you still end up with a list of numbers.

Now let's try a new approach. With Tableau and Hortonworks, you can connect to the data directly and visualize the latest data – without a programmer. With just a few clicks in Tableau, you end up with the following visualization of the retail store data:

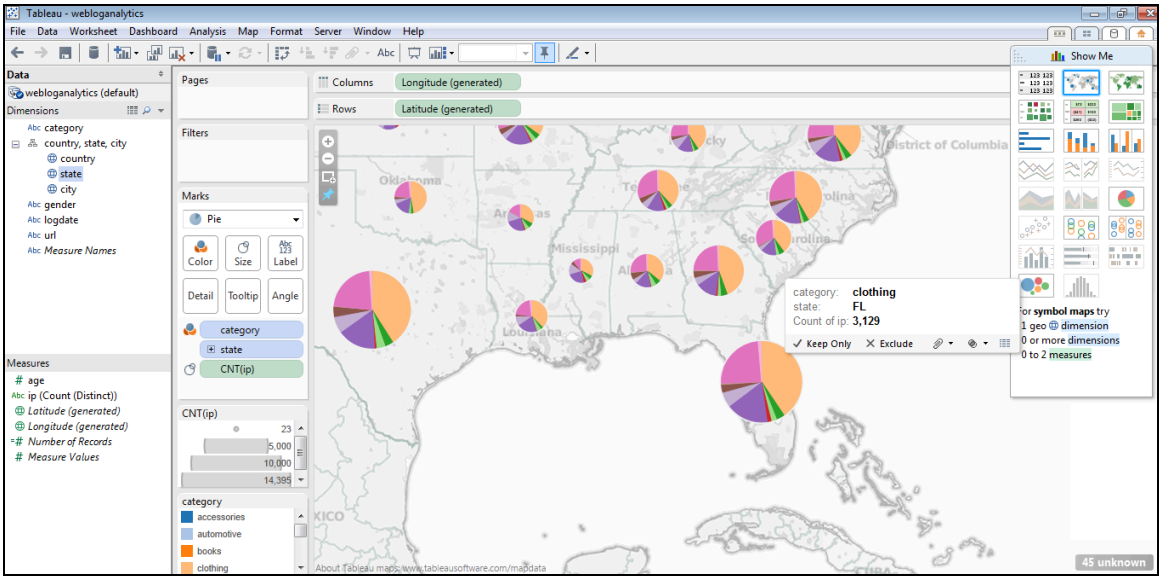


Figure 7: Sample Retail Store Data in Tableau

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 to access data. Tableau also lets you 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 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.

- Data Discovery, Visualization and Apache Hadoop – A recording of a joint Hortonworks-Tableau webinar (registration required).