datanami
BIG DATA • BIG ANALYTICS • BIG INSIGHTS

The Five Types of Hadoop Data

August 23, 2013 | Isaac Lopez

Hadoop is big and getting bigger. One recent estimate says that Hadoop currently had a market worth $1.5 billion dollars in 2012, and is projecting a compound annual growth rate of 54.7% between 2012 and 2018 for a global market size of $20.9 billion.

Still plenty of confusion exists about what exactly Hadoop can do for an organization. Questions abound from what data organizations can get use of, to what they can do to get business value and increase their competitive edge. I recently had lunch with Dave McJannett and Jim Walker at Hadoop distro vendor, Hortonworks, and we discussed the challenges that organizations face as they consider if Hadoop is right for them. They shared the five major types of Hadoop data that organizations are using to increase business value, including everything from decreasing costs, optimizing security processes, and building value.

While Hadoop is not the end-all-be-all of big data, it's clear that it is going to play an increasingly important role in the future of organizations of all types. We'll continue to focus on big data implementations of all types, but it's hard to ignore the elephant in the room.

Let's dig in and start with a use case that everyone in the Web 2.0 world can get use out of.

START – Data Type #1: Clickstream Data >>>

Getting to Know Users with Clickstream Data

A clickstream is exactly as it sounds – the stream of clicks that a user takes as they path through a website. As a user navigates through a company's website, they leave clickstream data that is captured in weblogs. Clickstreams were one of the earliest use cases for Hadoop in its original inception at Yahoo! as the company used the framework to store and process the enormous amount of data coming in from their users.

There is a tremendous amount of useful information that can be gleaned from clickstream data. Here are some of the ways that clickstream data can provide benefit to an organization:

- **Path Optimization –** Path optimization aims at reducing bounce rates and improving conversions.
- **Basket Analysis –** This aims at understanding aggregate customer purchasing behavior by examining such things as customer interests, and paths to purchase - when customers bought Product X, what common paths did they take to get there.
- **Next Product to Buy Analysis –** Related to basket analysis, this type of analysis looks at correlation in purchases, and what can be offered next to help provide more immediate value to the customer, and increase the likelihood of another sale.
- **Allocation of Website Resources –** Having clickstream data on hand, a company will know what their hottest and

coldest paths on the site are and can assign development resources accordingly, optimizing resource allocation.

- **Granular Customer Segmentation** – With clickstream and correlated user data, a company can discover and gain insight on how particular segments and micro-segments of customers are using the site, and how to best cater to them.

Clickstream analysis is all about measuring users and their behavior. With the data collected, the hot and cold spots on a website can be located and actions taken to ensure that users are seeing the most valuable offers.

NEXT – Data Type #2: Sentiment Data >>>

Understanding Your Customers Thoughts Using Sentiment Data

While clickstreams will give a sort of Boolean understanding of customers and their actions, this is the era of social media where organizations can get a more subjective appreciation of what their customers are thinking about them (and their competitors), and take actions to respond to these sentiments.

From Twitter, to Facebook, to blogs, to a never-ending array websites with comment sections and other ways to socially interact on the Internet, there is the potential to be a massive amount of sentiment information available about any given company. That unstructured data creates problems for traditional databases; however, unstructured data is where Hadoop shines.

An organization can collect all of these data streams and track how their customers and prospects feel about its products, the company itself, issues important to the company, competitors and more. Correlating information about prevailing sentiments and their locations, a company can custom target their marketing and ad campaigns to either capitalize on or combat the sentiments.

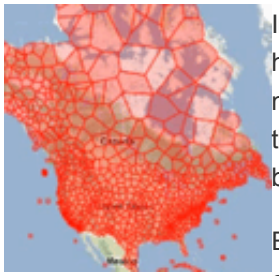
An example case study that Hortonworks gives for sentiment data using Hadoop involves the launch of Iron Man 3 this past May. Tracking the volume of tweets around the movie's launch, three clear spikes in global tweet volume can be identified, including the Friday premier, the Saturday matinee time, and the Saturday evening showing. Correlating location data with sentiment analysis, the studio is able to see that in Ireland, over half the tweets expressed positive sentiments about the movie, where in Mexico, negative sentiments about the movie outweighed the positive ones.

These sentiments can be tracked during the launch for real time marketing activities (i.e. increasing spends and extending campaigns in a particular market), as well as planning future product launches.

NEXT – Data Type #3: Geolocation Data >>>

Improving Processes and Operations Through Geolocation Data

Geolocation data gives organizations the ability to track every moving aspect of their business, be they objects or individuals.



In the age of smart phones, companies now have the ability to track their customers in ways that have previously been unimaginable. Recently we covered the story about how retailers are using moving Wi-Fi signals from phones to track their customer's movements through their stores for the purposes of improving store layouts. This is just one example of how geo-location data can be used to improve efficiency.

Earlier this year, we learned how shipping giant, UPS, is experimenting with a Hadoop cluster in order to find ways to streamline the enormous amount of data they have, including sensor data on vehicles out in the field. Big Brown says they are working to achieve "prescriptive analytics," – understanding what they did yesterday to better predict what will happen tomorrow, and thus devise new strategies around their existing processes.

Another recent example of using geolocation data was given to us earlier this year by Fujitsu, who created a real-time criminal activity map using the geolocation data from Twitter streams. Collecting data on regional happenings, the group was able to create an algorithm that synthesized the information into trouble events that they could map to a location, creating a map that people can use to see where trouble areas are at any given time.

Every day more sensors are coming online providing additional opportunities for gaining insights through geolocation data. Hadoop will continue to shine as a cheap and effective place to store and process that data.

NEXT – Data Type #3: Server Log Data >>>

Unlocking the Potential of Server Log Data

Some estimates say that global internet traffic will reach 1 zettabyte (the equivalent of 1 billion terabytes) by 2015. All of this data, as well as all assorted flavors of network traffic, winds up in server logs and is considered “exhaust data.”

One of the chief uses for server log data is for security analysis on a network. Admins are able to load their server logs into Hadoop using applications like Apache Flume, building a repository that they can use for analysis in order to identify and repair vulnerabilities.

Server log data can be used for an array of purposes to give organizations insights on everything from network usage, security threats, and compliance. Hadoop will be a central player in staging and analyzing this type of data for some time to come.

NEXT – Data Type #3: Machine and Sensor Data >>>

Unlocking Predictive Analytics with Sensor Data

Sensor data is among the fastest growing data types, with data collectors being put on everything under the sun. These sensors monitor and track very specific things, such as temperature, speed, location – if it can be tracked, there’s a sensor to track it. People are carrying sensors on them regularly (smart phones), and more come online every day.

Recently, we covered a case study where Hadoop was being used in manufacturing, with sensor data from the machine controllers being staged in Hadoop, where it was sliced, diced and analyzed in order to provide correlations that gave the shop manager a better understanding of what was happening with their tools. By analyzing the massive amounts of historical data, they were able to predictively analyze when machines were headed for problems and quickly get people where the trouble was at in order to minimize down time and maximize

operation.

Hadoop is a very attractive store for sensor data due to the ability to dump so much of the juice into the framework and then use analysis tools to extract correlations that give insights on operations, and how things change when conditions change.

Sensor data promises to be huge, as the desire to monitor the moment-to-moment status of things increases as businesses look for ways to increase efficiency and cut costs in their operations. With every second, another data point needs to be stored, creating a challenge for existing traditional database paradigms. The relative cheapness of Hadoop makes it a very attractive candidate for sensor data storage.

Related items:

Big Data Dispelling Preconceived Notions in the NFL

Providing Hidden Benefits With Predictive Analytics

Predictive Analytics Prevailing on the Wind Farm
