Are You Big Data Informed?

A Retail Case Study using Hadoop, Pig and Tableau

Jules Malin
Northwestern University
Jules.malin2015@u.northwestern.edu

Abstract-- Organizations that understand competing on Big Data analytics is a new minimum competitive requirement are implementing the relevant tools and technologies to extract actionable insights. Whether an organization is exploring, piloting or operationalizing Big Data solutions individuals and teams can facilitate this inflective technology adoption by proposing and piloting use cases for their organizations. This paper investigates a use case for analyzing retail order data using Hadoop, Pig and Tableau to store, process and analyze the data with the goal of discovering actionable insights. Using these tools and technologies enable organizations to scale and automate their analytics. Note that the data used for this paper is a random and disguised sample of Q1 2014 order data.

Keywords: Hadoop, PIG, Tableau, geospatial analysis, retail

Shutterfly now stores over 120 petabytes of customer data with the volume, velocity, variety, and veracity of that data increasing daily. What Big Data systems and tools can we implement to effectively manage this data and to extract customer and product insights, enabling Shutterfly to compete on analytics? One proven and hyper scalable solution is deploying Hadoop as an enterprise platform and integrating it with a traditional relational database infrastructure.

This paper will summarize Big Data concepts, challenges and explore a use case enabled by Hadoop, Pig and Tableau. The use case illustrates how these platforms, working together can produce actionable insights related to Shutterfly customers and products. The insights discovered include which products are top sellers and where they are popular by state and city using geospatial analysis. In addition, the analysis provides insights into hot spots in the United States where the highest concentration of orders originate. With these geospatial demand insights we can use targeted marketing to further increase sales in regions where demand is strong and work to open new markets where demand is lite or non-existent.

Big Data and Shutterfly

In 2001 Doug Laney articulated the concept of Big Data as the massive growth of data volume, velocity, and variety. Additional V’s or characteristics have been added to the definition, including veracity and value. I will use the first four V’s to define Big Data as it relates to Shutterfly. Value is less related to the definition and more related to how Big Data can be exploited (SAS, 2001).

Volume

Data volumes are increasing in line with the increasing amount of structured and unstructured data being generated and stored. For example the increasing amount of unstructured transaction data being generated on our Facebook page, on Twitter and in our customer service logs. Also driving this data volume increase are the increases of sensor data generated by the RFIDs being used to track product in production through to shipping.

Velocity

Refers to the rate at which data volume is increasing or speeding up from one moment to the next and also speeding up between connections. For example, Shutterfly data velocity has increased as we acquire new companies and grow organically. Also as we deploy machine learning and additional analytics these drive our data velocity increases through machine generated data.

Variety

Describes the increase in the type of data being generated including, unstructured data generated by social media like text, email, video, and images. For
example, Shutterfly is collecting customer order data, photos, videos, Facebook, Twitter and customer service text logs. In addition to these data types we are also collecting new RFID data from production, printer press data, external shipping data from FedEx and UPS and other third party logistics partners.

**Veracity**

Refers to the messiness or trustworthiness of data. For example, customer service text logs originate from chats, phone conversations and emails, all of which generate different levels of messiness including but not limited to misspellings, symbols and web and tool generated characters.

These four concepts of increasing data volume, velocity, variety and veracity illustrate the data challenges Shutterfly is facing as well as the opportunities that Big Data enables. These data challenges can be met by deploying Big Data ecosystems like Hadoop and its associated tools. Hadoop can scale horizontally to keep pace with the increasing data growth and enable analysis of that data which is difficult and in many cases impossible with a traditional relational database system alone.

**Case Study: Discovering Customer and Product Insights**

What were the top 10 products sold by Shutterfly in Q1 2014 by order count? Answering this question will provide insights into which products will be successful in Q1 2015 and which products need to be improved or eliminated. The data used for this analysis contains a sample of 100,000 orders from Q1 2014. It includes userid, userzipcode, orderdate, prodsku, ordercount, unitstotal and rating.

Figure 1 below is the Pig script for finding the top 10 products of Q1 2014 by order count, followed by the output table in Table 1. In addition to order count we could also explore the data by quantity ordered as well.

**Table 1**

<table>
<thead>
<tr>
<th>prodsku</th>
<th>ordercount</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB25</td>
<td>9810</td>
</tr>
<tr>
<td>CD31</td>
<td>3808</td>
</tr>
<tr>
<td>AR52</td>
<td>3188</td>
</tr>
<tr>
<td>AR54</td>
<td>2366</td>
</tr>
<tr>
<td>CR23</td>
<td>1812</td>
</tr>
<tr>
<td>PB21</td>
<td>1275</td>
</tr>
<tr>
<td>AP45</td>
<td>970</td>
</tr>
<tr>
<td>MA03</td>
<td>712</td>
</tr>
<tr>
<td>AB01</td>
<td>600</td>
</tr>
<tr>
<td>FG88</td>
<td>582</td>
</tr>
</tbody>
</table>

Next, is a bar chart using the output from the Pig script which is then generated by Tableau, illustrating the top ten products sold in Q1 2014. Tableau enables ad hoc and dashboard visualizations of our data. See Figure 2.

**Figure 2.** Tableau bar chart generated using Pig script output data of top ten products.

**Product Geospatial Analysis**

Now that we know what the top 10 products are we can use Pig to discover what the top 100 zip codes are and where demand for these top 10 products are the strongest. Figure 3 below is the Pig script used for finding the top 100 zip codes where most orders originated.

**Figure 3.** Pig script used for finding the top 100 zip codes.
Using the data output from the Pig script in combination with Tableau to perform geospatial analysis we can determine which states drive the highest percentage of orders. See Figure 4. 100 zip codes may not be enough, so I will look at the top 800 next.

Figure 4. Tableau visualization of Pig script output used for finding top 100 zip codes.

Now let us look at the top 800 zip codes. Figure 5 shows the Pig script for finding these orders. Figure 6 shows the Tableau visualization of the top 800 zip codes with the highest concentration of orders including all products. In this analysis we can see that the highest concentration of orders are on the west coast in the SF Bay Area, Southern California and in the North East of the US.

Figure 5. Pig script used for finding top 100 zip codes.

Next, since we know product AB25 are the top seller we can use the same data set and Pig to analyze which regions are ordering these by looking at the top 800 customers and where they live in order to determine which states and zip codes should be targeted for marketing promotions. See Figure 7.

Figure 6. Tableau visualization of Pig script output used for finding top 800 zip codes.

Analyzing the output data in Tableau we can see where customers are ordering AB25. In Figure 8 the darker the circle the higher the order count density and therefore the locations marketing should focus their efforts. The images below show us that customers from most states order AB25, even Alaska.

Figure 7. Pig script used for finding top 800 AB25’s by zip codes.
Using Tableau we can take this analysis further by zooming in on California orders by area code. In Figure 9 below we can see that the SF Bay area, Los Angeles and San Diego customers order the most AB25’s.

Because we have zip code data we can even zoom in to analyze which cities in the SF Area customers order this product. San Mateo, Palo Alto, Walnut Creek and Dublin customers really like AB25’s. See Figure 10.

**Conclusion**

In conclusion, Hadoop, Pig and Tableau can be used to store and analyze Shutterfly retail data to improve our analytics capability. Also, by using a Hadoop infrastructure, Shutterfly can continue to scale well beyond the 120 petabytes of data currently being stored and analyzed. As demonstrated in this paper Pig can be used in conjunction with Tableau to automate geospatial reporting and analysis. The key benefits of Hadoop, Pig and Tableau are their ability to scale, process and automate analysis on ever larger data sets and data pipelines, thus complimenting the capability of traditional relational data bases and business intelligence tools.

**Reference**