unifi
Enterprise-class self-service data integration

WHITEPAPER

Finding the Needle in the Semi-Structured Hay Stack

By Ayush Parashar,

Co-Founder & VP Engineering, Unifi Software

DIY Data Integration for Business Insight Gurus

Business Insight Guru. You know who you are. Data is your life. The more the merrier, keep it coming. It's your job to derive critical business insights from all the data available to you.

Before you installed Unifi Software, your job was challenging. The guys in IT are great, super helpful but pulling all the data together and visualizing the results on your BI tool was a long and painful process. Often it meant that insight deadlines were missed and you took the grief from the office with the window. Now you've got Unifi on your side business insights could not be easier – that's why you're the BIG in your office.

Unifi Software

- Enterprise-class data integration built natively on Hadoop
- Built by business people, for business people
- Auto Data Integration™ eliminates the need to write a single line of code
- Integrates structured and unstructured data from the widest range of sources
- Allows easy pursuit of "what if" scenarios by business users without calling IT
- Transform In Place™ processing runs natively on Hadoop to give faster results
- Open platform seamlessly integrates with a wide range of analytics and data visualization tools

Monday Morning Situation Report

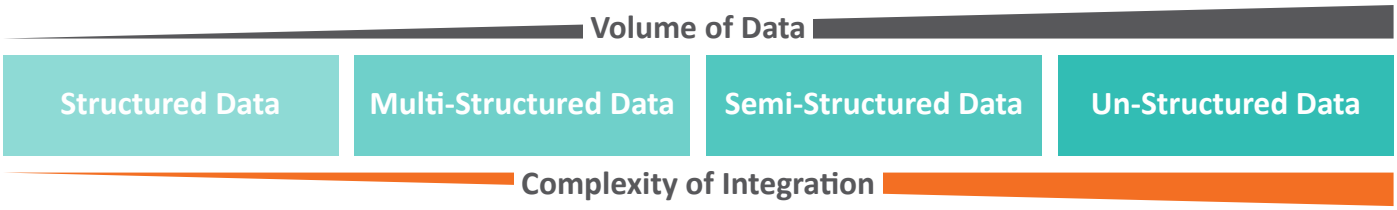
The boss wants to know if the business analyst can help the merchandising department forecast the demand for the company's Halloween costumes based on social media trend data when combined with other supply chain variables.

The answer is needed by the end of the week to meet the final ERP schedule or the company may be faced with rush charges in manufacturing or higher distribution costs to ensure supplies are ready to meet customer demand.

The request seems obvious and logical; if lots of people, especially those with buying power or influence, are talking about the company's products in the second quarter, the manufacturing department would have some indication whether the company has one of the "Must Haves" for this year's trick-or-treating.

Social Media Data—Lots of Promise, Lots of Headaches

Data about social network members and their usage habits is a treasure trove of insight. From detailed demographic profiles to interests and activities, there's nothing private about a person's life if they choose to be



active on one or more of the growing number of social media networks.

The landscape of social networks is changing and so, too, is the demographic that each attracts. Now that parents and grandparents have Facebook accounts, most of the younger generation has moved on. They have their conversations on Instagram, Yik Yak or Snapchat. Businesses trying to get a total picture of their customer base must acquire these new data sets. Developers of these networks – looking for early revenue streams that won't alienate users with advertising – sell usage data directly or license it to third parties.

The good news is there's a ton of data; the bad news is there's a ton of data. And for most business analysts and the IT departments that support them, the real challenge is not only acquiring the data, but also transforming the data so that it can be combined with existing data and analyzed using business intelligence and visualization tools to extract business insights.

The Structure of Things

If you're reading this, you probably already know that data can be defined in a number of ways: Structured, multi-structured, semi-structured, and unstructured.

Generally, the volume of data and the complexity of integration increase as the structure decreases. Before any analysis can occur, the structure of the data must be normalized or flattened into a tabular form consisting of rows and columns so the end user can understand context and enable their visualization tool of choice to display the results.


sample_twitter.json

{"id":"tag:search.twitter.com,2005:251727978931834880","objectType":"activity","actor":{"objectType":"person","id":"id:twitter.com:
36916922","link":"http://www.twitter.com/BlackBerryHelp","displayName":"Research In
Motion","postedTime":"2009-05-01T10:01:43.000Z","image":"http://a0.twimg.com/profile_images/1993678305/
r3_a1_twitter_profile_88x88__2__normal.jpg","summary":"Official Twitter support account of Research In Motion.\r\nWe offer DM support.
Follow us and send us a message. We're staffed from Monday to Saturday.","links":[{"href":"http://www.blackberry.com/
support","rel":"me"}],"friendsCount":32115,"followersCount":962895,"listedCount":7677,"statusesCount":19757,"twitterTimeZone":"Eastern
Time (US & Canada)","verified":true,"utcOffset":"-18000","preferredUsername":"BlackBerryHelp","languages":["en"],"location":
{"objectType":"place","displayName":"Waterloo, Ontario, Canada"}},"verb":"post","postedTime":"2012-09-28T17:00:01.000Z","generator":
{"displayName":"SocialEngage","link":"http://www.exacttarget.com/social"},"provider":
{"objectType":"service","displayName":"Twitter","link":"http://www.twitter.com"},"link":"http://twitter.com/BlackBerryHelp/statuses/
251727978931834880","body":"How to back up data on a BlackBerry smartphone http://t.co/gm7IDyjr ^MD","object":
{"objectType":"note","id":"object:search.twitter.com,2005:251727978931834880","summary":"How to back up data on a BlackBerry
smartphone http://t.co/gm7IDyjr ^MD","link":"http://twitter.com/BlackBerryHelp/statuses/
251727978931834880","postedTime":"2012-09-28T17:00:01.000Z"},"twitter_entities":{"urls":[{"expanded_url":"http://bbry.lv/
9uez3X","indices":[47,67],"url":"http://t.co/gm7IDyjr","display_url":"bbry.lv/9uez3X"}],"user_mentions":[],"hashtags":
[]},"retweetCount":0,"gnip":{"language":{"value":"en"},"matching_rules":[{"value":"RIM OR blackberry OR bberry OR bbpin OR bbm OR
crackberry OR whiteberry","tag":null},{"value":"from:Blackberry OR from:crackberry OR from:BlackBerryHelp OR from:BlackberryDev OR
from:BlackBerryBlog OR from:BBtips OR from:Blackberry_News OR from:bbgeeks OR from:BBerryEmpire OR
from:BlackberryCool","tag":null}],"klout_score":78,"urls":[{"url":"http://t.co/gm7IDyjr","expanded_url":"http://
btsc.webapps.blackberry.com/btsc/viewdocument.do;jsessionid=3F8997945733333EDB0E67067348EE708?
externalId=KB12487&sliceId=2&cmd=displayKC&docType=kc&noCount=true&ViewedDocsListHelper=com.kanisa.apps.common.BaseViewedDocsListHelpe
rImpl"}]}}
{"id":"tag:search.twitter.com,2005:251727982895435777","objectType":"activity","actor":{"objectType":"person","id":"id:twitter.com:
122810303","link":"http://www.twitter.com/HikeOurPlanet","displayName":"Lee Hiller-
London","postedTime":"2010-03-14T00:38:46.000Z","image":"http://a0.twimg.com/profile_images/751168208/
HikeOurPlanetLogosmallnobackground_normal.jpg","summary":"March 14th Interenational Hike Our Planet Day Love Nature, Go Hiking and Get
Intouch with Your Inner Child - Vegan ","links":[{"href":"http://www.HikeOurPlanet.com","rel":"me"}],"friendsCount":
5651,"followersCount":5155,"listedCount":77,"statusesCount":26658,"twitterTimeZone":"Mountain Time (US &
Canada)","verified":false,"utcOffset":"-25200","preferredUsername":"HikeOurPlanet","languages":["en"],"location":
{"objectType":"place","displayName":"On The Trail in Arkansas"}},"verb":"post","postedTime":"2012-09-28T17:00:02.000Z","generator":
{"displayName":"web","link":"http://twitter.com"},"provider":{"objectType":"service","displayName":"Twitter","link":"http://
www.twitter.com"},"link":"http://twitter.com/HikeOurPlanet/statuses/251727982895435777","body":"Portable Affordable Art
#LeeHillerDesigns iPhone Samsung BlackBerry Cases http://t.co/sfkF06EB Nature, Goth, Modern, Atomic & More.","object":
{"objectType":"note","id":"object:search.twitter.com,2005:251727982895435777","summary":"Portable Affordable Art #LeeHillerDesigns
iPhone Samsung BlackBerry Cases http://t.co/sfkF06EB Nature, Goth, Modern, Atomic & More.","link":"http://twitter.com/
HikeOurPlanet/statuses/251727982895435777","postedTime":"2012-09-28T17:00:02.000Z"},"twitter_entities":{"urls":
[{"expanded_url":"http://bit.ly/LeeHillerCases","indices":[74,94],"url":"http://t.co/sfkF06EB","display_url":"bit.ly/
LeeHillerCases"}],"user_mentions":[],"hashtags":[{"text":"LeeHillerDesigns","indices":[24,41]}]},"retweetCount":0,"gnip":{"language":
{"value":"en"},"matching_rules":[{"value":"RIM OR blackberry OR bberry OR bbpin OR bbm OR crackberry OR
whiteberry","tag":null}],"klout_score":49,"urls":[{"url":"http://t.co/sfkF06EB","expanded_url":"http://www.zazzle.com/
leehillerloveadvice/cases"}]}}
{"id":"tag:search.twitter.com,2005:251727983478448128","objectType":"activity","actor":{"objectType":"person","id":"id:twitter.com:

Semi-Structured Data as Received from Twitter

Data Sources	Understand what data is available to the organization	**Data Discovery**
Locate Data	Where in the organization does the data reside? Local storage, remote storage, personal desktop?	
Aquire Data	License new data as needed from third-party vendors. This must factor: Cost, timeline, sign-off process	**Data Acquisition**
Desired Outcome	Once the data sets have been identified, the analyst must explain the business requirements to the IT programmer	
Sample Data	A sample of each data set must be acquired and reviewed	
Data Structure	The structure of each data set must then be determined	**Transform**
Data Normalizing	A series of normalizing processes must be applied to the sample data	
Attributes of Interest	For each data set Attributes of Interest must be extracted from the normalized process by writing code in very technology-oriented tools native to the Hadoop eco-system such as Pig, Hive or Java map-reduce programs	
Data Aggregation	Aggregation code must be written to combine multiple data sets into a single normalized view	
Sample Testing	A derived normalized, flat file must be tested against the corporation's BI visualization tools	**Test & Fix**
Analytic Tests	A series of analytic tests are performed to test the functionality of the combined data and find edge cases	
Load Raw Data	Raw data for each data set is imported and the code is applied so the analyst can view the data holistically	
Data Errors	Discrepancies in the results caused by incomplete data, data not normalized or aggregated correctly, requires code change	

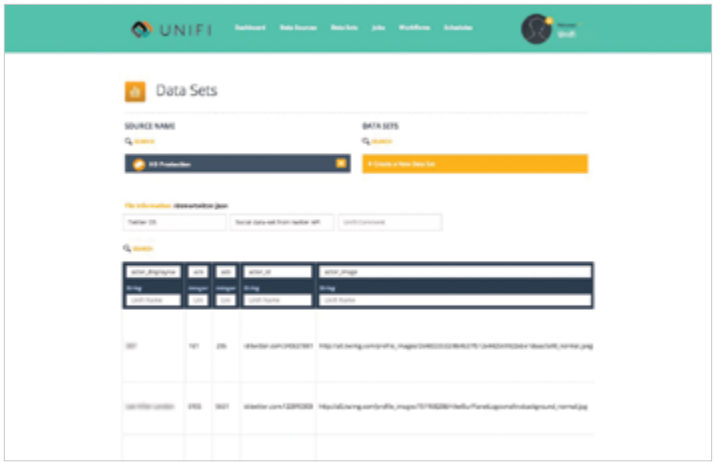
Business Insight

The Nature of Things

Social media data can be classified as semi- or multi-structured data. That is, some structure exists, but much of the data must be parsed and normalized before any analysis can occur. The data often consists of multiple levels of nesting based on the various linear and non-linear parameters being reported. For example, time-based events can be considered linear and structured, whereas trend data based on Tweets, Retweets, Follows, etc., is more unstructured, happening randomly from a time perspective and with no obvious behavior pattern.

Normalizing the Data

Before any data can be analyzed, the business analyst must explain to a programmer in the IT department what question she is trying to answer or what problem she is trying to solve. This is a huge challenge—the analyst may not know exactly which data sources are available and therefore may not even know about some pivotal information she could request to have included. The IT professional is familiar with the data sources and has the technical expertise to structure the data into usable files, but may not know enough about the analyst's specific business case to include all of the relevant sources and attributes. This can mean incomplete data and frustrating, time-consuming iterations.



Twitter data processed and normalized and viewable by the visualization tools

Before any analysis on the data can occur the business analyst and the IT programmer must follow most or all of the following steps.

Even with this overly simplified description of the process that business analysts face daily, it's no surprise that many insights are delivered late or not at all. The process is challenging—make a non-technical analyst explain in business terms, to a non-business technical person, what the end result of the analysis should be. One senior data analyst at a Fortune 500 company explained that often analysts will not even try to explain to the IT department what it is they are trying to accomplish because it is just too difficult.

Normal Operations Will Resume Shortly

The good news is that there is help in sight. Two key developments are occurring in the analytics space that will dramatically improve the life of both the business analyst and the IT department that supports them. First, less expensive, centralized storage and the processing of semi-structured and multi-structured data natively in Hadoop and second, substantially more capable visualization tools that help analysts glean the nuggets of insight from the mountains of raw data.

New emerging companies such as Unifi Software, Tamr, and Trifacta, along with evolving industry stalwarts like Syncsort and Teradata, are helping companies integrate their data more quickly by applying complex data parsers and data logic where software developers previously had to hand code the integration with very technology-oriented tools native to the Hadoop eco-system such as Pig, Hive or Java map-reduce programs.

Unifi Software has taken integration even further and offers a suite of tools that let business analysts select and integrate their data sets in a self-service fashion without writing a single line of code. This allows the analyst to pursue "what if" scenarios with their data and develop the business insights they need in a fraction of the time of traditional hand-coded programming.

About the Author

Ayush Parashar
Co-Founder & VP, Engineering UNIFi Software, Inc.

Ayush has deep software engineering expertise around big data solutions & has strong domain knowledge around Hadoop, MPP database & Systems, Performance Engineering & Data Integration. Before UNIFI he was part of the founding engineering team at Greenplum.