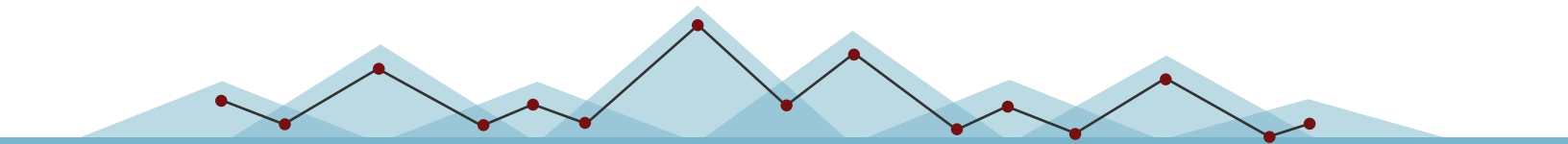
Open Analytics



AN OPEN FUTURE

TRENDS, TIPS, SOLUTIONS AND METHODOLOGIES FOR OPEN ANALYTICS

sponsored by: IKANOW kognitio

INTRODUCTION

A little more than a year ago, we started our first Open Analytics Meetup group in Washington, DC.

Within the first day we had 100+ members and a few weeks later organized our first Meetup, which had close to 100 attendees. New to the Meetup scene, it was a somewhat stressful but overall fantastic experience, and it became clear we were onto something special.

Six cities (DC, NYC, Boston, San Fran, Chicago, and London), 15 Meetups, and 2,500+ members later, we've realized there is huge interest in Open Analytics. Working for an open source, analytical software startup, this was encouraging news. The whole purpose of creating these groups was to bring like-minded individuals together to discuss open source technologies, data analytics, and the benefits, outcomes, and impact these solutions have on organizations.

Over the last few months we've been collecting input, feedback, and suggestions from the Open Analytics community, which helped us create this Open Analytics Guide. This will be an ongoing project, so please continue to provide input and feedback as we'll be releasing future eBooks.

We hope you enjoy this eBook and look forward to many future events together.

If you have something to share, we want to hear it. Visit OpenAnalyticsSummit.com/advice to contribute to the project.

Regards,

Scott Raspa

@sraspa *http://www.linkedin.com/in/scottraspa*

AN OPEN FUTURE:

CURRENT STATE OF OPEN ANALYTICS

by Cassie Lancellotti–Young, VP of Client Optimization & Analytics *@ Sailthru*

@dukecass *http://www.linkedin.com/in/cassyoung*

> No single solution covers all the data, from all the sources, for all the answers you need. Read widely and often.
>
> —Anonymous

The recent proliferation of Big Data has had good, bad and ugly effects with regards to the state of open analytics.

The good news: data is everywhere, meaning we have more data about our businesses than ever before and moreover, that we can access that data faster than ever – whether we're talking about customer insights, site analytics, in-app behavior, email metrics, purchase trends or even customer service touch points.

Additionally, we are living in the era of the "always addressable" customer. Indeed, Forrester recently confirmed that two-thirds of people under 45 in the US are always addressable, meaning they utilize multiple devices to frequently access email, apps and more from multiple locations; in other words, there are unrivaled new opportunities to get information in front of users.

So what's the problem, you ask? With big data have come big silos, which often lead to a very ugly outcome: ignorant marketers. I recently received an email from American Express with subject line "Cassie, check out our updated mobile app." In addition to giving me calls-to-action to download either the iOS or Android app (despite the fact I have the iOS app, use it regularly and thus find the Android messaging completely irrelevant), I had actually already downloaded the most recent version of the app, so when I tried to click through to download

this "updated" app, there was nothing there for the taking. This problem is not isolated to American Express; a 2012 Columbia Business School study around marketing in the era of Big Data found that 39% of marketers have difficulty turning data into action.

Open analytics help to paint a 360-degree portrait of businesses and can thereby help marketers and other stakeholders alike improve the signal-to-noise ratio in their big data. At Sailthru, we call this big data alchemy of sorts Smart Data and look to provide businesses with a single platform for housing all pertinent data points on a customer so that they can easily develop actionable insights. Our approach is possible thanks to the API economy; indeed, the availability of open APIs has climbed steadily year-over-year since 2005 and has quite frankly exploded since 2011.

Open analytics help to paint a 360-degree portrait of businesses and can thereby help marketers and other stakeholders alike improve the signal-to-noise ratio in their big data.

APIs are the backbone of open analytics; they allow businesses to stitch together key data from a variety of sources to paint a comprehensive portrait of what's happening at the customer level or in the business ecosystem. That said, the future of open analytics is likely to rely heavily on API aggregation tools (that are still mostly in fairly nascent stages).

What do you know about the current state of analytics?
Click to share or tweet #oatips

AN OPEN FUTURE:

HOW TO IMPLEMENT OPEN ANALYTICS

by Charlie Greenbacker, Principal Data Scientist *@ Berico Technologies*

@greenbacker *http://www.linkedin.com/in/greenbacker*

> Never ever advocate for one methodology! You don't want to be a hammer looking for nails if you truly want to solve problems. Instead, be a tool kit with various tools and methodologies and build relationships with other people that have a variety of tools and methodologies and attack a problem together.
>
> — Jen Leong, *Booz Allen Hamilton*

To successfully implement open analytics, three main ingredients are fundamentally indispensable: the right people, the right data, and the right tools.

Good help is hard enough to find; finding good data scientists is even harder. When building data science teams, it's important to look for a mix of complementary skills distributed among multiple team members. There is no such creature as "the quintessential data scientist." Most data scientists are "T-shaped" individuals, with a breadth of knowledge across many different technical areas and deep expertise in a few. They need to know how to ask the right questions, how to assemble the resources necessary to answer them, and how to communicate their answers to the appropriate audience. Don't forget to make sure your team includes folks with elite software engineering skills, as well as those with relevant subject matter expertise.

Everything depends on having the right data. The best people and the best tools in the world are useless without adequate data on which to run their analytics. Data may be collected from many sources, including internal systems, third-party applications, and very often directly from the Internet itself. Wherever it comes

from, this data must be relevant to the questions you're trying to answer, it must be stored in a way that facilitates the development and execution of analytics, and it must be easily accessible by the scientists & engineers who need to work with it.

Although there are many high-quality commercial tools for big data analytics, most analytic capabilities can be implemented using freely-available open source software and a bit of custom development effort. Any organization that cares about minimizing costs and maximizing flexibility should consider developing an open analytics platform based on open source tools. The following is a very brief and incomplete survey of some of the many such tools that are widely used to power open analytics.

Distributed Computing

The term "big data" analytics implies you're working with sets of data that are at least too large to be processed by a single machine, so you'll need a distributed computing platform upon which to build your analytic tools. Hadoop is the reigning king of this space, combining a scalable file system (HDFS) with a programming model for parallel computing (MapReduce). However, Hadoop is really only designed for batch processing, so if you need the ability to analyze streaming data in real time, you might want to use a platform like Storm instead.

Data Storage

No matter what your data looks like, you have to store it somewhere. MySQL and PostgreSQL are standard open source options for relational databases to store structured data. For a more flexible data model, and to take advantage of certain scalability & performance gains, many organizations are now opting for so-called "NoSQL" databases. These include column-oriented stores like HBase and Cassandra, document databases like MongoDB and CouchDB, key-value stores like Redis and Riak, and graph databases like Neo4j.

HOW TO IMPLEMENT OPEN ANALYTICS

Quantitative Analysis

Analytics and data science require a lot of math – often really complicated math involving advanced numerical & statistical analysis. Fortunately, many open source tools are available to help get the mathematics under control. R is a programming language for statistical computing and graphics that is widely used by statisticians and data scientists for statistical data analysis. NumPy and SciPy are extensions to the Python programming language that together provide a large library of high-level math functions and tools for scientific programming.

Machine Learning

Many analytic tools are based on machine learning approaches, including classification, clustering, and recommender systems. Mahout is a collection of distributed and scalable machine learning algorithms designed to run in a Hadoop cluster. Scikit-learn is a machine learning library for Python featuring support vector machines, logistic regression, naïve Bayes, and k-means clustering. Other machine learning toolkits include Weka, Dlib-ml, MALLET, Shark, and Waffles.

Natural Language Processing

A massive amount of potentially useful business information originates as unstructured text documents. The value of this unstructured data is just waiting to be unlocked by tools & techniques related to natural language processing. Stanford CoreNLP and Apache OpenNLP offer Java-based tools for parsing, tagging, and extracting entities like people, places, and organizations from raw text data. NLTK is the natural language toolkit for Python, providing a suite of text processing libraries.

Geospatial Analysis

Our community is seeing a surge in demand for geospatial analytics, as "where?" becomes an increasingly important question within the modern mobile society.

HOW TO IMPLEMENT OPEN ANALYTICS

PostGIS is a spatial extension to the PostgreSQL, adding support for geographic objects and location queries. GeoServer provides standardized access to GIS data and maps, allowing users to share and edit geospatial data. OpenLayers is a JavaScript library for displaying map data in web browsers. CLAVIN, created by Berico Technologies, is a package for geotagging and geoparsing unstructured text documents that employs context-based geographic entity resolution.

Data Visualization

No analysis is complete without effectively conveying a message to the intended audience. Most human beings prefer intuitive visual representations of data, such as infographics, rather than to stare at numbers in a table or watch them scroll across a screen. D3.js is a powerful JavaScript library for displaying data in visually-stunning graphical forms inside web browsers. Gephi is an interactive visualization and data exploration software platform for network analysis and graph manipulation.

What advice do you have on implementing open analytics?
Click to share or tweet #oatips

AN OPEN FUTURE:

THE IMPACT TO ORGANIZATIONS

by Alex Poon, co-Founder & VP of Engineering @ *Visual Revenue*, an Outbrain company

@Alexpoon06 *http://www.linkedin.com/in/alexchibunpoon*

> It may be wise to segment different participants based on their industry, in addition to a stated desired focus for analytics.
>
> —Anonymous

Open source software and particularly Open Analytics tools and software packages are dramatically transforming modern organizations.

As more companies dive head first into exploring Big Data, OpenSource Analytics tools have become a critical part of the journey. Over the past 10 years, the open source movement has cultivated amazing tools to gather, clean, transform, visualize, and model data. Engineers today truly stand on top of shoulders of giants. A few skilled engineers can perform the same, if not better work that used to take hundreds of people. These dramatic changes affect how new products are developed, decisions are made and teams are created.

Open Analytics tools enable modern organizations to build scalable and powerful products faster with less resources. Yet, these benefits come with their own unique set of challenges.

When engineering a new product or feature in the open world, it is usually worth the time to research the current progress the community has already made in the subject matter. This is counterintuitive to traditional engineering teams. Engineers, hackers, and tinkerers love to build stuff and most have emotional ties to their own code and have trouble building on top of other people's work.

THE IMPACT TO ORGANIZATIONS

A successful organization must proactively battle this "not built here" mentality. To take this a step further, one could say that a modern organization must embrace the open community. We must treat the open development community as an extension of our internal teams.

Building on top of the community's progress should only be the first step. Our teams should actively participate in the conversation and when appropriate, contribute bug fixes and/or features to better the community. Companies like Google, Twitter, Yahoo, all have large teams contributing to opensource projects daily. Some of the top open analytics projects such as Hadoop, Hive, and Storm are direct results of their contributions. They have done a great service to the community but also reap the benefit of having a large pool of talented developers to constantly improve some of their core services for free.

The shift in mindset doesn't apply only to developers, it also applies to users.

a modern organization must embrace the open community. We must treat the open development community as an extension of our internal teams.

Managers, analysts or anyone who wishes to gain insight into their surrounding for that matter, need to move away from a deterministic world view to a statistical/probabilistic one. As better tools enable us to gather and process more and more data, many of our decisions can be quantified. The challenge is that these quantified results are usually not black and white. Sentences like "we are x% confident in the results" or "we foresee events happening with y% probability" are common. One must understand the implication of these statements and change the decision process accordingly.

Well, what do all these changes mean to a development team? Instead of placing strong emphasis on the ability to build everything from the ground up, skill sets such as Googling and hacking are becoming more relevant. For most use cases one can think of, there are at least one library or GitHub project for them. To build something amazing, the best path is often to research what is already

available within the community and figure out how to configure and "glue" different technologies together.

Another critical skill is the ability to evaluate the strength and potentials of open source projects. The maturity, development momentum, and community support play a huge role in whether a particular project would be successful in the long run. Choose wisely.

For example, at Visual Revenue, we have positively struggled with scaling our data pipeline for the whole history of the company since we gained initial traction with the product. We were handling billions of monthly web traffic with our homegrown solution but significant engineering effort was required to constantly battle scale. A long long time ago, in a Galaxy far far away, we decided to migrate to a Kafka/Storm stream data process infrastructure (thanks LinkedIn & Twitter.) Since then, we have been able to double and triple our traffic load with minor configuration changes and renting more machines from Amazon. Engineers who worked on the homegrown solution get to refocus themselves on other technical challenges and exciting features. It was simply beautiful.

Open Analytics brings tremendous opportunities to better utilize the data around us. Teams can run a magnitude faster and build solutions that were previously impossible. To take advantage of this opportunity, organizations must change their development mentality, decision making process, and the make up of the overall team. It is a brave new world out there for us to explore.

CONCLUSION

We would like to thank the Open Analytics community for their contributions to this eBook. We plan to create additional eBooks so keep the suggestions, input, and feedback coming by clicking here.

This eBook was sponsored by:

Ikanow is focused on removing the current gaps in technology and knowledge to help open-source approaches gain mass appeal and adoption among enterprises and technology vendors alike. Ikanow solves the underlying issues of synthesizing structured and unstructured data for analytical consumption. Their goal is to enable mission agility by reducing the time required for analytical business value.

Ikanow's Infinit.e platform is today's leading Open Source intelligence analysis platform. Ikanow's commitment to Open Source means their users are never locked into closed and proprietary code. The platform and source code are transparent, so other departments and organizations can make and publish changes to meet their specific needs. Infinit.e is based on the most popular Open Source, Big Data technologies - like Hadoop, Lucene, and MongoDB. Therefore, its engine is continually being improved by the large and rich communities who work on and with those platforms.

For more information on Ikanow or our product offerings,
visit http://ikanow.com/.

CONCLUSION

For more than a generation, Kognitio has been a pioneer in software for advanced analytics, helping companies gain greater insight from large and complex volumes of data with low latency and limitless scalability for competitive business advantage. Sitting at the nexus of Big Data, in-memory analytics and cloud computing, Kognitio extends existing data and BI investments as an analytical accelerator, providing a foundation for data scientists and analytical information services. The Kognitio Analytical Platform can be used as a data science lab or to power comprehensive digital marketing analytics; it runs on industry-standard servers, as an appliance, or in Kognitio Cloud, a ready-to-use analytical Platform-as-a-Service (PaaS) in a public or private cloud environment.

To learn more, visit www.kognitio.com and follow us on Facebook, LinkedIn and Twitter.