MOBILE

BIG DATA
REPUBLIC | Transform Your Business With Data

ABOUT US   REGISTER   LOGIN

HOME   BLOGGERS   MESSAGES   POLLS   WEBINARS   RESOURCES   VIDEOS | STRATEGY   **ANALYTICS**   TECHNOLOGY   FB   TW   LI   G+   RSS

FINANCIAL SERVICES   HEALTHCARE   GOVERNMENT   EDUCATION   RETAIL

BIG DATA ANALYTICS

Case Study: Kaggle's Data Scientist as a Service Appears

Robert Plant, Associate Professor, School of Business Administration, University of Miami
9/18/2013
Comment
13 comments

Login

👍 👎
50% 50%

For many organizations the gap between a conceptual understanding of big data to the possession of a capability for execution is limited by their desire to embrace the challenge of managing the process.

In part this may be driven by the internal limitations of their organizational structures, their existing technology functions, and their desires to limit expenditures, both in terms of human capital and technology capex. This is understandable. The world of big data can be opaque. The technology, such as Hadoop is divergent from the traditional ERP-type environments upon which businesses run. The talent, in terms of the data scientist, could be seen as a potential asset, but the capital outlay and ROI may not be acceptable. The management structures to undertake big data project deployment are for many a black box of uncertainty. With these elements combined, there is a hard business case for a manager to make and sell to the senior management.

Data science as a service
However, there is an alternative, one that holds significant value for firms wishing to grow their big data knowledge base, limit their cost liability, and manage the risks. That is the Data-Scientist-as-a-Service model. One firm that specializes in providing this service is Kaggle. Founded in 2010 by Anthony Goldbloom, Kaggle began life in Melbourne, Australia, as a platform for hosting public-data science competitions, where data scientists compete for the prize of developing the most effective solution for a given problem. Kaggle recognized both the demand for data scientists and firms in need of a methodology for understanding big data problems. Backed with $11M in venture funding, Kaggle moved to San Francisco, where it evolved into a two-sided marketplace that matches data scientists with clients in need of their services.

The company provides two vital roles: It provides a structured methodology for companies inexperienced in the processes around big data, and it provides data scientists to solve their analytical problems. The data scientists are drawn from a talent pool of over 115,000 members, many of whom are attracted to Kaggle's continuing role as a competition platform. The Kaggle methodology has four stages:

1. Data collection. The client organization is directed in terms of data gathering. The resultant data is examined by Kaggle's data scientists, who make determinations of its relevance. The company stresses the need to present the data in its raw form, rather than after being cleaned up. This transformation process, if left in unprepared hands, can lead to information loss and to reducing the deductive capabilities of the algorithmic analysis.
2. Problem definition. In this phase the client organization has to specify the goals of the project as well as expectant deliverables. These include topics such as predicting (customer) lifetime value, wallet share estimation, (sales) lead prioritization, fraud detection, and resume screening.
3. Purchase analytical resources. Having defined the problem, the next task is understanding what resources are required in order to execute a solution for that problem definition. Kaggle matches the client with one of their data scientists, who enters into a problem elicitation dialogue with the client to determine an estimation of the resource requirements to solve the problem. (Kaggle only uses the top 0.5% of its members in its analytical group.)
4. Data sharing. The data scientist acts as an extension of the client organization's team and works with them to solve the problem.

Finally, the Kaggle Data Scientist presents the client with any intellectual property assets created, including analytical source code, data transformation processes, cleaned and enriched data, and the ultimate prize, the analytical results.

MORE BLOGS FROM ROBERT PLANT

A Review: Big Data in 2013
39 comments
Data scientists, NSA, C-suite buy in... Big data changed a lot in the last 12 months.

What Would Seymour Cray Think of Big Data?
13 comments
How would the tech legend respond to the 4 Vs?

The 360 Degree View of a Customer
56 comments
Big data is helping companies like Netflix see what you really want (and even when you go to sleep).

Lessons for Big Data From President Obama's Healthcare Implementation
24 comments
Three big data lessons for CEOs from HealthCare.gov's bumpy start.

Is Big Data Ruining Sports?
29 comments
From sailing to the NBA -- big data could increase the gap between sporting haves and have-nots (while ruining the magic in the process).

More from Robert Plant

FLASH POLL

All Polls

DIGITAL AUDIO

Big Data Analytics Radio
Latest Archived Broadcast
Your Big Data Initiative Starts With a Great Team
The Big Data Conference and Nemertes Research have teamed up to provide expert advice on building a big data team and finding the right leader. Join us on Tuesday, Aug. 20.
Listen

Full schedule   Archived shows   Download to calendar

DATA VISUALIZATION SHOWCASE

Visualizing Sporting Value: Football Player Transfers

⚽ Player Transfers (1900 - 2013)
Number of players vs Transfer Fees (transfer from country)

A great visualization of transfer fees leads us to think of deeper sporting analytics.
Explore this data here.

More Data Visualization Showcase

Who outsources big data?
This outsourcing approach to big data is not merely the province of small firms finding their methodological way towards a new analytical insight, but also by firms such as Tesco who have their own analytical firms, yet look to improve lifetime customer value from Kaggle's data science approach. These firms include Microsoft, which identified broader categories of gesture recognition for their Xbox Kinect; Ford Motor Company, for earlier detection of driver drowsiness; NASA, for a more accurate imaging of dark matter; and Boehringer Igelheim, for better predictability for drug targets.

Clearly, these companies contain the talent and resources to develop their own big data solutions. However, the Kaggle approach is useful because data scientists can be specifically selected for and matched with a target project. Each scientist has his or her own specific talent set that can be applied to the problem. As the resultant work products are provided to the company, that organization benefits from, not only the resultant solutions, but also the ability to learn from experienced data scientists, and then feed this knowledge back into their own organization's big data knowledge pool.

Perhaps the solution to big data is not to dedicate internal big data groups but to look outward for nimble, agile solutions from marketplace vendors and then let the market provide the best talent for a specific problem domain. The result would be a unique individual solution to each big data problem. This is not a solution model that would naturally result from an internal group within a firm. This also has the potential of producing a lower overall total cost-fee basis through the Data Scientist-as-a-Service solution model -- a potential win-win for both sides.

Related posts:

- Michael J. Fox Foundation Points Big Data at Parkinson's
- Finding Unexpected Value in Data
- Machine Learning Case Studies

Email This   Print   Comment

COMMENTS

Newest First | Oldest First | Threaded View

PAGE 1 / 2 > >>

AlphaEdge, User Rank: Exabyte Executive
9/30/2013 | 9:56:12 PM

Re: The process
Kaggle indeed leads the data as a service trend.

Login

50%   50%   Reply | Post Message | Edit/Delete | Messages List | Start a Board

Pradeepta Mishra, User Rank: Exabyte Executive
9/30/2013 | 1:37:31 PM

Re: The process
Kaggle has adopted the data as a service model to mobilize innovation and bringing best knowledge on board. It not only provides best in class solution it also educates aspiring data scientists. Hence it educates and spread knowledge around implementation of big data projects.

Login

50%   50%   Reply | Post Message | Edit/Delete | Messages List | Start a Board

Ariella, User Rank: Blogger
9/21/2013 | 9:41:20 PM

Re: Old model, new solution.
@Daniel that's true, though usually the prizes are just 5 figures, and often in the low 5 figures.

Login

50%   50%   Reply | Post Message | Edit/Delete | Messages List | Start a Board

Numair Mansur, User Rank: Megabyte Messenger
9/21/2013 | 3:27:33 PM

Re: The process
No @James -They don't send people on site, Kaglle is an online competition that helps an organization to create the best predictive model for their specific problem. Players (Data Scientists) develop different predictive models to that specific problem as there are many different ways a data mining problem can be solved (KNN, Neural Networks, Naive Bayes). Lots of data scientists attacking the same problem using different approaches means that best approach has to outperform all the different approaches. And the best part is, at the end of the competition, the winning model gets a prize ! (typically cash)

Reply | Post Message | Edit/Delete | Messages List | Start a Board

Daniel Gutierrez, User Rank: Blogger
9/20/2013 | 7:14:11 PM

Re: Old model, new solution.
Then there was the situation centered around Kaggle's Heritage Health Network $3 million competition last year. It was rumored that a New York financial services firm "hired" one or more machine learning specialists just to work on the Heritage project. I heard they were paid $300,000 per annum, which still would be pretty good profit for the employer if their entry won.

Login

50%   50%   Reply | Post Message | Edit/Delete | Messages List | Start a Board

Saul Sherry, User Rank: Blogger
9/20/2013 | 4:38:26 AM

Re: Old model, new solution.
Makes sense @legalcio - but I guess if Kaggle can connect you with individual analytical geniuses, you can take your pick.

Login

50%   50%   Reply | Post Message | Edit/Delete | Messages List | Start a Board

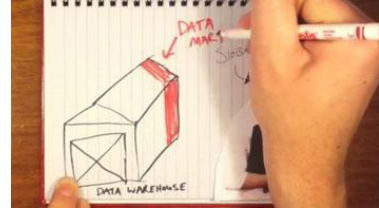**legalcio**, User Rank: Exabyte Executive
9/19/2013 | 9:30:55 AM

BDR IN YOUR INBOX

Enter email   [SUBMIT]

FEATURED VIDEO

Big Data Explained: What is a Data Mart?   18

Use this video to explain quickly to someone the difference between a data warehouse and a data mart.
Watch This Video

More Video Blogs

FOLLOW US ON TWITTER

Big Data Republic@BigDataRepublic   15 Apr
AllJoyn: A Common Language For Internet Of Things - Open-source lingua franca aims to let IoT devices and services... ow.ly/2FHbY7
Retweeted by Yakamoto
Expand

Hajime Ozaki@hajozaki   22h
BDR Top 20 #BigData David Smith David Smith List of top tweeters on Big Data from @BigDataRepublic wp.me/p4Aavi-2
Show Summary

Marcus Borba@marcusborba   28 Apr
#BigData #Analytics #BusinessIntelligence #FS @informationweek @AllAnalytics @BigDataRepublic @BigDataNetwork @BigDataClub

LIKE US ON FACEBOOK

Big Data Republic
Like

4,168 people like Big Data Republic.

Facebook social plugin

ACCOLADES

MEDIA PIONEER AWARDS WINNER 2013   Media Pioneer Award Winner

Re: Old model, new solution.
That depends Saul. The try and buy approach works well if I've engaged a contractor either on my own or through a third party. If the individual works for a company that I've outsourced work to then I'd be reluctant to poach them. This is a business of relationships and if I've found an outsourcer I work will with I'd rather not screw that up.

Login

50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

James M. Connolly, User Rank: Blogger
9/19/2013 | 9:27:58 AM

The process
@Robert. I'm curious, does Kaggle provide their service only on a remote basis, or do they send people onsite to work directly with the client on the data collection, problem definition, and buildout aspects? I ask because it sounds like in-person would be so much more interactive and effective, compared with trying to do it from 3,000 or 6,000 miles away.

Login

50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

James M. Connolly, User Rank: Blogger
9/19/2013 | 9:24:52 AM

Re: Old model, new solution.
I like this concept. I'm sure other readers came into the blog thinking that it would be good for small organizations, but, as Robert notes, there are big outfits using it too. (However, it does appear that they are using it for fairly niche purposes, possibly because they are venturing into areas outside their core expertise).

It makes no sense for a company to go out and build an infrastructure for big data and hire a team if they have no basic knowledge of big data. In this case they hire a guide who can walk them through the process. Then, they can move up the learning curve and build out additional BD capabilities inhouse, or they can stick with Kaggle on an as-needed basis.

Great idea.

Login

50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

Saul Sherry, User Rank: Blogger
9/19/2013 | 5:50:33 AM

Re: Old model, new solution.
@legalcio - as a CIO do you also look at this as a potential 'Try before you buy' approach to hiring?

Keep your personality tests - you know that this girl has the chops as a data scientist, understood your business and delivered results.

Should a fulltime postition come up, surely she's first in the queue?

Login

50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

PAGE 1 / 2 > >>

BIG DATA REPUBLIC