

SPONSORED BY:

MOBILE



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](#)

**HADOOP**

**BYOBD: Bring Your Own Big Data**



**Robert Plant**, Associate Professor, School of Business Administration, University of Miami  
 12/12/2012  
 Comment  
 49 comments

[Login](#)  
 50% 50%

The hack/reduce model of experimenting with data should be embraced within the organization, because it could lead to a combination of four rewarding results.

Many organizations would like to experiment with big data, but they are put off by the setup costs, a lack of internal expertise, and an unclear understanding of the outcomes they may expect to receive. Clearly, this is a poor business case to present to the CFO signing off on the project and its budget.

One alternative is to look for external entities that can provide the necessary components on an outsourcing arrangement. Unfortunately, not too many environments support this type of computing need. As such, perhaps now is a perfect time for you and your organization to promote the idea of developing a collaborative big data venture -- in essence, establishing a facility with partners to leverage the technology, people, and resources needed to succeed in big data.



**The hack/reduce model**  
 One model that could be used as a framework is the nonprofit [hack/reduce](#) organization based in Cambridge, Mass., whose tagline is "Code big or go home." Hack/reduce is a partnership involving the state of Massachusetts, local firms, MIT, Harvard, and several other universities. It links participants with a host of industry sponsors (Google, IBM, Microsoft, Dell, etc.) and specialist firms such as GoGrid, a provider of cloud hosting

and automated provisioning architectures, and Greenplum, a division of EMC that offers analytics solutions. There are also venture capital partners such as Bessemer and Bain Capital, making that potential for a big data venture even more accessible.

Though hack/reduce's membership is based on a meritocratic system (Fellows are gurus, and Resident Hackers are programmers working on "something incredibly cool"), there is room for more mortal types (as Contributors) who want to learn and collaborate on big data projects.

By following the hack/reduce model, big data development mashups could be created anywhere, but the best target locations are within technology centers. In the UK, tech clusters such as London, Reading, Manchester, Oxford, and Cambridge are obvious targets. However, they could be established anywhere with the potential for building a cluster of industry players, academic computing, and data-driven business. It would also be ideal if these ventures were supported by private equity or venture capital investments.

**The payoff**

What would the potential payoff be? Four outcomes would result.

1. In the BYOD model, the partners would share the facility costs. Many organizations already have high-performance computing clusters and systems. The costs of establishing and running them can be reduced by pooling resources into a common datacenter. Software licenses and operational costs would also be shared, and vendors supporting the project could be persuaded to act as collaborators, rather than purely as vendors.
2. The learning curve would be reduced for all participants. One of the biggest barriers to entry in the big data space is access to talent. The BYOBD approach would allow easier access to talent and training from vendors and partners. University resources could be leveraged in a centralized location, with students engaged in projects eventually becoming highly talented potential employees.
3. The environment could spin off new entities and businesses, creating a virtuous cycle of innovation that further builds the talent pool for big data. One way to achieve additional levels of ROI is to encourage innovation. By allowing startups to use their facility, large enterprises could buy in to these startups, develop them, and spin them off for a profit.
4. You can use these facilities to literally bring your own big data and exploit its capabilities.

In light of these advantages, executives need to act fast, move out of their silos, and collaborate. It has been done before in other industries: Covisint in the automotive supply chain, GS1 in B2B data,

**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**

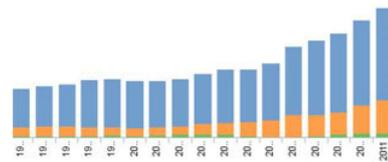


**Hadoop Radio**  
 Latest Archived Broadcast  
**Stop Kidding Yourself: Cloud Realities You Need to Embrace**  
 Join this radio show to truly understand what a CIO needs to do to build a successful private cloud and what skills and values the IT team will need to embody.  
[Listen](#)

[Full schedule](#) [Archived shows](#) [Download to calendar](#)

**DATA VISUALIZATION SHOWCASE**

**Anatomy of a Visualization**



This Tableau visualization of international debt demonstrates how simple visualizations can give great insight. Explore this data [here](#).

[More Data Visualization Showcase](#)

**BDR IN YOUR INBOX**

Enter email

healthcare exchanges in the insurance industry, and Orbitz in the airline industry. The first movers in these collaborative spaces got the best ROI, an early-mover advantage, and leverage. The laggards lost out, and many are fighting to recoup lost marketshare.

Success in big data requires thinking outside the traditional datacenter box, and the hack/reduce model may well be the place to house those thought processes. At the end of the day, big data is not traditional data processing; it's more akin to technology innovation, and that requires an entrepreneur's mindset. Where better to begin than by building a big data startup? Who knows? It may grow to be bigger than your current business.

Related posts:

- [Free Big Data Education: A Management Perspective](#)
- [How to Hire a Qualified Data Scientist](#)
- [Hadoop Won't Turn Your Dump Into a Data Warehouse](#)

— Robert Plant, Associate Professor, School of Business Administration, University of Miami

[Email This](#) [Print](#) [Comment](#)

COMMENTS

[Newest First](#) | [Oldest First](#) | [Threaded View](#)

PAGE 1 / 5 > >>



Saul Sherry, User Rank: Blogger  
3/21/2013 | 9:51:28 AM

**Re: steps to gain benefits**

Hi Ahmed... have you made any progress with this project?

[Login](#)



Saul

[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)



Keith Grinstead, User Rank: Petabyte Pathfinder  
2/28/2013 | 8:37:12 PM

**Re: A collaborative future**

@Robert picking up on your entrepreneurial angle it has been identified in other discussions that there are solutions needed out there and certainly someone is required to co-ordinate pools of big data and merge them into seas / oceans of big data that can then be interrogated to provide solutions from this collaborative solution.

[Login](#)



[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)



ahmed farag, User Rank: Bit Player  
2/1/2013 | 3:23:21 PM

**Re: steps to gain benefits**

Hi @Saul Sherry, how are you? I've a question and i hope to help me.

[Login](#)



How to acquire , organize , analyze big data using Hadoop. I mean what are the tools used in each step. And how to transform the unstructured data with the structured data.

[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)



ahmed farag, User Rank: Bit Player  
1/25/2013 | 5:46:38 PM

**Re: steps to gain benefits**

Thank you guys, i find an open source c# code that crawl "garbage" plenty html tags that need more and more efforts to parse these tags into meaningful data. you mentioned legalities but i want to crawl the public data only, by the way i'll take care of this point.

[Login](#)



I hope i can found a short way to crawl and parse the data. @AlphaEdge why you said that JAVA or Python are the best ? could you give me more explanations please? also if you know any kind of open source code written by JAVA or Python share it to me

[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)



AlphaEdge, User Rank: Exabyte Executive  
1/23/2013 | 6:02:45 PM

**Re: steps to gain benefits**

Agree. It might involve copyright infringement related issues. I think you guys may have to seriously consider this issue. It seems like either JAVA or Python would be the best tools for webcrawler. Expecially given that JAVA integrates with Hadoop so well. Maybe that can be the tool you choose? Any other comments from other experts?

[Login](#)



[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)



Saul Sherry, User Rank: Blogger  
1/23/2013 | 6:23:16 AM

**Re: steps to gain benefits**

Hi @Ahmed. There are plenty of crawler guides out there, that all allow you to do some interesting stuff and pull interesting data - I recommend checking out pages like [stackoverflow](#) to see what people like you have done with similar needs.

[Login](#)



A caveat though - the legalities on your doing this are never particularly clear cut. According to this great article on [PeteSearch](#). They nearly got sued by Facebook for accessing such data, being told that "the only legal way to access any web site with a crawler was to obtain prior written permission."

So i would make sure you get clarity on the legal side of what you are doign first.

As from there, I would recommend taking this one step at a time - for instance, you won't know the best kind of storage system until you know the size and types of data your crawling efforts throw up. Anyone have thoughts to add?

[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)

FEATURED VIDEO

Big Data Explained: What Is YARN? 21



Getting to grips with Hadoop 2.0? Here's a simple video explaining YARN.  
[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



ahmed farag, User Rank: Bit Player  
1/22/2013 | 4:12:28 PM

**Re: steps to gain benifits**

Let me introduce myself to you, i'm ahmed farag from egypt currently i'm in a training program in business analytics and working with a team on the project that i mentioned it to you. I need to say that i just start to take my first steps in big data analytics. So please try to help me as much as you can.

[Login](#)  
50% 50%

[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)



ahmed farag, User Rank: Bit Player  
1/22/2013 | 4:04:05 PM

**Re: steps to gain benifits**

Hi @Saul, let me explain to you what i want to do and what is the case again to find out how can you and the guys help me.  
I'm working on a project that we need first to crawl data from social websites (facebook, twitter), we don't want to use the API's of facebook or twitter, we need to extract exactly the data related to the company, for example from the company official page, we need to know all of the members on the page, thier names, comments on the page per post, their likes, go to thier profiles and know their ages, their contacts if possible, and so on. Also we need to search all of the pages on facebook that speaks about the company who is thier author, his names, contacts, all possible details inside every page, and so on. Sure it will be the public data only. Also we need to apply the same concept with some other sites(one news website, and blog) we need to extract the data related to the company. Hopefully after we have these data we will merge it with some structured data from the company and apply analytics and insights in these data.

[Login](#)  
50% 50%

- so i think the scope of the project should be:
- Build the crawling & parsing engine to extract and parse the crawled data.
  - Store the unstructured data in Hbase for example.
  - Merge it with the data from inside the company.
  - Analyse the new valuable data that we have.

Now i need to know if these steps is okay? and if their is a way to crawl the data from the perspective i mentioned before ? simply how can go ahead with this project?

Any suggestions please.

[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)



Saul Sherry, User Rank: Blogger  
1/22/2013 | 9:21:34 AM

**Re: steps to gain benifits**

@Ahmed - sounds to me like this approach is to pull data from a FB page you are administrator on - is this what you were after or were you thinking of a broader pull from Facebook?

[Login](#)  
50% 50%

[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)



ahmed farag, User Rank: Bit Player  
1/18/2013 | 1:25:53 PM

**Re: steps to gain benifits**

Hi @Saul, realy i'm very thankful to you & AlphaEdge, your link is great, but i have another issue will this crawler work with facebook?. Could you take a look to these links and tell me what you understand from:

[Login](#)  
50% 50%

<http://onestopdotnet.wordpress.com/2012/06/25/talk-to-facebook-graph-api-via-fql/>  
<https://developers.facebook.com/docs/getting-started/graphapi/>

i feel like it seems to be a pip or bridge from facebook to developers

take a look and tell me what you found...

[Reply](#) | [Post Message](#) | [Edit/Delete](#) | [Messages List](#) | [Start a Board](#)

PAGE 1 / 5 > >>

**BIG DATA REPUBLIC**

[ABOUT US](#) | [CONTACT US](#) | [HELP](#) | [REGISTER](#)  
[TWITTER](#) | [FACEBOOK](#) | [LINKED IN](#) | [GOOGLE+](#) | [RSS](#)



bigdatarepublic : /bigdatarepublic/section/2635 : /bigdatarepublic/section/2635/255820



UBM TECH

OUR MARKETS: [Business Technology](#) | [Electronics](#) | [Game & App Development](#)

Working With Us: [Advertising Contacts](#) | [Event Calendar](#) | [Tech Marketing Solutions](#) | [Corporate Site](#) | [Contact Us](#) / [Feedback](#)

[Terms of Service](#) | [Privacy Statement](#) | [Copyright © 2014 UBM Tech, All rights reserved](#)