CITO Research
Advancing the Craft of Technology Leadership

Published on *CITO Research* (https://citoresearch.com)

[1]

How Vertica Was the Star of the Obama Campaign, and Other Revelations [1]

The 2012 Obama re-election campaign has important implications for organizations that want to make better use of big data. The hype about its use of data is certainly justified, but a lesser-noticed aspect of the campaign ran against another kind of data hype we've all heard: the Silicon Valley hype around Hadoop that goes too far and claims an unreasonably large role for Hadoop. One of the most critical contributors to the Obama campaign's success was the direct access it had to a massive database of voter data stored in Vertica [2].

The Obama campaign did have Hadoop running in the background, doing the noble work of aggregating huge amounts of data, but the biggest win came from good old SQL on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed.

Hadoop fans please note: The point of this story is not that Hadoop sucks and Vertica rules, but rather that there are many scenarios for putting big data to use, and in most of them Hadoop will play a role as part of an ensemble of technologies.

Josh Hendler, CTO of H & K Strategies, a veteran of the John Kerry 2004 campaign, was a consultant to the Obama 2008 campaign, and was a technology chief for the Democratic National Committee and Organizing for America, the President's campaign arm within the White House. Mike Conlow was the deputy CTO of the Obama 2012 campaign. The two presented a post-election debrief about the campaign's tech strategy to the New York CTO Club in mid-November.

"From the beginning, coming straight down from Jim Messina, the campaign manager, technology was going to be at the very center of the campaign," Conlow said.

Their story contains lessons about how organized access to a central database can be a powerful marketing force.

The Collection Process

The Obama campaign process was organized around making careful selections about telephoning or knocking on the doors of prospective voters. The central object of the strategy was to refine the

strategy so as to eliminate individuals who were not worth pursuing, identify “middle of the road” favorable voters to ask for more donations, and further engage strongly supportive voters to take specific actions, such as volunteering to canvass.

That meant collecting 10 terabytes of data, beginning with all registered voters in the nation. That data is not normalized. Each state has slightly different recordkeeping requirements. Some will display party affiliation and some don’t, for example. It’s also not centralized, and often has to be collected at the county or municipal level, Hendler said.

The next big dataset covered supporters and volunteers, all of whom received highly specific and targeted emails from Barack and Michelle Obama, Vice President Joe Biden, and Messina, among others.

The third large dataset was the voter contact list, which detailed who was phoned, who was visited, and how they responded to detailed survey questions, to be used for future analytics and targeting.

These large batches were then suffused and supported with real-time updates from field workers and surveys of social media to determine who among undecided voters was likely to turn in favor of Obama, where advertising was then targeted.

One particularly clever use of social media was connecting a video called “The Life of Julia,” which depicted how the President’s policies would help a citizen throughout her life, to the viewer’s Facebook account. By the time the viewer finished watching the show, BarackObama.com had trawled through all of the viewer’s Facebook friends, associating their names and birthdates with voting records, gender, and state of residence.

“At the end of the video, when you’re presented with four people that you should share this with,” Conlow said. “Those are not four random people. It is a very calculated list of the four best people that we want you to share this with, ranked by however we’ve decided is the best way to target them. That data currently lives with Facebook, but I think there is a ton of web data out there like that that we are only beginning to use.”

The Need for a Scalable Data Warehouse

“There’s a vast amount of data cleansing that needs to go on after the collection process,” Conlow said.

To take on this massive job in 2012, the Obama team scaled from eight total engineers to 35 software engineers, a 50-person analytics team and a 100-person digital content team, for a budget of approximately $20 million, Conlow said.

The data warehouse had to scale, too. In 2008, the team had a 3 TB Netezza warehouse, but recognized that it would need something bigger to handle its ambitions for 2012. In 2010, the team did several proofs-of-concept and selected a 10 TB Vertica massively parallel processing (MPP) data warehouse for the next campaign.

“We beat the hell out of everything we use,” Conlow said, so all the big data eggs did not go in one basket. The team complemented Vertica with Hadoop, using the open-source system to support asynchronous, ad-hoc queries.

“The decision was made to have Hadoop do the aggregate generations and anything not real-time, but then have Vertica to answer sort of ‘speed-of-thought’ queries about all the data,” Hendler added. “This is really important, because it used to be the case that only one or two people would

have access to the entire dataset. And even having 50 people who could run queries against all of it was really quite significant."

Lessons Learned

Conlow learned some critical lessons at the helm of the Obama tech team.

1. **Empower product managers.** The campaign consciously aped the product-team organization of successful Internet companies such as Google, but politics is, well, political. "There's too many people on a campaign who want to make decisions from the top," he said. "Because of that, we never let the product managers own their products in the campaign, but to do product right in business, you have to empower product managers."
2. **Employ a "query policeman."** It was important to the campaign to empower everyone to ask questions of its massive data store, but it also had to take safety precautions. Using Vertica as its data warehouse, Conlow's office needed to make sure that inexperienced SQL programmers on the team did not write a query so massive that it would crash the system. "There was a Vertica policeman who would watch the queries that were running, and if your query was too bad, it would get kicked out," Conlow said. "But we had 50 people in the analytics department who could just go at all of this data and find whatever—if they were doing voter analysis, they didn't have to be a statistician, they didn't have to build their own model. They could just explore the data and come up with whatever they were looking for."
3. **Keep the money machine on at all costs.** As vital as the engagement strategy was to the Obama campaign, it would all have been for naught if the ability to collect money went down for even a few seconds. A big part of the team technology spend went to ensuring redundancy and 100 percent uptime for the campaign's financial tools. One payment gateway was hosted on Amazon and the other was on dedicated hardware, each of which could charge to two different merchants, PayPal and CyberSource. These supported the Quick Donate function, which would queue donations in orderly fashion. The investment paid off—the campaign collected $1 billion, or $19 million per day—almost its entire technology spend.

Forward—to the Cloud

The Obama 2012 campaign's use of data did have precedents. Despite its demise, the Howard Dean 2004 campaign raised more money online than any previous campaign and was the first to recognize that technology would be important to all future campaigns, Hendler said. The Obama 2008 campaign was much more successful and raised the majority of its money online.

"The 2008 campaign was notable because there were six to eight full-time engineers," Hendler said. "In 2004, I was the only engineer on the Kerry campaign who was actually at headquarters."

In 2008, the Obama campaign had its first real reckoning with big data. The campaign used the public cloud in the form of Amazon Web Services, but the majority of its data was stored and processed on physical infrastructure in-house. Much of the campaign engineering staff was involved in setting up, running, then dismantling a data center.

In 2012, with a few exceptions, it was almost all about the cloud.

With the exception of Vertica, which ran on dedicated hardware, virtually all of the infrastructure used by the campaign was in the cloud in the form of Amazon Web Services Elastic Cloud 2 (EC2) and Amazon Relational Database Service (RDS).

“Amazon has made incredible progress in just the last 18 months,” Conlow said. “There’s a good chance our data warehouse would end up on Amazon if we had to make that decision today.”

Subscribe [3]

Source URL: https://citoresearch.com/data-science/how-vertica-was-star-obama-campaign-and-other-revelations

Links:
[1] https://citoresearch.com/data-science/how-vertica-was-star-obama-campaign-and-other-revelations
[2] https://vertica.com
[3] https://citoresearch.com/crss/node/1813