

Dems score with better data

DNC's Linux warehousing project delivered on '50-state strategy'

November 15, 2006

InfoWorld

Behind every big success these days, there's probably some darned good IT making it happen. That appears to be the case in the surprising electoral victory by the Democratic Party last week.

New data warehouse solutions commissioned by the Democratic National Committee (DNC) and also by Catalist, a for-profit group backed by a faction of leading Democratic players, are being credited for their part in the Party's strong performance in nationwide midterm elections. Those solutions may have helped Democrats close the gap with tech-savvy Republicans, according to a people involved with the projects and with the party's countrywide get-out-the-vote operation.

The DNC solution, which was commissioned one year ago by DNC Chairman Howard Dean, tapped a new generation of low-cost, Linux-based data warehouse technology to improve the quantity, quality, and availability of voter information used by state Democratic parties during the election turn-out effort. Those close to the project say the new system, part of Dean's so-called 50-state strategy, helped tip close races in the House and Senate in favor of the Democrats.

The solution was developed by Intelligent Integration Systems (IISi) of Boston, a company that develops datacenter solutions and uses a Netezza Performance Server data warehouse appliance to integrate information provided by 45 state-level Democratic parties on about 200 million voters, according to Paul Davis, IISi's CEO.

In addition to the Netezza back end and IISi code, the system uses data quality and cleansing tools from FirstLogic and enterprise integration software vendor Sunopsis, as well as data modeling tools from SPSS, according to a Netezza statement.

The new solution was hosted at a datacenter in Virginia and allowed the DNC to rapidly update so-called "voter files" as state-level party workers provided them with new information. The data was then cleaned up by comparing it to lists of known phone numbers and addresses. The DNC was also able to "overlay" the information and match it to data about individuals in the lists culled from various consumer data stores, Davis said.

Netezza, which makes the technology used by the DNC, is part of a new generation of data warehousing companies that are using commodity hardware such as Seagate hard drives, Intel processors, and hardened Linux operating systems to create low-cost, fast data warehouse appliances, according to Donald Feinberg, of Gartner.

Like incumbent data warehouse players such as Teradata (part of NCR), Netezza uses distributed database intelligence, in which data filtering, processing, and analysis is done on the same device that stores the data.

"They have code running on the hard drive, so you can parallelize the queries and do them as fast as you can lift the data off the hard drive. Fundamentally it results in a two order of magnitude improvement in speed," said Rich Zimmerman, IISi's CTO.

Parallelizing queries to databases is nothing new. However, running parallel queries on inexpensive hardware and software, like Linux and PostgreSQL, and being able to match what high-end vendors like Teradata can offer is new, said Feinberg. Appliance-based products like Netezza's Performance Server are also easier to maintain, requiring less staff and keeping the cost to implement and run the data warehouse low, he said.

Motivating the DNC's data warehouse project was an effort to improve on the organization's 2004 voter targeting project, which was roundly criticized for providing state-level organizations with inaccurate data in a close race against a well organized Republican opposition.

Gus Bickford, a DNC National Committeeman and voter file expert, says that in the 2004 contest, the DNC had reams of data on voters, but had done little "modeling" to piece it together. "It was like buying all the pieces for a jet engine, but not telling anyone how to put it together."

For example, the DNC outsourced data "cleansing" of state-level voter information in 2004, but got shoddy results, with many records having incorrect phone numbers and lacking multiple addresses that are often necessary to locate voters. State parties that tried to use that data afterwards, in some cases abandoned it altogether because it was unreliable, he said.

Following the DNC playbook for that election year, the group also stopped cleansing data for all but about 18 "swing states," writing off the rest to focus resources where the party felt they mattered most, according to Bickford.

In comparison to 2004, the new system was fast enough to digest updated voter information from all 45 participating states and cleanse it two or three times before Election Day. That meant that state-level operatives got useable data for all 45 participating states, said Bickford and Sullivan.

"The thing I noticed while I was driving around from state to state is that the volunteers were so much happier," said Sullivan. "The difference between phone bank volunteers having, say, 45 percent of the phone numbers accurate versus 70 percent of the phone numbers being accurate is enormous."

Better voter file data from the DNC made a huge difference in the final days of the campaign, said Mark Sullivan, founder of Voter Activation Network (VAN), in Cambridge, Mass., which makes voter file tools that are used by Democrats and other progressive groups.

"There were vast improvements in phone number quality, address quality, and large amounts of consumer data," Sullivan said

While it's unclear whether the DNC data warehouse was a deciding factor in any race, Davis and others cite at least one example of where it came into play: The Florida State Democratic party's efforts to target "Snow Bird" voters from New York and New England states were a direct product of having more detailed voter information with multiple addresses.

"They were able to communicate with Florida voters earlier in the cycle and not wait until people came back in October," said Davis. They could search for people who had addresses in northern states. They never would have been able to do that before," he said.

"[The DNC] put people in areas where we didn't expect big battles to take place, and gave them the tools and [voter] modeling data," said Sullivan. "That played no small part in what happened, especially in some of those states that weren't on the radar."

The DNC database has been a source of controversy in recent months. Following the 2004 debacle, factions within the Democratic Party decided to start their own voter information project, dubbed Data Warehouse and now known as Catalist, under the guidance of former Clinton administration deputy chief of staff Harold Ickes. Ickes was beaten out by Dean to head up the Democratic National Committee.

The Catalist project also bore fruit last week, according to its CTO Vijay Ravindran, who worked for Amazon.com before joining Catalist.

The system, which relies on EMC storage hardware on the back end, was built using open source components such as Linux and MySQL and development frameworks such as Hibernate and Spring. Like the DNC voter file project, Catalist's Data Warehouse is used and is designed to support third-party applications such as VAN, he said.

As for the role Catalist's Data Warehouse played in last week's midterm election, Ravindran said the database of 150 million people was used for a variety of activities including mailings, phone banks, and get-out-the-vote and canvassing efforts in 12 states, affecting around 60 million voters.

Page 3 of 3 « [Previous Page](#)

Groups using the Catalist data included Emily's List, The Sierra Club, Moveon.org, labor organizations, and America Votes, an umbrella group representing 250 different organizations. Like the DNC, Ravindran pointed to races where the data may have swung outcomes: A group called Women's Voices, Women Vote used modeled data on single, unregistered women in Missouri to target a voter drive in support of Claire McCaskill's successful candidacy against Congressman Jim Talent.

The Sierra Club used the organization's data to target 310,000 infrequent voters who support environmental issues in 33 races across the country, including the successful effort to unseat California Congressman Richard Pombo, Ravindran said.

In general, the system performed well, though Catalist's CTO saw room for improvement.

"Having done this before at Amazon, I know that it doesn't mean anything until you go through your first Christmas. We did scale up for [2006], and we learned how to improve the system for 2007 and 2008," Ravindran said.

In published reports, Ickes has said the Data Warehouse project amounted to a vote of "no confidence" in the DNC effort, and that the newly funded Catalist organization would provide Democrats with voter targeting and data mining capabilities on par with the Republican Party's program (Findit/4689).

Despite the history between powerful figures like Ickes and Dean, Ravindran said that he doesn't pay much attention to the "political dynamic" or history between Catalist and the DNC. He anticipates cooperation in the future.

"We've done the first-tier job of getting [Data Warehouse] off the ground. My fond hope is that the DNC and other progressive groups with valuable data share that data," he said.

Bickford also said he doesn't see the two projects competing with each other. Two voter databases -- one associated with the Democratic Party, the other commercial -- may be fine.

"There are like-minded democratic organizations that will always be extremely happy to have organizations like [Catalist's] Data Warehouse, but would not be able to get information directly from the Democratic Party."

What all those involved with the Democrats get-out-the-vote effort agree on is that the Party has closed the technology gap that previously existed between their party and Republicans.

"The quality of the data has substantially improved. It was a huge step up. And with the overlay data, the state parties can do more, but it's not automated. What we want in the future is to automate that and be able to make intelligent decisions about who to contact," Davis said.

Sullivan agrees, and said that the media underreported the sophistication of the Democratic turnout effort in 2006, and overestimated the abilities of the GOP.

While Democrats now have detailed and accurate enough information to look at sub-areas within individual precincts, they lack the ability to target individuals in the way that Republicans can -- a process called "micro targeting," Davis said.

"The idea that Democrats are doing micro targeting is a myth," Davis said. "If you look at the close races, Republicans were able to do things to narrow the margins, even in this cycle. Their performance was impressive."

However, micro targeting is within reach, and Davis said that the data warehousing solution his company helped develop could work as a platform for it -- for example: using profiles of known voters to match up with other individuals who may be sympathetic, but infrequent poll-goers.

As both Democrats and Republicans reach parity on the technology front, the battle will shift to integrating the various data sources in a seamless manner, said Sullivan.

"We'd like to get the data as soon as its refreshed at DNC, then migrate it into our systems. That would reduce the number of human interventions from what we currently have," he said.

While platforms like Netezza are great for extracting data from huge numbers of records, they aren't well suited to the variety of tasks that enterprise databases perform, Gartner's Feinberg said, noting that Democrats may want to harmonize their competing data warehouse projects.

"The DNC does a lot more than identify voters in Florida or DC, or run a program every two years for an election," he said. "If you have a complete enterprise data warehouse, maybe you can take a subset of that data out and put it on the Netezza box for special functions, like doing targeted campaigns. That way you can run queries against it all day long and not hurt the DNC," he said.

"We're really scratching the surface with what we've done with techniques and technologies," said Catalyst's Ravidran. "A decade from now, we're going to look at these first few years where we cleaned up voter lists just so we could do simple queries as the stone age. And they are."