Forbes
APRIL 3, 2012 | WWW.FORBES.COM

Amplifying Big Data Analyst Productivity

By Dan Woods, Contributor

While much of the attention paid to big data focuses on Hadoop capabilities such as Map/Reduce, HDFS, HBase, Hive, and Pig, (in that order) several companies have realized that the practice of analytics is essentially a team sport and are focusing on capturing and sharing knowledge rather than just figuring out how to crunch the largest dataset possible.

Imagine you are a CIO who has been given $1 million to spend to make big data useful. How do you get the most bang for the buck? Would it make sense to build a Hadoop cluster and then be faced with the task of writing Map/Reduce programs to answer every question? No.

I recommend enabling the largest amount of people with the cheapest and easiest tools to perform five tasks:

1. Explore big data sets to understand what can be learned from them.
2. Clean and distill big data sets into smaller collections that can be enhanced with additional contextual information to support analysis.
3. Allow analytics to be applied to the cleaned and enhanced data.
4. Share what has been found out.
5. Create mash-ups and simple applications.

In earlier stories (Does Big Data Always Require Big Money and Designing Scalable and Agile Big Data Platform), I pointed to cheap solutions for requirements 1 and 2, where Hadoop is most relevant. Starting with this story, I'm going to review a diverse set of technology that can create a Hadoop-less big data analysis environment for requirements 2, 3, 4, and 5.

Each of the technology products and platforms I will be covering have a different strategy for meeting my requirements. While no one technology meets all requirements, each of them offers a bridge to a wider group of users, each has a sweet spot. By the end of this series, a CIO with $1 million to spend will have lots of ideas for how to effectively create business value from big data.

Alteryx: A Platform for the Data Artisan

I am starting my journey by looking at a company called Alteryx, which focuses primarily on requirements 2, 3, 4, and 5. I chose Alteryx as a starting point because the company is focusing on one of the largest unacknowledged challenges of making use of Big Data, the shift in responsibility for what happens in the Extraction, Transformation, and Load process, or ETL as it is commonly known.

ETL is the process of taking data from many sources, cleaning it up, combining it, and doing whatever is needed so that the data can be analyzed.

In the era of data warehouses, much of ETL was concerned with preparing star schemas and data cubes that allowed questions anticipated in advance to be answered quickly. In the world of big data technology, this step is no longer necessary. Technologies like Hadoop, Splunk, Pervasive Data Rush, SAP HANA, 1010data, and most of the NoSQL databases can all return answers fast enough so advance preparation of data cubes is not required.

Cleaning and combining is still required, a task which now must be performed by the analyst. Of course, analysts have always cleaned and combined data. But now, they are doing much more of it with larger volumes of data, and this is a new responsibility. In addition, the amount of structured and unstructured data has vastly grown. So the combination process is vastly more challenging. Hadoop doesn't help much with this step, unless you find it easy and fun to write Map/Reduce programs.

"For a good analytic decision to be made today, there are five variants of external analytic content," said George Mathew, President & COO of Alteryx. "You need spatial, you need population, you need demographics, you need segmentation and you need firmographics, which is the business level data. But those five data sources aren't going to come from a data warehouse."

Alteryx is creating an environment in which analysts, which it calls "data ar-

tisans" can use a variety of components to create reusable data objects that can be the foundation for applications. Here's how it works:

- Alteryx offers connectors to systems like Hadoop that crunch and distill data. These connectors can bring in subsets of data into the Alteryx environment. We'll call these subsets data objects.
- Alteryx offers connectors to many other sources of data that may not be in the big data realm but provide the missing dimensions mentioned above that are needed to fully support an analysis.
- Alteryx then offers a visual environment to combine and synthesize data objects into other data objects and load the synthesized information into an in-memory database.
- This in-memory, consolidated view of the information can be packaged as an Analytic Application or used to provide information to other visualization and data discovery tools such as Tableau, QlikView, or TIBCO Spotfire.
- Alteryx also has its own application development environment that allows data objects to be combined into applications.

To Mathew, ease of use and design-time experience is crucial, more important that raw data crunching power or advanced analytics. "Data really needs to be contextualized in the hands of the business owner," said Mathew. "These are people who aren't necessarily coming in with PhDs in data science and statistics, but know the business, know and understand the data that surrounds their business, and are smart about what they need to get done. They need very agile solutions to make it possible."

Mathew is right that making data easy to munge will have a big impact. The other thing that Alteryx offers is a way to capture the intermediate work product that is usually tossed aside after an analysis is done. It won't happen automatically, but it seems that a company should be able to go beyond the idea of a data mart and create a set of reusable data objects that have been properly enhanced.

Alteryx is priming the pump for this sort of reuse by creating templates that have data objects and sample applications targeted for specific industry scenarios such as retail trade area analytics & telecommunications churn analytics.

Figuring out how to get a company sharing ideas about data would clearly be a big win but technology only enables a culture of sharing, which is difficult to build. Alteryx provides one set of mechanisms for sharing. EMC Greenplum said last week that it is going to release its Chorus environment for collaboration around analytics as open source, which may provide a way for multiple vendors to allow sharing on the same platform. IBM or Jive Software I'm sure would insist you can share lots of ideas about data in their general purpose collaboration environments.

Using its ideas of connectors, reusable objects that can be enhanced, and templates for applications, Alteryx provides a coherent system for meeting requirements 2,3,4,5, but the fact is that there are many technologies, some that are already in place, that can also get there if integrated properly.

A CIO with $1 million to spend would do well to create an environment that has the properties of Alteryx and encourages reuse, sharing, and captures patterns for successful apps in templates. The right way to do that will vary widely.

Dan Woods is CTO and editor of CITO Research, a firm focused on advancing the craft of technology leadership. He consults for many of the companies he writes about. F