› Conclusions Paper

Fast and Furious: Big Data Analytics Meets Hadoop

Featuring:

Wayne Thompson, Chief Data Scientist, SAS

Paul Kent, Vice President of Big Data Research and Development, SAS

sas
THE POWER TO KNOW®

Contents

Introduction

Charles Babbage's 19th century mechanical computer could store 1,000 numbers of 40 decimal digits each. That's about 16.7 kilobytes, equivalent to four pages of today's digitized novel.

In the mid 1980s, a state-of-the-art PC had a 20-megabyte hard disk and one megabyte of random access memory (RAM). That would have been enough capacity to store seven digital songs (if iTunes® had existed back then), but not enough RAM to play one of them.

Today's Kindle Fire tablet has 8 gigabytes of internal storage, enough to support 80 applications and store 10 movies or 800 songs or 6,000 books – all in a slim handheld device.

Scale that capacity more than 1,000 times over, and you start to approach the entry-level point for an enterprise Hadoop cluster. Hadoop is an open-source software framework for storing and processing huge data sets on a large cluster of commodity hardware. Hadoop delivers distributed processing power at a remarkably low cost, making it an effective complement to a traditional enterprise data infrastructure.

Non-Internet companies typically start with a cluster of fewer than 100 nodes and expand to 200 or more nodes in the production stages, working with anywhere from 100 gigabytes to petabytes of data. Some advanced adopters have clusters with more than 1,000 nodes.[1] The largest Hadoop clusters in production as of early 2012 contained about 4,000 nodes with about 15 petabytes of storage in each cluster. Yahoo runs thousands of clusters – more than 42,000 Hadoop nodes altogether – storing more than 200 petabytes of data.

You get the idea. Hadoop deals in volume, which makes it ideal for exploring those fast-growing streams of data from transactional systems, sensors, social media and more. You can dump large amounts of data into a Hadoop cluster, then use an analytics tool to explore it and find relationships. Or use a data management tool to transform and aggregate the data and export aggregated values into a data warehouse or another data source.

[1] *Cluster Sizes Reveal Hadoop Maturity Curve,* Timothy Morgan, EnterpriseTech Systems Edition, Nov. 8, 2013.

SAS® and Hadoop

"Think of Hadoop as analogous to the operating system and C drive on your PC, but for your enterprise data hub," said Paul Kent, Vice President for Platform R&D at SAS. "Hadoop will store your big data for you, keep it safe, and provide the CPU cycles to work with that data.

"Two years ago it was all about [the native Hadoop computational approach] MapReduce. MapReduce is still a great technique if your problem fits that style of processing, but it's not the only technique that can be used on a pile of data." SAS brings advanced analytic algorithms to the relationship – made faster with in-memory and in-database processing, and made more accessible with graphical user interfaces and data visualizations.

"Hadoop is a very efficient way to store data in a very parallel way to manage not just big data but complex data," said Wayne Thompson, Manager of Data Science Technologies at SAS. "The SAS® LASR™ Analytic Server environment, collocated in the Hadoop cluster, enables you to run very iterative statistical and machine learning algorithms.

SAS is an example of a new-age Hadoop application that runs directly on the cluster and does not require MapReduce as the proxy for submitting work requests for calculation. Hadoop is used to manage the data and distribute it across the cluster. SAS loads the data into memory and directly performs the calculations, multitasking to do explorations, predictive modeling and machine learning.

Since in-memory processing is very fast, tasks that took hours to run can now be done in minutes or seconds. With that kind of speed, you can do more iterations, create and test more models on the fly, and ultimately develop stronger models.

> "Since in-memory processing is so fast, the time to process advanced analytics on big data is reduced. This frees up more time to actually think differently, experiment with different approaches, fine-tune your champion model, and eventually increase predictive power."
>
> Fern Halper, Research Director for Advanced Analytics, TDWI

This interaction with Hadoop data through SAS can be point-and-click, drag-and-drop easy, if you want it to be. Or you can have programming-based flexibility, if that's your preference. In a presentation at the Strata 2014 big data conference in Santa Clara, CA, Kent and Thompson demonstrated both approaches, using SAS Visual Statistics and SAS In-Memory Statistics for Hadoop.

Exploring Big Data the Visual Way

Thompson demonstrated how easy it is to interactively build models – in this case, to better understand contributors to a charitable cause, so as to understand how to maximize donations.

The interface is intuitive – and fast. Drag and drop a variable onto the desktop and see what effect it has. Grab other variables to see how they might be correlated with donation amount. Drag and drop to do autocharting. Zoom to see details in a pop-up window.

"This is very fast and furious," said Thompson. "Working on the fly, we can drag and drop candidate explanatory variables onto the desktop, perhaps first a histogram, then a correlation matrix. The predictor model shows which variables are strongly or weakly correlated. We can do lots more exploratory analysis, very quickly. As a data scientist, I would do a lot of data wrangling, data dredging and discovery, and this tool lets me do it very interactively."

Thompson goes on to express a correlation matrix, showing donation amount as a function of the other selected variables as a multiple linear regression model. He highlights the row and selects a predictive model from a pop-up menu. The system automatically develops a regression model. He then grabs from the left pane a few more variables that might be of interest, drops them onto the desktop, and once again a regression model is automatically set up and developed.

"Bam – just that fast, I can refit my model," said Thompson. "It's very easy, very interactive, and just about anybody can do it. Very quickly the data is loaded into memory. The data is only read one time, then the computations are done across the grid, and I can work very interactively in an exploratory manner. The software does not drop the data back to disk; the data stays suspended in memory while I interactively hack away at it to fine-tune my model."

An autogenerated lift chart compares predicted to actual donations across various bins by deciles. A residual plot is also autogenerated to evaluate further model fit at the observation level. The software automatically uses a heat map for the residuals to accommodate preserving the distribution pattern for big data. Interactively add interaction or exclude selected observations and refit the model on the fly.

"With in-memory analytic processing, you can build models faster – even build dozens of base models almost in a model factory approach using the group-by capability," said Thompson. "There's no need to write data to disk or perform intermediate data shuffling, and you can instantly see the impact of changes, such as adding new variables or removing outliers."

"It's very easy, very interactive, and just about anybody can do it. Very quickly the data is loaded into memory, it is only read one time, then the computations are done across the grid, and I can work very interactively in an exploratory manner."

Wayne Thompson, Manager of Data Science Technologies, SAS

Exploring Big Data in a SAS® Programming Environment

A visual environment is quick and easy, but it might not be everybody's first choice, said Thompson. "We data scientists don't always like to clickety-click; we like to roll code. For us, there is SAS In-Memory Statistics for Hadoop, which provides a single interactive programming environment to do analytical data preparation, variable transformations, exploratory analysis, modeling, integrated model comparison and scoring - all inside the Hadoop environment. As with SAS Visual Statistics, the process is interactive and visual – adding and dropping terms – but I'm writing code."

Thompson demonstrated with a hypothetical business problem: what constitutes a lemon vehicle and how to avoid buying one at the auto auction. The demo database has more

than 11 million observations, such as odometer reading, price, buyer number and whether or not the vehicle was an online purchase – joined with additional car information from a dimensional table. From this data, Thompson builds classification models such as logistic regression and random forests.

The process starts with exploratory analysis to understand what's there. Data is loaded into memory, explorations are interactive, and responses come back almost instantly. He can see how the data is distributed across each data node on the Hadoop cluster. The head node manages all communication across the data nodes. He computes distinct counts for all variables in the data set to evaluate cardinality almost instantaneously.

"As I'm analyzing the data, rather than writing to disk, temp tables are being created, and I can add new columns to these temp tables on the fly," said Thompson. He creates new variables – vehicle age and average odometer reading, both computed from other variables – then targets the exploration to vehicles of a certain age and use pattern. Click to run it. In seconds, we see that average odometer reading is higher for 'bad' older cars – no surprise there – but it's higher for 'good' newer cars. This finding points the way to further investigation.

"It's very easy to work in this interactive environment," said Thompson. "This is the way a lot of data scientists work, but it's very easy to work with the language and get back detailed information. It's very simple to look at, and you get results within seconds."

Thompson then gets a summary, creates a few new attributes to add to the model, strips out other variables, joins the car information dimension table with the detailed car data table, and displays an analysis-ready table. The results come back in about four seconds. "That's fast and furious," Thompson says. "We can build a lot of models. We can do k-means clustering, density based clustering, generalized models, and we can fit decision trees. We can do a lot of things, all within seconds or minutes."

He then runs a multipass algorithm – a logistic regression in this case – and gets results back in eight seconds. Next he computes assessment statistics, such as lift, to see how well the model is performing for each decile. Finally, Thompson creates a random forest containing 20 decision trees – but it could be thousands of them. The algorithm randomly swaps in candidate variables within and across the 20 trees, and combines the trees into an often stronger classifier. Model fit can be evaluated using bootstrap samples.

> "Data scientists prefer to analyze data in an ad hoc, interactive way to isolate anomalies, patterns and signals in complex big data. It is also important to get back results quickly so models can be fine-tuned. SAS In-Memory Analytics collocated with Hadoop makes this possible."
>
> Wayne Thompson, Manager of Data Science Technologies, SAS

Using Hadoop to Fuel a Recommendation Engine

If you liked those shoes, you'd probably like these shoes too. And if you liked that book/movie/product/service/experience, you would probably find this one interesting as well.

If you have ever shopped online – even if you have never purchased – you have been touched by a recommendation engine. Think Pandora.com, Amazon.com or Groupon.com, all offering up what you might like. Think of personalized ads that eerily mirror websites you have recently visited. With enough data about what you (and lots of people like you) liked and disliked in the past, recommendation engines can predict what you are likely to like in the future.

A recommender system can generate a user profile explicitly (by querying the user) and implicitly (by observing the user's behavior) – then compares this profile to reference characteristics (observations from an entire community of users) to provide relevant recommendations.

SAS provides a number of techniques and algorithms for creating a recommendation system, ranging from basic distance measures to matrix factorization and collaborative filtering – all of which can be done within Hadoop.

"In this demo, we're trying to build a movie recommendation system based on explicit user ratings of the movies," said Thompson. He creates and names a project, and loads multiple tables (user item table, profile information, ratings, etc.) into memory using the SAS procedure PROC recommend. This particular project incorporates a nearest-neighbor model based

on the Pearson correlation (a local measure of the linear correlation between two variables) and a matrix factorization model. The recommendation system automatically handles cold-starting (making inferences about new visitors based on multiuser context) using weighted averages.

"As with everything in machine learning, you jambalaya it," said Thompson. "By jambalaying it, you ensemble it, which inevitably provides a stronger solution. Just as with the model-building demo, I'm loading these things into memory. Then I can use a predict statement to predict new items for users based on nearest neighbors, latent factors or model ensembles. It's easy to do, it's all self-contained, and it's all part of this little bit of code right here."

In about 10 seconds, the recommender system delivers multiple recommendations for each user - in this case, movies the user might enjoy - based on what is known from the user's profile and historical preferences shown by a community of similar users.

Closing Thoughts

SAS and Hadoop are natural complements. Hadoop provides distributed storage and processing power. SAS treats Hadoop as just another data source and complements it with data management, data discovery and advanced analytics. The integration of the two provides important benefits for extracting the most value from your big data assets:

- **Accuracy.** You can apply SAS advanced analytical algorithms and machine-learning techniques to Hadoop data to produce the best business results.
- **Scalability.** Solve enterprise-sized problems with faster time to insights and minimized processing times – seconds rather than hours or days.
- **Productivity.** Multiple users can concurrently and interactively analyze big data in Hadoop using the fast, in-memory analytical programming language.

You can work with Hadoop data from within a familiar and interactive SAS Analytics environment – either with the graphical user interface of SAS Visual Statistics or with streamlined programming via SAS In-Memory Statistics for Hadoop. Either way, your processes will run at fast-and-furious speed, enabling broader and deeper investigation – and quickly freeing up Hadoop resources for the next person's inquiries.

Using SAS® Within a Hadoop Cluster

Reveal insights in your big data.

Redefine how your organization solves complex problems.

Prepare, explore and model multiple scenarios in minutes.

Model tasks interactively and in real time.

Ask what-if questions on all the data.

Instantly add or drop variables into a model and see their influence.

Assess the predictive power of your models.

Understand model fit with model diagnostics on the fly.

Use a scalable recommendation system to improve the customer experience.

About the Presenters

Wayne Thompson, Manager of Data Science Technologies, SAS
Over the course of his 20-year tenure at SAS, Wayne Thompson has been credited with bringing to market analytics technologies such as SAS Text Miner, Credit Scoring for SAS® Enterprise Miner™, SAS Model Manager, SAS Rapid Predictive Modeler, SAS Scoring Accelerator for Teradata, and SAS Analytics Accelerator for Teradata. Thompson received his PhD and MS from the University of Tennessee and was a visiting scientist at the Institute Superieur d'Agriculture in Lille, France.

Paul Kent, Vice President of Big Data Research and Development, SAS
As a leader of big data initiatives at SAS, Paul Kent spends his time with customers, partners and SAS R&D teams to discuss, evangelize and develop software at the confluence of big data and high-performance computing.

Learn More

Download the TDWI Best Practices Report *Integrating Hadoop Into Business Intelligence and Data Warehousing* – Philip Russom, 2Q2013: sas.com/reg/gen/corp/2243106

Learn more about SAS and Hadoop: sas.com/hadoop

Follow us on Twitter: @sasanalytics

Like us on Facebook: SAS Software

To contact your local SAS office, please visit: sas.com/offices

107107_S121362.0614

THE POWER TO KNOW.