Paper PA01

Survey of Big Data Solutions using SAS® Technologies

Jack Shoemaker, d-Wise Technologies, Inc., Morrisville, NC

ABSTRACT

When used in the context of using SAS® technologies, the term "big data" means one of two things: handling large volumes of data and performing statistical analysis in near real time, or distributing a large processing problem over multiple servers. SAS® offers a grid solution for the latter and two solutions for former: in data base and in memory. This paper offers a brief survey of each solution and what sorts of business problems each can solve.

HOW BIG ARE THE DATA

At a time when most readers of this paper have a portable computer in their pocket with approximately 250,000 times the compute power of the AGC in use by Apollo 8 when Jim Lovell read from the Book of Genesis on Christmas Eve in 1968, it is worth considering what “big data” has meant over time. We now employ this immense power to hurl pigs at rickety wood structures and send each other pictures of our cats and latest culinary creations; however, according to a recent article in the *Wall Street Journal*, back in the 1950s, computers were the province of large data-intensive industries like John Hancock which boasted a 600MB Univac platform that could store data for its 2M policy holders. By the 1960s, the big-data champ was the American Airlines Sabre system which boasted an 807MB platform on the largest iron that IBM had to offer. In the 1970s, it was FedEx’s Cosmos system which sported 80GB platform – a 100-fold increase in a decade. Two decades later, in the 1990s, Wal-Mart ran an 180TB platform for its sprawling retail operations. Then things really took off. In the 2000’s, Google measured its footprint in Petabytes – 25 to be exact. And now, the cat-sharing platform known as Facebook boasts a 100PB stash of content which generates a whopping 500TB daily based on the analysis of those pictures and pithy status updates. Just for reference, here’s what 100PB looks like as a decimal number: 100,000,000,000,000,000. In 1968, the Apollo 8 crew ran a computer with 76,000 bytes of storage.

BIG DATA AND ANALYTIC OPPORTUNITIES

Setting aside the sheer awesomeness of the data bases now all around us, the growth of big data has led to some truly novel analytic opportunities. Big data is fueling growth in the fields of advanced and high-performance analytics. This, in turn, is changing the culture and practice of management by empowering data-driven decisions. This is leading to better pricing decisions, more effective risk management, improved fraud detection, and ultimately an increase on the return on investment in these new disruptive technologies. Consider what is happening in the property-casualty industry. The ability to model risks like tornadoes and ice storms in hours instead of weeks or months provides those insurers who employ these techniques a strong competitive advantage. And as any regular viewer of television knows from the ubiquitous Geico ads, the harvesting and analysis of vehicle-use data is changing the landscape of auto insurance. In general, it is high-performance analytics that gives big data its oomph by allowing organizations to glean new insights from previously unemployed data assets.

IN-DATA BASE SOLUTIONS

The in data base solution means handing off computing work to the underlying DBMS and returning the summarized results. This solution is likely the most familiar to existing SAS users who have used SAS/ACCESS engines to access data stored in DB2, Oracle, and MS/SQL tables. Accordingly, you can use SAS LIBNAME engines to access data stored in Teradata, Pivotal (formerly Greenplum), Exadata (Oracle), and Netezza (IBM). These access engines allow you to treat data in these appliances as if they were native SAS® objects. The advantage of using these access engines is that you can leave the data in place or at the very least, move just what is needed for your application.

Although SAS® supports all of the appliances listed in the previous paragraph; Teradata is certainly the most supported. There are a dozen or so procedures that SAS can execute directly on the Teradata appliance. That is, the SAS® system passes the guts of the procedures to the DBMS in the form of complex SQL queries and the DBMS returns the result sets to SAS® for further processing. The advantage here is that you take advantage of the massive parallel processing (MPP) that is at the core of the Teradata architecture. There are in-database versions of the CORR, CANCORR, FACTOR, PRINCOMP, REG, SCORE, and VARCLUS procedures in SAS/STAT® as well as the TIMESERIES procedure in SAS/ETS®. These computationally-intensive statistical procedures benefit immensely from the computing power of the Teradata appliance. Users of SAS® Enterprise Miner™ will find that the DMDB, DMINE, and DMREG procedures have in-database versions available to run on the Teradata platform. In addition, the sampling and binning macros enjoy in-database support.

IN MEMORY SOLUTIONS

The in memory solution means doing the work inside the RAM resident on a multi-core and multi-processor compute server. In particular, SAS® Visual Analytics Explorer (VAE) and High Performance Analytics (HPA) use Hadoop as file storage to load large data sets into a SAS® LASR server running on blade servers. Under this topology, performance is nearly linearly scalable to the memory and CPU resources available. Although this solution currently requires special hardware, in the near future it will run on standard multi-core SMP PCs.

There are half dozen “High Performance” platforms for statistics, optimization, econometrics, and text mining. The statistics platform has a data-mining companion and the econometrics platform has a forecasting companion. That is, to use either of these companion applications you need to have the parent application in place first. As of the writing of this paper, the following compute appliances are supported: Teradata, Pivotal, Exalogic, Apache Hadoop, and Cloudera.

GRID SOLUTIONS

Though not strictly a “big data” solution like the two mentioned above, SAS/tkGrid provides a topology for establishing an enormous and flexible computing platform. The grid solution distributes the computing workload associated with SAS/BI products across different servers using SAS/tkGrid. This solution uses a shared disk store as the central repository that all grids access. For truly huge data, this data store becomes a bottle neck. However, for CPU-bound SAS® jobs, the solution is ideal because it allows load balancing, failover, and is easily scalable - all you do is add another node.

When using SAS Data Integration Studio or SAS Enterprise Miner, the SAS grid environment provides parallel processing of jobs. Metadata and other SAS options are maintained at the grid level as well. This allows Sas to apply these definitions and options to jobs based on the user’s identity. In principle, this simplifies SAS administration by centralizing policies and the associated metadata that drives those policies.

When using SAS Business Intelligence clients, grid-enabled code is stored as a SAS Stored Process. This effectively decouples the computing environment from the actual processing. In turn this leads to a shared environment than can automatically and dynamically allocate resources as needed to meet varying demand loads.

WHICH SOLUTION IS RIGHT?

Determining which of these solutions is right for you depends on your business needs, budget for hardware and software, and the degree to which you need to (or want to) manage the SAS environment. If you already have one of the appliances that support in-database processing (particularly Teradata) and your workload mostly consists of the dozen or so SAS statistical procedures that support in-database process, then that is the route to go. The entry point is familiar and you will be able to take full advantage of the superior processing power of the underlying data base appliance. Some vendors, like Pivotal have made great strides in providing a true SQL interface which opens up all sorts of possibilities for normal pass-through processing.

Keep in mind that when you invoke the in-database solutions, you are still essentially running a batch process. That is, there will be a certain amount of time devoted to generating a plan which is then executed against the database. If you are running one of the supported statistical procedures, this planning stage is certainly worth the effort. Notwithstanding, don’t expect immediate, interactive results and responses.

If you desire more interactive responses in a high-performance environment, the in-memory solutions are for you. The tradeoff is that at the moment, these solutions require specialized and therefore relatively expensive hardware. Also, you will be out there on the bleeding edge of technology. The plan is to support more standard computing platforms in the near future, but until that comes along you will be operating on specialized hardware which will require specialized skills to administer and maintain. SAS can host such an environment for you so you can take advantage of their expertise; however, your organization may not embrace such an in-the-cloud solution.

If you’d like to use commodity hardware to create a managed and shared environment, then the grid solution is the best fit. Think of the grid as a gigantic SAS/CONNECT application that offers dynamic resource-based load balancing. Although this topology makes possible the execution of multiple applications by multiple users, it comes with the price of increased administration. The SAS administration team – and it will be a team, not a single person – can prioritize jobs across the entire enterprise using a rules-based governance model, but maintaining and monitoring those is a full-time activity. A robust comfort level with the SAS metadata repositories is essential for effective grid management. If you find the SAS metadata model cumbersome or unworkable you should think twice about going down the grid road.

The grid solution allows administrators to create a set of enterprise-wide metadata for the grid. This, in turn provides a centralized environment to manage and maintain access policies. You can use a grid node as a hot standby for

failover. In general, the grid provides a high-availability environment using commodity computing resources. If you need a high-availability, fault tolerant and optimally load-balanced SAS environment, then the grid environment is an excellent choice provided you are ready to invest in the increased burden of administration. This flexible infrastructure allows an organization to add computing resources as needed to meet changing and increasing business demands.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jack Shoemaker
d-Wise Technologies, Inc.
Suite 150, 1500 Perimeter Park Drive
Morrisville, NC 27560
(919) 397-9066

jack.shoemaker@d-wise.com
www.d-wise.com