Intel IT Center

BIG DATA ANALYTICS

Data Accelerator: Speeding Your Time to Big Data Value

MetaScale leader shares strategies and tips on capturing value from big data analytics

Scott LaCosse
Director of Technology Operations, MetaScale LLC

MetaScale

Apache Hadoop* software and other elements of the Hadoop ecosystem offer new ways to improve data analytics and capture value, ranging from tighter operations to higher sales. But where do you start? How do you fit Hadoop into your enterprise business intelligence (BI) strategy? And how can you avoid potholes on the road to big data value?

Here are some suggestions based on my experiences implementing distributed analytics at Sears Holdings and working with a broad spectrum of enterprises at MetaScale. MetaScale is a data accelerator that provides technology, solutions, and expertise to help enterprises capture value quickly from big data.

Finding Value with Big Data

- **Put Hadoop's flexibility to work for you.** Hadoop isn't a traditional analytics platform. As a simple yet powerful framework for parallel distributed analytics, Hadoop gives you great flexibility to work with new data types, larger data volumes, and rapidly changing business requirements—all while reducing costs. It doesn't replace your existing analytics environment, but nearly every enterprise can find value in using it.
- **Explore new data sources and time frames.** When is Hadoop the right choice from the analytics toolbox? The answer that gets the most buzz is to use Hadoop when you want to analyze data in real time or analyze unstructured data streams. For example, social media conversations or chat scripts of call center interactions can help your company troubleshoot a product launch that's not going as well as you'd like, or identify and fix a confusing product feature. Real-time analysis of streaming machine data can alert you to events of interest—from potential security threats to optimal patterns of energy use.
- **Target problem BI processes. Hadoop isn't just for unstructured data.** Processes that are outgrowing the traditional data warehouse approaches can be great candidates for distributed analytics. Look at traditional analytics processes that are starting to miss their service-level agreement (SLA) targets—or those where costs are rising as data volumes grow. Sears Holdings migrated key aspects in the price optimization pipeline from a mainframe data warehouse into a Hadoop environment. We found we could process 100 times the number of products in one-fourth the time.

Data as an Asset

- **Embrace the simplicity.** The Hadoop environment removes many constraints that mainframe analytics environments have forced us to live with, particularly around cost, performance, and functionality. Traditional environments imposed major start-up penalties on data collection. Before you could collect data, you had to ask yourself, "What data source do I have? How do I strip it and put it into third normal form (3NF) and put some structure around it?" With Hadoop, you don't need to determine the data scheme in advance. You can ask, "What data source do I have?" and be off and running. You still need to ensure that the data you're capturing is clean and accurate, but you can figure out the next steps as you go.
- **Make data collection part of every application, service, operational checkpoint, and touchpoint.** Data is a crucial asset. The corollary of Hadoop's simplicity, flexibility, and economy is that you can gather and store potentially valuable data while you determine whether or how you want to use it. This means that as you

What We Did

- Created a subsidiary of Sears Holdings to share the expertise we developed as we modernized Sears Holdings' legacy infrastructure and analytics environment. MetaScale offers infrastructure and services to develop your enterprise competency with Hadoop, support your legacy modernization, and more.

What We've Learned

- Data is an asset in its own right. Planning for new services and applications should include a review of data collection options.
- The Hadoop environment provides simple, flexible, and affordable ways to manage new data types, larger data volumes, and real-time analytics requirements while reducing costs.
- Enterprise IT teams can optimize success by working with a big data accelerator, modernizing their infrastructure, and choosing a performance-tuned Hadoop distribution that follows open source guidelines and meets management and security requirements.

design any new service or application, you need to examine what data you can collect and where it can provide value. That data may turn out to be as valuable to the enterprise as the application's or service's targeted use.

Practical Matters

- **Start with a single use case.** We recommend selecting a simple, straightforward use case that will deliver value for the enterprise. Articulate the problem clearly, and then take advantage of this use case to experience the power of Hadoop, demonstrate success, and start building your IT team's skills.
- **Stay focused.** There are so many bright, shiny baubles surrounding big data that it's easy to get distracted. Keep your focus on that initial use case as you implement it and work to deliver value. Then you can branch out.
- **Don't be afraid of retraining.** Apache Pig*, Apache Hive*, and other Hadoop programming tools share many similarities with SQL* and other data programming languages. Mainframe programmers are often pleasantly surprised at their ability to move easily into Hadoop environments.

Technologies and Collaborators

- **Seek expertise.** A data accelerator such as MetaScale can help you identify target projects, develop in-house expertise, deliver value quickly, and create a long-term strategy and road map for data analytics.
- **Modernize your infrastructure.** Hadoop is well suited to scalable, modular infrastructure with high performance, bandwidth, and reliability. We use Dell PowerEdge* R720xd servers with the Intel® Xeon® processor E5 family. We isolate our Hadoop cluster behind a front-end batch access layer to improve security. And we use the Intel® Solid-State Drive (Intel® SSD) Data Center Family at this access layer to speed the ingestion of very large files.
- **Choose a Hadoop distribution that matches your requirements.** A number of MetaScale's customers choose the Intel® Distribution for Apache* Hadoop software (Intel® Distribution) for its performance optimizations and adherence to open source standards. Customers in security-sensitive industries like the built-in encryption support and other security features of the Intel Distribution. Customers who manage their own Hadoop environments appreciate the robust management capabilities of the Intel® Manager for Apache Hadoop software.

1113/LJ/TDA/XX/PDF Please Recycle 329742-001US