Your checklist for getting started with Hadoop

Eight things you need to know before analyzing big data in Hadoop

by Fern Halper, Director of TDWI Research for Advanced Analytics

As more organizations set out to compete on analytics, several technology factors are coming together to form the fabric of an evolving analytics ecosystem for Hadoop.

Of course, embracing new technologies always leads to numerous questions. Today's questions include: "How can I deal with data preparation on Hadoop?" "How does utilizing Hadoop affect visualization and other kinds of analysis?" "What kind of analytical techniques are available to analyze Hadoop data?" "How do I use Hadoop with in-memory processing?"

The following checklist focuses on these questions and provides the information you need to start exploring big data analytics.

Business value can only be

1. **Understand Hadoop.** Hadoop includes two components: a low-cost system for storing data called the Hadoop distributed file system (HDFS) and a processing engine that

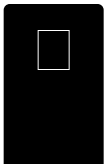
Read More

- This article has been excerpted from a longer TDWI Checklist Report. Read the full report today.
- What is Hadoop? Find out now.
- Get more big data insights.

> created from big data analytics if the model results are integrated into business processes to help improve decision making.

distributes data into manageable sections for consumption or processing, called MapReduce. Hadoop is an inexpensive solution for storing and processing big data, especially semistructured and unstructured data. However, there are limitations to Hadoop, especially when it comes to advanced analytics. As a result, a large ecosystem of advanced tools and connectors are being built up around Hadoop. Keep reading to better understand what to look for in this ecosystem.

2. **Consider in-memory analytics**. In-memory analytics processes data and mathematical computations in RAM rather than on disk, avoiding time-consuming I/O. Advanced analytical techniques such as advanced statistics, data mining, machine learning, text mining and recommendation systems can especially benefit from in-memory processing. Improvements include faster analysis and improved interactivity with data. MapReduce is not suited for iterative analytics. As a result, many vendors offer in-memory processing for Hadoop. In most cases, the in-memory processing capability sits outside of Hadoop. Some vendors lift the data from Hadoop and put it into an in-memory engine for iterative analysis.



3. **Change the data preparation process.** Big data analytics requires sophisticated analytics techniques, which in turn requires efficient data exploration and preparation to determine variables of interest for prediction, missing values, outliers, data transformations and so on. This might require a different mindset from that of someone using a data warehouse for reporting, where the data is predetermined. The mainstays of data preparation and integration, such as data quality or metadata, don't go away.
4. **Explore data for new insights.** You can use it as part of data preparation (as mentioned earlier) and also for insight discovery. For instance, you may want to perform simple visualizations or use descriptive statistics to determine what's in the data or identify variables of interest for more advanced analysis. Look for a vendor who can provide functionality for querying, visualization and descriptive statistics.
5. **Understand advanced analytics.** With big data and in-memory processing

there are no limits on the type of analytics you can perform on your data. To really move beyond simple descriptive analysis, develop a program that includes data mining, text mining and machine learning. The most popular application use cases include pattern detection, classification, prediction, recommendation and optimization.

6. **Don't ignore text data.** Much of the data in a typical Hadoop cluster is text data. This makes sense because HDFS is a file system, so it is used to store semistructured and unstructured (including text) data. A key benefit is to use all the data to your advantage for a more complete picture of what is happening with your customers, operations and more. Some companies write custom code to extract pieces of information from text data. Others use commercial text analytics, including natural language processing and statistical techniques, to extract and structure text data so it can be combined with existing structured data for advanced analytics techniques such as predictive modeling. The information extracted from text often provides substantial lift to these models.
7. **Operationalize analytics.** Business value can only be created from big data analytics if the model results are integrated into business processes to help improve decision making. This is a critical step in any analytical project. The most efficient way to operationalize predictive analytics is to integrate the models directly in the operational data store, known as "in-Hadoop scoring." As new data enters Hadoop, the stored-model scoring files are used by MapReduce functions to run the scoring model and generate timely results.
8. **Evaluate your skill set.** The people dimension can be just as important as technologies selected for extracting value from Hadoop. You need a range of talents for successful big data analytics. The data scientist has recently emerged as a role that combines the different types of skills needed for big data and big data analytics. Data scientists possess the necessary skills to process, analyze, operationalize and communicate complex data. They have the right mix of technical skills, including computer science, modeling, creative thinking and communications. If you cannot (understandably) find all of these skills in one person, try to make sure you have them spread among a few members of your team.

Fern Halper, Ph.D., is director of TDWI Research for advanced analytics, focusing on predictive analytics, social media analysis, text analytics, cloud computing, and other "big data" analytics approaches. She has more than 20 years of experience in data and business analysis, and is co-author of "Dummies" books on cloud computing, hybrid cloud, service-oriented architecture, service management, and big data. You can reach her at fhalper@tdwi.org, or follow her on Twitter: @fhalper.