TDWI RESEARCH

TDWI CHECKLIST REPORT

# Eight Considerations for Utilizing Big Data Analytics with Hadoop

By Fern Halper

Sponsored by:

sas

tdwi.org

tdwi

MARCH 2014

TDWI CHECKLIST REPORT

# EIGHT CONSIDERATIONS FOR UTILIZING BIG DATA ANALYTICS WITH HADOOP

By Fern Halper

TABLE OF CONTENTS



555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295

**T** 425.277.9126
**F** 425.687.2842
**E** info@tdwi.org

**tdwi.org**

## FOREWORD

As companies seek to gain competitive advantage using advanced analytics, a sea change is occurring in terms of the data and the infrastructure that supports it. Several technology factors are coming together to form the fabric of an evolving analytics ecosystem. These include:

- **Big data.** Companies have been dealing with increasing amounts of diverse and often high-velocity data for some time. Some of this big data is new, such as data generated from smartphones or sensors. Much of it is unstructured, including machine-generated data (e.g., satellite images) or human-generated data (e.g., text data, social media data, or website content). Big data is putting a strain on current analytical processes.
- **Hadoop.** As big data continues to get bigger, companies are seeking out new technologies to help them cope. One of these technologies is the Hadoop file system (HDFS) and the ecosystem of tools surrounding it. Hadoop is an inexpensive solution for storing and processing big data, especially semi-structured and unstructured data. It is rapidly becoming an important part of the big data ecosystem.
- **Advanced analytics.** At the same time, there have been advances in analytics algorithms and analytics processing. Visualization has helped companies explore data to discover insights—even with big data. Analytics algorithms such as machine learning and predictive analytics have matured to support the distributed processing needed for big data analytics. Text analytics is helping people derive new meaning from unstructured data.

Data preparation and staging technologies are evolving to support big data. In addition, advances such as in-memory analytics and in-database analytics have accelerated analytics performance, which has helped organizations analyze data more effectively in order to compete.

As enterprises look to embrace big data and Hadoop, they have numerous questions: "How can I deal with data preparation on Hadoop?" "How does utilizing Hadoop impact 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?"

This Checklist Report focuses on these questions and provides information to help you explore big data analytics.

## NUMBER ONE
## UNDERSTANDING HADOOP

Hadoop is an open source project managed by the Apache Software Foundation. At its core, it has two components:

- The Hadoop distributed file system (HDFS), which is a low-cost, high-bandwidth data storage cluster.
- The MapReduce engine, which is a high-performance distributed/parallel processing implementation. It helps to break data into manageable chunks and then makes it available for either consumption or additional processing.

The power of Hadoop is that it utilizes schema on read. With a data warehouse, you often have to know what the tables look like before loading data. With Hadoop, you can pull data from any source or type and then figure out how to organize it. Organizations are beginning to use Hadoop as a dumping ground for all kinds of data because it is inexpensive and doesn't require a schema on write. Such storage is often referred to as a Hadoop "data lake." On the flip side, the Hadoop/MapReduce engine is not optimized for the iterative processing that analytics often requires. It is best suited to batch processing.

To increase the adoption of Hadoop, a whole ecosystem of tools is growing around it. These include:

- YARN (Yet Another Resource Negotiator), a flexible scheduler
- HiveQL, which is not SQL but provides users who know SQL with a tool to get SQL-like access to data
- HBase, a non-relational columnar database that uses HDFS as its persistent store
- Pig, a script-based language for interacting with large data sets

Both Hive and Pig can generate Java that MapReduce can run. Because using Hadoop in its open source native form (that is, via Apache) requires programming skills, commercial distributions and suites are being enhanced to provide tooling and commercial support that can make Hadoop easier to set up. Still, programming is involved.

Although Hadoop doesn't replace the data warehouse, it can complement it, especially for storing disparate data types. Hadoop can be an important part of the analytics ecosystem, but open source or as-is Hadoop and MapReduce may not be the best choice for advanced analytics.

## NUMBER TWO
### CONSIDERING IN-MEMORY ANALYTICS

In-memory analytics processes data and mathematical computations in RAM rather than on disk and avoids time-consuming I/O. This can be a boon for analytics against big data. Theoretically, in-memory processing can be thousands of times faster than data access from disk, which is beneficial for advanced analytics, where iteration is often required to build models. In-memory distributed processing can handle multi-pass-through data and iterative analytic workloads, and some vendors even provide communication among independent units of work to take real advantage of massively parallel processing architecture.

Advanced analytical techniques such as advanced statistics, data mining, machine learning, text mining, and recommendation systems can especially benefit from in-memory processing. These advantages include:

- **Better performance for analysis.** Because in-memory processing is so fast, the time required 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. For example, a training set for a predictive model that might have taken hours to complete one iteration now takes minutes utilizing in-memory techniques. This means that more and better models can be built, which helps to derive previously unknown insights from big data. This in turn often results in competitive advantage.
- **Better interactivity.** Once data is in memory, it can be accessed quickly and interacted with more effectively. For example, if someone builds a model that is able to run faster, they can share intermediate results with others and interact with the model more quickly. The model can almost be changed on the fly, if needed, as others look at it and make suggestions. This supports the iterative process of building an analytical model with maximum accuracy and business benefit.

Various vendors offer in-memory processing with Hadoop. In most cases, the in-memory capability sits outside of Hadoop. Some vendors lift the data from Hadoop and put it into an in-memory engine for iterative analysis. Some vendors leverage MapReduce to do the processing; others don't. MapReduce is best suited for single-pass analytics (descriptive, non-instant results), although this may change in the future.

## NUMBER THREE
### CHANGING THE DATA PREPARATION PROCESS

There is a debate raging in the market about data preparation, including ETL, for big data analytics. On the one hand, some people argue that the beauty of big data analysis is the ability to manage and explore data in its unconstrained, native form. Data can be extracted from source systems and put into Hadoop, where it can be transformed and analyzed (this is the ELT argument: extract, load, then transform). In fact, one Hadoop use case is to preprocess data in Hadoop and then bring relevant data into the warehouse or to an in-memory server or other platform for analysis. Others argue that unstandardized, inconsistent data leads to poor decisions and that data quality is fundamental. The answer for your organization will depend on your specific business problem.

The reality is that big data analysis requires sophisticated analytics techniques, which in turn require exploration and preparation to determine variables of interest for prediction, missing values, outliers, and so on. This might require a different mind-set from that of someone using a data warehouse for reporting, where the data is predetermined.

Of course, the mainstays of data preparation and integration, such as data quality or metadata, don't go away. High-quality data is necessary for companies to make sound business decisions when dealing with big data as well as traditional data. Cleansing data without moving it and facilitating business user engagement for effective governance and context are even more essential with big data initiatives. Metadata comes into play when ensuring that the data source lineage used in model preparation is available in operational systems. HCatalog (now merged with HiveQL) provides some facility for this in Hadoop, but HiveQL is slow (see next section). Finally, leveraging logical data warehouses to create virtual views of data from relational and big data sources without data movement accelerates time to insight and reduces IT workloads.

## NUMBER FOUR

EXAMINING BIG DATA EXPLORATION AND INSIGHT DISCOVERY

Data exploration is important for big data. 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. A business analyst or modeler might want to build reports or models as a next step. Some useful techniques include:

- **Query it.** Querying the data is often a prerequisite for insight discovery on big data. HiveQL is part of the Hadoop ecosystem. It supports many of the SQL primitives, such as select, join, aggregate, and union. It can work with MapReduce to distribute the running of a query. However, HiveQL does not perform instantly. In fact, it can take minutes or even hours to get query responses. This does not lend itself to "speed of thought" exploration and discovery. The interactive query engine Cloudera Impala may speed up query times, although it is only now entering the market.
- **Visualize it.** Often the best way to explore data is to visualize it in an interactive, intuitive, and fast manner (see Figure 1). Visualization is an iterative process, and with big data, you might have to explore hundreds of thousands or even millions of observations with thousands of attributes (variables and features), or huge, unstructured data sets. Insightful analytic visuals such as box plots, scatterplots, word clouds, concept network diagrams, and heat maps provide meaningful views and provide a path for further analysis.
- **Perform descriptive statistics.** Another useful way to summarize and explore data is to apply descriptive statistics and present the results in simple-to-understand graphs to gain a quick sense of a particular measure. These include mean, median, range, summarizations, clustering, and associations, among others.

**Figure 1**

Click to enlarge

**Figure 2**

Click to enlarge

## NUMBER FIVE

GRASPING ADVANCES IN ANALYTICS

Advanced analytics provides algorithms for complex analysis of either structured or unstructured data. It includes sophisticated statistical techniques, machine learning, text analytics, and other advanced data mining techniques (see Figure 2). The most popular application use cases include pattern detection, classification, prediction, optimization, recommendation, and forecasting.

Many advanced analytics algorithms have been around for decades, although big data has helped to increase awareness and has prompted reengineering to take advantage of massive distributed in-memory computing environments. Primary techniques include:

- **Data mining and machine learning.** Data mining utilizes algorithms for detecting patterns and hidden relationships in often vast amounts of data. It draws on well-established techniques such as regression and principal component analysis. Machine learning is another related interdisciplinary field used on large, diverse sets of data for making predictions. Machine learning means a computer automatically learns insights from past observations via either supervised or unsupervised training.
  - **Supervised approaches.** Here, an algorithm is given a set of inputs and makes predictions for a set of corresponding outcomes or target variables. The target attributes can be classes or numeric values. For instance, in a churn classification model, the target variable might be class "Stay" or class "Leave." The algorithm uses historical data to extract patterns of attributes that relate to outcomes labeled "Stay." This is the learning or training phase. The patterns are then used to predict the outcome of labels on future data; this is the application or scoring phase. Popular machine learning techniques include decision trees, neural networks, and support vector machines.
  - **Unsupervised approaches.** In unsupervised learning, an algorithm is given a set of input variables but no outcome variables. The algorithm searches automatically for distinct patterns in the input data and groups the data into mutually exclusive segments based on similarity. Dimensionality reduction is another example; the goal is to reduce the number of input variables.
- **Optimization.** Optimization uses mathematical programming techniques to find the best solution given a mix of factors and a set of constraints. It is used in revenue management, marketing campaigns, simulations, manufacturing and supply chains, and other areas.

## NUMBER SIX

APPRECIATING HOW TEXT DATA FITS INTO THE ANALYTICS MIX

Text data is found in e-mail messages, call center notes, tweets, blogs, and a myriad of other sources. This "unstructured data" often contains the *why* behind the *what* in terms of particular actions. For example, "Why is there an increase in the number of returns?" Increasingly, companies are using text data for analysis—in fact, this is a big piece of the big data equation.

Much of the data in a typical Hadoop cluster is text data. This makes sense because HDFS is a file system and, as such, is used to store semi-structured 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. This can provide a clear competitive advantage.

Some companies write custom code to extract pieces of information from text data. Others use commercial text analytics methods to transform text data into usable data for analysis. These techniques often combine natural language processing and statistical techniques to extract entities (such as person, place, or thing), concepts (sets of words that convey an idea), themes (groups of co-occurring concepts), and sentiments from text data and use it for analysis. For instance, most text analytics engines will parse the sentences that make up a document to extract important dimensions, entities, concepts, and so on. Some use statistical techniques such as support vector machines to further reduce the dimensions of the unstructured data. Problem-specific taxonomies help to sharpen the automated extraction of important data pieces from unstructured data. Once the data is extracted and structured, 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.

Some vendors are providing ways to utilize text analytics as part of a scripting environment to run against text data stored in Hadoop. Although this is in early development, it will ultimately allow users to structure unstructured text and use it in different kinds of analysis.

## NUMBER SEVEN

OPERATIONALIZING MODEL DEPLOYMENT

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. You can build the world's best model, but it is useless if it is not deployed or operationalized against new data and monitored regularly for usefulness. For instance, a fraud model would be operationalized in the production transaction authorization process to automatically identify potentially fraudulent claims for special action, such as referral to an investigation unit. Machine-generated data can be automatically monitored and scored to predict when a part in a remote device might fail. Customer buying behavior can be analyzed to automatically create individualized recommendations.

The most efficient way to operationalize predictive analytics is to integrate the models directly in the operational data store—so-called "in-database scoring." The major benefit is that processing occurs directly on the data store, eliminating data movement, which is especially time-consuming and resource-intensive with big data. Putting analytical models into production with manual processes requires skilled human resources and increases the probability of errors. Automating the model deployment and execution step will help streamline the migration of big data analytics from research to production.

In-database scoring has been deployed on all major data platforms. Although Hadoop is not a database, vendors are working to put in-database scoring into Hadoop. As new data enters Hadoop, the stored model scoring files are used by MapReduce functions to run the scoring model and generate timely results.

## NUMBER EIGHT
EVALUATING THE SKILL SET

Much of the discussion around investments for big data has focused on selecting the right set of technologies for extracting value from Hadoop. However, big data and big data analytics are not just about technology. The people dimension—staff with the right skills—is equally important to derive business benefits. A range of talents is needed for successful big data analytics. These include roles traditionally associated with business analysts, statisticians, data miners, business intelligence practitioners, data management professionals, and computer scientists.

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 and the right mind-set and exhibit these characteristics:

- **Computer science/data hacker/developer.** The data scientist needs to have solid technical skills and a foundation in computer science in order to understand the technology infrastructure for big data.
- **Analytical modeling.** The data scientist needs to understand data and have a solid basis in analytics and modeling. Critical thinking is key, as is a disciplined (yet flexible) approach to problem solving.
- **Creative thinker who showcases curiosity.** The data scientist needs to appreciate data and be good at asking questions about it. Some organizations look for people who get a "gleam in their eye" when discussing data.
- **Communicator and trusted advisor about business results.** The data scientist needs to be able to weave a story around data and the results of an analysis so that a business person can understand it.

Of course, it can be hard to find one person with all of these skills, which is why some organizations have assembled special teams to fill this role. Other organizations have created centers of excellence where data scientists can work with and train others in big data analysis. Universities are also offering data science courses to fill in skills gaps and partnering with organizations to recruit talented individuals.

## ABOUT OUR SPONSOR

**sas.com/hadoop**

**Get complete data-to-decision support for Hadoop**

Combining the power of SAS® Analytics with distributed processing technologies like Hadoop helps transform your big data into big knowledge so you can make better decisions. From data preparation and exploration to model development and deployment, we've got you covered.

**Manage Data**

- Easily access data stored in Hadoop and execute Pig, Hive, and MapReduce data transformations.
- Execute data quality jobs inside Hadoop to generate clean data.

**Explore and Visualize**

- Quickly visualize your data stored in Hadoop, discover new patterns, and publish reports.

**Analyze and Model**

- Apply domain-specific, high-performance analytics to data stored in Hadoop.
- Uncover patterns and trends in Hadoop data with an interactive and visual environment for analytics.

**Deploy and Execute**

- Automatically deploy analytic models to score data stored inside Hadoop. Reduce data movement and get results faster.

For data scientists, SAS also offers a highly interactive programming solution for the entire data-to-decision process with Hadoop. In addition to procedures for preparing and exploring data, it includes predictive modeling and machine-learning techniques. Multiple users can concurrently analyze large amounts of data stored in Hadoop using in-memory processing.

Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.

To learn more, visit sas.com/hadoop.

## ABOUT THE AUTHOR

**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 has published numerous articles on data mining and information technology. Halper is co-author of "Dummies" books on cloud computing, hybrid cloud, service-oriented architecture, service management, and big data. She has been a partner at industry analyst firm Hurwitz & Associates and a lead analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her at fhalper@tdwi.org, or follow her on Twitter: @fhalper.

## ABOUT TDWI RESEARCH

TDWI Research provides research and advice for business intelligence and data warehousing professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence and data warehousing solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

## ABOUT THE TDWI CHECKLIST REPORT SERIES

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.