Apache Hadoop
ECAR-WG Technology Spotlight

This bulletin is one of a series of papers from ECAR working groups designed to help institutional leaders learn about and understand the implications of emerging technologies in higher education. These technologies have been identified as the “Top 10 Confusing Technologies” in the ECAR report *Higher Education’s Top 10 Strategic Technologies for 2015*. Other papers and related resources are available at the research hub for *Higher Education’s Top 10 Strategic Technologies for 2015*.

Apache Hadoop is an open-source software project for reliable, scalable, distributed computing. It is a “framework that allows for the distributed processing of large data sets across clusters of computers” and is “designed to scale up from single servers to thousands of machines, each offering local computation and storage.”¹ Hadoop is really a couple of technologies: the file system (HDFS, which stores the data—spreading it across many servers), and MapReduce—the technology that does the parallel processing.² Essentially, Hadoop enables big-data storage and processing by dividing the work across computer clusters. Large data sets can be broken into pieces for cheaper storage on readily available commodity hardware, allowing also for faster parallel processing.

Typically, big-data NoSQL (“Not only SQL”) distributed databases,³ such as those that support Hadoop, are used for unstructured data (e.g., web clicks, measurement information, cash-register data, etc.). NoSQL databases approach data storage and retrieval in a nonrelational way, allowing for high performance and highly scalable data management that can handle massive data.⁴ As a result, these databases aren’t very good at combining data from multiple tables. As with any data application, you really need to understand the data (context, cleanliness, source, etc.); to be cognizant of the data, format, and quality; and to think through the way the data will be created and stored so that they can be tied to the queries you’ll be programming.

The third piece needed to put Hadoop to use is creating the queries that MapReduce uses; because this isn’t rows and columns as in a relational database, it uses key/value pairs and is programmed using various languages (e.g., Java, Python, Pig, etc.). Right now, not many people can do this kind of programming, although a lot of people do know SQL programming, and so vendors are starting to develop tools that will translate SQL into these other languages.

**Importance to Higher Education**

The role of data has emerged at the forefront of higher education in the past several years. Big data—both research and administrative data—is no stranger to our campuses, and concerns about data management, storage, and use (particularly in analytics) are top of mind to many, particularly in IT. Hadoop provides one solution in this space, allowing institutions to store and process data locally—
structured or not—regardless of format. By giving colleges and universities an affordable, scalable, flexible, and reliable place to hold their data, Hadoop can then go the next step and begin finding connections and identifying relationships to help inform business decisions. Hadoop allows organizations to manage massive amounts of data at a faster rate of analysis than can be done using SQL.\(^5\) Not only is this capacity a boon for administrative and learning analytics, but it also presents real opportunities for research-data computations and visualizations that make use of big data (e.g., scientific modeling).

Looking at the larger context of NoSQL databases, there is even greater reason to expect that higher education will be seeing these kinds of databases across the institution. Just some examples of ways that NoSQL databases can be used include:

- To improve search and retrieval of library materials and for the university website
- To harvest and store web and social network data
- For predictive analytics
- For use with real-time web apps\(^6\)

### Current Landscape

Currently, Hadoop is primarily used in the commercial sector, with expectations that it will be widely adopted there in the next five years.\(^7\) Numerous vendors (e.g., Cloudera, Hortonworks, MapR, and others) support Hadoop implementations in cases where technical and analytical expertise and support may be limited.\(^8\) In addition, most data warehouse and management system vendors are working to incorporate or otherwise support Hadoop.\(^9\) In March 2014, it was announced that several Hadoop start-ups received a collective $1 billion from investors;\(^10\) in June 2014, MapR announced that it received $110 million in funding;\(^11\) and in December 2014, even more funding announcements were made, along with the Hortonworks IPO,\(^12\) setting expectations that “2015 could be the year that Hadoop finally goes mainstream.”\(^13\) Forrester has published a report predicting that Hadoop will become a “cornerstone of business technology agenda” in 2015, not only for its capabilities surrounding data analysis but also as a future application platform, with the potential of enhancing enterprise applications.\(^14\)

However, this is by no means a stable space, and there are challenges to adoption in higher education. Right now, most Hadoop implementations seem to be focused on storage and ETL (extract, transform, load) rather than the analysis potential,\(^15\) prompting one executive to refer to Hadoop as his company’s “unsupervised landfill.”\(^16\) In addition, several other technologies are emerging in this space, including Apache Spark\(^17\) out of UC Berkeley,\(^18\) though this may in fact “drive Hadoop adoption even faster.”\(^19\)

Another shortfall today is that these kinds of tools don’t typically have the control and consistency of relational databases (e.g., lock out, recovery, and other features that relational databases have built over time). As a result, the use of these tools does need to be limited to analysis rather than, for instance, transactional or financial information.

Finally, as noted earlier, many of the specific skills needed for Hadoop are currently difficult to find, particularly when it comes to coding queries. For higher education in particular, this gap is a major impediment to being able to take advantage of Hadoop. In the 2015 Top 10 Technologies Survey, when asked about Hadoop, only 2% of respondents noted that they have it in place, with another 8% planning or implementing it now. An additional 16% are tracking it for potential future implementation (see figure 1).
What we will probably see in the foreseeable future is that Hadoop and other tools that rely on NoSQL databases won’t be supporting applications but will be used to enhance applications. Today higher education institutions frequently employ data warehouses and business intelligence; the application of tools like Hadoop is likely to be seen in adding to and integrating with existing intelligence (i.e., to first look at existing big data and then bring them into data warehouses).

When You Should Expect It

With the commercial sector making predictions for wide-scale adoption of Hadoop in the very near future, we can expect to see it filter into higher education soon thereafter, particularly for early adopters who place a high priority on data retention, management, and analysis for big data. This is particularly true for the larger realm of NoSQL databases generally.

<table>
<thead>
<tr>
<th>When</th>
<th>Early Adopters</th>
<th>Mainstream</th>
</tr>
</thead>
<tbody>
<tr>
<td>Now</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–2 years</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>2–3 years</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>3–5 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than 5 years</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Reviewer

ECAR would like to thank Michael Fary, Enterprise Data Architect, University of Chicago, for his helpful and significant feedback to this paper.

Notes

1. See What Is Apache Hadoop?

2. MapReduce is the algorithm that Google first used to process its data. In June 2014, however, Google announced that it was no longer using MapReduce, finding that it had difficulty with data sets of a few petabytes or larger.


4. Some well-known NoSQL databases include MongoDB, Cassandra, CouchDB, and HBase (which is integrated with the Hadoop ecosystem and developed by the same community). Each has advantages and disadvantages and would need to be considered based on need. For more information, see Matt Asay, “MongoDB, Cassandra, and HBase—The Three NoSQL Databases to Watch,” InfoWorld, November 19, 2014. Essentially, Hadoop is typically for offline batch processing, while HBase, for instance, runs on top of Hadoop to allow for real-time read/write access to your data: “Difference between HBase and Hadoop,” StackExchange.

5. See “Hadoop Positioned to Displace Many SQL Database Implementations,” SwishData.


8. For a comparison of nine vendors in this space, see Mike Gualtieri and Noel Yuhanna, “The Forrester Wave: Big Data Hadoop Solutions, Q1 2014,” Forrester Research, February 27, 2014.


15. Matt Asay, “Hadoop Adoption Accelerates, But Not for Data Analytics,” ReadWrite, May 10, 2103.


17. Indeed, in November 2014 it was reported that Spark passed Hadoop in popularity, based on search terms as analyzed by Google Trends. See Alex Woodie, “Spark Just Passed Hadoop in Popularity on the Web—Here’s Why,” Datanami, November 21, 2014.

18. Storm and GraphLab are two other alternatives that are discussed in Vijay Srinivas Agneeswaran, Big Data Analytics Beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives (Pearson 2014).

19. Woodie, “Hadoop Hits the Big Time.”