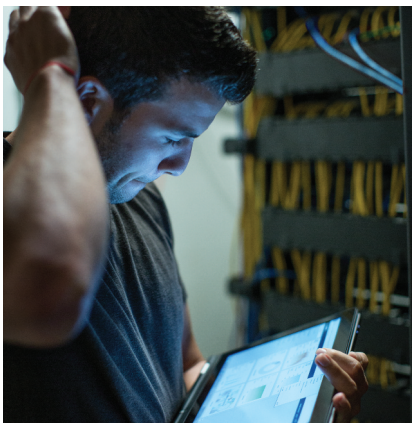
CASE STUDY
Cloudera's Distribution Including Apache Hadoop* (CDH*)
Intel® Xeon® Processor E5-2630 v2 Product Family
Big Data Analytics
Education

Data in the desert

Ben-Gurion University of the Negev opens a Big Data Lab running Cloudera's Distribution Including Apache Hadoop* (CDH*) on the Intel® Xeon® processor



אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

"Cloudera's Distribution Including Apache Hadoop* (CDH*) is optimized to run on Intel® Xeon® processors. This superior performance is magnified with the addition of Apache Spark*. By removing I/O bottlenecks with distributed RAM, Apache Spark secures even better utilization of the Intel® processors. The performance gains realized through these three elements working together are tremendous."

Lior Rokach,
Professor, Department of Information Systems Engineering, Ben-Gurion University

Intel worked with Ben-Gurion University of the Negev to set up a Big Data Analytics Lab, enabling information systems engineering students to better understand and develop complex machine-learning algorithms. The Lab is one of the first in the world running Cloudera's Distribution Including Apache Hadoop* 5.2 (CDH* 5.2) together with Apache Spark* 1.1, on servers powered by the Intel® Xeon® processor E5-2630 v2 product family. By removing I/O bottlenecks with distributed RAM, Apache Spark offers much better utilization of the Intel® processors. This project showcases the performance gains enabled through these three elements working together.

Challenges

- **Peak performance.** The Department of Information Systems Engineering at Ben-Gurion University wanted to roll out a more powerful computing cluster to enable its students to mine massive datasets and work on research projects akin to those faced in industry

Solutions

- **Big Data Analytics Lab.** The university implemented a powerful new computing cluster running CDH 5.2 together with Apache Spark 1.1 on seven HP ProLiant* DL360p Gen 8 servers powered by the Intel Xeon processor E5-2630 v2 product family and running a Red Hat CentOS*

Impact

- **Keeping pace.** The Big Data Analytics Lab allows students to mine even larger datasets than before and develop more complex and distributed algorithms, enabling them to continue to work with industry in the resolution of real-world Web mining and cyber security problems
- **New horizons.** It has also made it possible for the university to run a masters course in mining massive datasets focused on implementing distributed machine-learning algorithms. Together with strong, continued ties with industry, this properly prepares students for careers in the real world and helps the university to attract the best engineering students.

Ben-Gurion University

Ben-Gurion University was established in 1969 as the University of the Negev with the aim of promoting the development of the Negev desert that comprises more than 60 percent of Israel. The University was inspired by the vision of Israel's founder and first prime minister, David Ben-Gurion, who believed that the future of the country lay in this region. After Ben-Gurion's death in 1973, the University was renamed Ben-Gurion University of the Negev.

Today, Ben-Gurion University is a major center for teaching and research with close to 20,000 students. The Department of Information Systems Engineering, which sits within the Faculty of Engineering, prepares students for careers in the analysis, design, development, use and management of information systems.

The department has an excellent research program and strong ties with industry, which enables its fourth-year undergraduate and postgraduate students to work on real-world information system problems. The department also excels in filing patent applications.

Big data analytics

Many of the students' research projects require analyzing very large datasets, especially those focused on Web mining and cyber security. The emphasis is on developing novel, sound and theoretically-motivated algorithms for accomplishing large-scale tasks in data mining, such as high-dimensional clustering, classification algorithms and anomaly direction.

Lior Rokach, professor in the Department of Information Systems Engineering, explains: "With the proliferation of user-generated content on websites, subjective information in the form of reviews, blogs, and bulletin boards is more widely available and accessible. Such information includes critiques of pretty much everything from products to politicians, movies and medications. This constitutes fertile ground for sentiment analysis to support decision-making. A manual assessment of this information is often impractical. However, distributed computing now enables us to mine and analyze millions of texts to provide new knowledge and insight into the world."

Enabling research students to mine even larger datasets

In the past, the research projects students were able to undertake were limited by the department's computing resources. Generally, students had access to just one server for their research, which meant they were mostly only able to work on developing non-scalable algorithms. If its research programs were to keep pace with developments in the real world, the department realized it needed to roll out a more powerful computing cluster.

Computing clusters

Intel has an experienced Advanced Analytics Department with extensive knowledge of machine learning and big data analytics platforms. Intel Advanced Analytics delivered a presentation to the university explaining the benefits of CDH in mining massive datasets. The Apache Hadoop software library is a framework that allows for the distributed processing of large datasets across clusters of computers using simple programming models.

The university immediately recognized the benefits of CDH for better understanding and developing machine-learning algorithms. Machine learning is a branch of artificial intelligence concerning the construction and study of systems that can learn from data.

The university and Intel then worked together to roll out a seven-node computing cluster. Initially, just the Department of Information Systems Engineering was involved in the project. But when the benefits of the cluster became apparent, other departments stepped up to donate budget to extend the cluster into a shared resource. By pooling resources, the university was able to roll out a much more powerful computing cluster.

Intel assisted the university with the technical rollout, including installation and configuration, and advised students on the best steps and tools to meet their research objectives.

The Big Data Analytics Lab

The Big Data Analytics Lab is one of the first in the world running CDH 5.2 together with Apache Spark 1.1 on seven HP ProLiant DL360p Gen 8 servers powered by the Intel Xeon processor E5-2630 v2 product family and running a Red Hat CentOS. These machines also run the YARN*, HDFS*, HBase* and Hive* daemons. With a total capacity of over 220TB of disk space, over 448GB RAM and 84 cores, it is considered one of the largest Hadoop clusters operating in an Israeli academic institution.

Professor Rokach said: "Cloudera's Distribution Including Apache Hadoop is optimized to run on Intel Xeon processors. This superior performance is magnified with the addition of Apache Spark. By removing I/O bottlenecks with distributed RAM, Apache Spark secures even better utilization of the Intel processors. The performance gains realized through these three elements working together are tremendous. Ongoing management of the cluster is also relatively simple[1]."

"Once you understand the basics, it is very easy to adapt your previous code to a new setting using any of the tools that support this model over big data, thus gaining distributed computing capabilities."

Professor Rokach concludes: "Previously, we dedicated servers to each different service – for example, a server for SQL* processing, a server for Git* and so on. Now, the services are ubiquitous. Now we can analyze more reviews. We have more co-occurrences of concepts. Since we use the knowledge we obtain about one concept to produce knowledge about other concepts, the more co-occurrences we analyze, the more accurate the inference process. Evaluations are made directly on the correct meaning of newly discovered concepts. In short, the more data we can process, the more accurate the meaning of our new concepts. Also, we can now easily share this data with other universities."

Working with industry

The Big Data Analytics Lab is enabling research students to mine even larger datasets. This means they can better understand and develop more complex machine-learning algorithms, enabling them to continue to work with industry to resolve real-world Web mining and cyber security problems.

Lessons Learned

Cloudera's Distribution Including Apache Hadoop* (CDH*) is the only distribution built from the silicon up to enable the widest range of data analysis on Apache Hadoop. It is the first with hardware-enhanced performance and security capabilities and is optimized to run on Intel Xeon processors. Adding Apache Spark, which removes I/O bottlenecks with distributed RAM offers even better processor utilization. Looking to the future, Intel is committed to further developing a platform on which the entire ecosystem can build next-generation analytics solutions.

For example, information systems engineering students are working with several ISPs to develop a machine-learning algorithm to improve the page ranking system so it is based on actual traffic to each site rather than links within those sites. They are harnessing the Hadoop cluster to mine over 100TB of data delivered to them by the ISPs.

Another project is creating a recommendation system for hotels by analyzing thousands of hotel reviews collected from the Internet. Using the Hadoop cluster, they are able to extract specific features from unstructured texts – for example, does the hotel have a restaurant and, if so, is it any good? They are able to aggregate this data and use it to answer users' search queries.

The Big Data Analytics Lab has made it possible for the university to run a masters course in mining massive datasets, which counts towards the MSc degree with specialization in data mining and business intelligence. Together with strong, continued ties with industry, this properly prepares Ben-Gurion students for careers in the real world. It also helps the university to attract the best engineering students, since it is able to offer interesting research projects that other universities cannot.

Find the solution that's right for your organization. View **success stories from your peers**, learn more about **server products for business** and check out the **IT Center**, Intel's resource for the IT Industry.

0315/JNW/RLC/XX/PDF 332139-001EN