CASE STUDY
Intel® Distribution for Apache Hadoop* Software
Intel® Xeon® Processor E5-2630 v2 Product Family
Big Data Analytics
Education

Data in the desert

Ben-Gurion University of the Negev opens Big Data Lab running Intel® Distribution for Apache Hadoop* software on the Intel® Xeon® processor

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

"Intel® Distribution for Apache Hadoop* software 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 Intel® Distribution for Apache Hadoop* software together with Apache Spark*, 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.** Implemented a powerful new computing cluster running Intel Distribution for Apache Hadoop software together with Apache Spark 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 which focus 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. In 2011, more than 20 percent of patents filed by Ben-Gurion University were from students in the Information Systems Engineering Department.

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.

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.

Enabling research students to mine even larger datasets

Computing clusters

Intel has an experienced Advanced Analytics Department with extensive knowledge of machine learning and big data analytics platforms such as Apache Hadoop and Apache Spark. Intel Advanced Analytics delivered a presentation to the University explaining the benefits of Intel Distribution for Apache Hadoop software 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 Intel Distribution for Apache Hadoop software 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.

The Big Data Analytics Lab

The Big Data Analytics Lab is one of the first in the world running Intel Distribution for Apache Hadoop software 3.0.2, which is based on YARN* together with Apache Spark 0.81, 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. 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.

Lior Rokach, professor in the Department of Information Systems Engineering, said: "Intel Distribution for Apache Hadoop software 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."

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.

For example, information systems engineering students are working with several Internet service providers (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.

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.

Ben-Gurion University will announce the opening of the Big Data Analytics Lab at the Second Annual Data Mining for Business Intelligence Conference, of which Intel is a sponsor.

Lessons Learned

Intel® Distribution for Apache Hadoop* software is the only distribution built from silicon up to enable with the widest range of data analysis on Apache Hadoop. It is the first with hardware-enhanced performance and security capabilities. Optimized to run on Intel® Xeon® processors, it offers great performance. 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.

Find the solution that's right for your organization. View success stories from your peers (**www.intel.co.uk/Itcasestudies**), learn more about server products for business (**www.intel.com/xeon**) and check out the IT Center, Intel's resource for the IT Industry (**www.intel.co.uk/itcenter**).

This document and the information given are for the convenience of Intel's customer base and are provided "AS IS" WITH NO WARRANTIES WHATSOEVER, EXPRESS OR IMPLIED, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NONINFRINGEMENT OF INTELLECTUAL PROPERTY RIGHTS. Receipt or possession of this document does not grant any license to any of the intellectual property described, displayed, or contained herein. Intel® products are not intended for use in medical, lifesaving, life-sustaining, critical control, or safety systems, or in nuclear facility applications.

Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance.

Copyright © 2014, Intel Corporation. All rights reserved. Intel, the Intel logo, Intel Distribution for Apache Hadoop software, Intel Xeon and Xeon inside are trademarks of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others.

0314/JNW/RLC/XX/PDF 330300-001EN