PROOF OF CONCEPT:
INTEL® DISTRIBUTION FOR APACHE HADOOP* SOFTWARE

Pecan Street, Inc.: Smart Grid, Smart Meter, Big Data

Intel® Distribution for Apache Hadoop* software with the Intel® Xeon® processor E5 family shows speed and scale for energy analytics and the Internet of Things

Bert Haskell
Chief Technology Officer,
Pecan Street, Inc.

Utilities and energy companies are investing billions of dollars in upgrading and digitizing their electrical grids and metering systems. In a textbook example of the Internet of Things, these smart grids and smart meters use embedded, intelligent sensors to report massive amounts of real-time data about their operations and environments.

Companies that can process and analyze this data in near-real time—and handle it with appropriate privacy protections—are positioned to optimize operations, minimize service disruptions, and reduce costs. These companies can use data-driven insights to create innovative products and services, encourage appropriate energy usage, and increase customer satisfaction.

With near real-time insights into consumer and grid data, companies can incorporate smarter decision-support capabilities into demand-response solutions that align electricity consumption with supply. Companies can also use their analytic insights to show leadership on popular initiatives such as the industry-led Green Button* effort to provide consumers with energy-consumption information and help them turn it into wiser energy use.

To make it all happen, the hardware and software infrastructure—including the database—must be able to handle the volume, velocity, and variety of the data, and help maintain its security. In many cases, the explosive growth and influx of data far exceed the capabilities of traditional relational databases and are taxing the ability of even massively parallel databases to handle at a reasonable cost.

Experts at Pecan Street, Inc. and The Texas Advanced Computing Center (TACC) say Intel® Distribution for Apache Hadoop* software (Intel® Distribution) offers a viable approach to meeting these needs. They performed a proof of concept using Intel Distribution with

> *"The optimizations for Intel processors, the ability to support interactive on-demand queries, the ease of integration with SQL* and Oracle*, the ability to have the management infrastructure above the normal Hadoop stack—these are all really attractive to us. These are the areas where the Intel stack really has promise to us, above and beyond open source Apache Hadoop, and they are the reasons we are looking to make the switch."*
>
> Paul A. Navrátil, PhD,
> Manager of Scalable Visualization Technology,
> Texas Advanced Computing Center

Intel® Xeon® processor E5 family-based servers, and found the solution provided rapid response times for interactive queries and showed minimal performance degradation as the data set increased. Pecan Street and TACC researchers say the performance optimizations, practical enhancements, and management capabilities of Intel Distribution make it an outstanding option for Hadoop processing in the enterprise.

At a Glance

Project

- Explore the interactive performance and scalability of Intel Distribution for Apache Hadoop software for Pecan Street's big data analytics workloads

Accomplishments

- Conducted a proof of concept on a four-node, 32-core test system

Lessons Learned

- Look to Intel Distribution to handle data analytics workloads that go beyond what traditional databases can provide.
- Reduce migration costs by choosing a solution that lets you continue using existing queries.
- To facilitate large-scale enterprise deployment, choose a solution with comprehensive tools to help you set up, manage, secure, and troubleshoot your Hadoop clusters.
- Develop a device-to-cloud architecture that can handle high-frequency data generation, high quantities of data, and variable types of data, with redundancy and scalability throughout.

High-Frequency Data Generation

Pecan Street is a nonprofit consortium and a leader in testing, piloting, and commercializing smart grid technologies. Headquartered at the University of Texas, Pecan Street has gathered more than two years of energy consumption data from smart meters that capture electrical usage data every five seconds from hundreds of homes in the Mueller community of Austin, Texas. Researchers are using the data to study issues ranging from consumer incentives to the best positions for solar panels.

TACC provides data hosting, visualization, and data analysis services for Pecan Street using Intel Xeon processor-based servers and a PostGreSQL* database. Pecan Street scientists were concerned that as the project grew, they would need to slow down the beat rate of data sampling, even though that could mean missing out on insights they could gain through a more fine-grained view of the data.

"As we ramp up the data feed, we've realized that traditional SQL databases just really weren't going to work," says Paul A. Navrátil, manager of scalable visualization technologies at TACC. "Traditional Apache Hadoop approaches handle big data really well, but they don't supply us with the interactivity and on-demand processing that we need."

Interactive, On-Demand Performance

Navrátil worked with Pecan Street and Intel to conduct a proof of concept to explore whether Intel Distribution for Apache Hadoop software could provide the performance for interactive, on-demand processing and querying of Pecan Street's data.

Intel Distribution includes the full distribution from the Apache Hadoop open source project, as well as elements such as Hive*, HBase*, and Hadoop Distributed File System* (HDFS*). The Intel offering is designed from the silicon up to increase throughput, performance, and security on Intel Xeon processor-based servers, Intel® Solid-State Drives (Intel® SSDs), and Intel® 10 Gigabit Ethernet solutions. A central management console, Intel® Manager for Apache Hadoop software (Intel® Manager), provides tools to enhance securing, configuring, monitoring, and managing Hadoop clusters.

Navrátil worked with Intel and TACC engineers to load 3.6 TB of Pecan Street's data onto a Hadoop cluster at an Intel office in Chicago, and develop a script to test interactive, on-demand performance using the Intel solution stack.

The results were exciting. "We started with a test set of 17.5 million rows of data, and were able to query for a specific home source of data in that set in tens of milliseconds," Navrátil recalls. "If you're sitting on a Web site and it can return an answer in under a second, that's great."

The initial 17.5 million-row test set contained 0.15 TB of data and provided responses in 47 ms. The team also wanted to test scalability, *Pecan Street and TACC scientists wanted interactive query performance. Intel Distribution for Apache Hadoop software provided it, churning through 418 million rows of data and providing answers in 53 ms.*

Key Technologies

- Intel Distribution for Apache Hadoop software with HBase and Hive
- Four-node cluster with dual Intel Xeon processors E5-2680 (8-core) at 2.70 GHz and 64 GB of RAM, hyper-threading enabled
- Three Seagate Constellation* 1 TB internal hard drives
- 12 TB total disk space

so it ran two larger sets: one with 209 million rows comprising 1.8 TB of data, and the other containing 418 million rows and 3.6 TB of data. Both the scale sets delivered query responses in 53 ms.

"The cool thing is that in the first scale set, we gave it 12 times more data, and the response took only 6 ms longer," says Navrátil. "Then we doubled the data again, and it didn't slow down at all. There is some tremendous efficiency there, and we were able to get into the sweet spot. That really is encouraging, because it tells us we can expand the data maybe 100-fold and still get acceptable database performance. When we're getting new measurements every five seconds, that really helps."

Lowering the Cost of Adoption

In addition to its interactive performance, Navrátil says Intel Distribution provides important practical benefits for organizations moving into distributed data analytics.

"Many organizations and project teams have invested considerable time and effort to get their SQL queries the way they want them, and they don't want to have to redo that work," he says. "Intel provides a SQL interface into its stack, so you can switch from a traditional Oracle or SQL database into Intel Distribution without changing your higher-level workflow. That's a huge benefit."

The SQL interface, part of the Interactive Hive capability in the Intel solution stack, works to reduce the costs of adopting distributed analytics. "One of the costs for moving into a traditional Apache Hadoop-style database is that the old querying mechanism is no longer working," says Navrátil. "With the Intel stack, that cost goes away. You can still use the SQL query directly. You get the same benefit if you're interfacing with SQL-based third-party products. If you want to work with an analysis package that expects to query the database using SQL, you can use the Intel Distribution software stack. You couldn't do that with the conventional Apache Hadoop stack unless you wrote some middleware to handle the translation."

Improving Performance and Infrastructure Utilization

Whether a utility company focuses on consumer energy initiatives or grid management, distributed data analytics will be essential for making the most of the company's investments in smart grid and smart meter technologies. As utilities plan for big data and the Internet of Things, Bert Haskell, chief technology officer for Pecan Street, says they'll need to implement a distributed architecture with redundancy at each layer.

"You want to create a very robust device-to-cloud architecture with a significant amount of computing, storage, and backup at every layer of the infrastructure, so you can process data close to where it originates, and if a system becomes isolated, it can still function," Haskell says. "This is true from the home or office building, to the regional utility level, up to the high-performance network of microgrids."

Workload analysis should guide the infrastructure planning to ensure you provide the CPU performance for near real-time analytics, as well as the storage and network capacity to support your workflows.

"The great thing about Intel Distribution is that it contains optimizations for Intel processors, so you'll get additional performance benefits for having the latest Intel technologies—but it still pays to know your workloads," says Navrátil. "For example, Intel Distribution takes advantage of advanced processor features, like the Intel® QuickPath Interconnect (Intel® QPI), which brings data into each processor more efficiently, or the SSE vectorizing instructions. Depending on your particular data and use case, Intel Distribution is also optimized for caching hot data with SSD disk technology rather than traditional spinning disks. It's definitely worthwhile to characterize your particular problem."

"The cool thing is that in the first scale set, we gave it 12 times more data, and the response took only 6 ms longer. Then we doubled the data again, and it didn't slow down at all. There is some tremendous efficiency there, and we were able to get into the sweet spot. That really is encouraging because it tells us we can expand the data maybe 100-fold and still get acceptable database performance."

Paul A. Navrátil, PhD,
Manager of Scalable Visualization Technology,
Texas Advanced Computing Center

> *"You want to create a very robust device-to-cloud architecture with a significant amount of computing, storage, and backup at every layer of the infrastructure, so you can process data close to where it originates, and if a system becomes isolated, it can still function. This is true from the home or office building, to the regional utility level, up to the high-performance network of microgrids."*
>
> Bert Haskell,
> Chief Technology Officer,
> Pecan Street, Inc.

Weijia Xu, manager of the Data Mining and Statistics Group at TACC, also emphasizes the importance of storage architecture. Xu uses Hadoop software in a TACC research project developing new text-mining algorithms for the U.S. National Archives and Record Administration. "In general, the more memory you have, the more hard drives you have, and the faster the hard drives, the better," he says. "Solid-state drives are a big benefit to many Hadoop workloads, especially if you're doing a lot of reads and writes."

Network requirements will vary depending on the implementation of distributed analytic models. "The principle of Hadoop is to localize the processing and minimize the transfer of information over the network, but many organizations will process a variety of workloads on their cluster," Xu comments. "If you're going to be sharing information across the network or running a dynamic Hadoop cluster that supports multiple workloads from different users, you'll need very fast communication between nodes and fast networks to swap data in and out to other storage devices."

Scaling Success onto TACC's Stampede* Cluster

Pecan Street and TACC researchers are looking forward to implementing Intel Distribution and the full Pecan Street data set on TACC's new flagship Stampede cluster. The massive cluster, which ranked seventh on the November 2012 Top500* list, uses thousands of Dell PowerEdge* C8220 compute nodes powered by the Intel Xeon processor E5 family and Intel® Xeon Phi™ coprocessors.[1]

"The optimizations for Intel processors, the ability to support interactive on-demand queries, the ease of integration with SQL and Oracle, the ability to have the management infrastructure above the normal Hadoop stack—these are all really attractive to us," says Navrátil. "These are the areas where the Intel stack really has promise to us, above and beyond open-source Apache Hadoop, and they are the reasons we are looking to make the switch."

Since its founding in 2009, Pecan Street has expanded its research and development focus from home energy management to electric vehicles, distribution systems, big data analytics, and information privacy. Its Pike Powers Lab offers facilities for consortium members and other companies to test their innovations. With deployment of Intel Distribution for Apache Hadoop software, Pecan Street is well equipped to accelerate innovation for a greener world.

"The utilities are putting all the smart grid and smart metering technology in place, and they're wrestling with how they're going to deal with all the data," says Haskell. "How do they collect the data and keep it secure? How do they demonstrate that the money was well spent? How do they use the data they're collecting to run the network more efficiently and provide better service and reliability? Big data analytics and advances like Intel Distribution are important enabling technologies for this work, and ultimately can contribute to a greener and more sustainable environment."

Pecan Street Inc.

TACC

[1] http://www.top500.org/lists/2012/11/

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.

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.

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.