White Paper
Intel® Xeon® Processors
Intel® Distribution for Apache Hadoop* software
Big Data

DuPont Proof of Concept Explores Hadoop* Environment for Big Data and Big Science

In a proof of concept (POC), Intel® Distribution for Apache Hadoop* software (Intel® Distribution) on Intel® Xeon® processors shows potential to shrink a 30-day job to an estimated four days with linear scalability

Executive Summary

Scientists and IT engineers from DuPont Research IT groups in DuPont Central Research and Development (CR&D) and DuPont Pioneer worked with Hadoop experts from the Intel Datacenter Software Division to explore Hadoop's feasibility as a next-generation platform for data- and processing-intensive workloads at DuPont R&D.

The team established a Hadoop* and HBase* cluster environment and used it to assess data I/O and analytics performance for actual DuPont R&D workloads. They demonstrated linear scalability on one significant workload, and reduced a major part of the workload, which takes 30 days on the current platform, to four days.

Introduction: Big Data at DuPont R&D

Founded in 1802, DuPont is a science and engineering leader that focuses on solving some of the world's biggest problems—from reducing dependence on fossil fuels to protecting the environment and providing healthy food for a global population of seven billion.

Massive data sets are essential to many of DuPont's research projects, and several critical biological databases—driven by both field research and the explosion of public genomics data repositories—are growing beyond the capabilities of current database technology.

"Some of our databases have expanded by several orders of magnitude in the past few years," explains John Strobel, technical manager for comparative genomics research at DuPont Pioneer. "We are always looking ahead to anticipate future needs. If the amount of data expands by another order or two of magnitude, which we expect it to, we would have concerns about the ability to scale our current methods and relational databases."

To stay ahead of rapid data growth, scientists and computational experts at DuPont Research IT groups worked with Intel on a POC to see if the Hadoop environment could provide a next-generation solution for big data scalability. DuPont R&D scientists and research IT specialists selected two actual workloads where data growth, often in tandem with computational complexity, called for new approaches.

POC Technologies for Scalable Analytics

DuPont Research IT groups worked with Intel to establish a five-server testbed cluster and explore how specific workloads could benefit the native parallelism in the Hadoop environment. Hadoop is an open source framework that uses a simple programming model

to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms, and an application layer that manages distributed processing, parallel computation, workflow, and configuration management.

The POC used Intel Distribution for Apache Hadoop software, a comprehensive solution that contains the full distribution from the Apache Hadoop open source project, including MapReduce*, the Hadoop Distributed File System* (HDFS*), and components such as HBase and Hive*. Intel Distribution is optimized for Intel servers and networking platforms. It also incorporates hardware-aided security technologies and includes Intel® Manager for Apache Hadoop software, a management console to help enterprises deploy, configure, tune, monitor, and secure their Hadoop environments. The POC team also used Basic Cube* and Import TSV* for data loading.

The physical cluster consisted of four IBM System x3650 M4* servers based on the Intel Xeon processor E5 family and one IBM System x3860 M4 server based on the Intel Xeon processor E7 family. Servers were configured with 64 GB of RAM and ran Red Hat Enterprise Linux* 6.1. The cluster used direct-attached SATA disks and Intel® Ethernet 10 Gb Server Adapters.

To prepare for the POC, Intel and the DuPont Research IT groups conducted a hands-on workshop that provided DuPont R&D scientists with training and immersion in the Hadoop/HBase technology tied directly to their use cases. The training exposed the data modeling within the Hadoop technology ecosystem by creating programming routines using MapReduce and HBase for data-intensive tasks.

Use Case 1: Scaling and Speeding a Critical Comparative Genomics Workflow

In one POC use case, a multistep comparative genomics process looks for similarities and differences among thousands of genomes, which consist of over four million proteins.

The process starts with an all-versus-all comparison of all four million proteins using BLAST*, and even running as a parallel job on a 1,000-core cluster, this initial step takes two weeks to complete. The results from the initial all-versus-all comparison are then loaded into a relational database, which produces a terabyte-sized table—a full two-day process. This table is then partitioned and indexed accordingly so it can be queried within a reasonable response time. In the last step, a process runs for 30 days, including calling scripts that perform computationally intensive comparisons to find the best bidirectional matches between every pair of genomes, and loading the results into the relational database.

Since comparative genomics is on the critical path for many important research projects and initiatives and the number of completely sequenced genomes grows rapidly every year, DuPont bioinformatics experts were eager to see if the Hadoop environment could help them scale and accelerate this workload. The answer was "Yes" on both counts.

Using Hadoop, MapReduce, and HBase on the Intel cluster, the DuPont bioinformatics scientists parallelized the loading process. Using the data set previously generated by the all-versus-all comparison, the team loaded 1 percent and 5 percent of the entire 12-trillion-record data set into HBase. They observed linear scalability as they added more nodes and increased the size of the data set, and the bulk loading process was very fast.

The team also proposed that instead of having to run the initial all-versus-all comparison to completion and then spending two days to load the results into a relational database table, they could use MapReduce to run the all-versus-all comparison and feed the results directly into HBase as Hadoop generated them. This would eliminate the two-day loading step and let the team use Hive to immediately query and search the distributed data in HBase. Using HBase indexing and keys, team members achieved millisecond response times to their queries.

The team also examined the scalability of computing the best bidirectional matches by simulating the SQL query runs in HBase with the 1 percent and 5 percent of the data set loaded in the previous step, and also observed linear scalability with additional hardware. Extrapolating from those results, they projected that the SQL querying phase for the computation of the best bidirectional matches on the full data set would be done in hours running on the POC cluster.

On the current platform, the entire process takes 30 days, including the phases of loading data into the relational database, SQL querying, and some other post-processing steps. The team concluded that the solution could easily scale to handle future multi-terabyte data sets and could complete the best bidirectional match analysis for a set of 14 million proteins in just four days.

Use Case 2: Querying Multiple Data Sets for a Bioinformatics Workload

A second team of DuPont bioinformatics scientists used the POC technologies, including Apache Hive and HBase, to see if they could compare data across multiple samples by querying across different types of data sets in different databases. This is an important task that is difficult to accomplish with current database technologies. The team also wanted to get a sense of what kind of schema would best facilitate such operations. Again, the problem is compounded by large data volumes—in this case, 300 to 400 million data points per sample. Each sample ordinarily took several hours to load, and there are hundreds of samples.

The team found Apache Hive and HBase were well suited to the problem, enabling them to easily design optimized HBase tables with the same key for each data set and create a simple Java* program to query each table and extract related information. Flexible column names allowed data to be referenced using the name of the sample, which is not possible with most databases. The data from one sample was loaded within minutes, and HBase performed in near-real time for the query samples used in the POC.

These results also showed HBase and MapReduce to be an effective combination that allows the DuPont bioinformatics team to quickly analyze data with parallel computations on the cluster and rapidly access relevant data. HBase appears optimal for marker searching using row keys, while MapReduce can analyze data within a single row to compute various characteristics across DNA samples.

Practical to Deploy and Support

DuPont Research IT groups also looked at practical issues and found that the Hadoop environment could provide cost and reliability benefits compared to traditional relational databases. The scale-out architecture and features such as data replication help increase reliability, reduce the risk of data loss, and reduce costs. Capabilities such as Intel Manager simplify deployment by providing a dashboard-style interface for configuring, monitoring, and maintaining the Hadoop environment.

"We are excited about the Hadoop environment," says Guna Gurazada, a senior research associate in bioinformatics at DuPont. "Scalability is a huge factor for us, and that was impressive. The Hadoop environment appears to be suitable for some of our research problems. It helps with both data storage and computing, and it looks very feasible to deploy and support."

The POC also led to the following takeaways:

- **Choose the right tool for the right problem.** Among other use cases, the Hadoop environment can be valuable for very large, varied data sets that are expanding rapidly and growing beyond traditional relational databases. HBase can allow for cross-queries of multiple databases with simple Java programming.
- **To optimize scalability and performance,** start your HBase table design from the perspective of the application and the queries you'll want to ask. This is a shift from traditional table design, but it's feasible because HBase tables are relatively easy to set up—and the specialized table design can allow for extreme efficiency. The DuPont R&D teams created HBase keys and storage strategies with the target application and queries in mind.
- **Provide sufficient physical memory and local storage** to match the workloads, and consider solid-state drives (SSDs) for storage. Some applications will have significant benefit.
- **In choosing a Hadoop deployment,** look for adherence to open source standards, along with performance and ease of management. If you'll be handling proprietary information, consider a Hadoop solution that supports data encryption and other security technologies.

Path Forward

DuPont Research IT groups are continuing the POC with larger data sets and a broader range of problem sets. The groups are already using Hadoop for parallel computing, and see it as a logical step to add HBase and other elements of the Hadoop environment for parallel I/O.

 Printed in USA 0813/LJ/TDA/XX/PDF ♻ Please Recycle 329554-001