Accelerate Big Data Insights

ACCELERATE TIME-TO-INSIGHT FROM HOURS TO MINUTES

Cancun Systems MemoryLake™ demonstrated significantly faster time-to-insight while also eliminating infrastructure inefficiencies.

— David Vennergrund
Director, Data Science, CSRA

Executive Summary

An abundance of information isn't always helpful when time is of the essence. In the world of big data, the ability to accelerate time-to-insight can not only provide businesses with immediate competitive advantage, it can also allow mission-critical systems to better protect life, property and country. Cancun Systems provides an in-memory SDML (software-defined memory lake) platform that delivers massive acceleration, cost efficiency and deployment flexibility to big data workloads. By employing Cancun MemoryLake™ technology, organizations can get insights considerably faster to improve decision making, minimize risk, and increase profits.

Challenges of Large Datasets

Whether a MapReduce, Hive, or Spark cluster, most big data sets are much larger than the physical memory capacity of the cluster, causing a bottleneck in memory and storage I/O. This leads to poor application performance, inefficient architectures, and expensive scaling requirements. Cancun's MemoryLake™ SDML platform makes intelligent usage of available resources across memory and storage and allows analytics workloads to access data at the speed of memory but at the cost efficiency of disk. This allows organizations to query more data faster and more efficiently.

Cancun MemoryLake™: An In-Memory Software Platform for Accelerated Insights

Cancun's MemoryLake™ software platform delivers an SDML that enables applications to run up to 10X faster, allowing customers to accelerate time to insights and enjoy tremendous infrastrucutre efficiencies. Cancun's MemoryLake™ provides immediate benefits in three areas:

Faster time to Insights: By pooling and virtualizing available memory and storage resources within or across nodes, Cancun can create a software-defined memory lake. In-memory applications like Spark can now run significantly faster by accelerating and pipelining applications at memory speed, enabling workflows to complete in a fraction of the time.

Infrastructure Efficiency and Savings: Whether deployed on-premises or in the cloud, Cancun's MemoryLake™ software delivers unprecedented infrastructure efficiency. Existing build-outs can run more jobs and query more data without having to purchase additional infrastructure. New build-outs only require a fraction of the expected infrastructure. For cloud deployments, customers can experience both faster insights and immediate savings because they are able to complete jobs and decommission clusters much faster

Deployment Simplicity and Flexibility: Cancun enables businesses to deploy MemoryLake™ software in private, public, or hybrid cloud environments, and ingest data directly from various sources (e.g. HDFS, NAS, cloud object stores) for richer insights.. Installation is simple and takes only minutes. And deployment is frictionless, requiring no changes to application code or underlying infrastructures.

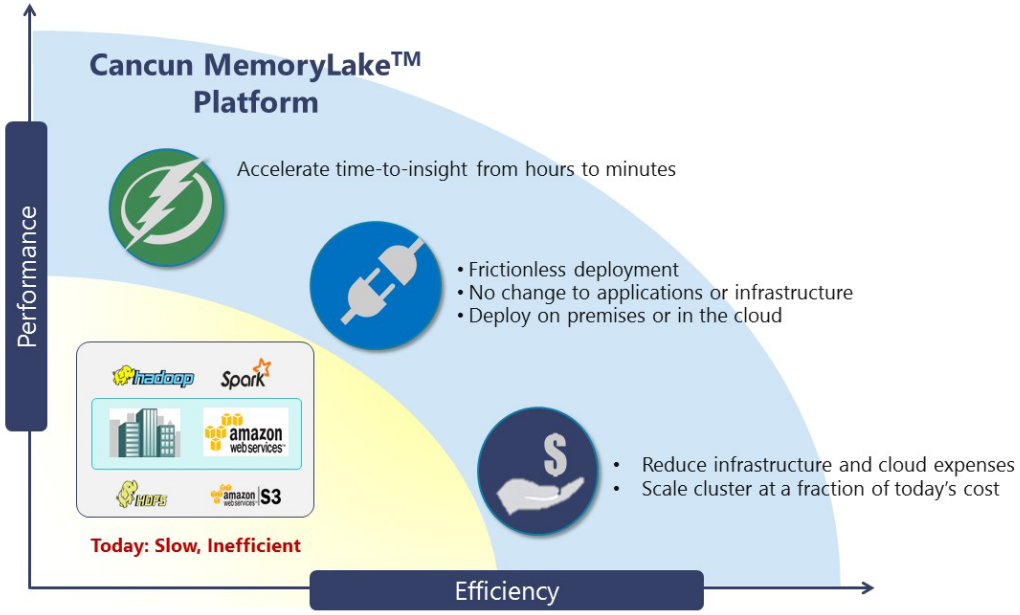


Figure 1 – Cancun MemoryLake™ technology delivers massive speed, agility, and cost efficiency to existing Big Data frameworks

Virtualizing Multiple Tiers of Memory

Cancun abstracts physical memory and storage resources resident in a node to give the impression of a very large memory pool available for memory-speed data access. It can also pool memory and storage from a remote node which makes deployment very easy. For example, in existing deployments, customers can add a new memory/SSD-dense node and dramatically improve the performance of the entire cluster at once.

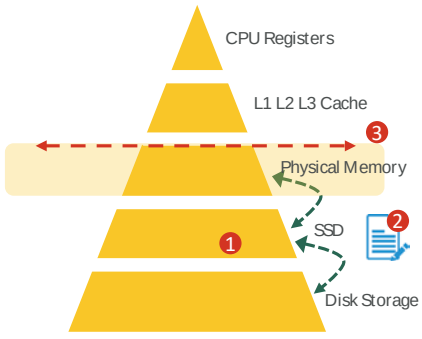


Figure 2 – Cancun MemoryLake leverages memory and different classes of storage (1) and uses simple policies (2) to manage data so applications access data at memory speed (3)

The Cancun MemoryLake™ platform automatically caches or evicts data using simple policies so that applications see orders-of-magnitude larger memory footprint. Cancun supports RAM, NVMe, SSD, HDD, and has built-in support for upcoming 3D Xpoint for even faster acceleration.

Off Heap Memory Management for Big Data

When large amounts of data are involved, issues with Java memory management can arise resulting in a significant hit to performance. Java's inability to handle large data sizes in JVM results in frequent, expensive garbage collection during which there is a significant drop in performance. In addition, if the JVM crashes, all data in memory is lost and must be rebuilt from disk – a slow and cumbersome process.

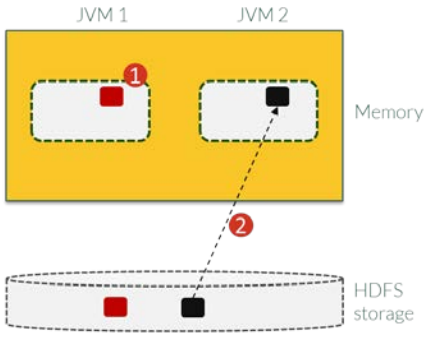


Figure 3 – Garbage collection (1) slows applications; if JVM crashes (2), it must be rebuilt from disk

Cancun MemoryLake™ technology significantly speeds up jobs by avoiding garbage collection. Data blocks are moved off heap to remove the load on garbage collection and a persistent distributed cache ensures that data can quickly be fetched.

In addition, if the JVM crashes, data is quickly retrieved from off-heap memory without having to read from slow HDFS, because with Cancun MemoryLake™ the data remains in memory, making crash recovery much faster.

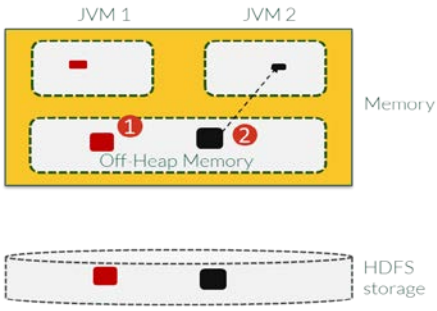


Figure 4 – Cancun avoids garbage collection (1) and data is quickly retrieved (2) if JVM crashes

Cancun MemoryLake™ Accomplishes Data Transfer via Fast, In-Memory File System

Big data workloads are typically built as pipelines. The output of one stage is fed into the next stage and this output is written to disk, which becomes a chokepoint.

Using Cancun MemoryLake™, data transfer across stages is done via in-memory file system (see notation 1 in Figure 5), which is an order of magnitude faster than disk-based file systems. For disk access within a stage (see notation 2 in Figure 5), Cancun allows numerous intermediate writes to disk be done via in-memory file system so that pipelined jobs are completed dramatically faster.

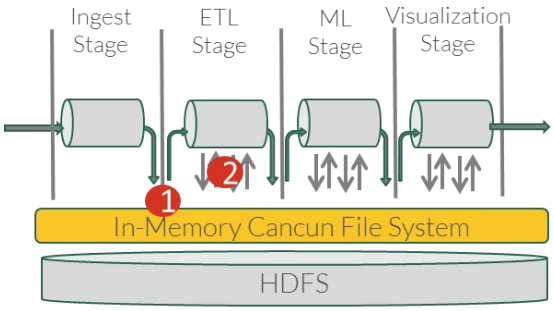


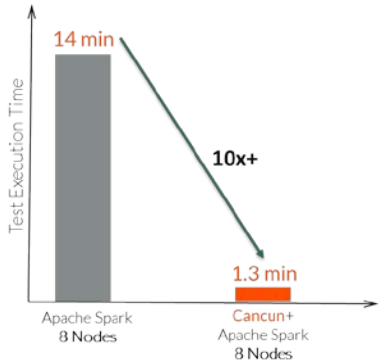
Figure 5 – Disk access is done via in-memory file system for blazing performance

Proof Points

Shown below are the results of key Spark and Hadoop tests that were run with and without Cancun MemoryLake™ technology.

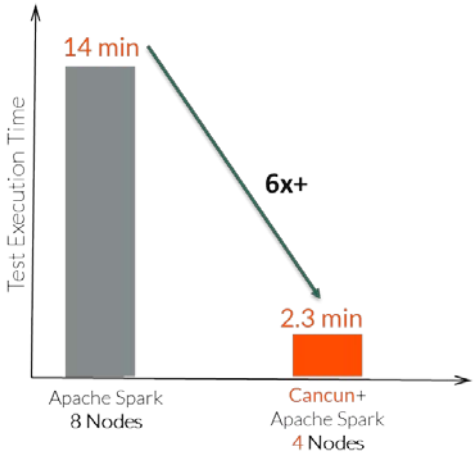
1. Real Customer ETL Scenario

Performance



The customer scenario was to JOIN 1B+ rows, spread in two data files, and then use the 'join'-ed file for downstream analysis. The test was run in both "without Cancun" and "with Cancun" scenarios. The performance on an infrastructure consisting of 8 nodes without Cancun was 14 minutes, while the run time with Cancun was 1.3 minutes – a more than 10x improvement in performance.

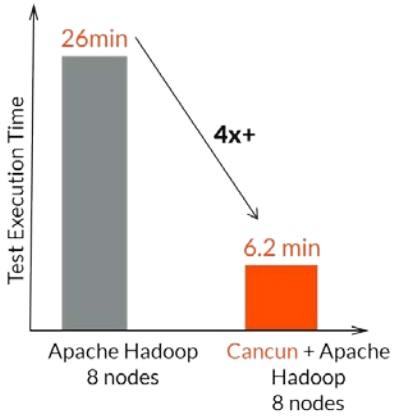
Infrastructure Efficiency



Efficiency testing on this scenario also demonstrated significant improvements. The run time on an infrastructure consisting of **8 nodes** without Cancun was 14 minutes, while the run time on an infrastructure consisting of only **4 nodes** with Cancun was 2.3 minutes. In essence, the Cancun environment dramatically cut run time while using only a small portion of the existing infrastructure footprint.

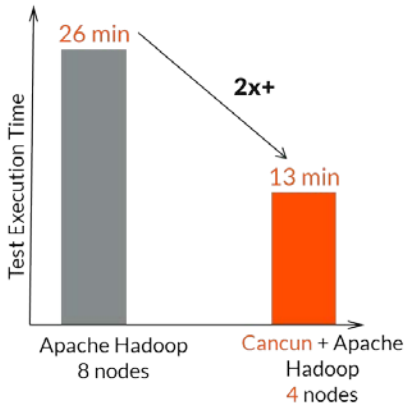
2. HiBench TeraSort Benchmark

Performance



The HiBench TeraSort test was run in both "without Cancun" and "with Cancun" scenarios. The performance on an infrastructure consisting of 8 nodes without Cancun was 26 minutes, while the run time with Cancun was 6.2 minutes – a more than 4x improvement in performance.

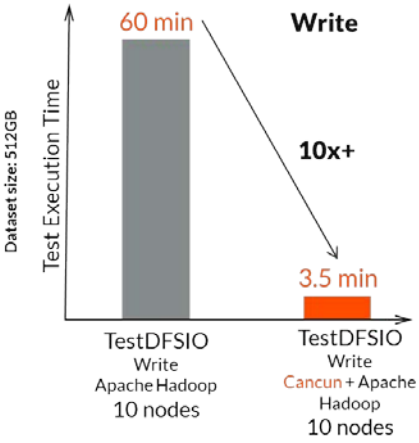
Infrastructure Efficiency

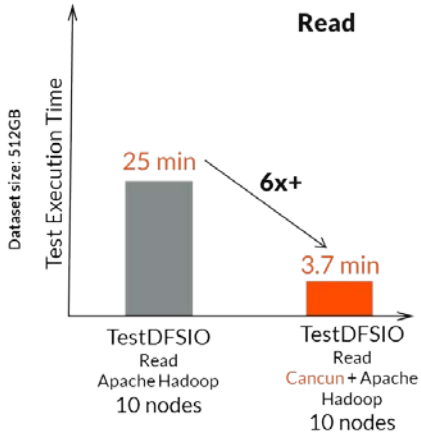


Efficiency testing on the HiBench benchmark also showed significant improvements. The run time on an infrastructure consisting of **8 nodes** without Cancun was 26 minutes, while the run time on an infrastructure consisting of only **4 nodes** with Cancun was 13 minutes. In essence, the Cancun environment dramatically cut run time while only using a small portion of the existing infrastructure footprint.

2. TestDFSIO Benchmark

Performance

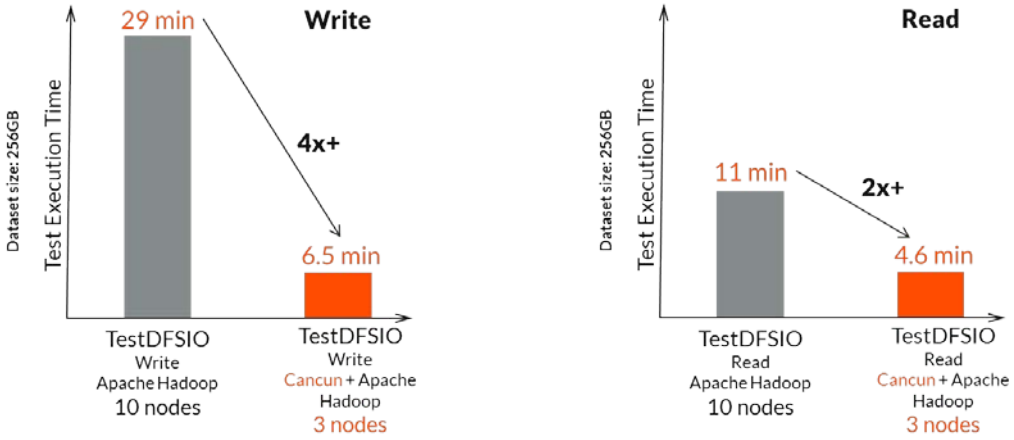




The Hadoop TestDFSIO test was run in various "without Cancun" and "with Cancun" scenarios. For a DFSIO write test on an infrastructure consisting of 10 nodes and a 512GB dataset, the run time without Cancun was 60 minutes, while the run time with Cancun was 3.5 minutes, which represents a more than 10x improvement in performance. For reads, the run time was recorded at 25 minutes and 3.7 minutes,

respectively, demonstrating a more than 6x improvement when using the Cancun MemoryLake™ platform.

Infrastructure Efficiency



Similarly, tests run on a smaller (3 node) Cancun-enabled infrastructure demonstrated much faster results than in a larger (10 node) non-Cancun environment. For writes, run time without Cancun was 29 minutes versus 6.5 minutes with Cancun, which represents a more than 4x improvement. For reads, run time without Cancun was 11 minutes versus 4.6 minutes with Cancun, which is a 2x improvement.

Conclusion

Optimized for big data workflows, Cancun MemoryLake™ solves the inefficiency of memory management in today's big data infrastructures. The innovative Cancun platform pools memory resources across nodes so jobs can take advantage of fast, in-memory resources to deliver stunning performance that enables organizations to improve decision making, minimize risk, and increase profits.

For more information or to request a demo, visit us at www.cancunsystems.net