Case Study

Intel® Optane™ Persistent Memory
Baidu BigSQL Data Processing Platform
Optimized Analytics Package for Spark

# Baidu BigSQL Delivers Faster Spark Interactive Queries with OAP & Intel® Optane™ Persistent Memory

Over the past few years, the world's data volume has grown almost exponentially, which means companies, especially tech companies, are facing greater challenges in meeting service time requirements. Apache Spark, a unified analytics engine for large-scale and high-performance data processing, is designed to meet this challenge. One module of Apache Spark - Spark SQL is widely used for working with structured data in large data centers. Baidu's BigSQL data processing platform is based on Spark SQL and has many features and performance enhancements that improve on it.

One important enhancement pertains to meeting sub-second performance requirements for interactive queries. This is where Intel and Baidu collaborated to create the Optimized Analytics Package (OAP) for Spark Platform project. OAP is designed to leverage the columnar data format and user-defined indexes built over selected columns, leading to improved data scanning efficiency. It also adopts a fine-grained in-memory data caching strategy to remove I/O bottlenecks in disks and networks, maximizing performance to sub-seconds.

As Baidu's business expands, the scale of hot data grows rapidly. Memory scaling is needed to deliver the same level of performance that users demand. However, the high cost of Dynamic Random-Access Memory (DRAM) adds increasing pressure to the Total Cost of Ownership (TCO). To lower TCO while ensuring satisfactory performance, Baidu and Intel collaborated and introduced Intel® Optane™ persistent memory (PMem) as a more cost-efficient solution to replace DRAM.

Baidu's internal testing has demonstrated that Intel Optane  PMem improves OAP cache performance and performance-per-dollar output when compared to solutions without PMem, leading to direct business impacts such as the optimization of its ad hoc query service, Tuling, by offloading its workload and reducing average query latency.

> "In order for Baidu Big SQL to provide users with high-performance ad hoc query services, large memory is needed to cache hot data locally on compute nodes to avoid DFS I / O slowing performance down. With Intel Optane persistent memory, we managed to ensure outstanding cache performance, while at the same time greatly improving cluster processing and achieving significant TCO benefits."
>
> **LI Shiyong**
> **Senior System Engineer**
> **Baidu**

## Baidu BigSQL with OAP

One fundamental characteristic of Spark SQL is that it is designed to deliver optimized performance for batch processing. However, some of Baidu's service queries have totally different characteristics. They are called interactive queries. Usually, they query over a large dataset with specific filtering conditions, serving the dedicated purpose of identifying a relatively small amount of data. Users expect this small amount of queried data to be returned in seconds or even sub-seconds, instead of the usual minutes or hours seen in batch processing, which is usually not possible for the current Spark SQL implementation.

To solve this problem, Baidu and Intel collaborated and implemented OAP, which uses index and caching techniques to accelerate interactive query response. By integrating OAP, Baidu BigSQL successfully achieved the desired level of interactive query performance.

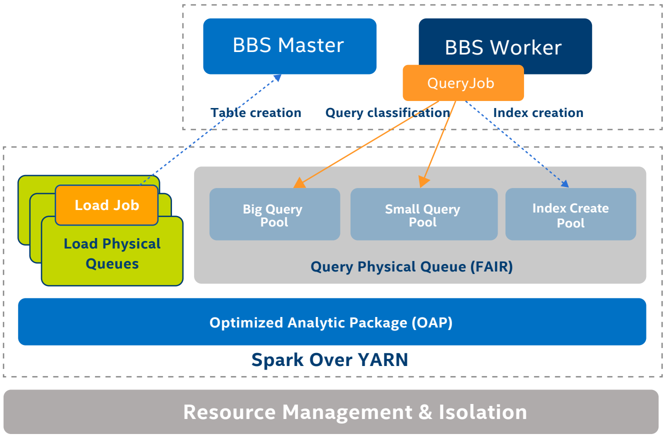

Figure 1. Baidu BigSQL and OAP Integration

When a query has specific filtering conditions, indexes can be created over the columns with such conditions. By creating and storing a full B+ Tree index side-by-side with the columnar data file, OAP can identify target rows by quickly searching through the B+ Tree index, and skip unnecessary data scans over backend storage such as HDFS. Furthermore, the index file is separated from the original data file. This makes it possible to create or drop indexes without the need to rewrite the original data files.

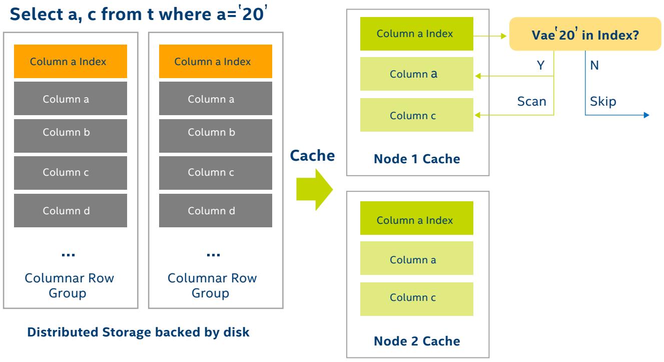

Figure 2. OAP Cache & Index Concept

To further reduce query response time from seconds to sub-seconds, OAP optimizes index and data access with cache. By caching the index and data in memory, index loading and data scanning get orders of magnitude faster, avoiding disk and network I/O overhead when reading from distributed file systems. What's more, index and data can be configured with separate caches, enabling independent eviction and memory space management for both.

Additionally, now that the cache is at the column level, it is possible to cache the columns required for the query exclusively. And based on the Least-Recently-Used (LRU) policy, those least-recently-used data items will be evicted from the cache if maximum capacity is reached, allowing more recent data items to be cached. Guided by this policy, an advanced cache manager is implemented in Baidu BigSQL to proactively populate hot columns, and retire columns no longer required in cache.

## Baidu BigSQL Optimization with Intel Optane Persistent Memory

When the data scale is small, Baidu BigSQL can deliver optimal performance by caching index or data in DRAM. However, as Baidu's business continues to grow, datasets are rapidly evolving in size. When cache space becomes too small to accommodate large amount of hot data, performance will suffer.

The simple solution is to add more DRAM, but there are several disadvantages. First, the price-per-GB is high, putting great pressure on TCO. Second, memory is a precious resource for computation, especially so in Spark's environment where the total DRAM capacity that can be configured on each node is limited. Third, even though DRAM has higher random-access bandwidth and lower latency, such benefits will be wasted when it is used for caching large data blocks and characterizing sequential access. To find more cost-effective alternatives, Baidu and Intel worked together to integrate Intel Optane PMem.

Intel Optane PMem is an innovative technology that delivers a unique and affordable combination of large memory capacity and persistence. It represents a new class of memory and storage technology, explicitly architected for data centers. It offers several key benefits that match the specific requirements of Baidu BigSQL:

- High bandwidth for sequential read
- Large capacity and affordable cost

Intel Optane PMem supports two operating modes. When configured for Memory Mode, the applications perceive a pool of volatile memory no differently than they do on DRAM-only systems; when configured in App Direct Mode, the application can direct how to use available space. Since OAP cache has the specific purpose of indexing and inputting data, App Direct Mode is used to ensure the application has full control of how to use the device. In addition, the cache can be repopulated from backend storage and does not need to be persistent. OAP uses the memkind library to access PMem without persistency and corresponding performance penalties.

To use PMem in place of DRAM, Intel extended OAP to allow memory manager plugins, and implemented a PMem-based memory manager to allow the allocation of cache space in PMem. Users can switch between DRAM and PMem, or even mix the two, for instance using DRAM to cache index while using PMem to cache data.

Additionally, to fully integrate PMem with Baidu's specific OS environment, Baidu and Intel carried out further wide-ranging collaborations in areas including hardware, operating system and libraries.

To validate the performance and benefits of Intel Optane PMem in OAP, Baidu conducted several evaluations and internal tests, first with decision support benchmark queries and then with Baidu's real workload queries. The main objective was to test and understand the cost-efficiency of PMem.

In the case of testing with decision support benchmark queries, firstly the dataset size is capped at 1TB, and DRAM and PMem are configured at the same capacity. Test results show that they are both able to cache all the data, and PMem is only slightly behind DRAM in performance (11.7%), while its cost is a lot lower[1]. When the dataset reaches 3TB, and DRAM and PMem are at the same cost, DRAM can no longer cache all the data due to its lower capacity. In comparison, PMem does not only have higher capacity to cache all the data, it shows much better performance – 6 times better[1]. DRAM has poor performance in the second scenario because when data size greatly exceeds cache size, DRAM needs to read data from backend storage frequently which delays the response time. Decision support benchmark query tests show clear evidence that when at the same cost level, Intel Optane PMem can provide larger capacity and higher performance than DRAM.

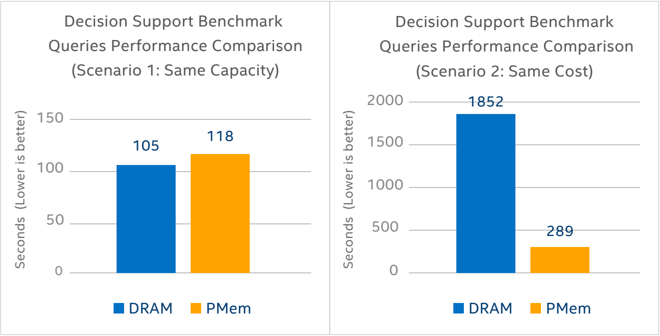

**Figure 3. DRAM and Intel Optane PMem Comparison Tests[1]**

The next stage of testing is based on the same two scenarios, but with Baidu's actual workload and a slightly different approach. In the first scenario, both DRAM and PMem are tested to cache 50% of the frequently used columns. Results show that the PMem caching speed is only about 12% lower than DRAM[1]. And since its cost is disproportionally lower, it is the more cost-efficient solution. In the second scenario (DRAM and PMem at same cost), only PMem has the capacity to cache all the hot data columns and it demonstrates a 22% performance improvement, while avoiding 30% of I/O requests to underlying systems[1].

Based on these test results, Baidu concluded that Intel Optane PMem can replace DRAM in BigSQL as a more cost-efficient cache solution. Since then, Baidu deployed PMem in BigSQL, and used it to optimize its ad hoc query service – Tuling. Supported by Intel Optane PMem, the cluster offloaded more than 30% of the workload fromTuling[2]. Additionally, after deploying PMem, the average query latency reduced by 20%[2]. The Spark/OAP performance per PMem server instance improved by 50% on Tuling Spark SQL workload, at an additional cost of only 20%[2].

## Outlook

Emerging trends are driving big data technologies to change and evolve. The focus is shifting from providing key functionalities to cloud based solutions, with in-depth optimizations to meet performance targets and reduce cost. In the future, as Baidu's BigSQL becomes cloud based, Intel Optane PMem will bring to it more significant advantages in terms of performance and TCO.

And beyond input data cache acceleration for Spark SQL, with its high capacity and high bandwidth, PMem has an even bigger role to play in Spark-based machine learning and deep learning scenarios which require many computational iterations in order to process very large volumes of data. Furthermore, Spark shuffle can be optimized to access PMem through RDMA and utilize it as shuffle storage, further reducing shuffle latency and improving performance.

Going forward, Baidu and Intel will continue working together to optimize Spark. As Intel Optane PMem and 2nd Generation Intel® Xeon® Scalable Processors become more advanced, Baidu and Intel will be able to leverage them to introduce more acceleration features to Spark, pushing performance and cost-efficiency to the next level.

1. The evaluation performance data was given on January 31, 2019. For more complete information about these test results, please contact Baidu. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

2. The production performance data was given on August 16, 2019. For more complete information about these test results, please contact Baidu. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

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 complete information visit www.intel.com/benchmarks.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available security updates. No product can be absolutely secure.
See configuration disclosures for details.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.