WHITE PAPER

Data Center
Advanced Analytics

Tame the Data Deluge

Harness the power of your data to enable deeper business insights and pave the way to AI

Table of Contents

Executive summary

The world's data is doubling every two years, driving an expected 50-fold increase in data between 2010 and 2020[1]. As these volumes grow, the variety of data types and the number of data sources are also rapidly expanding, creating more data silos and increased complexity in IT infrastructures.

A first step in bringing order to this complexity is to unite multiple data siloes into a centralized data lake. However, without the right checks and balances in place, this could quickly become a data swamp – impenetrable and opaque. To tame the data deluge facing your organization, you must start with a robust holistic, and well-defined data strategy for capturing, managing, sharing and understanding your increasingly complex and valuable data assets. This must be coupled with stringent data policies that govern the entire data lifecycle.

This is easier said than done for many organizations. In fact, less than one percent of the data available today is actually analyzed and used[2], because it is not easily accessible. This creates a significant challenge for those wanting to adopt advanced analytics and artificial intelligence (AI) to enhance business operations and strategy. These technologies depend on large volumes of clean, high-quality data to build robust and accurate models and to derive timely insights, so the need to address this issue is more urgent than ever.

This white paper outlines ways that you can turn your data into a business advantage, delivering impactful value and opportunity across the enterprise.

Taming the data deluge will require an intelligent data strategy that is tightly coupled with a modernized data infrastructure. To do this, enterprises need to address some key areas:

1. Prep your data
2. Ensure you are using the right media for the right data with effective data tiering
3. Optimize your data protection and redundancy
4. Take advantage of accelerators to optimize data ingestion
5. Develop and deploy data governance and security policies

1. Prep your data

The first step in taming the data deluge involves modernizing your data landscape and moving away from data silos. Breaking away from fragmented systems and older data storage models that keep your data trapped will give you the drive and flexibility to implement the infrastructure you need to foster innovation.

Montefiore Health System

This leading health system in New York had fragmented systems and data silos, which hindered its ability to drive innovation in healthcare practices.

It deployed a modern data hub built on Intel® Xeon® processors.

As a result, it is now able to identify high-risk patients in need of critical, time-sensitive intervention. Accurate prediction of prolonged ventilation detects patients with more than 70 percent likelihood of an event, 48 hours in advance of a fatal episode[3].

Gaining a full view of what data you have, and where it is, if not impossible, is likely to be extremely expensive, thanks to costs associated with storing the data, accessing it in a timely manner and then preparing it for use.

As an alternative to this approach, many organizations are now implementing a modern data lake model, which can enable you to aggregate, store and analyze data of any type, in any volume and of any variety in real or near real-time.

You can then build an end-to-end analytics infrastructure on top of this foundation, enabling you to access both depth and breadth of data and map the right analytics engine and AI approach to the right workload.

Caesars Entertainment

This international casino-entertainment company implemented a new data environment using Cloudera* Enterprise (Hadoop* cluster) running on the Intel® Xeon® Processor. This has enabled it to expand data analysis to include both unstructured and semi-structured data, accelerating processing for analytics and marketing campaigns.

It can now process key jobs in 45 minutes, compared to six hours previously. This means it can process more than 3 million records per hour and carry out fine-grained customer segmentation to improve its marketing results[4].

When building your data lake, it is essential to have a well-defined data strategy in place (from data ingestion, storage, and data transformation through to data modeling and visualization). This should be coupled with stringent data governance, lineage and security policies. And of course all this will be underpinned by a modern, agile technology stack, including hardware, software, storage and networking.

2. Use the right media for the right data

With any data modernization endeavor, data storage and tiering will form an important part of your technology stack and should be a key component of your strategy. Not all data is created equal so it's important to understand the types and sources of data that are most valuable for your organization. This will then enable you to define your data tiering strategy and determine the different levels of data and its use (for example, 'hot' data that is constantly in use and business critical, through 'warm' and 'cold' to 'frozen' data which may never need to be accessed but is kept for compliance or business policy reasons).

It is then important to assess your existing storage architecture and ensure you have a tiered storage model that matches your data tiers in order to reduce total storage cost while optimizing performance. Tiers are determined by performance and cost, and data is ranked by how often it is accessed. This approach places the most critical (or 'hot') data closest to the systems where it will be needed, reducing the latency to access that information. Hot data often sits on relatively more powerful, scalable and high-performance systems. Less important, 'cold' data can be placed on cheaper, less powerful systems.

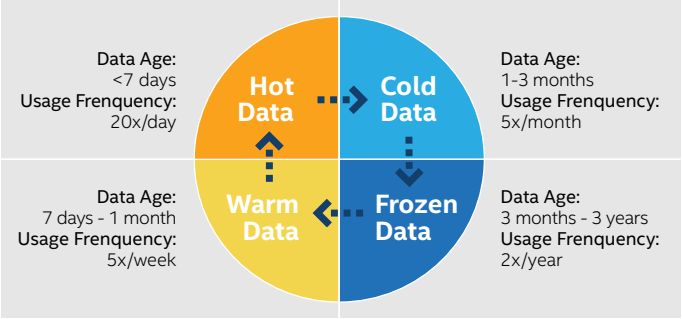

Figure 1. Four tiers of data

In the past, a gap has existed between fast, in-memory and slower storage technologies, forcing organizations to make trade-offs between placing their data on cost-effective but slow hard disk drives, or on much faster but costly dynamic random-access memory (DRAM). As more and more real-time data processing and analytics use cases emerge, it is now more important than ever that the data needed for these mission-critical workloads be kept near to the processor to reduce latency.

A shift from data silos, fragmented systems, and older data storage models to a modern data management infrastructure is essential.

The development of 3D NAND technology, which stacks memory vertically on the chip, enables more layers of memory cells to be held on the same surface areas. This increased density brings greater performance while keeping costs low.

At Intel, we have taken this concept further and developed a revolutionary non-volatile memory (NVM) technology that provides heightened throughput, low latency, high quality of service and high endurance. Based on this technology, Intel® Optane™ SSDs are optimized to handle analytics and AI workloads at high speed, breaking through storage bottlenecks by disrupting the traditional data tiering model. They can be applied in the data center to enable fast caching and storage for applications while reducing transaction costs for latency-sensitive workloads and increasing scale per server. This means organizations can achieve more memory for the same cost as DRAM, or the same memory for significantly lower cost.

Not all data is created equal. Prioritize the data that is most valuable to your business to guide your data tiering strategy.

As Intel Optane technology is also persistent, it also removes the need to wait for data to load as is the case with non-persistent DRAM. Data centers based on the latest Intel® Xeon® processors can now deploy bigger and more affordable datasets to gain new insights from larger memory pools. Using Intel Optane technology and 3D NAND SSDs helps move these larger, more complicated data sets with more speed, boosting the performance of large-scale data and analytics systems.

The addition of this option to your storage portfolio means you can create richer, more flexible data tiering strategies that optimize cost, performance and throughput to support large-scale advanced analytics workloads.

3. Optimize your data protection and redundancy

To achieve the growth and performance you need for your analytics and AI workloads, it is essential to optimize the way you protect your data and manage redundancy. By not taking this step, you may be leaving significant performance untapped while also incurring higher costs.

A number of techniques are available to help here, so it is worthwhile to research the ones that will be most appropriate for your data environment.

Data saving and encoding approaches often involve the use of Hadoop Distributed File System (HDFS) RAID for fault tolerance in the event of a disk failure. This is generally done using 3X replication, which is expensive. The supplemental storage needed for data redundancy proves costly. The HDFS process can require a 200 percent cost overhead when accounting for storage space, network bandwidth, and other

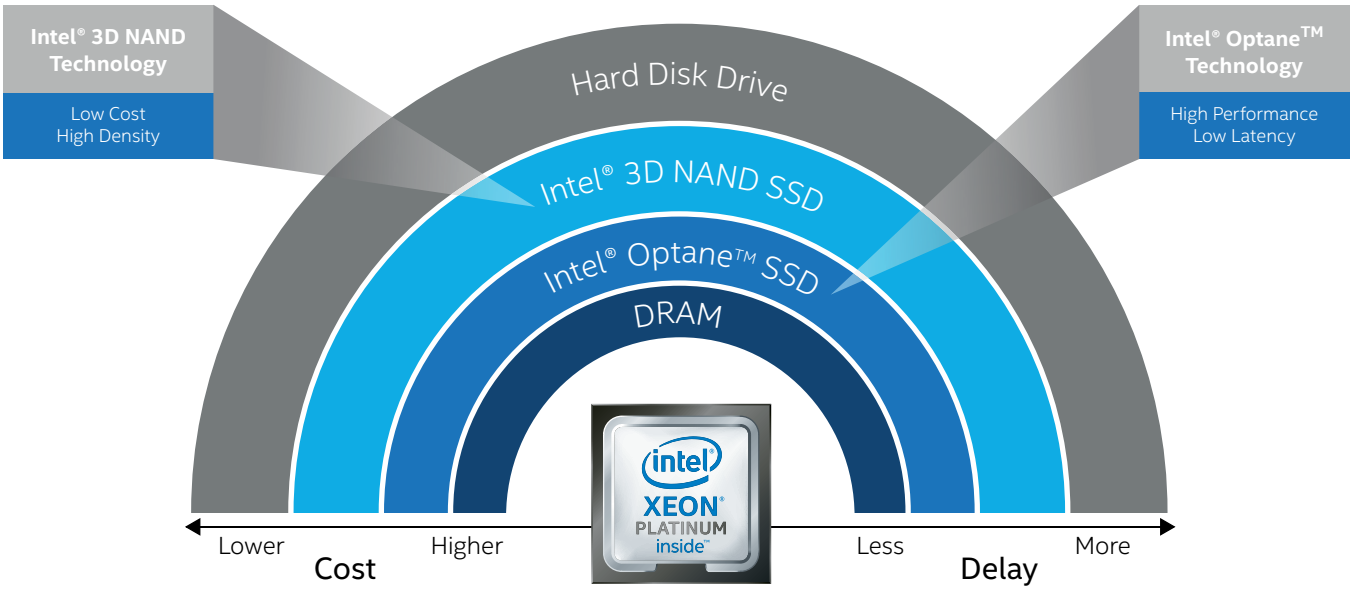

Figure 2. Cost/latency comparison of data storage media

Less HDFS storage space with erasure coding[6]

resources. In cases where an organization uses HDFS to preserve rarely accessed datasets, the infrastructure requirements can waste capital expenditures which could provide greater value when used elsewhere in an organization.

Today, enterprises have the alternative option of tapping the erasure coding (EC) process. EC breaks larger data blocks into smaller pieces, then preserves them in a very space-efficient way using erasure coding groups which combine both data and parity cells. When a cell is lost, the other cells in the group contain the information to restore the missing bits. Often, this approach can reduce storage cost by approximately 50 percent compared with HDFS.

Faster hash computation - accelerating storage de-dupe performance[6]

Because of the demanding algorithms used by EC for coding and decoding data, slower processors can create a bottleneck which leads to data latency. To address this challenge, Intel® Intelligent Storage Acceleration Library (Intel® ISA-L) works in parallel with the processors to speed EC data I/O, enhance security, and increase resilience. Because the open source Intel® ISA-L features optimizations for hashing functions, the library evaluates data for storage on-the-fly. If a particular data set is stored already, Intel ISA-L recognizes it, and prevents unnecessary duplication which hampers performance and hoards storage space.

The acceleration delivered by Intel ISA-L can increase de-duplication speeds by 200 percent[5], helping increase storage efficiency and enable more discrete data points to be stored for analytics. Faster data access also helps shorten time to results or allows for more nuanced analysis within a given response time.

Optimize your data to ensure strong performance, scalable growth and lower costs.

Michael J. Fox Foundation

This leading research institute is trialing the use of data from wearable monitors, apps, and other sources to accelerate a cure for Parkinson's disease. It deployed a modern data infrastructure built on the Cloudera* distribution of Hadoop* and Intel® Xeon® processors and applied Intel algorithms to analyze workloads from these multiple data sources. Although the immediate goal is to improve the quality of life for Parkinson's sufferers and lead clinical research scientists to potential cures, the information from these trials will undoubtedly also help people with other Parkinsonian disorders[7].

4. Take advantage of accelerators to optimize ingestion

Intel® FPGAs are streaming, parallel accelerators that attach directly to copper, fiber and optical wires. They can move any data in any format from wire to memory in nanoseconds without the need for a Network Interface Card (NIC). FPGAs make it easier for enterprises to optimize data ingest to handle spikes in data for a relatively low cost of ownership, making them a key component of the modern data center infrastructure. As they can be re-programmed to accommodate changing needs, they are also important for future-proofing your infrastructure investments and ensuring new use cases can be adopted using existing hardware.

For example, Intel Stratix 10 MX FPGAs – the industry's first FPGAs with integrated High Bandwidth Memory DRAM (HBM2) – offer up to 10x the memory bandwidth when compared with standalone DDR memory solutions[8]. With High Performance Data Analytics (HPDA) environments, streaming data pipeline frameworks like Apache Kafka* and Apache Spark* Streaming require real-time hardware acceleration. Intel® Stratix® 10 MX FPGAs can simultaneously read/write data and encrypt/decrypt data in real-time without burdening the host CPU resources.

5. Develop and deploy data governance and security policies

In today's rapidly moving marketplace, data is akin to a new form of currency for enterprises. As such, it should be preserved, protected, and shared in a way commensurate with the value it brings to an organization.

In the past, enterprises considered data management the responsibility of the IT team. Today, governance surrounding data collection and usage impacts all levels of the organization. Those companies successful in adopting a data-driven culture have defined and agreed upon ways to store, organize, manage, analyze, and share valuable data

across the enterprise. By breaking down silos isolating "departmental" data within an organization, and exploring broader data trends, bigger-picture insights can better inform individuals making critical business decisions. For example, a finance department with real-time information from sales programs can better forecast corporate revenue. Trends in social media commentary may suggest development improvements for current products or services. Real-time website data regarding customer preferences can identify new ways to provide value now and in the future. Ultimately, a data-driven company which enforces governance gains a significant competitive advantage over those organizations that do not. You should research your specific policy requirements (which can vary by industry, government requirements, legal and company privacy rules etc.) and establish policies with integrated checks and audits to ensure adherence.

Strong foundations for advanced analytics

The first step to accelerating insights and ultimately enabling AI in the enterprise is getting your data house in order. While this may seem like a daunting undertaking, it is worth the effort. By implementing processes and tools now to help tier, optimize and control your data, you will create solid foundations upon which your data-driven business can grow – equipped with deeper insights and stronger capabilities.

Intel® Xeon® Scalable Processors

The latest Intel® Xeon® Scalable processors, combined with Intel® FPGAs, provide acceleration and flexibility to address the performance requirements as your analytics initiatives expand to include advanced analytics and artificial intelligence workloads - from machine to deep learning.

Explore your opportunities to embrace advanced analytics and AI further:

- White Paper: Five Steps to Delivering the Data Driven Business
- Solution Brief: Solving the Big Data Analytics Riddle
- Solution Brief: Breaking Down Big Data Barriers with BlueData*
- Solution Brief: Intel® Select Solutions for Microsoft SQL Server* Business Operations Solution Brief

[1] https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/

[2] https://www.forbes.com/sites/bernardmarr/2016/11/01/20-mind-boggling-facts-every-business-leader-must-reflect-on-now/#1088273720dc

[3] https://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/montefiore-advaning-patient-care-solution-brief.pdf

[4] https://www.intel.com/content/www/us/en/big-data/xeon-entertainment-caesars-case-study.html

[5] https://blog.cloudera.com/blog/2015/09/introduction-to-hdfs-erasure-coding-in-apache-hadoop/

[6] https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSErasureCoding.html

[7] https://newsroom.intel.com/news-releases/the-michael-j-fox-foundation-and-intel-join-forces-to-improve-parkinsons-disease-monitoring-and-treatment-through-advanced-technologies/

[8] As compared to a standard DDR 2400 DIMM. See the "Intel® Stratix® 10 MX Devices Solve the Memory Bandwidth Challenge White Paper" at https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/wp/wp-01264-stratix10mx-devices-solve-memory-bandwidth-challenge.pdf for more information.

0418/JS/CAT/PDF 337399-002EN