INFINIDAT

STORING THE FUTURE

Storage for Big Data and Analytics Challenges

White Paper

Abstract

Big Data and analytics workloads represent a new frontier for organizations. Data is being collected from sources that did not exist 10 years ago. Mobile phone data, machine-generated data, and web site interaction data are all being collected and analyzed. In addition, as IT budgets are already being pressured down, Big Data footprints are getting larger and posing a huge storage challenge.

This paper provides information on the issues that Big Data applications pose for storage systems and how choosing the correct storage infrastructure can streamline and consolidate Big Data and analytics applications without breaking the bank.

Introduction: Big Data Stresses Storage Infrastructure

Big Data applications have caused an explosion in the amount of data that an organization needs to keep online and analyze. This has caused the cost of storage as a percent of the overall IT budget to explode. A study recently performed among Big Data and analytics-driven organizations discovered these top 5 use cases for this data explosion:

1. Enhanced customer service and support
2. Digital security, intrusion detection, fraud detection and prevention
3. Operational analysis
4. Big Data exploration
5. Data warehouse augmentation

Enhanced customer service is always on the minds of organizations, both large and small. "How can I help improve the relationship with my customer? I know my customer will choose to buy from me, and buy more often, if I cultivate and enhance the relationship." Typically this is preference data gathered from various sources, online navigation, and search criteria. Collected through the lifecycle of the customer, this data is mined and written alongside customer records.

Digital security and intrusion detection is very important to customers. This data is collected and analyzed in real time, and is typically machine generated. The analytic results must be returned immediately for this activity to be relevant. This requires fast storage with lots of capacity, as machine-generated and sensor data can consume large capacities.

Operational analysis involves collecting data, many times sensor-based from other machines, and using that data to identify areas of operational improvement and conduct fault isolation and break-fix analysis. Manufacturing firms collect up to the second data about robotic activities on their shop floor and want to know not just the status, but how they can improve the process. Like intrusion detection, this data is generated and analyzed in real time, and the results must be stored and sent back up the chain to be actionable. Unlike intrusion detection, all data is interesting and shows machine and process trends, which can be of use later.

Big Data exploration: How do you know what Big Data is until you find out what you are collecting, what you are not, and identify what is missing. Normally this is done by collecting more and more data.

Data Warehouse augmentation: "How do I take my existing analytics data, typically in a Warehouse or Mart form, and augment the data feeds from outside sources to improve accuracy, reduce execution time, and give me the answers I need without reinventing the wheel?" Data warehouse adoption is becoming widespread, even for smaller organizations, as transactional data analysis is becoming a requirement for any organization at any level.

All of these use cases require more storage and more compute power. Big Data is now considered production data, so availability, recoverability and performance are just as important as for the transactional systems within the organization. And as stated before, the trend is for IT budgets to get smaller, not bigger. These diametrically opposed forces are creating a change in the storage industry. How can you do more with less while taking into account that you also don't want to compromise on system and application reliability, efficiency or performance.

The demands that Big Data and analytic workloads place on enterprise storage can be summed up as follows:

- Must have excellent performance
- Must have extremely high density
- Must have excellent uptime, high reliability
- Must be easy to use, easy to manage, easy to provision
- Must have attractively low total cost of ownership

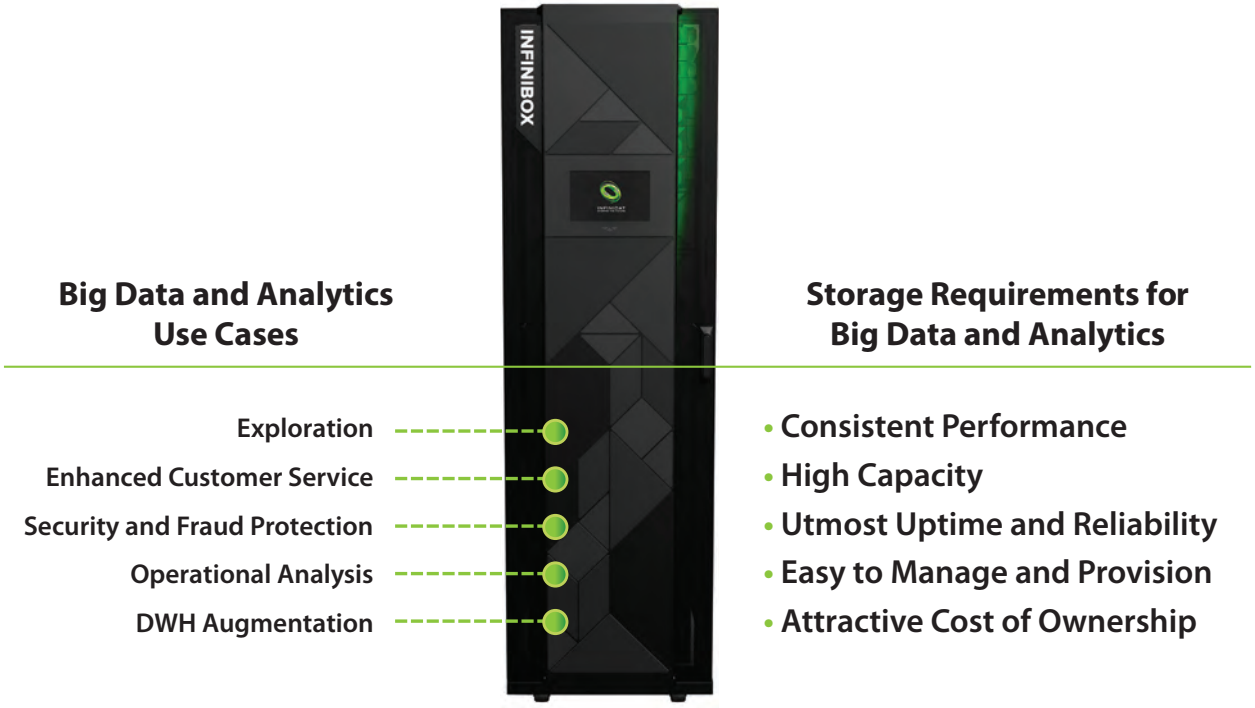
This is where a brand-new storage architecture such as INFINIDAT's InfiniBox™ can help.

High Performance

Large data sets, heavyweight analytics applications, and time-sensitive demands for results make high performance of the underlying storage a key criteria. In InfiniBox, maximum performance is achieved with no tuning or optimization. InfiniBox uses standard "off the shelf" hardware components (CPU/Memory/HDDs/SSDs) wrapped in sophisticated storage to extract the maximum performance from the 480 Near-Line SAS drives used in the InfiniBox architecture. One of the key elements developed in the core system code is the ability to analyze real application profiles and surgically define cache pre-fetch and de-stage algorithms. The system design specifically targets real-life profiles and provides optimum performance under those conditions. This capability is at the core of the InfiniBox architecture.

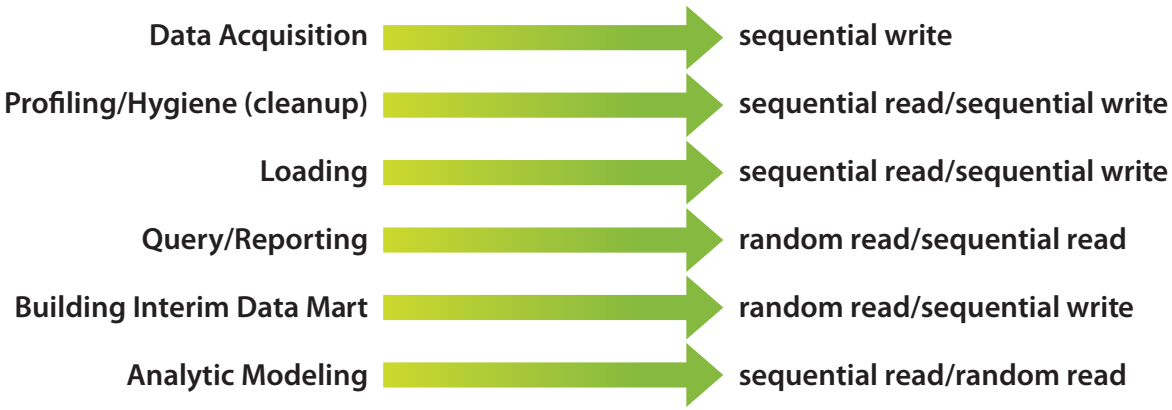
Large Data Sets

Large data sets pose a unique and daunting challenge to enterprise storage arrays by providing an I/O profile that is unpredictable, and often overwhelms the storage frame. This results in high latencies, which increase the run time of analytic workloads. Some analytic activities are very latency sensitive, and in many cases will affect the end-user population that the application supports. Many of these workloads will overwhelm storage platforms with limited cache sizes, but not the InfiniBox. InfiniBox uses high-speed L1 DIMM cache and L2 SSD cache to improve cache hits and reduce latency.



Diverse I/O Profiles and Patterns Used in Big Data

We have seen many analytic environments exhibit I/O profiles containing all of the following characteristics:



Many of these characteristics can be seen occurring simultaneously, while others are driven by specific activities such as backups or data load/ETL. InfiniBox thrives on supporting a wide range of I/O types, all at the same time. The data architecture virtualizes the storage for each volume by populating each of the 480 spindles in the InfiniBox frame, all in parallel, using a sophisticated data parity dispersed layout.

In addition, InfiniBox utilizes very advanced capabilities to improve writes. Using a unique and patented multi-modal log write mechanism, INFINIDAT significantly improves the effciency of write I/O de-staged from cache. This is very important for the Data Acquisition and ETL phases of this example.

Block Size Management Matters

Many analytic workloads can change the I/O profile on the fly. But in general, the vast majority of Big Data and Analytic applications use large block I/O, loading data in from storage, reducing, sorting, comparing, then writing out aggregate data. Large blocks can historically give traditional storage platforms problems because most storage environments are not designed with large-block support in mind.

High Density

INFINIDAT has the ability to configure a system with multiple petabytes of effective capacity in a single 19-inch rack. The InfiniBox storage system is a modern, fully symmetric grid, all-active controller (node) system with an advanced multi-layer caching architecture. The data architecture encompasses a double parity (wide-stripe) data distribution model. This model uses a unique combination of random data distribution and parity protection. This ensures maximum data availability while minimizing data footprint. Each and every volume created on a single InfiniBox frame will store small pieces of data on each of the 480 drives in the frame. InfiniBox usable storage per frame is the highest in the storage industry.

High Availability and Reliability

Keeping analytics available is critical for every storage system. The InfiniBox architecture provides a robust, highly available storage environment, providing 99.99999% uptime. That equates to less than 3 seconds of downtime a year! Our customers report no loss of data, even upon multiple disk failures. InfiniBox offers end-to-end business continuity features, including asynchronous remote mirroring and snapshots. Using snapshots, recovery of a database can be reduced to the amount of time it takes to map the volumes to hosts, minutes instead of hours using a more traditional backup and recovery process.

Easy to Use, Automated Provisioning and Management

The InfiniBox architecture, along with the elegant simplicity of its web-based GUI and built-in Command Line Interface, allows for easy, fast deployment and management of the storage system. The amount of time saved in performing traditional storage administration tasks is huge. Also, because of the InfiniBox open architecture and aggressive support for RESTful API, platforms such as OpenStack, and Docker allow storage administration tasks to be performed at the application level, without the need to use the InfiniBox GUI.

InfiniBox provides a management system that can isolate storage pools and volumes to specific users. Multi-tenancy features are supported so that application users, such as those deployed in a private cloud environment, can see and manage the storage that has been granted to that user community.

Very Low Total Cost of Ownership

High performance, extreme availability, highest data density and ease of use all point to an unmatched TCO. This is important for environments where there is a need to consolidate mission-critical databases into smaller and smaller physical footprints.

Hadoop and InfiniBox

One of the many Big Data use cases is support for Hadoop clusters. Today, the default deployment model for Hadoop clusters is to build a cluster from many, inexpensive nodes. Each node has equal amounts of compute, memory and dedicated storage. The initial design phase of most customers launching their first Hadoop environment is to find the right mixture of resources to come up with this node-level design. Then they string a number of these nodes together to create a large compute environment for their favorite MapReduce application. The problem most customers run into is that they run out of storage long before they run short of compute cycles in the cluster. The only option available for customers using dedicated storage is to keep adding more nodes to the cluster. This is fine, except now you are adding more compute power as well, which may not be needed, and in many cases is where this dedicated storage model breaks.

By adopting InfiniBox block-based SAN storage instead of dedicated hard drives per node, you are no longer limited by how much storage each node is capable of providing. You can dynamically add more LUNs, or add more space per LUN for each cluster node. There is no longer a need to add compute power until needed. The level of Hadoop data redundancy can be reduced, as the data will be fully protected with 99.99999% uptime.

Conclusions

InfiniBox bridges the gap between high performance and high capacity for Big Data applications. InfiniBox allows an organization implementing Big Data and Analytics projects to truly attain its business goals: cost reduction, continual and deep capacity scaling, and simple and effective management — and without any compromises in performance or reliability. All of this to effectively and effciently support Big Data applications at a disruptive price point.