WHITEPAPER / MAY 3, 2019 / QL573950

Comparing IBM Integrated Analytics System and Teradata IntelliFlex

Sponsored by: IBM

Market Situation

In this era of digital disruptions, successful organizations must be adaptive and agile or risk losing their competitive advantage. The rapid pace of technological advances has fundamentally changed the way users interact with technology. Data is a critical commodity for most organizations, providing key business insights that can drive transformation and innovation.

Driven by the Internet of Things (IoT), new devices and services collect data at an ever increasing rate. Proper use of advanced analytics technologies to harness this abundance of data is of great value to an organization, providing tools that may lead to greater success. Big data analytics systems utilizing tools such as distributed computing and machine learning can help companies increase operating efficiency and serve customers more effectively. Cloud computing, mobile devices and sensors, as well as social media platforms have helped even the competitive landscape for emerging firms, enabling innovative small and midsize businesses to more effectively compete with larger organizations.

The emergence of cloud services ranging from software- to infrastructure-as-a-service from vendors such as Amazon, Google, IBM, Microsoft, and Oracle have become an important equalizer for smaller companies that cannot support the capital cost of building their own IT infrastructure. Larger organizations with agile endeavors are also considering using cloud services for flexibility. Organizations can choose to implement standalone cloud solutions or integrate with existing on-premises infrastructure. Similarly, for most companies, on-premise solutions that are cloud-ready or open source compatible are more appealing than those which lock-in a customer to a specific vendor.

Whether on-premise or cloud-based, analytics are increasingly deployed to provide near real-time insights for decision makers in all departments, whether marketing, operations, or communication, thus enabling companies to create value from data faster than ever before. However, as organizations become more dependent on the intelligence delivered by these systems, the demands on performance, security, and reliability increase as well.

Mobile applications, another business critical technology, enable companies to rapidly communicate with customers, facilitate transactions, and provide services. Mobile transactions account for over a third of all retail ecommerce in 2018, and are projected to grow steadily over the next few years. Other industries, including banking and healthcare, as well as government services are also participating in, and contributing to, mobile ecosystems. These ecosystems, in conjunction with social media platforms, are significant channels that can be used for collecting customer data and deriving insights.

In addition, artificial intelligence (AI) and cognitive computing are increasingly being adopted by businesses to harness the benefits of technologies, such machine learning, natural language processing, and virtual agents. As more and more businesses rely on the intelligence generated from cognitive computing via the cloud, on-premises, or hybrid environments, use of a flexible, stable, and secure platform to access these innovative technologies becomes important than ever. Solutions that do not support next-generation technologies will not be able to meet the needs of modern organizations.

IoT, another vehicle through which companies collect data, provides unprecedented automation and access to real-time information. From smart home assistants to traffic optimization sensors, the availability of IoT data, for example, can increase the efficiency of a single-home or a metropolitan area-based monitoring and automation system.

As organizations leverage data for additional insights, traditional data warehouses may no longer meet the demands for high performance computing involving large data sets. These organizations are embracing open source distributed computing and machine learning platforms such as Spark, Hadoop, and TensorFlow. Leading cloud vendors, such as Amazon, IBM, and Microsoft, have taken note of these customer needs and integrated or bundled big data frameworks with their cloud offerings.

Enterprise solution vendors offering data warehouse solutions must adapt to the latest data science tools or risk losing their customer base, especially those that heavily focus on using data science to achieve strategic goals. Industry leaders such as IBM, Microsoft, Oracle, SAP, and Teradata are increasingly touting support and integration with data science platforms to appeal to customers. These platforms attempt to solve traditional problem of balancing cost, performance, capacity, and reliability while offering modern advanced capabilities to customers such as machine learning, cloud enablement, and open source technologies.

This paper compares two of these high performance data warehousing platforms, IBM Integrated Analytics System (IIAS) and Teradata IntelliFlex. The latest evolution of the Netezza and PureData systems, IIAS is built on IBM's Common SQL Engine, offering extreme workload portability across multiple data sources. IIAS also offers high performance build-in machine learning capabilities and a single point of management via IBM Data Server Manager.

As part of the Hybrid Data Management platform, IIAS provides elasticity and flexibility, allowing organizations to scale workloads on-premises and/or migrate to and from the cloud with ease. The built in Db2 engine offers advanced database functionalities such as hybrid columnar and row processing and massively parallel processing (MPP).

In comparison, IntelliFlex is a high-performance, scalable data warehouse solution with high capacity. IntelliFlex leverages Teradata Virtual Storage to dynamically move data into and from memory for high-performance analytics. IntelliFlex parallel processing takes advantage of Intel's parallel processing instruction sets to increase CPU throughput.

Key technology differentiators between the two platforms include support for open source data science, system scalability, deployment flexibility, and infrastructure enhancements to increase performance for analytics workloads.

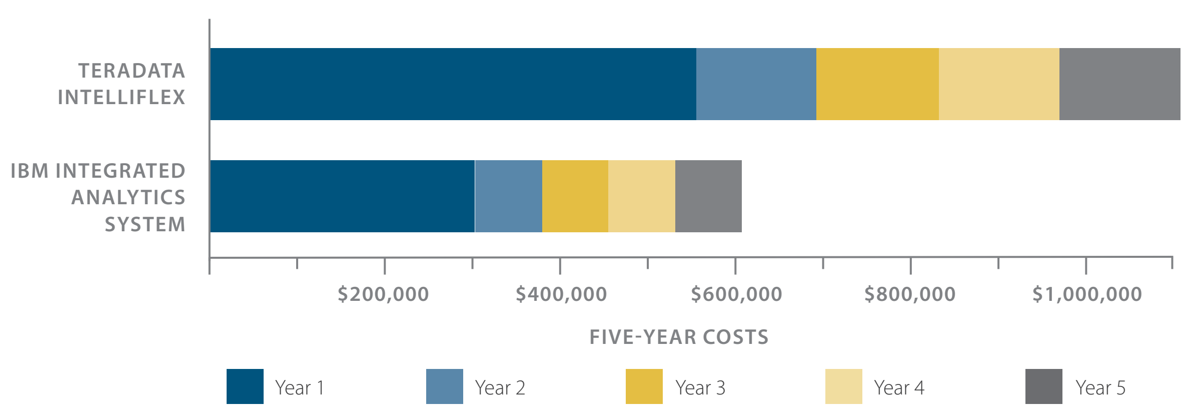
Due to smaller data center footprints, ease of management, and system elasticity, five year costs of ownership for IIAS average 45 percent less than Teradata IntelliFlex configurations (Figure 1). Cost estimates are based on data warehouse configurations of similar sizes with industry standard maintenance and support costs.

Next Generation Analytics

Any organization that aims to modernize its data analytics or begin a foray into data science and AI will consider performance a key factor in purchase decisions. Legacy data warehousing systems that cannot process modern workloads efficiently will become increasingly obsolete as business requirements grow. Data intake for organizations will only expand in volume, velocity, and variety over time, and businesses should only consider solutions that can support these expanding workloads.

Challenges faced by organizations with traditional data warehouses primarily include scalability and performance. When organizations purchase proprietary solutions such as large scale-up data warehousing appliances, they make a substantial capital investment based on limited knowledge of what their future needs will be. Workloads often exceed the capabilities of current solutions and require additional capital outlay. An easily scalable solution allows companies to purchase computing power and storage more granularly, and more cost effectively as requirements evolve.

FIGURE 1: Five Year Total Cost of Ownership for IBM Integrated Analytics System and Comparable Teradata IntelliFlex Configurations



SOURCE: Quark + Lepton (April 2019)

In the current data analytics landscape, many organizations still rely on data warehouses composed of disparate hardware and databases tightly coupled together. While this integration aggregates from multiple data sources, it tends to be more cumbersome and costly than loosely coupled systems.

Teradata IntelliFlex is an example of this type of traditional data warehouse. With its complex architecture, high performance compute nodes paired with ample storage, IntelliFlex systems are typically costly to purchase and costly to maintain. Contributing to costs are the personnel requirements necessary to maintain the interoperability of highly, and oftentimes manually, configurable systems. Additionally, one of the major performance limiting factors of big data workloads involves moving large amounts of data from where it is generated to where it will be processed. These large I/O operations tend to introduce latency and inefficiency.

Compared to IntelliFlex, IIAS is a next generation analytics platform that is not hindered by such drawbacks. IIAS's strengths in big data processing stem from its built in data science functionality, massively parallel processing hardware architecture, innovative database capabilities, and next-generation all flash storage. These strengths complement the distributed computing enhancements brought by integration with the Spark framework as well as compatibility with the Hadoop platform. Additionally, ease of management as well as the smaller infrastructure footprint of IIAS contribute to lower ownership and personnel costs.

Integrated Big Data and Machine Learning

Data asset growth across all industries and organizations is primarily driven by the potential revenue generated by insights gleaned from data. However, only when data is processed and analyzed efficiently and effectively will it produce value for an organization. For companies that are not harnessing the power of their data, increasing data footprints will only increase costs.

Organizations collecting large amounts of data at high velocity can be overwhelmed by the volume and variety of that data. Cleaning, processing, and analyzing large, diverse data sets require solutions capable of running high performance workloads. Often times, this results in a need for distributed, parallel processing. Distributed computing requires nodes to be tightly integrated to efficiently coordinate workloads and reduce latency in order to rapidly deliver user requests.

These business challenges have propelled open source data science and distributed computing platforms such as Spark and Hadoop to the forefront. These platforms are available through the Apache Software Foundation or through vendors such as Cloudera and MapR. Table 1 provides an overview of these platforms and the technologies they offer. Figure 2 depicts the architecture of distributed workloads on Spark and Hadoop.

The essential strategy behind these big data platforms consists of deconstructing large volumes of data and processing these smaller batches independently and in parallel on a cluster of compute nodes. These nodes return compute results that may be joined and/or aggregated to provide the final result. This strategy is also deployed in MPP data warehousing environments where a cluster of high performance computing nodes process data in parallel.

The similarities between the technology behind open source big data platforms and MPP data warehousing result in synergistic environments where organizations can leverage the benefits of each simultaneously. As such, traditional enterprise technology vendors without an open source compatible big data offering in their portfolio will lose customers as more and more organizations shift towards these data science solutions.

Although similar, open source technologies such as Hadoop and Spark and traditional MPP data warehouses can meet different needs. While MPP data warehouses are queried using SQL, Hadoop implementations must be controlled using MapReduce jobs written in Java. For standard business intelligence extraction, querying data warehouses may be more efficient. Large volume workloads with data source and type variety may be more effectively processed using Hadoop. Moderate volume workloads that prioritize processing speed for near real-time insight generation should be processed using Spark.

However, organizations do not have to choose and lock into any of these platforms. More than ever, the big data solutions landscape is witnessing a merging of MPP data warehouses with data science technologies such as Hadoop and Spark. These hybrid data solutions maximize flexibility for organizations with diverse data workloads while providing high-end hardware to support these workloads.

TABLE 1: Supported Libraries and Modules for Hadoop and Spark

PLATFORM	SUPPORTED LIBRARIES & MODULES
HADOOP **Implementations generally consist of:** **• Hadoop Distributed File System (HDFS)** **• Yet Another Resource Negotiator (YARN) application runtime scheduler & resource manager** **• MapReduce parallel processing engine runs on disk** **Hadoop only runs on disk & is extremely powerful for processing large amounts of data**	• Ambari, Avro, Cassandra, HBase, Oozie, Flume, Sqoop • Mahout for machine learning • Integrates through REST API • Supported languages: Java, PHP, Hive, Pig, Python, Ruby, R
SPARK **Newer parallel processing technology that runs in memory powered by the Spark Core engine** **Does not have its own distributed file system & often uses HDFS** **Runs in memory or on disk, enabling processing of both batch & streaming workloads**	• MLLib for machine learning, GraphX, Spark SQL, Spark Streaming • Integrates with Hadoop, Cassandra, HIVE, HBase, PostgreSQL, JSON, MySQL, Lucene • Supported languages: Scala, Java, Python, R

SOURCE: Quark + Lepton (April 2019)

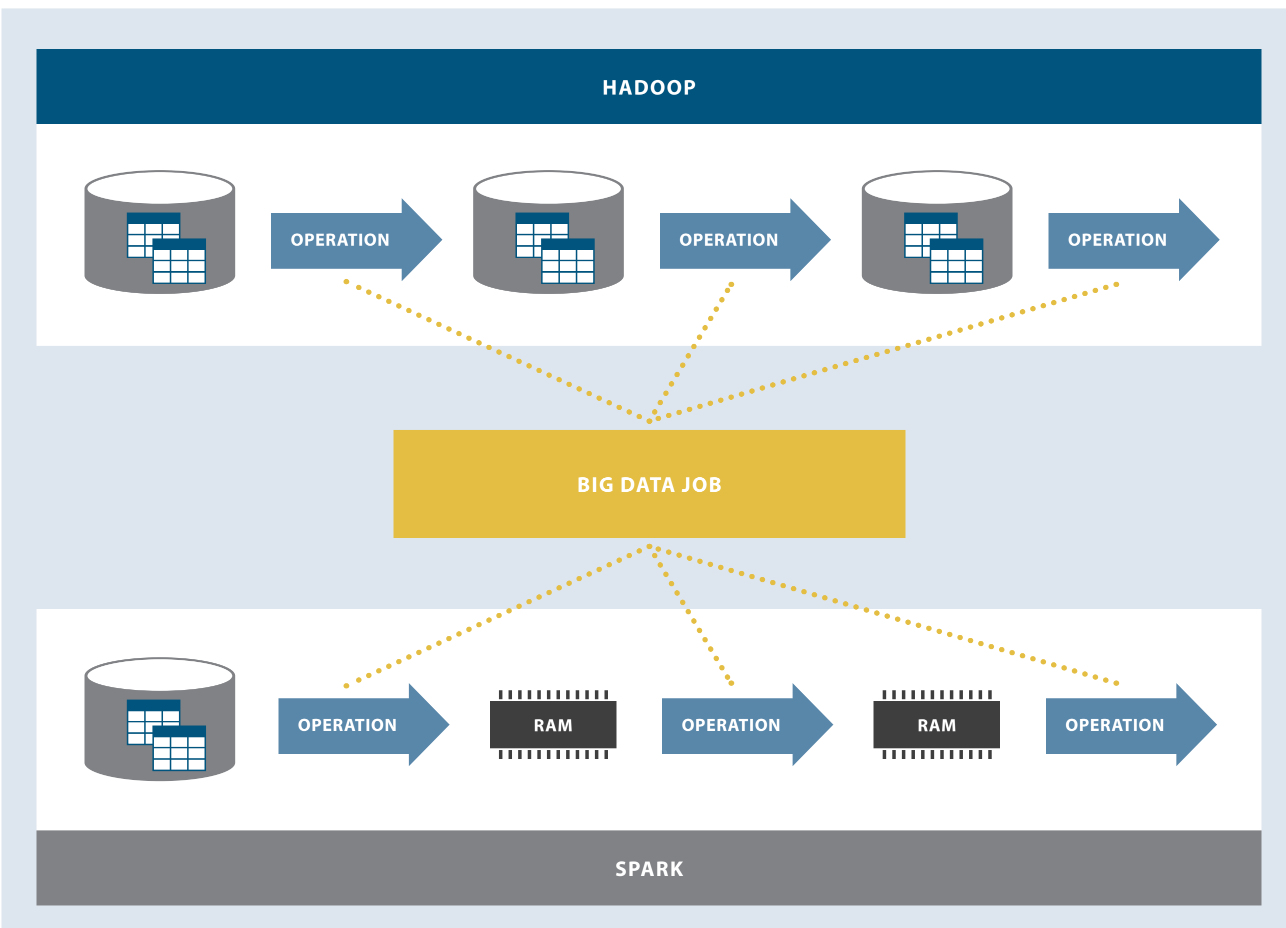
DATABASE AND MASSIVELY PARALLEL PROCESSING INNOVATION

In building the IIAS platform, IBM has integrated the strengths of its Netezza data warehousing platform with recent innovations in IBM's database product, Db2, to create a high performance analytics system capable of leveraging MPP hardware with minimal I/O overhead.

This is accomplished through use of Db2's MPP engine and query pushdown optimizations. Pushdown analysis first checks to see if operations or functions can be executed locally where pertinent data resides. If possible, the query is functionally deconstructed and processed at the data sources, only returning the results of operations and eliminating latency caused by moving data to a centralized compute location. This type of parallel workload distribution increases the efficiency of compute resource utilization and can improve performance significantly.

IBM's Data Science Experience (DSX) is also included in out-of-the-box on-premises IIAS environments. This offering provides a set of data science tools as well as integration with R Studio, Spark, and Jupyter Notebook. Embedded Spark in IIAS also means users can natively use Spark libraries such as Spark Streaming, MLib, and GraphX for their data science needs. Data scientists can not only capitalize on

FIGURE 2: Hadoop MapReduce Processing Compared to Spark In Memory Processing

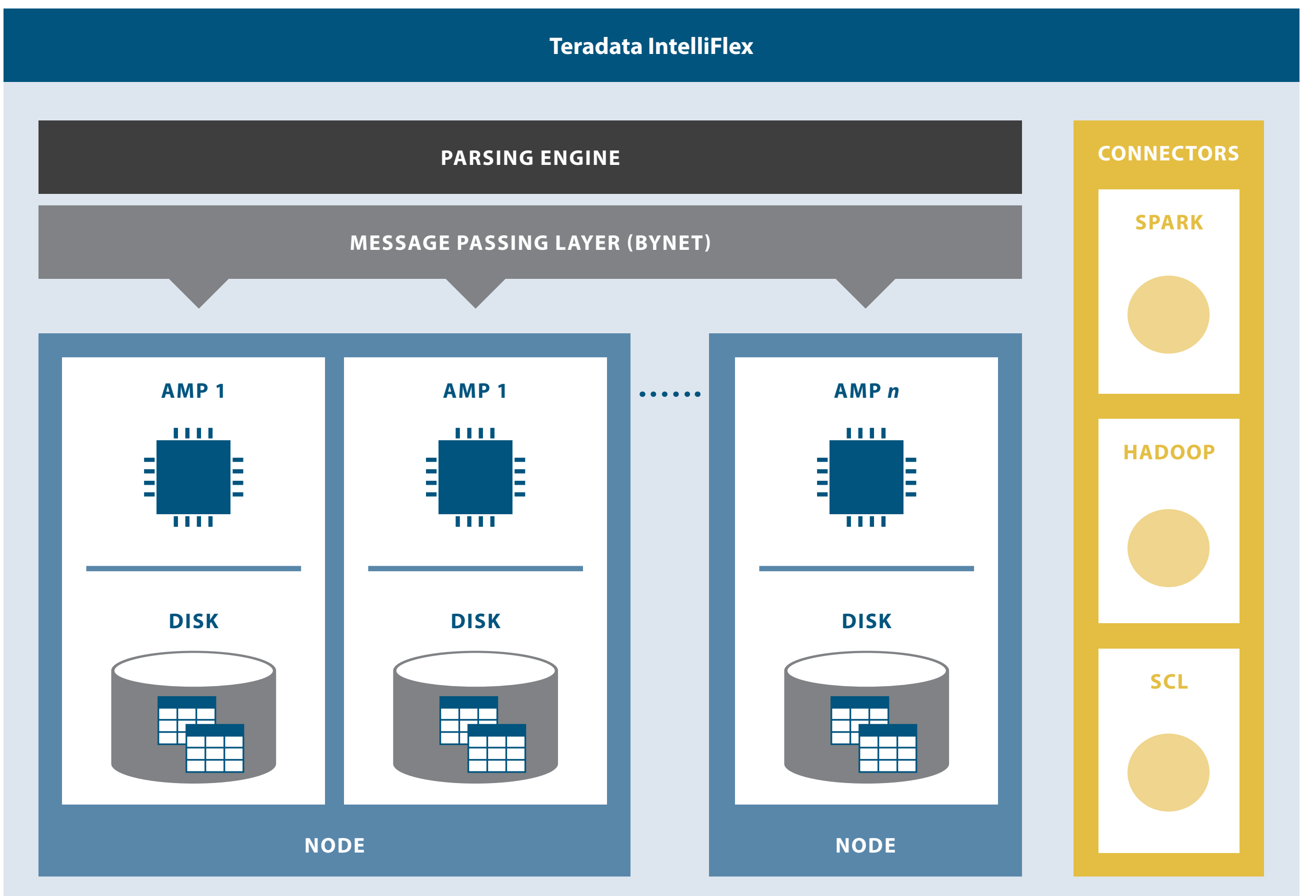


Spark performance enhancements, they can access and write to Db2 warehouses in R and dynamically document and share findings in Jupyter.

In contrast, IntelliFlex lacks the native support for big data analytics and machine learning. IntelliFlex systems tend to require migration of data to various engines, such as Hadoop or Spark, via connectors, which adds additional overhead to large queries and data science jobs.

Teradata's architecture consists of compute nodes that contain independent resources such as CPUs, memory, and a copy of the relational database management software. Initiated queries are first processed by a parsing engine (PE), which dissolves the query into processing operations and checks if the query can be executed. The PE also creates a query execution plan and sends it to the Message Passing Layer, also known as the BYNET. BYNET is the networking layer that coordinates communications between the PE and nodes as well as between nodes. Access module processors (AMPs), virtual processors in each node, receive data and query execution directions from the PE via BYNET and execute the directions based on the data on the disks controlled by each AMP. Each Teradata node typically consists of multiple AMPs and data sets that correspond to each AMP (Figure 3).

FIGURE 3: Architecture of Teradata IntelliFlex



SOURCE: Quark + Lepton (April 2019)

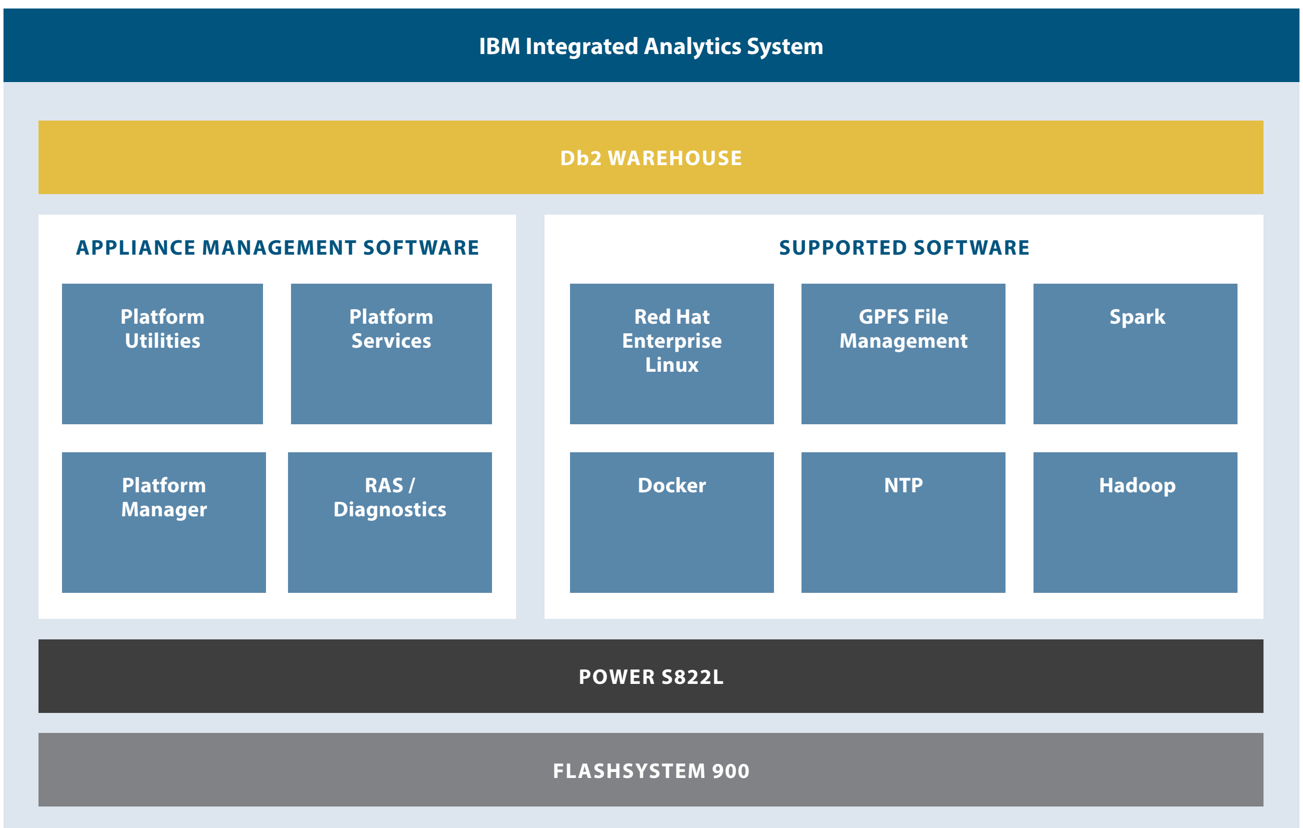
The IIAS platform, enabled by IBM's Common SQL Engine (CSE), allows organizations to easily aggregate all of their data from Db2 as well as non-IBM and open source databases for big data processing and analysis. IBM Data Server Manager simplifies data management by providing a centralized tool for administration, alerting, monitoring, federation, and SQL execution support (Figure 4).

Data management in Teradata tends to be highly manual and time consuming. Administrators are encouraged to migrate data to Teradata or use bloated connectors that can significantly limit performance.

Additionally, Db2's MPP enhancements are tightly integrated and enable high performance analytics processing across clustered implementations. Db2's MPP functionality greatly increases the scalability of workloads by harnessing the power of all compute and storage nodes and by supporting horizontal scale out of both computing power and storage capacity.

Although it is possible to run Db2 on a single machine with hundreds of CPU cores and terabytes of RAM, By leveraging the resources provided by multiple servers for a multi-partition environment, such as increased total buffer pool space, IIAS can improve the efficiency of the entire system. This flexibility

FIGURE 4: Architecture of IBM Integrated Analytics System



SOURCE: Quark + Lepton (April 2019)

allows organizations to self determine how they want to grow their system in the most sustainable way. Table 2 lists the various configurations and specifications for IIAS systems.

Integration with Db2 BLU Acceleration also brings other database performance enhancements to IIAS offerings, such as in-memory columnar processing, data skipping, and Single Instruction Multiple Data (SIMD) processing. Table 3 provides an overview of these features.

In-memory, column-based databases have become a standard solution for analytics workloads, drastically improving query processing performance compared to traditional row-based databases. Db2 columnar tables can coexist with traditional row-based data and improve analytical performance, without the need for database administrators (DBAs) to create additional indexes, partitioned tables, or aggregates. Db2 columnar technology also dynamically optimizes performance, minimizing latency associated with storage access by intelligently and automatically streamlining memory usage through data skipping and actionable compression techniques.

Db2's column-organized tables can scale beyond the limitations of RAM because Db2 can dynamically loads data required by queries into memory. Data skipping in Db2 improves query performance, and reduces memory requirements, by automatically detecting and skipping sections of data that is not

TABLE 2: IIAS Configuration Comparison

MODEL	Tier 1 (Flash) *(User data-approx 4x compression)*	Tier 2 (Disk) *(User data-approx 3x compression)*	PROCESSOR CORES	NODES	PHYSICAL RACKS	HIGH AVAILABILITY
M4001-003	64 TB	—	72	3	1	Yes
M4001-006	128 TB	—	120	5	1	Yes
M4001-010	192 TB	—	168	7	1	Yes
M4002-003	108 TB	—	72	3	1	Yes
M4002-006	216 TB	—	120	5	1	Yes
M4002-010	324 TB	—	168	7	1	Yes
M4002-020	648 TB	—	336	13 + 1 standby	2	Yes
M4002-040	1,296 TB	—	672	25 + 3 standby	4	Yes
M4002-003S	236 TB	399 TB	72	3	2	Yes
M4002-006S	344 TB	720 TB	120	5	2	Yes
M4002-010S	452 TB	1,041 TB	168	7	2	Yes
M4002-020S	904 TB	2,004 TB	336	13 + 1 standby	4	Yes
M4002-040S	1,808 TB	3,920 TB	672	25 + 3 standby	8	Yes

SOURCE: Quark + Lepton (April 2019)

relevant for queries. Actionable compression in Db2 further maximizes the capacity of memory and storage resources , and speed up query processing without decompressing the data. Figure 5 depicts how various Db2 data reduction capabilities synergize to improve query performance and reduce memory requirements.

DEPLOYMENT OPTIONS AND SYSTEM MANAGEMENT

The high performance and ease of management of IIAS result from the Power Systems infrastructure that resides in each rack. Hardware architecture, derived from the Netezza platform, was designed with simplicity and flexibility in mind. IBM's design philosophy aims to enable organizations to spend most of their resources on continuous improvement and not routine system maintenance and management.

TABLE 3: Performance Optimization Features of Db2 BLU

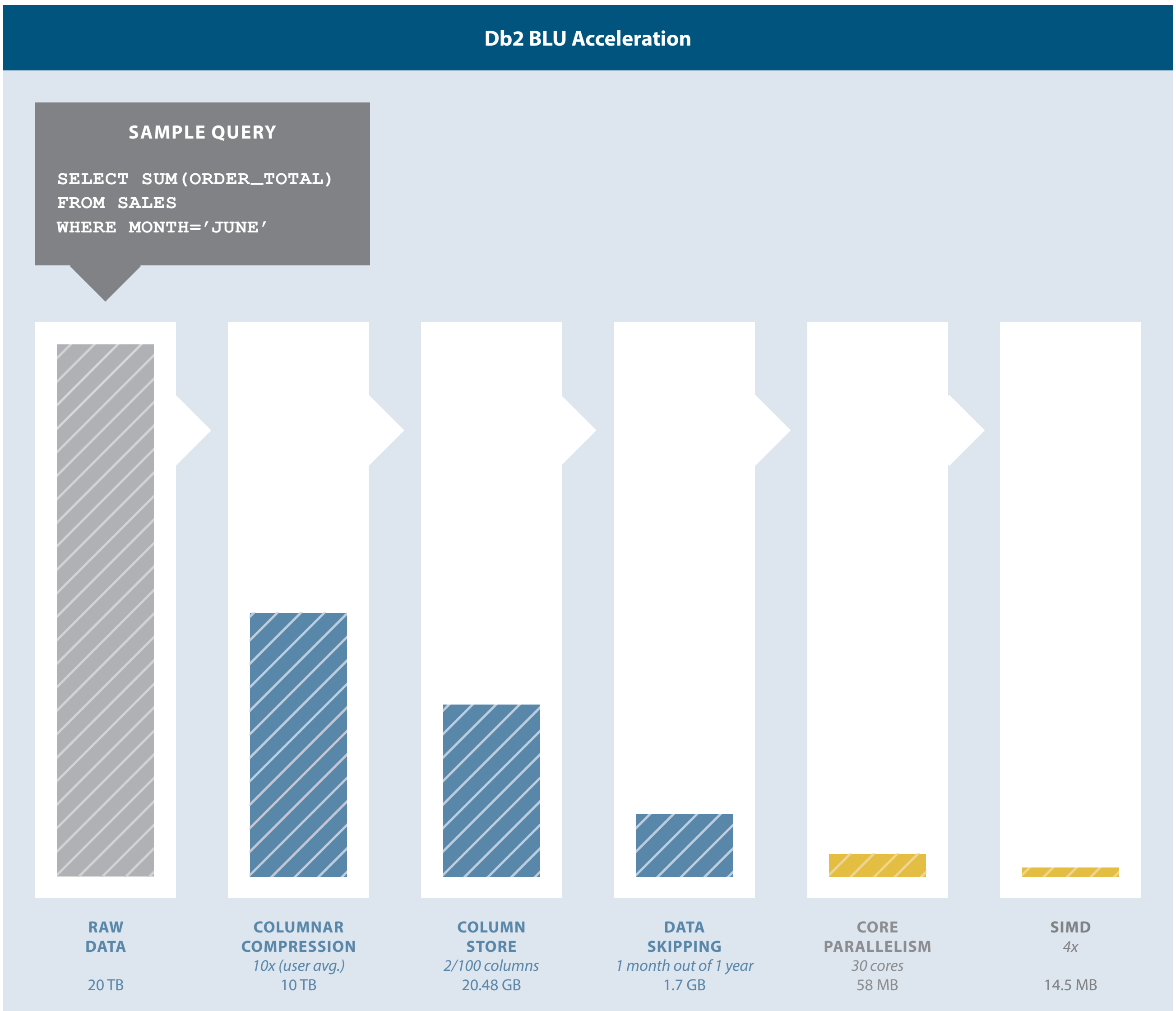
PERFORMANCE OPTIMIZATION FEATURE	DESCRIPTION
In-Memory Columnar Technology	Column-organized tables can be created using BLU Acceleration. Db2 BLU is in-memory optimized, & is not limited by RAM limitations such as those imposed by other in-memory technologies, including Oracle's. BLU dynamically loads data into memory & moves colder data to storage, optimizing use of memory resources based on query workload.
Actionable Compression	Enables automatic storage saving without the need for DBA intervention. Also, improve query performance through use a growing library of sort, join, & other operations that can be applied to encoded columns without needing to decompress the data.
Data Skipping	Contributes to faster, actionable compression analytics by detecting which encoded data are not required for processing a query, & ignoring those pages. Results in lower I/O & CPU requirements.
Massively Parallel Processing (MPP)	Compatibility with Database Partition Feature (DPF) environments enables Db2 BLU to scale near-linearly with hardware. Parallel processing across partitions improves query performance by leveraging horizontal scaling of resources.
Single Instruction Multiple Data (SIMD)	Hardware instructions are leveraged by BLU to apply a single instruction (e.g., predicate evaluation, join, grouping, arithmetic, R script) simultaneously to multiple data elements.
Continuous Data Ingest (CDI)	Allows users to move large amounts of data into tables without requiring a repeated unlocking & locking of the tables during the process. The ingest utility can continuously process data output from extract, transform, load (ETL) tools to populate large databases stored in partitioned environments. It uses row-level locking, populating tables without affecting other user activities conducted on the same tables.

SOURCE: Quark + Lepton (April 2019)

The compute nodes for each IIAS system consists of the Power S822L Server, each consisting of 12-core Power8 CPUs with 512 GB of RAM. The primary storage nodes of each IIAS rack are FlashSystem 900 all-flash storage arrays. A second tier of Storwize V5000 disk storage supplements the primary FlashSystem storage.

Within each new generation of Power Systems, IBM expertise in processor manufacturing, supercomputer and mainframe designs, and operating system and compiler optimization has focused on providing broad support for mixed workloads with industry-leading, high-performance computing speeds and throughput.

FIGURE 5: Data Reduction Capabilities Featured in Db2 BLU Acceleration for a Sample Query



SOURCE: Quark + Lepton (April 2019)

The POWER8 processor has been designed to handle a variety of high-performance computing workloads such as AI, in-memory processing, and big data analytics. POWER8 processors, for example, support larger numbers of threads per core (by a factor of four), greater memory bandwidth, a higher clock speed, and more cache compared to the Xeon 6140 processors used in IntelliFlex (Table 4). In the POWER8 design, L1 through L3 cache are on-chip, while up to 128 megabytes (MB) of L4 cache is added off-chip. Intel offers no equivalent to POWER8's L4 cache.

In a mixed workload environment, the amount of work a server can perform over time depends not only on processor, memory, and I/O throughput, but also on the mechanisms that allocate and reallocate system resources as demands change. For analytics workloads specifically, the increased number of POWER8 cores supported by higher I/O bandwidths provide industry-leading performance. POWER8-based Power Systems servers can deliver I/O bandwidth of up to 192 gigabytes per second (GB/s). Additional POWER8 features that reinforce efficient resource usage include the following:

- *Intelligent Threading* allows workloads to be executed using from one to eight threads per core. The system can automatically determine how many threads per core to use for optimum performance, or system administrators may select the number of threads to be employed.
- *Intelligent Cache* allows systems to dynamically vary cache utilization to minimize overall cache latency as workload characteristics change. The system may automatically determine the appropriate level of cache for specific workloads.

TABLE 4: Comparison of CPUs used in IIAS and Teradata IntelliFlex

FEATURES	IBM POWER8	INTEL XEON GOLD 6140
Cores/chip	12	18
Max threads/core	8	2
Max clock rate	4.2 GHz	3.7 GHz
Max main memory	1 TB	768 GB
Max L-1 cache	64 KB	1.125 MB
Max L-2 cache	512 KB	18 MB
Max L-3 cache	96 MB	24.75 MB
Max L-4 cache	128 MB	N/A
Max memory bandwidth	230 GB/s	119 GB/s
Max sockets	16	8
Process technology	22 nm	14 nm

SOURCE: Quark + Lepton (April 2019)

For high performance storage, IIAS systems utilize IBM FlashSystem 900 arrays. IBM introduced the FlashSystem 840 with new reliability, availability, and serviceable (RAS) features in 2014. This was the first release of reengineered technology IBM obtained in its acquisition of Texas Memory Systems (TMS) in 2012. The IBM FlashSystem 900, introduced in February 2015, is designed for Tier 0 high-end performance storage. The FlashSystem 900 can be configured to deliver RAID5 redundant disk array striping with up to 57 terabytes (TB) per array, ensuring its proven ability to deliver 99.999 percent availability.

Key differentiators for the IBM FlashSystem family are the ability to harness the high speed and performance of IBM MicroLatency Modules and its custom hardware controller architecture. Using IBM FlashCore, a proprietary, hardware-only controller, the FlashSystem family reduces the overhead generated by software data management produced by other types of flash storage. Although IntelliFlex contains all-flash storage, it uses commodity storage and third-party disk arrays.

For growing organizations with increasing storage or computing requirements, IIAS systems are horizontally elastic and additional nodes can be integrated with ease. Due to the Common SQL Engine, IBM Hybrid Data Management offerings such as Db2 Warehouse on IBM cloud or on-premises, IIAS, and Big SQL via Hadoop are interoperable and can federate with various other popular non-IBM database solutions.

Organizations also are not required, and sometimes not advised, to migrate data to IIAS. Through consultation with clients, custom solutions, on-premises or hosted, are proposed and delivered instead of encouraging all customers to migrate. This eliminates risks and costs associated with large scale migrations unless customers are ready to take the plunge.

In contrast, Teradata offerings are in the form of proprietary appliances with limited interoperability with the latest data science and big data platforms. Hosted solutions are on third-party cloud infrastructure such as Amazon Web Services (AWS) and Microsoft Azure. This complicates support and recovery procedures when issues arise.

Conclusions

Current technological capabilities have enabled companies of all sizes to take advantage of data in unprecedented ways. Engaging in data science, in the form of very costly data warehouse appliances, is no longer something only the largest organizations can afford. The latest and most advanced tools for big data are open source and encourage interoperability with other open source technologies, resulting in freedom for users to pick and choose the software they want to use. This philosophy markedly deviates from the philosophies of traditional data warehouse vendors, who tended to favor strategies that locked customers into a platform for as long as possible.

This shift in paradigm has driven significant changes in offerings from enterprise analytics vendors. Solutions that do not interoperate or integrate with the latest open source platforms in data science will be left in the dust by the new generation of software engineers and data scientists who have embraced technologies such as Hadoop and Spark and open source ecosystems in general.

The IBM Integrated Analytics System and IntelliFlex are both data warehouse appliances with roots in legacy data warehousing. However, IIAS was designed with modern users in mind, promoting simplicity, scalability, and open source and cloud readiness. Built-in Spark and Common SQL Engine means data scientists can consolidate data rapidly and generate insights in near real time on IIAS or any of IBM's data offerings, on-premises, cloud, or in a hybrid environment.

In contrast, Teradata tends to recommend, or even require, large data migrations to its IntelliFlex platform, increasing costs, risks, and deployment time for customers. Customers with disparate data sources may also need to use bulky connectors to federate their data, potentially significantly increasing latency for analytics workloads. Teradata also lacks native cloud offerings, and partners with third-party cloud vendors such as Amazon and Microsoft, further increasing platform complexity.

Index

LIST OF FIGURES

LIST OF TABLES

LIST OF REFERENCES

U.S. mobile retail commerce sales share 2017-2021 may be found at www.statista.com/statistics/249863/us-mobile-retail-commerce-sales-as-percentage-of-e-commerce-sales/

Intel Xeon 6140 specifications may be found at ark.intel.com/compare/120485

Additional Intel Xeon 6140 specifications may be found at en.wikichip.org/wiki/intel/xeon_gold/6140

Quark + Lepton®

CORPORATE OFFICE
Boulder, Colorado USA

www.quarkandlepton.com
info@quarkandlepton.com

IBM sponsored this publication, however, the information and conclusions contained in this publication do not necessarily represent the positions of IBM or other referenced sources.