White Paper

Delivering Hybrid Analytics at the Speed of Business: IBM Integrated Analytics System

Sponsored by: IBM

Carl W. Olofson
October 2018

IN THIS WHITE PAPER

This white paper considers the pressures that enterprises face as the volume, variety, and velocity of relevant data mount and the time to insight seems unacceptably long. Most IT environments seeking to leverage statistical data in a useful way for analysis that can power decision making must glean that data from many sources, put it together in a relational database that requires special configuration and tuning, and only then make it available for data scientists to build models that are useful for business analysts. The complexity of all this is further compounded by the need to collect and analyze data that may reside in a classic datacenter on the premises as well as in private and public cloud systems. This need demands that the configuration support a hybrid cloud environment. After describing these issues, we consider the usefulness of a purpose-built database system that can accelerate access to and management of relevant data and is designed to deliver high performance for the kind of queries that can lead to powerful analyses and rapid decision making. We also look at how the IBM Integrated Analytics System (IIAS) delivers on these hybrid capabilities, whether on-premises, in the public cloud, or in the private cloud.

SITUATION OVERVIEW

Enterprises are faced with new challenges as they confront the move to embrace data from a wide range of often unmanaged sources that may represent the operational state of the enterprise or may be external data relevant to marketing or operational needs. They need to bring together the data, transform the data into useful structures, and then use the data to drive insights and operations through both human- and machine-based decisions, often driven by machine learning (ML). The result is a complete change in business management through what is called digital transformation. In addition, the pressing requirement for on-demand scalable resources is driving many enterprises to the cloud.

Technical Challenges of Data Collection, Curation, and Integration

As enterprises consider how they may address the challenges of digital transformation, they must seek technologies capable of addressing a stunning range and variety of applications, use cases, and business needs. These technologies must have the ability to collect data from a wide range of sources, formats, and types, including external streaming data, Internet of Things (IoT) data, and data collected from internal sources. They must be able to bring this data together; define, cleanse, and order the data; and then harvest from that raw data the specific modeled data that may be used for deep analytics, machine learning, and business planning.

The technologies can include the NoSQL systems used to manage data at the edge of the enterprise and in systems of engagement; the data lake used for data collection, organization, and initial analysis; and the relational data warehouse for business analytics and decision support. These technologies tend to come from a variety of suppliers, including open source communities, managed subscriptions for open source–based products, and licensed software. The needs of the software technologies can vary as well and may demand flexibly deployed clusters of servers to optimize data storage and processing.

Managing, coordinating, and integrating these systems and their operations over time can be a daunting challenge, requiring a wide range of skills and great care in tracking version compatibility as software is acquired, upgraded, and patched. When these various technologies are acquired from many different sources and managed manually, the result can often be cost overruns and project failure.

The elements involved often include combinations of the following:

- A data lake management platform, usually based on Hadoop
- Technology for filtering, sorting, deduplicating, defining, and securing the data
- A streaming data ingestion and routing engine
- A high-performance, generalized data analysis platform, such as Spark
- A data integration technology that can move select data from a data lake or stream store to the enterprise data warehouse
- A data warehouse managed by a scalable, high-performance relational database management system
- Analytic tools for query, discovery, and visualization of the data
- Data-driven artificial intelligence and ML software that can be used to empower smart systems
- Servers, storage, and networking well configured to support all of the previously mentioned elements

Managing These Elements as a System

Bringing together and managing the elements as a system require a project of substantial proportion. The technologies must be carefully selected and an architecture developed to unite them. Processes need to be put into place for managing all elements of the system, adjusting and tuning the elements, and enhancing the elements as business requirements change. Such projects can run well over a year with no guarantee of success at the end.

A better plan is to find a well-designed combination of these technologies that is offered as a product, ideally by a single vendor. The product should not only include all the functional elements, well integrated, but also provide a hardware platform designed to run them all together with simple and straightforward design and management capabilities. One such offering is the IBM Integrated Analytics System.

The IBM Integrated Analytics System

The IBM Integrated Analytics System incorporates all the elements described herein and blends them as a single system. Up to now, IBM has offered two key analytics systems: PureData System for Enterprise Analytics and PureData System for Operational Analytics. The IBM Integrated Analytics System combines the capabilities of these two systems, including BLU Acceleration, into a converged solution. At its core is the next generation of technology that drove PureData System for Operational Analytics, formerly known as Netezza, which is backed by over 20 years of Netezza technology development.

It is intended for use in all structured data analytic contexts and to be applicable to the analysis of very low latency data — ingesting and processing the data within minutes of its creation. It is also designed to feature subsecond response times even in processing high-volume machine-generated data, such as cell phone call data records (CDRs). Yet it is designed to do so while supporting high concurrency, enabling thousands of users to slice and dice the data and to see timely data through such mechanisms as dashboards and alerts. With the embedding of Spark, and support for Python and R, the IBM Integrated Analytics System is meant to deliver first-class data science capability powered by the IBM Data Science Experience (DSX), enabling customers to do data science and machine learning at scale.

The purpose of the IBM Integrated Analytics System is to support deployment on-premises, in the cloud, or in a hybrid cloud configuration, incorporating relational and NoSQL database technology and supporting both structured and unstructured data. The IBM Integrated Analytics System enables seamless query across structured and semistructured data platforms in the cloud and on-premises, anchored by the Common SQL Engine.

The IBM Common SQL Engine is at the heart of IIAS, providing flexibility, application compatibility and portability, strong data integration, virtualization, and seamless query on-premises or in the cloud. Typically, data deployments require rewriting or restructuring queries and application/management schemes that use diverse data. The Common SQL Engine also includes an Oracle Application Compatibility Layer, allowing Oracle applications to integrate with the IBM Db2 family of offerings as well as IIAS. Typically, more than 98% of existing Oracle application code can run as is. Some manual work is usually required, but as IBM notes, "whenever changing a database deployment, vendor, or format, rewriting applications or licensing new software, or both, may be required."

In conjunction with the IBM Common SQL Engine, IIAS leverages the IBM Db2 Analytics Accelerator (IDAA) to significantly speed the execution of queries and delivery of key information to the users and applications that need it. Together, IDAA and the IBM Common SQL Engine deliver unparalleled mixed workload performance for complex business analytics, resulting in the ability to:

- Analyze large quantities of data quickly
- Eliminate DBA and data design work in analysis, index design, query rewriting, and so on
- Reduce monthly licensing charges by offloading complex query workloads

IIAS ensures that data written anywhere in the system is easily accessible. The elements of this system include the following:

- A managed public cloud DBaaS
- Software-defined warehouse on-premises or in the cloud
- A dedicated analytics appliance
- Ability to federate with Hadoop and move the query to the data using BigSQL, which can also integrate access to HDFS, NoSQL databases, object stores, and WebHDFS
- A custom-deployable database
- Open source Hadoop, backed by Hortonworks

The system is designed to accelerate development and deployment times for data scientists with a high-performance, optimized, and cloud-ready data platform. It also can be integrated seamlessly with other IBM systems, including applications and databases on z systems. Streaming data, captured by the IBM Db2 Event Store, can also be leveraged.

The hardware components of the IBM Integrated Analytics System include the following:

- IBM Power 8 S822L 24-core server with 3.02GHz processors and an integrated IBM FlashSystem 900
- In-place expansion tiered storage that may be configured to fit into or integrate with a hybrid cloud or logical data warehouse configuration
- Mellanox 10G Ethernet switches and Brocade SAN switches

The number of servers and cores and the amount of memory and storage vary based on the user's size requirements, with configurations ranging from one-third of a rack to four racks.

The appliance format described previously is not required to use IIAS. What is required is the MPP architecture that uses IBM Power and all IBM Flash Storage with in-memory BLU Acceleration. IBM has indicated that this technology delivers the following benefits:

- **Performance:** Unprecedented response times enable "train of thought" analyses frequently blocked by poor query performance.
- **Integration:** Deep integration with Db2 provides transparency to all applications.
- **Self-managed workloads:** Queries are executed in the most efficient location.
- **Transparency:** Applications connected to Db2 are entirely unaware of the presence of BLU Acceleration.
- **Simplified administration:** The hands-free operations of the appliance eliminate most database tuning tasks. IIAS can be integrated with System Z to deliver a hybrid computing platform that delivers dramatically faster business analysis for users of Db2 Z. It supports a combination of operational transaction processing with complex analytics that features high performance, deep integration, self-managing workloads, and simplified administration. It's also transparent. Applications connected to Db2 Z are unaware of the integration.

FUTURE OUTLOOK

Enterprises are looking at ways to integrate analytics systems for handling hybrid combinations of relational and NoSQL data with straightforward access as needed by nontechnical users. Such solutions will become available in the cloud from a variety of vendors, but for the time being, the question becomes how to implement such a system on-premises without incurring great cost and risk. Having a system such as the IBM Integrated Analytics System, which has a user datacenter deployment and corresponding IBM Cloud deployment, certainly makes that transition smoother and less risky.

CHALLENGES/OPPORTUNITIES

Many technologies are emerging that address aspects of the kind of integrated data collection and analysis described in this white paper. They may represent reasonable alternatives to the IBM Integrated Analytics System, at least for some workloads. Although most work well, they are generally limited in scope in terms of either the range of data supported or the range of systems that may be integrated easily. Also, few offer the level of integration with applications and databases on z systems that are offered by this product.

IBM's opportunity is to show that the range of data and analytics supported, combined with the ability to integrate with z systems and the compatibility and connectivity between these systems on-premises and in the IBM Cloud, sets this product apart from the alternatives.

CONCLUSION

The challenges that enterprises face in dealing with mounting volumes of data that must be ingested, correlated, and then acted upon are mounting. Enterprises must cope with growing numbers and types of data management systems and ever more complicated schemes for processing the data while also analyzing it and using it for ML-driven operations to exploit that data to maximum advantage. One approach to addressing this problem is to simplify it through an integrated configuration that covers all the bases, is supported by a trusted technology supplier, and can deliver data ingestion, data organization, transaction processing, and analytics, including ML, in a single coherent system. The IBM Integrated Analytics System serves just this purpose.

Enterprises faced with the challenges outlined previously should consider the following:

- Take an inventory of the range of facilities, including servers and storage, currently used to ingest, filter, sort, and order data; to combine data into useful structures; to execute transactions on the data; to perform analytics on the data; and to support ML driven by the data.
- Determine how much staff time is required to manage all of the systems, including routine maintenance, patching, and the manual steps involved in moving data from system to system as required.
- Calculate the business cost represented in the latency of data movement, the time required for data reformatting and movement, and the occasional incidents of human error, all of which result in delays affecting the practical leveraging of the full value of the data.
- Consider how much might be saved, and how much more might be gained, by moving to an integrated system of elements that addresses the various requirements of data ingest, transformation, transaction processing, analytics, ML, and so on. Such a calculation should be calculable in equipment cost, staff time, and the opportunity cost of delay.
- Investigate the availability of integrated data management and analytics systems that address these issues and include in that list the IBM Integrated Analytics System.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com