White Paper

The UniConnect^{TM} In-Memory Data Virtualization and Integration Appliance: The Next Revolution in Big Data Processing

Hewlett Packard
Enterprise

intel

Percipient
intelligent data engineering

This short paper introduces a revolutionary In-Memory Data Virtualization and Integration Appliance architecture built on HPE ProLiant™ systems and UniConnect™ — an industry-leading flagship product from Percipient. The partnership of optimized hardware-software architecture is the first of its kind for unifying Big Data aggregation and analytics across unlimited data streams and unprecedented efficiencies in power, scalability, availability, reliability, physical footprint, and total cost of ownership. Such a NextGen Data Integration ecosystem heralds the arrival of innovative solutions for application/report delivery reliant on varied data sources, data virtualization requirements, unified data access needs and archival scenarios. Productivity will be enhanced across costly OLTP systems, accelerated Extract, Transform Load (ETL) needs, and centrally-hosted infrastructures.

Table of Contents

Current Data Warehousing Landscape Challenges

Currently, the global surge in data collection trends and technologies has presented numerous challenges to how this growing explosion of Big Data can be mined at high speed and with cost-effectiveness:

- **Inefficient Processes** to virtualize and/or unify data from multiple data sources, including structured, unstructured and real-time data through in-memory parallel processing.
- **Data Has To Be Physically Moved To A Central Location** before it could go through the data discovery stage. With Big Data, this approach is too expensive and impractical. A platform to discover data "in place" is sorely needed. It has to be able to support the indexing, searching, and navigation of different sources of Big Data. It has to be able to facilitate discovery of a diverse set of data sources, such as databases, flat files, and content management systems. The security profile of the underlying data systems needs to be strictly adhered to and preserved.
- **Traditional Architectures Decouple Analytical Environments From Data Environments.** Analytics software now runs on its own infrastructure and retrieves data from back-end data warehouses or other systems to perform complex analytics. Hence, analytics are treated as a distinct workload that has to be managed in a separate infrastructure. This architecture is expensive to manage and operate, creates data redundancy, and performs poorly with increasing data volumes. The challenge is to have an architecture run both data processing and complex analytics on the same platform. It needs to deliver petabyte-scale performance throughput by seamlessly executing analytic models inside the platform, against the entire data set, without replicating or sampling data. It must enable data scientists to iterate through different models more quickly to facilitate discovery and experimentation with a "best fit" yield.
- **Current Challenges In Performing Analytics On An Activity As It Unfolds** present a huge bottleneck for the analytic enterprise. Historically, analytic models and computations run on data that is already stored in databases. Even the best performing storage drives have unacceptable latencies for real-time processing. Enterprises that want to boost their Big Data IQ need the capability to analyze data as it is being generated, and then to take appropriate action. It is about deriving insight before the data gets stored on physical disks. Since the volume of the data stream can vary dramatically, Big Data platforms not only have to be able to support data analytics in motion, but they currently experience inefficiencies to scale effectively to manage increasing volumes of data streams.
- **Challenges To Reduce The Analytics Cycle Time,** that is, to cut the amount of time needed to discover and transform data, develop and score models, and analyze and publish results. There are challenges in improving developer productivity. Most companies lack the luxury of employing hundreds of developers on hand who are skilled in new age technologies. To democratize Big Data across the enterprise, challenges abound in flattening the time-to-analysis curve with a rich set of accelerators, libraries of analytic functions, and a tool set that accelerates the development and visualization process. Because analytics is an emerging discipline, it is not uncommon to find data scientists having their own preferred mechanisms for creating and visualizing models. Creating a restrictive development environment curtails their productivity. A Big Data platform thus needs to support interaction with the most commonly available analytic packages, with deep integration that facilitates pushing computationally-intensive activities from those packages into the platform. It needs to have a rich and proven set of "parallelizable" algorithms to run on Big Data, and have specific capabilities for unstructured data analytics, and a framework for developing additional algorithms. It must also provide the ability to visualize and publish results in an intuitive and easy-to-use manner.

How UniConnect™ Provides Revolutionary Solutions

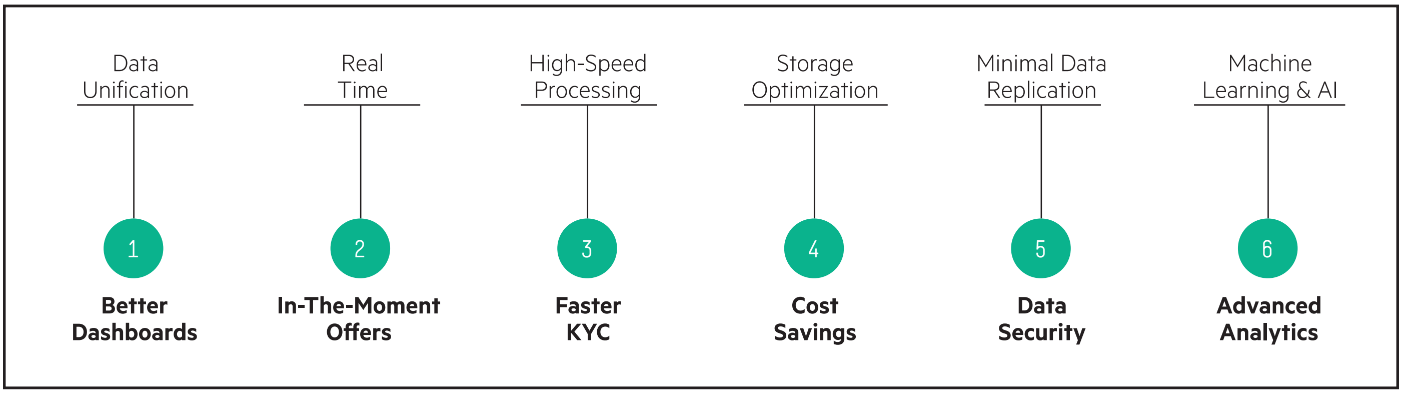
To address all the aforementioned challenges, Percipient's UniConnect™ platform amalgamates, integrates with, and enhances select open source software, including Hadoop, Presto, Spark, Kafka and Cassandra, and delivers five key capabilities that define the platform's unique proposition:

1. In-Place Data Discovery and Exploration
2. In-Memory Processing and Real-time Analytics for extreme performance
3. Parallel processing of Distributed Data, including unstructured Big Data
4. Powerful and universally compatible development and hardware infrastructure
5. Integration and governance of all data sources

The UniConnect^{TM} Ecosystem: Key Functions/Benefits

Today, the bulk of enterprise data remains trapped in legacy or spaghetti-like architecture. By unlocking data, UniConnect™ helps solve some of an enterprise's most complex, but also most rewarding, data challenges.

Designed for analytics: UniConnect™ offers organizations an advanced and efficient discovery platform offering advantages in six key areas (see chart) with which to productively implement business intelligence, financial and risk reporting, customer 360 campaigns and many other analytics-powered digital applications, without the need for a costly data warehouse investments. UniConnect™ is able to overcome the shortcomings of traditional databases that are designed to deliver ACID (Atomicity, Consistency, Isolation and Durability) properties. These properties ensure the integrity of transaction data and require developers to hard-code queries in the business layer. However, in the new age of big data analytics, databases need to also accommodate ever-changing queries, and complex joins across datasets.



As a simplified integration layer seamlessly inserted into an organisation's existing data infrastructure, UniConnect™ provides organizations with the flexibility to query as much data and as frequently as required. Data can be queried directly from a source system, database, data warehouse, Hadoop or any other data generating/storage mechanism, and unified via a single interface. The platform is also able to eliminate the Map Reduce process when accessing data from a Hadoop platform, thereby shortening processing time by more than 15 times for the same volume of data. All queries and processing functions are easily tracked via the UniConnect™ Admin Portal.

To enable predictive analytics, UniConnect™ facilitates connectivity from open source analytical tools e.g., Spark, R and Weka, as well as proprietary tools e.g., SAS, using a JDBC interface. This connectivity gives data scientists the ability to use UniConnect™ to access a full range of algorithms and statistical packages, including linear and non-liner modeling, time-series analysis, classification, and clustering. Results can be exposed as APIs for external reporting applications.

High Speed and Secure: The UniConnect™ platform is also able to materially lift an organization's data processing speed and security by embedding the latest in-memory computing capabilities. RAM is roughly 5,000x faster than standard disk drives, and while traditional databases operate by persisting every transaction through a logging process, UniConnect™ uses RAM to access only the datasets that are relevant to the query. The memory needed for such jobs is not reliant on a single server. Rather, parallel distributed processing via a basic ethernet networking infrastructure makes it easy to partition a large dataset across the RAM of several individual computers. In this way, a computer's memory capacity can be easily and cheaply scaled. This allows the platform to accomplish groundbreaking processing speeds, cutting data mining time by 4 to 8 times, for about a tenth of the cost of traditional solutions. For example, under test conditions, a large bank's risk analytics query of 62 million rows was benchmarked at 36 minutes, using a traditional data integration tool. The UniConnect™ platform was shown capable of the same query in 8 minutes, using 4 commodity machines that cost less than US$10,000 to acquire.

However, UniConnect™ platform's business value lies not just in the increased ROI (by enhancing the performance of the tools that organizations have already invested in), and reduced TCO (by limiting both hardware expenditure and operational overheads). As important are the security benefits that the platform's in-memory features can bring. Unlike almost all other data integration solutions, the data queries processed by UniConnect™ are not written to disk, and therefore do not exist in a physical form. The platform ensures that data is re-retrieved from the source rather than a local cache in the event of a system fault or system crash.

Seamless Deployment: The UniConnect™ platform is an advanced middleware that slips seamlessly into an organization's existing data infrastructure. Implementation of the platform does not require the replacement or removal of any existing tools. Instead, the platform is mostly agnostic to data formats and types, and capable of ingesting data from a large range of existing and new sources, including CRM and transaction systems, databases, APIs, and blockchain. In particular, UniConnect™ is able to access data from real-time sources, such as from real-time apps, IoTs (sensors and devices), and social media. On the other hand, the platform exposes its own APIs to allow integration with a host of consuming applications. This means that there is almost no data that an organization cannot unify. The UniConnect™ discovery layer does so by untangling complex enterprise data architecture from a spaghetti orientation to a data flow that is simplified, well-organised and effective. By doing so, UniConnect™ can help organizations construct real-time management dashboards, deliver elaborate regulatory reports, effect highly personalized customer services, and build highly impressive customer-facing applications. These require substantial processing stability in the face of large and evergrowing data volumes.

Business Requirements	UniConnect Functionalities
Operational Efficiency	• Unifies data across multiple sources without copying • Direct access to HDFS/Hive • Supports in-memory processing • Data compression and access to compressed data • Cloud-based deployment for each LOB, if required
Scale-out Data Storage	• Scalable and expandable using commodity machines • Extensible licensing model
Data Security	• User restricted access • Maintains the user authorisation of the underlying platforms • Admin user interface supportive of audit protocols
Real-Time User Engagement	• Able to integrate real-time messages with structured & unstructured data • Reads real-time URL data (JSON format)
Reporting & Advanced Analytics	• Supports SQL queries • Exposes APIs for external reporting applications • Integrated with Spark for processing power, machine learning algorithms and graph computations • Connectivity with R statistical computing and graphics environment • Data retrieval and ability to create tables from unified data

Infrastructure Configurations

The HPE-Percipient UniConnect™ ecosystem can be deployed in two different ways on top of existing data warehouse or Big Data platforms for joint query processing: Option 1 with In-memory Analytics, and Option 2 with In-memory Analytics and Big Data RA.

Appliance layer	Functionality	Minimum Configuration
Option 1	**HPE-Percipient In-memory Data Unification Appliance**	
Queen Node	Responsible for parsing statements, planning queries, and managing UniConnect™ worker nodes. It is the "brain" and also the node to which a client connects to submit statements for execution	Min 1 node
Worker Node	Responsible for executing tasks and processing data. Worker nodes fetch data from connectors and exchange intermediate data with each other	Min 2 nodes
Interconnection	HPE Aruba Networking	Min 10 Gbps throughput
Clustering	For high availability	HPE Serviceguard recommended
Option 2	**HPE-Percipient In-memory Data Unification Solution with Big Data RA**	
Queen and Worker Node	Responsible for executing tasks and processing data. Worker nodes fetch data from connectors and exchange intermediate data with each other	Min 3 nodes
Hive Storage	Hadoop storage for unstructured data	Min 3 nodes
Interconnection	HPE Aruba Networking	Min 10 Gbps throughput
Clustering	For high availability	HPE Serviceguard recommended
Monitoring Tool	For cluster management	HPE Cluster Management Utility
	Recommended platform for Auto Scaling (Auto-Deployment of new nodes)	
Composability	HPE ProLiant Synergy Infrastructure with Intel® Xeon® processor and Image Streamer Integration	

Node description for Queen and Worker: 8 Core Intel® Xeon® /256GB RAM

Note: For larger deployment scenarios where multiple terabytes of data from multiple data sources need to be processed for executing joint queries, HPE-Percipient recommends an in-memory platform by leveraging on the seamless scalability capabilities of the HPE Synergy Composability infrastructure & HPE Big Data reference architectures.

HPE References

HPE ProLiant Gen 10 Servers
https://www.hpe.com/us/en/servers/gen10-servers.html

HPE Synergy
https://www.hpe.com/in/en/integrated-systems/synergy.html

HPE Apollo Systems
https://www.hpe.com/in/en/servers/apollo.html

HPE Servers for Big Data Analytics and Hadoop
https://www.hpe.com/in/en/servers/big-data-analytics.html#Deployment

HPE Aruba Networking
https://www.hpe.com/us/en/networking/switches.html

Benchmarking Overview

The purpose of a benchmark evaluation is to validate and analyze the functionality and the scalability of UniConnect™ v1.0 running on HPE hardware to process various reports from multiple data sources at the backend by using SQL queries; and to arrive at an optimum sizing for actual customer production environment.

The TPC Benchmark™H (TPC-H) is a decision support benchmark consisting of a suite of business-oriented complex queries and concurrent large volume data modifications. Using this benchmark and just a single node installation of the Intel® Xeon® Processor v7 Family, the UniConnect™ platform was able to demonstrate average query CPU times ranging from under half a minute for 1 GB to 10 hours for 3,000 GB of data. The unification of CSV and Hive data across 2.3 billion rows took 30 minutes. For every node added, these performance outcomes can be lifted by several multiples.

After several weeks of study to understand and then put Percipient's flagship in-memory data unification product UniConnect™ through the most stringent TPC-H benchmarking performance testing at HPE Labs in Bangalore, unifying Oracle (1TB data size), MySQL Server (1TB data size) and Hive (3TB data size) in-memory was easy and quick. UniConnect™ was stable with TPC-H benchmark queries, and demonstrated stability in the unifying of over 5TB data set from three different data sources which were different in the way they operated on the data. More details can be found in the full white paper available from HPE and Percipient: Technical White Paper on Percipient UniConnect™ Appliance Deployment on HPE Infrastructure.

It is thus established that UniConnect™ has been vigorously tested to confirm high system speed and stability.

Epilog

This short paper has presented a quick overview of a radical reference architecture deployed on HPE hardware that will revolutionize the way data is unified and analyzed for gleaning unprecedented levels of business intelligence. The UniConnect™ ecosystem architecture is the first of its kind to deliver unprecedented benefits to small, medium and large clients alike, and offers compelling benefits in terms of enhanced business intelligence, lower TCO, higher ROI and all-round technical flexibility, security, productivity and reliability. The key attributes of UniConnect™ can be summarized as follows:

1. In-place data discovery/exploration and integration with traditional as well as digital data sources
2. Extreme performance by integrating data processing and complex analytics onto the same platform and with Cloud-scale capability leveraging on the HPE Composable Infrastructure
3. Game-changing management, storage, retrieval, exploration and analysis of both unstructured and structured data
4. In-memory, real-time analytics of (big) data in motion
5. Integration and governance of all data sources for quality, integration, security, master data maintenance and information life-cycle management, on an open-source platform with no vendor lock-in

Detailed information on Percipient and HPE partnership details are available online at **http://www.percipientcx.com**.

Full-scale technical and marketing papers and case studies on the HPE-UniConnect™ ecosystem are available upon request through HPE and Percipient representatives.

Brought to you by:

November 2017