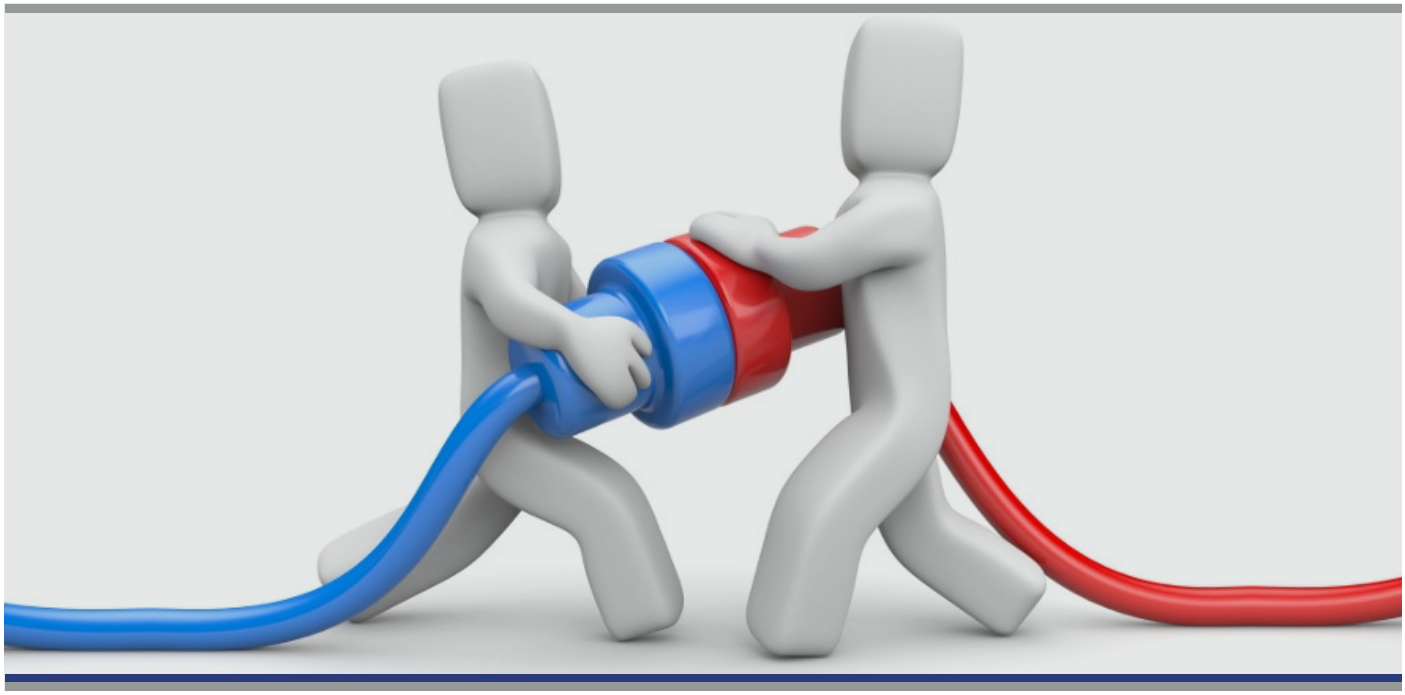
Percipient©
intelligent data engineering

UniConnect Integrate™



A Percipient Technology White Paper

Author: Gayathri Dwaraknath
Chief Solutions Officer
Updated Aug 2018

Content

1. Introduction

As many enterprises are discovering, big data analytics does not start with skilled data scientists or sophisticated algorithms. The journey to implement real time, personalised customer applications or advanced business intelligence begins with the optimal integration of data, regardless of its source, format or volume.

For the past 30 years, the most popular method of integrating data has been by means of a common data storage mechanism. This means that data is physically replicated and stored independent of its point of origination, typically in a Data Warehouse (DWH). By means of an ETL (Extract-Load-Transform) process, this data is pulled from source systems, normalized, stored in a single repository, and made query-able via a common interface. In this way, DWHs make it possible for enterprises to have a unified view of their data across heterogeneous sources and formats.

Despite (and sometimes, as a result of) advances in data generation and digital content, the efficiency and speed of data processing in many enterprises remains highly sub-optimal.

However, this physical integration of data has several drawbacks. The deployment of a separate system to house data that is continually expanding costs enterprises from hundreds of thousands to several million dollars to maintain. This is because DWHs require enterprises to 'scale-up' for every block increase of data by ramping up CPUs, hard-disks, network cards, and of course, license fees.

Enterprises are also continually seeking to embed new and emerging technologies. For example, alongside a physical data warehouse, many enterprises have turned to cloud storage services. These cloud services are sometimes preferred because they represent an operating, rather than capital, expense. Similarly, many enterprises are adopting Data Lakes as a more cost effective way to parallel process and store large volume of structured and unstructured data. However, the shift towards these modern and more cost effective technologies have exacerbated the problem of data siloes.

In addition to cost is the increasing need for enterprises to conquer data volume and processing performance amid the explosion of digital channels. Given the complex processes required for the unification and management of data within an unwieldy data technology stack, organisations have generally reconciled themselves to data delivery delays that can last hours, days or weeks. Despite (and sometimes, as a result of) advances in data generation and digital content, the efficiency and speed of data processing in many enterprises remains highly sub-optimal.

This is where Percipient's products, UniConnect Integrate™ and UniConnect Transact™, come to the rescue. By enabling FIs to access, unify, publish and consume data, the UniConnect suite of products helps bridge the chasm between line-of-business siloes, internal and third party data, and between modern and legacy technologies. Through the entire spectrum of small to big data, UniConnect helps enterprises turn innovative ideas into workable solutions, and accelerate the deployment of these solutions into a variety of game-changing applications.

This whitepaper focuses on UniConnect Integrate™, which aims to unify highly diverse data for the purpose of implementing flexible, high speed and scalable data analytics and data applications.

2. Product Goals

2.1. On-the-fly data discovery and exploration

Data discovery is the process of determining what data is available within a particular source, and getting a sense of its quality and its relationship with other data elements. This process enables data scientists to create the right analytic model and computational strategy

The analytic architecture of the future needs to run both data processing and complex analytics on the same platform

Traditional approaches require data to be physically moved to a central location before it can be discovered. With Big Data, this approach is too expensive and impractical unless this can discovered indexed, searched, and navigated "in place" across a diverse set of data sources. This includes databases, flat files or content management systems, and any other persistent data store for structured, semi-structured or unstructured data.

The security profile of the underlying data systems needs to be strictly adhered to and preserved. These capabilities benefit analysts and data scientists by helping them to quickly incorporate or discover new data sources in their analytic applications.

2.2. Run analytics closer to the data

Traditional architectures decoupled analytical environments from data environments. Analytical software would run on its own infrastructure, based on data retrieved from back-end data repositories. The rationale behind this was that data environments were optimized for faster data access, but not necessarily for advanced mathematical computations.

The analytic architecture of the future needs to 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, thereby enabling data scientists to iterate through different models with minimum delays.

2.3. Integrate data-at-rest and data-in-motion

Performing analytics on activity as it unfolds presents a huge untapped opportunity for the analytics enterprise. Historically, analytic models and computations have run on data that must first be batched, typically overnight.

Enterprises that want to boost their Big Data IQ need the capability to analyze static as well as streaming data, ie as and when the data is generated. For example, consider a data stream from a stock trading system that has swelled significantly above its normal volume. A bank wants to understand the cause of this by identifying the profiles of the traders involved.

To join static and streaming data, an integration platform must transform the data to a common language ie enable SQL querying, easy to assemble, deliver high performance and fault tolerance.

2.4. ELT, not ETL

ETL processes require enterprises to Extract from source systems, Transform into target data formats in a staging area and Load in the target system. Such processes enable data analysis, but enterprises struggle when confronted with requirements of unstructured, voluminous and volatile data.

In contrast, ELT processes first Extract and Load as per source format, with filtering if required, and then Transform on demand or at target. ELT models can provide much needed relief when enterprises are faced with an avalanche of ever-increasing data of interest to their internal applications. By allowing faster turnarounds in use of data, and greater flexibility and agility in incorporating new sources and changing definitions, ELT models are highly cost-efficient

Also by processing data on-the-fly, making full use of the available compute resources over a cluster of processing nodes, the T of ELT can happen at much higher speeds, even for higher volume data.

3. Percipient Overview

UniConnect Integrate™, is designed to facilitate the development of high performance applications by enabling on-demand aggregation of multi-sourced data

Percipient is a data technology company founded in December 2014. The company helps enterprises to integrate their data across both traditional and modern systems. The company's flagship in-memory data unification platform, UniConnect Integrate™, is designed to facilitate the development of high performance applications by enabling on-demand aggregation of multi-sourced data. Implementations of UniConnect Integrate™ are already underway in India and the US.

Percipient was named by EY to their prestigious Global Top 30 innovative tech-focused companies under their 2018 Accelerating Entrepreneurs program. Percipient's next-gen data virtualization & API management platform was labeled as 'the next revolution in big data processing' by Hewlett Packard Enterprise in 2017.

More details are available at: **http://www.percipientcx.com**

4. Architecture Principles

4.1. The need for speed

Given the complex processes required for data to be unified in a DWH, organisations have generally reconciled themselves to data availability that is lagging by several hours, days or weeks. This is because data must be reconciled and then loaded into the DWH according to a scheduled downtime, typically overnight.

The introduction of Hadoop platforms has revolutionised the ability to store and process very large volumes of unstructured and provisional data. However,its MapReduce model still relies on data being processed in batches rather than as a continuous stream. Therefore, while Hadoop has greatly enhanced large scale data processing, it does not support high speed data analytics and applications.

In fact, some studies suggest that for a variety of analytical tasks, and assuming a fixed number of nodes, a MapReduce framework may actually be slower than two parallel database systems by a factor of 3.1 to 6.5. UniConnect Integrate™ is 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.

4.2. Efficient processing

UniConnect IntegrateTM is 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.

The UniConnect Integrate™ platform is also able to materially lift an organisation's data processing speed and security by embedding the latest in-memory computing capabilities. RAM is roughly 5,000 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 basic ethernet networking infrastruture makes it easy to partition a large dateset across the memories of several individual computers. In this way, a computer's memory capacity can be easily and cheaply scaled.

This allows the platform to accomplish ground-breaking processing speeds, bringing data mining jobs down 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 Integrate™ platform was shown capable of the same query in 8 minutes, using 4 commodity machines which cost less than USD 10,000 to acquire.

4.3. High security standards

The UniConnect Integrate™ platform has inbuilt features to support only authenticated and authorised access. These controls take place at three levels:

- **Who –** user access is restricted by integrating with an enterprise's existing LDAP server. LDAP protocols are the enterprise standard for user authentication. Where more fine-grained users authentication is required, UniConnect Integrate™ is also integrated with the open source Shiro framework that recognises not just persons, but also roles, permission holders or even computer processes.

- **What** – The functional authorization is also maintained for each of the data sources. This means that UniConnect Integrate™ maintains the access control of the underlying system or database, which can extend to the specific schema or tables within the database. Shiro adds further restrictions, by limiting access to an instance and functionality-level.
- **How** – The UniConnect Integrate™ admin user dashboard is supportive of audit requirements. The dashboard enables users to keep track of the queries made, query owners, the data sources queried, and all finished, failed or erroneous queries. This is in line with most audit protocols.

5. High Level Architecture

5.1. Product Overview

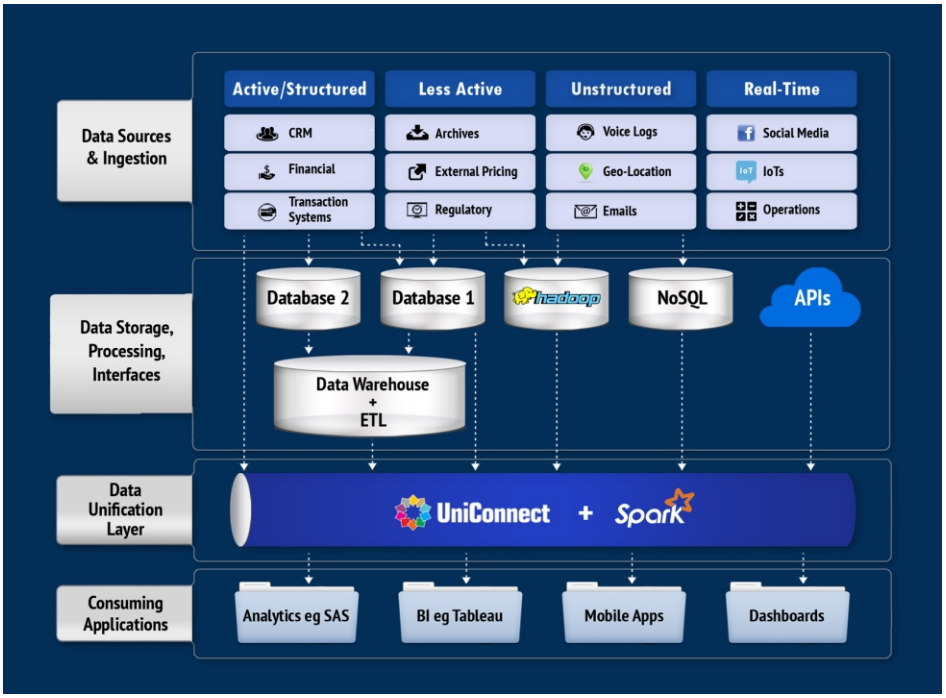
The platform is largely agnostic to data formats and types, and capable of ingesting data from a large range of existing and new sources.

The UniConnect Integrate™ platform is an advanced middleware that slips seamlessly into an organisation's existing data infrastructure. Implementation of the platform does not require the replacement or removal of any existing tools.

In addition, the platform is largely 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. The platform is also able to consume data from real time sources, such as from real time apps, IoTs (sensors and devices), and social media.

Meanwhile, the platform exposes its own APIs to allow integration with consuming applications, as shown in the diagram below.

Figure 1: UniConnect Integrate™ reference architecture



5.2. Product Details

The UniConnect Integrate™ platform amalgmates, integrates with, and enhances select open source software, including Hadoop, Presto, Spark, Kafka and Cassandra, to deliver four key capabilities:

- **Data connectors**

  UniConnect Integrate™ enables interactive analytical querying of data via connectors to data sources. Regardless of its origination, format or type, data can be queried and consumed.To date, the platform offers a large variety of connectors to almost all standard databases, transaction systems, and real time sources including IOTs and web streaming.

- **Data virtualization**

  UniConnect Integrate™ provides inbuilt support for a rich library of functions that can operate on the underlying raw data in high speed memory. Unlike traditional ETL processing, queries are dynamically pushed down to the sources of the data, and only the data relevant to the query is selected and processed in the data virtualization layer.This means less data is transferred through the network, enabling much higher ETL speed and efficiency.

Only the data relevant to the query is selected and processed in the data virtualization layer

- **In memory processing**

  UniConnect Integrate™ processes data by loading it into a computer's main memory, also known as RAM (Random Access Memory) instead of its hard disk, or CPU. Data travels to CPUs via wired connections that slows down processing. In memory processing means that data can be queried on-the-fly, and is also more secure as data is not physically copied.

- **Distributed data**

  In UniConnect Integrate™ ,queries are parsed and then distributed among a cluster of compute nodes, rather than relying on a single centralised server. These nodes enable the data to be processed not sequentially but in parallel, with the outcome then reassembled for the final result. Distributed data processing is more reliable and delivers high performance even for large data volumes.

- **SQL-based queries**

  The UniConnect Integrate™ platformenables heterogeneous data can be queried using the standard programming language, ANSI SQL. This overcomes the issue of poor portability of query codes across different systems. The ability to issue a single query against disparate sources not only facilitates data integration but also reduces ETL workloads.

6. Benchmarking

Enterprises require substantial processing stability in the face of large and ever-growing data volumes. UniConnect Integrate™ has been vigorously tested to ensure such stability based on theTPC-H industry standard. 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 3000 GB of data. The unification of CSV and Hive data across 2.3 billion rows took 30 minutes. For every node added, these performances can be lifted by several multiples.

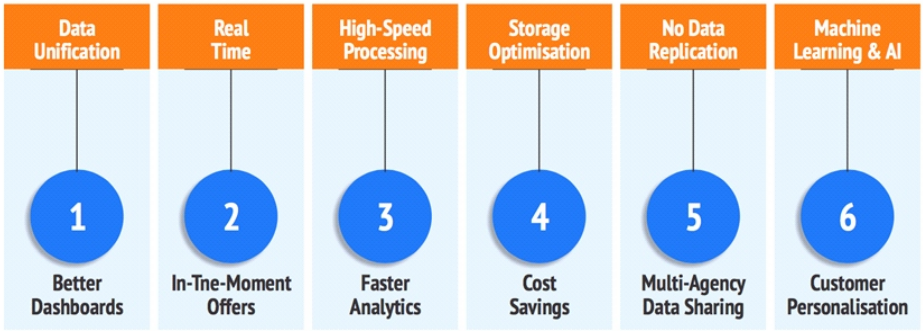
7. Use Cases

There is almost no data that an enterprise cannot unify with UniConnect Integrate™. The platform does so by untangling complex enterprise data architecture from a spaghetti orientation to a data process that is seamless and agile.

This result of using UniConnect Integrate™ is multifold: significant cost savings, process efficiencies and improved performance.

This result of using UniConnect Integrate™ is multifold: significant cost savings, process efficiencies and improved performance.

This result of using UniConnect Integrate™ is multifold: significant cost savings, process efficiencies and improved performance. By making siloed data available for discovery and analytics, UniConnect Integrate™ makes it possible for enterprises to make strong strides in their digital transformation journey.



Here are some value-add applications that can be implemented using UniConnect Integrate™.

7.1. Managed migration to newer technologies

Here is a specific financial sector use case. A large bank with several lines of business used an OLTP database to manage and store 60 TBs of data relating to its merchants, corporate clients and retail customers. Some of the data was compressed to save on storage costs. The database supported several online applications accessed by over 5 million users. Data growth averaged 200 GB daily.

However, this single, large database suffered from performance issues. In addition, data was distributed in the database on the basis of pre-defined queries, greatly limiting user flexibility. A single database also resulted in unacceptable process interdepencies affecting the individual lines of business.

UniConnect Integrate™ made it possible for the bank to separate the data into different smaller instances without sacrificing the bank's data integration capabilities. Furthermore, a next gen Hadoop Data Lake wasintroduced to archive data for regulatory reporting. This archived data continued to be easily queried alongside the bank's current, more active data in order to support on-demand bank statementing.

8. Summary

Here is a easy reference summary of UniConnect's key functionalities:

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