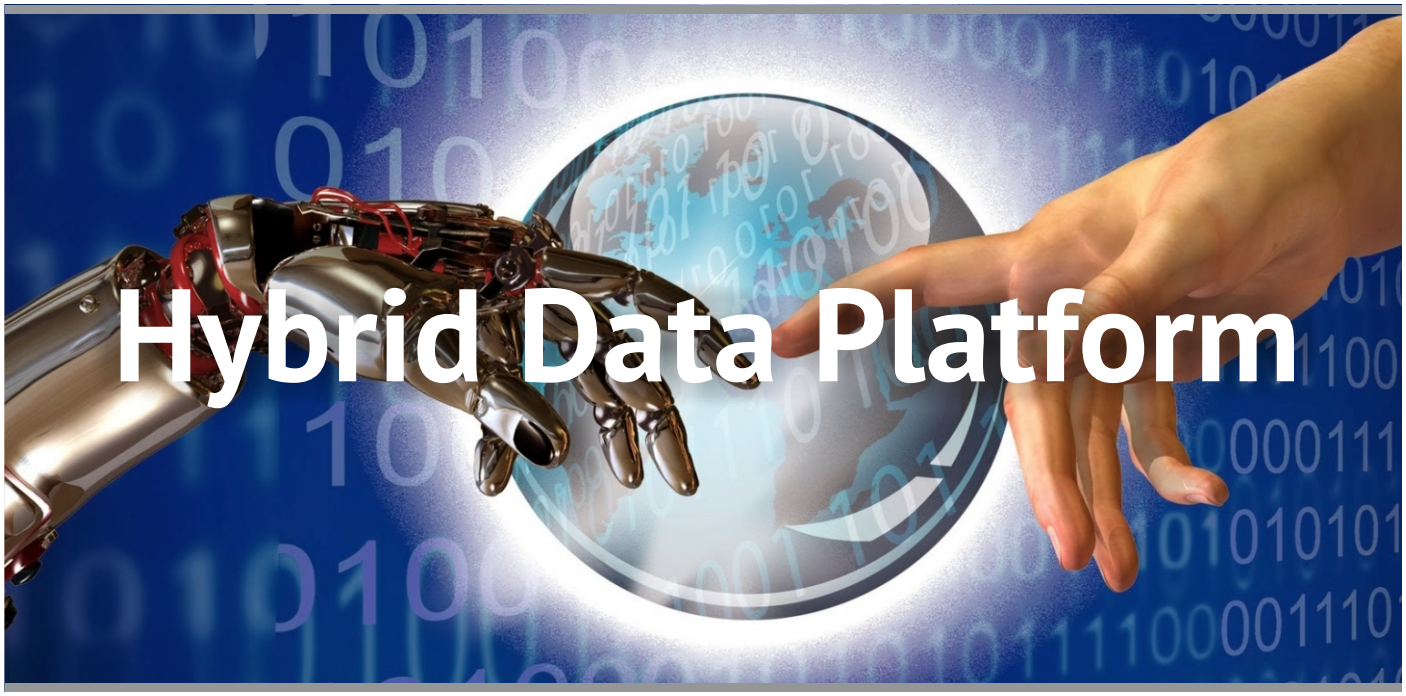
Percipient
intelligent data engineering

UniConnect-Powered Data Aggregation Across Enterprise Data Warehouses and Big Data Storage Platforms



A Percipient Technology White Paper

Author: Ai Meun Lim
Chief Product Officer
Updated Aug 2017

Overview

The concept of hybrid data storage platforms has gained traction over the past few years. By combining existing enterprise data warehouses (EDWs), and new age data lakes, hybrid data platforms help break down data silos, thereby enabling business analysts, data scientists, and marketing teams to gain useful insights and improve customer services.

These hybrids offer several advantages over traditional data warehouses. They create disruptive possibilities that help enterprises:

- enhance business decision support at various levels by integrating data from various data silos to respond faster to business imperatives.
- overcome the performance and functional limitations of existing legacy systems while reduce existing server and licensing costs.
- identify hidden patterns in customer data to help devise targeted marketing strategies.
- leverage the scalability and fault resistance offered by modern technologies in order to enhance digital initiatives
- take advantage of cloud storage mechanisms

Data Lakes vs EDWs

Percipient believes that data lakes and EDWs are complementary and can happily co-exist.

A data lake, often associated with Hadoop-based storage mechanisms, is a repository for holding very large amounts of data in its native format and using a flat architecture unlike the hierarchical and relational approach adopted by EDWs. This means that not only are data lakes more scalable, the development time needed to introduce a new data set to a data lake is much shorter than that needed for a EDW. Data warehouses are also more expensive to maintain than data lakes owing to the high cost of hardware and licenses. As a result, enterprises have understandably sought to take advantage of capabilities that data lakes bring, yet are reluctant to completely forsake the fixed schema model and ACID (Atomicity, Consistency, Isolation, Durability) properties that EDWs bring.

Nor should they have to. Percipient believes that data lakes and EDWs are complementary and can happily co-exist. By deploying both systems in an optimal fashion, it is possible to leverage on the best of both worlds. Percipient can help enterprises create an enterprise data lake (either on-premises or on-cloud) for storage of their unstructured, high volume or less active data, while continuing to maintain a more limited EDW framework within an optimised data eco-system. This hybrid data platform is enabled by Percipient's UniConnect platform, which is designed to ensure that, despite their disparate technologies, the data from both sources remain easily accessible, can be seamlessly aggregated, and is available to power both big data and traditional applications.

Feature Rich

Percipient's hybrid data platform offers the following capabilities:

1. Data Discovery and Exploration:

The process of data analysis begins with understanding data sources, figuring out what data is available within a particular source, and getting a sense of its quality and its relationship to other data elements. This process, known as data discovery, enables data scientists to then develop the right analytic model and computational strategy for maximum insight. Traditional approaches required data to be physically moved to a central location before it could be discovered, but the rise of Big Data has rendered this approach too expensive and impractical.

These capabilities benefit analysts and data scientists by helping them to quickly incorporate or discover new data sources in their analytic applications.

To facilitate data discovery and unlock resident value within Big Data, the platform must be able to discover data "where it lives" regardless whether this is a EDW or data lake. In fact, the platform should be able to support the indexing, searching, and navigation of all other data sources, including data marts, flat files, content management systems, and any persistent data store of 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. Extreme Performance:

To achieve performance, it is necessary to run analytics closer to the data. Traditional architectures decoupled analytical environments from data environments. Instead, analytics is treated as a distinct workload and analytical software runs on its own infrastructure. Data for analytics is copied from back-end data warehouses where the data had previously been extracted from different transactions systems, transformed and loaded. The rationale behind this was that data warehouse environments were suitable for access and storage of data, but not necessarily for advanced mathematical computations.

UniConnect solves this very efficiently – it's in memory unification allows real time assimilation of data from multiple sources and we can expose that directly providing very unique value to clients.

This architecture was expensive to manage and operate, created data redundancy, and performed poorly with increasing data volumes. Percipient's hybrid architecture runs both data processing and complex analytics on the same platform using data gather from multiple systems. It delivers peta byte scale performance throughput by executing analytic models inside the platform. These models are run against either the entire, or selected, data sets without replicating or sampling the data. It enables data scientists to iterate through different models to facilitate discovery and experimentation with a "best fit" yield.

3. Multi-Structure Data Unification:

For a long time, data has been classified on the basis of its type—structured, semi-structured, or structured. Existing enterprise infrastructures typically have barriers that prevent the seamless correlation and holistic analysis of this data; for example, EDW is used for structured data and data lakes for unstructured or semi-structured data. These are managed as independent systems with very limited ability to co-mingle their data on demand.

UniConnect retrieves unstructured and structured data onto a single interface, and provides the tools to explore, join and transform datasets as required.

Unfortunately, organizational processes don't distinguish between data types. Take for example, the need to analyze customer support effectiveness at a call centre. This typically entails structured call information such as call time, duration, general outcome, and customer satisfaction survey responses. However, as important is unstructured information gleaned from the conversation, such as customer sentiment, specific concerns raised, and the service agent's responses. By combining structured and unstructured data, it is possible to analyse customer interactions in

context and thereby accurately identify ways to improve the service. A game-changing analytics platform like UniConnect retrieves unstructured and structured data onto a single interface, and provides the tools to explore, join and transform datasets as required.

4. Data Analytics in Real Time:

Performing analytics on activity as it unfolds presents a huge untapped opportunity for enterprises. Historically, analytical models and computations ran on data that was stored in EDWs. This worked well for transpired events, typically batched from the previous day, with disk drives used to store and retrieve data. However, even the best performing disk drives have unacceptable latencies when needing to react to events in real time. Enterprises wanting to boost their Big Data IQ need the capability to analyze data as it is being generated, and then to take the appropriate action immediately.

Percipient's UniConnect platform is not only able to support analytics of data in motion, but can also be used to manage the real time integration of data-in-motion and data-at-rest.

This means deriving insight before the data gets stored on physical disks. We refer to this type of data as streaming data, and it is the analysis of this data in motion that offers enterprises some of the most innovative ways to engage with their customers. However, part of the problem of analysing data in motion is the inconsistency of flows. For example, depending on the time of day, the volume of the data stream carrying stock trades in an Exchange can quickly swell from 10 to 100 times its normal volume. The Stock Exchange may also want to analyse this streaming data against static company data held in the exchange's data warehouse. Percipient's UniConnect platform is not only able to support analytics of data in motion, but can also be used to manage the real time integration of data-in-motion and data-at-rest.

5. Analytical Functions and Tool Sets:

One of the key goals of a Big Data hybrid platform should be to reduce the analytic cycle time, i.e. the amount of time that it takes to discover and transform data, develop and score models, and analyze and publish results. A platform that supports fast data access and exploration means data analysts are empowered to run multiple analytic iterations and speed up model development as part of a virtuous cycle.

However, consumability is key to democratizing Big Data across the enterprise. Regardless of the processing potential of the platform, a flattening of the time-to-analysis curve requires a rich set of accelerators, a vast library of analytic functions, and a tool set for enhancing the reporting and visualization process.

The UniConnect platform undertakes the computationally intensive activities.

The UniConnect platform offers analytics users the flexibility to mine their hybrid data platforms by using their own preferred mechanisms for model creation and visualisation. This includes packaged applications, open source libraries of "parallelizable" algorithms, or an entirely customised approach using procedural languages. By deeply integrating with commonly available analytic packages, the UniConnect platform undertakes the computationally intensive activities from those packages, such as model scoring, while providing a framework for developing additional algorithms. It also supports the visualisation and publication of analytical results in an intuitive and easy-to-use manner.

6. Governance

Over the last few years, the information management community has made enormous progress in developing sound data management principles. UniConnect enables these principles to be applied to a hybrid data platform, including the policies, tools, and technologies for data quality, security, governance, master data management, data integration, and information life cycle management. These are required to establish veracity and trust in the data regardless of its source, and are extremely critical to the success of any analytics program.

7. Building a Hybrid Data Platform

To become more agile, but also open up new business opportunities, it is now necessary for enterprises to move beyond their EDWs without necessarily abandoning them altogether. Enterprises are therefore exploring the advantages offered by cloud platforms and Big Data ecosystems, alongside next-generation technologies such as distributed loading, parallel processing and NoSQL databases.

8. Platform architecture

A small-scale MPP engine can reside on premise for the calculation and aggregation of high volume analytical queries.

To implement a hybrid data platform, a Hadoop-based data storage platform, complemented by an MPP database, can be used. Built on the 'shared nothing' concept, MPP allows efficient queries involving large tables. The entire operation comprises three steps: optimization, consolidation of results, and administration or coordination of activities among nodes. It enables real-time balancing of queries across nodes where large tables are split and stored. Each node processes the locally stored data, and coordinates with other nodes to consolidate and return the results. It runs every operation in parallel to provide the scalability required for Big Data systems.

As a result, the fault tolerance and scalability achieved with MPP is very high. And while Hadoop can handle much of the workload on premise, pre-calculated and aggregated data can be hosted on the cloud in an MPP database. In some cases, a small-scale MPP engine can reside on premise for the calculation and aggregation of high volume analytical queries, before it is uploaded onto the cloud for consumption by business users.