Percipient™
Connecting the world's data

Big data easily, efficiently, affordably

The UniConnect platform is designed to unify data in a highly scalable and seamless manner, by building on an organisation's existing tools, processes and skills. It enables organisations to meet their most pressing data challenges, including those of cost and inefficiency, while ensuring that they are future-proofed for the revolutionary potential that big data can bring.

A Percipient Technology White Paper

Author: Ravi Shankar Nair
Chief Technology Officer
Updated May 2017

Content

UniConnect 2.3

Introduction

As many organisations are discovering, big data does not start with skilled data scientists or sophisticated algorithms. The journey to implement real time 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, normalised, stored in a single repository, and made queryable via a common interface. In this way, DWHs make it possible for organisations to have a unified view of their data across heterogenous sources and formats.

Costs and Copies Abound

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

The deployment of a seperate system to house data that is continually expanding costs organisations from hundreds of thousands to several million dollars to maintain.

These skyhigh costs are further exacerbated by data management processes that often require data to be copied many times. For example, disparate data from multiple systems, once extracted, is copied to a staging area where it is transformed and aligned before being copied again and stored in the DWH. Besides the additional disk space needed, this replication process adds to the problems of latency, complexity, failure and security risks.

More recently, instead of a physical data warehouse, many organisations have turned to cloud storage services. By leveraging a distributed and virtualised cloud computing infrastructure, these services enable organisations to pay only for the storage they actually consume. Although not necessarily cheaper than traditional DWHs, cloud services are sometimes preferred because they represent an operating, rather than capital, expense. Importantly, cloud services result in data copies being stored in even more locations, potentially worsening the problem of unauthorised access.

Scalability Is Key

While high data storage costs have always been a problem, the extreme proliferation of data over the past decade has caused this to reach crisis proportions. Due to ongoing advancements in digitisation, data is now estimated

to be growing at 2.5 exabytes (2.5 billion gigabytes) per day, which means 90% of the world's data was generated only in the last two years.

At an enterprise level, the State Bank of India, for example, currently generates about four terabytes (4,000 gigabytes) per day, and Walmart's one million online customer transactions every hour has resulted in over 2.5 petabytes (2.5 million gigabytes) of stored data. However, newer forms of unstructured data is clearly outpacing the growth of traditional structured data. A Forrester survey of large companies (1000 or more employees) found that while traditional structured data grew by 15% in the 2010 -2012 period, unstructured data from enterprise content management grew by 54%, and digital and web content repositories grew by 56%.

This data architecture is capable of being 'scaled-out' rather than 'scaled-up', thereby enabling large amounts of data to be processed without an associated rise in costs.

In 2006, Apache Hadoop, an open source project offering distributed storage and cluster computing, burst onto the stage. Hadoop's MapReduce programming model enabled very large amounts of data to be processed parallelly, while its Hadoop Distributed File System (HDFS) enabled data to be stored on commodity machines. This data architecture is capable of being 'scaled-out' rather than 'scaled-up', thereby enabling large amounts of data to be processed without an associated rise in costs. Savings relative to a traditional DWH are estimated to range from 10 to 100 times.

The Need For Speed

The need to conquer data volume has been accompanied by the need to conquer processing performance. Given the complex ETL 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, rather than feeding data to the DHW on a continuous basis, data must be reconciled and then loaded into the DWH according to a scheduled downtime, typically overnight.

Spark is today used by some of the world's largest companies, including Alibaba, Amazon, Elsevier, eBay, TripAdvisor, and Samsung.

While Hadoop has revolutionised the ability to store and process very large volumes of unstructured and provisional data, its MapReduce model also relies on data being processed in batches rather than as a continuous stream. Therefore, while Hadoop has greatly enhanced the efficiency and fault tolerance of 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.

It was amid these speed challenges that the IT world welcomed Spark, created by Matei Zaharia at UC Berkeley's AMPLab, and open sourced in 2010. Billed as "lightning-fast cluster computing", Spark is said to be 100 times faster than Hadoop and in 2014, formally demonstrated its ability to sort 100 TB of data in 23 minutes using 206 nodes, officially a new world record. While Spark users continue to face memory problems, especially when attempting more complex join queries across multiple data sources, the advancements brought by Spark have overturned

traditional benchmarks for data processing. Spark is today used by some of the world's largest companies, including Alibaba, Amazon, Elsevier, eBay, TripAdvisor, and Samsung.

Other Stumbling Blocks: Data Variety, Platform Lock-ins, and Data Security

The innovations brought by Hadoop and Spark, together with a host of other open source software releases over the past five years, have transformed how organisations view data. Data today is potentially abundant, varied, and fast, and capable of powering business and public sector applications in ways that were previously unthinkable.

However, pieces of the puzzle are still lacking, and these typically come to the fore only in the process of implementation. Importantly, the task to integrate heterogeneous data across systems and innovations remains a challenging one. For example, while Hadoop solves the problem of big data storage, organisations are unable to jointly query the data stored in Hadoop and their existing databases, without resorting to yet more data replication.

Big data innovations also often fail to make the final cut due to platform lock-ins. Organisations find it difficult to introduce new technologies given the need to introduce new skills, enforce changes to existing processes and tools, and overcome challenges to interoperability. So-called thick enterprise stacks (ie applications and databases offered by a single vendor) and Enterprise Level Agreements also limit the short term financial benefits of an alternative platform.

Finally, the amassing of so much data intensifies the need to tighten up on data security and authorization. Meanwhile, data privacy and ownership concerns have hindered the sharing of data across organisations, and the tapping of externally-sourced (eg social media) data, despite the potential to enhance customer insights and to re-imagine the entire customer experience.

UniConnect: Next-Gen Data Integration

Advanced Features

Percipient's UniConnect platform amalgamates, integrates with, and enhances select open source software, including Hadoop, Presto, Spark, Kafka and Cassandra, to deliver four key capabilities:

- **Data connectors**

  UniConnect enables interactive analytical querying of data via a connector. In this way, multiple data tables, regardless of origination, format or type, can be accessed via a single UniConnect query. To date, the platform offers over 250 connectors for connection to almost all standard databases, transaction systems, and real time sources including IOTs and web streaming.

- **In-Memory Processing**

  UniConnect 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 slow down processing. In memory processing means not only that data can be queried live and on-the-fly, but is also more secure as data is not physically copied.

- **Distributed data**

  In UniConnect, queries are parsed and then distributed among a cluster of compute nodes, rather than relying on a single centralised server. These nodes process their data in parallel, rather than sequentially. The outcome is then reassembled for the final result. Distributed data processing delivers cost savings, is more reliable, and offers high performance even for large data volumes.

- **SQL-based queries**

  UniConnect enables heterogeneous data can be queried using the recognised 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.

Here is a 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

Designed for Analytics

Drawing on these capabilities, UniConnect offers organisations an advanced and efficient discovery platform with which to implement business intelligence, financial and risk reporting, customer 360 campaigns and many other analytics-powered digital applications, without the need for a costly DWH.

UniConnect provides organisations with unparalleled flexibility to query as much data and as frequently as required

By doing so, 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 infrastruture, UniConnect provides organisations with unparalleled 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 eg Spark, R and Weka, as well as proprietary tools eg 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.

Ultra-Fast and Secure

The UniConnect 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 rougly 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 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

The memory needed for such jobs is not reliant on a single server. Rather, parallel distributed processing via basic ethernet networking infrastructure makes it easy to partition a large dataset 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 platform was shown capable of the same query in 8 minutes, using 4 commodity machines which cost less than USD 10,000 to acquire.

However, UniConnect's business value lies not just in the increased ROI (by enhancing the performance of the tools that organisations 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 does 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.

In addition, the Uniconnect platform has inbuilt features to support only authenticated and authorised access. These controls take place at three levels, ie who, what and how:

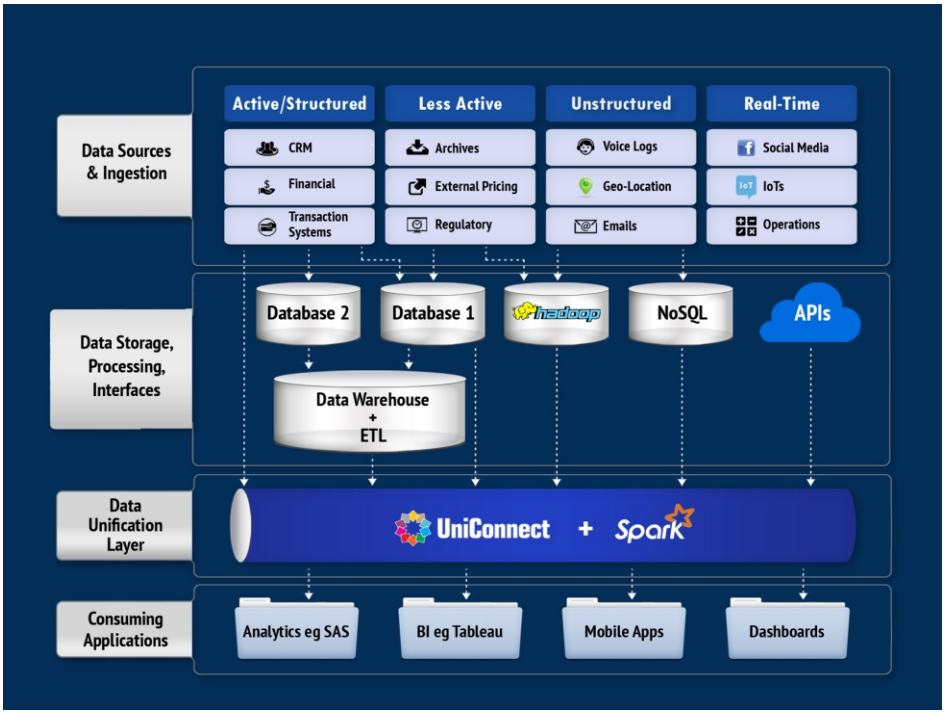
- Who – user access is restricted by integrating with an enterprise's existing LDAP server. LDAP protocols are the enterprise standard for user authentication. Where more fine-grained users authentication is required, UniConnect is also integrated with the open source Shiro framework that recognises 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 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 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.

Seamless Deployment

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

Instead, 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. 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, as shown in the diagram below.

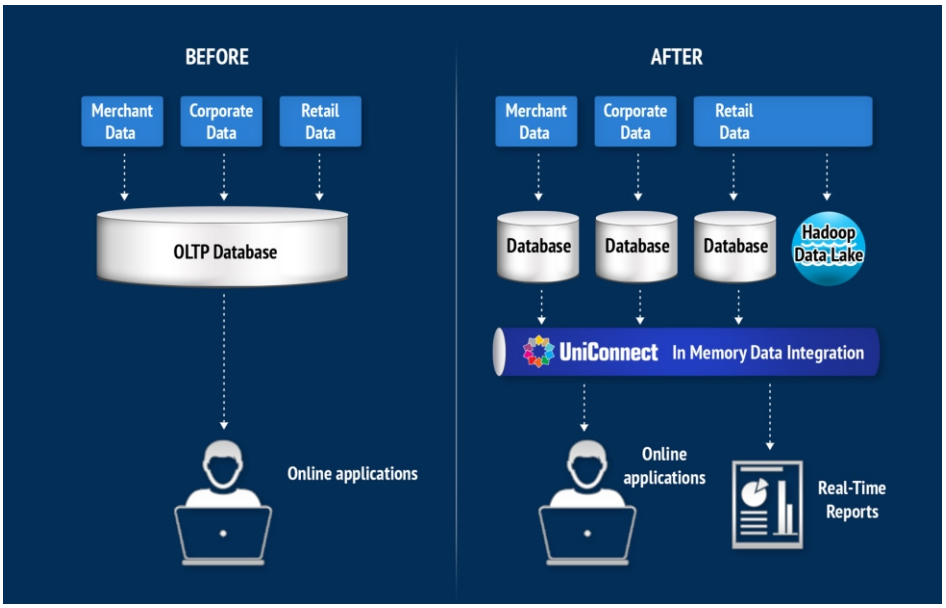


This means that there is almost no data that an organisation cannot unify. The UniConnect discovery layer does so by untangling complex enterprise data architecture from a spaghetti orientation to a data flow that is simplified, well-organised and effective. By doing so, UniConnect can help organisations construct real time management dashboards, deliver elaborate regulatory reports, effect highly personalised customer services, and build highly impressive customer-facing applications.

These require substantial processing stability in the face of large and ever-growing data volumes. UniConnect has been vigourously tested to ensure such stability. 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 installationof 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.

Case Study: More flexible and cost effective data storage

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 200GB 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 inter-dependencies affecting the individual lines of business.

By deploying UniConnect, it is possible to seperate the data into different smaller instances without sacrificing the bank's data integration capabilities. Similarly, given the need to archive data for regulatory reporting, a Hadoop Data Lake can be included within the bank's IT infrastructure, while ensuring that this data continues to be easily queried alongside the bank's existing, more active data.

This solution is able to deliver significant cost savings, process efficiencies and better security against hacking risks. The higher performance also means the bank is able to use the data for other purposes besides its existing web applications, such as real time branch level reports.