intertrust®

Governed data analytics environments on virtualized datasets

Guide

Platform

Data warehousing as a service

There is a lot to sift through when evaluating vendors offering data warehousing-as-a-service. Let's take a look at what is being promised, what is being delivered, and how it stacks up to a modern approach like the Intertrust Platform.

Some businesses claim to have created "completely new SQL data warehouses" to differentiate themselves from conventional data warehouses, by offering data warehousing as a service (DWaaS). Data warehousing-as-a-service proponents claim legacy data warehouses are too complex, too costly, and inflexible. DWaaS vendors also claim that 'Big Data' platforms such as Hadoop are merely toolkits, not complete solutions. They say these platforms require massive resources to build and maintain and are even more complex, require specialized skills, and weren't designed for data warehousing.

To solve these issues, some say they have designed unique architectures to preclude the limitations of existing architecture and software products. A few forgo on-premises solutions altogether, claiming they are too complex or too expensive, and instead offer it as a cloud-based service.

Data virtualization and data lakes

Data virtualization represents an agile approach to data integration. It provides the data consumer an abstraction layer that hides most of the technical aspects of how and where data is stored, processed, and how it's accessed.

It allows accessing resources without delving into details like:

- Where data is stored
- What technology or platform is used to store data
- What technologies are used to process and store data, what interfaces are needed to access data

Data virtualization allows integration of data from various sources, keeping the data where it is. Reports and dashboards can be created to gain business value from the data without moving it (and adding cost and risk). It is an alternative to ingesting data into a data warehouse, where data is literally collected from various sources and a copy of the data gets stored in a new location or data store.

The main advantage of data virtualization is time-to-market: a solution can be built in a fraction of the time it takes to set up a new data warehouse. This is because you don't need to design and build the data warehouse and the Extract, Transform, Load (ETL) operations required to copy the data into it. It also doesn't require as much testing. Copying the data means more hardware costs, more software licenses, more ETL flows to build and maintain, more data inconsistencies, and more data governance costs, so using data virtualization can also save a lot of money.

A data lake is a repository that holds a vast amount of raw data in its native format until it is needed. It's a good option for storing, exploring, experimenting, and refining data, in addition to archiving data. Data lakes are becoming popular as there is a vast number of fast-growing data sources, from social sources to weather to IoT/sensor data. Some characteristics of a data lake include:

- A place to store unlimited amounts of long-term data in any format inexpensively
- Allows for easy integration of structured data, semi-structured data (e.g., XML, HTML), unstructured data (e.g., text, audio, video), and machine-generated data (e.g., sensor data)
- A way to describe any large data pool in which the schema and data requirements are not defined until the data is queried: 'just in time' or 'schema on read'
- Complements an enterprise data warehouse (EDW) and can be seen as a data source for the EDW—capturing all data, but only passing relevant data to the EDW
- Allows for data exploration to be performed without waiting for the EDW team to model and load the data

Secure data virtualization

Using secure data virtualization, the Intertrust Platform ensures there is no movement of data, adding value to any kind of data management asset (be it an on-premises data warehouse, or a cloud-based data management platform).

The Intertrust Platform works agnostically to provide an end-to-end governed data asset. There is no need to aggregate data in a data lake or a warehouse. Transferring data is when you incorporate risk and costs, because cloud services charge hefty data egress fees. Add to this the cost and complexities involved from extensive ETL (Extract, Transform, Load) work required, and this turns into a very expensive problem indeed.

Data management, or getting data ready to make it analyzable or actionable itself is part of the problem. Large enterprises do not have the time nor should they be spending any of it on the precursor to data management: the act of just getting their data organized, scrubbed, and ready.

Intertrust's approach is to instead eliminate the need to consolidate data. Far more efficient is to query, process, and analyze the data where it already is, in place, while making sure that the data is processed in a controlled environment and cannot be copied.

Let's look at how the Intertrust Platform is different from a typical or even "brand new" data warehouse using three different parameters: data orchestration, security, and analytics.

Intertrust Platform™

The Platform leverages container orchestration technologies such as Kubernetes and Docker to make deployments cloud-agnostic.

Identity and access management

Device and user identity, authentication, and authorization; maintains platform objects and their relationships.

Secure execution environment

Secure network-isolatable environments for workload execution and controlled, interactive data exploration.

Data virtualization

Data object definitions, permissions, restrictions. Provides data interfaces, manages DBs and virtualized datasets.

Time series database

Scalable, efficient, high performance database designed for time series data.

Data Orchestration

Even a modern data warehouse is essentially a database that requires data to be ingested before it can be used. In other words, data has to be moved first into the warehouse. The Intertrust Platform operates with existing data stores using data virtualization, without moving the data. For companies seeking to modernize their existing data warehouse assets with modern features like dynamic rights management, security, and multi-party data integration, this is a compelling alternative. The Platform provides governance and fine-grained access controls on top of an existing data warehouse, enabling the option of moving away from data warehouses completely in the future, if required.

Modern cloud-based data warehouses support deployments in multiple cloud platforms like Azure, AWS, or Google Cloud, but do not support on-premises deployments. The Intertrust Platform supports both on-premises and cloud-based deployments.

All cloud infrastructure providers make it very easy to move data into the cloud, but very expensive to get it back out again. If there is a need to export large amounts of data to a different cloud or local processing environment, things will become very expensive very quickly. With the Intertrust Platform, enterprises do not have to export data for every action performed on the data, saving money.

More importantly, moving data to a proprietary format in a cloud service also means taking the business risk of having the entity managing the service going out of business or getting acquired, etc. Copying and consolidating all data in a data lake or a data warehouse creates delays and often needs ETL processes or ETL system deployments. These systems add cost, complexity, and require ongoing operational support. This happens whenever a data source is updated or the schema changes.

It is very difficult to get point-in-time consistency across a number of databases when copying things around. As an example, a retailer might be working with two different datasets. The first dataset, 'Customers,' is a table of all customer records: customers' names, addresses, and credit card information. The second dataset, 'Orders,' is a table containing details on all customer purchases. At the end of the business day (or at other defined intervals), these tables would typically be uploaded to the cloud, in a particular order, to be backed up and for further usage by other cloud-native applications.

It might take an hour for a copy of the Customers table to be uploaded to a data warehouse (or data lake). After this first operation is completed, the routine would call for an upload of the Orders table. This time lag creates the possibility of error. For example, if a new entry were to be made in the retailer's master Orders table during the Customer table copy process, that could lead to a situation within the data warehouse where the two tables did not match. The Orders table subsequently uploaded to the data warehouse/lake would contain references to a customer that did not exist in the customers table in the data warehouse/lake. If a report or query were run on the records in the data warehouse, the results would be wrong.

This is a simplistic case. In reality, you could have multiple tables with dependencies on other datasets; they could also be really large. A manual process to fix all this becomes very complex indeed. The Intertrust Platform uses Secure Data Virtualization instead. Virtualizing data creates a dynamic abstraction layer between the data source and the consumer. This enables and allows changes to be made in a schema on the data side without affecting the consumer, simply by modifying the abstraction instead.

Data virtualization has many immediate benefits. Using data virtualization enables moving data to more suitable data storage or database technologies, also without affecting consumers. It can do this seamlessly, by simply changing the data source definition to read from a new source. Virtualizing data also makes it immediately available for analysis: data can be organized dynamically, again without the need for predefined schemas. In situations where a data warehouse or a data lake is necessary, Intertrust can support these via data virtualization by treating them yet as another data source within the system.

Finally, Secure Data Virtualization allows the integration of security without affecting existing systems. The Intertrust Platform securely and seamlessly fits between existing systems, as the input and output protocols and formats can remain the same as before, but with an added security layer. The Platform also provides a secure execution environment. By offering integrated, secure computing capabilities, Intertrust allows advanced analytics development without copying or moving data and does not require trusting partners or vendors with your data.

By offering integrated, secure computing capabilities, Intertrust provides a secure execution environment that allows advanced analytics development without copying or moving data and does not require trusting partners or vendors with your data.

Security

Many data warehouses have fairly standard authorization systems that are similar to those found in relational database management systems—table level access control for tables created in their warehouse. Because it is often useful to have data in a single table, the Intertrust Platform provides row-level access control, but allows some users to access only some parts of it. An example would be a table that lists all the customers serviced by an energy utility, throughout a state or over many cities. A user inside that utility may only be authorized to see the users in a specific city, appropriate for their role only, and nothing more. A data warehouse would not be able to support such a use-case. Intertrust does.

Data warehouses manage security within their databases. Intertrust centralizes governance over any connected data source, including data warehouses, allowing an organization to manage data access with fine-grained permissions across its entire data architecture. With the Intertrust Platform, you can govern data where it resides, without the data warehouse overhead of ETL workloads and rigid predefined schemas. Furthermore, virtual datasets can be composed on demand from data within multiple data sources, allowing you to respond quickly to new downstream requirements from analysts and data scientists.

Analytics

Some data warehouses advertise that they enable "sharing" of data. The Platform works on the principle that data does not need to be shared; instead, workloads may be deployed within a secure execution environment. The moment you let someone transfer data out of your systems, you lose control over it. The Intertrust Platform precludes that problem by not moving data in the first place. Some data warehouses operate by moving the data to the analytics. The Platform operates on the principle of moving the analytics to the data. As datasets grow, moving them becomes expensive and time-consuming. Most data warehouses do not have the concept of an analytics sandbox, nor do they allow running untrusted code securely, while guaranteeing no data leakage.

Lastly, some data warehouses claim they enable "data exchanges, without moving the data." What they actually do is create a database with some ETL capabilities that ingests data from various sources, to which they add querying capabilities: they enable running a SQL query against the datasets.

Capabilities summary

Data warehouse as a service	The Intertrust Platform
Copies data	Virtualizes data
Ingests and copies data to a warehouse	No copying of data
Traditional data warehouses are deployed on-premises DWaaS deployments are cloud-based only	Can be deployed on-prem or in the cloud
Risk of vendor lock-in Difficult and costly to migrate	No lock-in risk Migrate at any time
Data egress introduces costs and risk of data loss	Doesn't require data egress and no risk of data loss
Coarse-grained access control: table-level	Fine-grained access control: row and column-level
Permissions management on own databases	Permissions levels on existing databases or on big-data tools. For example, Intertrust can implement access controls on HDFS or on No/SQL databases.
Moves data to the analytics	Moves analytics to the data
Cannot provide isolated, secure, governed execution environments/ sandboxes	Provides isolated, secure, governed execution environments/sandboxes
Data exchanges require all data producers to load their data into the warehouse	Data exchanges can be implemented across existing data producer data stores

The bottom line

Businesses today depend on data to improve workflows, run more efficiently, and discover new opportunities. They need secure, governed access to diverse data ecosystems to scale and remain flexible as markets become increasingly data driven. Using secure data virtualization, a fully integrated, modern, governed, efficient time-series data store, and a secure execution environment that brings analytics to the data, the Intertrust Platform facilitates multi-organization collaboration so you can easily share data, whether stored on-premises or in a cloud service.

Unlike cloud-based data warehouses or data lakes, Intertrust eliminates the need to copy or move data, ensuring data always remains protected. There is no ingestion of large amounts of data as a precursor to the setup of operations. This translates into savings on storage, processing, maintenance, and operations on data.

intertrust®

Building trust for the connected world.

Learn more at: intertrust.com/platform
Contact us at: +1 408 616 1600 | dataplatform@intertrust.com

Intertrust Technologies Corporation
400 N McCarthy Blvd, Suite 220, Milpitas, CA 95035