

**Whitepaper**

# Implementing Virtual Databases in Your Enterprise Data Warehouse (EDW)

[sales@astera.com](mailto:sales@astera.com) | 888-77-ASTERA

**Astera**  
Enabling Data-Driven Innovation

# Table of Contents

1. The Evolution of the Traditional EDW	01
2. Traditional EDW Limitations	02
3. How Virtual Databases Enable the Modern EDW	03
4. Managing Virtual Databases in Astera Data Virtualization	05
5. Database Caching in Astera Data Virtualization	08
6. What's Next for Astera Data Virtualization?	16

# The Evolution of the Traditional EDW

In the modern operating environment, every organization relies on accurate and timely reporting to drive decision-making across its business processes. From sales and finance to marketing and HR all business functions produce a wealth of data that must be collected and analyzed on a consistent basis so that emerging opportunities and potential vulnerabilities can be handled in a proactive manner. However, as the volume of enterprise data has grown, so too has the complexity of BI infrastructures.

Today, business users draw insights from a wide variety of sources, including relational databases, NoSQL systems, cloud applications, social media, spreadsheets, CSV files, and more. In most cases, the sources are scattered across disparate purpose-built repositories that are only accessible to a small segment of employees. These data silos are often a source of redundancy and inefficiency as the same data may be stored in multiple systems.

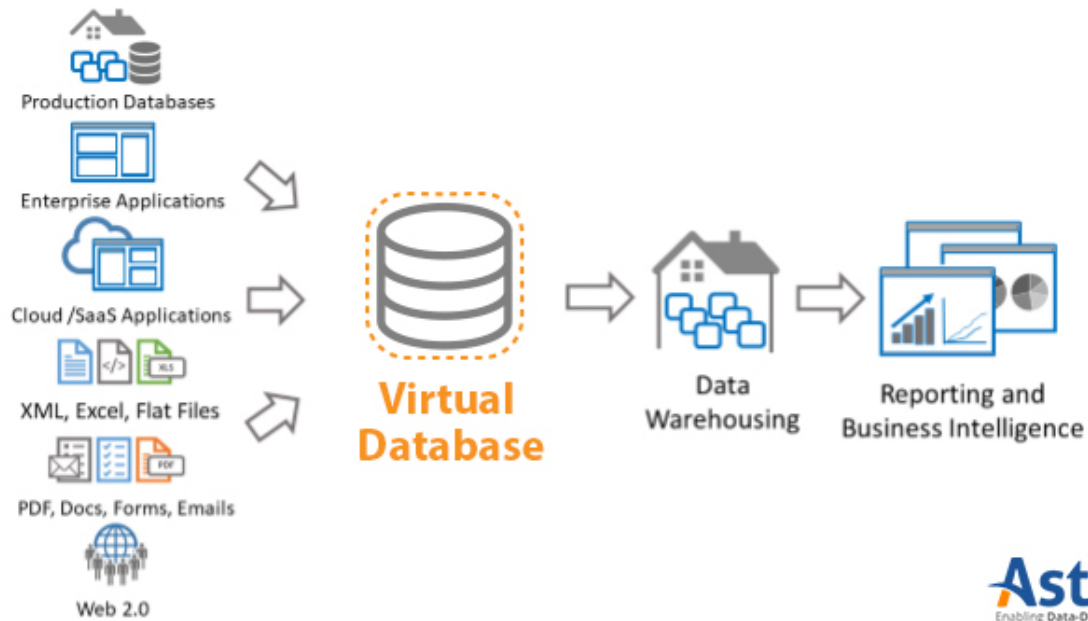
To make matters even more complicated, different versions of a dataset may be used by various departments leading to inconsistent decision-making due to the presence of outdated or inaccurate records within their systems. Indeed, the segmented nature of many of these BI infrastructures can keep key information that would be useful to a particular group of users hidden from them altogether.

Traditionally, organizations have looked to resolve these issues by using ETL tools to extract their disparate data and load it into physical data marts that feed an enterprise data warehouse (EDW). However, there are a few drawbacks to this approach.

## Traditional EDW Limitations

- Physical data marts are expensive to set up and maintain. First, all sources for the data mart must be identified. Next, a separate database server needs be allocated to store relevant data. The database server and storage structure must be optimized and tuned for performance. Finally, backup and recovery processes must be implemented to ensure continuity in the event of a disaster.
- If additional data is required for reporting and analytics purposes, these databases will need to be unloaded, repopulated, and re-tuned. According to The Data Warehousing Institute (2016 BI Benchmark Report), it takes organizations an average of 7.1 weeks to add a new data source to their data warehouses, and 27% of the organizations surveyed take more than 9 weeks for the same task.
- When updates are made to source systems, the changes must be replicated in the data mart. This is a continuous process that often results in several copies of the same data ending up in the warehouse. Unless strict policies for synchronization and data consistency are maintained, many of these replicated records will become outdated, further hampering the quality of data in the data warehouse.
- The EDW generally follows a load first mentality that emphasizes data storage with little consideration given to the relevance, purpose, formatting, and security of this data.

# How Virtual Databases Enable the Modern EDW



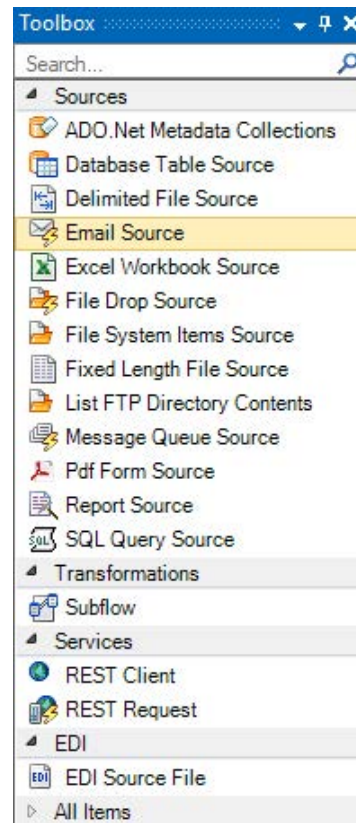
Data virtualization derives data from multiple sources in a variety of formats without any kind of replication. An intermediary virtual layer is introduced between the source systems and the EDW. This layer functions as a virtual DBMS that can access and view data from any source system as if they were all contained within a single entity.

In such a deployment, data sources are simply plugged into the virtual DBMS and can be made available for querying in minutes. With virtualization in place, consumers no longer need to develop complex scripts to search physical repositories for the specific records they need, nor must they secure credentials for each database that they need to view. Instead, any queries made against the virtual database will be matched to the source system that contains the requested data, and the resulting views will be showcased in the virtual layer without any physical movement of data (unless caching is employed).

This approach makes the traditional EDW far more agile, while still providing necessary accessibility to users across the enterprise. Additionally, advanced virtualization solutions such as Astera Data Virtualization offer in-built dataflow and workflow validation capabilities which allow administrators to ensure that all data is extracted, processed, loaded according to business requirements. These tools also enable administrators to assign permissions for source databases based on user roles, so that each individual's views are limited to what is relevant under the scope of their responsibilities.

# Managing Virtual Databases in Astera Data Virtualization

Using Astera Data Virtualization, users can bring together heterogeneous data from a diverse array of sources and present them in a single virtual database for querying and analysis.



*Supported Sources in Astera Data Virtualization*

The first step to building such a model is identifying the data marts that are currently in use across your existing architecture and the systems that it draws from.

Astera Data Virtualization supports drag-and-drop integration of a wide array of sources including relational databases (sourced from one or more servers), Excel workbooks, delimited files, PDF files, emails, as well as web services delivered through REST and SOAP APIs. When brought together, these internal and external streams of data will act as a single source of truth for reporting and analytics purposes.

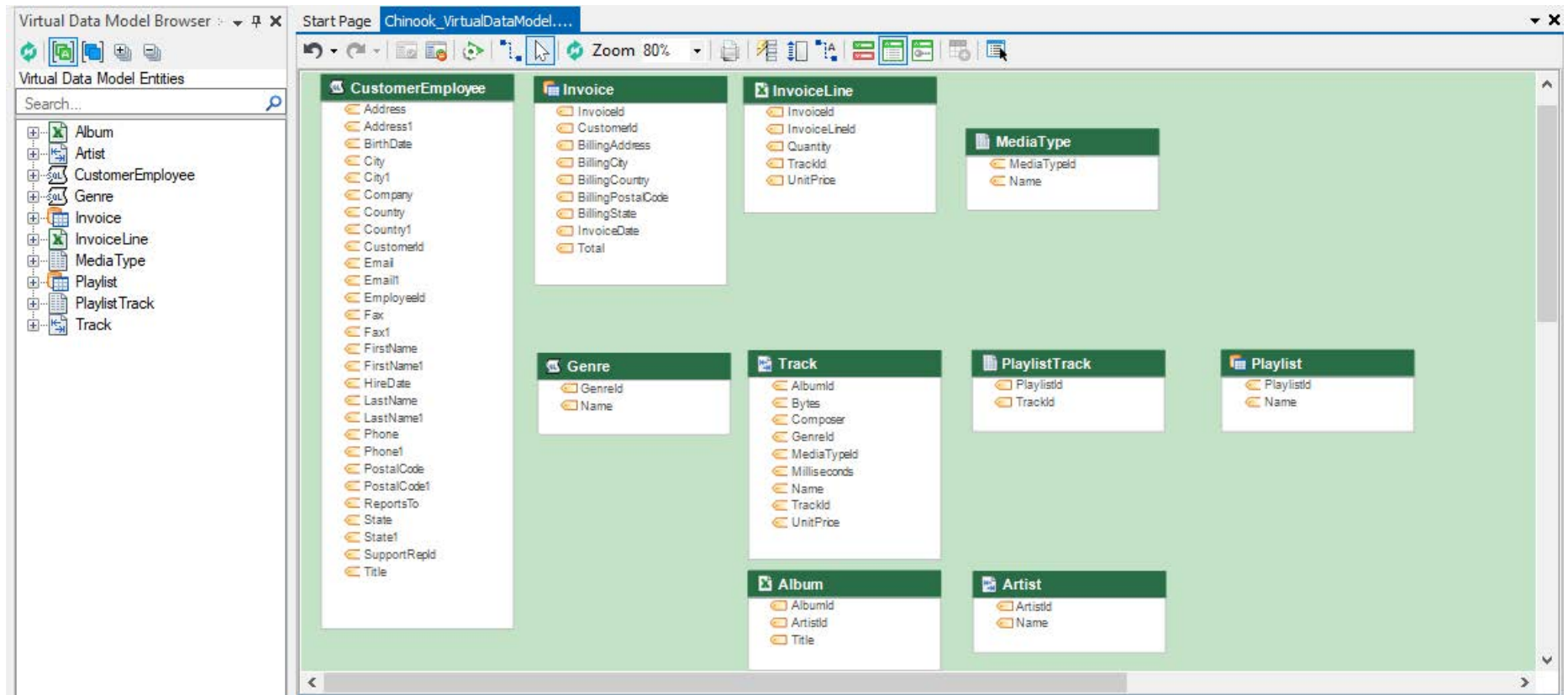
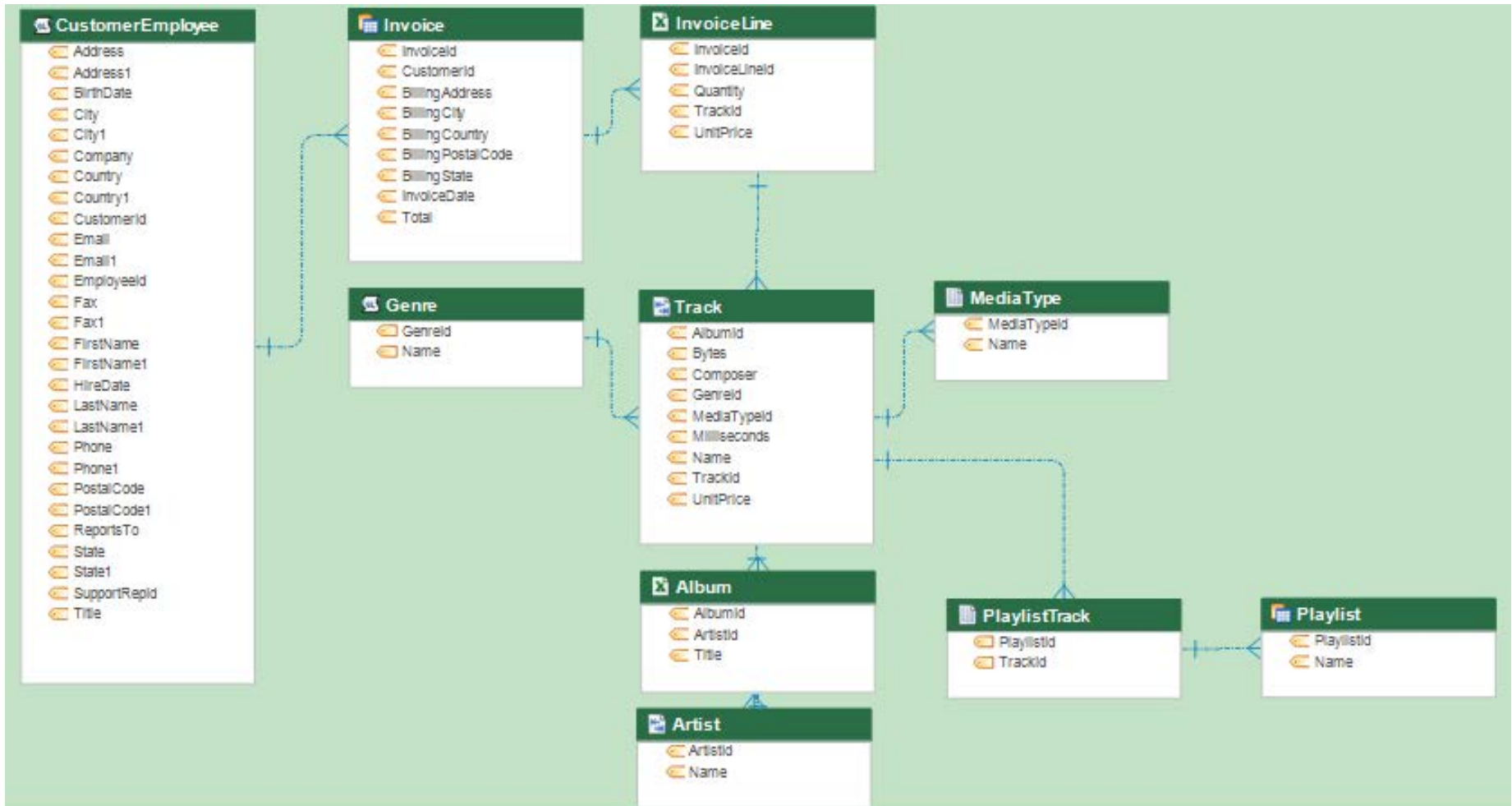


Fig. 1



Once sources have been brought into the virtual data model and appropriate parameters have been defined for each one, we can start connecting tables based on shared fields using the linked nodes setting. These reference relationships will improve query performance in the data warehouse.



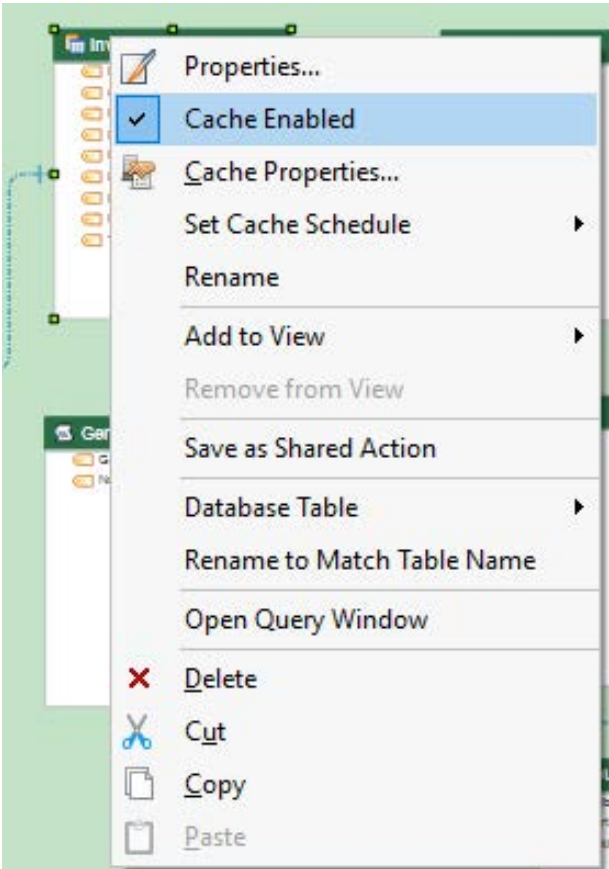
*Linked Nodes Applied to a Multi-Source Table*

# Database Caching in Astera Data Virtualization

Another key aspect to performance tuning your virtual database is implementing caching across the abstraction layer. This process creates a temporary copy of source records in server-accessible database, which in turn saves analytics and reporting systems the cost of retrieving and recomputing data each time a query is executed. Caching provides several benefits for the virtual database:

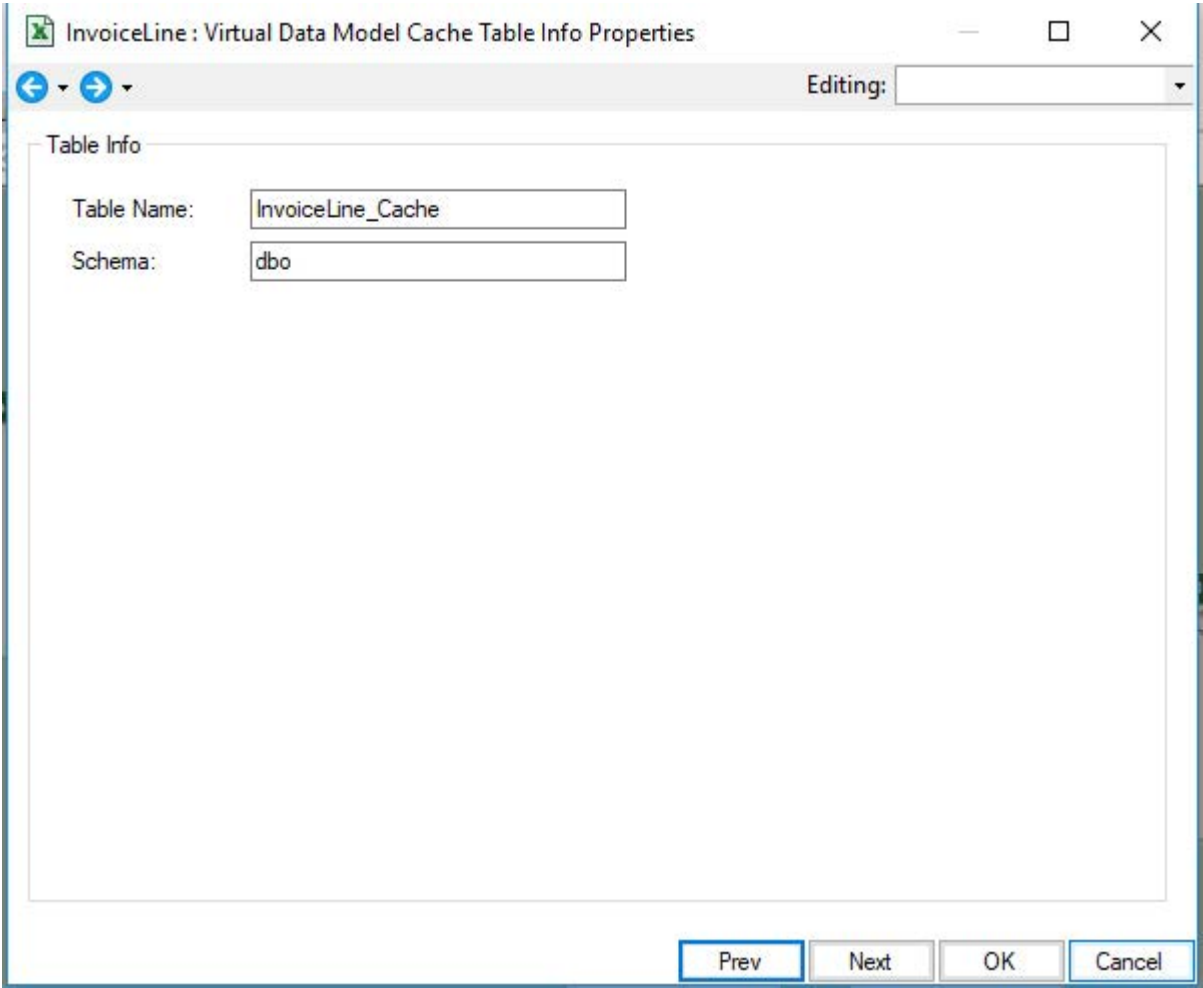
- Source data is housed across disparate systems in a variety of formats some of which are built for efficient query performance (relational databases) and some which are not (flat files, web applications). Without caching, any queries that seek to bring together data from slower data sources will take longer to run and may consume more server resources, if data needs to be parsed before it is brought into the virtual database. By caching these sources, you ensure that the cost of data retrieval remains consistent.
- Certain types of queries may occur more frequently than others. For example, in the model shown above, an executive may require reporting on their most successful genres and artists on a monthly basis. In such cases, sales performance for each business unit can be aggregated and stored as a derived view in the cache to enable more efficient query performance.
- If a large number of query requests are made to particular source systems, they may become over-utilized. This could compromise their ability to quickly process and record operational transactions. Caching can prevent these issues from occurring.
- Some of the data contained within the virtual database will be imported from external sources which are not owned or controlled by the business. This raises the issue of availability as external data may not always be available for querying. Again, caching will help to ensure the constant availability of virtual database records regardless of the current state of the source system.

Caching can be enabled through the right-click menu for each source table in the virtual model builder.



**Cache Enable Option**

Next, go to the cache properties and assign appropriate fields for caching as well as the name and format for the cached table.



*Virtual Data Model Properties*

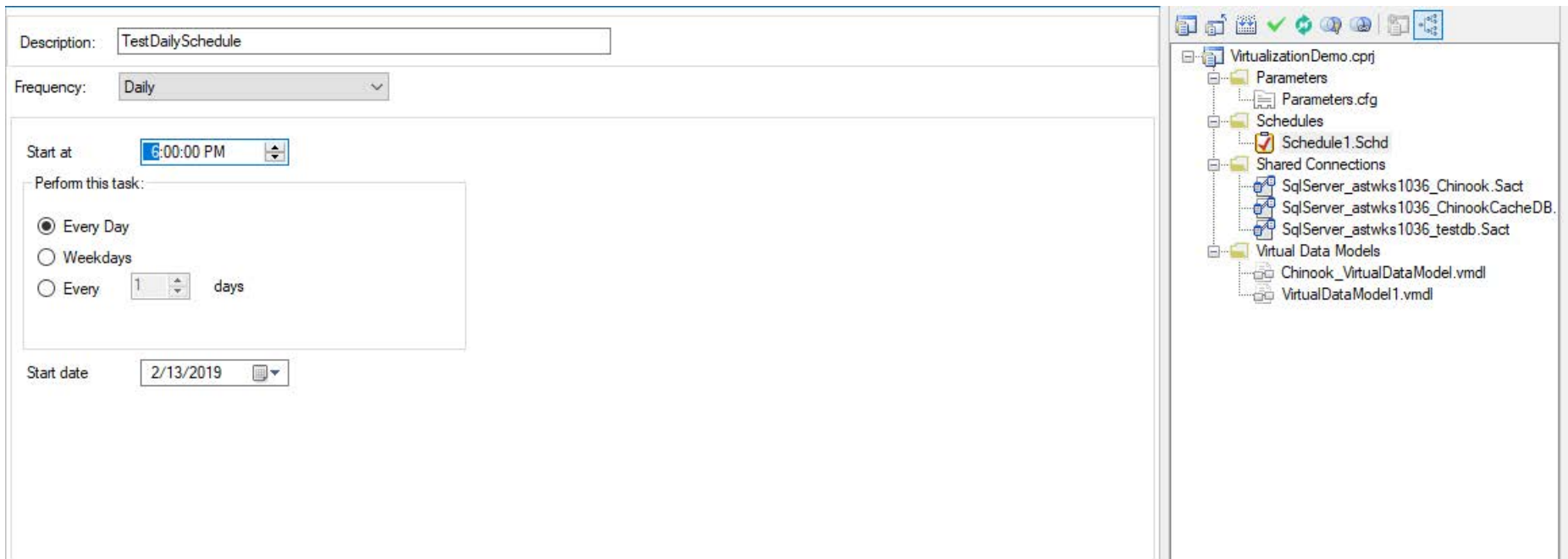
	Name	Column Name	Data Type	Db Type	Length	Prec
▶ 1	InvoiceLineId	InvoiceLineId	Integer	INT	0	0
2	InvoiceId	InvoiceId	Integer	INT	0	0
3	TrackId	TrackId	Integer	INT	0	0
4	UnitPrice	UnitPrice	Real	FLOAT	0	0
5	Quantity	Quantity	Integer	INT	0	0

**Layout Builder for Virtual Data Model Table**

Once the cache layout has been set up, an appropriate cache schedule must be determined for the virtual database. Keeping the cache current is necessary to ensure the accuracy and timeliness of query

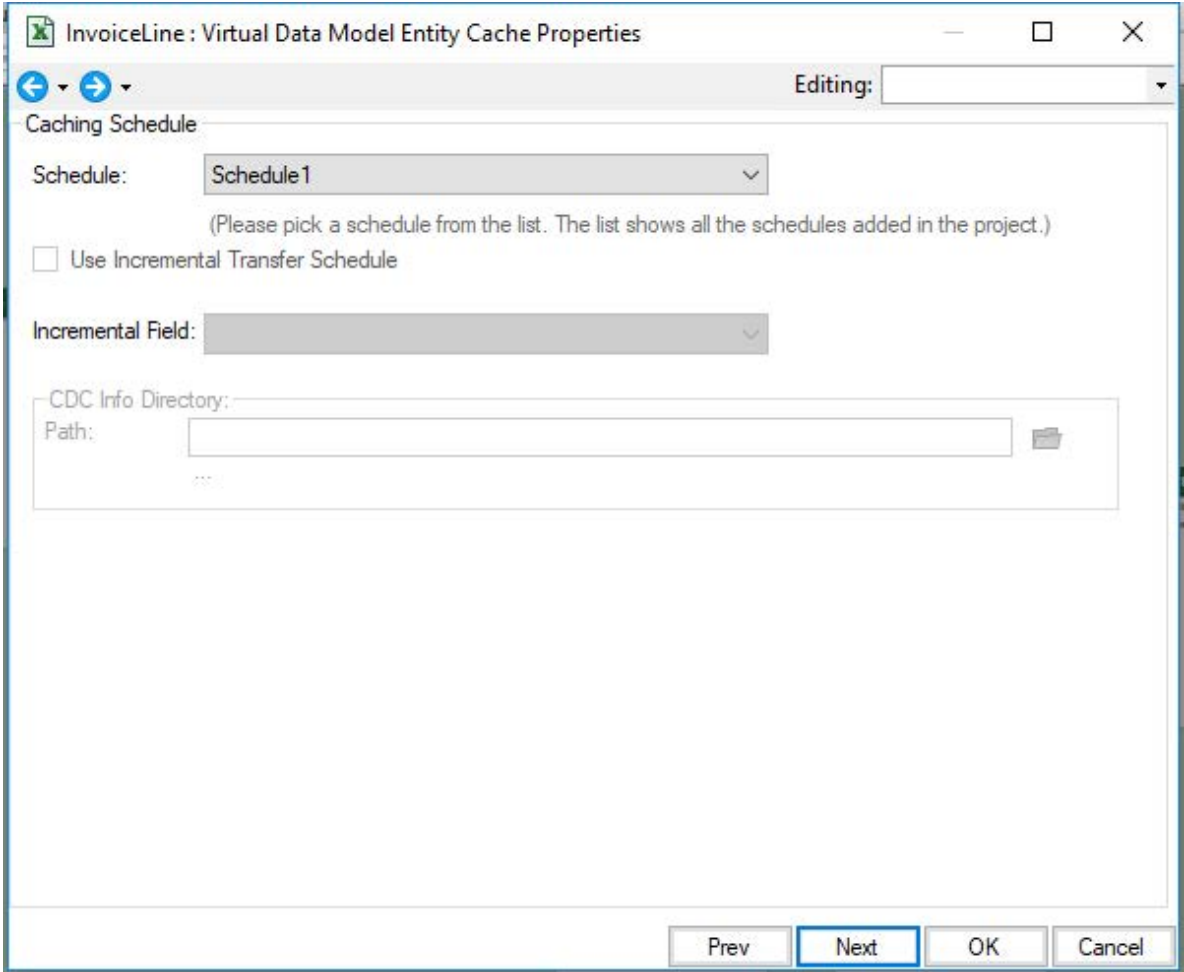
Once the cache layout has been set up, an appropriate cache schedule must be determined for the virtual database. Keeping the cache current is necessary to ensure the accuracy and timeliness of query results. Cache schedules can be built in the schedule builder window which is accessible through the project explorer.

In our example, the company has decided to refresh the cache on a daily basis to keep it aligned with their transactional systems.



**Cache Scheduling in Astera Data Virtualization**

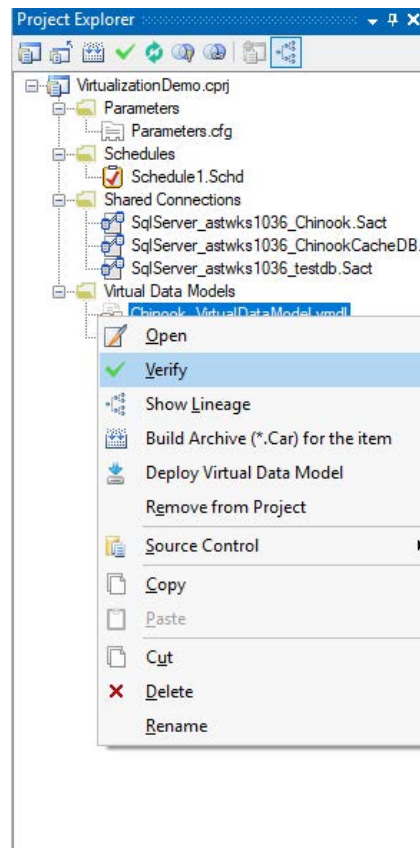
You can then assign this schedule to the relevant virtual database table in its cache properties.



*Incremental Load Setting in Cache Properties*

As you can see, the properties window also offers users the option to implement incremental caching for certain fields. This setting ensures that only data that has been modified or added data since the last load is updated in the virtual database cache. This will save a significant portion of the network load involved in replicating large datasets from the source system. Note that this loading option is only available for database sources.

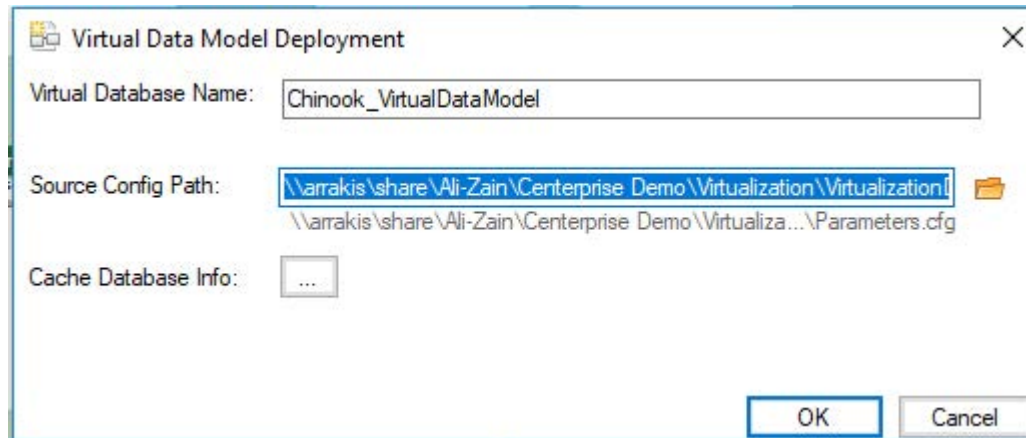
Now that the mapping and properties for the virtual database have been completed, the user can verify their model to ensure that there are no errors.



**Verify Virtual Model Option**



With that confirmed, they can go ahead and deploy their virtual model. At the deployment stage, the user must create a name for their virtual database, define parameters, and assign a database for cached data.



Virtual Data Model Deployment

Virtual Database Name: Chinook\_VirtualDataModel

Source Config Path: \\narrakis\share\Ali-Zain\Centerprise Demo\Virtualization\Virtualization\Parameters.cfg

Cache Database Info: ...

OK Cancel

### *Virtual Model Deployment Settings*

This virtual model can now be used as a source in any data warehousing project in Centerprise's data warehousing component, Astera Data Warehouse Builder.

## What's Next for Astera Data Virtualization?

Over the coming months, we're planning on building on our existing database virtualization tools with additional auditing and security features that will help administrators manage privileges across the virtual layer while monitoring user activities across their data warehouse architecture. For more information on Astera Data Virtualization, contact our sales and support team, or schedule a private demonstration to see our product in action.

## About Centerprise Data Integrator

Centerprise Data Integrator is a robust end-to-end data integration platform that facilitates business users by providing a code-free development environment. The data integration tool is equipped with ETL and ELT capabilities and provides native connectivity to a wide range of commonly used databases.





[www.astera.com](http://www.astera.com) Contact us for more information or to request a free trial  
[sales@astera.com](mailto:sales@astera.com) | 888-77-ASTERA

Copyright © 2018 Asteria Software Incorporated. All rights reserved. Asteria and Centerprise are registered trademarks of Asteria Software Incorporated in the United States and / or other countries. Other marks are the property of their respective owners.