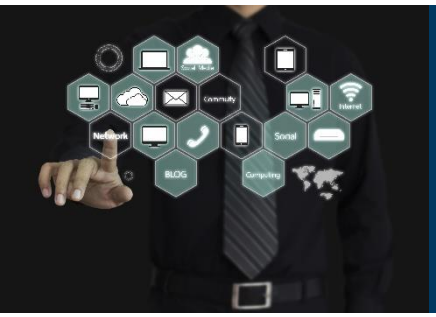
DIYOTTA



Unified Data Integration

Across Big Data Platforms

Whitepaper

DIYOTTA

Table of contents

The Business Problem

In today's increasingly fast-paced business environment, organizations need to use more specialized and fit-for-purpose applications to accelerate deployment of functionality and data availability to meet rapidly changing requirements and voluminous data growth. They also need to ensure the coexistence of these applications across heterogeneous data warehouse platforms and clusters of distributed computing systems, while guaranteeing the ability to share data between all applications and systems.

Imagine that, instead of having disparate data stores of structured and unstructured data, you can implement a logical data warehouse combining traditional data warehouses with Big Data platforms like Hadoop, where the technology is driven by the concept and not vice versa. This is quite important as many times we find enterprises inadvertently locking themselves in to technologies which then drive the concept. Organizations should not be limited to trying to solve all their data challenges with a single technology, and there are many solutions on the market today that provide overarching analytical capabilities across platforms. However there is also the need for a unified bus which can steer through these heterogeneous platforms and provide data movement, transformations, aggregations and consolidation of data in the right place to support timely and effective analysis and insight.

Let’s say you have variety of internal and external data sources to extract from and you store the results in a Hadoop cluster for cost efficiencies and full visibility across your data assets. However, this platform today does not provide the low-latency analytic and reporting capabilities you need to meet business requirements.

As a result, you need the right data on your high performing platforms. If you could take this sea of data and readily perform selection and Map/Reduce functions to transform, segment, and then push your high-value structured data into a MPP Data Warehouse platform, you could meet both the needs. You would maintain the entirety of your data on Hadoop, have the critical information you need in your warehouse, and could also move data not immediately needed on the MPP Data Warehouse back to the Hadoop cluster for archival and high-latency analysis, as well as bring back to the warehouse when and if needed.

What is missing, yet required to make this a reality is to have a comprehensive data integration capability and metadata backbone which understands and can readily communicate with all these components, performing necessary data-related operations in a rapid, unified and seamless manner. It’s not just about technology, it’s about an architectural concept which allows for this seamless data integration across a variety of heterogeneous platforms without introducing its own data latency issues.

As Hadoop is quickly becoming mainstream technology for the ingestion of a variety of data, along with data refinement, analytical analysis and high-volume data storage, it must also work in conjunction with existing data warehouse platforms.

Unified Big Data Integration

Data Integration is a key component of the enterprise business intelligence stack and many organizations which have adopted Massively Parallel Processing based (MPP) data warehouse platforms have long recognized the benefits of an ELT (Extract, Load, Transform) approach to data integration through taking advantage of complete in-database processing as an alternative to the traditional ETL (Extract, Transform, Load) approach used for data integration in the past.

As a relatively recent alternative to conventional data integration however, this ELT approach suffered from the lack of sophisticated development tools available on the market. Most companies that are taking advantage of the benefits of ELT have done so using custom SQL scripts, which introduces all the challenges of in-house software development.

While some conventional ETL tool vendors have recently incorporated limited ELT capabilities in their products through the use of SQL pushdown optimization, these tools still have many limitations and do not completely support in-database processing, can't take advantage of native functions such as user-defined functions or analytical function libraries such as Fuzzy Logix' DBlytics or IBM Netezza Analytics, in addition to incurring added licensing, infrastructure and maintenance costs.

Conventional ETL tools:

- Can't fully leverage MPP/distributed computing platforms
- Introduce delays in data movement and processing
- Are overly complex, maintenance-heavy and expensive
- Don't scale well as data volumes grow
- Have become bloated over time as a result of incorporating disparate technologies
- Require a significant investment in additional hardware
- Limit the use of native analytical functions and in-database processing
- Will require substantial reengineering to effectively perform in this new environment

ETL using script-based SQL is relatively simple to develop and deploy, but also has major disadvantages:

- Not a metadata driven solution as maintaining metadata at the attribute level is near impossible due to the use of hand-crafted SQL.
- Difficult to understand and modify SQL scripts when there are changes in business rules.
- Data lineage cannot be maintained as transformation rules are buried in SQL scripts.
- End to end impact analysis is difficult if not impossible.
- Highly skilled SQL resources are required to build the solution.

- It is difficult and time consuming to maintain the solution.
- Job monitoring, alerts, metrics, stop and restart capabilities are missing or limited.

By taking advantage of a platform purpose-built for true ELT, these challenges are abated and all the benefits of this approach such as reduced data latency, speed to deployment and full data lineage are realized.

The Modern DI Suite, a Unique and Innovative Solution

The Modern Data Integration Suite provides a unique design approach to defining data transformation and integration processes in this environment, effectively leveraging the power of your Big Data platforms and resulting in reduced data latency, simpler development, deployment time and costs.

Unlike conventional data integrations solutions which have been patched together with a variety of various components to facilitate enterprise-wide data integration (particularly with the advent of Hadoop), Diyotta's Modern Data Integration Suite is based on an innovative unified architecture purpose-built for seamless Big Data integration.

Designed for MPP-based data warehouse appliances such as Netezza and Teradata as well as large-scale distributed computing platforms using Hadoop, the Modern DI Suite not only delivers the highest level of performance possible for the execution of data transformation and validation processes, but has also shown to be the most cost-effective solution available.

By eliminating the intermediate ETL infrastructure, essentially getting out of the way of your data, the Modern DI Suite employs a seamless approach to data movement across enterprise systems so data moves at wire speed, significantly reducing data latency and has shown to reduce the time required to take data from source systems, perform the necessary transformations and get the data to its final location from hours to minutes.

The Modern DI Suite:

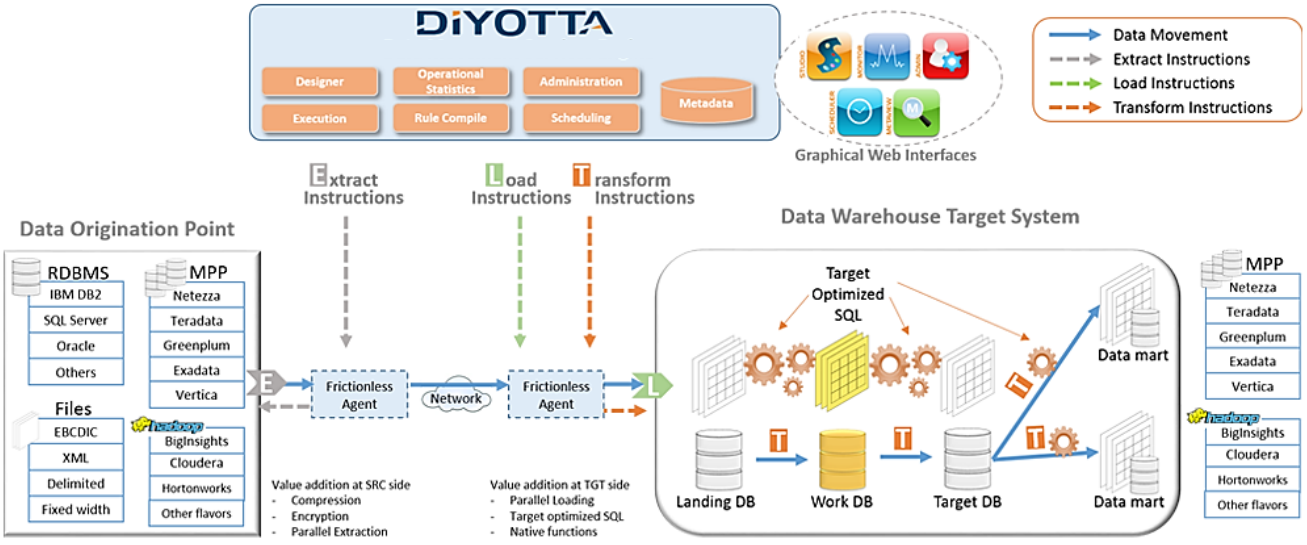
- Takes full advantage of MPP/HDFS platforms
- Eliminates intermediate ETL products and hardware
- Dramatically reduces data latency
- Simplifies data stream development
- Supports a unified development environment
- Maintains end to end data lineage supporting full impact analysis
- Provides ease of implementation
- Readily scales as data volumes increase
- Significantly reduces data integration costs

How the Modern DI Suite is Different

Modern Data Integration Suite	Other ETL Tools
Purpose built for ELT : 100% in database processing reduces data latency from hours to minutes	Limited SQL pushdown still requires intermediate processing with back-and-forth data movement
Browser-based development environment simplifies deployment and management	Desktop development environments require installation, administration and ongoing maintenance
Ready for Cloud with full- fledged Di features and functionally	Cloud solution with only limited features are available
Unified data integration solution supporting multiple platforms including Cloudera, BigInsight, Hortonworks, Netezza, Teradata, Pivotal HD, Vertica, Paraccel and more	Separate tools required to manage data integration on Hadoop and other distributed platforms
Diyotta is available as a ready-to-deploy DI appliance preconfigured with all necessary components to ease and speed deployment or an easy to install SW package	Require complex and expensive hardware infrastructure (servers, storage, networking) and introduce integration and maintenance challenges
No intermediate data processing infrastructure is required thus reducing costs significantly and supporting unlimited scalability	High-priced core software, costly add-on options, indivisual products for multiple sources and targets

The Modern Data Integration Suite is purpose-built on a standardized architecture with unified metadata to manage data integration across multiple big data platforms. The MDI Suite incorporates a pure ELT approach to effectively leverage your big data platform infrastructure as the data transformation engine, and uses native functionality to transfer the data where it is required across multiple data platforms.

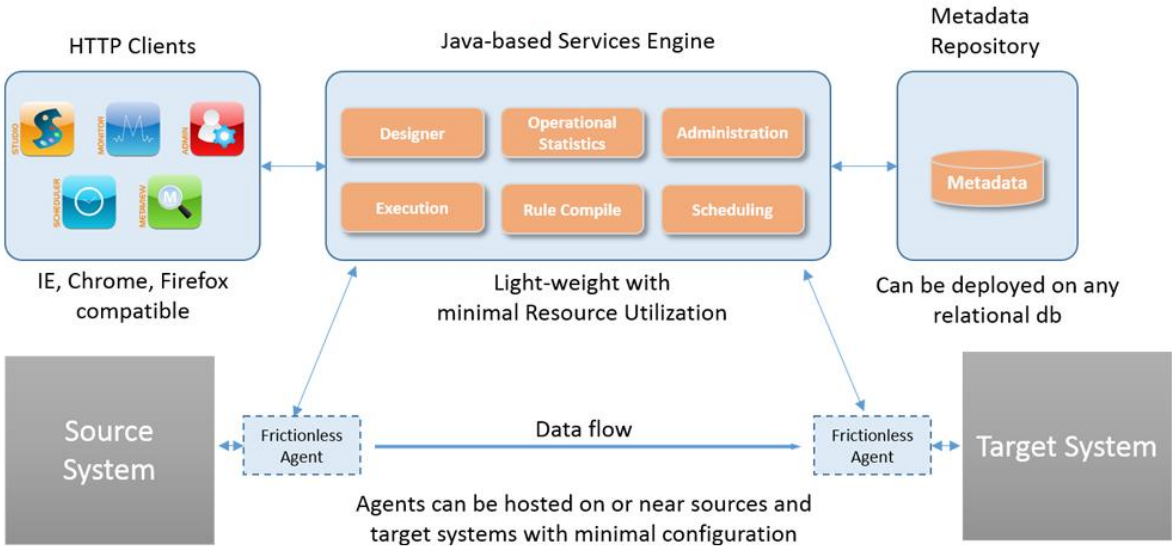
The MDI Suite Solution Architecture



- Data is acquired from multiple sources and loaded into a landing area on the MPP Warehouse or Hadoop platform without any major change to the data.

- The data is then cleansed and integrated with existing data and loaded into a staging area.
- Additional data will be loaded into the Data Warehouse in the required format based on business rules
- Finally, data is aggregated with rules applied and the final results are loaded into target data marts.

Modern DI Suite Component Architecture



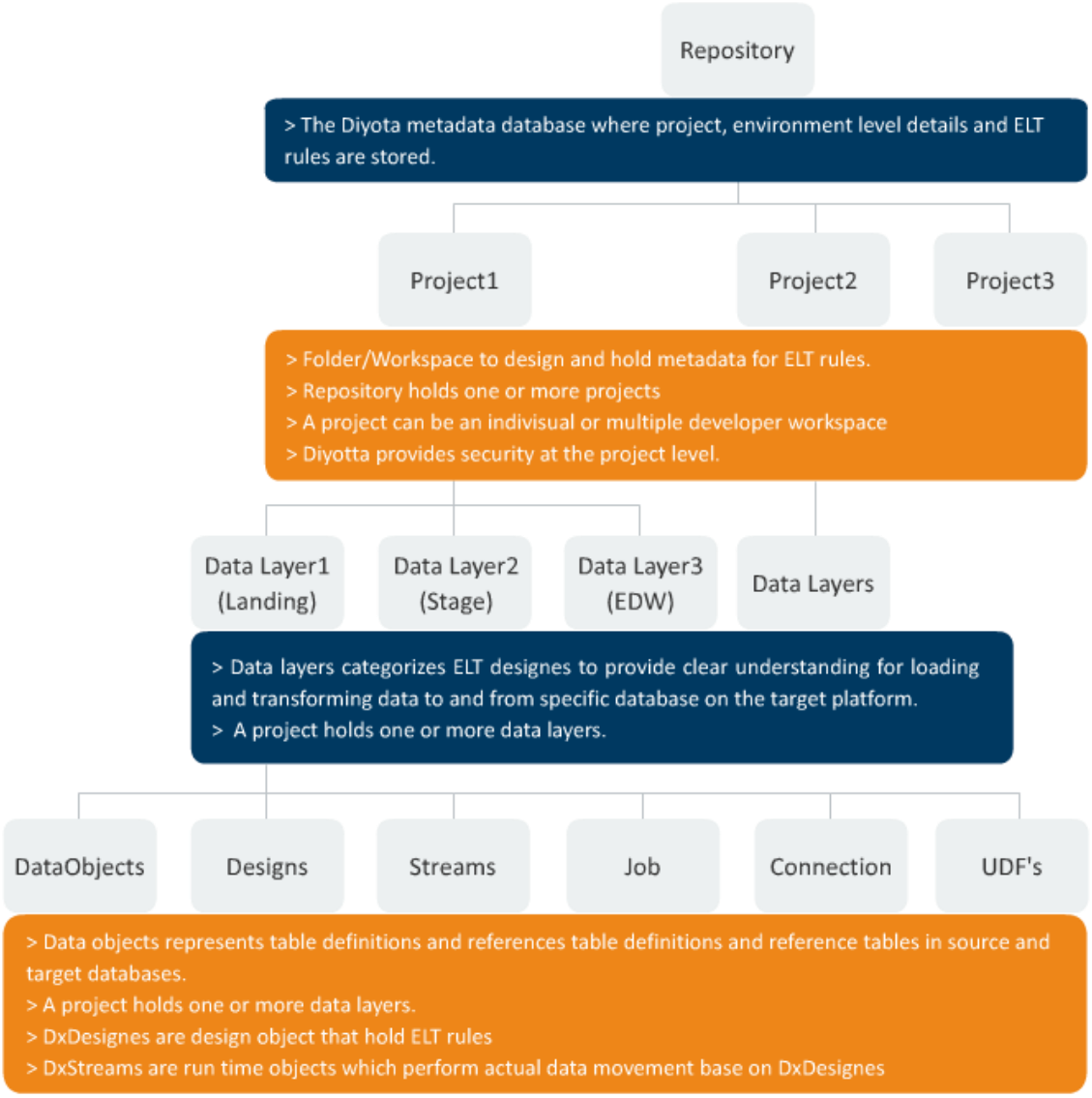
The Modern Data Integration suite consists of a browser-based integrated graphical development environment which allows you to:

- Incorporate and configure multiple source and targets systems, create and manager users, roles and security within the ADMIN component
- Import objects and design Data Streams across multiple platforms within the same projects using the STUDIO component
- Schedule, execute and monitor execution of these Data Streams via SCHEDULER and MONITOR
- Manage and analyze the associated metadata using METAVIEW
- Frictionless agents are lightweight services which can be independently deployed on data origination points or target systems. These take instructions and move data point to point

The Integration Server provides a set of services including a rules compile engine, execution manager, scheduling manager, metadata manager and the metadata repository

MDI Suite Metadata Management

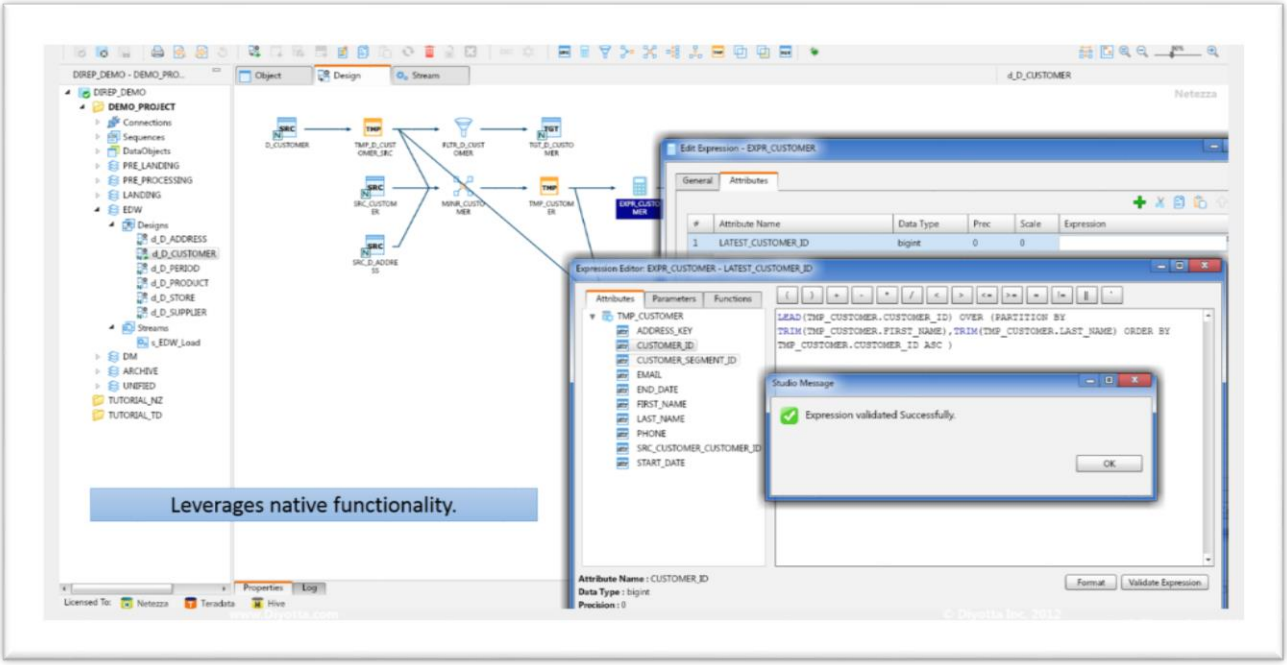
Fit-for-use metadata object hierarchy for ELT implementations:



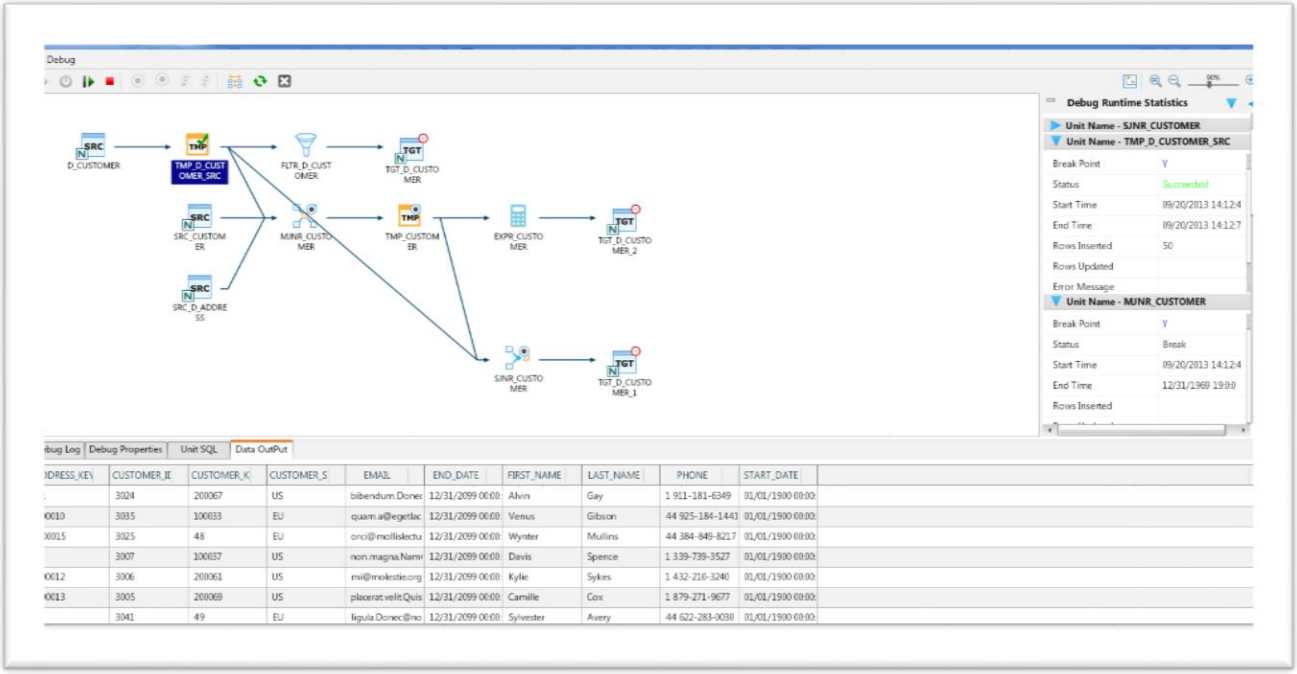
Expression Editor

The Expression Editor provides an interface to import, define and use all native database functions including row-level and analytical set-based window functions like RANK, LEAD, LAG, etc.

Ability to validate and debug expressions.



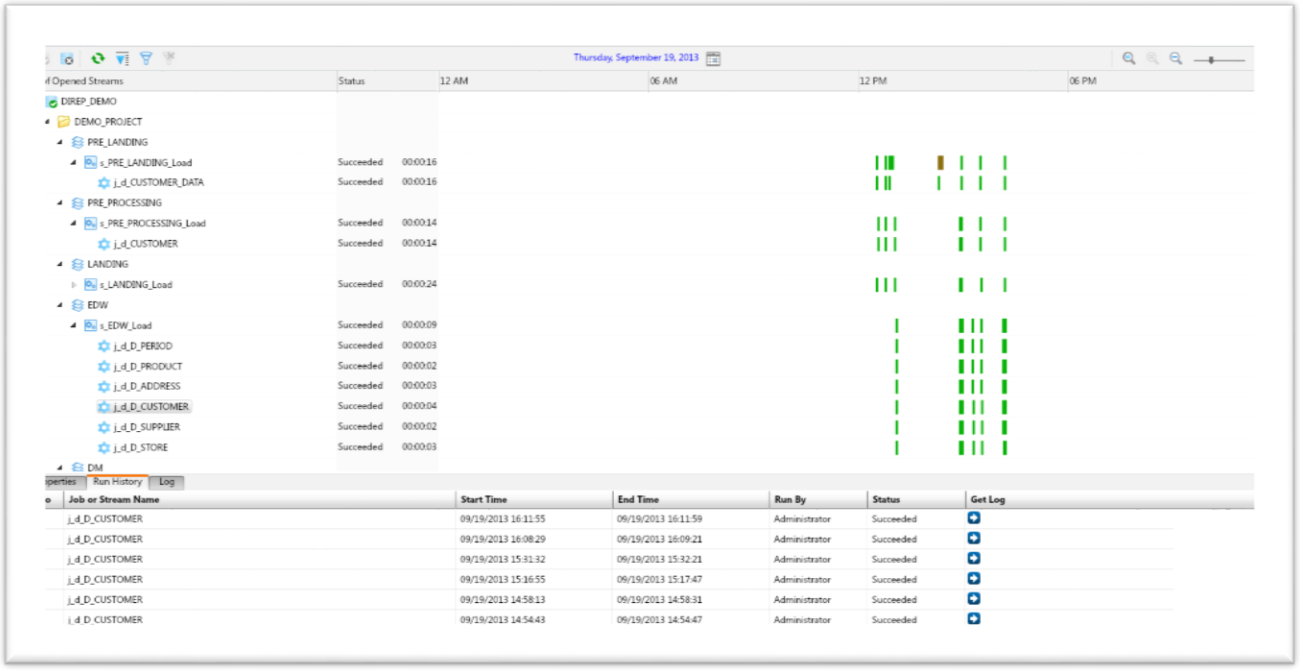
Data Stream Debugger



- Intuitive and unique UI for debugging Data Stream designs in set-based manner.
- Allows break points at all logical units
- Provides complete runtime statistics on each unit/break point including start time, end time, number of records processed, etc.

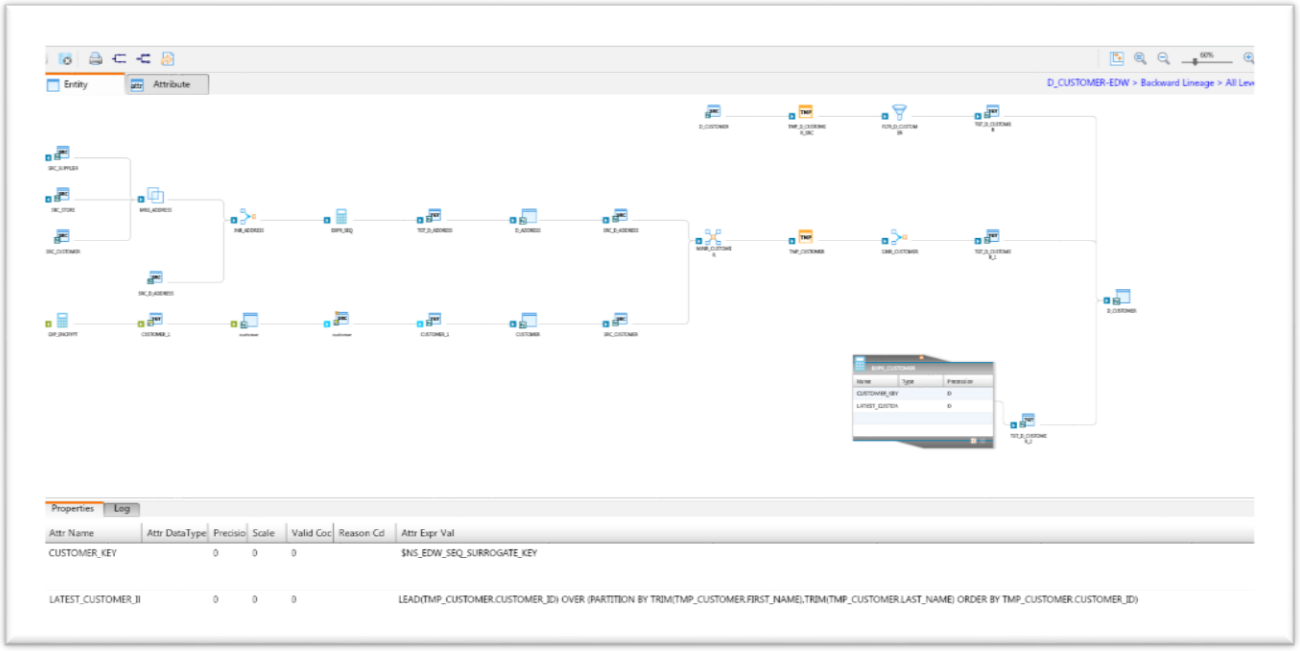
- Ability to resume and start debug from any break point
- Allows viewing SQL and Logs at each logical unit/break point level.

Monitor



The Monitor component UI allows users to monitor execution of streams and jobs, restarting failed data streams, viewing processed data on each logical unit and viewing detailed job logs.

Metaview



Diyotta Metaview allows users to view and analyze MDI platform metadata.

- Comprehensive data lineage with ELT rules.
- Forward and Backward data lineage at entity and attribute level.
- Provides impact analysis at the entity and attribute level.
- Design reuse is supported through parameterization for multiple Data Streams by passing parameter values at runtime using parameter files.
- Diyotta Objects can be exported as XML and maintained in any external version control system.
- Object versions can be maintained at the Project, Layer, Stream, Design or Data Object level as the system supports XML exports at any of these levels. This facilitates promotion from DEV to TEST to PROD.
- An intuitive Import wizard allows selection of child objects to reuse, or rename if they already exist in target environments.
- Allows you to choose target-related connection information while importing.
- The Command Line Interface (CLI) supports import and export for automated change control and deployment processes.
- The Scheduler component is an integrated UI which allows you to schedule execution of data streams and external programs or scripts, while the CLI gives you the flexibility to employ other enterprise scheduling tools such as AutoSys, Control-M or IBM TWS among others.

Modern DI Suite Capabilities

The Modern DI Suite architecture offers a complete unified data integration environment leveraging a pure ELT approach and seamless architecture.

- A completely metadata-driven solution that generates optimized SQL and performs 100% pushdown to execute transformation rules in-database
- Collaborative development features with check-out/check-in functionality to prevent overlap on development efforts
- Role based project level security to manage development efficiently
- Ability to extract data from heterogeneous external sources to pipe into native bulk load utilities to achieve optimized load performance
- Ability to define different types of transformations like Source, Single Joiner, Multi Joiner, Rollup, Filter, Splitter, Stage, Union and Minus to build rules for in-database processing
- Ability to view generated SQL for each logical unit the system defines for a transformation table.

Conclusion

The Modern Data Integration Suite is purpose-built on a standardized architecture with unified development and metadata to manage data integration across multiple Big Data platforms. The MDI Suite incorporates a pure ELT approach to effectively leverage your big data platform infrastructure as the data transformation engine.

The MDI Suite eliminates the need for standalone ETL servers, instead fully leveraging the inherent power of parallel processing data engines. This provides the fastest movement of data in your environment, the highest performance in executing data transformations, and the lowest data latency possible.

The MDI Suite is an intuitive, graphical, collaborative, and metadata-driven integrated development environment that increases productivity, reusability, and traceability of your data integration processes.

End to end data lineage maintains the heritage and veracity of your data – you know where it came from, what you did to it along the way, and where it ended up. This is key to regulatory compliance and impact analysis.

Besides the MDI Suite, there is no true metadata-driven pure ELT platform purpose-built for MPP with full data lineage exists on the market today. With our 100% SQL pushdown approach for complete in-database processing, end-to-end data lineage and metadata management, and sophisticated scheduling and monitoring tools in a unified architecture, you can finally leverage the power of your Big Data platforms.

Diyotta, Inc.3700 Arco CorporateDrive,Suite #410, Charlotte, NC 28273. +1-704-817-4646,
+1-888-365-4230, +1-877-813-1846