Intelligent and automated transformation of your enterprise data warehouse to Microsoft Azure

Modernizing legacy ETL and analytics to a cloud-based modern data platform like Azure Synapse tech stack has become a strategic imperative for enterprises struggling with petabytes of unstructured and fast data from multiple sources, high cost of ownership, and operation.

Microsoft Azure is one of the preferred choices of enterprises moving to the cloud as it offers a fully managed modern data architecture with advanced analytics solutions. With deep integration of SQL, machine learning, and other cloud-native services, Azure also helps enterprises lower operational costs and enhance analytics capabilities.
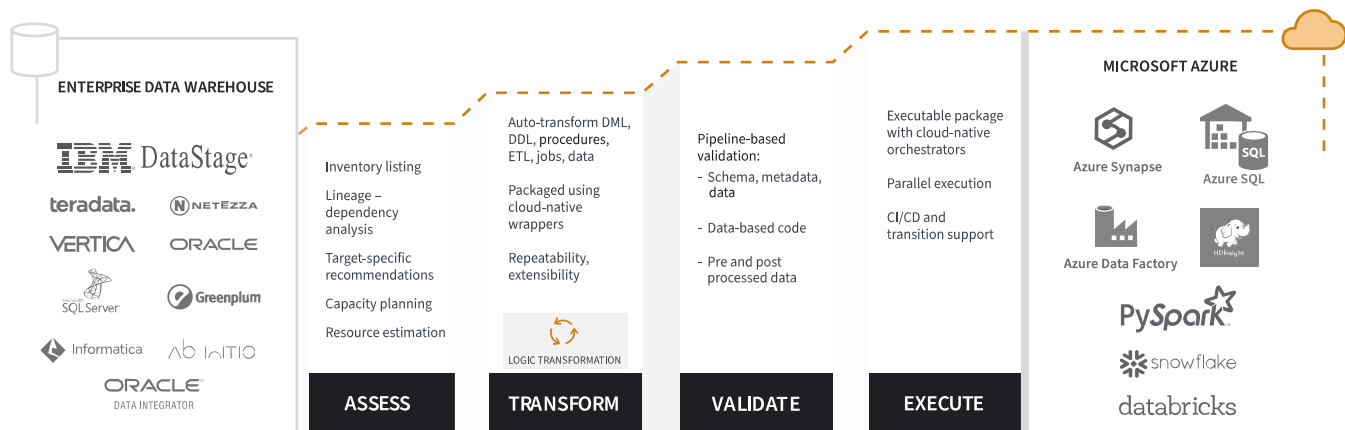
However, enterprises are still skeptical about moving to the cloud because of several concerns:

- Will there be any business downtime?
- How do I transform years of complex business logic and code?
- Will my workloads be optimized to address the nuances of the new environment?
- How do I prioritize the workloads for transformation?
- How do I handle data types in schema conversion?
- How do I ensure seamless operationalization of ETL and analytics workloads on the target environment?
- How do I validate the transformed code and migrated data?

LeapLogic, an Impetus product for automated workload transformation addresses all these concerns. Its intelligent grammar engine enables end-to-end transformation, operationalization, and transitioning of your legacy warehouse and analytical workloads to Azure tech stack for data storage, processing, and orchestration such as SQL Data Warehouse, HDInsight/Databricks, Data Factory, Data Lake Storage, or Synapse Analytics in four steps:

KEY BENEFITS

- 4x faster
- 50% cheaper than manual migration
- Zero business disruption
- Azure Stack-compliant conversions
- Complete visibility of DDL and DML conversions
- Preserve years of business logic, workflows, and execution rules
- Legacy workloads and ETL transformation and validation to Synapse compatible code
- Transition of procedural constructs (loops, cursors, etc.) to cloud-native wrappers

ENTERPRISE DATA WAREHOUSE

IBM DataStage®
teradata.   NETEZZA
VERTICA   ORACLE
SQL Server   Greenplum
Informatica   AB INITIO
ORACLE
DATA INTEGRATOR

| ASSESS | TRANSFORM | VALIDATE | EXECUTE |

ASSESS
Inventory listing
Lineage – dependency analysis
Target-specific recommendations
Capacity planning
Resource estimation

TRANSFORM
Auto-transform DML, DDL, procedures, ETL, jobs, data
Packaged using cloud-native wrappers
Repeatability, extensibility
LOGIC TRANSFORMATION

VALIDATE
Pipeline-based validation:
- Schema, metadata, data
- Data-based code
- Pre and post processed data

EXECUTE
Executable package with cloud-native orchestrators
Parallel execution
CI/CD and transition support

MICROSOFT AZURE
Azure Synapse   Azure SQL
Azure Data Factory   HDInsight
PySpark
snowflake
databricks

The 4-step approach to workload transformation

STEP 1: Assessment

- Automates inventory listing and profiling of legacy data warehouse

  - Assesses ETL scripts, DML and DDL scripts, procedures, scheduler/orchestrator scripts (jobs), shell scripts, analytical workloads (such as SAS scripts), etc.

  - Provides actionable insights and prescriptive recommendations

- Identifies complex interdependencies to group workloads for offload

  - Plots end-to-end lineage showing interdependencies between different kinds of workloads

  - Provides advanced filters according to workload type and an interactive graphical interface to deep dive into certain flows

  - Identifies technical debts

- Provides an advanced blueprint of the target architecture

  - Creates optimized schema and data distribution (appropriate format/file size for Azure Blob/ADLS uploads, etc.) to improve CPU usage, memory usage, cache hit ratio, and disk I/O

    o Strategize partitioning, bucketing, clustering, sorting, distribution keys, storage (columnar storage, columnar ordering, table partitioning, hash distribution, non-clustered indexes), table indexes, clustered columnstore index (default primary), clustered index (primary), heap (primary), non-clustered indexes (secondary)], and table distributions (round-robin distributed, hash distributed, and replicated)

  - Provides comprehensive, configurable recommendations for workload parallelism to ensure optimum performance on target

- Categorizes workloads as candidates for as-is migration, optimization, and total re-engineering

- ML-based assessment to provide actionable recommendations for future-state functional component architecture and tech stack components

STEP 2: Transformation

- Transforms diverse workloads and migrates schema and data to Azure

  - Up to 80% automated code conversion to Python and assisted transformation for the rest

  - DML scripts, DDL scripts, ETL scripts, scheduler scripts, stored procedures, etc.

  - Optimized and parallel data ingestion with PolyBase/T-SQL usage

  - Ensures automated schema conversion and data migration

  - Converts complex ETL scripts automatically

- Ensures end-to-end packaging

  - Transforms core business logic to Azure-native wrappers or orchestrators

  - Ensures end-to-end transformation of scheduler/orchestrator scripts to production-ready jobs on target

  - Ensures end-to-end execution on staging and production environments after system integration testing

  - Handles data types, nested views, intervals, loops, UDFs, procedures, complex constructs, etc.

  - Tunes performance and optimizes queries for Azure storage

  - Auto-generate patterns for newer migration

STEP 3: Validation

- Pipeline-based automated validation of the transformed code and migrated data on Azure
  - Ensures syntactical and data-based validation of scripts and queries on automatically generated or user-provided datasets
  - Validates code at the row and cell-level and reports errors
  - Instantly verifies the transformed code with a pluggable validation tool
- Data-based validation of the transformed code
  - Auto-generates sample dataset based on complex query conditions – ideal for unit testing of the transformed queries
  - Feeds the customer-provided dataset for testing on real datasets – suitable for integration testing of the transformed queries

STEP 4: Operationalization

- Delivers a target-specific executable package
  - Cloud-native orchestration and execution on production
- Ensures optimal performance through parallel execution
  - Provides parallel execution recommendations
  - Generates the required artifacts in the transformation output
  - Executes the generated artifacts in parallel on production
- Ensures end-to-end operationalization
- Supports end-to-end transitioning into production and operationalization
  - Minimal parallel-run period for stabilization
  - Azure Data Factory-driven pipelines
  - On-demand capacity planning for optimized price-performance
  - Auto setup Azure pipelines for continuous integration and delivery (CI/CD)
  - Data governance and compliance using Azure Purview/Data Catalog
  - Operational monitoring of Azure

LeapLogic, an Impetus product for automated workload transformation is built to efficiently assess, migrate, validate, and move complex workloads from legacy data warehouse systems to the Azure tech stack. Based on a sophisticated grammar engine capable of translating the most complex functions, this automated conversion tool helps you kick-start your digital transformation journey while opening a world of big data analytics for your business.

To start your end-to-end automated transformation to Azure, write to us at

inquiry@impetus.com