# Table of Contents

Overview 3

Dataflow Design 3
  Modularity and Reusability 3
    Subflows 4
    Shared Actions 4
    Shared Connections 4
    Detached Transformations 5

Performance 5
  Data Sources 5
    Filtering in Database 5
    Avoid Mapping Extra Fields 5
    Sorting for Succeeding Joins 5
    Partitioning 6
    Change Data Capture 6

Joins 6
  Lookups 7
  Destinations 7
  General Guidelines 7

Input Parameters and Output Variables 7
  Input Parameters 7
  Output Variables 7

Additional Resources 8
Overview

Centerprise represents a new generation in data integration platforms designed for superior usability, productivity, and performance. Astera has worked closely with customers from a variety of industries, including financial services, healthcare, government, nonprofits, and others to build and continually refine Centerprise with the specific goal of creating and maintaining complex hierarchical dataflows and workflows. Over the years, we have assisted customers in successfully deploying Centerprise in a variety of usage scenarios. The Centerprise Best Practices Series represents a synthesis of these experiences and each part will cover a specific key area of Centerprise technologies and processes. The objective is to assist customers in developing world-class solutions by establishing and communicating these best practices.

Dataflow Design

Dataflows are the cornerstone of any integration project. The Centerprise Dataflow Designer leads the industry in power and ease of use. Our customers consistently give high marks to the Centerprise visual interface, drag-and-drop capabilities, instant preview, and full complement of sources, targets, and transformations.

The dataflow best practices outlined herein are intended to facilitate initial creation and ongoing maintenance of Centerprise dataflows. Whether you are developing small dataflows with a few steps or large ones comprising scores of steps, these practices will help you manage your integration project more effectively.

Modularity and Reusability

Modularity is a key design principle in the development of integration projects. Modularity enhances maintainability of your dataflows by making them easier to read and understand. It also promotes reusability by isolating frequently-used logic into individual components that can be leveraged as "black boxes" by other flows. Centerprise supports multiple types of reusable components. These components and their usage recommendations are discussed in the following subsections.
Subflows
Subflows are reusable blocks of dataflow steps that have inputs and outputs. Subflows enable users to isolate frequently-used logic in reusable components. Once created, subflows can be used just like built-in Centerprise transformations. Examples of reusable logic that can be housed in subflows include:

- Validations that are applied to data coming from multiple sources, frequently in incompatible formats
- Transformation sequences such as a combination of lookup, expression, and function transformations that occur in multiple places in the project
- Processing of incoming data that arrives in different formats but must be normalized, validated, and boarded

Shared Actions
Shared Actions are similar to subflows but only contain a single action. Shared actions should be employed when a source or destination is used in multiple places within the project. This way, if a field is added or removed from this source, all the projects inherit that change automatically.

Shared Connections
Shared Connections contain database connection information that can be shared by multiple actions within a dataflow. Shared connections can also be used to enforce transaction management across a number of database destinations. Use shared connections whenever multiple actions in a dataflow use the same database connection information.
Detached Transformations

Detached Transformations are a capability within Centerprise developed for scenarios where a lookup or expression is used in multiple places within a dataflow. Detached Transformations enable you to create a single instance and use it in multiple places. They are available in expressions as callable functions, enabling you to use them in multiple expressions. Additionally, Detached Transformations allow you to use lookups conditionally. An example of a conditional lookup would be, “if party type is insurer, perform lookup on insurer table else perform lookup on provider table.”

Performance

Centerprise has been designed as a parallel-processing platform to deliver superior speed and performance, so designing dataflows to take advantage of the software’s abilities can significantly affect your data integration performance. The performance best practices discussed here can result in a major performance boost in many common situations.

Data Sources

Frequently, a dataflow can be optimized by some fine-tuning at the data source. Some of the optimization techniques are discussed in this section.

Filtering in Database

When loading data from a database, enter a where clause to filter data at the source. Loading data in Centerprise and then filtering using the Centerprise Filter Transformation can significantly degrade performance.

Avoid Mapping Extra Fields

The Database Table Source automatically creates a query to load only the fields that are mapped. To take advantage of this optimization, map only the fields that are used in subsequent actions.

Sorting for Succeeding Joins

The performance of Centerprise joins improves by orders of magnitude when working with previously-sorted data. Where possible, avoid sorting data in Centerprise and sort instead in your database query by adding order by clause.
**Partitioning**

Centerprise Database and File Sources enable data partitioning, which speeds up reading by breaking a dataset into chunks and reading these chunks in parallel. Use partitioning if you are moving a large data table.

**Change Data Capture**

If you periodically transfer incremental data changes, consider using one of Centerprise’s change data capture (CDC) patterns to ensure your data is as up to the minute as you need it to be. Centerprise supports a variety of CDC strategies enabling you to select the appropriate strategy to fit your environment and requirements. Refer to Additional Resources at the end of this document for more information on Centerprise change data capture.

**Joins**

Centerprise Join Transformations enable you to join multiple data sources. Joining often involves sorting incoming data streams, making it the most common reason for performance issues. Here are the practices to keep in mind when using joins:

- Joining data in Centerprise is time consuming. Use Database Joins wherever possible.
- Sort data in the source where appropriate. Joining sorted data streams is much faster.
- When joining data from the same database, use the Database Join option in the Join Transformation. When the Database Join option is specified, Centerprise builds and runs a single query joining multiple tables. This can enhance performance in most situations.
- Reducing the number of fields in the Join Transformation also improves performance, so be sure to remove any unnecessary fields.

**Lookups**

Lookups are another source of performance issues in integration jobs. Centerprise offers several caching techniques to improve the performance of lookups. Experiment with lookup caching options and select the options that work best for specific situations. Some tips are:

- If you work with a large dataset that does not change often, consider using the new Centerprise Persistent Lookup Cache, which stores a snapshot of the lookup table on the server’s local drive and uses it in subsequent runs. In situations where the lookup table is updated daily, a snapshot can be taken on the first run after update and can be used throughout the day to process incremental data.
• If you work with a large dataset but use only a small fraction of items in a single run, consider using the Cache on First Use option
• If a lookup is used in multiple places within the same flow, consider using a detached lookup
• Where appropriate, use a database join instead of a lookup function

Destinations
Centerprise supports bulk load for popular databases. Where possible, use bulk load for database destinations and experiment with bulk load batch sizes to fine-tune performance. Centerprise offers a Diff Processor Transformation that can be used to compare an incoming data stream with existing data in the table and apply differences. In certain situations, this transformation can substantially speed up dataflows.

General Guidelines
• Avoid unnecessary steps. Some steps such as Record Level Log incur major performance overhead and should only be used during development to isolate data issues.
• The job monitoring window shows execution time for each step. These are not precise numbers but provide a rough approximation of time taken by each step and can be used to optimize specific steps.

Input Parameters and Output Variables
Input parameters and output variables make it possible to supply values to dataflows at runtime and return output results from dataflows. Well-designed parameter and output variable structures promote reusability and reduce ongoing maintenance costs.

Input Parameters
Input parameter values can be supplied from a calling workflow using data mapping. When designing dataflows, analyze the values that could change between different runs of the flow and define these values as input parameters. Input parameters can include file paths, database connection information, and other data values.

Output Variables
If you would like to make decisions about subsequent execution paths based on the result of a dataflow run, define output variables and use an expression transformation to store values into output variables. These output variables can be used in subsequent workflow actions to control the execution path.
Additional Resources:

Astera Website:

Brochures and Presentations:
- Centerprise 6.0 Presentation - http://www.slideshare.net/gloverastera/whats-new-in-centerprise-60

Videos:
- Introduction to Centerprise 6 Webinar - http://www.youtube.com/watch?v=e9q4mh8p1gc
- Getting Started with Centerprise Dataflow
  - Part 1 - http://www.youtube.com/watch?v=UaU2VqmLD6Y
  - Part II - http://www.youtube.com/watch?v=yRlXdA3GAS0
- Dataflow Concepts in Centerprise - http://www.youtube.com/watch?v=cQ7qz5lIvaQ

Documentation:
- Parameterizing Dataflows - https://astera.zendesk.com/entries/21530938-Parameterizing-Dataflows
- Setting Up Destinations - https://astera.zendesk.com/entries/21383441-Setting-Up-Destinations