DATASHEET

BLUEPRINT FOR BIG DATA SUCCESS: A BEST PRACTICE SOLUTION PATTERN

Filling the Data Lake

Simplify and Accelerate Hadoop Data Ingestion with a Scalable Approach

What is it?

As organizations scale up data onboarding from just a few sources going into Hadoop to hundreds or more, IT time and resources can be monopolized, creating hundreds of hard-coded data movement procedures – and the process is often highly manual and error-prone. The Pentaho Filling the Data Lake blueprint provides a template-based approach to solving these challenges, and is comprised of:

- A flexible, scalable, and repeatable process to onboard a growing number of data sources into Hadoop data lakes
- Streamlined data ingestion from hundreds or thousands of disparate CSV files or database tables into Hadoop
- An automated, template-based approach to data work-flow creation
- Simplified regular data movement at scale into Hadoop in the AVRO format

Why do it?

- Reduce IT time and cost spent building and maintaining repetitive big data ingestion jobs, allowing valuable staff to dedicate time to more strategic projects
- Minimize risk of manual errors by decreasing dependence on hard-coded data ingestion procedures
- Automate business processes for efficiency and speed, while maintaining data governance
- Enable more sophisticated analysis by business users with new and emerging data sources

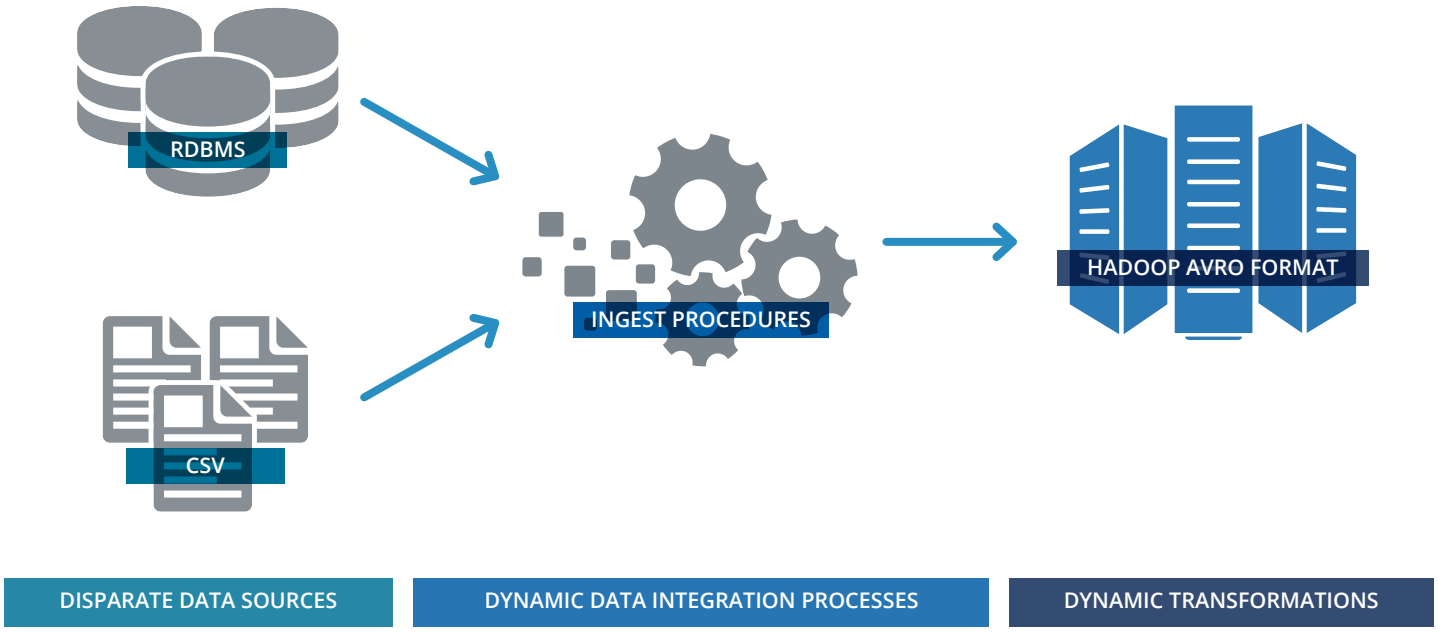
Value of Pentaho

- Unique metadata injection capability accelerates time-to-value by automating many onboarding jobs with just a few templates
- Intuitive graphical user interface for big data integration means existing ETL developers can create repeatable data movement flows without coding – in minutes, not hours
- Ability to architect a governed process that is highly reus-able
- Robust integration with the broader Hadoop ecosystem and semi-structured data

Example of how a Filling the Data Lake blueprint implementation may look in a financial organization

This company uses metadata injection to move thousands of data sources into Hadoop in a streamlined, dynamic integration process.

- Large financial services organization with thousands of input sources
- Reduce number of ingest processes through Metadata Injection
- Deliver transformed data directly into Hadoop in the AVRO Format



pentaho.com/contact | +1 (866) 660-7555