Pentaho Data Integration’s Key Component: The Adaptive Big Data Layer

Driving the Flexibility and Power of Pentaho Data Integration (PDI) to Future-Proof You Against Big Data Technology Advancements

The Not-So-Hidden Intricacies of Transforming Big Data into Business Insights

As any IT or developer department knows, there are growing demands to turn data into business insights for multiple business departments, all with heterogeneous data sources. In order to fulfill these demands with high-impact results, it is critical that data projects are executed in a timely and accurate fashion. With the multitude of ever-changing big data sources, such as Hadoop, building analytic solutions that can blend data from disparate sources is a daunting task. Further, many companies do not have the highly skilled resources for difficult-to-architect big data projects.

THE ADAPTIVE BIG DATA LAYER: A PROTECTIVE BUFFER BETWEEN TECHNICAL TEAMS AND BIG DATA COMPLEXITY

PDI’s Adaptive Big Data Layer (ABDL) gives Pentaho users the ability to work with any big data source, providing a fully functioning buffer to insulate IT analysts and developers from data complexity. Created with common data integration challenges in mind, the tool enables technical teams to operationalize existing and emerging big data implementations in an agile manner. The ABDL accounts for the specific and unique intricacies of each data source to provide flexible, fast, and accurate access to all data sources.

EMBEDDED PENTAHO DATA INTEGRATION

Step Plugins
Sort
Cloudera

Adaptive Big Data Layer
MongoDB
Amazon EMR

Perspective Plugins
Schedule
Hadoop
Hortonworks

and More
and More
MapR
and More
How it Works

As a core component within Pentaho Data Integration, the ABDL insulates the data developer from the shifting sands of big data and enables "create once, run anywhere" transformation processes to work against any big data store (Hadoop, analytic database, NoSQL store, and others) allowing immediate communication with minimal effort. Enriching and blending data across big data sources with the assistance of PDI's ABDL becomes a simplified process of drag and drop.

**BENEFITS OF ABDL**

ABDL gives IT/Developers access to the broadest, deepest, and easiest to use big data integration in the industry, allowing for:

- Insulation from changing big data sources: versions, vendors, data stores
- Reduced risk from proven and tested PDI transformations
- Rapid time to value with real-time updated transformation options based on vendor updates
- Native integration into the big data ecosystem, whether it's an analytic database (HP Vertica, Amazon Redshift), Hadoop distribution (Amazon EMR, Cloudera, Hortonworks, MapR), NoSQL (MongoDB, Cassandra), or specialized technology (Splunk, Spark)
- Simplified blending of data across multiple big data sources to provide the richest insights
- Unparalleled flexibility to work with any big data source or combination of sources
- Ability to report and analyze directly from Big Data sources

**PROVEN ROI FROM PENTAHO CUSTOMERS**

The Adaptive Big Data Layer has achieved significant returns with:

- **Faster development:** Drag-and-drop ETL through Pentaho MapReduce is up to 15X faster than manual coding, meaning major time and cost savings on development
- **Faster execution:** Run queries up to 40X faster resulting in lower project turnaround time
- **Minimal Training:** Easy-to-use ETL tools allow users to get up to speed quickly, with little retraining required, equating to less staff overhead
- **Deployment flexibility:** Simple movement and integration of Big Data deployed on-premise or in the cloud

---

**A PENTAHO MAPREDUCE CASE STUDY**

**MAJOR FINANCIAL INSTITUTION**

**BUSINESS CHALLENGE**

- Gain competitive advantage through intraday balance reporting for commercial customers
- Technical proof of concept
  - Job 1: Extract, clean/manipulate, load in Hadoop HDFS (10 lookups to HDFS)
  - Job 2: Load proceeded results into RDBMS

---

**JAVA MAPREDUCE**

- 2 Java developers from major SI
- 1+ month
- Job 1: 5 min. 43 sec
- Job 2: 7 min. 45 sec
- 13 min 45 sec
- Incorrect data due to coding errors
- Results

**PENTAHO MAPREDUCE**

- 1 ETL developer (50% the resource)
- 2 days (15X faster)
- Job 1: 20 sec
- Job 2: 27 sec
- 47 sec (17X faster)
- Accurate Results