



Pervasive DataMatcher

Pervasive DataMatcher™ Powered by Pervasive DataRush™ Improves Data Quality

Pervasive DataMatcher (PDM) is a highly accurate and lightning fast, high-performance data-matching solution that delivers the power of a cluster in a single SMP node. It is designed to help many types of organizations that suffer huge costs from the combinations of fraud, corruption and non-compliance issues. These organizations regularly need to process staggering amounts of data from multiple data sources that are often inaccurate, inconsistent and contain duplicate records. PDM detects data redundancies and correlates records to deliver fast, precise analytic results with rapid ROI.

The seemingly straightforward task of data matching on large disparate datasets is a daunting challenge for

many IT organizations. Even with a fairly small dataset, comparing each record to every other record can generate an overwhelming amount of data. For example, comparing every record of a 100,000-row dataset would involve nearly 5 billion record comparisons.

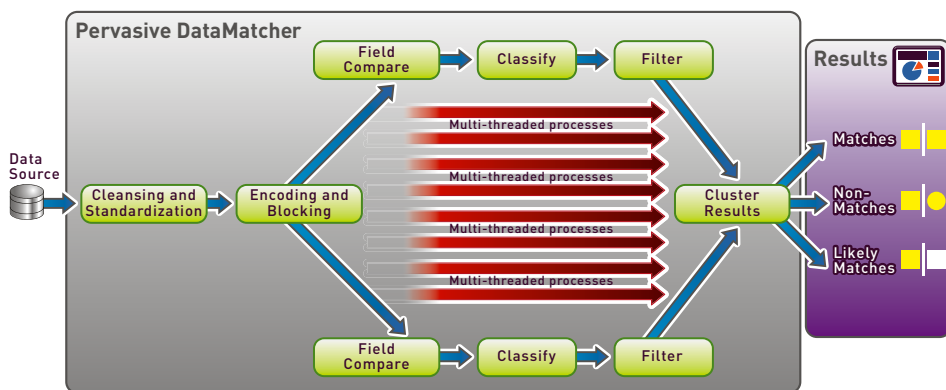
PDM scales seamlessly on large, complex datasets with the ability to match on any or all fields in a dataset, including fuzzy matching. PDM is built on top of the Pervasive DataRush (PDR) parallel dataflow platform, enabling the solution to crunch through massive datasets quickly and accurately on commodity multicore hardware.

Pervasive DataMatcher Use Case

Pervasive helped a global intelligence solutions provider reduce the amount of time to perform duplicate record identification from several hours to a few minutes! The company helps its customers to find new business opportunities and understand international trade trends by collecting bills of lading and translating raw data into meaningful intelligence. It took the company **4-5 hours** to identify duplicate data in a dataset containing tens of millions of source records prior to using Pervasive DataMatcher (PDM). Using the PDM application based on the DataRush platform, the same duplicate identification took just **over 10 minutes** on an 8-core server. This performance boost enabled more iterative business process development leading to better quality matches.

Benefits

- Lower operational expenses – quick ROI
- Improved decision support and organizational operations – faster and more accurate knowledge are available for decision support and organization operations
- Long-running processes can be reduced to near real-time
- New paradigm opens up additional business opportunities
- Flexible design allows for easy integration into existing or new solutions



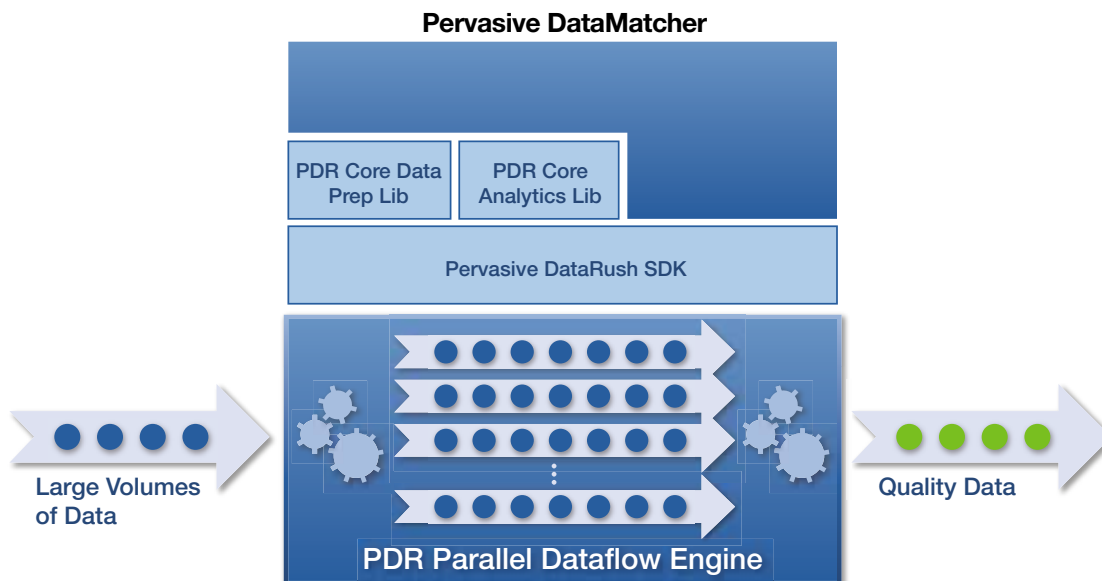
Architecture

PDM uses the PDR parallel, dataflow-based architecture incorporating an innovative combination of pipelining, vertical and horizontal partitioning to rapidly crunch through massive datasets on off-the-shelf multicore servers. PDM utilizes PDR partitioning to block the input data into groups of like data for deeper comparison. This allows PDM to execute fuzzy, compute-intensive field comparisons in parallel. Taking advantage of multicore processing, PDM is able to provide rich matching functionality with best of breed performance.

Features

- Fuzzy matching to detect duplicate records in single or multiple datasets
- Multi-thread performance provides unprecedented response times for large datasets

- Ability to match on any combination of fields in datasets
- Ability to handle multilingual and international character sets with full Unicode support
- Data partitioning to segregate data into like groups for deeper comparison, delivering high scalability and better application performance as more processor core resources are available, using encodings including Soundex, Refined Soundex, Metaphone, Double Metaphone and sub-string
- Fields comparisons using matching algorithms including Levenshtein Edit Distance, Jaro, Jaro-Winkler, Jaro-Hef, Damerau-Levenshtein, Q-gram, Positional Q-gram, prefix, suffix and exact match



About Pervasive Software

For over two decades, Pervasive has helped thousands of customers around the world in every industry to manage, integrate and analyze their data. Pervasive DataRush is a parallel dataflow platform enabling data-intensive applications such as claims processing, risk analysis, fraud detection, data mining, predictive analytics, bioinformatics and life sciences.

Contact Us

www.pervasivedatarush.com
1-866-980-RUSH (7874)
1-512-231-6818 (Main Office)

info@pervasivedatarush.com

PLATFORM SUPPORT

Design-Time Platforms and Runtime Operating System Prerequisites

- Windows
- Windows XP
- Windows Vista
- Solaris
- IBM AIX
- Red Hat Enterprise Linux
- SUSE Linux Enterprise
- HP-UX

Processors

- Any 32- or 64-bit Intel or AMD CPU
- Any Sun SPARC 32/64-bit CPU
- Any IBM POWER Architecture 32/64-bit CPU

Java 6 JVMs