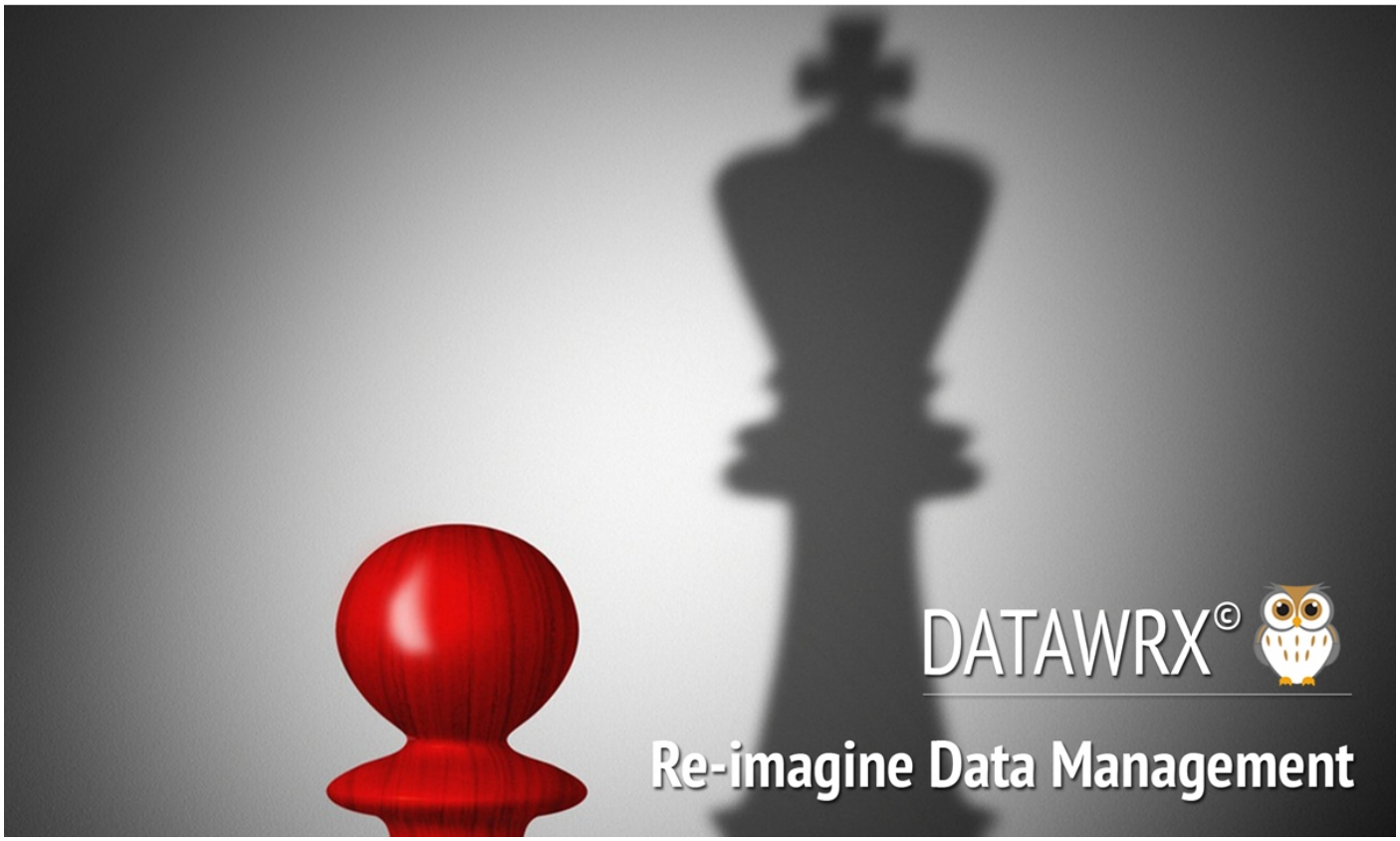
Percipient©
intelligent data engineering

A Complete Open Source-Based Data Management Platform

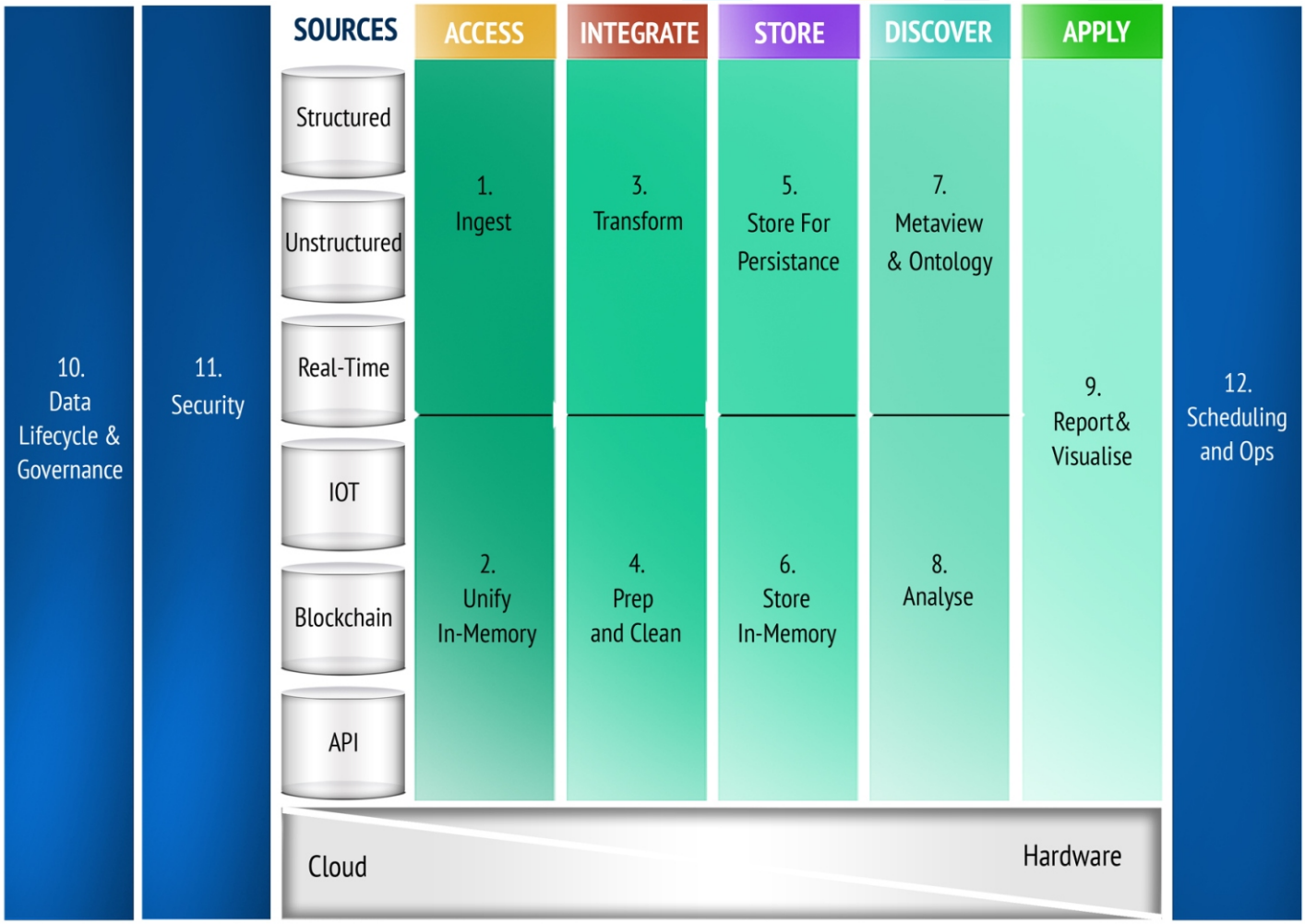


Author: Gayathri Dwaraknath
Chief Solutions Officer
July 2017

Summary

DataWRX© solutions enable mix-and-match enterprise data stacks that stay in step with business growth, digital trends and technological innovation. By carefully selecting the most suitable open source projects, and continually replacing outdated components, DataWRX© helps enterprises build a data infrastructure that remains fit-for-purpose, now and in the future. Expensive proprietary lock-ins with legacy vendors can be eliminated.

DataWRX© is comprised of 12 building blocks



1. Data Ingestion

What is it?

Typically in an enterprise, data originates from an array of legacy and new data sources. These can be in discrete batches or streaming in real time.

To access these sources can take weeks of effort, and routing it to a desired destination means even more delays.

How do we do it?

The DataWRX© platform has pre-built connectors to 250 data sources, including the conventional eg OLAP warehouses, OLTP application databases and CSV files, and unconventional eg APIs, blockchain, Apache Hive, IoTs and social media.

Connections can be established in minutes. Custom connectors can be built for standalone data sources.

What's the business value?

- ✓ Ability to preserve and access legacy sources
- ✓ Change management and incorporation of new sources
- ✓ Easy and fast connections

2. In-Memory Unification

What is it?

Traditionally, before multi-sourced data can be used, it must first be staged and stored in a central location like an Enterprise Data Warehouses (EDW).

When an enterprise's data volumes increase, EDWs must be scaled out at great expense. In-memory offer a compelling alternative to these on-disk technologies.

How do we do it?

The DataWRX© platform enables data, regardless of the source, to be directly available at its final destination. Memory, parallel processing and node clusters are used to ensure this data unification is achieved at high speed and limitless scalability.

The data can then be discovered, explored and queried from a single interface, using SQL, and delivered via JDBC to any visualisation or BI tool.

What's the business value?

- ✓ No need to invest in an expensive DWH if not already in place
- ✓ Ability to contain data growth costs
- ✓ High speed data discovery and exploration

3. Transformation

What is it?

Data from a source system must be transformed if its structure or format does not align with that required at its destination. ETL (Extract, Transform, Load) tools, eg Ab Initio, Datastage, Informatica do this by applying a series of rules or functions to the original data.

This requires extra proprietary code maintenance, learning and resources.

How do we do it?

The DataWRX© platform uses in-built SQL functions, procedures and packages to facilitate 70% of all required data transformations. Customised SQL functions can be added to meet enterprise-specific requirements, while specialised functions, eg Lambda for real time data, are fully supported.

Where highly complex transformations are needed, DataWRX© offers sophisticated and scalable open source software such as Apache Nifi, which is loss tolerant, highly scalable and uniquely, enables flows to be modified at runtime.

What's the business value?

- ✓ Reduced ETL complexity and infrastructure
- ✓ Advanced features are supported using SQL functions
- ✓ Highly scalable transformations

4. Preparation and Cleaning

What is it?

The task of cleaning and prepping data is important for achieving meaningful data outcomes. Data can be inaccurate or incomplete, but can also be in a messy or unusable form.

This requires enterprises to engage in the detection of data inconsistencies, dataset transformations and determining of data linkages, typically onerous processes.

How do we do it?

The DataWRX© platform enables data to be easily filtered and partitioned using regular expressions. Basic and advanced cell transformations are possible, even for cells with multiple values, to achieve the desired granularity.

These are achieved without the complex coding required by traditional tools.

What's the business value?

- ✓ Data which usually requires weeks to prepare can be explored in seconds
- ✓ Import of unified, rather than individually-sourced, data, greatly enhances processing efficiency.
- ✓ No extra coding means there is more user control

5. Persistant Data Storage

What is it?

Where data persistance is required, the data is traditionally stored in monolithic servers using a tabular and relational structure. This is now proving inadequate or too expensive where large volumes or new sources, such as realtime streams, web applications and voice recordings, are involved.

In such cases, distributed computing and schema-free mechanisms are more appropriate.

How do we do it?

The DataWRX© platform offers enterprises the option of using affordable open source data storage platforms. These can be relational databases, eg MySQL, or non-relational databases such as NoSQL, or data lakes eg Hive. Built on top of Hadoop, Hive supports cluster-based distributed processing of very large datasets at low cost.

For even greater capacity, data can be further compressed without affecting its ready availability.

What's the business value?

- ✓ Flexible way to store data
- ✓ Large cost arbitrage between traditional databases and new age data lakes
- ✓ Caters for exponential data growth and new data types

6. Virtual Data Storage

What is it?

Data can be temporarily stored in a computer's cache memory, as an alternative to moving data to a monolithic data warehouse.

This can help to overcome latency problems due to high volumes of data being processed, high frequency of data movement, a large number of concurrent users, or limited network bandwidth.

How do we do it?

The DataWRX© platform draws on cached data to improve on processing performance, even under stressful conditions. This means that, for example, data in disparate formats, stored across several clouds, can be mirrored locally to facilitate fast retrieval and unification.

Also, instead of numerous team-based copies, memory units of data can be shared by multiple teams.

What's the business value?

- ✓ Data is available to large numbers of concurrent users
- ✓ Enables frequent re-querying of large and varied datasets
- ✓ Fast data processing across locations and geographies

7. Metaview & Ontology

What is it?

In order to easily discover what datasets reside within an enterprise's central or multiple repositories, it is necessary to have a high level metaview of the data, plus the ability to drill down as required.

Ontologies help to order and make sense of these datasets.

How do we do it?

The DataWRX© platform enables linkages between datasets to be created instantaneously. Datasets can be expanded using interactive web services and lookups.

The easy-to-use interface is designed to make data discovery across multiple systems easier not only for data scientists, but also business users.

What's the business value?

- ✓ One comprehensive view of an enterprise's data
- ✓ Easy search for data patterns
- ✓ Pre-built intelligence to help identify interrelationships across datasets

8. Advanced Analytics

What is it?

Once data is readily available, it can be put to good use.

Sophisticated statistical tools and techniques are available to examine very large amounts of heterogeneos data in order to forecast customer behaviour and derive new business insights.

How do we do it?

The DataWRX© platform delivers consolidated data to standard analytics engines like SAS, but is also seamlessly integrated with new age, open source engines such as Apache Spark.

This integration enables DataWRX© to offer a vast library of machine learning algorithms, but also frees up more of Spark's memory for analytics rather than data management.

Support for Lambda and Kappa architectures also enables analytics on batch plus realtime data.

What's the business value?

- ✓ Availability of cutting edge analytical tools
- ✓ Existing R users can run their analytics on the high speed Spark platform
- ✓ Readily available algorithms

9. Reporting & Visualisation

What is it?

Once queried and analysed, enterprises seek to portray their data outcomes in a visual context to facilitate easy understanding.

Basic charting of trends, patterns, and correlations are commonplace using Excel, but a range of proprietary digital tools have evolved to enable interactive and sophisticated data visualisation. However, these typically carry a high price tag.

How do we do it?

The DataWRX© platform offers a number of open source reporting tools, including BIRT, Jasper Reports and Am Charts, that fulfill enterprise's data reporting and visualisation needs. These can be embedded into mobile and web applications, dashboards, regulatory reports, etc.

Dynamic forms and pivot charts are supported, together with the ability to collaborate notebooks, thereby eliminating the need for multiple proprietary tools.

What's the business value?

- ✓ Feature-rich data visualisation
- ✓ Reporting customisation to ensure alignment with regulatory requirements
- ✓ Significant cost savings

10. Data Lifecycle & Governance

What is it?

It is important in the management of data to keep sight of who is using the data and track what it is being used for.

It is also crucial to monitor how data flows within the enterprise's data stack, trace its lineage, and be able to remove data when it become obsolete or irrelevant.

How do we do it?

The DataWRX© platform offers a seamless user experience across the design, control, feedback and monitoring of an enterprise's data via a web interface. an interface to administer its use.

This notebook-style UI enables data administrators and users to track data flows from beginning to end. Complex queries can be saved and shared with other users.

What's the business value?

- ✓ Oversight of data usage
- ✓ No duplicated effort
- ✓ Supports most IT and data audit protocols

11. Security

What is it?

A top priority for any enterprise is to protect its data from the unwanted actions of unauthorised users.

Many enterprises, especially those with sensitive customer data such as financial institutions, have functional authentication frameworks in place.

How do we do it?

The DataWRX© platform engages a variety of methods and technologies, including SSL, SSH, HTTPS and content encryption, to ensure an enterprise's security standards are maintained or exceeded.

Besides providing controls at the level of the individual and function, the platform allows for even more fine-grained access controls, such as to specific data files, tables and columns.

What's the business value?

- ✓ Supports existing LDAP directories
- ✓ Offers role and data-based access controls
- ✓ Access tracking via the platform's Admin UI

12. Scheduling & Ops

What is it?

Moving from a manual to an automated process of consuming analysed data requires the use of scheduling software. Schedulers can be used to run specific data "jobs" or queries at regular predefined internals of seconds, days or weeks.

Meanwhile, existing hardware resources and platform operations must be effectively managed.

How do we do it?

The DataWRX© platform includes a scheduler to ensure data is computed automatically where required. It can handle complex scheduling involving thousands of jobs and large amounts of data.

Separate software is used to manage the many nodes required to run these jobs. Furthermore, DataWRX©'s integration with Apache Ambari helps users with the configuration, provisioning and monitoring of these node clusters.

What's the business value?

- ✓ Supports large e-commerce applications
- ✓ Supports regular risk or management reporting
- ✓ Available for use across multiple business units