

Infoworks

DataFoundry for
Databricks

ENTERPRISE DATA
OPERATIONS +
ORCHESTRATION

THE CHALLENGE

Leverage the Strategic Value of Your Company's Data

Businesses want to redefine their operations and customer experience to be competitive in an increasingly digital world. Winning in this environment is highly dependent on the ability of an organization to harness data effectively.

However, businesses struggle to keep up with the demand for new analytics use cases in support of their changing business models. The reality is that over 80% of big data projects fail

to deploy to production because project implementation is a complex, resource-intensive effort that takes months or even years. The technology is complicated and the people who have the necessary skills are either extremely expensive or difficult to find. In addition, the market is rapidly evolving with constantly changing technologies, while organizations simultaneously move from on-premise to cloud, multi-cloud or hybrid implementations, very often orchestrating their data and data pipelines across multiple environments.

THE SOLUTION

Enterprise Data Operations and Orchestration

Enterprise Data Operations and Orchestration (EDO2) refers to the systems and processes that enable businesses to organize and manage data from disparate sources and process the data for delivery to analytic applications. EDO2 systems aim at shorter development cycles, increased deployment frequency, and more dependable releases of data pipelines in close alignment with business objectives.

Infoworks DataFoundry is the only EDO2 software system that automates the development and operationalization of data pipelines from source to consumption in support of business intelligence (BI), machine learning (ML) and artificial intelligence (AI) analytics applications. Historically, data integration platforms have provided point solutions for each step in the development and management of data pipelines and workflows. In contrast, DataFoundry integrates these modules into a fully unified system running on the Databricks managed Spark platform that provides

a holistic and agile environment for delivering data, data pipelines and data workflows that scale elastically as your needs fluctuate.

DataFoundry for Databricks lays the foundation for the digital transformation of business, with a complete solution that provides:

- **Agility:** The fastest and most automated path to launch analytics use cases running on Databricks at scale.
- **Flexibility:** The only system that enables businesses to manage and orchestrate enterprise data operations in one venue and freely choose the best place to run specific applications without recoding.
- **End-to-end services:** The only integrated system to manage data operations and orchestration from data sources through to consumption by analytics applications.
- **Extensibility:** Architected to adapt to new business requirements and technologies.

INFOWORKS DATAFOUNDRY IN ACTION

Infoworks DataFoundry has been deployed in production by large enterprises to run business critical applications. Infoworks' customers have successfully implemented complex, large-scale analytics use cases in days instead of months with minimal resources. Some examples of these successes are:

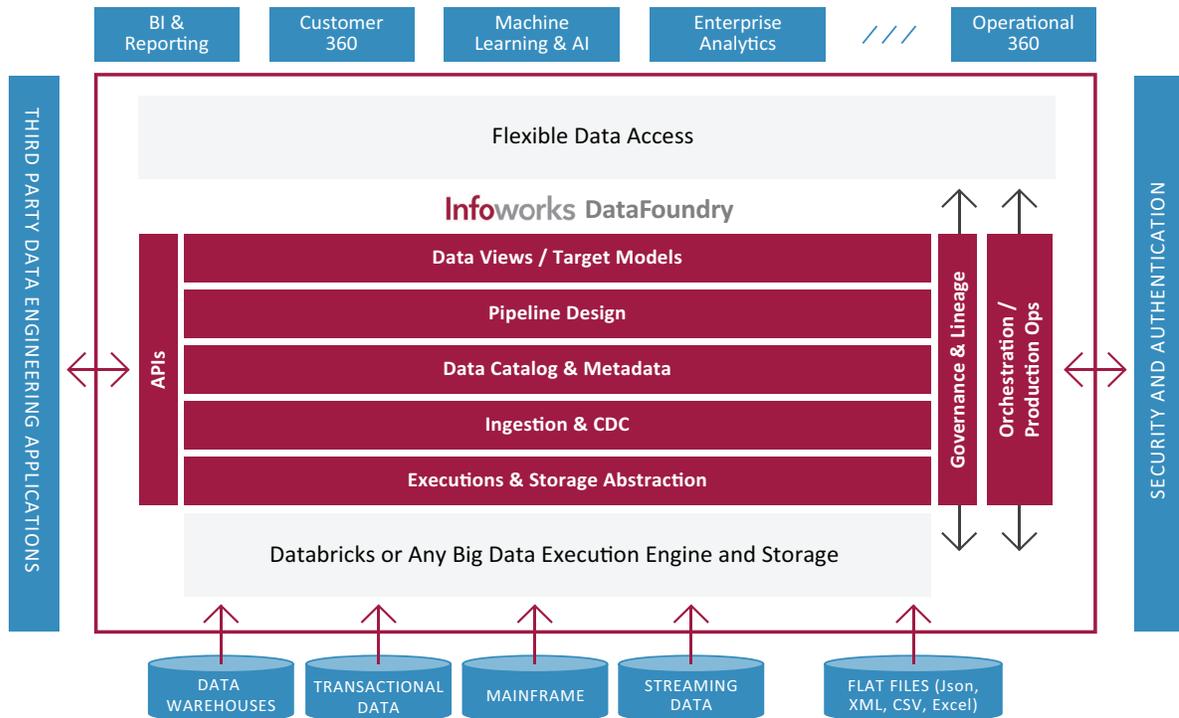
Fortune 10 Pharmacy: Claims Processing Implemented near-real-time, machine learning claims processing data flow in 19 days

- Synchronized business process data from Teradata every 10 minutes
- Achieved the data load window SLA of 15 minutes
- Implemented by 2 engineers in 19 days through production

Oil and Gas Exploration: Self-Service Analytics Implemented a streaming ingestion, near-real-time, predictive analytics solution in 2 weeks

- Streams data from oil wells in near real time
- Integrates streaming and historical data
- Built a real-time drilling dashboard with preventive maintenance analytics

ANY SOURCE, ANY BIG DATA PLATFORM, ANY ANALYTICS



DataFoundry Capabilities

Infoworks DataFoundry for Databricks provides a complete solution that automates end-to-end data workflows from source to consumption as well as the on-going operational management of those workflows. DataFoundry is an EDO2 system that is platform independent and delivers the following automated capabilities:

Data Ingestion and Synchronization

Data Ingestion

Ingests source data in a high-performance, parallel process, while automatically preserving data precision. DataFoundry provides a no-code environment for configuring the ingestion of data into your data lake via batch, change data capture (CDC), and data streaming.

Metadata Synchronization

Automatically crawls data sources, ranging from flat files, XML, and JSON to relational databases. Learns the metadata and infers data relationships for the ingested data from external data sources as well as for data sets created using Infoworks, making metadata searchable via a metadata repository.

Data Synchronization

Continuously synchronizes source data from enterprise databases, data warehouses, and file sources. Changing data is captured from the source systems using log-based and query-based methods. The changed data is merged with the base data in a high-performance continuous merge process.

- Automatically handles slow-changing data and schema changes and creates current and historical tables.
- Supports export functionality to other enterprise operational and data warehouse systems.

... cont'd on p.4

Data Catalog and Metadata Management

Provides the ability to search metadata about data sources, pipelines and workflows, relate business and technical metadata and identify the best artifacts for your data project.

Pipeline Design

Provides self-service data preparation using an interactive, drag-and-drop data transformation capability with support for SQL-based and other transformations. Users work with data in a collaborative, suggestion-based interface that reduces or eliminates dependence on IT skills. DataFoundry also makes it possible to integrate data pipelines with advanced analytics algorithms from libraries such as SparkML and R, without the need for coding. Build trained models or import pre-trained models into data pipelines.

SQL Migration

Automates migration of SQL workloads from legacy data warehouses to a big data environment. With automated data, schema and workload migration, DataFoundry accelerates data warehouse offload and migration and allows seamless migration across big data platforms, both on premise and in the cloud.

Orchestration and Production Operations Management

Design end-to-end workflows and orchestrate them in production with fault-tolerant, distributed execution that leverages Databricks' auto-scaling with overrideable parameterization. Migrate from development environments to production across big data or cloud platforms with single-click operation.

Execution and Storage Abstraction

Data workflows developed in DataFoundry can be run on Databricks and also have data stored natively in Databricks Deltalake or on any number of execution engines and storage platforms. Workflows can be migrated from an on-premise Hadoop platform to Databricks in the cloud, or from one cloud environment to another without recoding. DataFoundry automatically optimizes data workflows to run at scale on all supported execution engines.

Platform and Data APIs

All platform services and data are accessible via a wide variety of API access points.

Enterprise-Grade Security Integration

DataFoundry provides security integration for user authentication and role-based data security policies. It supports single-sign-on/LDAP integration, Kerberos based authorization as well as encryption for data in motion and at rest. DataFoundry also provides role and domain management for controlling access to data managed within the data fabric.

Governance and Lineage

DataFoundry tracks end-to-end data lineage so users can trace data elements back to the original source systems and perform downstream impact analysis. It also provides audit logs that track who has created or changed data pipelines, cubes and workflows as well as tracking what changes were made and when. DataFoundry supports the creation of users with different levels of user access, as well as domains, so administrators can control which users have access to specific data sets. Users within a domain can share data, pipelines, and workflows to enable team-based development of end-to-end data workflows and pipelines.

Infoworks



490 S. California Ave. Palo Alto, CA 94306



(650) 391-9306



info@infoworks.io

www.infoworks.io