Infoworks

ONBOARDING OVERVIEW

Infoworks DataFoundry for Databricks

Enterprise Data Operations and Orchestration for cloud analytics and machine learning at scale

Infoworks DataFoundry is the only automated Enterprise Data Operations and Orchestration (EDO2) system that runs natively on Databricks and leverages the full power of Databricks and Apache Spark to deliver the fastest and easiest solution to onboard data and launch analytics use cases on Databricks.

DataFoundry offers a comprehensive suite of functionality that covers the entire data workflow, enabling users to onboard, prepare, and operationalize data, and achieve unprecedented scale and agility in analytics.

Step 1: **Onboard Your Data**

Data onboarding is the critical first step in operationalizing your data lake. DataFoundry not only automates data ingestion, but also automates the key functionality that must accompany ingestion to establish a complete foundation for analytics. Data onboarding with DataFoundry automates:

1. **Data Ingestion** – from all enterprise and external data sources
2. **Data Synchronization** – CDC to keep data synchronized with the source
3. **Data Governance** – cataloging, data lineage, metadata management, audit, and history

Step 2: **Prepare Your Data**

DataFoundry automates preparing data for analytics and optimizing data pipelines for performance. Data preparation with DataFoundry applies intelligent automation to:

1. **Data Transformation** – Data pipeline design, optimization and incremental updates
2. **Data Modeling** – Use-case specific optimization of data models with incremental updates

Step 3: **Operationalize Your Data**

DataFoundry greatly simplifies deployment and management of analytics use cases in production by automating:

1. **Export and Migrate Data Pipelines** – from development to production
2. **Pipeline Orchestration** – Automated management of fault-tolerant analytic workflows
3. **Hybrid and Multi-Cloud Deployment** – Automated export or migrate data pipelines to target platforms on-premises or in the cloud

How Infoworks Onboarding Works

Data Ingestion

Infoworks DataFoundry automatically crawls data sources, ranging from relational databases and data warehouses such as Oracle, Teradata, SQLServer & others to flat files, XML, and JSON. 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.

Infoworks DataFoundry ingests source data in a high-performance, parallel process, while automatically creating type mapping to preserve source data precision. DataFoundry provides a no-code environment for configuring the ingestion of data into your delta lake via batch, change data capture and data streaming.

Data Synchronization

Infoworks DataFoundry 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 slowly changing data and schema changes and creates current and historical tables
- Supports incremental export functionality to other consumption data warehouse systems such as BigQuery, SQL Data warehouse and others

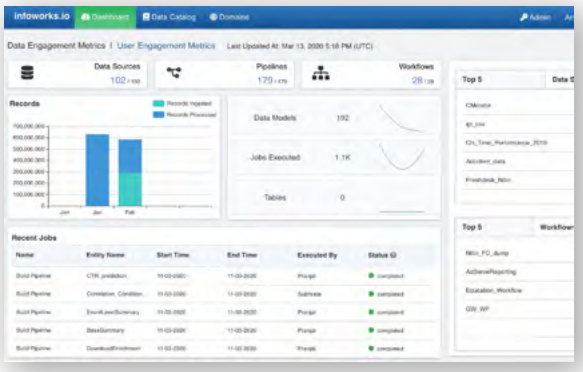
Governance and Lineage

Infoworks DataFoundry creates and synchronizes a Data Catalog that can be tagged and searched using the UI. It 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 raw data and semantic data. It also provides the ability to track changes to data pipelines and workflows that operate on the raw data.

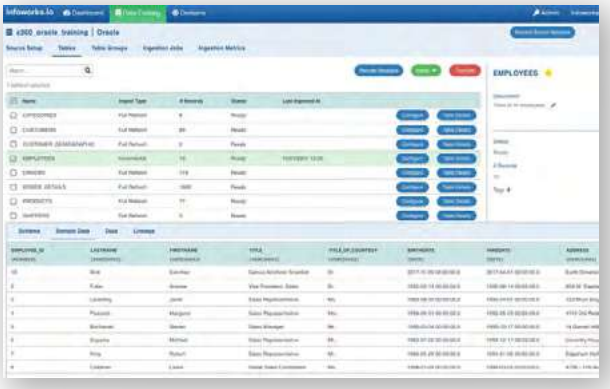
DataFoundry supports the creation of users with different levels of user access as well as domains, so administrators can control which users have access to specific data sets. Users within a domain can share data, pipelines, and workflows.

Infoworks Data Onboarding Runs Natively on Databricks

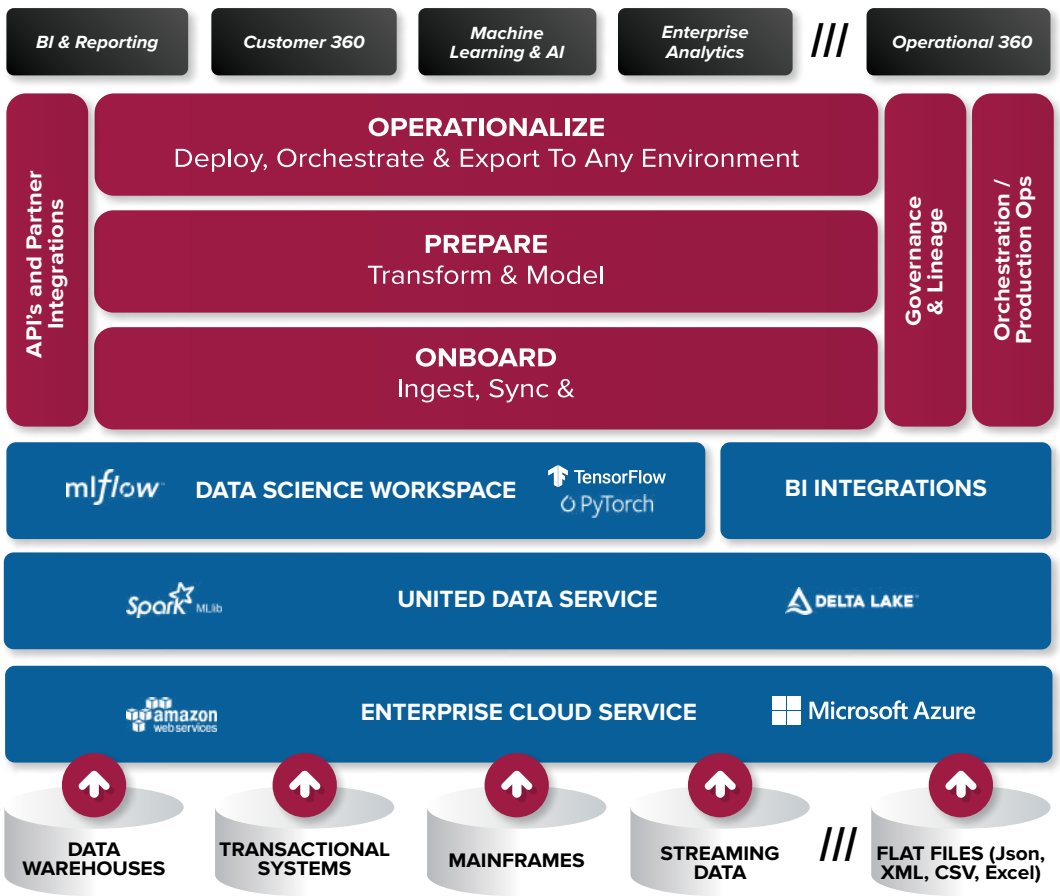
BENEFITS FOR DATABRICK USERS	DATA INGESTION & SYNCHRONIZATION CAPABILITIESA
Performance & Optimization	• All ingestion is run using Databricks Runtime processing (not JDBC), for better performance • Automated deployment of auto-scaled, on-demand clusters tailored for individual jobs and data sizes for easier optimization
Simplified Management	• Augments Delta Merge support, by automatically maintaining record versioning, for any time period • Data is natively stored in Delta Tables • Automatically prevents duplicate record errors in Delta merge auto-optimizes Delta Tables to • Overcome fragmentation due to multiple small files
Easy Access to Data	• All Delta tables created by Infoworks data onboarding are registered in the metastore, allowing easy SQL access via notebooks • Onboarded data is automatically cataloged



BENEFITS FOR DATABRICK USERS	DATA GOVERNANCE CAPABILITIES
Simplified Data Discoverability	• Automatically maintains a data catalog with business and technical metadata, for all ingested data, for easier data discovery by data engineers, data scientists and others
Built-in Data Governance	• Provides a single trusted view of all data assets in the Delta Lake • Avoids a data swamp, through automated metadata lineage, audit, and governance



Infoworks Onboarding Marchitecture



Infoworks DATAFOUNDRY
Enterprise Data Operations and Orchestration System

databricks
Unified Analytis Platform

Healthcare Company Reduces Operational Costs by 38%

PREVIOUS STATE

A leading healthcare company found it difficult to meet and prioritize service-level agreements (SLAs). Their Apache Hadoop-based analytics infrastructure was also causing operational costs to skyrocket.

WITH DATAFOUNDRY AND DATABRICKS

The healthcare company was able to migrate analytics workloads to Databricks Unified Analytics Platform rapidly and met all service level agreement (SLA) requirements and total cost of ownership (TCO) goals. Some of the key attributes of the migrated workloads and results delivered were:

- 42 source tables
- 24 data pipelines
- Migrated to Databricks from Azure HDInsights **in 7 hours**
- 38% reduction in ongoing operational costs
- 25% increase in query productivity
- Ability to control costs down to the individual data workflow
- Automated use of on-demand and elastic clusters
- Eliminated need for hand coding

An Extensive and Growing Set of Available Connectors

DATA INGESTION CONNECTORS

- Teradata
- Oracle
- MySQL
- Netezza
- MS SQL Server
- Salesforce.com
- Delimited Files (including mainframe copybooks)
- JSON
- DB2 (Linux, Unix, Windows and z/OS)
- Hive
- REST API (Generic and Custom)
- MariaDB
- MongoDB
- SAP Hana
- Apache Ignite
- Sybase IQ
- Redshift
- MapR DB
- Vertica
- BigQuery

MACHINE LEARNING LIBRARIES

- SparkML
- H2O.ai

EXPORT CONNECTORS

- Snowflake
- Azure Cosmos DB
- Azure SQL Data Warehouse
- Teradata
- Netezza
- Apache Ignite
- BigQuery
- Apache Phoenix
- Postgres
- CloudSQL

About Infoworks

Infoworks offers the most comprehensive and automated Enterprise Data Operations and Orchestration (EDO2) system. It is the only EDO2 system built to automate and accelerate deployment and orchestration of analytics projects at scale, in cloud, hybrid, multi-cloud, and premise-based environments. Through deep automation and a code-free environment, Infoworks empowers organizations to rapidly consolidate and organize enterprise data, create analytics workflows and deploy projects to production within days – dramatically increasing business agility and accelerating time-to-value. Infoworks counts some of the world's largest financial, retail, technology, healthcare, oil & gas, and manufacturing companies as its customers. To learn more, please visit infoworks.io.

ENTEPRISE DATA OPERATIONS AND ORCHESTRATION FOR BUSINESS TRANSFORMATION

sales@infoworks.io | (650) 391-9306 | infoworks.io

April 2020