How-To Library

Informatica

New Features and Enhancements in Big Data Management® 10.2.1

Abstract

This article describes new features and enhancements in Informatica Big Data Management 10.2.1. The new features and enhancements centered around three key areas: enterprise class, advanced Spark, and cloud and serverless.

Supported Versions

- Big Data Management 10.2.1

Table of Contents

Overview

Version 10.2.1 enhances the capabilities of Big Data Management to cover enterprise solutions, data integration, and cloud ecosystems. Informatica focuses on three major feature categories for Big Data Management 10.2.1:

Enterprise class

Informatica's focus is to improve the core Big Data Management platform to support scalability for large enterprise projects.

Advanced Spark

Informatica aims to enhance the quality and stability of the Spark engine to establish Spark as the primary engine for data processing. In version 10.2.1, Spark is the target engine for data science use cases, enhanced data integration use cases, and enhanced hierarchical data processing.

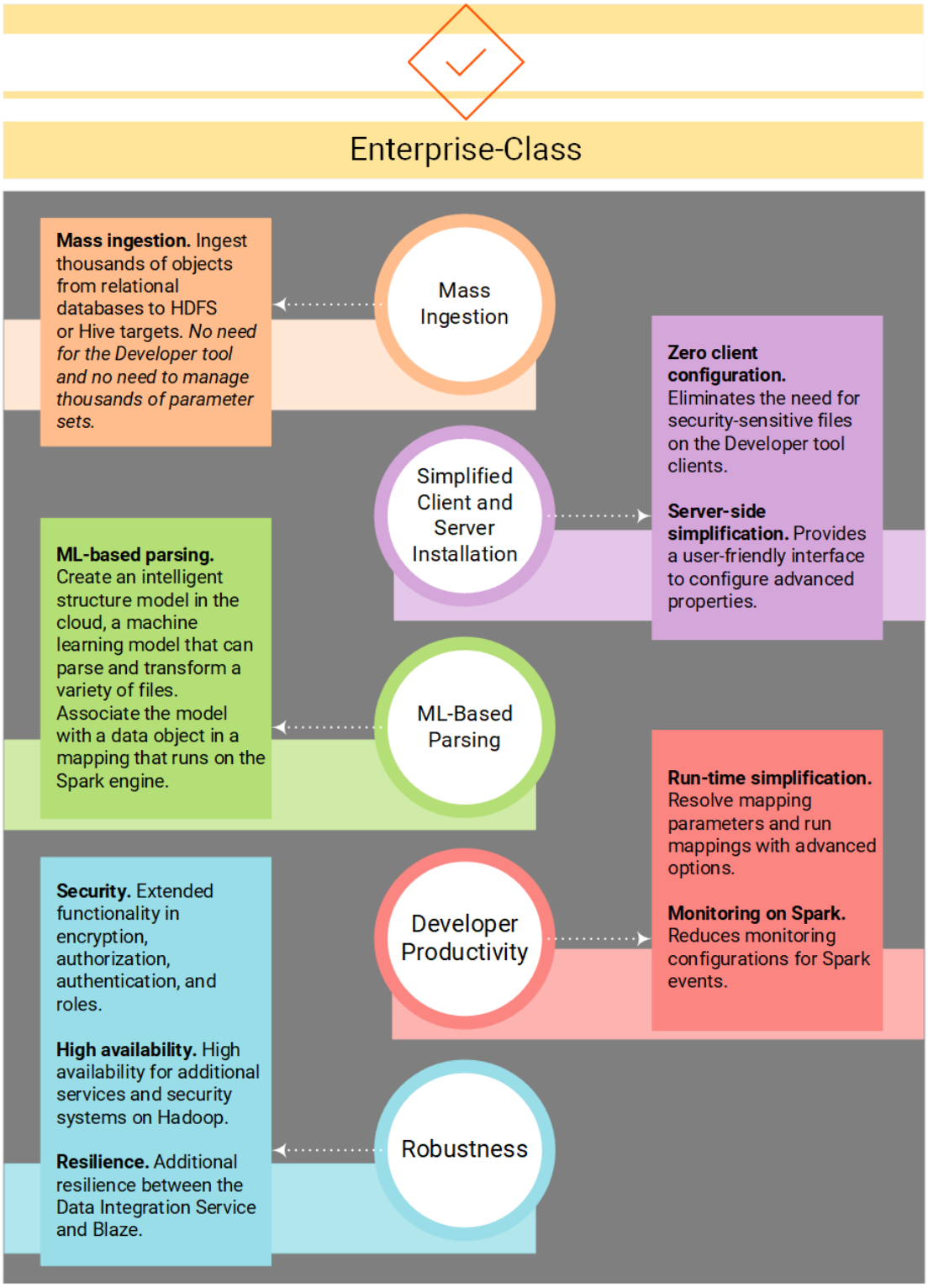
Cloud and serverless

Informatica is reinforcing Big Data Management as a key player in cloud ecosystems. Big Data Management has enabled ephemeral clusters on Amazon EMR and Azure HDInsight distributions and improved connectivity to Amazon and Microsoft Azure environments. Big Data Management has also introduced features that automate environment configurations to create serverless environments.

Enterprise-Class

This section describes new enterprise-class features in version 10.2.1.

The following diagram summarizes these features:



Mass Ingestion

Mass ingestion is a new big data solution in version 10.2.1. You can use mass ingestion to ingest thousands of relational tables from a relational database to a data lake such as Amazon S3 or Microsoft ADLS.

To perform mass ingestion jobs, you use the Mass Ingestion tool. The Mass Ingestion tool provides a non-technical interface that you can use to create a mass ingestion specification. In the mass ingestion specification, you can configure how you want to ingest the data and specify parameters to cleanse the data that you ingest.

A mass ingestion specification replaces the need to manually create and run mappings. You can create one mass ingestion specification that ingests all of the data at once and operationalizes the ingestion process.

For more information on mass ingestion, see the Informatica Big Data Management 10.2.1 Mass Ingestion Guide.

Simplified Client and Server Installation

Developers can leverage the Metadata Access Service to import metadata from Hadoop clusters without configuring Kerberos keytab files or configuration files on individual workstations.

Version 10.2.1 also provides enhancements to Hadoop connections and cluster configuration objects to improve usability and enable advanced configurations. The Hadoop connection is reworked to allow you to configure advanced properties on the Blaze, Spark, and Hive engines. You no longer have to configure the hadoopEnv.props file. You can also configure the Hadoop distribution directly in the Hadoop cluster configuration properties.

The changes eliminate the need to store sensitive files on client workstations, and they reduce client footprint

For more information, see the Informatica Big Data Management 10.2.1 Hadoop Integration Guide.

Machine Learning-Based Parsing

Version 10.2.1 extends support for the intelligent structure model into Big Data Management. An intelligent structure model is a machine learning model that can parse and transform a variety of files such as clickstream, log, text, CSV, and Excel files. You can incorporate the intelligent structure model in an Amazon S3, Microsoft Azure Blob, or complex file data object in a mapping that runs on the Spark engine.

For more information, see the "Processing Unstructured and Semi-Structured Data with an Intelligent Structure Model" chapter in the Informatica Big Data Management 10.2.1 User Guide.

Developer Productivity

Version 10.2.1 contains enhancements to simplify the effort for developers to create, run, and monitor mappings.

Run-time Simplification

Developers can resolve mapping parameters before running a mapping. The resolved parameter view helps developers manage dynamic mappings that run using many parameter sets or parameter files. In the resolved parameter view, you can preview the parameters and validate the mapping to ensure that parameters are valid at run time.

Developers can also run mappings using advanced options. As a developer, you can use advanced options to complete the following tasks:

- Run a mapping with a parameter set or a parameter file.
- Choose a reusable mapping configuration or create a custom mapping configuration.
- Set tracing and optimizer levels to run the mapping.
- Run the mapping on a specific Data Integration Service.

For more information, see the "Mapping Parameters" chapter in the Informatica 10.2.1 Developer Mapping Guide.

Monitoring on the Spark Engine

The Spark engine has fewer configurations, and you do not need to configure the Spark monitoring port. The Data Integration Service has a range of available ports, and the Spark executor selects a port from the available range. The Spark executor listens on the port for Spark events.

For more information on Spark monitoring, see the "Monitoring Mappings in the Hadoop Environment" chapter in the Informatica Big Data Management 10.2.1 User Guide.

Robustness

Version 10.2.1 provides enhancements in security, high availability, and resilience to harden your environment.

Security

The following new security features are available in version 10.2.1:

- Cloudera Navigator Encrypt. Use Cloudera Navigator Encrypt to secure data and implement transparent encryption of data at rest.
- EMR File System Authorization. Use EMR File System (EMRFS) authorization to access data in Amazon S3 on the Spark engine.
- IAM Roles. Use IAM roles for EMR File System to read and write data from the cluster to Amazon S3 in Amazon EMR cluster version 5.10.
- Kerberos. Enable Kerberos authentication for Amazon EMR clusters and Azure HDInsight clusters with WASB storage.
- LDAP. Configure LDAP authentication for Amazon EMR cluster version 5.10.
- Sqoop. Sqoop honors all high availability and security features such as Keberos keytab login and KMS encryption that the Spark engine supports.

High Availability

You can enable high availability for the following services and security systems in the Hadoop environment on Cloudera CDH, Hortonworks HDP, and MapR Hadoop distributions:

- Apache Ranger
- Apache Ranger KMS
- Apache Sentry
- Cloudera Navigator Encrypt
- HBase
- Hive Metastore
- HiveServer2
- Name node
- Resource Manager

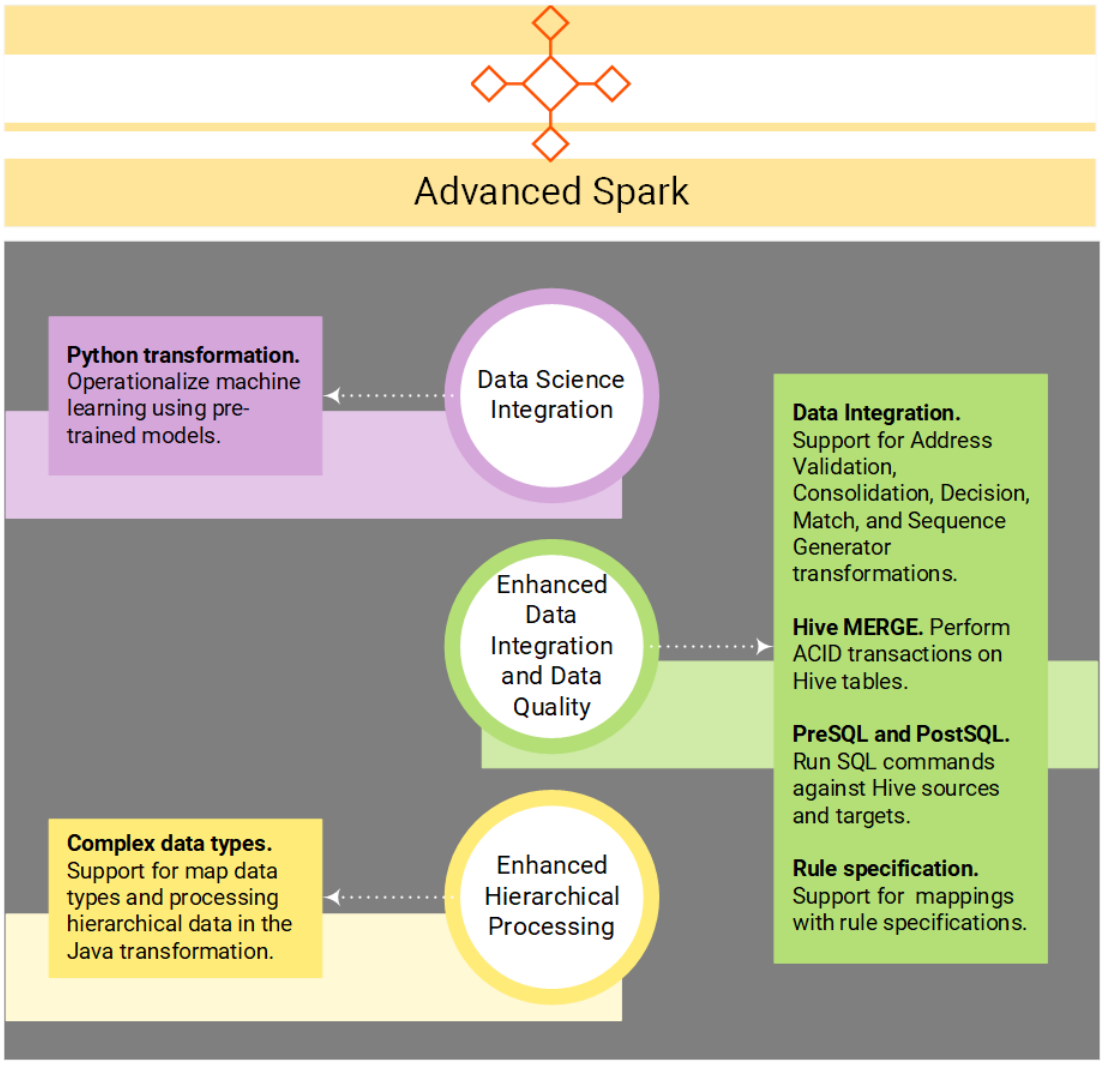
Resilience

There is additional resilience between the Data Integration Service and the Blaze Grid Manager and Blaze Orchestrator. The Data Integration Service can now reconnect to the Blaze services during temporary connection failures.

Advanced Spark

This section describes new advanced functionality on the Spark engine in version 10.2.1.

The following diagram summarizes these features:



Data Science Integration

Version 10.2.1 expands on the capabilities of the Developer tool to cover data science use cases by operationalizing machine learning.

To operationalize machine learning, you can use the new Python transformation. In the Python transformation, you can specify a pre-trained model and the Python code that calls the pre-trained model. You can execute the Python transformation in a mapping that runs on the Spark engine.

The pre-trained model can classify input data or create predictions based on the data that you pass to the Python transformation. For example, you can pass the classic Titanic data set to the transformation to predict the passengers that survived or did not survive and compare the predictions to the outcomes.

For more information on the Python transformation, see the "Python Transformation" chapter in the Informatica 10.2.1 Developer Transformation Guide.

Enhanced Data Integration and Data Quality

The Spark engine provides a variety of new functionality that enables end-to-end data integration and data quality use cases.

The Spark engine supports the following new functionality:

- Address Validator, Consolidation, Decision, Match, and Sequence Generator transformations.
- Hive MERGE statements to perform ACID transactions on Hive tables.
- PreSQL and PostSQL commands for Hive data sources.
- Rule specifications in mappings.

For more information, see the Informatica Big Data Management 10.2.1 User Guide.

Enhanced Hierarchical Data Processing

The Spark engine supports the following new functionality to process hierarchical data:

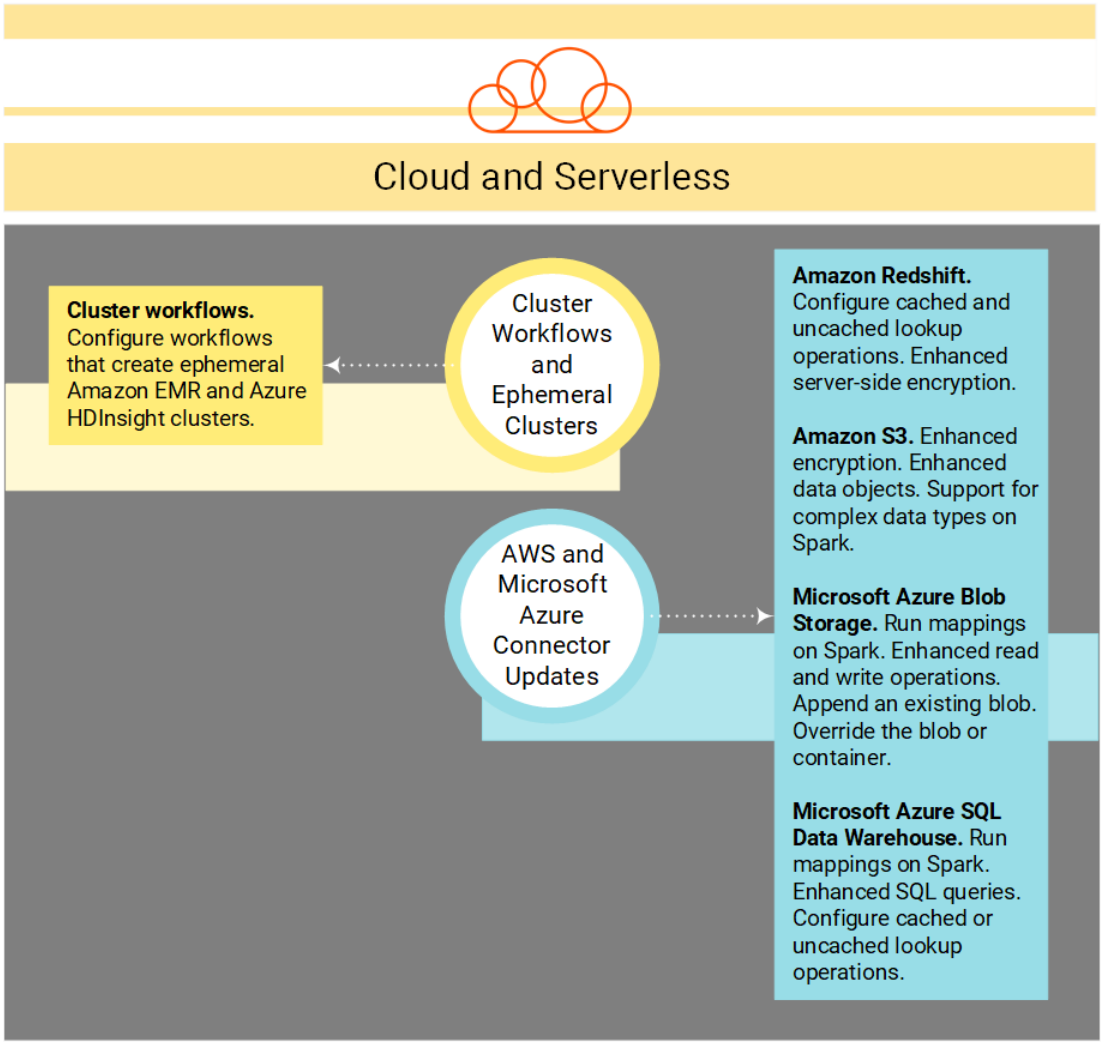
- Process map data types to generate and process map data in complex files.
- Use complex data types to read and write hierarchical data in Avro and Parquet files on Amazon S3.
- Process complex data types in the Java transformation.

For more information, see the "Processing Hierarchical Data on the Spark Engine" chapter in the Informatica Big Data Management 10.2.1 User Guide.

Cloud and Serverless

This section describes new features in cloud ecosystems in version 10.2.1.

The following diagram summarizes these features:



Cluster Workflows and Ephemeral Clusters

You can develop a cluster workflow that creates an ephemeral Amazon EMR or Azure HDInsight cluster on a cloud platform such as AWS or Microsoft Azure. You can choose to terminate and delete the cluster when workflow tasks are complete to save cluster resources.

When you run the cluster workflow on the Spark engine, you can also enable auto-scaling rules on the cluster. The auto-scaling rules add or remove nodes from the cluster depending on the amount of resources that workflow tasks require.

Cluster workflows set up a serverless architecture in the cloud environment that does not require you to provision, scale, or manage servers.

For more information, see the "Cluster Workflows" chapter in the Informatica Big Data Management 10.2.1 User Guide.

AWS and Microsoft Azure Connector Updates

The following adapters have functional, usability, and performance updates:

- Amazon Redshift
- Amazon S3
- Microsoft Azure Blob Storage
- Microsoft Azure SQL Data Warehouse

Amazon Redshift

PowerExchange for Amazon Redshift includes the following functionality:

- Configure a cached lookup operation to cache the lookup table on the Spark engine and an uncached lookup operation in the native environment.
- For a server-side encryption, configure the customer master key ID generated by AWS Key Management Service in the connection in the native environment and Spark engine.

For more information, see the Informatica PowerExchange for Amazon Redshift 10.2.1 User Guide.

Amazon S3

PowerExchange for Amazon S3 includes the following functionality:

- Configure the customer master key ID generated by AWS Key Management Service (KMS) in the connection in the native environment for a client-side or server-side encryption and on the Spark engine for a server-side encryption.
- Configure the Amazon S3-managed encryption key or AWS KMS-managed customer master key to encrypt the data while uploading the files to the buckets .
- Create an Amazon S3 file data objects from an intelligent structure model, a JSON file, or an ORC file.
- Compress an ORC data in the Zlib compression format when you write data to Amazon S3 in the native environment and Spark engine.
- Create an Amazon S3 target using the Create Target option in the target session properties.
- Use complex data types on the Spark engine to read and write hierarchical data in the Avro and Parquet file formats.
- Use Amazon S3 sources as dynamic sources in a mapping. Dynamic mapping support for PowerExchange for Amazon S3 sources is available for technical preview.

For more information, see the Informatica PowerExchange for Amazon S3 10.2.1 User Guide.

Microsoft Azure Blob Storage

PowerExchange for Microsoft Azure Blob Storage includes the following functionality:

- Run mappings on the Spark engine.
- Read and write CSV, Avro, and Parquet files when you run a mapping on the Spark engine and in the native environment.
- Read and write JSON and intelligent structure files when you run a mapping on the Spark engine.
- Read a directory when you run a mapping on the Spark engine.
- Generate or skip header rows when you run a mapping in the native environment. On the Spark engine, the header row is created by default.
- Append an existing blob. The append operation is applicable to only to the append blob and in the native environment.

- Override the blob or container name. In the Blob Container Override field, specify the container name or sub-folders in the root container with the absolute path.
- Read and write CSV files compressed in GZIP format.

For more information, see the Informatica PowerExchange for Microsoft Azure Blob Storage 10.2.1 User Guide.

Microsoft Azure SQL Data Warehouse

PowerExchange for Microsoft Azure SQL Data Warehouse includes the following features:

- Run mappings on the Spark engine.
- Configure key range partitioning when you read data from Microsoft Azure SQL Data Warehouse objects.
- Override the SQL query and define constraints when you read data from a Microsoft Azure SQL Data Warehouse object.
- Configure PreSQL and PostSQL queries for source and target objects in a mapping.
- Configure the native expression filter for the source data object operation.
- Perform update, upsert, and delete operations against Microsoft Azure SQL Data Warehouse tables.
- Configure a cached lookup operation to cache the lookup table on the Spark engine and an uncached lookup operation in the native environment.

For more information, see the
Informatica PowerExchange for Microsoft Azure SQL Data Warehouse 10.2.1 User Guide.

Authors

Margarita Pelyushenko