Aug 31, 2021

Intelligent Structure Discovery in CDI - Implementation and Use cases

Nalin Yadav

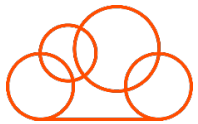
Solutions Architect, Customer Success Management

Informatica

Housekeeping Tips

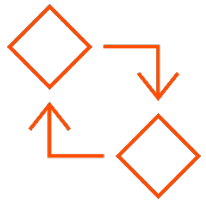
- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our INFASupport YouTube channel and Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Feature Rich Success Portal



Bootstrap trial and POC Customers

Enriched Customer Onboarding experience



Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

Informatica

More Information

Success Portal

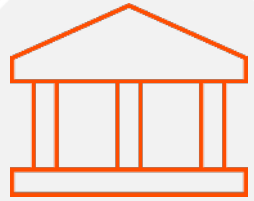
https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com



University

https://www.informatica.com/in/services-and-training/informatica-university.html

Informatica

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Agenda

- Introduction to Intelligent Structure Discovery (ISD)
- Intelligent structure model example with sample files
- Intelligent Structure Discovery process
- Benefits, Uses, Use Case Example, Mapping compatibility
- Data Drifts & Unassigned data
- Refine a discovered structure
- Demo
- Q&A

Introduction

Intelligent Structure Discovery (ISD)

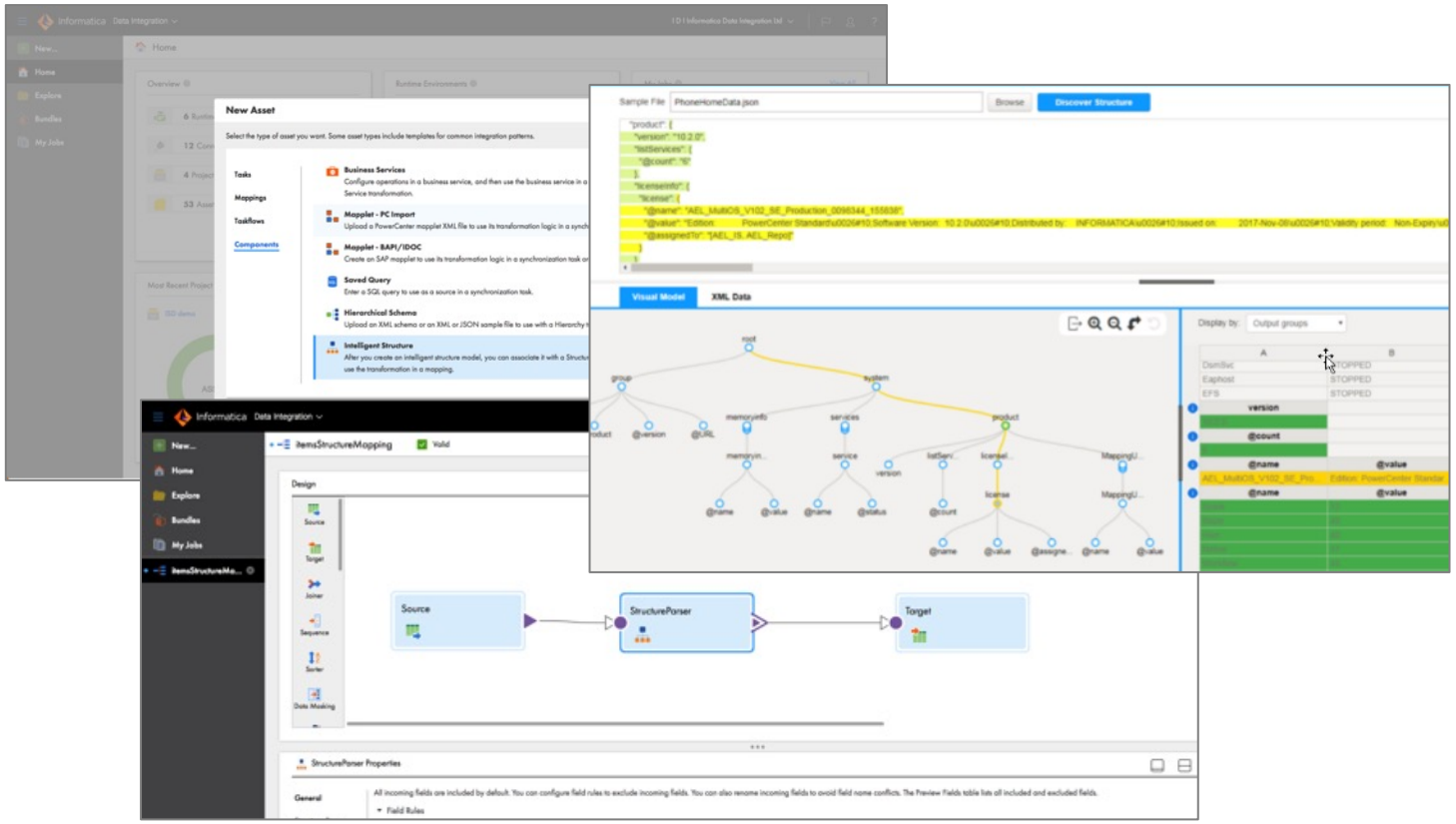
- Automates file ingestion processing and is powered by the Informatica CLAIRE® AI engine to discover and parse complex files
- Determines the underlying patterns and structures of the input that you provide for the model and creates a model that can be used to transform, parse, and generate output groups
- Provides out-of-the-box support for a variety of data file formats including clickstreams, IoT log, CSV, text delimited, XML, JSON, Excel, ORC, Parquet, Avro, PDF forms, and Word table files
- Long, complex files with little or no structure can be difficult to parse. Intelligent Structure Discovery can automatically decipher input data and discover the patterns, repetitions, relationships, and types of data in unstructured files
- After intelligent structure discovers the structure of the data, you can refine and test the structure, and then save or export it. When you save or export an intelligent structure, Intelligent Structure Discovery creates an **Intelligent Structure Model (ISM)** in an .amodel file

Intelligent Structure Discovery

Using CLAIRE™ to automate complex file processing and performance

POWERED BY CLAIRE™

- Machine learning to discover and parse complex files
- Simplify complex business processes
- High performance loading with advanced partitioning and push-down optimizations



Informatica

Intelligent Structure Model

Intelligent Structure Models

- A CLAIRE® intelligent structure model is an asset that Intelligent Structure Discovery creates based on input that represents the data that you expect the model to parse at run time
- Intelligent Structure Discovery creates a model that expresses the expected output data. You can use an intelligent structure model in mappings to parse unstructured, semi-structured, or structured data

Can create models from the following input types

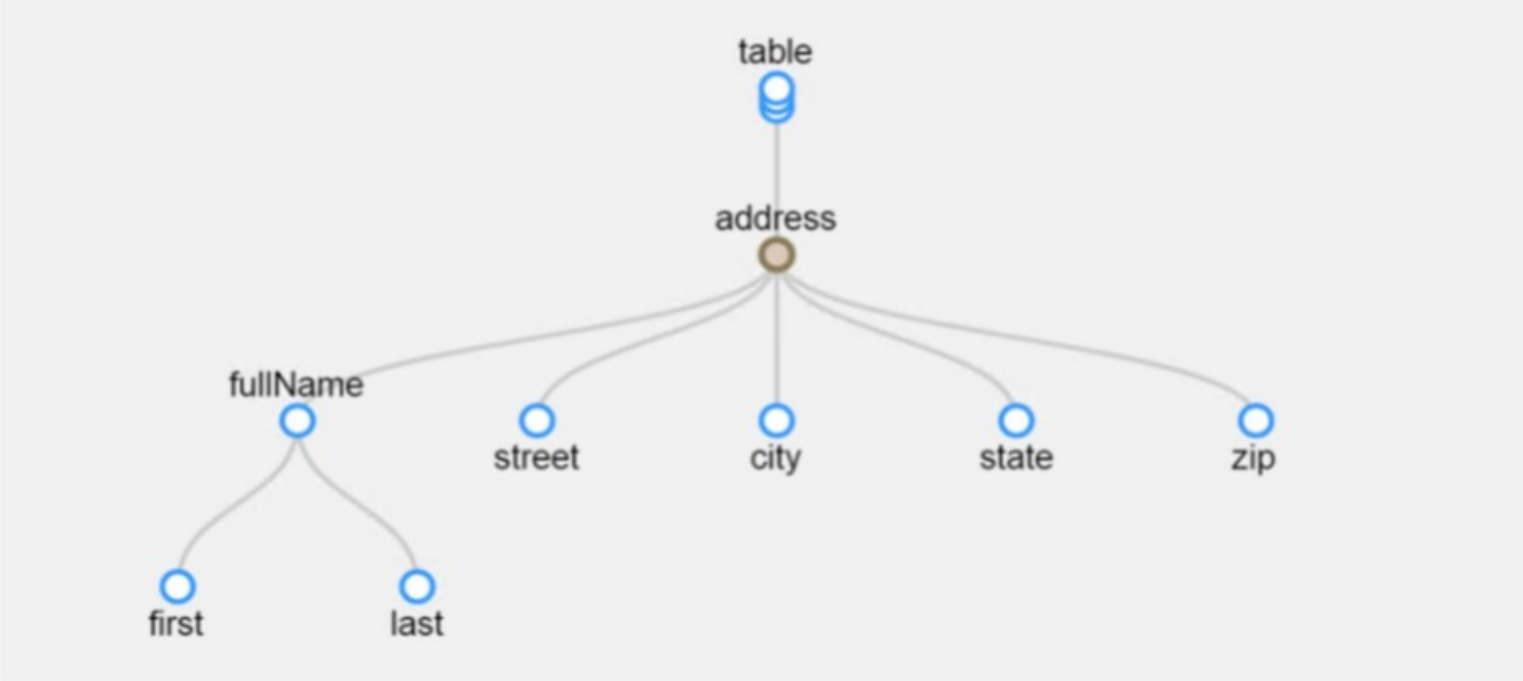
- Delimited files, for example, CSV files
- JSON files
- XML files
- ORC files
- Avro files
- Parquet files
- Microsoft Excel files
- Data within PDF form fields
- Data within Microsoft Word tables
- Machine generated files such as weblogs and clickstreams

Intelligent Structure Models: Example

As an example, you want to create a model for a CSV input file that contains the following content

```
first,last,street,city,state,zip
Carrine,Stone,17 Torrence Street,Livingston,PA,10173
Poona,Tillkup,52 Perez Avenue,Livingston,PA,10256
Tasha,Herrera,158 Shiraz Boulevard,Kensington,WA,33823
John,Washington,22A Zangville Drive,Tucson,AZ,20198
Jane Hochuli 4483 Central Street Suite 30 Phoenix PA 38721
```

The following image shows the structure that ISD discovers based on the input file

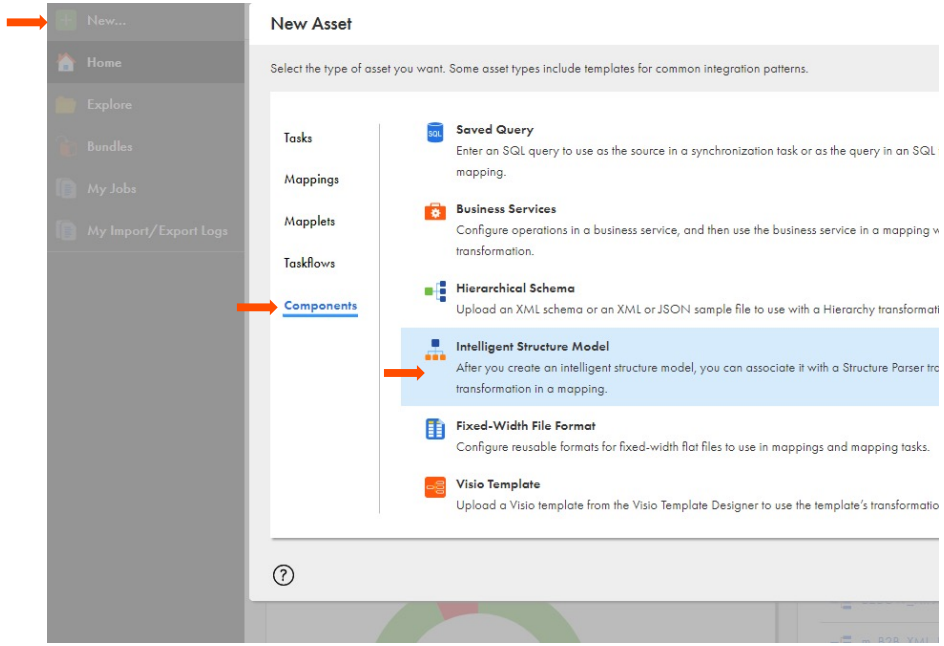


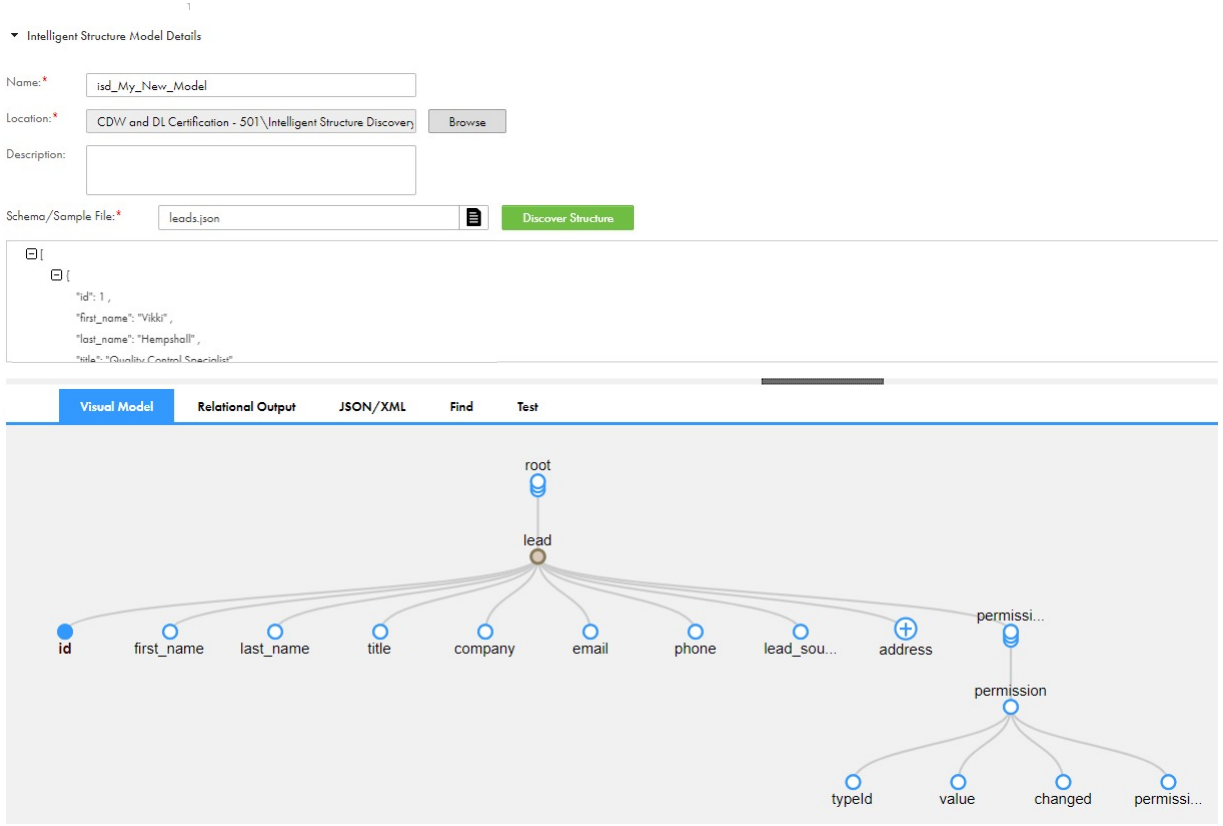
- ISD creates an intelligent structure model based on the structure of the input data that you provide
- ISD Created nodes representing the fields in the input file, such as first, last, street, city, state, and zip
- Intelligent Structure Discovery also recognized that the data as a whole represented addresses. The data is grouped under a parent node address

Informatica

Creating Intelligent Structure Models

How to create models with automatic structure detection

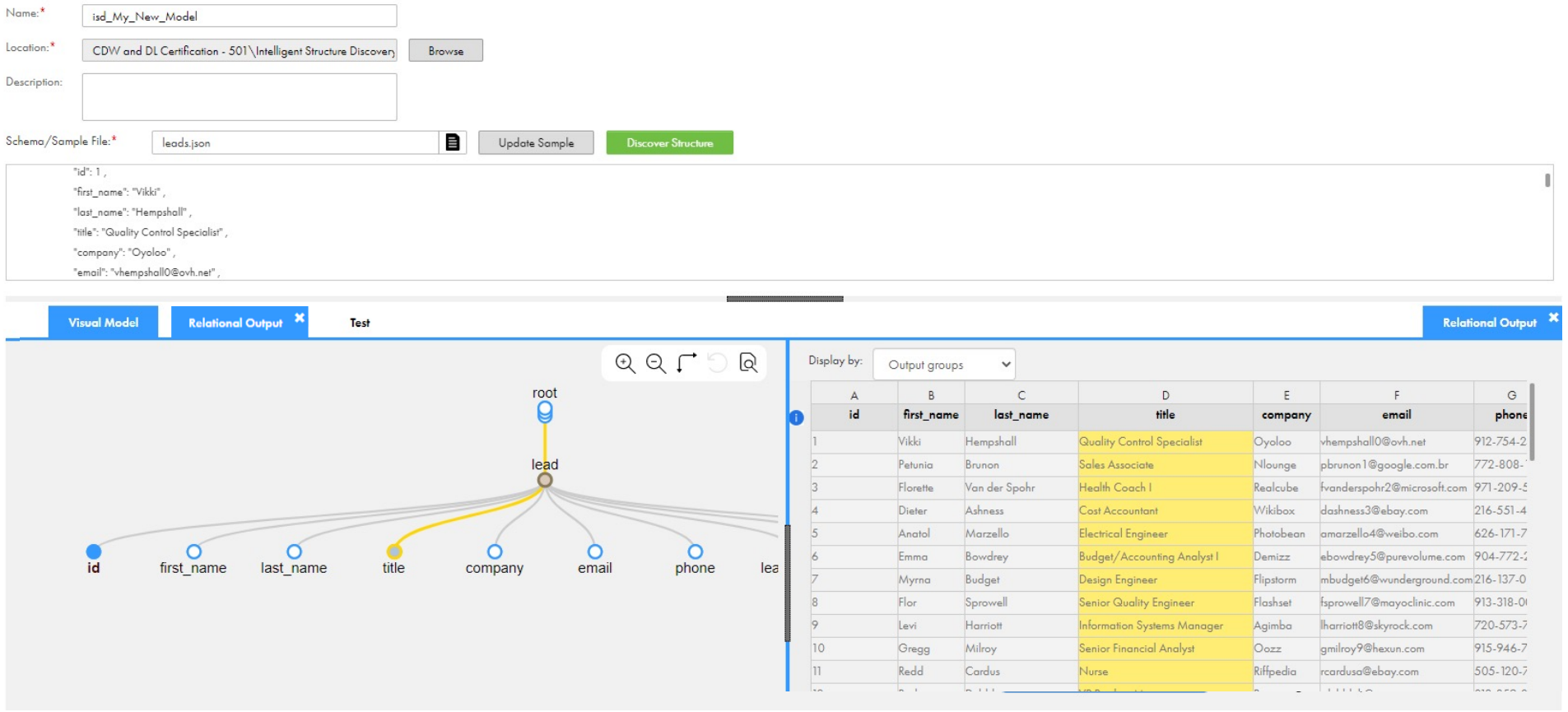




Informatica

Working with Intelligent Structure Models

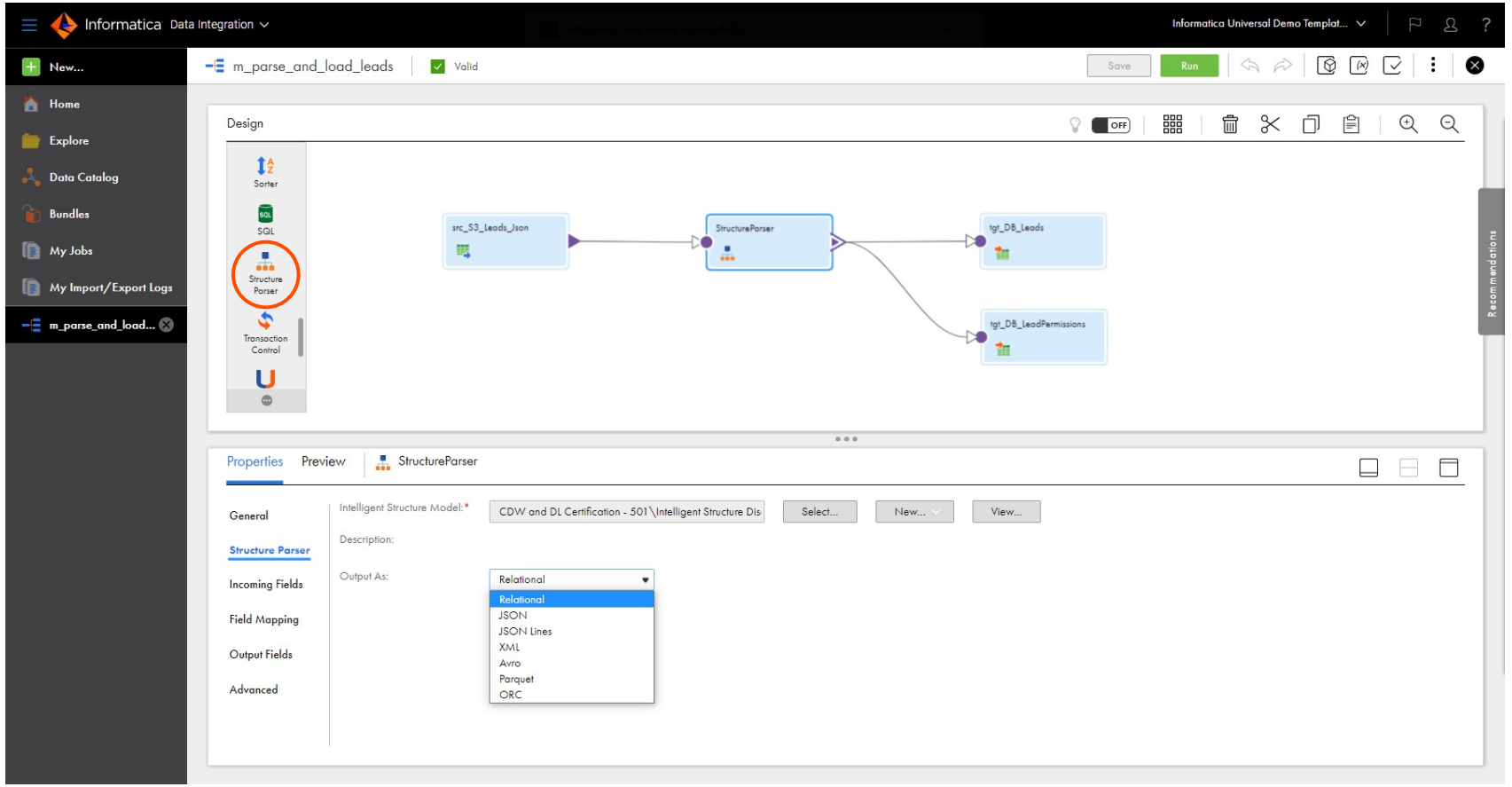
Modifying the automatically detected structure



Informatica

Deploying Intelligent Structure Models

Leverage ISD inside Informatica Cloud Mappings



Informatica

ISD: Process

Intelligent Structure Discovery process

- Once we provide an Input file, ISD Determines the underlying and repeating patterns of the data and creates a structure that represents the fields of data and their relationships
- We can quickly model data for files whose structure is very hard, time consuming, and costly to find, such as log files, clickstreams, customer web access, error text files, or other internet, sensor, or device data that does not follow industry standard

Benefits, Uses, Mapping compatibility, Use Case Example

Benefits

- Leverage AI, machine learning, and auto-generated visual models to more quickly understand machine data and prepare it
- Empower teams to be more agile with machine data through advanced tools that assist through automation
- Easily put auto-generated parsers for machine data and other log files to work in Informatica Intelligent Cloud Services
- Leverage high-performance loading with advanced partitioning and push-down optimization

Uses

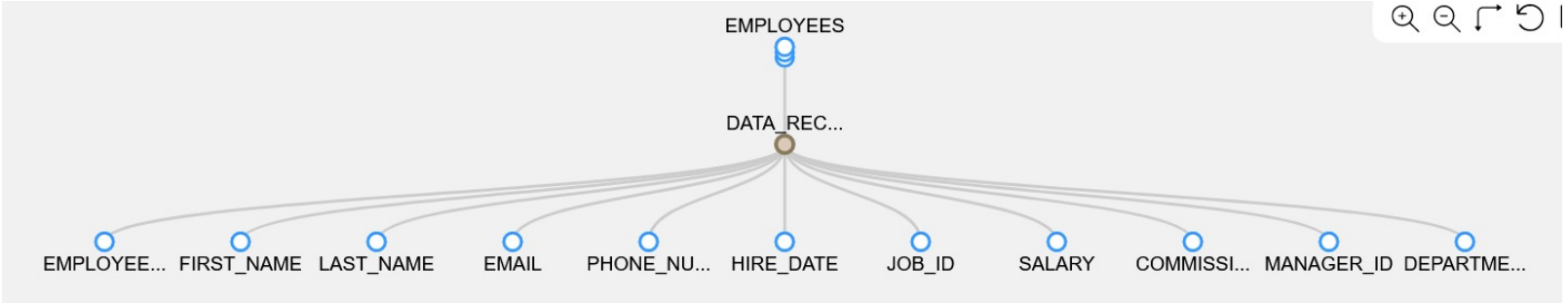
- Long, Complex files with little or no structure can be easily parsed
- Can automatically decipher input data and discover the patterns, repetitions, relationships and types of data in unstructured files
- Can use in mapping to parse unstructured, semi-structured or structured data

Use Case Example

Input Data

```
<?xml version="1.0" ?>
<EMPLOYEES>
  <DATA_RECORD>
    <EMPLOYEE_ID>100</EMPLOYEE_ID>
    <FIRST_NAME>Steven</FIRST_NAME>
    <LAST_NAME>King</LAST_NAME>
    <EMAIL>SKING@abc.com</EMAIL>
    <PHONE_NUMBER>515.123.4567</PHONE_NUMBER>
    <HIRE_DATE>2003-06-17 00:00:00</HIRE_DATE>
    <JOB_ID>AD_PRES</JOB_ID>
    <SALARY>24,000</SALARY>
    <COMMISSION_PCT></COMMISSION_PCT>
    <DEPARTMENT_ID>90</DEPARTMENT_ID>
  </DATA_RECORD>
  <DATA_RECORD>
    <EMPLOYEE_ID>101</EMPLOYEE_ID>
    <FIRST_NAME>Neena</FIRST_NAME>
    <LAST_NAME>Kochhar</LAST_NAME>
    <EMAIL>NKOCHHAR@abc.com</EMAIL>
    <PHONE_NUMBER>515.123.4568</PHONE_NUMBER>
    <HIRE_DATE>2005-09-21 00:00:00</HIRE_DATE>
    <JOB_ID>AD_VP</JOB_ID>
    <SALARY>17,000</SALARY>
    <COMMISSION_PCT></COMMISSION_PCT>
    <MANAGER_ID>100</MANAGER_ID>
    <DEPARTMENT_ID>90</DEPARTMENT_ID>
  </DATA_RECORD>
</EMPLOYEES>
```

Generated Model



Observation:

- MANAGER_ID is missing in first record but exist in second record. Intelligent Structure discovery identified the pattern and generated correct ISM for the scenario

Informatica

Mapping Compatibility

- Cloud Data Integration (CDI) Mapping
- Cloud Data Integration (CDI) Elastic Mapping
- Data Engineering mapping
- B2B Gateway

Data Drifts & Unassigned data

Data Drift

Sample Data use to create Intelligent structure model

05967|2014-09-19|04:49:50.476|51.88.6.206|custid=83834785|cntry=Tanzania|city=Mtwango|movie={b1027374-6eec-4568-8af6-6c037d828c66|"Touch of Evil"}|paid=true
01357|2014-11-13|18:07:57.441|88.2.218.236|custid=41834772|movie={01924cd3-87f4-4492-b26c-268342e87eaf|"The Good, the Bad and the Ugly"}|paid=true
00873|2014-06-14|09:16:14.522|134.254.152.84|custid=58770178|movie={cd381236-53bd-4119-b2ce-315dae932782|"Donnie Darko"}|paid=true

The input data that the model parses contains the following text

0448|2015-04-07|01:50:5.35|27.248.247.174|custid=613068|cntry=Iran|city=SarÄ•b|movie={50fb37b-621-484e-a565-2b5c1cbdc43|"Network"}|paid=false|ua=Mozilla/5.0 (Windows NT 5.1)
02780|2014-12-28|08:14:58.685|17.2.236.233|custid=731|cntry=Greece|city=NÃ©a RÃ³da|movie={1876aea0-3cb5-4c7a-22f-d33f233210|"Full Metal Jacket"}|paid=true|ua=Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_1)
03353|2015-04-20|21:02:40.532|143.48.11.171|custid=83736441|cntry=Russia|city=Mozhaysk|movie={67272f85-bfc-418a-82ea-a7c4ae6b028a|"Gangs of Wasseypur"}|paid=true|ua=Mozilla/5.0 (iPad; CPU OS 5_1 like Mac OS X)

Unassigned Data

Intelligent Structure Discovery assigns data to an Unassigned Data field in the following case

- When records exceed the maximum record size. The default maximum record size is 640,000 bytes
  We can increase the maximum record size by configuring one of the DTM JVM properties of the Data Integration Server service in Administrator

- When delimited files, for example CSV files or logs, contain more elements than expected

  Model created using below Data:
  computer ID, computer IP address, access URL, username, password, and access timestamp

  Input File has below Data:
  computer ID, computer name, computer IP address, country of origin, access URL, username, password, access code, and access timestamp

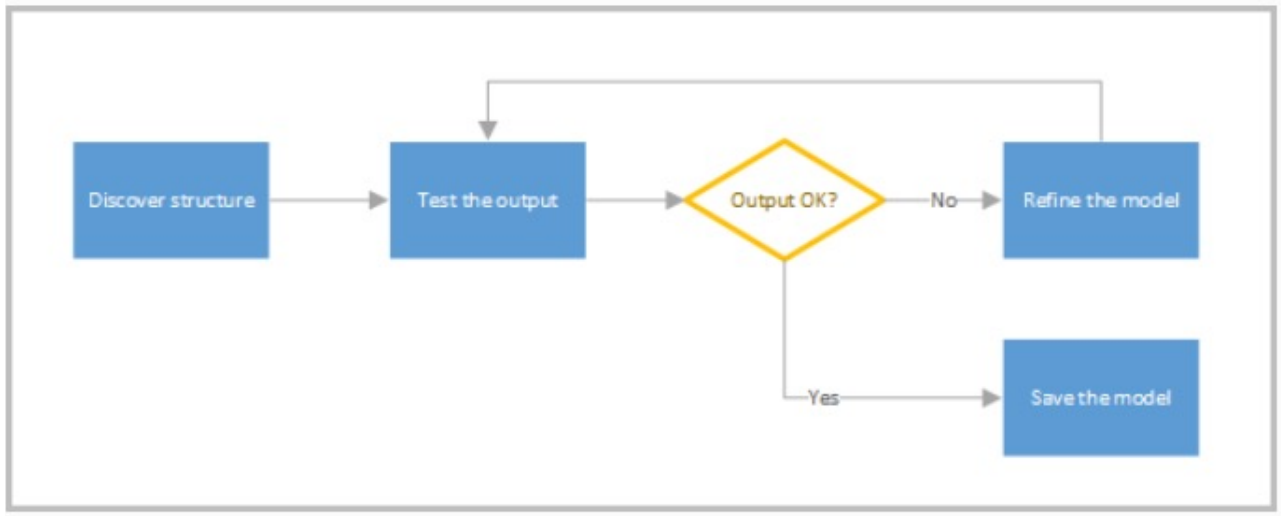
Refine a discovered structure

Refine a discovered structure

- After Intelligent Structure Discovery discovers the structure of the model input, you can refine the structure so that when you use the model in production the output meets your requirements

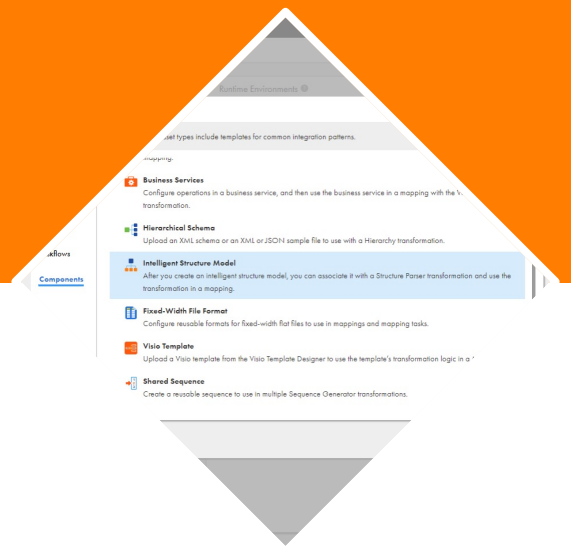
  For example, rename a node, add a prefix to file names, or define rows or columns in Microsoft Excel files as table headers.

- When you refine the discovered structure, you can test the output that Intelligent Structure Discovery generates based on the model in different output formats. Repeat refining and testing the model until you are satisfied that it generates the required output



Demonstration

Topics Covered

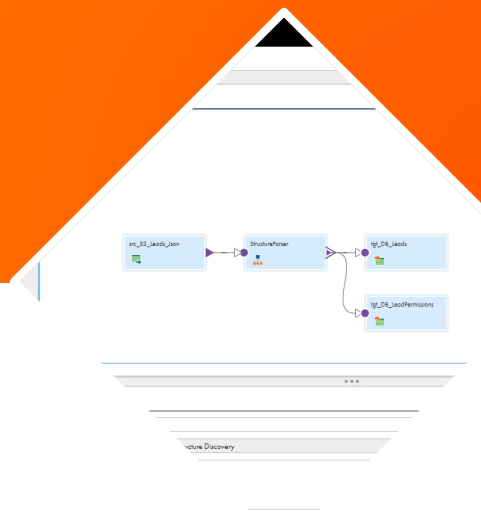


Creating ISD Models

- Upload a sample file
- A machine learning algorithm is run in the background to recognize the file structure and automatically generate a model to parse and analyze the data

Working with ISD Models

- View data in a visual structure
- Users can clearly see which elements are connected to real data
- Users can refine data
  - Normalize
  - Exclude
  - Rename



Deploying ISD Models

- After saving the model, a parser is automatically created
- Intelligent Parsers can be used in run-time to transform similar data files on an on-going basis.

Informatica

References

References

- Intelligent structure models
  https://docs.informatica.com/integration-cloud/cloud-data-integration/current-version/components/intelligent-structure-models.html
- Refining intelligent structure models
  https://docs.informatica.com/integration-cloud/cloud-data-integration/current-version/components/refining-intelligent-structure-models.html
- White Paper- Informatica Intelligent Structure Discovery
  https://www.informatica.com/content/dam/informatica-com/en/collateral/data-sheet/intelligent-data-discovery_data-sheet_3280en.pdf

Thank You

Informatica