

Say No to Stale Data!!

Implementation of automated GxP data pipelines using
KNIME

Mar 6th, 2023

Shiva Katepalli
Director, FITS Lab

Jeffrey Xiong
Associate Director, Clinical Data Programming

Sravani Koganti
Senior Manager, Clinical Data Programming



Innovation today, healthier tomorrows



Agenda

Explore **Traditional** Data Flow

Proposal of new data flow at Sunovion

Introduce KNIME

Use cases of KNIME at Sunovion

Future at Sunovion

DATA



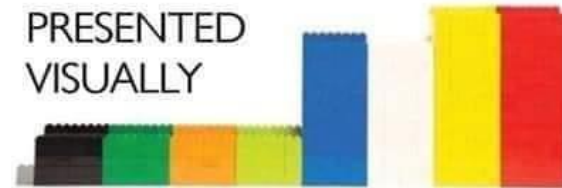
SORTED



ARRANGED



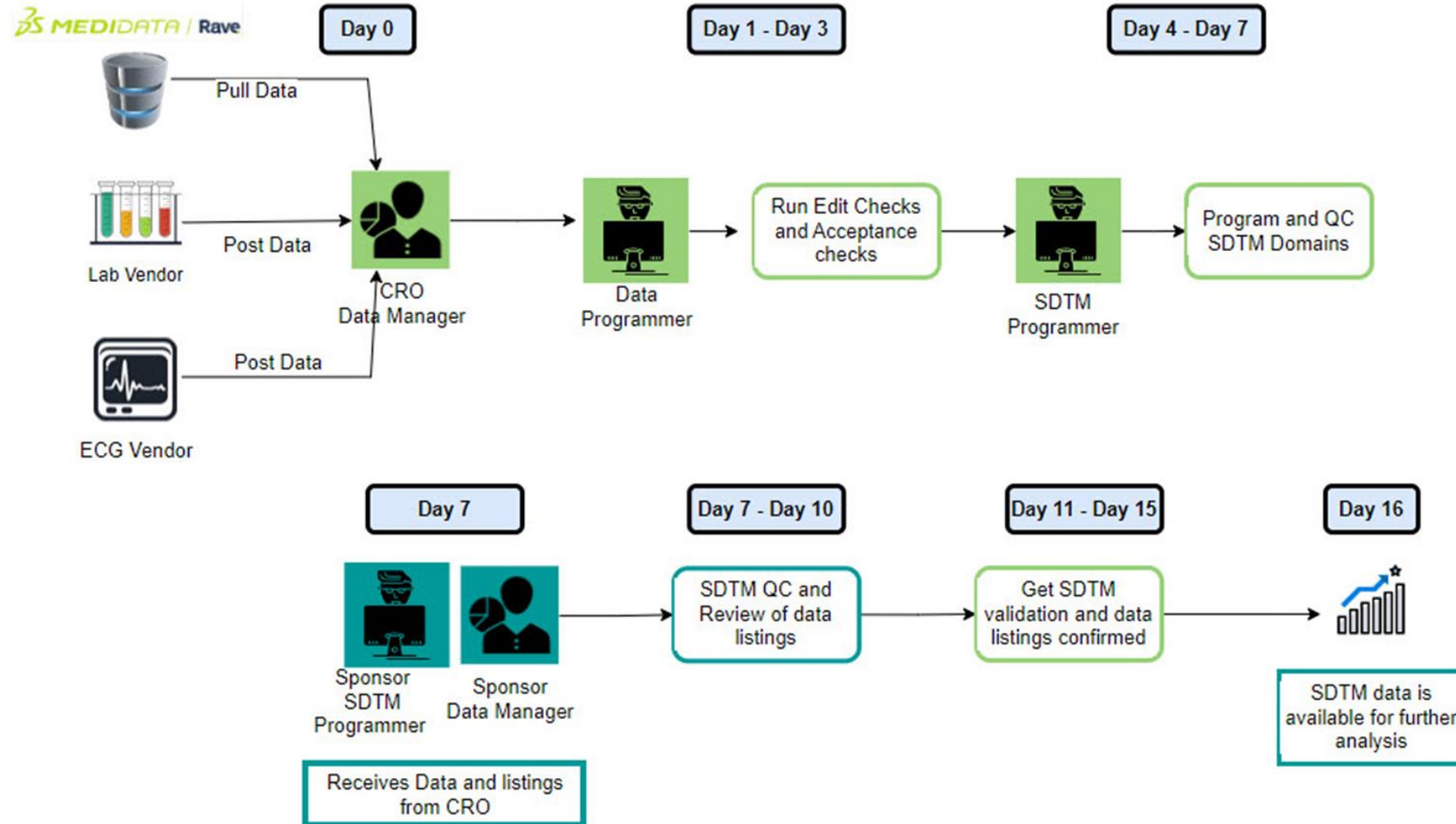
PRESENTED
VISUALLY



EXPLAINED
WITH A STORY



Traditional Data Flow



Challenges in Traditional Data Flow

No Fresh Data



Costs More Money



Very Fragile



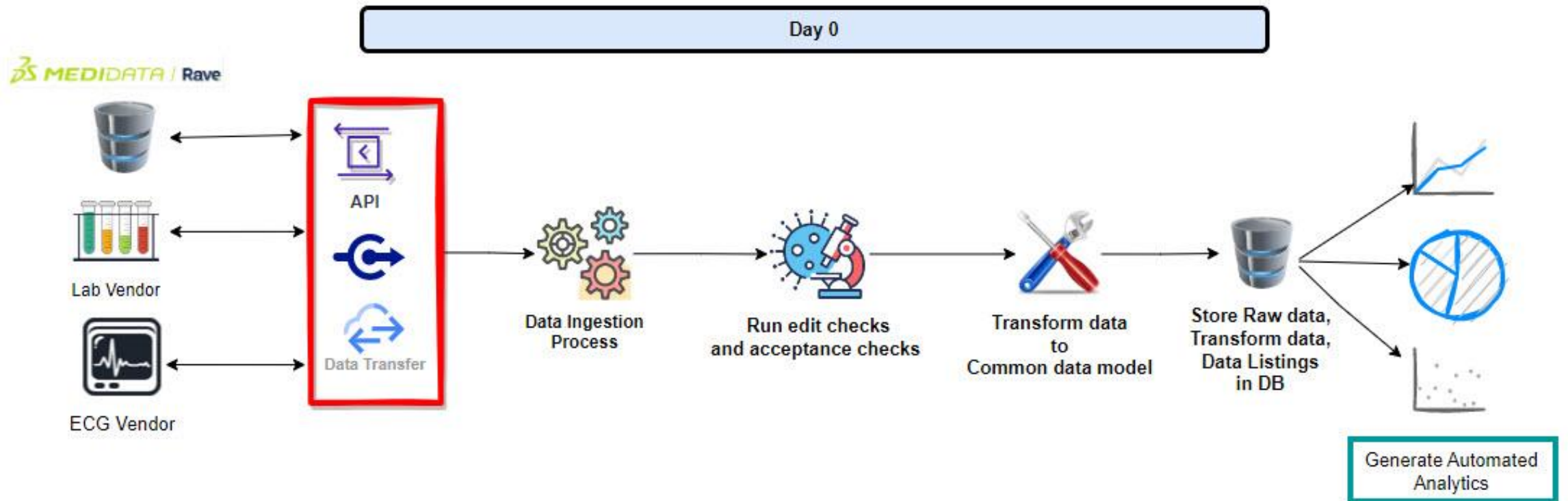
Need more time from people



Not Flexible

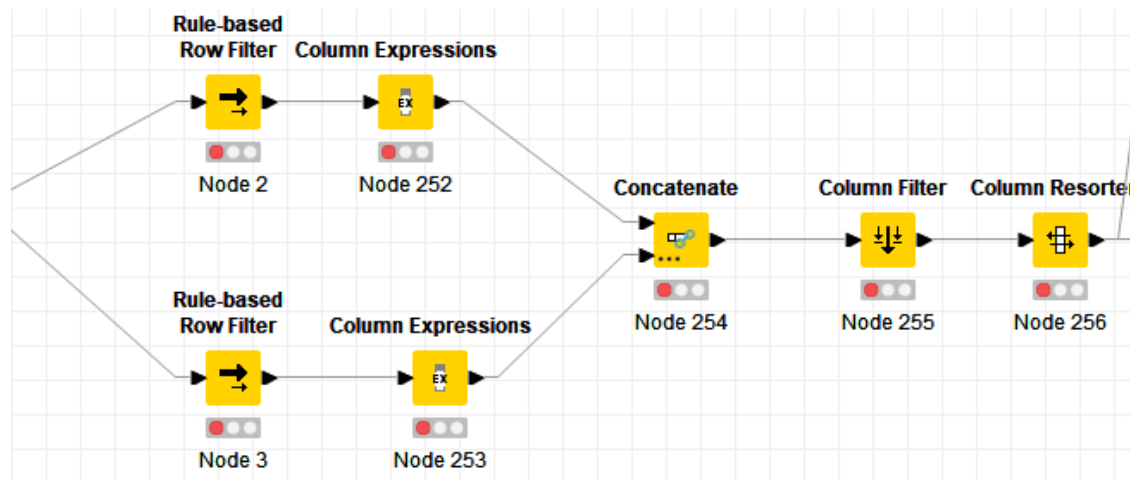


New Data flow at Sunovion for Raw data



KNIME (/naɪm/) - the Konstanz Information Miner, is a free and open-source data analytics, reporting and integration platform and it is based on **visual and low-code programming**.

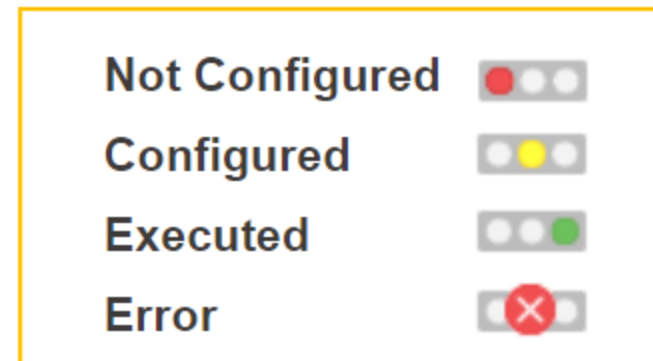
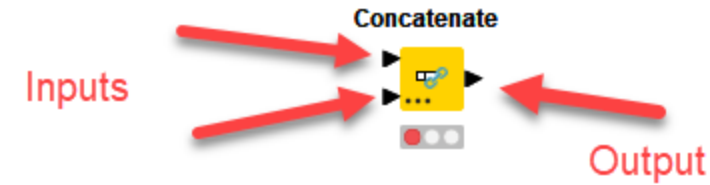
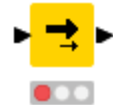
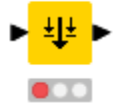
Visual programming is a programming language that lets humans describe processes **using illustration**.



In visual programming, a Graphical User Interface (GUI) guides you through all necessary steps to build a **pipeline (workflow) of dedicated blocks (nodes)**.

Visual Programming also makes the **learning curve much faster** than for code-based tools

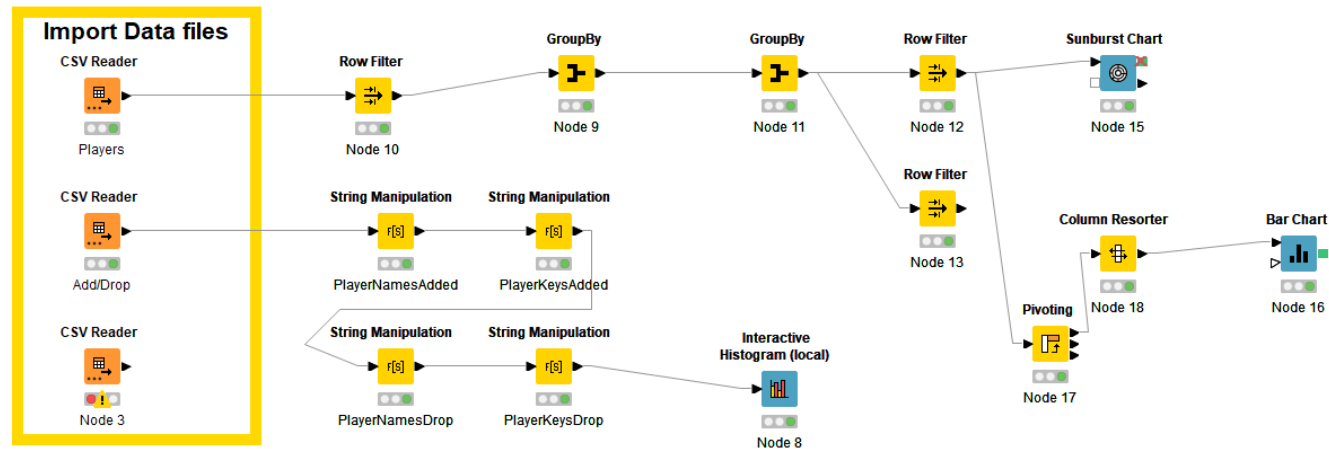
Users can configure the nodes and manage the node operations.



KNIME Terminology

Workflows

- Workflow describes how the data is processed from beginning to the end in a designed journey.
- KNIME nodes are connected to build workflows.

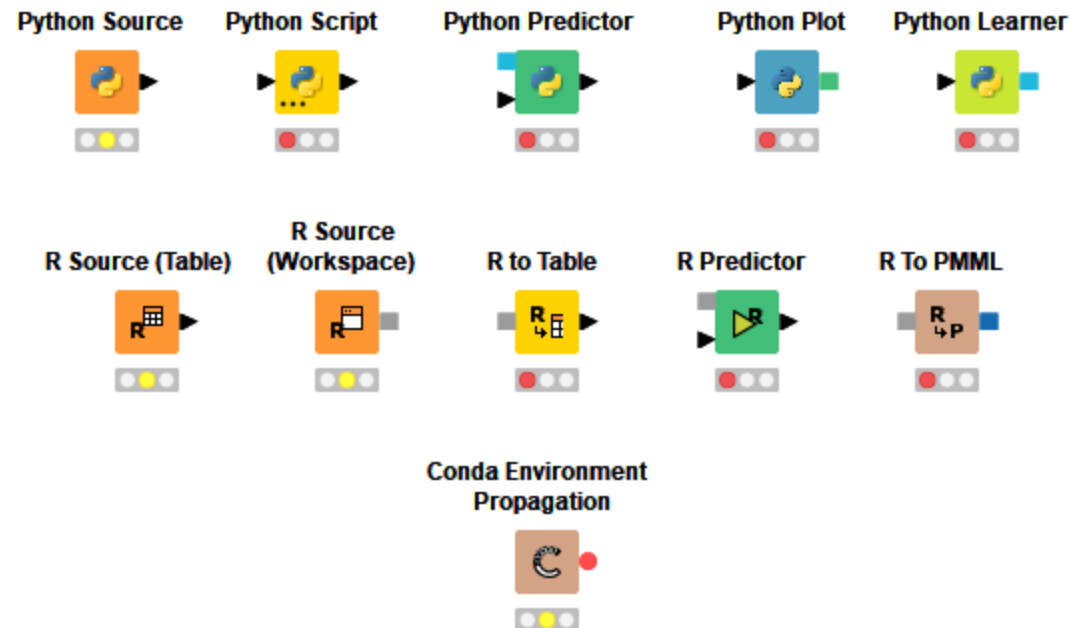


Components

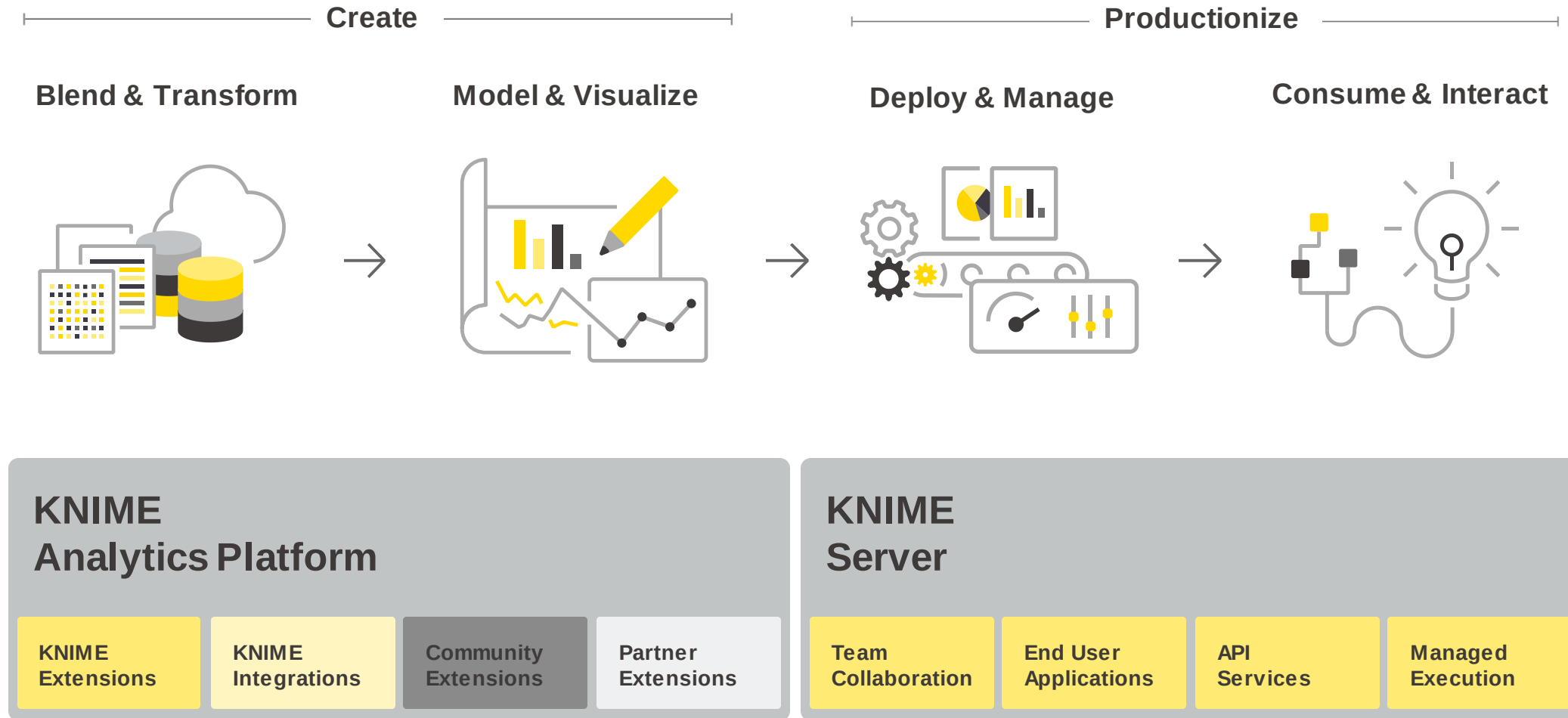
- Components are really KNIME nodes that you create with a KNIME workflow, enabling you to easily bundle, reuse, and share functionality.
- Components can be used along with nodes to build workflows

Python and R Integration

- We have nodes in KNIME that support Python and R code to be executed.
- We can do package management using the conda propagation node.
- Environments created locally can be recreated in any other system or platform.
- We can build new nodes using Python SDK.



KNIME Software – One Ecosystem



GxP – Features on KNIME Server

Authentication

- Users can be authenticated using existing systems and can be added into defined user groups

Versioning

- KNIME WF's can be versioned and restored to any previous versions on KNIME server
- Different versions of KNIME WF's can be compared and generate report

Access Control

- Folders and workflows can be provided with Read, Write, Execute permission to Users and User groups.

Audit

- All events on KNIME server including WF's ownerships, Folder and File access permissions, WF executions, WF updates are stored.
- All the audit data can be accessed and pulled out of server using predefined REST API.

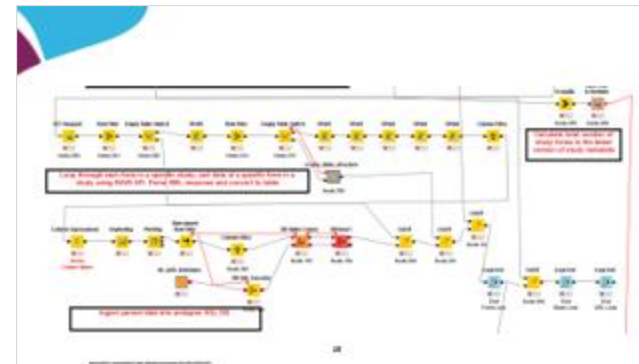
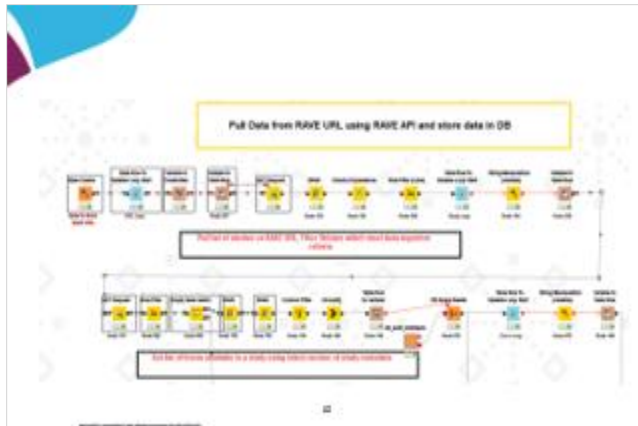


KNIME Extensions

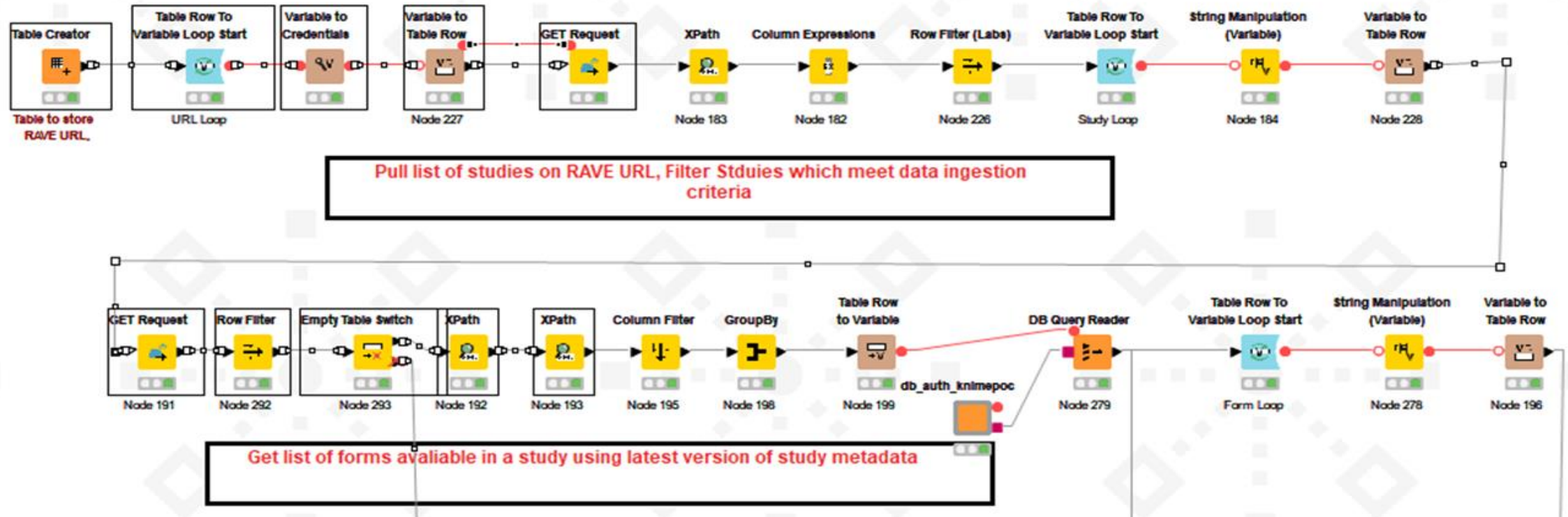
- KNIME extensions provide additional functionality to increase usability and KNIME operations
- KNIME extensions are developed by KNIME Institute, Verified User Community and KNIME Partners
- Some of the popular extensions are
- **KNIME Jupyter** – Allows user to import jupyter notebooks and use them in workflows
 - **KNIME AMAZON** – Provides users with nodes to use AWS services in the workflows
 - **KNIME DB Connectors** – We can use these nodes to connect to various DB in the workflows
 - **KNIME Statistics** – Provides nodes with advanced statistical algorithms
 - **Text processing** - Load documents from a variety of formats including PDF, Docx, Doc, PubMed, DML, and more
 - **Network mining** - Extract information from attributed graphs such as social networks, co-authorship networks, and so on

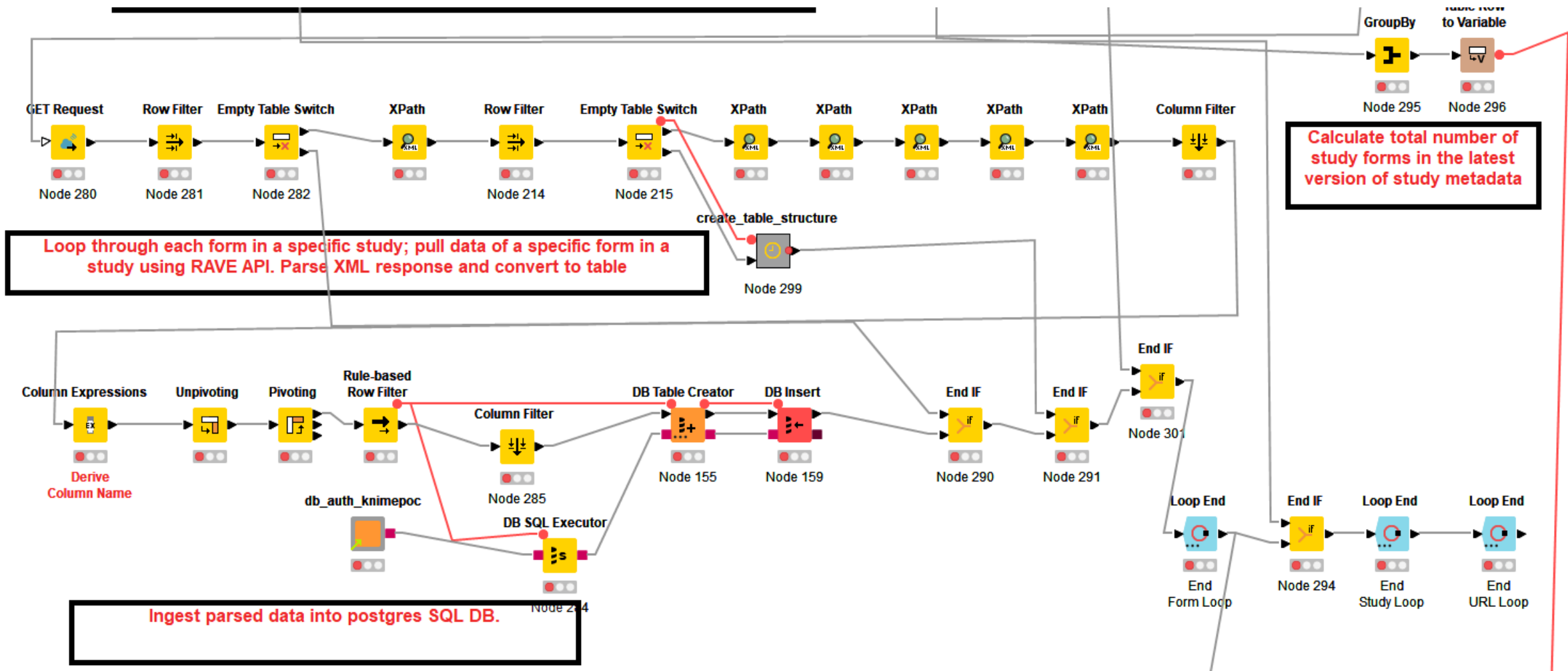
KNIME Use Case 1 – Ingest data from RAVE

- Built and scheduled a KNIME workflow to pull data from all Sunovion RAVE URLs for all the studies published on them on a daily frequency.
- RAVE web services provide data in XML format, workflow will transform data into tabular structure and store in Sunovion data lake



Pull Data from RAVE URL using RAVE API and store data in DB



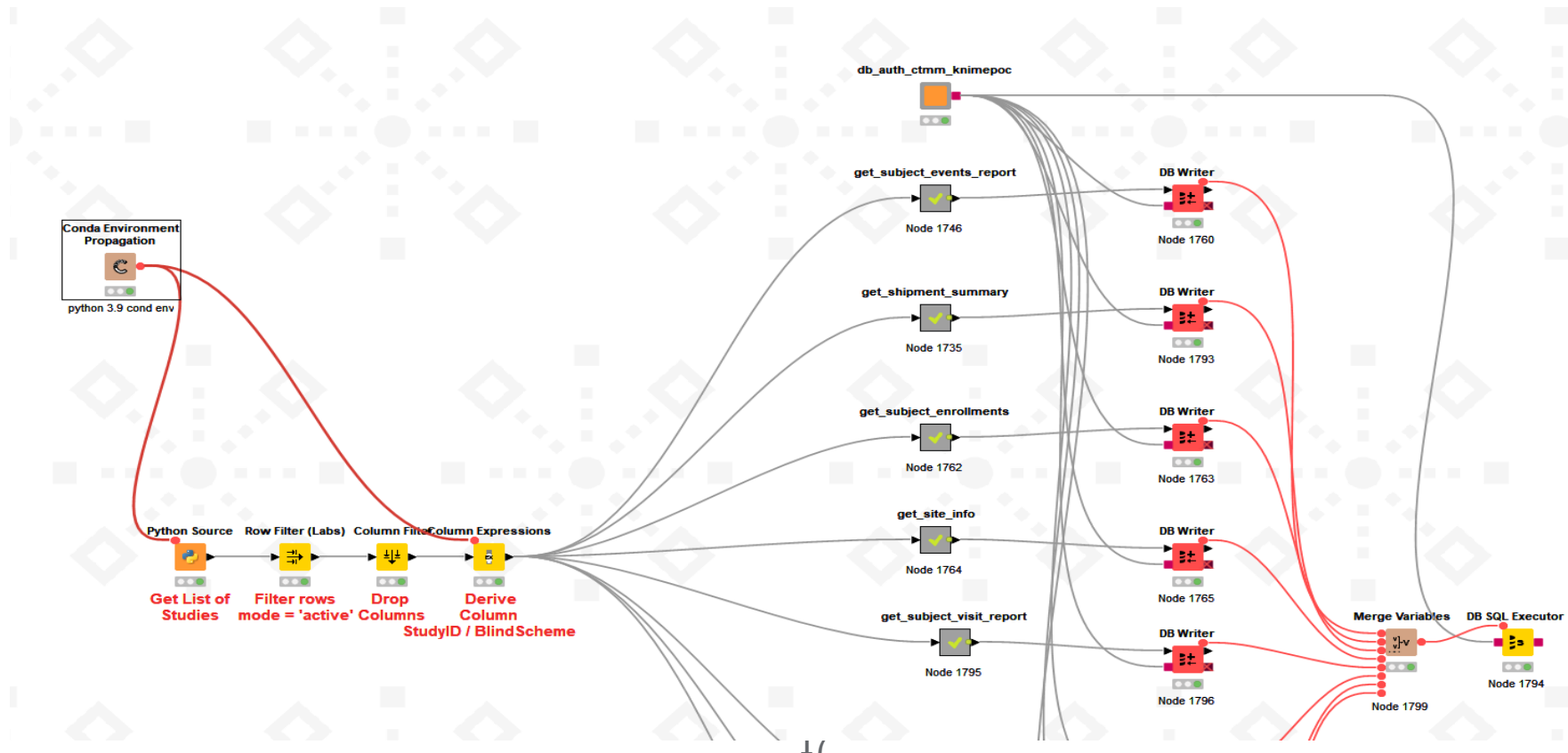


KNIME Use Case 2 – Ingest data Oracle Clinical One

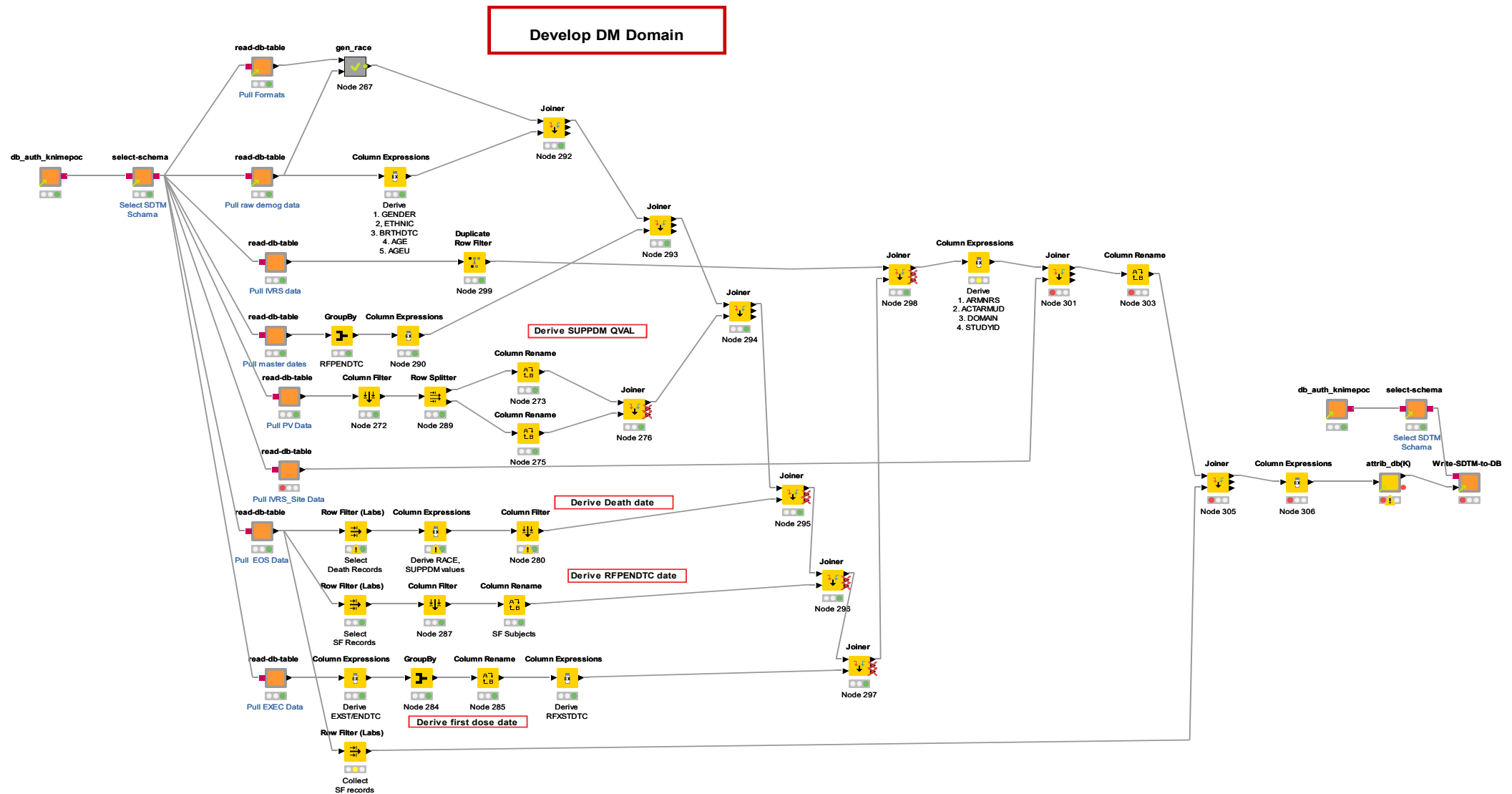
Built KNIME workflow to ingest data from “**Oracle Clinical One**” for all the Sunovion studies.

Data from about 15+ different source tables are being ingested daily.

Business rules and Data Quality checks are designed in the WF to support different business needs.

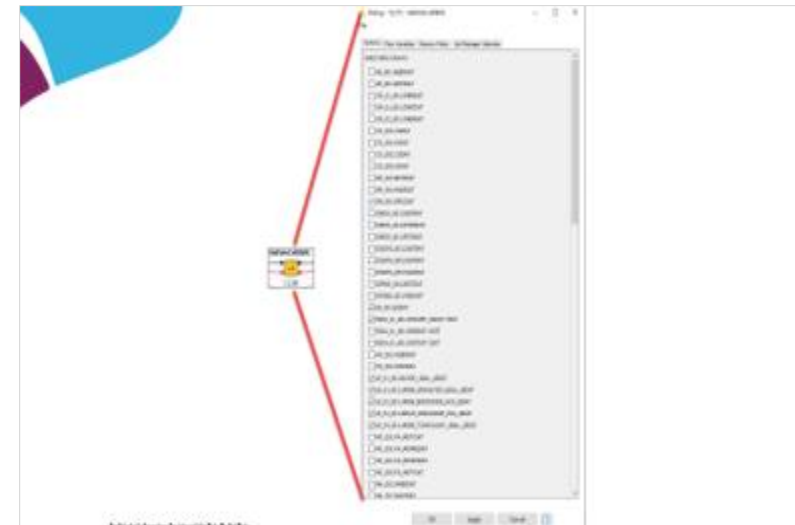
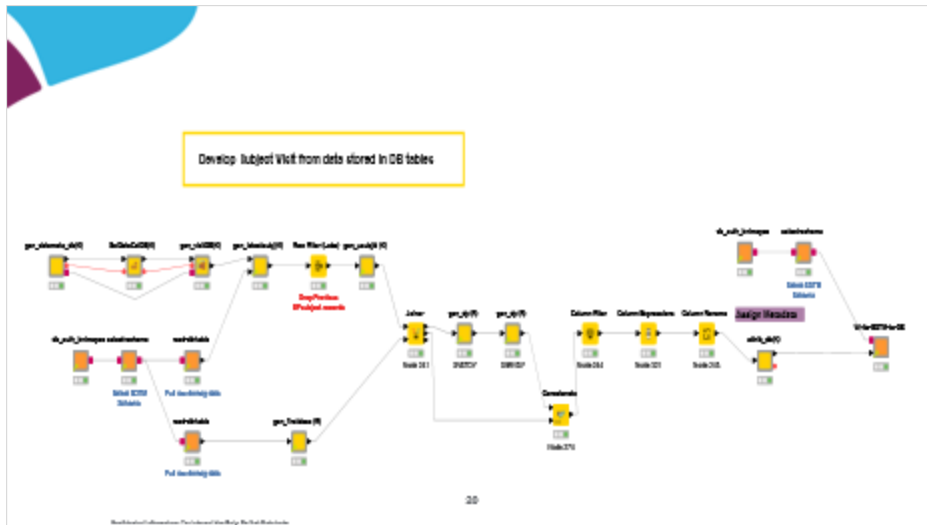


KNIME Use Case 3 – Generate SDTM DM Domain

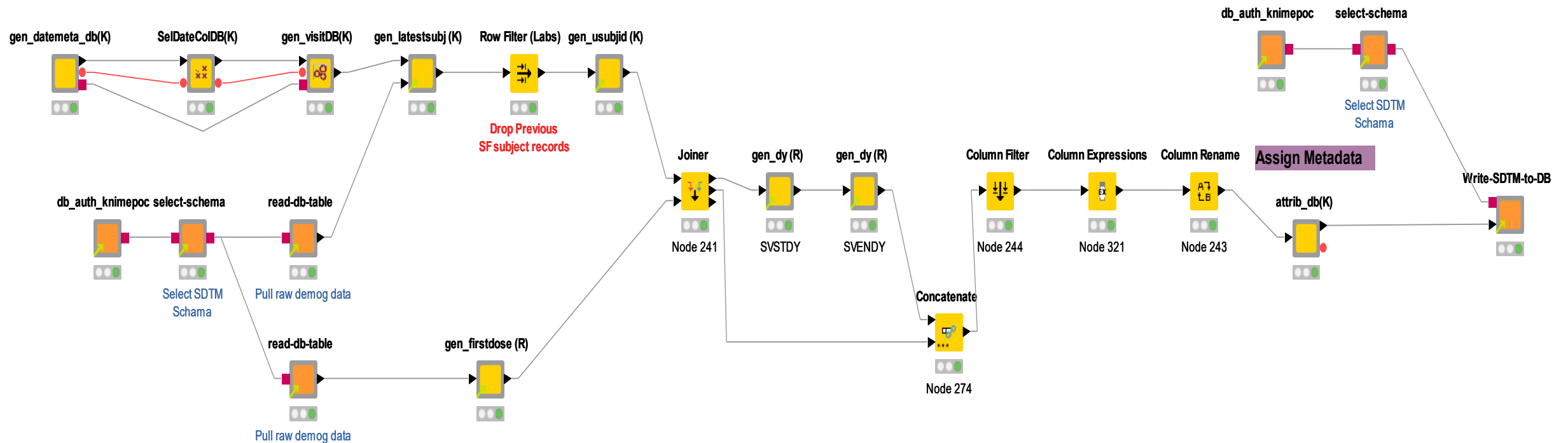


SV Domain

Developed SV domain workflow with user configuration that can be used across different studies



Develop Subject Visit from data stored in DB tables



Dialog - 12:272 - SelDateColDB(K)

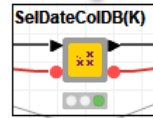
File

Options: Flow Variables Memory Policy Job Manager Selection

Select Date Columns

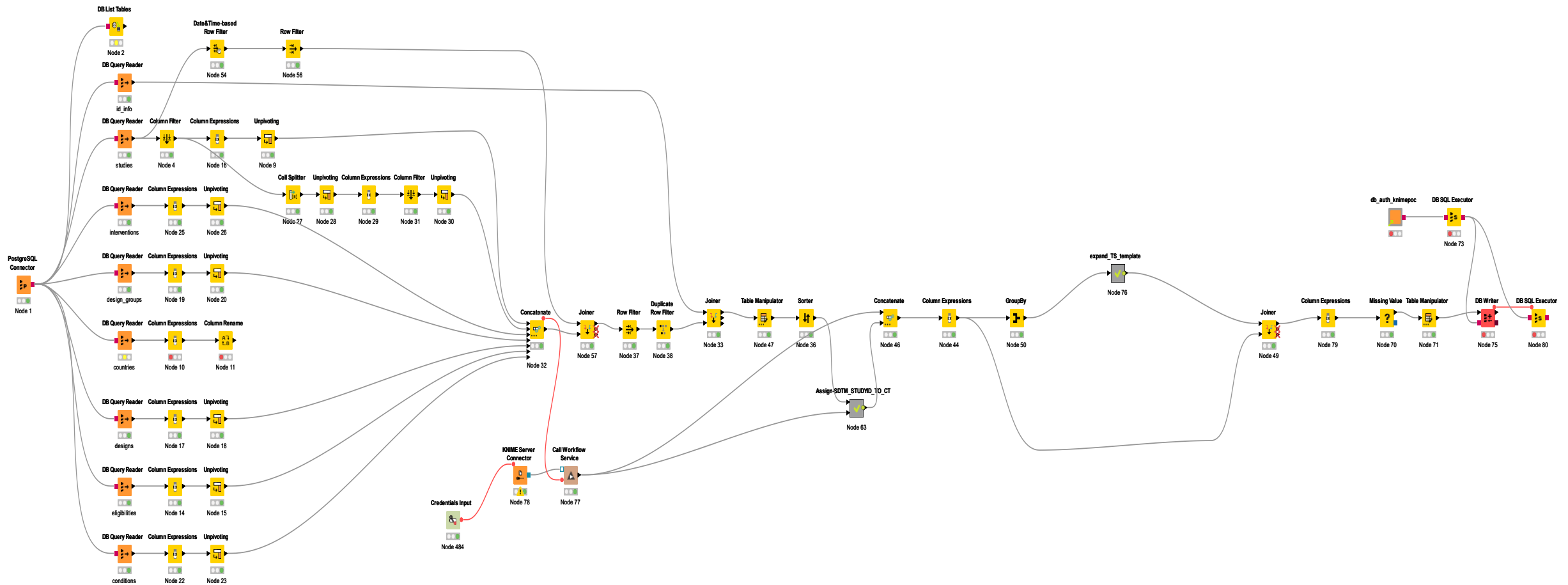
- ☐ AE_001.AEENDAT
- ☐ AE_001.AESTDAT
- ☐ CM_01_001.CMENDAT
- ☐ CM_01_001.CMSTDAT
- ☐ CM_02_001.CMENDAT
- ☐ CO_001.CODAT
- ☐ CO_001.SVDAT
- ☐ CO_002.CODAT
- ☐ CO_002.SVDAT
- ☐ DM_001.BRTHDAT
- ☐ DM_001.INASDAT
- ☒ DM_001.RFICDAT
- ☐ DSEOS_001.DSSTDAT
- ☐ DSEOS_001.RFXENDTC
- ☐ DSEOS_001.SFDTDAT
- ☐ DSICPG_001.DSSTDAT
- ☐ DSICPG_002.DSSTDAT
- ☐ DSICPG_002.PGASDAT
- ☐ DSRND_001.DSSTDAT
- ☐ ECMISS_001.MISSDAT
- ☒ EG_001.EGDAT
- ☒ EXDA_01_001.DISPAMT_DADAT.VISIT
- ☐ EXDA_01_001.EXENDAT.VISIT
- ☐ EXDA_01_001.EXSTDAT.VISIT
- ☐ HO_001.HOENDAT
- ☐ HO_001.HOSTDAT
- ☒ LB_01_001.BLOOD_LBALL_LBDAT
- ☒ LB_01_001.URINE_URINALYSIS_LBALL_LBDAT
- ☒ LB_02_001.URINE_ENDOCRINE_HCG_LBDAT
- ☒ LB_03_001.SERUM_ENDOCRINE_HCG_LBDAT
- ☒ LB_04_001.URINE_TOXICOLOGY_LBALL_LBDAT
- ☐ MC_001.FA_MCFYDAT
- ☐ MC_001.FA_MCMREDAT
- ☐ MC_001.FA_MCMRSDAT
- ☐ MC_002.FA_MCFYDAT
- ☐ MH_001.MHENDAT
- ☐ MH_001.MHSTDAT

OK Apply Cancel ?



Use Case – 5 Develop Automated TS domain

Developed draft TS Domain from data sourced from AACT (ClinicalTrials.GOV) and SDTM datasets



Sample List of Components built at Sunovion



Generate **USUBJID**

loinc2sdm



Populate **LBTEST** and **LBTESTCD** from **LOINC** leveraging **CDISC Library API**

gen_latestsubj (K)



Generate **Latest Subject Number** when multiple screenings are done in the study

gen_dy (R)



Generate **Study Day**

gen_firstdose (R)



Generate First Dose of the Subject
RFSTDTC and **RFXSTDTC**

gen_epoch(K)



Generate **EPOCH** based on dates and SE domain configuration

attrib_db(K)



Assign **SDTM metadata** to the final data table

Components may have **configurations** to be managed by user . Configurations will add **flexibility to component**

A screenshot of a configuration dialog box titled "Dialog - 5:240 - gen_dy (R) (SVSTDY)". The dialog has a "File" menu and four tabs: "Options", "Flow Variables", "Memory Policy", and "Job Manager Selection". The "Options" tab is selected. It contains three sections: "name of --DY" with a text field containing "SVSTDY", "date" with a dropdown menu showing "svstdtc", and "first dose date" with a dropdown menu showing "EXSTDAT". At the bottom are buttons for "OK", "Apply", "Cancel", and a help icon (a question mark in a circle).

Sample List of Components built at Sunovion

ucum based unit
conversion



Convert results in one unit to
another unit

search controlled terminology
in CDISC Library



Search controlled terminology

Search LOINC in
loinc.org and cdisc.org



Get LOINC metadata from loinc.org
and CDISC library

Dialog - 3:339 - ucum based unit conversion

File

Options | Flow Variables | Memory Policy

Source unit in UCUM (required)
pg/mL

Target unit in UCUM (required)
pmol/L

LOINC (optional)
83127-1

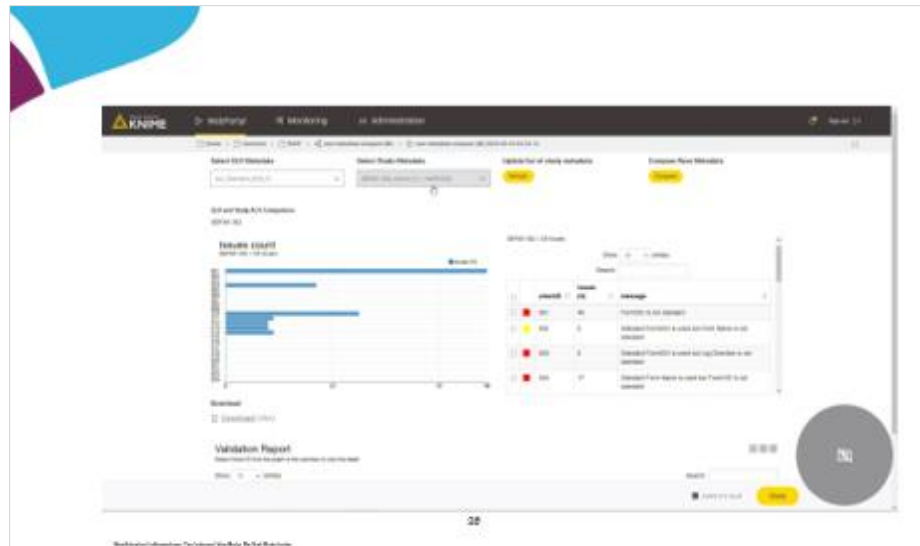
Molar weight (optional)

result (required)
3.99

OK Apply Cancel ?

Data Apps

- Data Apps provide a user interface to perform scalable and shareable data operations, such as visualization, data import, export, and preview
- Provide actionable insights to end users whenever they need them.
- Through KNIME Server you can also share your Data App with 5, 10, or 1000 end-users
- Data App to compare metadata of a study with GLV and generate discrepancy report





WebPortal
Monitoring
Administration

Sign out

Home
>
Sunovion
>
RAVE
>
rave metadata compare (db)
>
rave metadata compare (db) 2023-02-23 23.23.16

Select GLV Metadata

GLV_Standard_2022_01

Select Study Metadata

SEP361-302_Version_3.1_19APR2022

Update list of study metadata

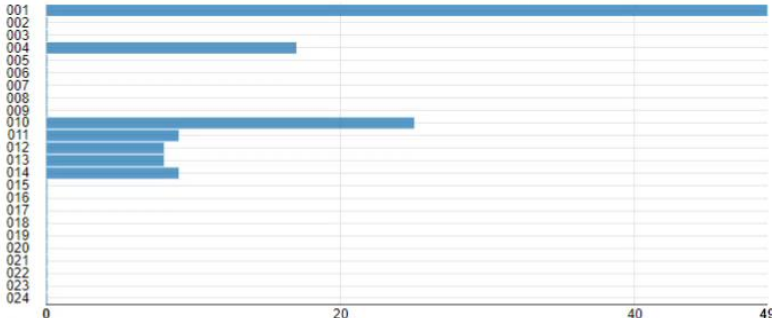
Refresh

Compare Rave Metadata

Compare

GLV and Study ALS Comparison
SEP361-302

Issues count
SEP361-302: (125 Issues)



SEP361-302: (125 Issues)

Show 25 entries
Search:

	checkID	Issues (N)	message
☐	001	49	FormOID is not standard
☐	002	0	Standard FormOID is used but Form Name is not standard
☐	003	0	Standard FormOID is used but Log Direction is not standard
☐	004	17	Standard Form Name is used but FormOID is not standard

Download

Download (xlsx)

Validation Report

Select Check ID from the graph or the summary to view the detail

Show 10 entries
Search:

☒ Delete this result

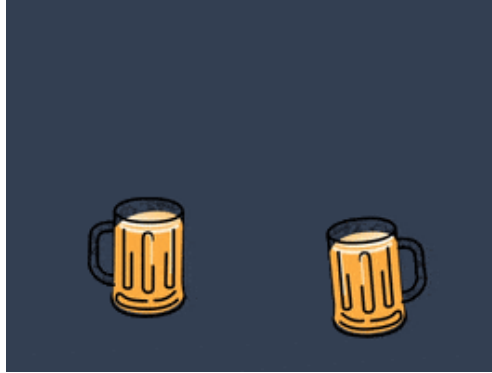
Close



Future of KNIME at Sunovion

- Complete required validation process to qualify system for GXP deliverables.
- Develop more components to perform more comprehensive data quality checks
- Develop workflows to ingest data from more external vendors to complete data ingestion for all the raw data.
- Perform POC by developing workflows to derive ADaM datasets.
- Develop automated ETL workflows from real world data sources to support exploratory analysis.
- Develop clinical data analysis specific nodes and contribute to KNIME community

Acknowledgements



- **Clinical Data Programming Sunovion** - Cindy Ma, Jyotsna Bhate, Doris Menke for their support in building workflows and components
- **Sunovion IT team** – For their extensive help in setting up infrastructure to support business needs
- **KNIME Team** - Norbert Sas, Martyna Pawletta for their continued guidance and support



References

KNIME Community Hub – Find workflows, nodes, components from user community

<https://hub.knime.com/>

KNIME Blog

<https://www.knime.com/blog?page=3>

Nodepit – Over 5K nodes, 19K workflows from user community and partners

<https://nodepit.com/>

<https://medium.com/low-code-for-advanced-data-science/why-knime-98c835afc186>

<https://medium.com/low-code-for-advanced-data-science>

