Webinar: How to analyze data on Databricks without coding

April 29, 2025

Speakers

Frank Heitkämper
Bosch

Tobias Kötter
Product Manager, KNIME
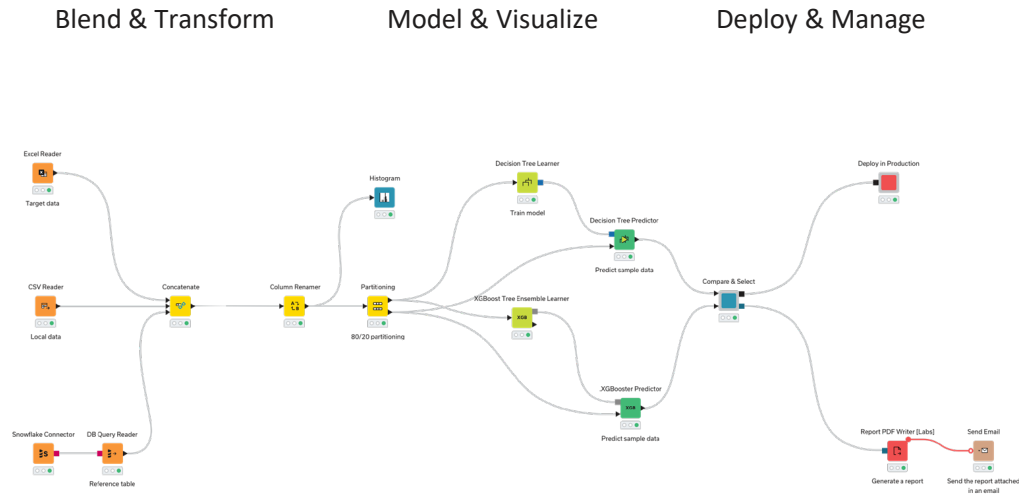
Swantje Schulze
VP Revenue EMEA, KNIME

Open for Innovation
KNIME

Drive innovation and save cost

Blend & Transform   Model & Visualize   Deploy & Manage

300+ Data Sources

Files, Docs
Audio, Images
Databases
Apps
Cloud
SAP
APIs
Sensor data

REST APIs
Edge
BI Tools
Docs & Emails

6

Databricks: A powerful platform for data experts

- Databricks Data Intelligence Platform is for:
  - Data engineers
  - Data scientists
  - Those comfortable with code

- KNIME supports the roles above, **plus**:
  - Business users
  - Domain experts,
  - Data analysts,
  - Decision-makers who need to work with data and get insights without scripting.
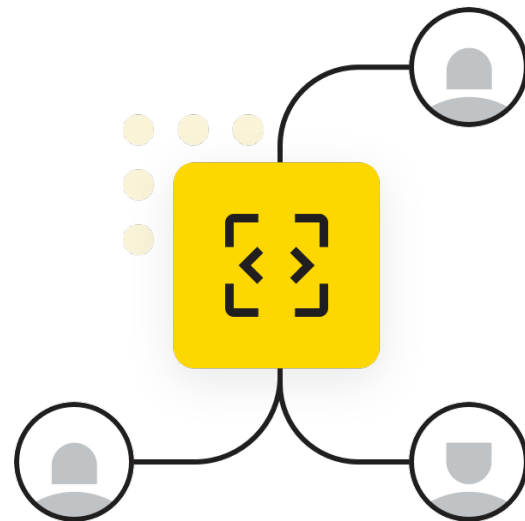
Why use KNIME with Databricks

- **Democratize data analysis**

  Visual workflows let everyone work with data on Databricks.

- **Capture frontline insights**

  Business users can explore data and derive insights directly — you can't afford to miss what they know.

- **Free up data teams**

  Keeps data teams focused on high-impact work, not just servicing requests.

- **Support governance and compliance**

  Audit trails help with debugging, compliance (including EU AI rules), and building trustworthy AI.

- **Orchestrate across systems**

  KNIME connects easily to numerous data sources and tools outside of Databricks — like Excel, accounting platforms, and IoT devices.

Open for Innovation

KNIME

Why users choose KNIME to work with data on Databricks

"Integrating KNIME UI into Databricks attracts users who rely heavily on corporate data in Databricks but are not technically inclined"

"It is really easy to use the compute power that is provided by Databricks in KNIME workflows"

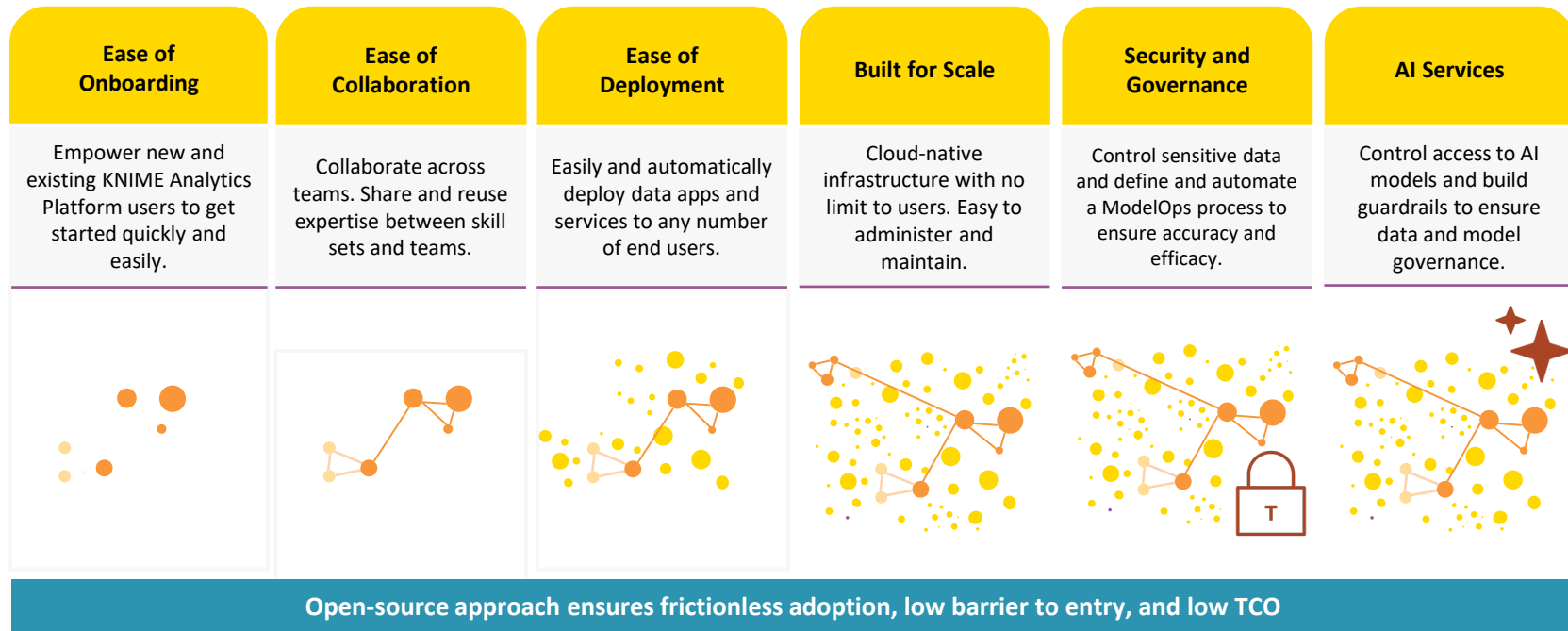"KNIME lets us join data from different sources that are in Databricks AND on-prem!"

Open for Innovation
KNIME

KNIME and Databricks
Better Together

KNIME visual workflows with Databricks

Databricks	Databricks + KNIME	KNIME
Big data storage, lakehouse architecture	Visual workflows for analysis	Visual workflows
Distributed compute	ML/AI app development	Data apps with browser-based front-end
Collaborative notebooks	Open source foundations	Data blending/data neutral/platform neutral
Delta lake	Agentic workflows powered by Databricks models	Connect to any tool/technology
Cloud native, SaaS	Markets: Horizontal	Customer-managed, placed closer to databases
Users: Data engineers, data scientists (Python,SQL)	Users: All data workers	Users: Citizen data scientists, business users, data engineers

Open for Innovation
KNIME

Scale the impact of data on Databricks with KNIME Business Hub

Ease of Onboarding	Ease of Collaboration	Ease of Deployment	Built for Scale	Security and Governance	AI Services
Empower new and existing KNIME Analytics Platform users to get started quickly and easily.	Collaborate across teams. Share and reuse expertise between skill sets and teams.	Easily and automatically deploy data apps and services to any number of end users.	Cloud-native infrastructure with no limit to users. Easy to administer and maintain.	Control sensitive data and define and automate a ModelOps process to ensure accuracy and efficacy.	Control access to AI models and build guardrails to ensure data and model governance.

Open-source approach ensures frictionless adoption, low barrier to entry, and low TCO

453% ROI — Forrester Total Economic Impact™ (TEI) study

Open for Innovation
KNIME

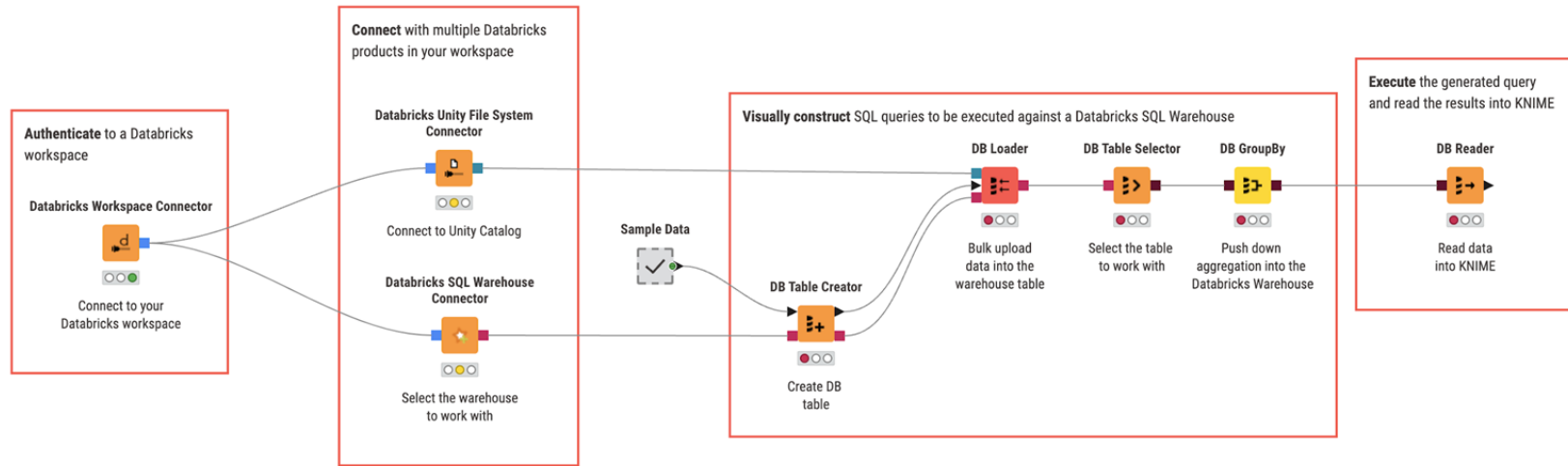# Use KNIME to analyze data on Databricks without coding
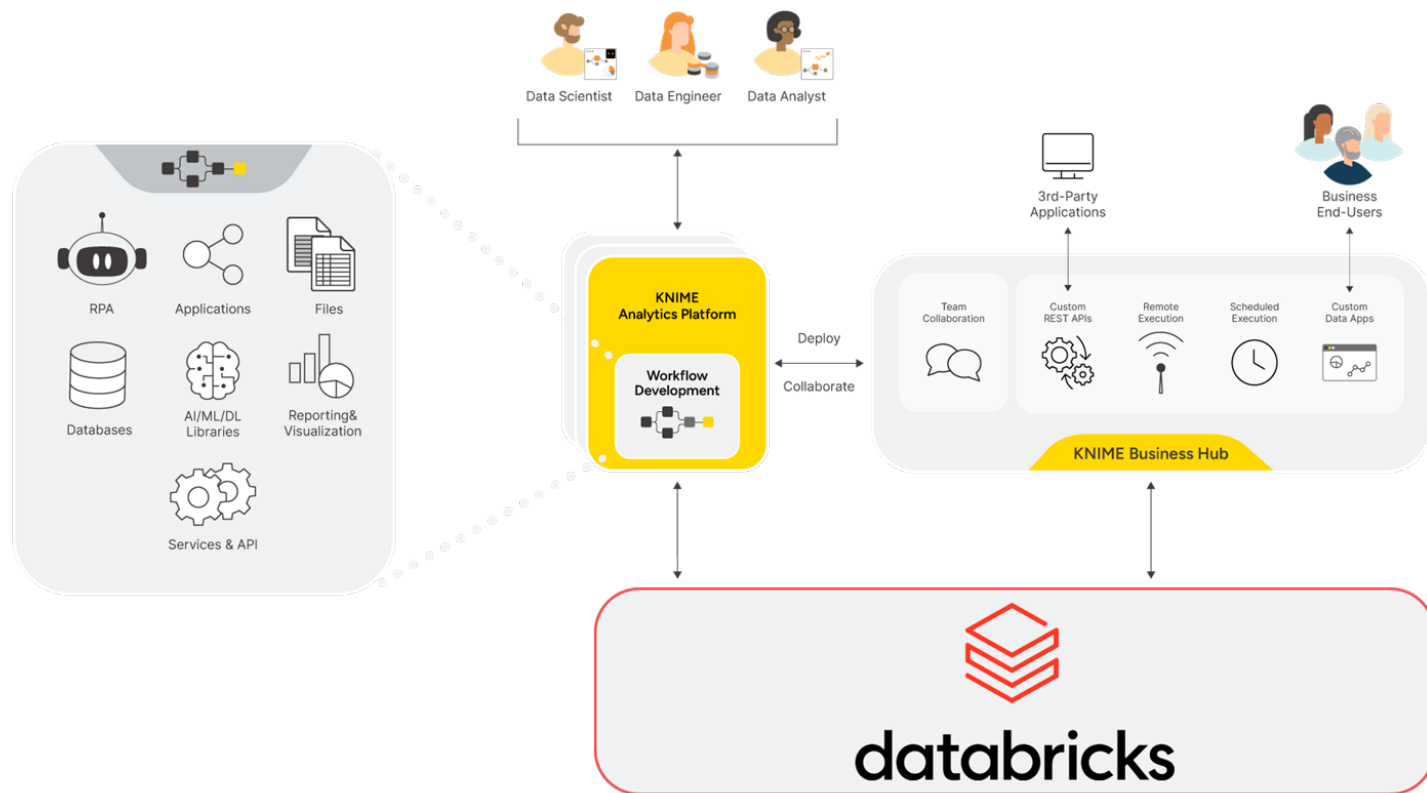
Working with different Databricks products

Working with Databricks Unity File System
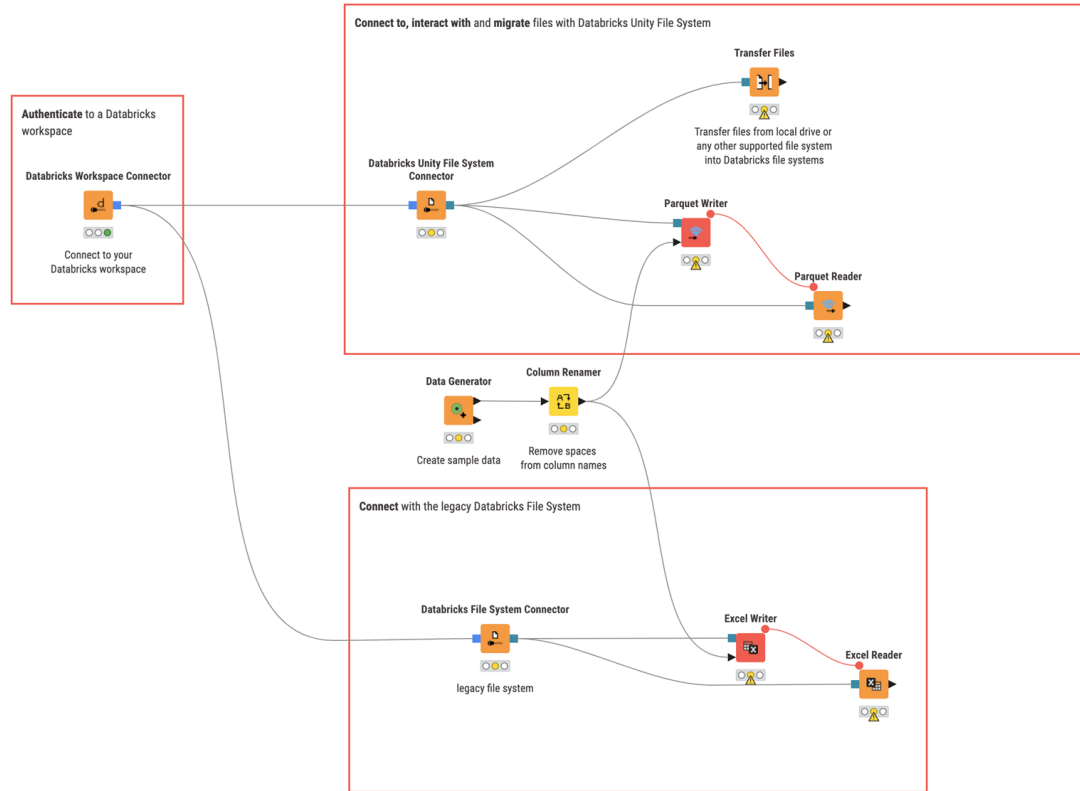
- Securely interact with files stored in Databricks Unity File System

- Uses the KNIME file-handling framework

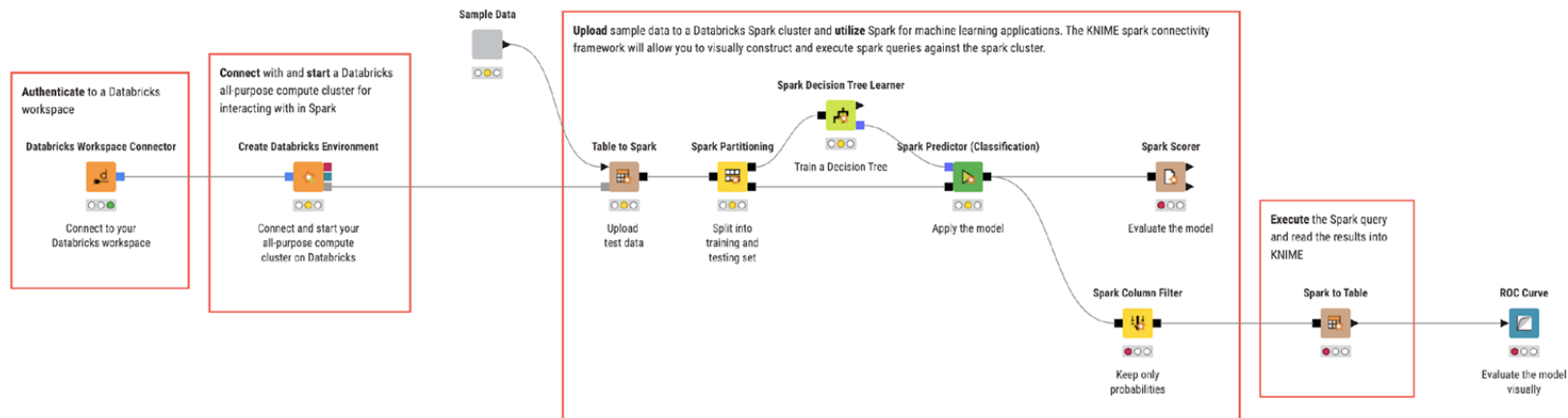- Migrate files to Databricks from other file systems

6

Working with Databricks SQL Warehouse

- Visually assemble SQL queries to be sent to and executed on Databricks SQL Warehouse

- Make use of the KNIME Database connectivity framework
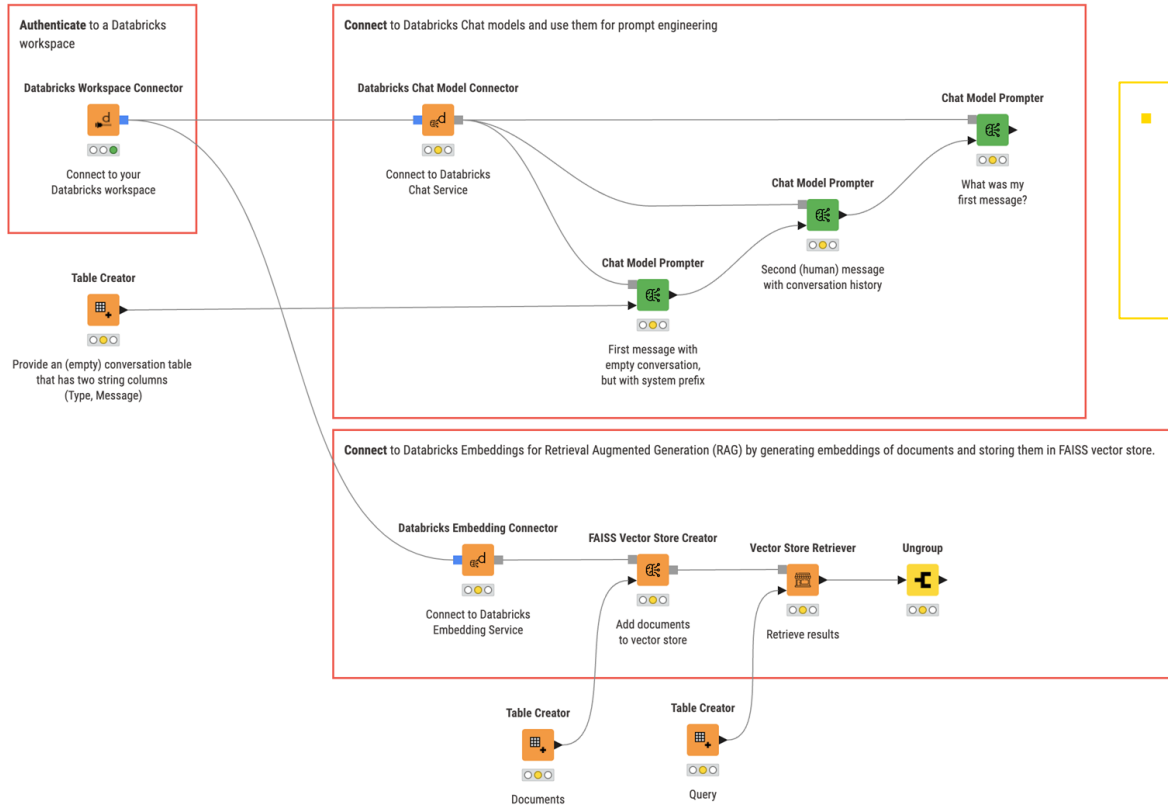
7

Working with Databricks Machine Learning

- Visually assemble Spark ETL and Machine learning jobs
- Send and execute at scale on Databricks all-purpose Compute Clusters

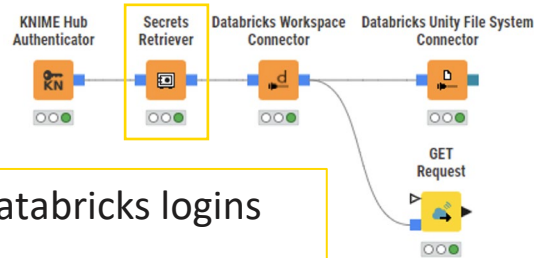- Make use of the KNIME Spark integration for large-scale data processing

Working with Databricks AI

- Design LLM-based solution using model endpoints served on Databricks for prompt engineering

- Enhance the accuracy and reliability of GenAI models by utilizing model embedding endpoints on Databricks for RAG

Databricks Secret Type

- Allows customers to manage their Databricks logins centrally and securely
- Databricks secret type with support for all available authentication methods
  - OAuth for service principals (OAuth M2M)
  - OAuth for users (OAuth U2M)
    (only available for personal but not team secrets)
  - Personal access token (PAT)
- Available for Business Hub and Community Hub customers

10

How to Analyze Data on DataBricks without Coding

Frank Heitkämper, FeP/TEF4

29.04.2025

BOSCH

How to Analyze Data on DataBricks without Coding
First Things First

Dependency on Cloud Environment:
- ▶ The use of Databricks is highly dependent on how your company account is set up in the underlying cloud environment.
- ▶ Similarly, the setup of your company and personal accounts in Databricks affects usability.

Bosch Environment:
- ▶ All illustrations in this presentation are based on the Bosch environment in Databricks.
- ▶ These experiences may work in your environment, but not necessarily.

KNIME Developers:
- ▶ KNIME developers cannot test their nodes and features in advance against your individual setup.

Possible Causes for Limitations:
- ▶ Security topics
- ▶ Data confidentiality
- ▶ Costs
- ▶ Network architecture
- ▶ Legal aspects
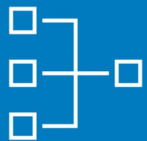- ▶ Initial setups not evaluated deeply for your use case

Be careful when You create a workaround!

BOSCH

How to Analyze Data on DataBricks without Coding
First Things first

Cloud Provider:

Aside of DataBricks the Cloud-Provider beneath can charge quite high Prices for so called „egress Traffic".

DataBricks:

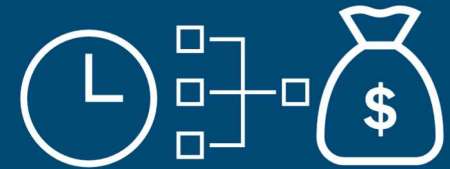Cluster-Configuration should be optimized between initial Costs and Run-time Costs.

3rd Party

Be careful when using 3rd Party Software that queries in high Frequency.

KNIME:

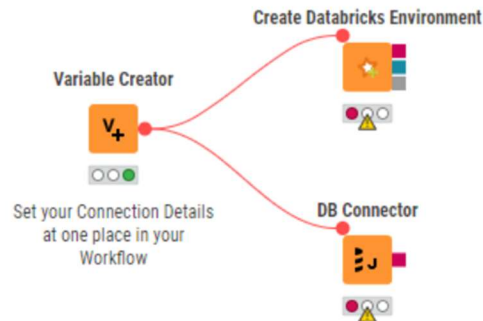By using KNIME you can create a huge Lever on all of that.

BOSCH

How to Analyze Data on DataBricks without Coding
How to connect to DataBricks

Depending on Network- / IT-Environment a lot of different Settings

(JDBC-Settings, Proxy-Settings, Account-Settings, Target-Server, etc.)

- We soon came to the conclusion that it is easier to handle all the Settings in a centralized way
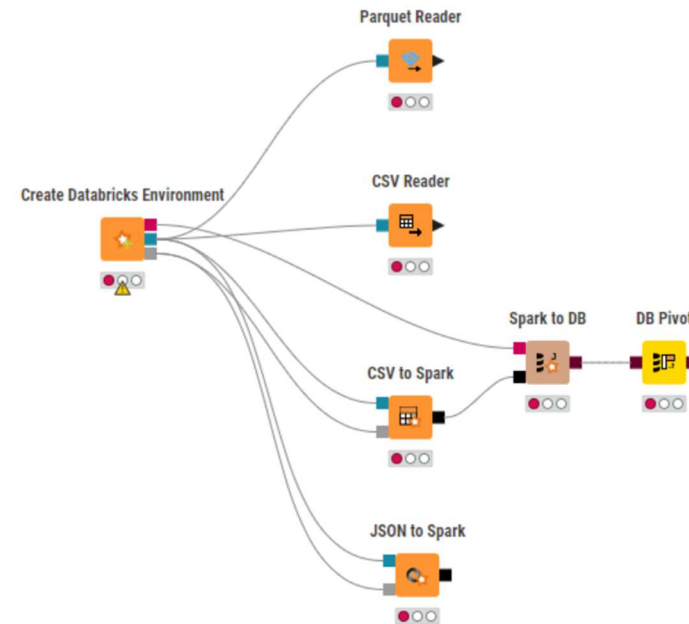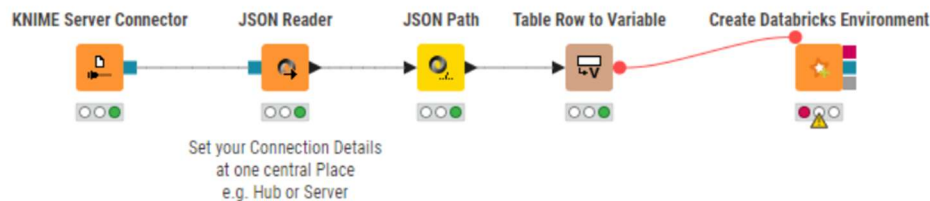- There are various ways to realize this

Variable Creator

Create Databricks Environment

DB Connector

Set your Connection Details at one place in your Workflow

KNIME Server Connector   JSON Reader   JSON Path   Table Row to Variable

Set your Connection Details at one central Place e.g. Hub or Server

Even when you do not use a KNIME-Server/KNIME-Hub. We strongly recomend to use such config-Files

It is up to you if you use, JSON, XML, Text, .csv

BOSCH

How to Analyze Data on DataBricks without Coding
„Create DataBricks Environment" Node

Provides a really convinient way to use
multiple Ports/Features
even simultaneously

Upper Port -> DB Connection (SQL-Warehouse)
Middle Port -> DBFS Connection
Lower Port -> Spark Context

KNIME Server Connector · JSON Reader · JSON Path · Table Row to Variable · Create Databricks Environment

Set your Connection Details
at one central Place
e.g. Hub or Server

Parquet Reader · CSV Reader · Create Databricks Environment · Spark to DB · DB Pivot · CSV to Spark · JSON to Spark

In our Environment we can´t use „Spark Context". Because we
can not spwan a Cluster setup with the needed rights Policy.
Remember Slide 2

BOSCH

How to Analyze Data on DataBricks without Coding
„Create DataBricks Environment" Node
DB-Connection <-> SQL-Warehouse
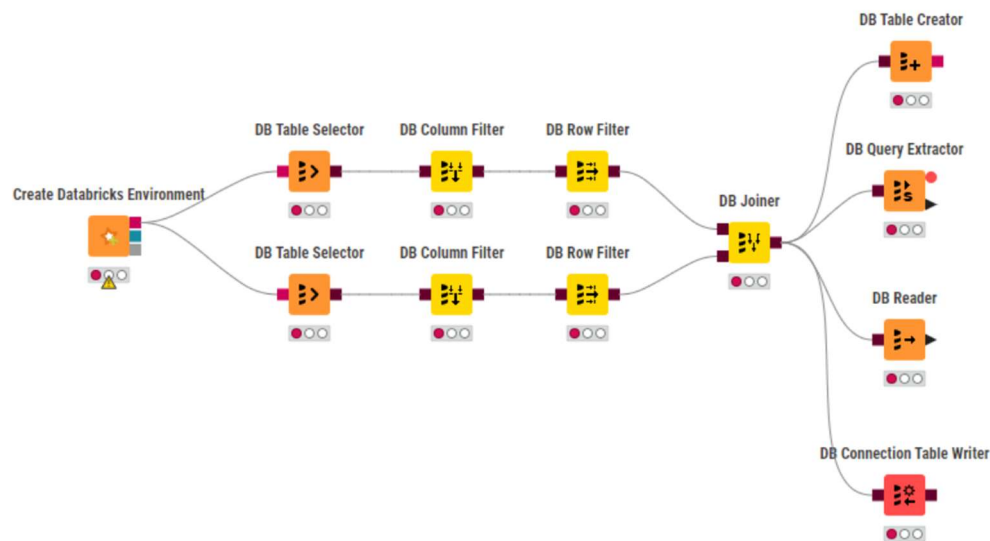
In general really well working
- „DB Column Filter"
- „DB Row Filter"
- „DB Joiner"

Depending on your Setup
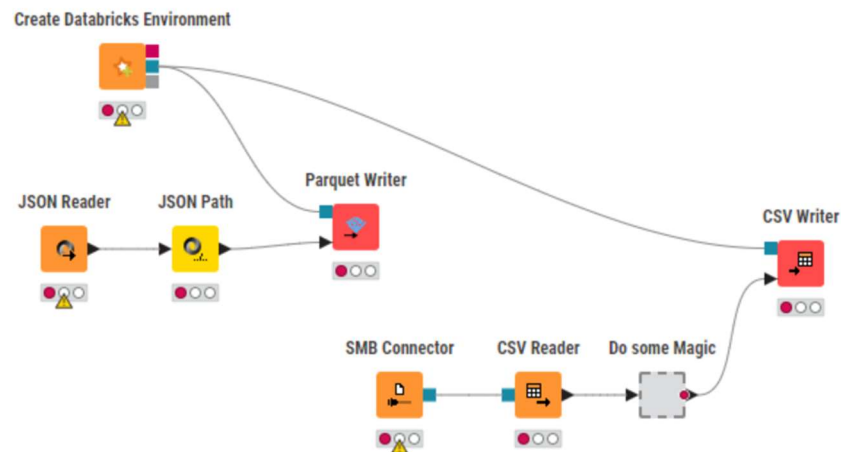- „Table Writer"
- „DB Update"
- „DB Load"
- „DB Merge"

BOSCH

How to Analyze Data on DataBricks without Coding
„Create DataBricks Environment" Node
DBFS-DataBricks File System

In User-Context, and according to your Setup maybe as a general Feature, you can access the DBFS

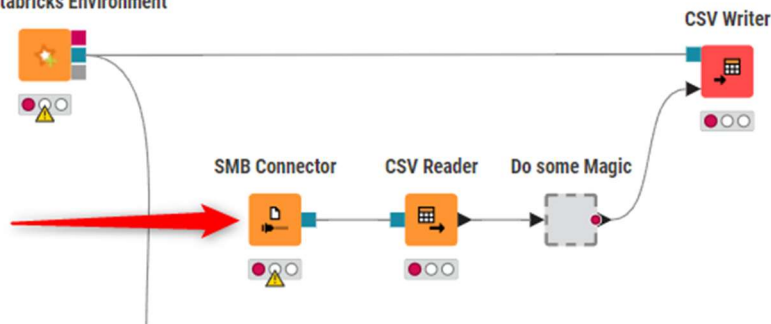There you can upload/download Files of many supported Formats, e.g. .csv; .parquet; etc.

BOSCH

How to Analyze Data on DataBricks without Coding
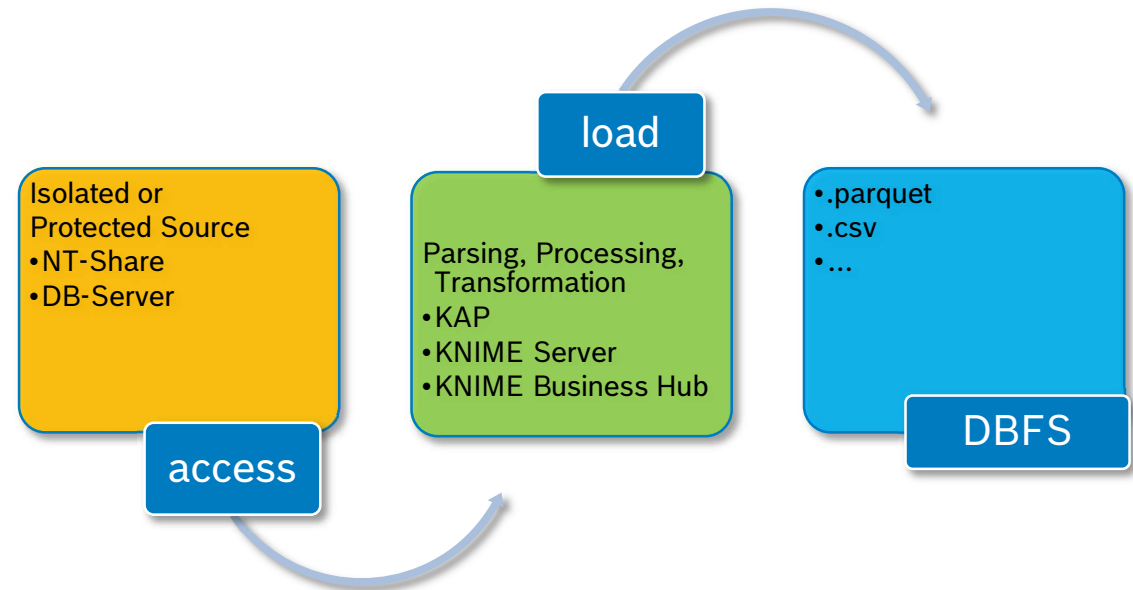## KNIME as local Instance to connect to Data from a trusted Device

It might can happen that you need to get Data to DataBricks that reside in protected Sources which are not reachable from DataBricks.

Because the Cloud Environment is in the Internet. And your Source is behind e.g. a Firewall to be protected.

Create Databricks Environment

SMB Connector CSV Reader Do some Magic CSV Writer

KNIME-Server, KNIME Business Hub, KAP Can run on trusted Machines and/or in trusted Network Segments.

Isolated or Protected Source
- NT-Share
- DB-Server

access

Parsing, Processing, Transformation
- KAP
- KNIME Server
- KNIME Business Hub

load

- .parquet
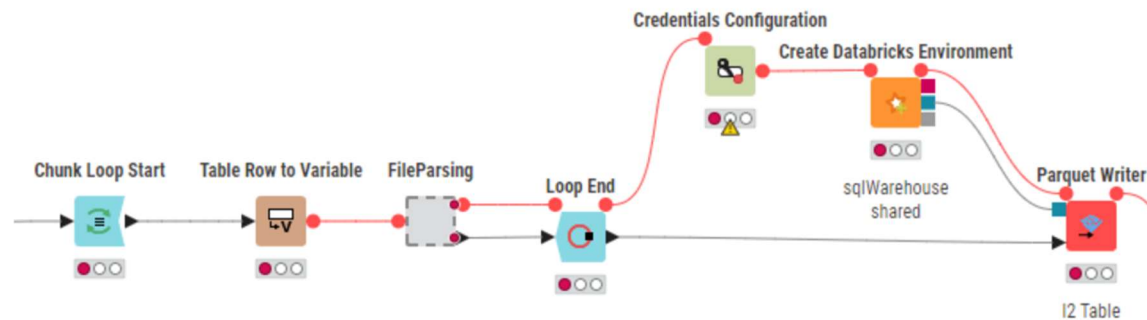- .csv
- ...

DBFS

BOSCH

How to Analyze Data on DataBricks without Coding
„Create DataBricks Environment" Node DBFS, Pipeline Tricks

This Screenshot is from an actual Project for an initial upload of Data. That was later on tranfered into a Pipeline

You can re-use large Parts of your Workflow to build the Pipeline.
By that you can ensure the consistency of Table-Specs, Data-Formats, etc

BOSCH

How to Analyze Data on DataBricks without Coding
Front-end for End-Users

Easy to use Interface for End-Users

Variable Date Range

Recursive Querry over large Delta-Table

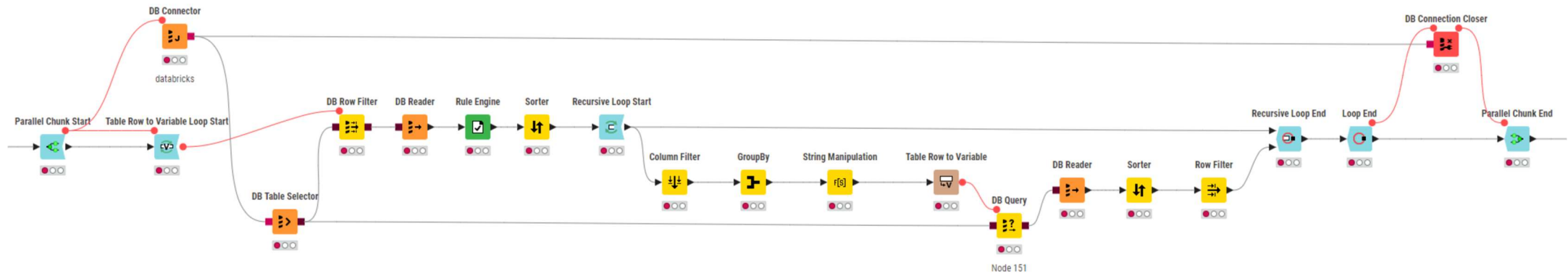**Input Validation for Format, Spaces, etc**

Various Search Directions

Components -> End-Product

End-Product -> Components

Output as Download in .csv File

- End-Users are totally unexperienced in SQL and Table Structure
- Workflow Deployed on KNIME-Server used via Web-Interface

BOSCH

How to Analyze Data on DataBricks without Coding
How Complex can it go

Parallel Chunk Loop
- Works on DataBricks really well -> we love it!
- Be care regarding Load & Costs!

DB-Reader in Loop
- Take care when using multiple DB-Reader in a Loop
- Filter Properly

DB-Table Selector
- Use a basic SQL to select only needed Columns.
- Then use that further in your Workflow

DB-Querry
- A further relatively simple SQL runs here. That can be modified by Variables created during loop Execution
- Each loop has ist own variable

DB-Connection-Closer
- As a good Practise closer your Connection if no longer needed

Loop in Loop in Loop in .....
- Yes it works to encapsule multiple loops various Types
- Meanwhile we would solve that by call of Workflow Services

BOSCH

How to Analyze Data on DataBricks without Coding
More Tipps Unity Catalog

When your Setup in DataBricks uses the so called „Unity Catalog"

Make sure in any given Connector which supports it, has in the „Advanced" Tab of Configuration, mostly at the very bottom, „Support multiple databases" is checked on!
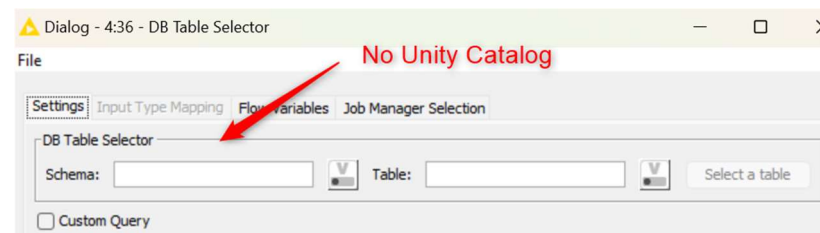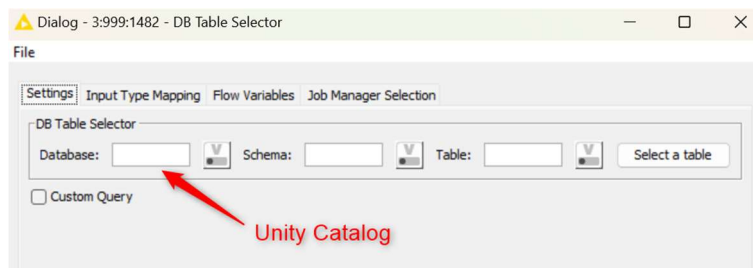
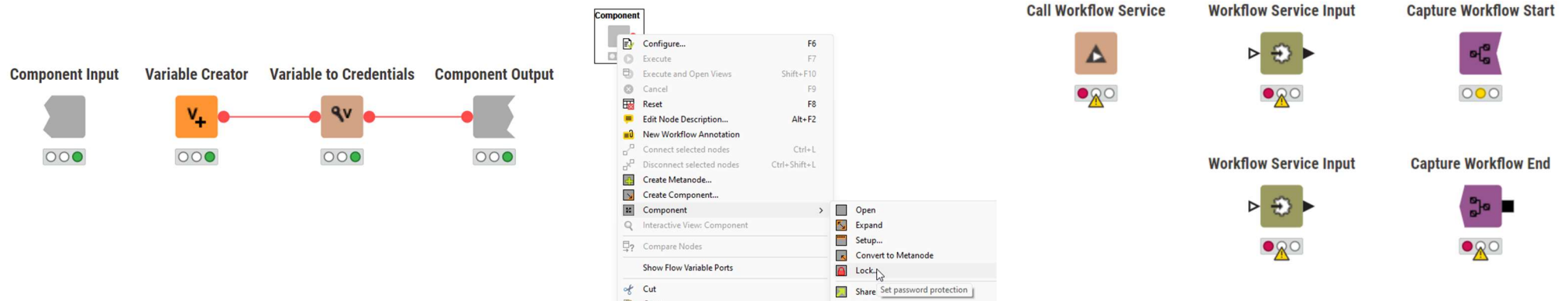# How to Analyze Data on DataBricks without Coding
More Tipps Misc

Handling the Access to various Backend Systems and even Shares. Is a bit painfull, to say the least. Do all of that without breaking your Companies Regulations or Legal Standards is solvable but not easy

Just to mention the obvious: Do not hardcode Credentials in your Workflows!

If you to handle Secrets in large Workflows it just ad complexity

- Get yourself familiar using Workflow Calls. It does not solves everything, but can make your live a lot easier.
- Maybe a locked Component can help you.
  Create 2 Variables-> User & Password, use a „Variables to Credentials", create a Component, Setup this Component that only the Credentials are exposed to outside, and lock that Component.
  Check with your Companies Security People and Guidelines if you can use this!

BOSCH

How to Analyze Data on DataBricks without Coding
More Tipps aside of KNIME

Be careful when you use Tools that generate their own querry towards backends. Or wrap your querry in their own Code.

Automatically generated Code by any given Tool is very likely not the best in Performance

We saw really ugly and resource hungry code fired from Tools to DataBricks Backend
15 Lines of SQL can became 4000 Lines!

Be care when blaming someone or something!

Analyze thoroughly!

BOSCH

How to Analyze Data on DataBricks without Coding
Thanks for your Attention

BOSCH