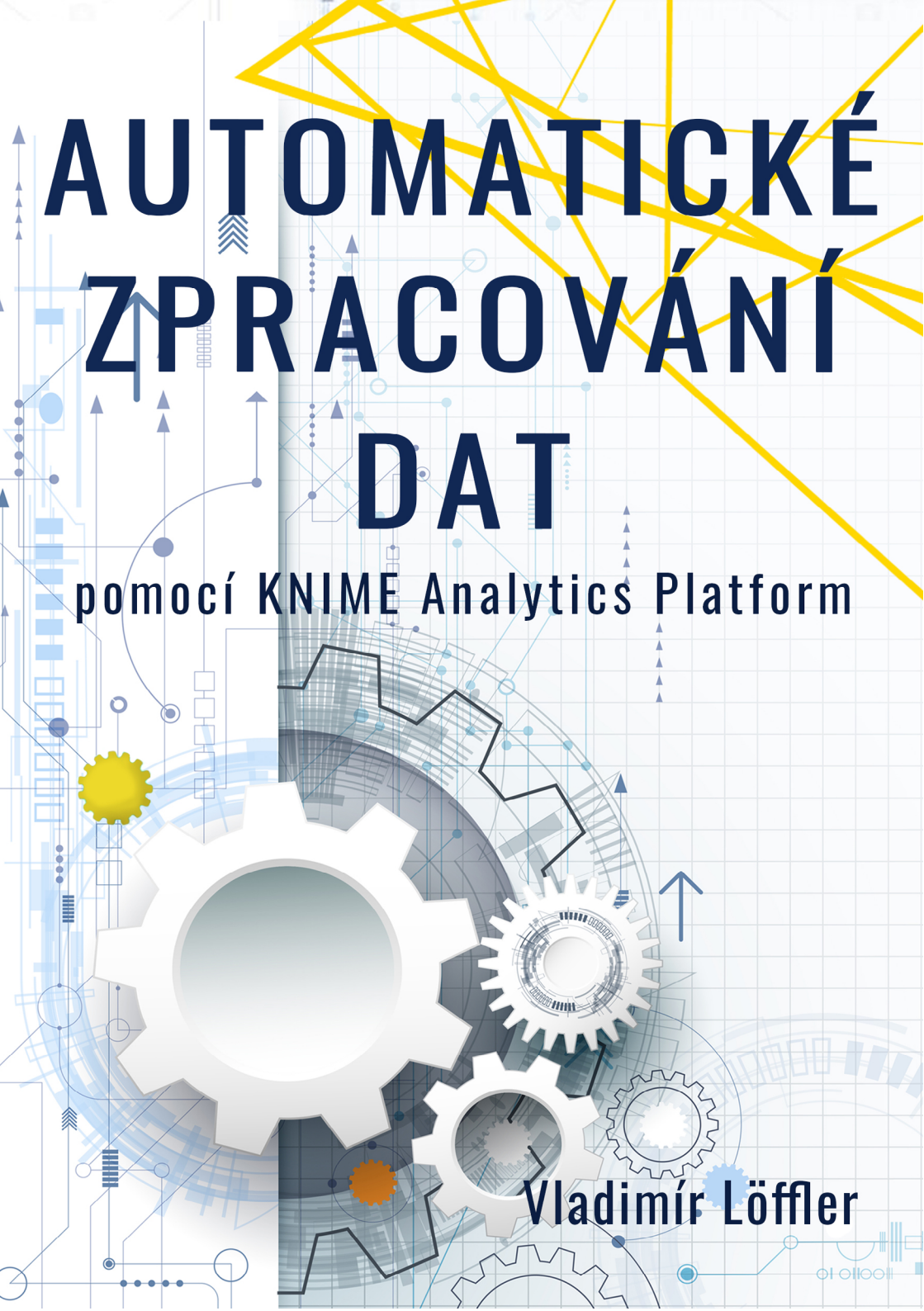
AUTOMATICKÉ ZPRACOVÁNÍ DAT

pomocí KNIME Analytics Platform



Vladimír Löffler

Automatické zpracování dat pomocí KNIME Analytics Platform

Ing. Vladimír Löffler

1. vydání © 2021, Lukáš Vik – E-knihy hned
ISBN ePub formátu: 978-80-7536-324-4 (ePub)
ISBN PDF formátu: 978-80-7536-325-1 (PDF)
ISBN MOBI formátu: 978-80-7536-326-8 (MOBI)
Konverze do elektronických formátů:
webdesignér Lukáš Vik
https://www.lukasvik.cz

Obsah

Counting Loop – smyčka s daným počtem opakování
Generic Loop – smyčka s podmínkou na konci
Recursive Loop – rekurzivní smyčka

5. KNIME – podmíněné větvení Workflow

KNIME – větvení typu IF

Uzly workflow – IF Switch – End IF
Příklad – sloučení denních a týdenních KPI reportů

KNIME – větvení typu CASE

Příklad – výpočet celkové týdenní potřeby pracovníků
Dokončení workflow
„Tunning“ workflow

6. KNIME – automatizace volání workflow

Asistovaná automatizace – poloautomatická workflow

Spouštění poloautomatických workflow

Autonomní automatizace – automatické workflow

Spouštění automatických workflow

7. KNIME – automatizace a přehlednost workflow

Přehlednost workflow

Popisy uzlů
Anotace
Metanode
Component
Meta informace

8. KNIME – automatizace a rychlost workflow

Rychlost

Nastavení prostředí

Uzel Cache

Parallel execution

Další možnost optimalizace výkonu workflow

1. Úvod

Začínají dvacátá léta 21. století. Nová industriální revoluce již běží na plné obrátky a tempo zavádění přelomových technologií bude v následujících letech jen zrychlovat. Rychlá adaptace na probíhající změny světa je klíčovým faktorem úspěchu firem i jednotlivců. Schopnost efektivně shromažďovat a vyhodnocovat všudypřítomná data pomalu přestává být konkurenční výhodou, ale spíše podmínkou nutnou pro fungování a přežití na současných turbulentních trzích. Tento fakt dramaticky zvyšuje nároky na pracovní sílu a její „datovou gramotnost“.

Aspekt odbornosti a produktivního využití potenciálu lidí

Roste počet oborů, u kterých je automaticky předpokládána schopnost zpracovávat data. Inzeráty typu „přijmeme skladníka, znalost MS Excelu podmínkou“ jsou již běžnou součástí pracovního trhu. Jsou dokonce obory, kde práce s daty časově převáží vlastní odbornou práci. Plánovači, účetní, personalisté, pracovníci controllingu, zásobovači i jiní odborníci dostávají svou mzdu z části za provádění svého hlavního oboru, a pak za svou „druhou odbornost“ - opakované načítání, spojování, rozdělování, transformace a ukládání dat. Podstatná část těchto operací s daty není kreativní, ale naopak „otrocká“, opakující se každý den, každý týden, každý měsíc…

Věříme, že otrocká práce, včetně sofistikované otrocké práce s daty, do 21. století nepatří. Věříme, že hodnota pracovníků je v jejich schopnosti provádět kreativní činnosti, a že odborník by měl mít možnost soustředit se na svou hlavní odbornost. Věříme, že pracovní doba by měla být využívána pro odbornou práci, a ne na „trápení se“ s tabulkami. Opakované vytváření tabulek a reportů však mnohdy tvoří značnou část pracovní náplně řady odborníků.

Proč by hodnota účetního, kontrolora kvality, plánovače výroby, logistika, nebo manažera měla být závislá na tom, jak umí vyrábět tabulky?

Aspekt času

Existují ovšem tabulky, reporty a analýzy, které jsou velmi užitečné, a informace v nich obsažené mohou vést ke snížení nákladů, ke zvýšení výnosů, nebo pomáhají managementu pružněji a efektivněji reagovat na každodenní situace. Jsou však k dispozici **jen občas** (týdně, měsíčně, ročně), protože je náročné je vytvářet každý den, nebo dokonce na požádání. Několikadenní příprava měsíčního reportu, kdy jsou stahována, spojována, čištěna a transformována data z několika zdrojů, nebývá ničím výjimečným.

Věříme, že hlavní přidaná hodnota podnikových reportů je v tom, že co nejvíce odráží aktuální stav firmy. Tak je možné pružně reagovat na změny světa a firmu efektivně řídit. Věříme, že potřebná data by měla být k dispozici v lidsky čitelné podobě v podstatě na požádání (a nikoli jednou týdně nebo měsíčně). Rovněž věříme, že zpracování dat má být automatizováno v nejvyšší možné míře, což nejen eliminuje chyby a prostoje, ale navíc umožní odborníkům soustředit se na svůj obor, a manažery podpoří v jejich snaze dobře vést svůj podnik.

Velmi efektivním nástrojem nejen pro automatizaci zpracování dat (od individuálního použití až po použití v nadnárodních korporacích) **je Knime Analytics Platform**.

Knime Analytics Platform

Knime Analytics Platform je aplikace původně vyvinutá na univerzitě ve švýcarské Kostnici (University of Konstanz) a řadu let patří mezi nejlepší světové platformy pro zpracování a analýzu dat, strojové učení a datovou vědu.

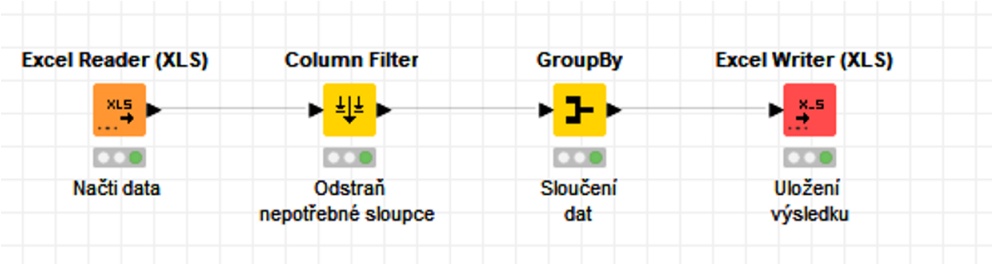
Knime patří mezi open source software, a k jeho používání není třeba komerční licence (tzn. Knime je k dispozici zdarma). Pro rozsáhlé použití Knimu existuje i komerční licence, která umožňuje spolupráci v týmu, obsahuje webový portál pro publikování automatických reportů (podobně jako např. Microsoft Power BI) a také Knime Server, pro pokročilé řízení automatických workflow.

(blíže viz Knime Open Source Story: https://www.knime.com/knime-open-source-story)

Jak Knime pracuje s daty?

Knime používá grafické rozhraní pro vytvoření tzv. *workflow*, kdy jednotlivé kroky workflow, tak zvané uzly, provádí zamýšlené operace s daty. (Pokud znáte prostředí MS Excelu, můžete si Knime workflow představit jako přehledné grafické Makro, nebo jako transformační proces v MS Excel Power Query.) Nastavení uzlů je **intuitivní a nevyžaduje znalosti programování**. Způsob práce s uzly je interaktivní a uživatelsky příjemný.

Příklad jednoduchého Knime workflow:



Workflow lze vytvářet pro individuální analytické úkoly, nebo pro opakovaná spouštění – pro asistovanou, nebo autonomní automatizaci. Kroky workflow můžeme spouštět postupně (s průběžnou kontrolou výsledků kroků), nebo celé – jedním kliknutím. Workflow je také možné volat externě (z jiného programu), nebo naplánovat pomocí plánovače s automatickým spuštěním ve vybraném čase.

Analytická cesta

Knime workflow umožňuje vytvořit a uchovat kompletní analytický postup, kterým ze surových, nezpracovaných dat postupně vznikají data organizovaná, zpracovaná přesně podle našich potřeb. To může být velmi výhodné, pokud nějakou analýzu provádíme jen jednou za delší období, nebo pokud se zadání pokaždé jen mírně odlišuje. Pak, i když takové analytické workflow nemusí být vhodné pro automatizaci,

nám vizualizovaná analytická cesta pomůže rychle se zorientovat a naši analýzu efektivně dokončit.

Co publikace obsahuje

V této publikaci bychom se rádi věnovali především automatizaci zpracování dat – poloautomatickým a autonomním datovým workflow. Postupně se seznámíme se základními stavebními kameny automatizace – proměnnými, smyčkami a větvením. Také si ukážeme, jak naše datová workflow optimalizovat, ošetřit proti chybám, a jak je automaticky spouštět.

Jednotlivé kapitoly se věnují teorii i praxi. Snažili jsme se o vyvážený přístup, proto je každý výklad doplněn příkladem, který je možné prakticky vyzkoušet.

Teorie

- Vysvětlení problematiky
- Příklady použití s detailním rozborem jednotlivých operací
- Poznámky a postřehy z praxe
- Odkazy na další příklady a zdroje informací

Praxe

Věříme, že interaktivní výuka je pro studenty nejpřínosnější. Připravili jsme proto sdílenou složku, ze které si můžete stáhnout balíček, který obsahuje všechna probíraná workflow (20 plně funkčních Knime workflow), včetně použitých dat.

Poznámka:
Žádný z příkladů nevyžaduje placenou verzi Knimu (tzn. nevyžaduje Knime Server).

Použité symboly (platí pro tištěnou a PDF verzi; formáty e-pub a mobi obsahují pouze text)

	Výklad
	Příklad
	Poznámka
	Technické jméno workflow
	Odkaz na další zdroje
	Soubor

Knime a operační systémy

Příklady uvedené v jednotlivých kapitolách byly připraveny převážně v Knime 4.2 pro MS Windows 10. Pracovat s Knimem je možné a podporované i pod operačními systémy Linux a macOS.

Testovali jsme Knime na Ubuntu 18.04 LTS (Linux) a na macOS Big Sur (Apple macOS), ale vrátili jsme se k verzi pro MS Windows 10. Důvod byl ten, že v prostředí MS Windows 10 byl Knime stabilní, což o Linuxové a macOS verzi nemůžeme s čistým svědomím potvrdit.

2. Začínáme

V následujících kapitolách se postupně věnujeme vysvětlení základních stavebních kamenů automatických datových workflow, což jsou:

- Globální a flow proměnné
- Aktivní prvky pro vytváření proměnných (uživatelské dialogy)
- Různé druhy smyček (cykly)
- Podmíněné větvení (možnost vybrat alternativní zpracování dat na základě vzniklé situace)
- Ošetření workflow proti chybám
- Výkon workflow
- Vlastní automatizace (možnost spouštět workflow automaticky, nebo poloautomaticky)

Předpokládáme, že základy práce s Knime Analytics Platform již ovládáte. Pokud tomu tak není, doporučujeme seznámení se základy práce v prostředí Knime.

Základům práce s Knimem se detailně věnujeme v publikaci „**Základy práce s Knime Analytics Platform, B. Štětinová, 2021**“, případně v angličtině existují další školení, blogy a fóra (viz odkazy na další zdroje v poslední kapitole).

V případě, že Knime ještě nemáte nainstalovaný, můžete použít automatické instalační programy (pro MS Windows, Linux a macOS) umístěné na stránkách Knimu.

Odkaz: https://www.knime.com/downloads/download-knime

Pracovní soubory

Složku se všemi probíranými workflow a příslušné datové soubory si můžete stáhnout zde (Heslo pro stažení je Knime2020):

Knime workflow

Stáhnout archiv – Knime_Automation_book

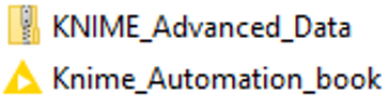
Datové soubory

Stáhnout archiv – Knime_Advanced_Data

Alternativní odkaz (Heslo pro stažení je Knime2020):

https://strojove-uceni.eu/downloads/

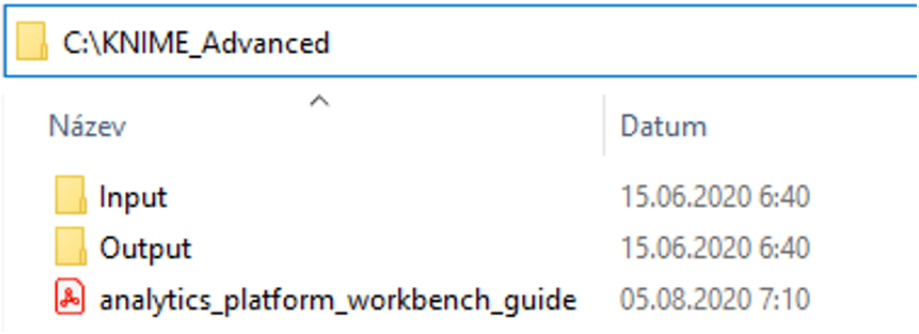
Po stažení byste měli mít k dispozici tyto soubory:



Kam umístit data

Stažený archiv *Knime_Advanced_Data* rozbalte. Složku s daty umístěte na disk C:\ (pokud máte disk jinak pojmenovaný, nebo preferujete jinou složku pro data, změňte v příslušných uzlech workflow cesty pro načtení, nebo uložení dat).

Složka s daty



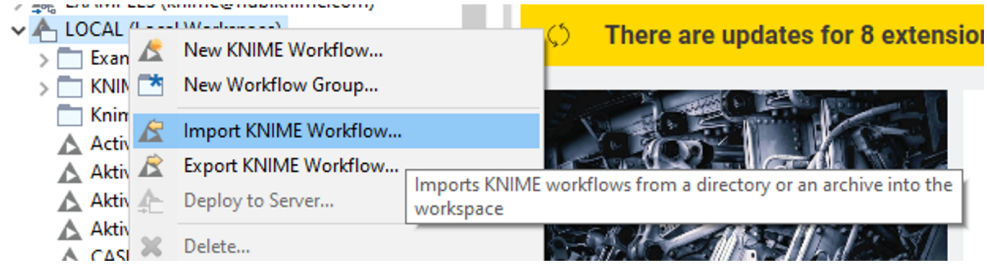
Poznámka:
ve složce output a jejich podsložkách je vždy umístěna speciální složka Archiv, která obsahuje připravené výsledné soubory (aby bylo možné porovnání vašich a našich výstupů).

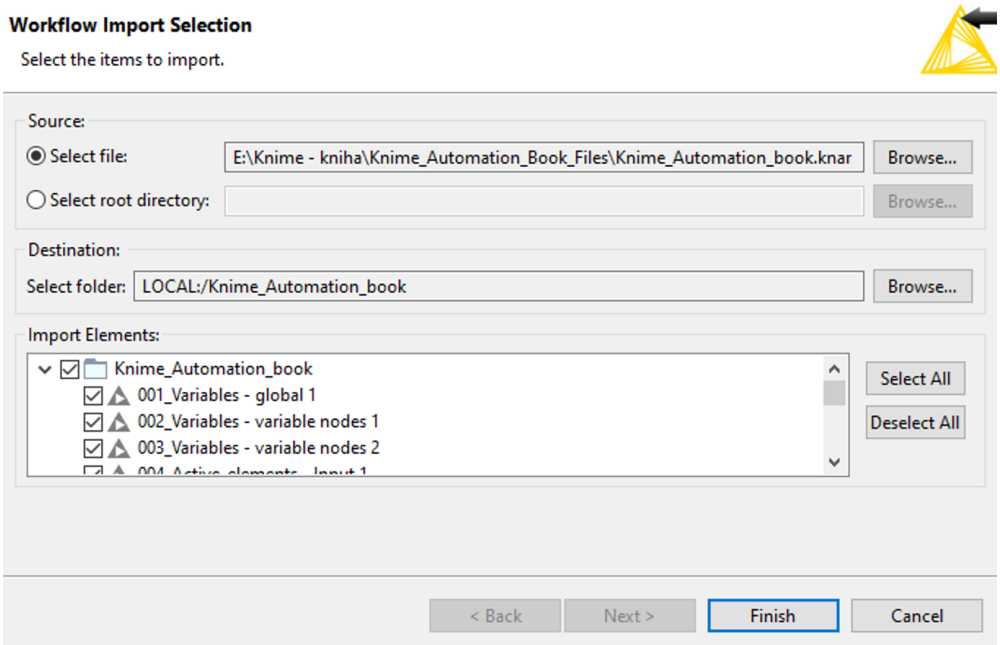
Poznámka k anglickému jazyku:
pro pojmenování souborů a datových workflow používáme anglický jazyk. Důvodem je především zvyk (angličtina je jazykem Knime komunity) a rovněž možnost sdílet workflow bez ohledu na hranice. Použití českého jazyka v prostředí Knimu je samozřejmě také možné, i když Knime na českou lokalizaci teprve čeká.

Import Knime workflow

Knime archiv *Knime_Automation_book* s pracovními workflow umístěte do vašeho Knime workspacu, pomocí volby „Import KNIME Workflow…“. Po dokončení importu by workflow měla být dostupná ve stromu lokálních workflow.

Import KNIME Workflow





Poznámka:
některá workflow vyžadují ke svému spuštění instalaci rozšiřujících balíčků. V rámci výkladu uvádíme vysvětlující kroky k jejich instalaci. Všechna workflow byla vytvořena ve verzi Knime 4.1 a 4.2 a byla otestována také ve verzi 4.3. Knime 4.3 přinesl nové uzly pro načítání dat (například z MS Excelu), proto některé uzly v Knime 4.3 mohou vybízet

k použití nové verze (označení Deprecated v názvu uzlu). Knime respektuje zpětnou kompatibilitu, takže funkčnost workflow vytvořených v předešlých verzích bude vždy zachována.

Automatizace

Co myslíme automatizací?

Úroveň automatizace v kontextu Knime volně definujeme takto:

Úroveň	Název	Stručný popis
0	Jednorázové použití	Workflow je vytvořené pro vyřešení jednorázové datové úlohy (snadné, nebo složité), a v budoucnu nepočítáme s jeho opětovným spouštěním.
1	Základní poloautomatické použití	Vytvoříme workflow pro opakované použití (například pro slučování a transformaci dat), ale nepotřebujeme prvky zobecnění a pokročilé automatizace (parametry uzlů workflow měníme ručně dle potřeby).
2	Poloautomatické použití s prvky automatizace	Jako úroveň 1, ale parametry uzlů řídíme pomocí proměnných a automaticky reagujeme na vzniklé situace (používáme větvení, smyčky, ošetření chybových stavů apod.).
3	Plně automatický běh	Workflow je postavené tak, aby nebylo nutné spouštět je ručně. Běh workflow řídíme pomocí parametrických tabulek a proměnných. Workflow obsahuje prvky zobecnění, ošetření chybových stavů a generuje chybový log v případě neočekávaných situací. Workflow lze externě plánovat a volat (pomocí plánovače, Knime Serveru, nebo pomocí vlastní aplikace, RPA robota apod.).

Příklady automatizace

Knime lze použít na automatizaci téměř jakékoli datové úlohy. Pro představu zde uvedeme několik typických příkladů. Konkrétní řešení vybraných úloh pak detailně popisujeme v rámci výkladu.

Sloučení souborů	Potřebujeme sloučit soubory, například ceníky, reporty ze všech oddělení, ankety, výsledkové listiny, chybové logy apod. do jednoho reportu. Souborů může být jen několik, nebo stovky či tisíce. Soubory mohou mít různé formáty (MS Excel, textový soubor, CSV, JSON apod.)
Data mining	Potřebujeme z různých datových zdrojů vytáhnout konkrétní informace a použít je ve výsledném reportu.
Sloučení listů MS Excelu	Máme data na několika listech v MS Excelu (např. 1 list = 1 den, nebo 1 list = pobočka apod.) a potřebuji je spojit do jedné tabulky pro další vyhodnocení.
Rozdělení dat	Máme velkou tabulku a potřebujeme jí rozdělit a distribuovat konkrétním lidem, např. dle produktu, nebo dle odpovědnosti.
Založení složek	Potřebujeme vytvořit strukturu složek v MS Windows dle nějakého vzoru (například projektovou složku na sdíleném disku dle vzoru v MS Excelu, nebo dle vzorové složky v MS Windows).
Příprava reportů	Máme například export dat z ERP systému, ten potřebujeme spojit s dalšími daty a poté data vyčistit a transformovat ve výsledný report, nebo data připravit jako zdroj pro nějaký management dashboard (v Microsoft Power BI apod.).
Pokročilá datová analytika	Knime lze použít i pro automatizaci strojového učení a umělé inteligence, včetně nasazení v produktivním provozu.

3. KNIME – proměnné – úvod

V prostředí Knime můžeme pracovat s proměnnými - tzv. *Flow proměnné*. Proměnné nám umožňují provádět složitější operace v rámci našich workflow. Do proměnných můžeme průběžně ukládat užitečné hodnoty, a ty pak používat v uzlech workflow, kdykoli je třeba.

Proměnnou definuje **jméno, datový typ** a **hodnota**. Například proměnná *city* může mít datový typ *String* (tzn. textový řetězec) a hodnotu *Tokyo*. Nebo proměnná *age* může mít datový typ *Integer* (tzn. celé číslo), a hodnotu *62*.

Tabulka základních typů proměnných:

Typ proměnné	Význam	Ikona	Příklad
String	Textový řetězec	S	Osaka
Integer	Celé číslo	I	1024
Double	Číslo s desetinnou čárkou	D	14.2345
List/Collection	Seznam hodnot	[...]	[Kyoto, Tokyo, Osaka]
Time	Čas		06:39:06
Date	Datum	31	2020-11-26
Date&Time	Datum a čas		2020-11-26T06:39:06
Data&Time	Datum a čas a zóna		2020-11-26T06:39:06+01:00[Europe/Prague]

Kdy Flow proměnné využijeme

Flow proměnné využijeme buď u složitějších jednorázových workflow (například u workflow, kde je třeba větvení na základě nějakých pravidel, případně workflow obsahujících cyklicky prováděné operace), nebo u dynamických workflow předpokládajících opakované použití.

Pomocí Flow proměnných je možné vytvořit poloautomatická workflow (pro scénáře asistované automatizace), a dokonce workflow, která budou zcela autonomní (spouštěná bez zásahu člověka).

Statická workflow, ve kterých nedochází ke změně žádných parametrů, a není třeba dynamicky reagovat na výsledky plynoucí z jednotlivých uzlů, flow proměnné nepotřebují.

Druhy proměnných

Máme dva druhy proměnných:

- Proměnné vytvořené pro celé workflow – *globální proměnné*
- Proměnné vytvářené v uzlech workflow

Globální proměnné jsou definovány před spuštěním workflow a jsou viditelné pro všechny jeho uzly, včetně uzlu, který je volaný jako první.

Proměnné vytvářené v uzlech workflow jsou viditelné pouze uzlům, které navazují na uzel, ve kterém proměnná vznikla.

V uzlech workflow mohou proměnné vznikat několika způsoby:

- V rámci uzlu, definované ve stromu *Flow Variables*
- V rámci konfigurace uzlu, pomocí tlačítka pro zadání proměnné (jen některé uzly)
- Pomocí *Input* a *Widget* uzlů
  - definujeme proměnnou, a její hodnota pak vznikne pomocí uživatelského dialogu
- Pomocí speciálních uzlů pro proměnné
  - uzly přímo určené k vytváření proměnných, buď z hodnot tabulek, nebo podle pravidel
- Pomocí Java uzlů