SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
MATERIÁLOVOTECHNOLOGICKÁ FAKULTA SO SÍDLOM V TRNAVE

Ing. Peter Střelec

Autoreferát dizertačnej práce

Získavanie znalostí z databáz výrobných podnikov (štruktúrované dáta) pre potreby riadenia procesov

na získanie akademického titulu: doktor (philosophiae doctor, PhD.)

v doktorandskom študijnom programe: Automatizácia a informatizácia procesov

v študijnom odbore: Kybernetika

Forma štúdia: denná

Miesto a dátum: Trnava, 30.5.2023

STU

Dizertačná práca bola vypracovaná na Katedre aplikovanej informatiky a automatizácie a Ústave aplikovanej informatiky, automatizácie a matematiky

Predkladateľ: Ing. Peter Střelec
Ústav aplikovanej informatiky, automatizácie a mechatroniky (MTF)
Materiálovotechnologická fakulta so sídlom v Trnave,
Slovenská technická univerzita v Bratislave
Jána Bottu 2781/25
917 24 Trnava

Školiteľ: prof. Ing. Pavol Tanuška, PhD.
Ústav aplikovanej informatiky, automatizácie a mechatroniky (MTF)
Materiálovotechnologická fakulta so sídlom v Trnave,
Slovenská technická univerzita v Bratislave
Jána Bottu 2781/25
917 24 Trnava

Oponenti:

Autoreferát bol rozoslaný: .....

Obhajoba dizertačnej práce sa bude konať dňa ...... **o** .....**h**.

Na Materiálovotechnologickej fakulte fakulte STU so sídlom v Trnave, J. Bottu 25, 917 24, Trnava

.....

prof. Ing. Miloš Čambál, CSc.
dekan MTF STU

Obsah

Úvod

Slovo dáta predstavuje označenie pre základné informačné jednotky. Dáta - údaje predstavujú zaznamenané určité fakty o procesoch alebo prvkov reálneho sveta. Dáta môžu byť prezentované rôzne a môžu byť zaznamenávané ako písmená, čísla alebo ich kombinácie. Zápis teda môže byť rôzny, ale dáta majú jednu spoločnú vlastnosť a to fakt, že obvykle nesú zmysluplný obsah, ktorý nazývame informácia. Je tiež ľahko možné, že informáciu nie je možné jednoznačne identifikovať, alebo nepredstavuje nič nové. Vzhľadom na predchádzajúce údaje, môže byť predstavená teda ako informácia, ktorej obsah je nulový. Dáta predstavujú kvalitatívny a kvantitatívny popis skutočností. Vznikajú pozorovaním alebo sú generované vykonávaním určitých činností, ktoré je možné zaznamenávať. Dáta sa transformujú do informácií, keď sa na ne nazerá v kontexte alebo po vykonanej analýze. Medzi dátami a informáciami je presne definovaný rozdiel. Informácie, ktoré sú reprezentované v kontexte a sú určené na vymedzenie rámca pre kognitívne procesy sa nazývajú znalosti. Samotné dáta sú málokedy užitočné samostatne. Aby sa z dát stali informácie, je potrebné ich uviesť do súvislosti. Ak sú údaje spracovávané, interpretované, organizované, štruktúrované alebo prezentované tak, aby boli zmysluplné alebo užitočné, hovorí sa im informácia [1].

Tento proces cielenej analýzy viedol k vzniku vednej oblasti známej ako dátová veda (Data Science - DS). Dátová veda sa zaoberá skutočnosťou, že údaje nesú so sebou aj informácie, ktoré môžu byť užitočné a podporujú rozhodovanie alebo môžu ovplyvniť vývoj nových prístupov a technológií. Jedným z prístupov je aj získavanie znalostí z databáz (Knowledge Discovering in Database – KDD) a dolovanie údajov (Data Mining – DM). Dáta produkujeme denne vedome alebo nevedome a podľa výskumu bolo vyprodukovaných do roku 2020 až 40.900 EXABYTE dát [2].

Tento objem dát nie je možné spracovať klasickými technikami a daná problematika sa rieši technikami veľkých dát (Big Data - BD). Veľké dáta sa zaoberajú hľadaním a získavaním nových faktov, informácií a vedomostí z veľkého objemu dát.

1 Analýza a problematika súčasného stavu

Proces objavovania znalostí (KDP - The knowledge discovery process alebo aj Knowledge discovery in databases). Podľa Fayadda (1996) [3] predstavuje „netriviálny proces identifikácie (rozpoznávania) platných, nových, doposiaľ neznámych, potencionálne užitočných a jednoznačne zrozumiteľných vzorov z dát.“ Klasicky tento proces je vykonávaný nad dátami, ktoré sú uložené v databázach. Vzhľadom na vývoj čoraz viac vystupuje potreba analýzy dát, ktoré sú používané v komunikačnom procese počas všetkých aktivít, ktoré sú realizované. Podnetom k novému prístupu je digitálna revolúcia, ktorú predstavuje Industry 4.0.

Priemyselné odvetvie sa postupne snaží zavádzať štandardy a prispôsobovať alebo meniť svoje prostriedky tak, aby boli schopné sa čo najviac priblížiť štandardu Industry 4.0. Zmena na štandard Industry 4.0 je často prirovnávaná k revolúcií, ktorá významne zmení postavenie ľudí a predpokladá sa, že dôjde k takému dopadu pre ľudstvo, aký nevyvolala doteraz žiadna iná technologická zmena, ktorá sa udiala v minulosti [4].

Na dosiahnutie cieľa je potrebné zabezpečiť prepojenie medzi procesmi, ktoré prebiehajú počas tvorby výsledného produktu. Bol zavedený po jem Smart Factory (inteligentná továreň), kde je v ideálnom prípade tesné prepojenie produktov, výrobných zariadení, strojov, logistiky, monitorovacích systémov a ľudí [5].

Big Data nadväzujú na dáta vznikajúce pri výrobe a nasadzovaní štandardov Industry 4.0. Big Data sú oblasťou, ktorá sa zaoberá spôsobmi analýzy, systematického získavania informácií alebo iného zaobchádzania so súbormi údajov, ktoré sú príliš veľké alebo zložité na to, aby ich bolo možné zvládnuť tradičným aplikačným softvérom na spracovanie údajov. „Merať znamená vedieť“ vyhlásil lord Kelvin. Vo Francúzku v 19. storočí vznikol presne definovaný systém merných jednotiek, pomocou ktorých je možné definovať čas, priestor a všeobecne fyzikálne veličiny. O pol storočie neskôr v 20tich rokoch 20. storočia sa sen o dokonalom meraní zmenil objavom kvantovej mechaniky. V priemyselnej výrobe, pri počítačovo podporovanom riadení výroby, presnosť, kvalita a efektivita predpokladajú presné dáta a čo najmenej chýb. Mať k dispozícií presné dáta je nutné k technikám, ktoré sú používané na dolovanie údajov. V terminológií big data, sa na chyby pozerá odlišným spôsobom. Viktor Mayer-Schoenberger a Kenneth Cukier popisujú tento fakt. Ak pripustíme nepresnosti a chybovosť v dátach, môžeme informácie naopak získať a nie stratiť. Niečo za niečo: keď mierne zmiernime svoje nároky na prípustný počet chýb, môžeme zhromaždiť viac dát. Neplatí len „radšej viac dát ako menej“ ale platí „radšej viac horších dát ako menej lepších“ [6].

Cieľom dizertačnej práce je analyzovať dostupné metódy pre vyhľadávanie informácií v štruktúrovaných dátach výrobných podnikov. Navrhnúť metódy normalizácie dát, analyzovať a navrhnúť proces získavania a nahrávania dát a následne ich klasifikovať a vytvoriť metódy, ktoré umožnia identifikovať, kvantifikovať a vizualizovať dáta tak, aby bolo možné ich použiť pre lepšie rozhodovanie ako aj identifikáciu neznámych alebo skrytých faktov, informácií a vedomostí, ktoré umožnia skvalitniť a riadiť výrobný proces.

2 Ciele a východiská dizertačnej práce

Dizertačná práca sa venuje problematike návrhu platformy na získavanie nových znalostí, na základe nasadenia technológií Industry 4.0, analýzy komunikácie a zabezpečenia a ochrany bezpečnosti a tým aj stability výroby. Moderné výrobné systémy nie sú izolované jednotky a čoraz väčšmi je nasadzované rozšírenia a prostriedky, ktoré doplňujú a vylepšujú výrobu. Dochádza k nasadeniu IIoT zariadení a štandardizácií konceptu Smart Factoy, ktoré využívajú externé výpočtové kapacity v rámci cloud computingu.

Nasadenie doplňujúcich výrobných prostriedkov je potenciálne možné narušenie výroby. Dizertačná práca sa zaoberá návrhom platformy pre vyhľadávanie nových znalostí na základe prepojenia existujúcich dostupných dát a predstavuje návrh, ako implementovať a navrhnúť univerzálu platformu na získavanie dát v reálnom čase, a skombinovať dostupné dáta z databázových zdrojov tak, aby bolo možné získať nové znalosti na riadenie výroby.

Pre komplexnosť tejto problematiky je v práci vpracovaných viacero postupov, ktoré pokrývajú celú škálu situácií.

Základom pre analýzu dát je správne definovaný ETL procesom. Je predstavených viacero prístupov, ktoré umožňujú pokrzť komplexný zber údajov až po návrh adaptérov, ktoré sú určené pre spracovanie dát vo forme dátových tokov.

Z analytického hľadiska je urobený návrh a popis pre základné dávkové spracovanie údajov zo statických dávkovo spracovateľných zdrojov dát ako sú súbory, databázy a dátové skady. Nasleduje časť o spracovaní a príprave dát z dátových tokov – data streamov.

Získavanie znalostí z databáz výrobných podnikov (štruktúrované dáta) pre potreby riadenia procesov

1.Analyzujte problematiku získavania znalostí a spracujte teoretické východiská pre danú oblasť.

2.Navrhnite ETL proces so zameraním na štruktúrované dáta.

3.Navrhnite platformu na získavanie znalostí a vaše riešenie experimentálne overte.

4.Zovšeobecnite navrhované riešenie vo forme modelu pre potreby implementácie vo výrobnom podniku.

5.Zhodnoťte dosiahnuté výsledky a splnenie cieľov práce.

Hlavným cieľom dizertačnej práce je návrh platformy na spracovanie a analýzu bázy dát a dátových tokov so zameraním na objavovanie nových znalostí na základe technológií Industry 4.0 (prevažne Big Data a IIoT) pre potreby efektívnejšieho riadenia výrobných procesov, s prihliadnutím na atribúty bezpečnosti a spoľahlivosti.

Splnenie hlavného cieľa dizertačnej práce je podmienené naplnením stanovených čiastkových cieľov, ktoré sú sformulované nasledovným spôsobom:

1. **Analýza a spracovanie teoretických východísk, spôsoby komunikácie a integrácie. Skúmanie poznatkov z danej problematiky z domácej a zahraničnej literatúry a vedeckých štúdii.** Analýza a vyhodnotenie dostupných analytických metód, ich aplikácia a skúmanie možností ich určenia a využitia pre riešenie úloh a ich potenciálne nasadenie pre analýzu bázy dát alebo vhodnosť pre určenie aplikácie v dátových tokoch.
2. **Stanovenie cieľov a návrh platformy pre získanie dát potrebných na spracovanie.** Na základe vypracovania analýzy, vytvoriť návrh a proces a orientovaní sa v problematike stanoviť ciele a smerovanie skúmanej problematiky.
3. **Návrh ETL procesu, stanovanie typov komunikácie, ich dopad a návrh realizáciu procesov.** ETL proces má zabezpečiť spoľahlivé a rýchle doručenie dát pre potreby objavovania znalostí zo štruktúrovaných dát z heterogénnych zdrojov generovaných dát.
4. **Návrh a overenie hľadania nových znalostí na základe riadenia výroby a sieťovej komunikácie. Návrh analytických vrstiev a priradenie metód.** Na základe, ktorej je realizovaný experiment pre identifikáciu neželaného sieťovej prevádzky v reálnom čase, pri nasadení IIoT zariadenia.
5. **Návrh a implementácia aplikácie výsledkov pre nastavenie systémov a pre potreby monitorovania prevádzky z pohľadu zvýšenia bezpečnosti a zabezpečenia kvality výroby.** Na základe experimentu, je navrhnutá aplikácia dynamickej zmeny konfigurácie sieťovej prevádzky a tým ošetrená problematika dopadu na spoľahlivosť výrobného procesu a zamedzenie potenciálne nebezpečnej integrácie IIoT zariadení, alebo zariadení, ktoré nemajú ovplyvňovať riadenie výroby. Zároveň je možné takto identifikované sieťové komunikačné odchýlky zobraziť a poskytnúť obsluhe prehľad o dianí vo výrobe.
6. **Overenie a zhodnotenie navrhnutých riešení.** Implementáciou experimentálneho návrhu a aplikácie navrhnutej platformy na získavanie nových znalostí, overiť účinnosť v reálnej výrobnej infraštruktúre. Na základe výsledkov zhodnotiť efektivitu navrhnutých a implementovaných riešení.
7. **Posúdenie prínosov dizertačnej práce pre špecifikovanú oblasť.** Na základe analýzy návrhu, implementácie, overenia a zhodnotenia riešení špecifikovať prínosy dizertačnej práce.

3 Návrh analytických vrstiev a priradenie metód

Vo výrobnom procese je potrebné identifikovať systémy a ich úloha vo výrobe. Na základe analýzy jednotlivých vrstiev, je možné potvrdiť tvorbu dát na každej vrstve. Zozberaním a analýzou dát z jednotlivých vrstiev je možné vytvoriť návrh na objavovanie nových znalostí v systéme a umožniť proces pre podporu rozhodovania.

Na všetkých úrovniach výrobného podniku sa dajú položiť tieto základné otázky.

- Čo sa práve deje vo výrobe?
- Prečo sa to deje?
- Čo sa bude diať?
- Čo chystať a pripraviť pre budúcnosť?

Všetky tieto otázky, je možné kategorizovať do analytických vrstiev a priradiť im metódy, a analytické nástroje, ktoré ich dokážu zodpovedať ako je zobrazené na obrázku číslo 1. Z vykonaných analýz v teoretickej časti je schematicky znázornená závislosť štyroch typov analytických metód, ktoré môžu byť vzájomne prepojené. Podľa typu, je možné klasifikovať úroveň vedomostí alebo hodnotu informácií, ktoré sú potenciálne k dispozícii na danej úrovni. Návrh úrovní je zobrazený na obrázku 24. Úlohou navrhnutých úrovní je zabezpečiť výrobu, a umožniť určenie hodnoty údajov, podľa kategórii kvality. Môžu byť použité na monitorovanie a vyhodnotenie výroby alebo na ďalšie, významnejšie úlohu. Na základy priradených metód a dostupných dát a parametrov, je možné identifikovať problematické oblasti, monitorovať a indikovať chybové udalosti a lebo predikovať chovanie systémov. Nevýhodou je, že tieto hodnoty nemusia zobrazovať aktuálny priebeh. Obvykle sa jedná mierne oneskorený stav, podľa spôsobu a stupňa integrácie výrobných zariadení.

3.1 Objavovanie vedomostí v toku dát

Objavovanie vedomostí v data streame za posledné roky prešlo pozoruhodným vývojom a upútava čoraz väčšiu pozornosť. Po základnom identifikovaní dátových alebo na činnostných ohraničení oblasti je potrebné popísať postupy a algoritmy, ktoré sú vhodné na spracovanie údajov a objavovanie informácií v data streamoch [7]:

- Základné štatistické metódy (Basic statistics)
- Klasifikácia a regresia (Classification and regression)
- Kolaboratívne filtrovanie
- Zhlukovanie (Clustering)

- Redukcia dimenzií (Dimensionality reduction)
- Extrakcia a transformácia funkcií (Feature extraction and transformation)
- Objavovanie často vyskytovaných vzorov (Frequent pattern mining)
- Hodnotiace metriky (Evaluation metrics)
- Optimizácia (Optimization)

3.1.1 Základné štatistické metódy (Basic statistics)

Popisujú základné sumárne štatistiky dát.

- **Súhrnná štatistika**

Pomocou nej sa dá určiť minimálna a maximálna hodnota, identifikácia typov, nulových a nenulových hodnôt, celkový počet prvkov, opakovanie výskytu prvkov a podobne. Tieto metódy sú základnými a sú používané veľmi často a ich implementácia je veľmi často optimalizovaná pre dosiahnutie čo najlepšej výkonnosti.

- **Korelácia**

Výpočet korelácie medzi dvoma alebo viacerými množinami údajov. Táto operácia je bežnou operáciou v štatistike. Poskytuje flexibilitu na výpočet párových korelácií medzi mnohými radmi. Najčastejšie podporované korelačné metódy sú v súčasnosti Pearsonova a Spearmanova korelácia [8].

- *Stratifikovaný odber vzoriek (Stratified sampling)*

Používa sa na generovanie podmnožín dátových setov. Tieto môžu veľmi dobre reprezentovať vlastnosti pôvodných dát. Namiesto náhodného vzorkovania podmnožiny funkcií pre každú množinu dát komponentov, v tejto metóde sú najskôr zoskupené vlastnosti vysoko dimenzionálnych údajov do niekoľkých skupín funkcií, ktoré sa nazývajú vrstvy funkcií. Pomocou stratifikovaného vzorkovania náhodne vzorkujeme niektoré prvky z každej vrstvy a zlúčime vzorkované prvky z rôznych vrstiev, aby bol vygenerovaný súbor dát komponentov. Týmto spôsobom majú súbory dát komponentov lepšiu reprezentáciu klastrovej štruktúry ako v pôvodných údajoch [9].

- **Testovanie hypotéz (Hypothesis testing)**

Testovanie hypotéz je mocným nástrojom v štatistike na zisťovanie, či je výsledok štatisticky významný a nezáleží, či k nemu došlo náhodou alebo nie. Táto vlastnosť je dôležitá pre identifikáciu

informácií. Sada údajov sa modeluje ako hodnoty súboru náhodných premenných, ktoré majú spoločné rozdelenie pravdepodobnosti v niektorých množinách spoločných rozdelení. Pre rozdelenie pravdepodobnosti údajov sa navrhuje alternatívna hypotéza, a to buď výslovne, alebo iba neformálne. Porovnanie týchto dvoch modelov sa považuje za štatisticky významné, ak je pravdepodobné, že podľa prahovej pravdepodobnosti - úrovne významnosti - dôjde k pravdivosti nulovej hypotézy [10,11,12].

- **Generovanie náhodných dát (Random data generation)**

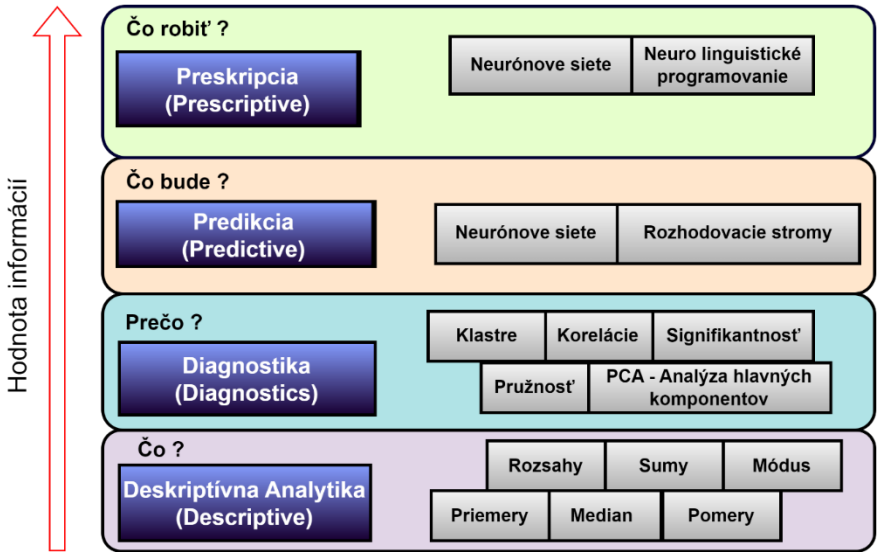
Náhodné generovanie údajov je užitočné pri testovaní algoritmov a riešení prototypov a pri testovaní výkonu.

- **Sumarizácia (Summarizer)**

Sumarizácia pre data stream predstavuje vektor, ktorý reprezentuje stĺpec dát. Dostupné hodnoty sú vždy aplikovateľné na celý stĺpec analyzovaných dát a poskytujú podobné dáta ako súhrnná štatistika – maximálna a minimálna hodnota, nulové a nenulové hodnoty, variabilita, sumy a celkový počet [13].

3.1.2 Klasifikácia a regresia (Classification and regression)

Problematika metód klasifikácie a regresie, ktorá sa spomína takmer vždy spoločne, ale predstavuje jeden dôležitý fakt, ktorý ich rozdeľuje a určuje ich použitie na iné oblasti riešenia. Klasifikácia je v zásade o predikcií nejakého popisu alebo označenia, prípadne určenia stavu true a false. Regresia je špecificky určená na predikciu množstva. Obe kategórie sú zamerané na prediktívne modelovanie a obe sú zamerané na učenia sa funkcie a mapovania jej vstupov na výstupy. Prediktívne modelovanie možno označiť ako matematický problém aproximácie mapovacej funkcie (f) zo vstupných premenných (X) na výstupné premenné (Y). Toto sa nazýva problém aproximácie funkcií. Úlohou algoritmu modelu je nájsť najlepšiu mapovaciu funkciu, ktorá je schopná poskytnúť vzhľadom na čas a zdroje, čo najlepšie výsledky, ktoré sú k dispozícii [14].



Obrázok 1. Štyri úrovne analytických vrstiev a ich aplikácia pre získavanie znalostí z výrobných procesov

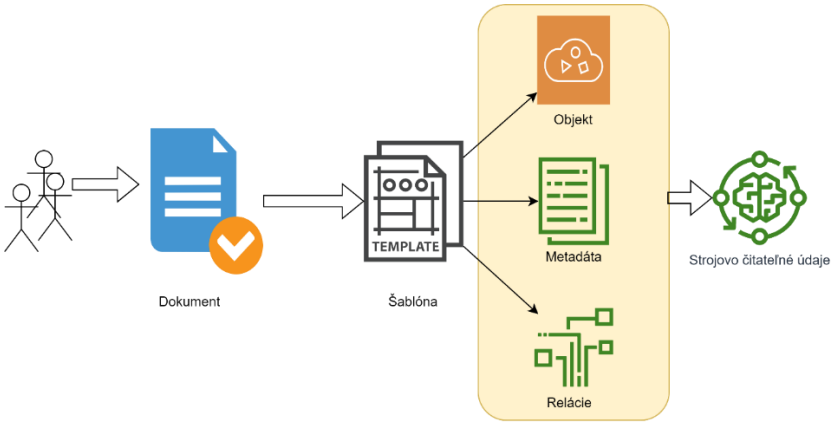
2.2 Návrh transformácie plánov a výrobnej dokumentácie do strojovo čitateľnej podoby

Dáta obsiahnuté v dokumente, alebo v návrhoch je potrebné transformovať tak aby boli vypĺňané napríklad do pripravenej typizovanej šablóny, ktorá logicky rozdeľuje dokument na časti tak aby bolo možné extrahovať hodnoty a použiť ich atribúty. Prirodzeným spôsobom je možné vyplniť a udržiavať šablóny formulárov a nastaviť proces dokumentácie tak, aby podľa potreby ich bolo možné sledovať zmeny, ktoré boli vykonané s časovými údajmi. Výsledkom takto zadaných údajov je existencia dát, ktoré môžu byť reprezentované ako

- objekt
- metadáta
- relácie medzi objektami

- Objekt zachytáva znaky a vlastnosti entity.
- Metadáta upresňujú opis objektu alebo umožňujú jeho doplňujúcu charakteristiku.
- Relácie ak existujú, môžu definovať prepojenia na iné objekty.
- Z hľadiska analytických nástrojov takto pripravené vstupné údaje sú vhodné na analýzu
- Navrhovaný postup vychádza a dopĺňa štandard OPC UA ISA-95 pre Job control (*OPC UA for ISA-95 Part 4: Job Control*).
- ERP a dokument manažment systémy, sa snažia pokryť zber a ukladanie údajov s použitím a zozbieraním opisu dokumentov na základe metadát. Rozšírením o aplikácie, ktoré sú schopné

transformovať dáta do objektov, doplnený o opisné metadáta a prípadne vytvorenie ich vzťahov a nadväzností ale predstavujú vyššiu úroveň integrácie dát, z ktorých použitie údajov pre analýzu je ľahšie a efektívnejšie. Nevýhodu môžu predstavovať náklady spojené z implementáciou systému. Existujú viacerý výrobcovia a systémy, ktoré zabezpečujú podobnú funkcionalitu. Schematické znázornenie je na obrázku číslo 2. Jedná sa o systémy ako napríklad Confluence, Nuclino, Obsidian a ďalšie.



Obrázok 2. Návrh transformácie dokumentácie na strojovo čitateľné dáta

Na základe integrácie a použitia protokolov a komunikačných prostriedkov z teoretickej časti je možné zostaviť všeobecný ETL proces. Zdroje dát môžu byť rozdelené do dvoch základných kategórií:

- Relačné a nerelačné databázy
- Štruktúrované súbory JSON, XML, YAML, CSV
- Aplikačné programové rozhrania – API, napr: MQTT, REST FULL, XML RPC a ďalšie
- Data streamy ktoré sú generované IoT zariadeniami v podobe JSON dokumentov alebo inak štruktúrovaných dátových objektov

4 Návrh ETL procesu

Cieľom ETL procesu je pripraviť dáta v čitateľnej forme na analýzu. Ako vhodný formát pre prípravu dát v dávkach, je možné nahrať a pripraviť dáta do databázy. Pritom záleží od použitej technológie a možnosti licencií. Je možné použiť komerčné riešenie alebo open source systémy. Splnené ale musia byť nasledovné podmienky použitia:

- Báza dát by mala podporovať zvolený formát ukladania dát

- Prepojiteľnosť pre analytickú časť

- Podpora pre analytickú časť v zmysle meta dát

- Podpora paralelne spracovateľných úloh

Podpora zvoleného formátu ukladania dát je kľúčová. Vzhľadom na svoje vlastnosti je pre tieto účely vhodný JSON formát. Je ľahko čitateľný strojovo ale aj pre ľudí. Zároveň má schopnosť obsiahnuť zložitejšie štruktúry. Skladá za z dvojíc názov/hodnota. Dokáže reprezentovať objekt, záznam, štruktúru, slovník, hašovaciu tabuľku, zoznam alebo asociatívne pole. Vo väčšine jazykov sa reprezentuje ako objekt, pole, vektor, zoznam alebo postupnosť. JSON formát je veľmi často požívaný v Big Data.

ETL proces má za úlohu pripraviť dáta. Treba zabezpečiť funkcionality pre dva typy zdrojov údajov. Dáta pre dávkové spracovanie a dáta pre real-time spracovanie, ktoré sú k dispozícii v podobe toku dát.

ETL proces musí zabezpečiť základné úlohy:

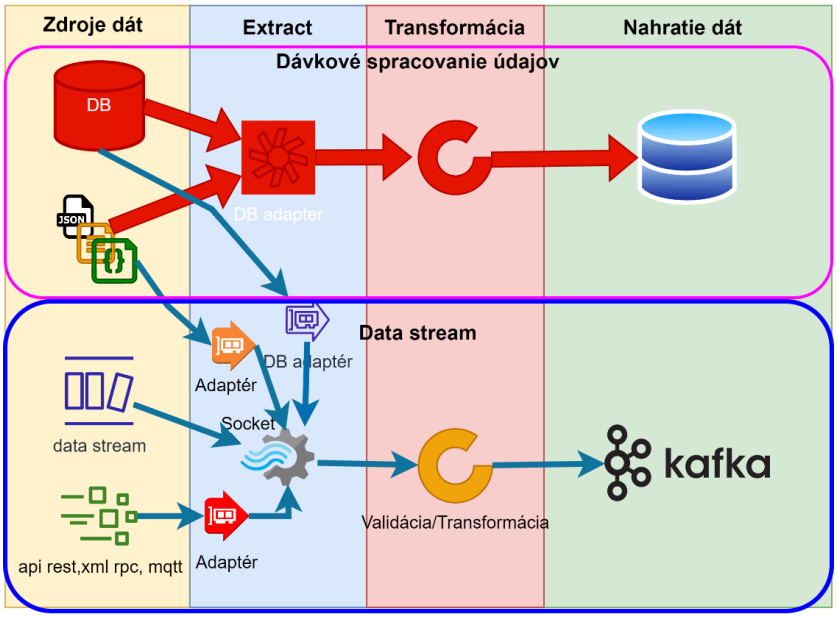
1. konzistentné dáta a ich dostupnosť
2. odstraňovanie chýb a doplňovanie chybných údajov
3. pri transformácií vznik nových metadát
4. ak je možné kvalitatívne označiť dôveryhodnosť dát
5. adaptácia a formalizácia rôznych formátovaných dát

Z týchto požiadaviek je možné navrhnúť model pre ETL proces, zabezpečujúci heterogénne zdroje dát. Bloková schéma ETL procesu ja zobrazená na obrázku 3. Návrh zabezpečuje univerzálny ETL proces pre dávkové spracovanie a aj pre dáta, ktoré je možné spracovávať v reálnom čase.

5 Návrh ETL procesu z bázy dát a dátových tokov

Pre dávkové spracovanie sa jedná o načítanie údajov z rôznych DB systémov alebo súborov. Táto fáza je kritická z pohľadu zdrojového systému, nakoľko načítavanie dát, môže ovplyvniť svojou aktivitou zdrojový systém, spomaliť ho alebo úplne znefunkčniť. Preto je vhodné precízne naplánovať a otestovať čas potrebný na extrakciu a overiť vplyv na zdrojový systém. Môže ísť o inkrementálny prístup alebo plné čítanie údajov. V praxi je zavedený postup inicializačného nahrania, a postupné inkrementálne spracovanie nových údajov.

Data stream extrakcia je otázka spojená s prepojiteľnosťou alebo schopnosťou konzumovať dáta z existujúcich API tak, aby volania a dáta boli dostupné pre pôvodné určenie a zároveň aby dáta z týchto volaní, správ alebo dátových tokov boli dostupné aj pre proces získavania dát. Pre API, ktoré používajú protokoly spomenuté v teoretickej časti, je možné obvykle implementovať adaptér alebo využiť techniky subscribe pre pripojenie sa na zvolené typy správ. Ideálne sú systémy založené na MQTT, Kafka, OPC UA, ktorých architektúra tento typ získavania dát umožňuje. Pre systémy zbernicového typu, je tak isto možné počúvať a odchytávať dáta, môže to ale ovplyvniť výkon systému, príklad môže byť TIBCO Rendezvous® zbernica od firmy TIBCO. Na tieto účely je na obrázku číslo 5. znázornený prvok Adaptér, ktorý podľa typu protokolu umožňuje zberať dáta a privádzať ich na vstup dátového toku.



Obrázok 3. Návrh ETL procesu vstupy definované pre dávky ako aj pre dátové toky

6 Identifikácia procesov integrácie a ich vyhodnotenie z technologického pohľadu.

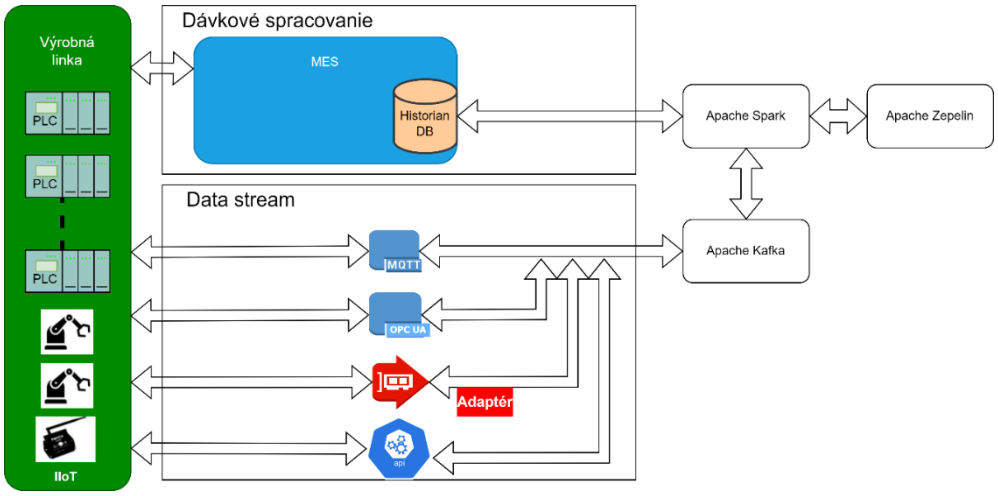
Poznanie výrobného procesu, riadenie a zostavenie prehľadu a závislostí je možné považovať za jednu z úloh systémovej a dátovej integrácie. Výrobné systémy riadené MES alebo MOM systémom, majú implementáciu podpory riadenia výroby. Každý element výrobnej linky, ktorý sa podieľa na riadení ovplyvňuje stroje, ich pohyby a riadi vykonávanie úloh podľa zostaveného výrobného procesu a programu riadenia. Zdroj údajov ako prebieha výroba je možné klasifikovať do systémov, ktorý sú riadené bez prvkov Industry 4.0 najčastejšie PLC a riadiacimi prvkami, ktoré sú schopné a podporujú Industry 4.0 štandardy na rôznych úrovniach. Zostava riadenia výroby pomocou PLC je obvykle vystavaná na naprogramovaní PLC, robotov na jednotlivé úkony. Tieto sú integrované do skupín, ktoré vykonávajú určitú časť výroby. Zostavenie krokov, ktoré sú potrebné pre dosiahnutie produktu alebo vykonania čiastkovej úlohy vychádza zo zadania výroby a je explicitne definované a PLC sú naprogramované na konkrétne typy riadiacich aktivít. Reprezentácia činností je buď úplne skrytá alebo je možné z PLC systémov načítavať dáta, ktoré predstavujú zmeny stavov podľa vykonávanej úlohy a tento typ získavania údajov je realizovaný na základe dostupnej dokumentácie. Zostavenie výrobného procesu takýmto spôsobom je pomerne zložité. Na druhej strane po nasadení komponentov Industry 4.0, sa situácia v tomto smere veľmi zjednodušuje. Hlavným rozdielom je možnosť a dostupnosť exportu riadiacich prvkov a ich fungovanie pomocou finálneho stavového automatu alebo schém. Jedným z príkladov je OPC UA architektúra, kde je dostupný model stavového automatu v strojovo čitateľnej forme, ktorý opisuje logiku a fungovanie riadiaceho prvku. Na základe týchto modelov je možné identifikovať nasledujúce parametre:

- Dáta, ktoré sú spracovávané riadiacim systémom
- Udalosti generované počas riadenia
- Vnútorné stavy riadiaceho systému
- Stavy vstupov a výstupov v ľubovoľnom čase
- Časový priebeh udalostí a k nim prislúchajúce stavy

7 Návrh platformy

Návrh architektúry pre získavanie dát vychádza z teoretickej časti, kde boli predstavené dva nástroje, ktoré sú vhodné na plnenie úloh. Apache Spark a Flink. Vzhľadom na povahu dát a identifikáciu systémov je výhodnejšie nasadenie Apache Spark, ktorý podporuje dávkové spracovanie údajov, spracovanie dátových tokov a aj ich kombináciu. Pre čisto prostredie s technológiou Industry 4.0 by bolo možné použiť aj Apache Flink. Tento ale nemá modul podpory pre strojové učenie, v čom je potenciálna nevýhoda a preto výber padol na systém Apache Spark.

Bloková schéma zapojenia je zobrazená na obrázku číslo 4. Model výrobnú linky obsahuje všetky prvky riadenia. MES systém a jeho časť Historian boli zdroje dát pre dávkové spracovanie. MQTT broker, OPC UA server a klient a Adaptér spoločne s API boli implementované ako samostatné časti. MQTT broker bol použitý pre zber údajov aj do MES systému. VerneMQ bol nakonfigurovaný pre získavanie dát z riadiacich členov výrobnej linky vrátane PLC S300. OPC UA server bol vytvorený a implementovaný pre doplňujúce IIoT zariadenia, ktoré boli zahrnuté do navrhovaného modelu, aby bolo možné overiť prvky Industry 4.0. API boli k dispozícií RESTULL API a ďalšie ktoré integrovali doplňujúce snímače.

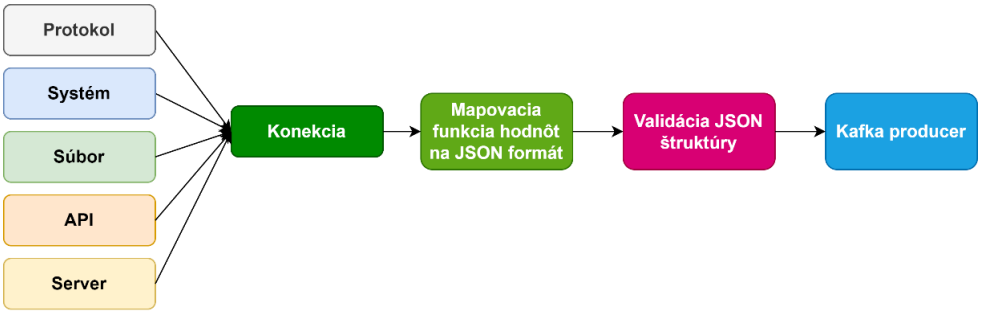


Obrázok 4. Bloková schéma zapojenia testovacej výrobnej linky

Ako vyplýva z blokovej schémy, pre niektoré komponenty je potrebné zabezpečiť prepojiteľnosť na systém Kafka. Niektoré systémy ako MQTT a OPC UA majú konektory dostupné, ale je možné, že je potrebné pripojiť systém, ktorý podporovaný nie je. Dostupné sú komerčné riešenia napríklad od firmy SentinelOne alebo od poskytovateľa cloud služieb, alebo je možné navrhnúť riešenia, ktoré sa nazývajú aj Adaptér.

7.1 Návrh adaptéra

Všeobecná schéma pri vytvorenie adaptéra je zobrazená na obrázku číslo 5.

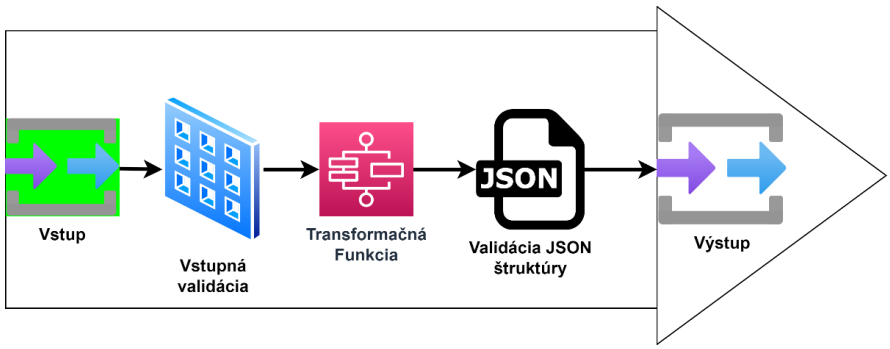


Obrázok 5. Model vytvorenia adaptéra pre rôzne zdroje

Vstup môže predstavovať systém, API, Sever alebo zariadenie. Záleží od integrovanej technológie, ktorá bude zdrojom dát a má byť integrovaná do analytického systému. Veľmi jednoduchým riešením je použiť data flow systém ako je napr. Node-Red alebo Apache Nifi.

7.2 Návrh Half Duplex Adaptéra

Adaptéry môžu byť half duplex alebo full duplex. Half duplex je zobrazený na obrázku číslo 6. prúdi len jedným smerom. Zo vstupu na výstup. Full duplex predstavuje systém dvoch adaptérov.

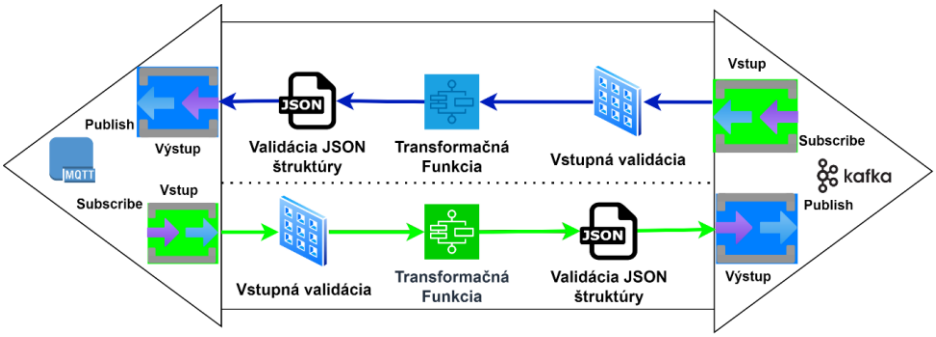


Obrázok 6. Návrh funkcionality pre half duplex adaptér

7.3 Návrh Full Duplex Adaptéra

Full duplex adaptér je vhodný pre nasadenie pre protokoly, ktoré sú typu publish/subscribe, alebo majú povinnú obojsmernú komunikáciu ako je napríklad REST Full API, MQTT alebo OPC UA. Bloková schéma je zobrazená na obrázku číslo 7.

V prípade zbernicového systému je tiež možné implementovať full duplex komunikáciu. Záleží od typu zbernice a požiadavky pre nasadenie prenosu dát v reálnom čase. V tomto prípade je potrebné definovať sieťovú vrstvu, ktorá zabezpečí potrebnú latenciu spracovania dát. Tu je vhodné zvoliť Ethernet systémy pre prenos dát v reálnom čase ako napríklad: **gigabit Ethernet bandwidth with Time Sensitive Networking (TSN), ktorý je priemysle často nasadzovaný v prevádzkach, kde je potrebná komunikácia v reálnom čase.** Tak isto implementácia adaptéra, musí splniť požiadavku pre časovo definovaný zbernicový systém, ak je vyžadovaný. Cieľom je minimalizácia času, ktorý potrebuje adaptér na transformáciu údajov.



Obrázok 6. Návrh funkcionality pre full duplex adaptér.

Vstup zabezpečuje prepojiteľnosť. Nasleduje validácia dát, ktoré sú privádzané na vstup. Nasleduje ich transformácia a vytvorenie JSON štruktúry. Vytvorená JSON štruktúra pre overenie správnosti je validovaná a odosielaná na výstup. Výstup zabezpečuje prepojiteľnosť na zamýšľaný systém. V tomto návrhu je použitý systém Kafka.

Adaptér pre nasadenie v reálnom čase alebo pre integráciu zbernicového systému je potrené navrhnúť tak, aby vyhovoval nasledovným parametrom.

- **Včasnosť** - adaptér musí byť schopný správne fungovať v dátovej a časovej oblasti nakoľko riadenie úloh s explicitnými časovými obmedzeniami môžu mať kritické nároky na tento parameter.
- **Predvídateľnosť** - všetky požiadavky na časovanie v úlohách v reálnom čase by mali byť zabezpečené tak, aby systém bol schopný spracovať a doručiť naspäť dáta načas. Platí tu opäť pravidlo, kde spracovanie dát a doručenie musí pracovať na dvojnásobnej frekvencii ako je rýchlosť očakávaných dát v rámci špecifikovaných časových obmedzení.
- **Efektívnosť** - väčšina systémov reálneho času sú systémy s nízkym množstvom zdrojov a sú orientované na dosiahnutie požadovaného výkonu.

- **Robustnosť** - adaptér musí byť schopný obslúžiť systémy počas zaťaženia
- **Modulárnosť** - pri navrhovaní adaptéru pre systém reálneho času by sa mala použiť modulárna štruktúra, aby bola zaručená jednoduchá implementácia a aj aktualizácia adaptéru.
- **Odolnosť voči poruchám** - časovo kritické adaptéry musia byť implementované tak, aby boli odolné voči poruchám. Systém by mal byť schopný behu aj keď vypadne jedna inštancia adaptéru, či už v dôsledku hardvérovej alebo softvérovej poruchy.

7.4 Doručovanie údajov z adaptérov pomocou Kafka systému so zameraním na udalosti a platforma na spracovanie streamov

V návrhu je výstup realizovaných adaptérov zabezpečený pomocou Kafka systému. Posledná časť v realizácií adaptéra je Kafka producer, ktorý je schopný vygenerované správy odosielať ako definované a validované JSON štruktúry. Pre správne fungovanie je potrebné nastaviť komunikačné výstupné parametre, ktoré zabezpečia prenos do systému Kafka. Kafka producer v tomto prípade posiela data stream do Kafka klasteru. Možné je definovať dva základné typy producera, podľa typu realizovanej komunikácie.

- Synchrónny
- Asynchrónny

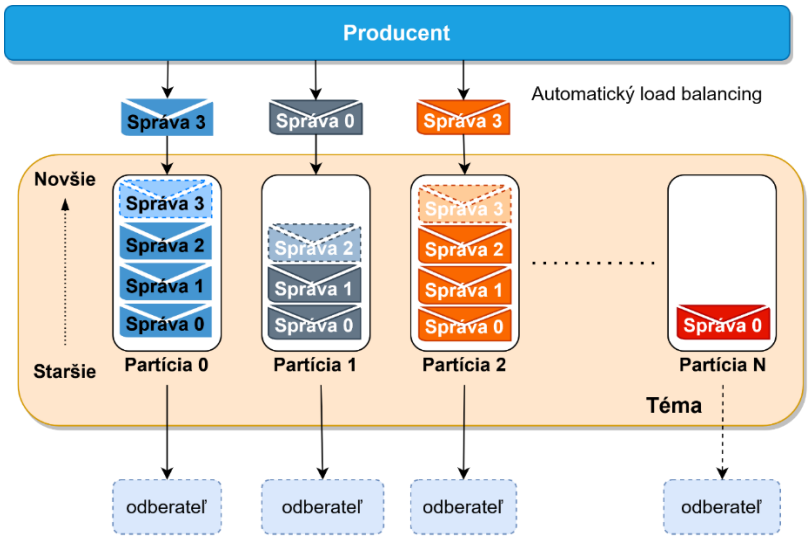
Synchrónny typ producera posiela správy priamo a je kontrolovaná ich konzumácia a spracovávaná odpoveď. Pre tento typ je nutné použiť full duplex adaptér.

Asynchrónny prenos je odporúčaný, a je schopný zabezpečiť väčšiu priepustnosť systému. Je vhodný na väčšiu časť zabezpečovanej komunikácie. V tomto prípade je možné použiť full duplex alebo dva half duplex adaptéry, kde pre každý smer komunikácie je možné implementovať jeden. Výhodnejšie z pohľadu obsluhy sú dva typy half duplex adaptéru, kde je možná identifikácia toku dát, ako aj možnosť lepšieho monitorovania systému.

Asynchrónna komunikácia ale nie vždy vyhovuje a pre dáta, ktoré sú kritické pre riadiace systémy, kde je kladený dôraz na vyhodnotenie a musí byť zaručené spracovanie údajov je vhodnejší synchrónny prenos, ako je zobrazené na obrázku číslo 8.

Správy, ktoré sú vytvárané Kafka producentom musia obsahovať minimálne základné parametre: topic, partition,, replication factor, key1, value1, callback.

- Topic - predstavuje tému, ktorú je možné vytvoriť pri odoslaní do systému.
- Parition – je partícia, do ktorej je možné tému zaradiť a dostať tak lepšiu priepustnosť systému.
- Replikačný faktor - určí koľko násobne je daná správa posielaná pre prípad, že jedna inštancia dát spracovania zlyhá. Druhá inštancia zabezpečí včastné dokončenie a doručiť dát na čas.
- Ostatné parametre ako key, value sú určené na samotný prenos dát. Štandardný používaný dátová Štruktúra je JSON.
- Callback – spätné volanie slúži na implementáciu rozhrania, ktoré môže používateľ implementovať, aby bolo umožnené vykonanie kódu po dokončení požiadavky. Je to spätné volanie a spravidla sa vykoná vo vlákne na pozadí a umožní tak rýchlejšiu notifikáciu a dokončenie úlohy.



Obrázok 8. Schéma komunikácie pomocou Kafka

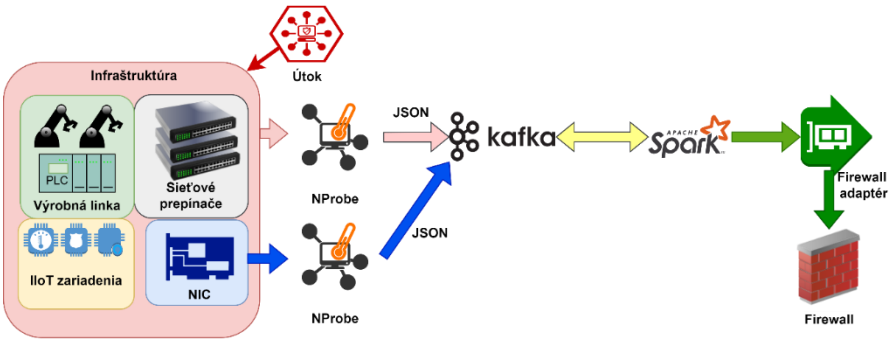
7.5 Návrh zberu dát pre zvýšenie bezpečnosti infraštruktúry a využitie objavovania znalostí o sieťovej prevádzky výrobných zariadení.

Špecifická oblasť pri výrobných technológiách je zabezpečenie IT infraštruktúry. Sieťová prevádzka, ktorá zabezpečuje komunikáciu medzi riadiacimi prvkami výroby produkuje obrovské množstvo dát. Základné monitorovanie tejto prevádzky dokáže pokryť spoľahlivostné aspekty, ako je identifikácia výpadkov, poruchy prepínačov alebo výpadky v sieti a ďalšie. Problematická je obvykle identifikácia bezpečnosti. Špeciálne pri integrácií a nasadení technológií, ktoré komunikujú mimo sieť, ktorá zabezpečuje výrobu. Nasadzovaním IIoT zariadení, využívanie cloud technológií, predstavujú potenciálne bezpečnostné riziko, ktoré je potrebné analyzovať. Z pohľadu ETL procesu

je možné zberať sieťové dáta o prevádzke a komunikácií zariadení. Je možné využiť štandardy sFlow a NetFlow, ktoré sú schopné zaznamenávať spôsob komunikácie. Sieťové dáta sú posielané v obrovských objemoch. Analytické nástroje Big Data sú vhodné na vyhodnotenie potenciálneho bezpečnostného rizika. Z pohľadu ETL procesu je to odlišnosť od riadiacich prvkov a to v povahe generovaných dát. Stroje a riadiace prvky v priemysle generujú udalosti a dáta, ktoré súvisia s výrobným procesom. Sieťové dáta opisujú komunikáciu a ich zaznamenávanie musí podporované komponentami infraštruktúry.

Je možné nasadiť adaptér, tak ako bolo navrhované pre priemyselné dáta. Zo sieťového pohľadu je ale jednoduchšie využiť NetFlow protokol, ktorý je súčasťou prepínačov, brán je možné ľahko zberať dáta ich presmerovaním na komponent, ktorý sa postará o ich transformáciu na formát JSON. Tento komponent sa nazýva NProbe. Je možné ho nainštalovať na rôzne platformy a často je súčasťou sieťových aktívnych komponentov.

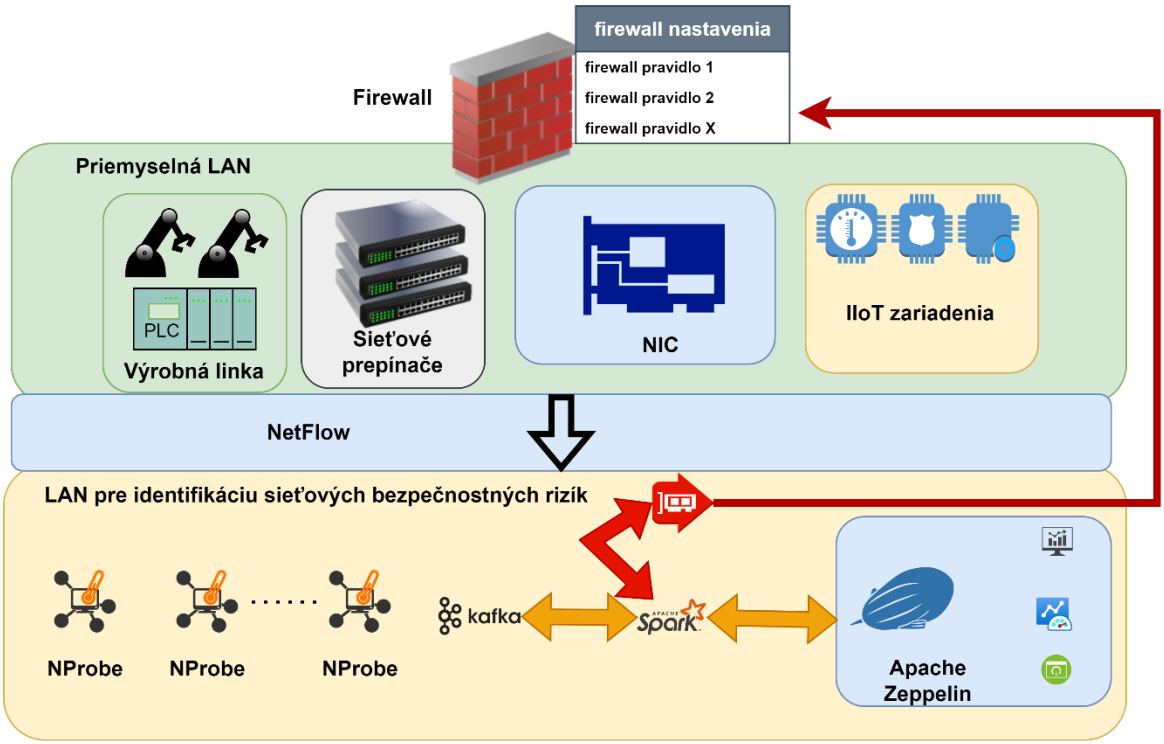
Pomocou NProbe je možné sledovať sieťovú prevádzku až na siedmej, to je aplikačnej vrstve ISO/OSI modelu. Je podporovaných viac ako 250 aplikačných protokolov. Podpora nižších vrstiev je platná až po MAC vrstvu. Realizácia ETL procesu je možná na základe spracovania údajov, kde je možné použiť štandard NetFlow verzie 5 až 9 ale aj IPFIX. Podpora zahŕňa implementáciu Cisco NetFlow-Lite. Aplikácia ETL procesu pomocou nProbe a protokolu NetFlow prístupu je zobrazená na obrázku číslo 9.



Obrázok 9. ETL proces a návrh analýzy údajov pre priemyselný infraštruktúru.

Dáta, ktoré sú posielané medzi komponentmi výrobnej linky, IIoT zariadeniami, sieťovými kartami NIC alebo aktívnymi sieťovými komponentami ako sú napríklad prepínače, sú zberané pomocou zaznamenávania alebo posielania kópie prenášaných dát na zariadenia s nainštalovaným programom nProbe. nProbe prijíma a spracovávajú informácie a transformujú do JSON štruktúry. Nevýhodou je vysoký dátový tok. Pre tento účel je vhodné oddelenie sietí pre výrobné dáta a pre

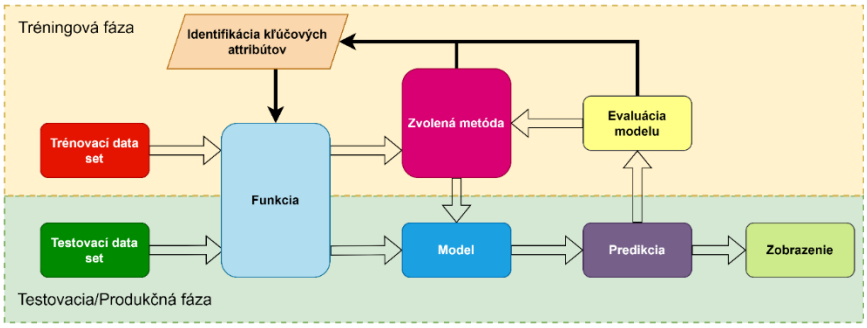
analytické spracovanie údajov zo sieťovej infraštruktúry. Rozdiel pre sieťové dáta oproti dátam z riadiacim komponentom, sú komponenty pre zber údajov umiestnené mimo priemyselnej sieťovej infraštruktúre, ako je zobrazené na obrázku číslo 10.



Obrázok 10. Rozdelenie komunikácie do dvoch nezávislých LAN

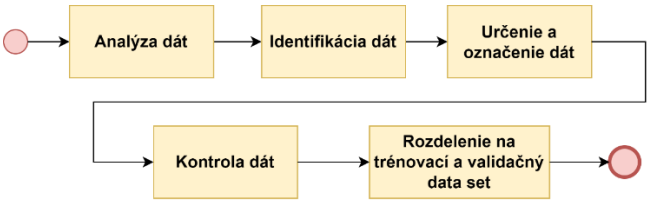
8 Návrh systému na získavanie znalostí a príprava testovacieho a validačného datasetu

Spracovanie dát pre klasický data mining je pokladaný za dávkové spracovanie údajov. Dávkové spracovanie údajov je základom aj pre spracovanie tokov dát alebo spracovanie dát v reálnom čase. Oba typy majú spoločný model spracovania údajov. Pre správnu klasifikáciu alebo regresiu je potrebné vybrať a identifikovať funkciu a atribúty, ktoré majú významný vplyv na vyhľadávaný vzor. Príprava modelu pre objavovanie dát oboch typov je znázornená na obrázku číslo 11.



Obrázok 111. Príprava modelu pre objavovanie znalostí.

Dáta prechádzajú prípravným procesom. Sú klasifikované tak, aby bolo možné ich vyhodnotenie a bolo možné natrénovať model. Príprava testovacích dát je zobrazená na obrázku číslo 12.

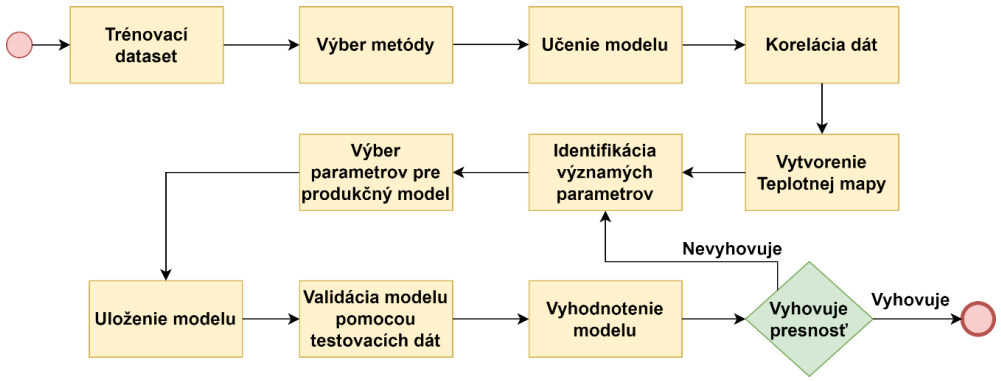


Obrázok 12. BPMN diagram pre prípravu trénovacej a validačnej sady dát.

Analýza dát je základným krokom, na základe ktorého je možné vytvoriť trénovací a validačný dátový set pre prípravu modelu. Nasleduje identifikácia dát, kde sa určujú dáta, ktoré môžu ovplyvniť správanie sa modelu. Nasleduje označenie dát v zmysle, ktoré dáta sú vyhovujúce a ktoré sú chybné z hľadiska hľadanej informácie. Nasleduje kontrola dát. V ideálnom prípade, je použiť dáta, ktoré boli dlhodobo zaznamenávané a predstavujú najrôznejšie existujúce situácie. Ak je k dispozícii málo vzoriek, je vhodné niektoré duplikovať a odvodiť nové hodnoty, tak aby boli blízke reálnemu príkladu. Rizikom tohto úkonu je ale potenciálna deformácia a väčšia nepresnosť pripravovaného modelu.

8.1 Návrh vytvárania modelu, validácia a vyhodnotenie presnosti modelu.

Klasifikované dáta sa náhodne rozdelia do dvoch skupín. Náhodné rozdelenie nie je celkom náhodné a je vhodné štatisticky zahrnúť všetky situácie, ktoré opisujú známe správanie sa systému. Trénovací dataset je obvykle 70 až 80 percent pripravených dát. Validačná množina predstavuje zvyšok. Vytvorenie modelu je zobrazené na obrázku 13.



Obrázok 13. BPMN diagram pre vytvorenie modelu na základe data setu

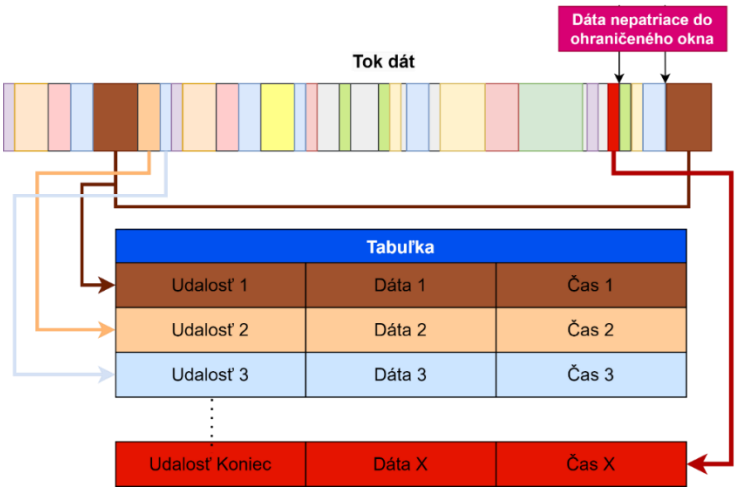
Vytvorenie modelu predstavuje veľmi častú činnosť. Pre predstavovaný návrh, po príprave trénovacieho datasetu je možné pristúpiť k vytvoreniu modelu. Správne trénovanie modelu je založené na dobre pripravenom dáta sete a vhodnom výberu metódy. Je možné otestovať viaceré metódy a porovnať ich efektivitu pre požadovanú úlohu. Učenie modelu zabezpečí základný prehľad a je možné identifikovať koreláciu dát. Koreláciou dát je možné zistiť, ktoré atribúty sú vhodný na určenie hľadaných vlastností a majú významný vplyv. Veľmi často je reprezentovaná teplotnou mapou, ktorú je vhodné vytvoriť. Určenie dát, ktoré majú významný vplyv sa dá vyjadriť aj tabuľkou s prehľadom percentuálneho vplyvu na pripravovaný model. Podľa vplyvu na výsledok po nahratí trénovanej množiny je možné vyhodnotiť limitovaný počet atribútov tak, aby model bol schopný pomocou ich hodnôt čo najpresnejšie vytvoriť predikciu na ich základe. Nasleduje uloženie modelu a jeho validácia pomocou validačného datasetu. Vyhodnotí sa presnosť modelu. Ak presnosť vyhovuje zadanému typu úlohy, je uložený a je možné ho použiť na nasadenie v produkcii. Model, ktorý dokáže určiť dáta na viac ako 95 percent je obvykle pokladaný za úspešný, ale vždy závisí od požiadaviek ktoré sú kladené na riadenie, aby nedochádzalo k nesprávnemu chovaniu systému. Takto je možné otestovať viaceré metódy strojového učenia a porovnať, ktorá metóda je najlepšia a dáva najpresnejšie výsledky. Výsledkom je model a pre ďalšie skúmanie je vhodné ukladať aj získaný dataset, ktorý môže poslúžiť na neskoršie vylepšovanie alebo porovnanie nových modelov.

9 Návrh spracovania dát v reálnom čase

Pre spracovanie dáv v reálnom čase, je možné definovať spracovanie dát ako mikrodávku. Mikrodávka umožňuje opakované spracovanie úloh a použitie výsledkov z predchádzajúcej mikrodávky. Od začiatku po koniec je možné online spracovanie dát až po ukončenie v mikrodávkach. Tento proces sa javí ako spracovanie dát v reálnom čase. Takto spracované dáta umožňujú medzi-výsledkový manažment. Už po spracovaní obmedzenej časti dát, je možné dostať výsledok alebo výsledky, ktoré reprezentujú hľadaný vzor alebo dáta. Na ich základe je možné aplikovať napríklad rozhodovanie alebo použiť dáta na rozhodovanie vo výrobnom procese.

Pre celkový prehľad je potom možná aplikácia výsledkov do predchádzajúcich dávkových výsledkov alebo je možné vytvárať celkový pohľad, ktorý obsahuje všetky doteraz spracované dáta.

Aby bolo možné jednoducho pracovať s dátami, je nutné definovať čas udalosti tak, že signál udalosti je transformovaný do podoby, ktorá reprezentuje dáta a udalosti ako riadky v tabuľke. Prenášané dáta a čas udalosti je hodnota stĺpca v riadku, ako je zobrazené na obrázku 14. Dáta je možné použiť na spracovanie a aplikovať transformačné funkcie na úrovni analytického systému, ktoré sú špecializované na operácie so stĺpcami. Tieto operácie sú nenáročné na výkon a preto je vhodné ich používať.



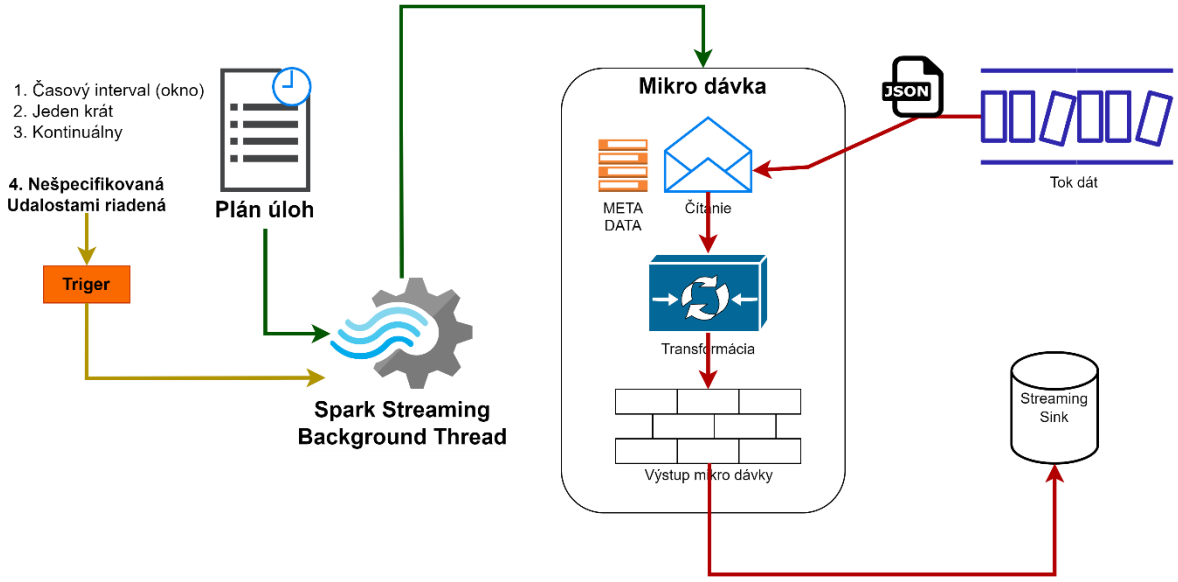
Obrázok 14. Transformácia dát na tabuľkovú reprezentáciu a vznik štruktúrovaného dátového toku.

Agregácia na stĺpci podľa času udalosti definuje nové časové okno. Každý riadok môže patriť do viacerých okien podľa toho, ako sú dáta ohraničené udalosťami. Jedny dáta môžu patriť do viacerých okien určených na analýzu. Takto je možné získať a definovať dotazy založené na udalosťami definovanom okne alebo aj časovo definovanom okne. Ak je možné ohraničiť dáta na základe

udalostí, je možné aplikovať a definovať akékoľvek metódy alebo machine learning metódy tak, ako keď sa spracovávajú statické dáta alebo dáta z dávkového súboru. Ak je uplatnené zovšeobecnenie, tak, je možné nahliadať na takto pripravené a ohraničené dáta, ako na dáta, ktoré sú reprezentované v klasickom relačnom modeli, kde sú dáta uložené v báze dát. Tento princíp je z používateľovho hľadiska spracovania dát výhodný, lebo uľahčuje pochopenie systému a umožňuje návrh realizácie, implementáciu a aplikáciu analytických metód. Výsledok tejto operácie je štruktúrovaný dátový tok a ohraničenie okna pre analýzu dát štruktúrovaného dátového toku, na ktorý sa dá nahliadať ako na tabuľku dát.

Špeciálny prípad sú dáta, ktoré sú medzi začiatkom a koncom a nepatria do vyhradeného okna, ako je znázornené na obr. 15. Tieto dáta môžu patriť do iného definovaného okna, ktoré môže obsahovať väčší objem udalostí a dát alebo ak je nezaručené doručenie, je možné zaradenie do predchádzajúceho alebo aj nasledujúceho okna a použiť tak systém watermark, a prepočítať výsledky mikrodávky.

Schéma spracovania údajov pomocou systému Apache Spark pre štruktúrované dátové toky je znázornená na obrázku 15.



Obrázok 15. Spracovanie dát v reálnom čase pomocou mikrodávok.

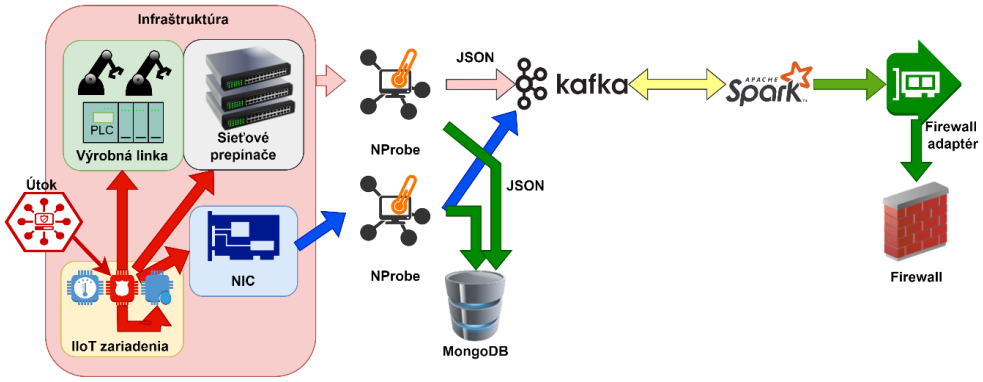
Dátový tok alebo Data Stream je privádzaný na sieťový port – socket. Tento je schopný prijímať dáta. Podľa operačného systému je vhodné skontrolovať systémové a sieťové nadstavenia, ktoré môžu limitovať priepustnosť systému. Je potrebné alokovať pamäť pre vstupný buffer a nastaviť správanie systému pre počet možných otvorených súborov.

Privádzané dáta sú v podobe toku dát. Ak je dátový tok ohraničený, je potrebné zadefinovať stav, a koniec dátového toku, inak dáta môžu byť prijímané do nekonečna, čo v praxi nie je reálne možné. Ak je nemožné identifikovať a zadefinovať stop signál, je vždy možné na základe podnetu z programu vykonať ukončenie procesu a jeho nové vytvorenie. Strata údajov je minimalizovaná tým, že na doručenie správ je použitý systém Kafka. Výnimka, kedy dochádza k prerušeniu prijímania toku dát je vedome nadstavená a je potrebné sledovať vstupné dáta a monitorovať použitie dostupných prostriedkov a pamäte, aby nebol ohrozený beh samotného systému.

Časť transformácie bola už vykonaná v príprave dát do formátu JSON. Ďalší krok predstavuje čítanie a transformácia údajov v rámci mikrodávky. Tu sa nahrádzajú nulové hodnoty, tak aby bolo možné vyhodnocovať dáta. Povolené sú aj všetky aritmetické operácie a operácie, ktoré sú vykonávané nad stĺpcami mikrodávky. Tieto operácie sú špecificky výhodné z pohľadu náročnosti na výkonu systému, pretože stoja málo systémových zdrojov a nie je ich možné aplikovať počas prípravy dát v predchádzajúcom ETL procese.

10 Experimentálne overenie návrhu získavania znalostí zo štruktúrovaných dát

Overenie návrhovej časti bolo testované viacerými spôsobmi. Výber padol na detegovanie útokov na výrobnú linku. Je vhodný pre vygenerované dostatočné množstvo dát a demonštráciu všetkých navrhovaných častí. Pripravený scenár bol založený na distribuovanom útoku DDoS za pomoci nezabezpečeného IIoT zariadenia. Tento dokázal šíriť falošné pakety do infraštruktúry výrobnej linky. Následok útoku bolo možné pozorovať náhodnými chybami vo výrobe alebo vypadávaním komunikácie komponentov výrobnej linky. Pozorované boli nedoručené obrazové informácie z kamier, neschopnosť robota dokončiť operáciu alebo nekorektná detekcia prechodu tovaru cez detekčné body, ktoré boli čítané pomoc RFID čítačiek. Generované sieťové pakety útoku boli zberané a zachytávané aj s bežnou bežnou prevádzkou. Príprava bola založená na analýze dát. Následne boli aplikované kroky potrebné pre klasifikáciu a vyhodnotenie datasetu. Dáta sú zaznamenávané pomocou JSON štruktúry. Príprava datasetu bola realizovaná pomocou noSQL databázy MongoDB, kde celá dátová komunikácia bola zberaná pomocou systému NProbe. Bloková schéma je na obrázku 16.



Obrázok 16. Bloková schéma experimentu pre overenie navrhovanej metódy

Dátové pakety, zachytávané na sieťových switchoch, boli zberané pomocou zvlášť vytvorenej infraštruktúry, aby bol zabezpečená nezmenená komunikácia výrobnej linky. Pakety, boli za pomoci NetFlow protokolu zozbierané a ukladané do JSON štruktúr. Ukážka jedného dokumentu, podľa návrhovej časti je zobrazená na obrázku 17.

```
[
 {
   "SOURCE": "FLOW",
   "AGENT_ADRESS": "192.168.0.1",
   "INPUTPORT": 2,
   "OUTPUTPORT": 0,
   "CREATETIME": "27.4.2025 9:00:01 332",
   "SRC_MAC": "001b1ba3afbf",
   "DST_MAC": "01005e010101",
   "ETHERNET_TYPE": 800,
   "SRC_IP": "192.168.10.253",
   "DST_IP": "231.1.1.1",
   "IP_PROTOCOL": 17,
   "IP_TOS": 0,
   "IP_TTL": 1,
   "SRC_PORT_OR_ICMP_TYPE": 4445,
   "DST_PORT_ICMP_CODE": 4446,
   "TCP_FLAGS": 0,
   "PACKET_SIZE": 95,
   "IP_SIZE": 73,
   "STATUS": 1,
   "QUALITY": 1
 }
]
```


Obrázok 17. JSON štruktúra opisujúca komunikáciu dvoch riadiacich prvkov zachytená pomocou nProble na základe protokolu NetFLOW

Okrem sieťových parametrov, ako sú verzia IP protokolu, zdrojové a cieľové adresy boli vytvorené doplňujúce parametre. Kľúčový identifikátor je CREATETIME, ktorý predstavuje časový údaj vzniku údajov. Stĺpec STATUS, ktorý v prípade korektnej komunikácie predstavuje hodnotu 1 a pre reprezentáciu útokov nadobúda hodnotu 2. Parameter STATUS je potrebný pre definovanie trénovacieho a validačného datasetu. Označovanie STATUS je možné vykonať počas zberu údajov alebo potom hromadne. V tomto prípade bol identifikátor zdrojová adresa posielania paketov a typ komunikácie, ktoré jednoznačne identifikovali problémovú komunikáciu.

Posledný parameter QUALITY je indikátorom dôveryhodnosti dát a jeho potenciálneho vplyvu. Opäť je možné ho vytvoriť pri zbere údajov. Kvalita by bola znížená len pri opakovaných prenosoch. Tieto operácie sú pomerne náročné na výkon sieťového prepínača.

Model bol vytváraný pomocou opísaného postupu na základe pripravených trénovacúch dát. Dáta boli ukladané v JSON formáte a bolo použité dávkové spracovanie údajov pre vytvorenie modelu. Výber funkcie padol na logistickú regresiu, ktorá je vhodná na klasifikáciu a predikciu dvoch stavov. V tomto prípade ide o klasifikáciu paketov. Úlohou modelu je predikcia na základe parametrov prebiehajúcej sieťovej komunikácie. Model má vyhodnotiť, či sa jedná o útok alebo reálny sieťovú komunikáciu výrobnej linky. Trénovanie v dávkovom spracovaní je jednoduché a model je možné

použiť pre dátové toky ale aj pre dávkovo orientované úlohy. Overenie modelu bolo realizované na validačnej vzorke a neskôr aj na nasadení v produkčnom prostredí.

Na základe zvolenej metódy je potrebné transformovať všetky stĺpce dát, ktoré je možné transformovať na číselné hodnoty. Následne je potrebné ošetriť nulové hodnoty, za pomoci stĺpcovej operácie a nahradiť ich buď veľmi malými aj desatinnými hodnotami alebo použiť NaN hodnotu, ktorá nie je nulová ale predstavuje neexistujúcu hodnotu. Každá transformácia stĺpca napríklad IP adresy, ktorá je dostupná v ľudsky čitateľnej podobe je možné transformovať na číselnú hodnotu. Model Logistic Gregresion pracuje len z číselnými hodnotami. Logistická regresia je metóda na predikciu kategorizácie. Je to špeciálny prípad zovšeobecnených lineárnych modelov, ktorý predpovedá pravdepodobnosť výsledkov. V Spark machine learning sa logistická regresia môže použiť na predikciu binárneho výsledku pomocou binomickej logistickej regresie. Aby vstupné dáta bolo možné zahrnúť do predikcie, je potrebné ďalšie stĺpce upraviť do číselnej podoby. V sieťovom dátovom toku je takmer každá hodnota definovaná ako číslo, takže táto aplikácia nepredstavuje náročnú operáciu. To isté platí od dátumovej hodnote, kde je možné mapovať timestamp na celočíselnú hodnotu. Na každú operáciu je možné napísať funkciu, ktorá vykoná transformáciu do požadovaného číselného tvaru. Napríklad IP adresu fe80:0000:0000:0000:19c7:e249:9d49:d39d je IPv6 adresa aje potrebné ju transformovať na celočíselný tvar 338288524927261089655876599018460664733. Pre transformáciu hexadecimálneho tvaru zdrojová mac adresa 0180c200000e po transformácií je číselná hodnota 1652522221582.

Pre typ Ethernet Type konverzia hexadecimálneho údaju

Ukážka konverzie funkcie pre ethernet type je zobrazená na obrázku 18.

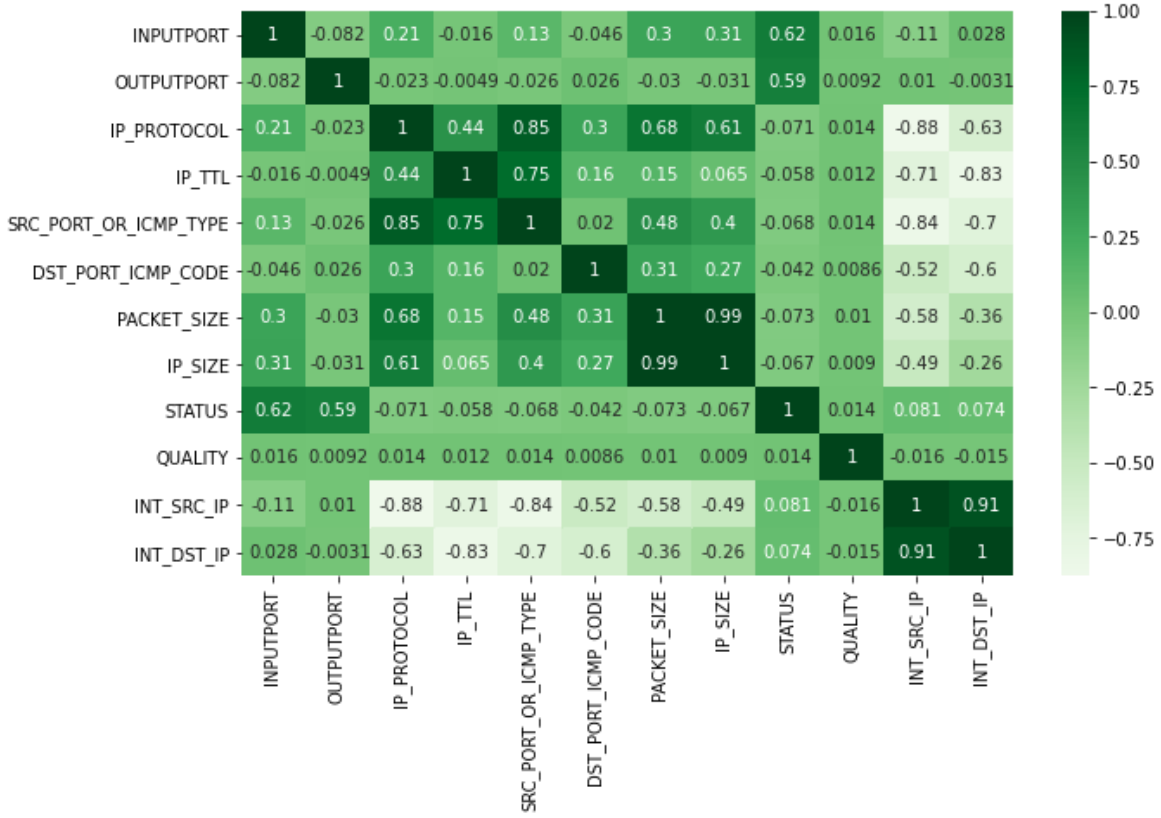
```
def conv_eth_type(ethtype):
    return int(ethtype, 0)

#Transform function to collumn udf pyspark
spark_conv_eth_type = F.udf(lambda data: conv_eth_type(data),IntegerType())
```

Obrázok 18. Funkcia konverzia Ethernet typu na číslo s úpravou konverzie pre stĺpce

Vstup tejto funkcie je hexadecimálne číslo napr. 0x0800, kde výsledok po konverzií je 2048. Celkový popis typov v sieťach je možné nájsť podľa normy ieee-802, ktorá opisuje typy Ethernetu [15]. Potom je potrebné v Sparku upraviť túto funkciu na prácu so stĺpcom pomocou metódy PySpark User Defined Function UDF. Na obrázku 18 je znázornená implementácie pomocou funkcie spark_conv_eth_type, ktorá zabezpečí prevod funkcie conv_eth_type na Spark funkciu

zabezpečujúcu hromadnú aplikáciu na stĺpce. Táto operácia ako bolo už spomínané, je oveľa lepšia z pohľadu výkonu vzhľadom na klasickú úpravu hodnôt individuálne po riadku.

Nakoľko dáta typu String alebo Byte alebo hodnoty, ktoré nie je možné vyhodnotiť, lebo nemusia predstavovať číselne hodnoty, môžu byť vynechané. Pre stĺpce, ktoré je možné transformovať je potrebné tieto operácie vykonať či už v rámci dávky alebo dátového toku, po transformácií mikrodávky na tabuľkovú reprezentáciu. Vytvárané povinné hodnoty typu status a quality je vhodné rovno zadefinovať číselné hodnoty a vyhnúť sa tak, zbytočnej konverzií. Korelácia dát po transformácií je zobrazená pomocou teplotnej mapy je zobrazenńa na obrázku číslo 19. Výrazné farby majú významný vplyv na predikciu a budú zaradené ako potenciálne parametre, ktoré zabezpečia učenie modelu a jeho rozhodovanie.



Obrázok 19. Vytvorená teplotná mapa pre údaje na základe trénovacej množiny

Z pripravenej teplotnej mapy je možné identifikovať parametre, ktoré majú význam pre učenie klasifikácie a ovplyvňujú presnosť predikcie. Medzi parametre, ktoré je sem možné zaradiť patria parametre v ohodnotenej korelačnej mape. Všetky plusové aj mínusové hodnoty blízke -1 a +1 sú vhodné zahrnúť do výberu. Je možné otestovať aký vplyv môžu mat hodnoty okolo -0,5 a + 0,5. Menšie hodnoty nemusia mať významný vplyv a je možné ich vynechať alebo urobiť viacero

pokusov s validačným datasetom overiť správanie sa modelu. Ako je vidno z teplotnej mapy významný vplyv majú INT_SRC_IP, INT_DST_IP IP_SIZE a PACKET_SIZE ale aj IP_PROTOCOL v kombinácií s SRC_PORT_OR_ICMP_TYPE.

Po výbere parametrov, je možné vytvoriť model a overiť jeho percentuálnu účinnosť na validačnej množine. Pri trénovaní množiny je vytváraný nový stĺpec features, ktorý nesie informáciu o vektore, ktorý je reprezentovaný poľom. Vektor určuje predikciu pre daný paket. Tento stĺpec určuje, či je paket klasifikovaný ako korektná dátová prevádzka na sieti alebo je to potenciálny nebezpečná dátová prevádzka. Ukážka výstupu pre status a stĺpec features je zobrazená na obrázku 20.

```
1 from pyspark.ml.classification import LogisticRegression
2 final_data = output_data.select('features','status')
3 final_data.show()
```

```
+--------------------+------+
|            features|status|
+--------------------+------+
|[2.0,0.0,4.637121...|     1|
|[2.0,0.0,4.637121...|     1|
|(14,[0,2,3,4,12,1...|     1|
|(14,[0,2,3,4,12,1...|     1|
|(14,[0,2,3,4,12,1...|     1|
```

Obrázok 20. Vektorová reprezentácia features podľa statusu.

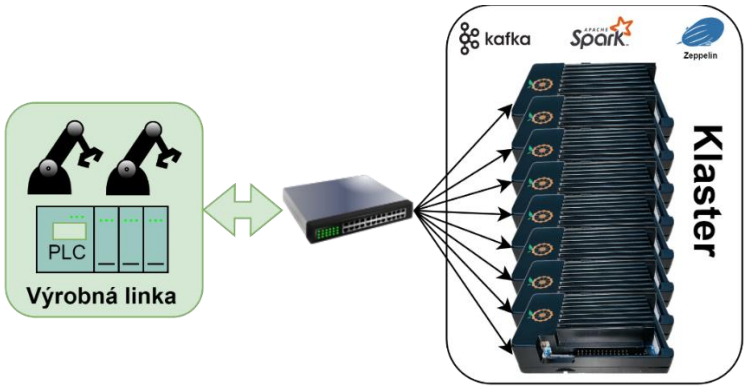
Parametre modelu z natrénovaných dát sú uvedené v tabuľke číslo 1. PySpark používa na vrátenie štandardnej odchýlky (standard deviation) hodnoty pre konkrétny stĺpec v datasete. Ak je status a predikcia rovnaká, tento indikátor reprezentuje 100% zhodu v predikcií pre validáciu. Tento výsledok bol dosiahnutý spoznaním sieťového trafiku a demonštráciou špecificky odlišných paketov. Pre reálnu prevádzku je ale bez väčšieho problému možné dostať výsledky podobné, teda kvalitu predikcie 100%, nakoľko bežná prevádzka presne určuje, čo s čím komunikuje. Pridaním nového IoT zariadenia, môže byť jeho komunikácia označená ako nebezpečná a zablokovaná.

Tabuľka 1. Parametre natrénovaného modelu

Sumár	status	predikcia
Počet jedinečných vzoriek	4547	4547
mean	1.0704225352112675	1.0704225352112675
stddev	0.25621455364847543	0.25621455364847543
Min hodnota predickie	1	1
Maximálna hodnota predikcie	2	2

Vytvorená štruktúra umožňuje identifikovať aj veľkosť a ohraničenie okna. V experimente sa jedná o vyhodnotenie jedného paketu takže ohraničenie je platné pre každý objekt JSON ako samostatná úloha.

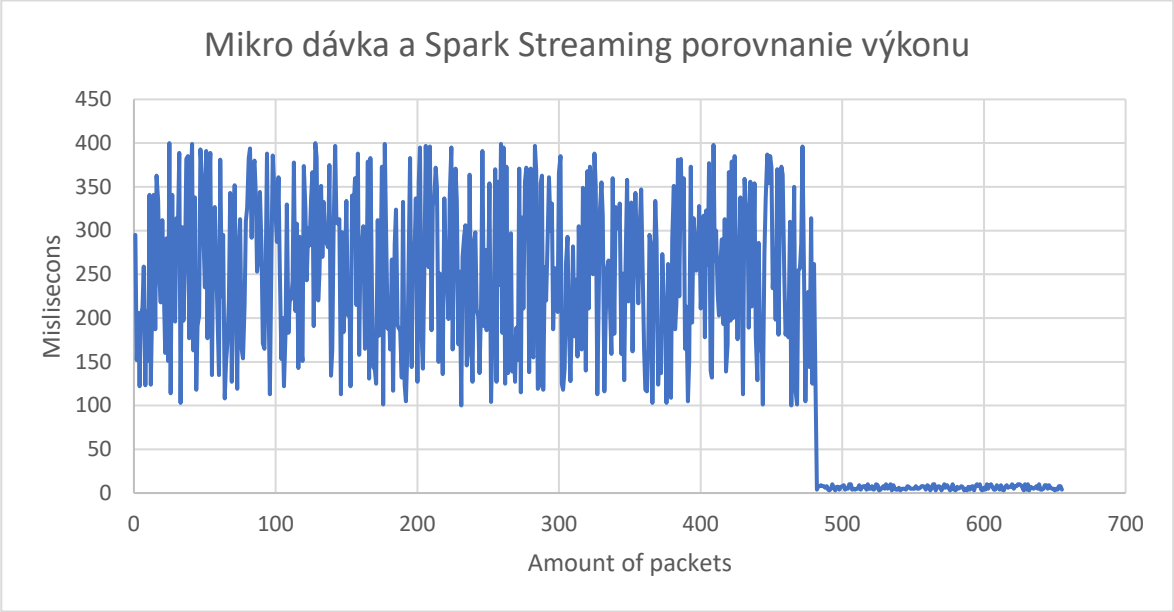
Nasleduje doručenie za pomoci systému Kafka. Bola overená synchrónna aj asynchrónna komunikácia typu half duplex. Dáta za pomoci synchrónnej aj asynchrónnej komunikácia vykazovali rovnakú spoľahlivosť a aj rýchlosť doručenia. Pri synchrónnom spojení je potrebné definovať neblokujúce spojenie v prípade posielania komunikácie. Blokujúce spojenie znamená, že je overované doručenie na každý uzol clusteru, ak je náodou potrebné spracovanie na inom uzle, je toto opäť validované a blokované. Preto môže dôjsť k spomaleniu. V našom testovacom prípade bolo zapojenie Kafka a Spark realizované pomocou 8mich mini počítačov OrangePI 4B s procesorom ARM, 256GB SD kartou a 8BG RAM. Nebolo možné vyťažiť systém tak, aby bolo možné objektívne posúdiť vplyv typu synchrónnej a asynchrónnej komunikácia na rýchlosť doručenia. Bloková schéma zapojenia je na obrázku 21.



Obrázok 21. Bloková schéma zapojenia klustru pre Kafka a Spark.

Synchrónnou komunikáciou boli dáta spoľahlivo doručované na vstup modelu, ktorý vykonával klasifikáciu paketov a podľa výsledku klasifikácie vygeneroval firewall pravidlo pre sieťový prepínač. Prepnutie medzi mikro batch spracovaním a online spracovaním je pomerne jednoduché.

V súčasnosti zvyčajne používajú vo svojich aplikáciách spracovanie mikrodávok len vtedy, ak prijali architektonické rozhodnutia, ktoré vylučujú prúdové spracovanie. Napríklad v Apache Spark sa môže používať Spark Streaming, ktorý je - napriek svojmu názvu a používaniu výpočtových zdrojov v pamäti - v skutočnosti rozšírením rozhrania Spark API pre spracovanie mikrodávok. Dátové toky a spracovanie namiesto mikrodávky umožňuje aplikáciám reagovať na nové dátové udalosti takmer okamžite takmer v času vzniku. Grafický porovnanie rýchlosti na obrázku 22.



Obrázok 22. Porovnanie výkonu spracovania mikrodávky a Spark Data Streamu

Na grafe je najskôr spracovanie pomocou mikrodávky. Pakety sú vyhodnotené pomocou mikrodávkového režimu. V bode 490 je viditeľná dôjde k režimu prepnutia do Spark streamingu a vyhodnotenie bolo zrýchlené a dĺžka spracovania je od 3 do 10 milisekúnd.

Namiesto zoskupovania údajov a ich zhromažďovania v určitom vopred stanovenom intervale prúdové spracovania zhromažďujú a spracúvajú údaje okamžite po ich vytvorení.

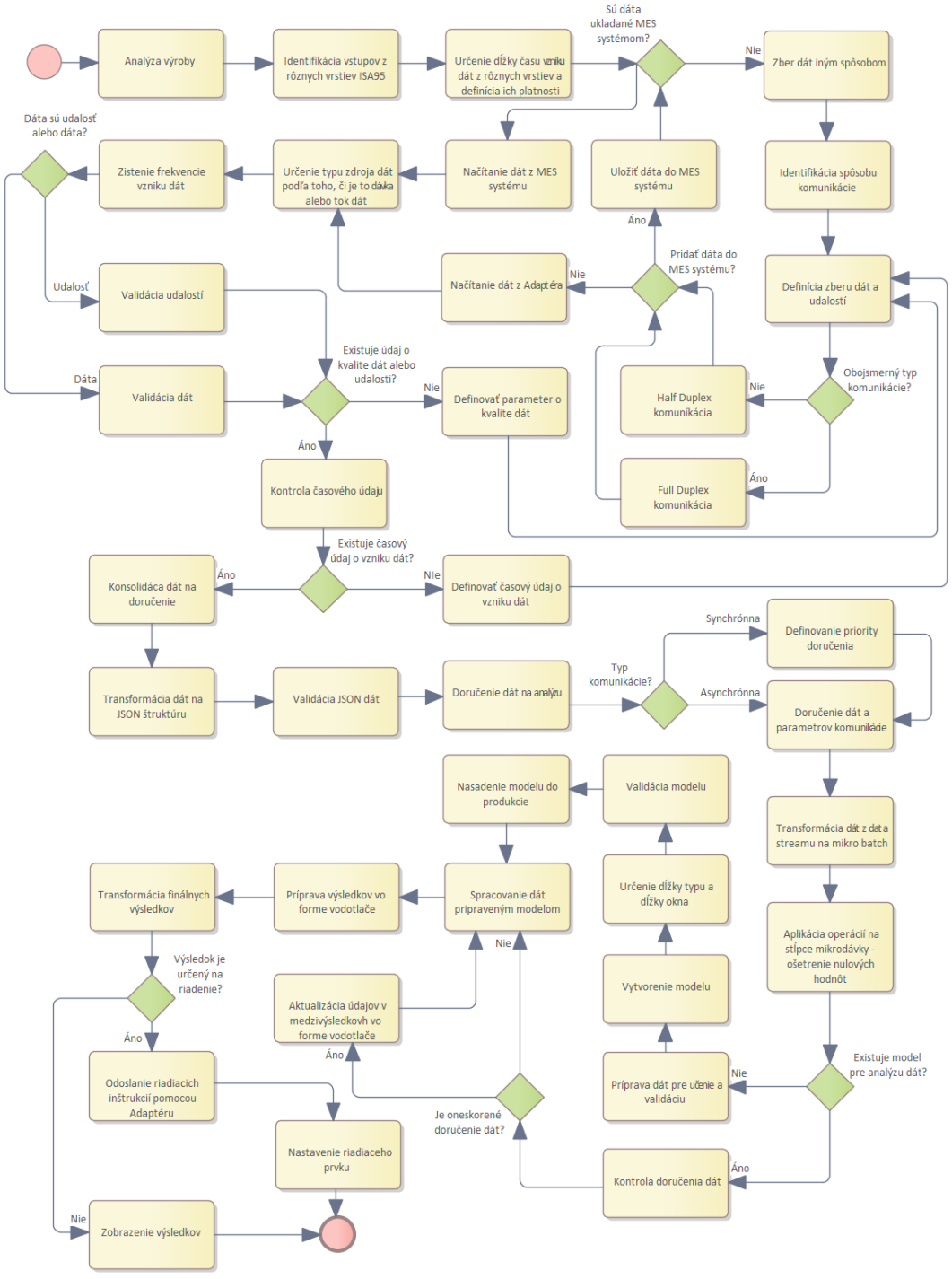
Porovnanie výsledkov modelu a nezávislé overenie správania sa výrobnej linky umožnilo vidieť vplyv útokov aj za pomoci MES systému. V histórií a zbere dát MES systému, bolo možné pozorovať výpadky komunikácie v časoch, kedy boli identifikované útoky a nebolo implementované zabezpečenie. Týmto postupom bolo overené, že vytvorené riešenie identifikácie a zabránenia nesprávnej sieťovej komunikácie je možné zvýšiť bezpečnosť výrobnej linky a tým aj spoľahlivosť výroby. Pre nesprávnu komunikáciu je možné vyhodnotiť, parameter spoľahlivosti, ktorý určuje potenciálne zlyhanie komunikácie medzi prístrojmi a riadiacimi členmi.

Overenie funkčnosti návrhu a modelu bolo ďalej vykonané na viacerých testovacích scenároch, ktoré boli publikované počas štúdia a publikované v karentovaných časopisoch.

Navrh a relizáciu experimentu, správanie sa modelu a proces komunikácie je možné vyzualizovať. Ako vhodný je opäť systém, ktorý podporuje doteraz vybrné prostriedky a patrí do skupiny Apache. Systém Apache Zeppeling je možné integrovať a vztvoriť prehľady a sledovať aplikované zmeny alebo aj monitorivať chovanie systému. Je vhodný a ľahko je možné za pomoci dashboardu preskúmať skúmať a odladiť prípadné nedostatky, ktoré môžu nastať počas realizácie. Táto časť nie je nevyhnutná ale môže napomôcť v jednoduchšej realizácií výsledného riešenia.Celkové zhrnutie a porovnanie navrhovaných riešení

11 Zovšeobecnenie návrhu modelu pre objavovanie znalostí

Z navrhnutých a vypracovaných čiastkových úloh je možné navrhnúť všeobecný BPMN diagram, ktorý umožní vytvorenie všeobecného návrhu modelu pre spracovanie a analýzu dát pre potreby objavovania nových znalostí z dát. Návrh je zobrazený na obrázku číslo 23.



Obrázok 23. BPMN diagram pre návrh a implementáciu modelu pre získavanie znalostí z databáz výrobných podnikov (štruktúrované dáta) pre potreby riadenia procesov

Na základe analýzy výroby je potrebné preskúmať všetky úrovne riadenia podľa ISA95. V ďalšom kroku je potrebné identifikovať potenciálne vstupy. V tejto časti je možné definovať systémy, ktoré budú zahrnuté do zberu dát. Nasledujúca aktivita je zameraná na určenie vzniku dát a dĺžky životného cyklu a platnosti dát.

V prípade, že je výroba riadená MES systémom je potrebné zistiť, či sú dáta ukladané MES systémom a ako dlho sú tieto dáta o výrobe k dispozícii. MES systém býva vybavený podsystémom, ktorý je určený na ukladanie a historizáciu dát. Ak dáta neexistujú v MES systéme a boli identifikované a označené ako potrebné, je potrebné zabezpečiť ich zber iným spôsobom. Je potrebné určiť spôsob komunikácie, ktorý môžu predstavovať protokoly spomenuté v teoretickej časti, alebo aj iné. Výsledkom je určenie typu komunikácie. Nasleduje zadefinovanie zberu dát a ak je systém schopný generovať udalosti, ktoré môžu byť indikátorom aktivity, je vhodné ich zahrnúť do výberu. Po určení typu komunikácie je vhodné zistiť, či sa jedná o obojsmernú alebo jednosmernú komunikáciu, a či je očakávaná odpoveď systému napríklad pre potvrdenie prijatia dát a podobne. Nasleduje rozhodnutie aký typ adaptéru bude nasadený podľa určenia komunikácie a to buď Full Duplex pre spomínaný obojsmerný typ komunikácie alebo Half Duplex pre jednosmerný typ komunikácie. Následne sú dáta klasifikované z pohľadu potreby pridať ich integráciu do MES systému alebo či je vhodnejšie nasadenie nového prvku Adaptéra, ktorý bude prostredníkom pre zber údajov zo zvoleného komponentu. Ak je rozhodnutie pridať dáta do MES systému, je potrebné priradiť tieto dáta, namapovať ich a zabezpečiť ich ukladanie s označením kvality, času vzniku, všetkými dostupnými dátami, ktoré sú načítané v dostupnom stave, prípadne ich transformovať na potrebnú veličinu. Táto časť je ale súčasťou integrácie dát MES systému. Ak je potrebné nasadenie adaptéra, je potrebné implementovať ho podľa modelov, ktoré sú opísané v kapitole 8.1. Oba spôsoby sú vhodné a ich výsledkom sú nové dáta, ktoré sú k dispozícii na ďalšie spracovanie. Nasleduje krok určenia typu zdroja dát podľa spôsobu ukladania alebo doručovania dát. Dáta sú k dispozícii v podobe dávky alebo dátového toku. Ďalšia aktivita je určenie frekvencie vzniku dát. Dávka môže byť k dispozící raz za určený čas, napríklad export dát beží raz za deň. Tok dát je takmer s určitosťou generovaný počas výroby alebo aktivity. Nasleduje klasifikácia dát a rozdelenie typu dát podľa toho, či sa jedná o udalosť alebo dáta, ktoré opisujú stav zberaného objektu. Ideálna je kombinácia oboch. Nasleduje aktivita, ktorá vykoná validáciu dát alebo udalosti. V prípade, že validácia dát a udalosti neobsahuje dva základné parametre.

- Prvým je kvalita dát alebo udalostí. Kvalita predstavuje hodnovernosť zberaných údajov.
- Druhým parametrom je čas vzniku dát alebo udalostí. Tento parameter, ak nie je k dispozícii je

možné odvodiť od doručenia dát. Lepšie je ale implementovať tento údaj pre každú operáciu, ktorá generuje dáta alebo udalosti. Tento typ informácie dokáže pomôcť identifikovať chybnú integráciu alebo uľahčí hľadanie potenciálnych chýb, ktoré môžu vzniknúť pri implementácií riadiaceho prvku a sú kritické z pohľadu času.

Ak je potrebné je nutné tieto dva parametre vytvoriť v časti, kde sa definuje, ktoré dáta majú byť zberané.

Dáta a udalosti, ktoré majú všetky povinné parametere a prešli validačným procesom, je možné spracovať v nasledujúcej aktivite nazvanej konsolidácia dát na doručenie. Táto aktivita je zodpovedná za prípravu a rozloženie dát do podoby, ktorá je potrebná na zostavenie štruktúry alebo objektu. Identifikuje dáta a na základe identifikátorov zdroja a povinných parametrov, ktoré sú očakávané pre daný element je možné pristúpiť k aktivite Transformácie dát na JSON štruktúru. JSON štruktúra je schopná opísať a obsiahnuť objekt a jeho parameter. Po vytvorení JSON štruktúry je táto opäť validovaná. Táto časť je odporúčaná, nakoľko závisí od konsolidácie a oba kroky môžu používať rovnakú definíciu overenia JSON štruktúry. Výsledkom validácie je JSON dokument, ktorý je možné doručiť na spracovanie. Je identifikovaný typ komunikácie podľa špecifických požiadaviek na doručenie dát. Môže ísť o synchrónny alebo asynchrónny typ komunikácie. Synchrónny typ je náročnejší na doručenie a musí obsahovať parametre, ktoré zabezpečia konfiguráciu synchrónneho prenosu, určenie priority a podobne, podľa zvoleného integračného nástroja. Asynchrónny typ komunikácie, definuje čas, ktorý je potrebný na odoslanie a následne je nezávislý na oboznámení sa o doručení alebo predpokladam nedoručení výsledku. Opäť závisí od špecifických požiadaviek podľa výberu integračného systému.

Po príprave dát, transformácií a doručení dát na spracovanie je potrebné pristúpiť k modifikácií dát. V tomto prípade sa jedná o transformáciu dát streamu a dávky na systém mikrodávky. Transformácia dátového toku je ohraničená udalosťami alebo dátami, ktoré predstavujú určitý stav, na základe ktorého je možné ohraničiť vykonávanú a skúmanú činnosť. Toto ohraničenie musí byť kratšie ako je definovaný čas na spracovanie jednej dávkovej činnosti. Ohraničenie nesmie byť kratšie ako plánovaná úloha ktorá je automaticky spúšťaná v čase Triger Time. Spracovanie viacerých ohraničených mikrodávok v rámci jedného spustenia pomôcou plánovača ale možné je a je nutné v prípade medzivýsledkov dostupných pod tzv. vodoznakom, alebo v prípade, že sú očakávané oneskorené doručenia dát. Spracovanie dátového toku spoločne v kombinácií so statickými alebo presnejšie dávkovým dátami, ktoré sú dostupné v databáze, je potrebné identifikovať kľúčové identifikátory, na základe ktorých je zabezpečený výber dát z databázy. Táto

aktivita je považovaná za spoľahlivú a preto nie je nutné implementovať medzivýsledky v podobe vodootlačku. Spracovanie viacerých dátových tokov, je považované za štandardný postup. Pre výber a definíciu mikrodávky pre dva alebo viac dátových tokov, je opäť určený kľučový identifikátor alebo identifikátory a mikrodávky sú spracovávané ako štandardná úloha. Opäť je tu povinnosť ošetriť situáciu s oneskoreným doručením dát a kontrolou medzivýsledkov vo forme vodoznaku. Existuje tu ale významná výnimka, pre spájanie dát pomocou metód outer join, ktoré môžu predstavovať systémovo náročnú operáciu.

Po príprave dát sú dáta pripravené v mikrodávke, a ich transformáciou na tabuľku je možné aplikovať stĺpcové operácie. Typickým príkladom je ošetrenie nulových hodnôt, doplenie chýbajúcich dát, odstránenie extŕemov a podobne. Tieto sú zo systémového hľadiska v tejto fáze málo náročné na systém a ich predchádzajúca transformácia a úprava nebola možná alebo bola neželaná.

Analytický systém je pripravený a nasleduje rozhodnutie, či existuje model, pre ktorý sú pripravené dáta na spracovanie. Ak nie, je potrebné model vytvoriť ak chceme aby dáta boli vyhodnotené. Je potrebné zabezpečiť zber údajov, následne ich klasifikáciu, prípravu trénovacieho a validačného datasetu. Nasleduje výber metódy a vytvorenie modelu a jeho testovanie. Kompletný postup je opísaný v príprave dát a návrhu modelu.

Nasleduje identifikácia dĺžky a typu okna. Podľa zvolených parameterov je definované okno, ktoré bude slúžiť na analýzu dát. Typ okna je najčastejšie ohraničený udalosťami pri riadení v priemysle, len odzrkadľuje skúmané deje počas výroby.

Nasleduje validácia modelu, ktorý je testovaný v reálnych podmienkach výroby alebo pomocou digitálneho dvojčaťa. Ak model vyhovuje, je možné jeho nasadenie do produkcie. Integráciou modelu do produkcie, je možné spracovanie dát z riadenia výroby. Na ich základe je potom možná príprava výsledkov vo forme vodotlače. Ako bolo spomínané je definícia medzivýsledku, ktorý po splnení podmienky časovej alebo dátovej, ktorá potvrdí správnosť je transformovaný do finálneho výsledku. V tejto aktivite dochádza k spojeniu čiastkových medzivýsledkov alebo potvrdeniu medzivýsledku. Ak je určený výsledok na zabezpečenie riadenia alebo zabezpečuje nadstavenie systému, je zabezpečené doručenie dát za pomoci adaptéru. Takto je zabezpečené vykonanie požadovaného nastavenie, alebo úprava hodnôt. Ďalje je možná aplikácia pravidla alebo spustenie alebo zastavenie operácie. V prípade, že je výsledok určený na zobrazenie, je možné ho vizualizovať.

12 Prínosy dizertačnej práce

V práci bola analyzovaná problematika súčasného stavu problémovej oblasti objavovania nových znalosti z bázy dát. Boli analyzované a vyhodnotené systémy doručovania zberu dát a ich uplatnenie pre nasadenie v priemysle ale aj uplatnenie pre integráciu všeobecne v informačných systémov. Bola vykonaná analýza sieťových protokolov, ich využitie so zameraním na presnosť a rýchlosť prenosu údajov, na základe ktorých boli určené požiadavky na komunikačné schopnosti navrhovaného riešenia. Zároveň táto analýza slúži ako základ pre zahrnutie sieťovej prevádzky pre dosiahnutie objavenia nových znalostí, zvýšenia komunikačnej bezpečnosti a výrobnej spoľahlivosti.

V práci je navrhnutý komplexný prístup pre získavanie dát, z rôznych vrstiev ISA95 architektúry, z heterogénnych úrovní riadenia. Úrovne boli klasifikované do časových zón, kde každá vrstva určuje platnosť dĺžky vytváraných dát. Podľa tejto kategorizácie je možne určiť, ako pristúpiť k ETL procesu a vhodne vybrať spôsob spracovania dát k následnej analýze. Dáta s dlhším časovým obdobím je možné pokladať za dáta vhodné pre spracovanie v podobe dávkového spracovania údajov. Nižšie vrstvy, kde sa časová úroveň riadenia pohybuje v jednotkách sekúnd alebo v milisekundách, je výhodnejšie spracovanie pomocou dátových tokov.

Proces získavania dát zahŕňa možnosti využitia dát z existujúcich riadiacich systémov ako je MOS a MES a ich archivačných častí ako napr. systém Historian, ktorý ukladá dáta a archivuje výrobných dát, ktoré vznikajú počas výroby. Na základe analýzy bolo zistené, že je potrebné počas vytvárania modelov doplniť nové údaje. Vzhľadom na riziko zmien integrácie v produkčnom prostredí, bol vytvorený návrh implementácie pomocou adaptérov, ktoré zabezpečia doplnenie údajov tak, aby bolo minimalizované riziko zmien výroby.

Podľa typu komunikácie adaptéry boli navrhnuté pre half duplex komunikáciu a pre full duplex komunikáciu. Oba typy adaptérov je možné kombinovať a je možné ich integrovať do výroby podľa potreby a hlavne podľa typu použitých technických a komunikačných prostriedkov. Bol navrhnutý spôsob a opísané výhody synchrónnej a asynchrónnej komunikácie a ich vplyv na výkon. Na základe požiadaviek návrhu bol zvolený nástroj Apache Kafka, ktorý vyhovuje požiadavkám z výkonnostného hľadiska ako aj možnosti širokej škálovateľnosti a zabezpečenia spoľahlivosti komunikácie a doručovania dát.

Cieľom dizertačnej práce je návrh platformy na spracovanie a analýzu bázy dát a dátových tokov so zameraním sa na objavovanie nových znalostí s využitím technológií Industry 4.0 (prevažne Big

Data a IIoT) pre potreby efektívnejšieho riadenia výrobných procesov, s prihliadnutím na atribúty bezpečnosti a spoľahlivosti. V práci sú demonštrované a opísané správanie sa systému, kategorizácia do vrstiev, na základe ktorých je možné identifikovať očakávané vedomosti, podľa typu použitých metód. Pre potreby zabezpečenia tohto cieľa bol spracovaný jednotný návrh na tvar doručovania dát, kde bol vybraný formát JSON, na základe vlastností, ktorý spĺňa komplexný opis dát a je možné jeho univerzálne použitie pre všetky vrstvy od ERP až po riadiace vrstvy.

Návrh testovacieho a validačného data setu vychádza z rozdielov dávkového spracovania a dátových tokov. Zjednodušenie je demonštrované na opísanom modely pripravovaných znalostí na dávkovom modely. V návrhu je opísaný všeobecný postup vytvárania modelu. Nasleduje všeobecný opis učenia modelu a jeho validácia a vyhodnotenie parametrov na presnosť modelu. Po príprave modelu, validácií a optimalizácií parametrov, nasleduje použitie a testovanie modelu v aplikáciách orientovaných na dátový tok. Aby bolo možné model použiť na spracovanie dátového toku je nutná identifikácia ohraničenia, teda okna, ktoré ohraničí spracovanie dátového toku na mikrodávku. Mikrodávka je následne transformovaná na tabuľkovú reprezentáciu a jej spracovanie prebieha klasicky ako pre dávkové spracovanie. Je tu definovaných viacero výnimiek, ktoré sú opísané ako špeciálne situácie, ktoré majú vplyv na výsledok alebo na výkon a stabilitu navrhovanej architektúry.

Vplyv plánovania úloh a dĺžka ich platnosti vzhľadom na doručenie dát, a dĺžku okna pre vymedzenú mikrodávku, prípadne operácie, ktoré sú pamäťovo náročné na spracovanie dát, môžu viesť k preťaženiu a nestabilite systému alebo aj k jeho kolapsu. Výber typu okna podľa dát je ďalej presnejšie opísaný s ohľadom na typ pripravovaných úloh. Z povahy dát sieťovej komunikácie a integrácie riadiacich systémov v priemysle je najčastejším ohraničenie okna definované udalosťami alebo objektom JSON, ktorý je doručený na vstup modelu. Nasleduje stavová alebo bezstavová transformácia, ktorá rieši problematiku oneskoreného doručenia dát a vytváranie medzivýsledkov vo forme vodoznaku. Po splnení podmienok a doplnení údajov a platnosti mikrodávky je dokončený finálny výsledok, ktorý je možné použiť na riadenie. Forma výsledku je rozdelená na viacero kategórií podľa typu možnej aktualizácie na kompletnú, režim výsledku vo forme doplnenia alebo režim aktualizácie. Pre najčastejšie aplikované úlohy je tu navrhnutý spôsob spracovania dát dátových tokov a dávok, alebo kombinácia viacerých dátových tokov. Obe majú spoločnú vlastnosť a to je výber kľúčových identifikačných parametrov pre určenie výsledkov.

Na základe realizovaného experimentu bol potom vytvorený zovšeobecnený proces návrhu objavovania nových vedomostí, ktorý je uplatniteľný nielen v priemysle ale aj všeobecne.

Navrhnutý BPMN diagram, zahŕňa všeobecný postup, kde postupnou realizáciou odporúčaných postupov, je možné dosiahnuť komplexné riešenie pre návrh ETL procesu, modelu, trénovacích a validačných data setov ako aj nasadenie a vylepšenie integrácie pomocou adaptérov. Ošetrené sú aj situácie a navrhnuté doplnenie kritických dát ako je čas vzniku údajov a informácia o kvalite zberaných dát. Nasleduje validácia modelu a vyhodnotenie a aplikácia výsledkov.

Naplnením všetkých čiastkových cieľov je možné prehlásiť, že hlavný cieľ dizertačnej práce bol úspešne splnený.

Nasledujúce body opisujú prínosy dizertačnej práce pre rôzne oblasti.

Prínosy pre teóriu:

- Analýza teoretických východísk, použitých komunikačných a integračných technológií z domácej aj zahraničnej literatúry a dostupných vedeckých štúdií a špecifikácií protokolov.
- Zhrnutie poznatkov týkajúcich sa ETL procesov, návrh adaptérov pre podporné systémy a komunikačné protokoly.
- Identifikované typy komunikácie podľa vhodnosti použitia, podľa typu úloh a riadiacich prvkov.
- Návrh získavania dát z databáz a dátových tokov, štandardizácia dátového popisu zberaných dát.

Prínosy pre prax:

- Návrh ETL procesu, aplikácia a overenie na modely výrobnej linky a na sieťovej komunikácií riadiacich prvkov s integrovaným MES systémom.
- Návrh zberu dát, definovanie nových atribútov pre analytické účely, transformácia dát na úrovni zberu alebo na úrovni modelu, podľa typu operácie a vhodnosti z hľadiska požadovaného výkonu.
- Prínosov integrácie IIoT zariadenia s podporou riadenia na základe udalostí a ich využitie v analytickom procese objavovania nových znalostí vo výrobnom procese.
- Zovšeobecnenie navrhnutého riešenia objavovania nových znalostí s orientáciou na využitie v priemyselných podnikoch ale aj pre všeobecné nasadenie pre informačné systémy a zabezpečenie sieťovej prevádzky.

Prínosy pre vedu:

- Výsledky dizertačnej práce boli publikované na medzinárodných vedeckých konferenciách a v karentovaných časopisoch typu Q1 až Q3.
- Výsledky skúmania dizertačnej práce sú široko uplatniteľné a boli akceptované aj pri tvorbe dokumentov OPC UA špecifikácie v rámci konceptu Industry 4.0 a môžu slúžiť ako podklad pre tvorbu ďalších dizertačných prác a vedeckých štúdií.
- Výsledky dizertačnej práce preukázali opodstatnenosť ďalšieho skúmania bezpečnosti IoT zariadení vo výrobnom procese.
- Výsledky dizertačnej práce boli publikované v karentovaných časopisoch Q1 a Q2 a prezentované na svetovom kongrese.
- Získané výsledky je možne využiť pre tvorbu vedeckých štúdii a príspevkov so zameraním na výskum problematiky dizertačnej práci.

Záver

Cieľom práce je návrh platformy na spracovanie a analýzu bázy dát a dátových tokov so zameraním na objavovanie nových znalostí na základe technológii Industry 4.0 (prevažne Big Data a IIoT) pre potreby efektívnejšieho riadenia výrobných procesov, s prihliadnutím na atribúty bezpečnosti a spoľahlivosti.

Boli identifikované a analyzované komponenty a normy Industry 4.0. Analyzované technológie vychádzajú z trendu, na základe ktorého sa odvíja realizácia integračných riešení. Ďalej boli analyzované typy komunikácie, ich forma, štruktúry a princípy, na základe ktorých je možné pristúpiť k analýze dát. Techniky Data miningu a metódy objavovania informácií v dátach a zhrnuté metódy, ktoré umožňujú zjednodušiť a správne identifikovať informácie. Bol pripravený detailný ETL proces, ktorý je univerzálne použiteľný pre bázu dát získavaných zo statických zdrojov, ako sú súbory, databázy alebo dátové sklady. Ďalej boli navrhnuté adaptéry, ktoré sú schopné zabezpečiť dátovú integráciu a komunikáciu v synchrónnom, asynchrónnom režime a sú schopné zabezpečiť spôsob komunikácie vo full duplex alebo aj half duplex režime. Návrhom ETL procesu sú pokryté potreby získavania dát, z ohľadom na pripravované scenáre použitia.

Výsledkom ETL procesu sú privádzané dáta normalizované na štruktúry typu JSON, ktoré predstavujú vstupné dáta pre analýzu. Bola vykonaná analýza metód, výstupom je identifikácia vhodnosti a rozobrané vhodné prípady použitia. Bola identifikovaná ich vhodnosť aplikácie na získavanie nových znalostí, ktoré vyplývajú zo stanovených a analytických vrstiev a ich prípadné

obmedzenia pre použitie na statické spracovanie údajov alebo ich vhodnosť pre nasadenie aj na dátové toky. Výsledkom je prehľad, ktorý umožňuje identifikovať, vhodnosť metód na spracovanie konkrétneho typu úloh.

V kapitole, ktorá sa venuje dáta streamu, boli analyzované metódy a nástroje, ktoré umožňujú spracovávať toky dát a opäť bola vykonaná analýza, ktoré známe metódy sú vhodné a určené na typy úloh podľa typu klasifikácie alebo regresie. Dáta streamy, výber vzoriek, prípadne riešenia, ktoré určujú rámce spracovávaných dát boli analyzované a vyhodnotené tak, aby bolo možné určiť potenciálnu stratu informácie alebo potenciálne zníženie kvality identifikácie možných vedomostí v toku dát. Strojové učenie a metódy, ktoré je možné aplikovať na analýzu dáta streamu sú ďalšou časťou a sú úzko prepojené s identifikáciou a stanovením rámca, ktorý slúži ako základná jednotka na rozdelenie a vytvorenie bloku dát, ktoré sú vyhodnocované. Podstatná časť bola venovaná výberu typu okna, na základe, ktorého je možné realizovať dátovú analýzu v tokoch dát.

Technológie, ktoré slúžia na integráciu boli ďalším predmetom skúmania a boli zhrnuté tak, aby bol jasný ich spôsob fungovania a spracovávania, doručovania dát a ich možnosť nasadenia v realtime prostredí. Na základe týchto poznatkov, je možné navrhnúť architektúru riešenia, ktoré dokáže zabezpečiť integráciu dát, objavovanie nových znalostí na základe historických dát alebo dát, ktoré sú generované počas výroby alebo sieťovej komunikácie riadiacich protokolov a integrovaných IIoT. V poslednej časti boli popísané nástroje, ktoré sa používajú na analýzu bázy dát a dátových. Experiment realizovaný za pomoci nástrojov Apache Spark a Kafka, ktoré sú zamerané na pracovanie dátových tokov a analýzu Big Data, overili funkčnosť navrhovanej architektúry. Na základe výsledkov bol zostavený univerzálny model a boli opísané funkcie a spôsoby fungovania, a identifikácia vlastností, podľa typu riadenia a integrácie. Výsledkom dizertačnej práce je navrhnutý univerzálny model, na základe ktorého je možné realizovať architektúru objavovania nových znalostí a aplikovať metódu prípravy analýzy dát nielen v priemyselných podnikoch. Model je všeobecne uplatniteľný pre analýzu dát v podnikoch, čo dokazuje aj nasadenie v telekomunikačnej firme, kde bol systém návrhu aplikovaný do systému kontroly kvality dát.

Zoznam publikačnej činnosti k 30.5.2023

V3 Vedecký výstup publikačnej činnosti z časopisu

V3_01 **STŘELEC, Peter** - HORÁK, Tibor - KOVÁČ, Szabolcs - NEMLAHA, Eduard - TANUŠKA, Pavol. IIoT Device Prototype Design Using State Machine According to OPC UA. In ***IEEE Access***. Vol. 10, (2022), s. 134004-134017. ISSN 2169-3536 **(2021: 3.476 - IF, Q2 - JCR Best Q, 0.927 - SJR, Q1 - SJR Best Q).** V databáze: DOI: 10.1109/ACCESS.2022.3232061 ; SCOPUS: 2-s2.0-85146224364 ; WOS: 000906231000001 ; CC: 000906231000001. **[Vnútrofakultná kategória: M].** Kategória publikácie do 2021: **ADC**

V3_02 VÁCLAVOVÁ, Andrea - **STŘELEC**, **Peter** - HORÁK, Tibor - KEBÍSEK, Michal - TANUŠKA, Pavol - HURAJ, Ladislav. Proposal for an IIoT Device Solution According to Industry 4.0 Concept. In ***Sensors***. Vol. 22, iss. 1 (2022), s. 1-27. ISSN 1424-8220 **(2021: 3.847 - IF, Q2 - JCR Best Q, 0.803 - SJR, Q1 - SJR Best Q).**V databáze: DOI: 10.3390/s22010325;SCOPUS: 2-s2.0-85121967894 ;WOS:000751041400001;CC:000751041400001.[**Vnútrofakultná kategória: M**]. Kategória publikácie do 2021: **ADC**

V3_03 HORÁK, Tibor - **STŘELEC, Peter** - HURAJ, Ladislav - TANUŠKA, Pavol - VÁCLAVOVÁ, Andrea - KEBÍSEK, Michal. The Vulnerability of the Production Line Using Industrial IoT Systems under DDoS Attack. In ***Electronics*.** Vol. 10, iss. 4 (2021), s. 1-32. ISSN 2079-9292 **(2021: 2.690 - IF, Q3 - JCR Best Q, 0.590 - SJR, Q2 - SJR Best Q)**. V databáze: DOI: 10.3390/electronics10040381 ; SCOPUS: 2-s2.0-85100446535 ; WOS: 000623355200001 ; CC: 000623355200001. [**Vnútrofakultná kategória: M*A**]. Kategória publikácie do 2021: **ADC**

V3_04 HORÁK, Tibor - **STŘELEC, Peter** - KEBÍSEK, Michal - TANUŠKA, Pavol - VÁCLAVOVÁ, Andrea. Data Integration from Heterogeneous Control Levels for the Purposes of Analysis within Industry 4.0 Concept. In ***Sensors*.** Vol. 22, iss. 24 (2022), s. 1-20. ISSN 1424-3210 **(2021: 3.847 - IF, Q2 - JCR Best Q, 0.803 - SJR, Q1 - SJR Best Q).** V databáze: DOI: 10.3390/s22249860 ; SCOPUS: 2-s2.0-85144503698;WOS:000904450900001;CC:000904450900001. **[Vnútrofakultná kategória: M]**. Kategória publikácie do 2021: **ADC**

V3_05 HURAJ, Ladislav - HORÁK, Tibor - **STŘELEC, Peter** - TANUŠKA, Pavol. Mitigation against DDoS Attacks on an IoT-Based Production Line Using Machine Learning. In ***Applied Sciences*.** Vol. 11, iss. 4 (2021), s. 1-18. ISSN 2076-3417 **(2021: 2.838 - IF, Q2 - JCR Best Q, 0.507 - SJR, Q2 - SJR Best Q).** V databáze: DOI: 10.3390/app11041847 ; SCOPUS: 2-s2.0-85101984331 ; CC: 000632078900001 ; WOS: 000632078900001. **[Vnútrofakultná kategória: M*A].** Kategória publikácie do 2021: **ADC**

V2 Vedecký výstup publikačnej činnosti ako časť editovanej knihy alebo zborníka

V2_01 **STŘELEC, Peter** - BARTOŇ, Martin - TANUŠKA, Pavol - KEBÍSEK, Michal - ŠPENDLA, Lukáš. Real Time Data Acquisition as a Part of Data Processing from Production Line According to. In ***World Congress on Industrial Control Systems Security*** *(WCICSS 2020) : Virtual Conference, December 8-10, 2020, London, UK*. 1. vyd. London, UK : Published by Infonomics Society, 2020, s. 139-144. ISBN 978-1-913572-26-6. **[Vnútrofakultná kategória: M*B].** Kategória publikácie do 2021: **AFC**

V2_02 **STŘELEC, Peter** - HORÁK, Tibor - KOVÁČ, Szabolcs - TANUŠKA, Pavol - NEMLAHA, Eduard. IoT Device Data Acquisition and Experimental Integration in Production Line Based on OPC UA Protocol. In *Software Engineering Perspectives in Systems [11th Computer Science On-line Conference 2022, Vol. 1]*. 1. vyd. Cham : Springer Nature, 2022, S. 215-223. ISSN 2367-3370. ISBN 978-3-031-09069-1. V databáze: DOI: 10.1007/978-3-031-09070-7_19 ; SCOPUS: 2-s2.0-85135041403 ; WOS: 000893645700019. [**Vnútrofakultná kategória: M**]. Kategória publikácie do 2021: **AFC**

V2_03 KEBÍSEK, Michal - TANUŠKA, Pavol - ŠPENDLA, Lukáš - KOTIANOVÁ, Janette - **STŘELEC, Peter.** Artificial Intelligence Platform Proposal for Paint Structure Quality Prediction within the Industry 4.0 Concept. In *IFAC-PapersOnLine*. Vol. 53, iss. 2: **IFAC World Congress, Berlín, Germany,** 12-17 July 2020 (2020), s. 11168-11174. ISSN 2405-8963 **(2020: 0.308 - SJR, Q3 - SJR Best Q).** V databáze: DOI: 10.1016/j.ifacol.2020.12.299 ; WOS: 000652593100380 ; SCOPUS: 2-s2.0-85103128034. [**Vnútrofakultná kategória: M*A]**. Kategória publikácie do 2021: **AFC**

V2_04 HORÁK, Tibor - **STŘELEC, Peter** - KOVÁČ, Szabolcs - NEMLAHA, Eduard - TANUŠKA, Pavol. IoT Device Using LoRaWAN for Data Transfer for Long Distances. In *Software Engineering Application in Systems Design : Proceedings of 6th Computational Methods in Systems and Software 2022 (CoMeSySo2022) conference, Volume 1*. 1. vyd. Cham : Springer, 2023, S. 491-500. ISSN 2367-3370. ISBN 978-3-031-21434-9. V databáze: DOI: 10.1007/978-3-031-21435-6_43 ; SCOPUS: 2-s2.0-85147996175. **[Vnútrofakultná kategória: M].** Kategória publikácie do 2021: **AFC**

V2_05 KOVÁČ, Szabolcs - **STŘELEC, Peter** - HORÁK, Tibor - MICHAĽČONOK, German - VAŽAN, Pavel. Forecasting Heat Production for a Large District Heating Network with NARX Neural Networks. In *Artificial Intelligence Trends in Systems : 11th Computer Science On-line Conference 2022 (CSOC 2022), Vol.2*. 1. vyd. Cham : Springer Nature, 2022, S. 131-139. ISSN 2367-3370. ISBN 978-3-031-09075-2. V databáze: DOI: 10.1007/978-3-031-09076-9_12 ; SCOPUS: 2-s2.0-85135059334 ; WOS: 000893642100012. **[Vnútrofakultná kategória: M].**Kategória publikácie do 2021: **AFC**

V2_06 NEMLAHA, Eduard - **STŘELEC, Peter** - HORÁK, Tibor - KOVÁČ, Szabolcs - TANUŠKA, Pavol. Suitability of MQTT and REST Communication Protocols for AIoT or IIoT Devices Based on ESP32S3. In *Software Engineering Application in Systems Design : Proceedings of 6th Computational Methods in Systems and Software 2022 (CoMeSySo2022) conference, Volume 1*. 1. vyd. Cham : Springer, 2023, S. 225-233. ISSN 2367-3370. ISBN 978-3-031-21434-9. V databáze: DOI: 10.1007/978-3-031-21435-6_19 ; SCOPUS: 2-s2.0-85148023204**.[Vnútrofakultná kategória: M].** Kategória publikácie do 2021: **AFC**

Štatistika: kategória publikačnej činnosti od 2022 (k 30.5.2023)

V2	Vedecký výstup publikačnej činnosti ako časť editovanej knihy alebo zborníka	6
V3	Vedecký výstup publikačnej činnosti z časopisu	5
Súčet		**11**

Štatistika: kategória publikačnej činnosti do 2021 (k 30.5.2023)

ADC	Vedecké práce v zahraničných karentovaných časopisoch	5
AFC	Publikované príspevky na zahraničných vedeckých konferenciách	6
Súčet		**11**

Štatistika: kategória ohlasov od 2022 (k 30.5.2023)

1 Citácia v publikácii registrovaná v citačných indexoch (WOS, SCOPUS)				**35**
	Zahraničné		35	
2 Citácia v publikácii vrátane citácie v publikácii registrovanej v iných databázach okrem citačných indexov				**10**
	Zahraničné		10	
Súčet				**45**

Zoznam bibliografických odkazov

1. Vazan, P., D. Janikova, P. Tanuska, M. Kebisek, and Z. Cervenanska, Using data mining methods for manufacturing process control, in 20th IFAC World Congress. 2017. pp.6178-6183
2. USAMA NOMAN: Data science the profession of the future? [cit. 2021-04-18]. Dostupné na internete: https://medium.com/@usamanoman/data-science-the-profession-of-the-future-d411215f6936
3. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37-37.
4. Lasi, Heiner, et al. "Industry 4.0." *Business & information systems engineering* 6.4 (2014): 239-242.
5. Shi, Zhan, et al. "Smart factory in Industry 4.0." *Systems Research and Behavioral Science* 37.4 (2020): 607-617.
6. V. Mayer-Schoenberger, K. Cukier, Big Data, ISBN: 978-80-251-4119-9
7. Gama, Joao. *Knowledge discovery from data streams*. CRC Press, 2010.
8. Ali, Zulfiqar, and S. Bala Bhaskar. "Basic statistical tools in research and data analysis." *Indian journal of anaesthesia* 60.9 (2016): 662.
9. Jing, Liping, Kuang Tian, and Joshua Z. Huang. "Stratified feature sampling method for ensemble clustering of high dimensional data." *Pattern Recognition* 48.11 (2015): 3688-3702.
10. Lancaster, Gemma, et al. "Surrogate data for hypothesis testing of physical systems." *Physics Reports* 748 (2018): 1-60.
11. Castillo-Davis, Cristian I., and Daniel L. Hartl. "GeneMerge—post-genomic analysis, data mining, and hypothesis testing." *Bioinformatics* 19.7 (2003): 891-892.
12. Nascimento, Abraao DC, Renato J. Cintra, and Alejandro C. Frery. "Hypothesis testing in speckled data with stochastic distances." *IEEE Transactions on geoscience and remote sensing* 48.1 (2009): 373-385.
13. Balzanella, Antonio, Lidia Rivoli, and Rosanna Verde. "Data stream summarization by histograms clustering." *Statistical models for data analysis*. Springer International Publishing, 2013.
14. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.
15. IANA: IEEE 802 Numbers [cit. 2023-02-11]. Dostupné na internete: https://www.iana.org/assignments/ieee-802-numbers/ieee-802-numbers.xhtml