PANEURÓPSKA VYSOKÁ ŠKOLA
FAKULTA INFORMATIKY

FI-185022-19043

ANALÝZA EKONOMICKÝCH ÚDAJOV S VYUŽITÍM PLATFORMY MICROSOFT AZURE

Diplomová práca

2026
Bc. Roderik Antol

PANEURÓPSKA VYSOKÁ ŠKOLA
FAKULTA INFORMATIKY

ANALÝZA EKONOMICKÝCH ÚDAJOV S VYUŽITÍM PLATFORMY MICROSOFT AZURE

Diplomová práca

Bc. Roderik Antol

Študijný program:	Aplikovaná informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Ústav aplikovanej informatiky
Školiteľ:	doc. RNDr. Eugen Ružický, CSc.

Bratislava 2026

PANEURÓPSKA VYSOKÁ ŠKOLA

Fakulta informatiky

ZADANIE DIPLOMOVEJ PRÁCE

Meno a priezvisko študenta:	**Bc. Roderik Antol**
Evidenčné číslo diplomovej práce:	**FI-185022-19043**
Študijný odbor:	**informatika**
Študijný program:	**Aplikovaná informatika**
Forma a metóda štúdia:	**externá prezenčná**
Vedúci diplomovej práce:	**doc. RNDr. Eugen Ružický, CSc.**
Ústav/katedra:	**Ústav aplikovanej informatiky**
Dátum zadania diplomovej práce:	**13. 09. 2025**

Názov: **Analýza ekonomických údajov s využitím platformy Microsoft Azure**

Anotácia: Cieľom diplomovej práce je spracovať a analyzovať ekonomické údaje s využitím platformy Microsoft Azure. Práca sa má zamerať na návrh a implementáciu integrovaného dátového riešenia v cloudovom prostredí Microsoft Azure, ktoré spája zber a spracovanie údajov, analytické metódy a vizualizačnú vrstvu do jednotného analytického systému. Navrhované riešenie bude predvedené na príklade analýzy verejných financií, pričom fiškálne a makroekonomické údaje budú slúžiť ako prípad použitia na overenie funkčnosti, škálovateľnosti a rozširovateľnosti riešenia.

.....

doc. RNDr. Eugen Ružický, CSc.
vedúci práce

.....

Ing. Juraj Štefanovič, PhD.
vedúci ústavu

Poďakovanie:

Pri vypracovaní tejto práce by som sa rád poďakoval za odbornú pomoc pánovi doc. RNDr. Eugen Ružickému, CSc., ktorý mi poskytol mnoho cenných rád a konzultácií.

Čestné prehlásenie:

Čestne vyhlasujem, že záverečnú prácu som vypracoval samostatne a že som uviedol všetku použitú literatúru.

. . . . . .

Bc. Roderik Antol

Abstrakt

Každoročne pribúda veľké množstvo dát, čo vedie k zvýšeným nárokom na informačné systémy. V akademickej literatúre však bývajú jednotlivé časti dátových riešení zväčša spracované izolovane. Táto práca sa zaoberá návrhom a implementáciou integrovaného cloudového dátového riešenia, ktoré prepája zber dát, ich spracovanie, analytické metódy a vizualizačnú vrstvu do jednotného uceleného analytického rámca.

Navrhnuté riešenie je implementované v prostredí Microsoft Azure s využitím služieb Azure Data Factory (ADF), Azure Storage Account a Azure Batch. Dátová vrstva je postavená na medailónovej architektúre, ktorá umožňuje oddeliť surové, očistené a analyticky pripravené dáta. Riešenie je konkrétne zamerané na štruktúrované fiškálne a makroekonomické údaje z webového portálu Eurostat a zároveň integruje ekonomické prognózy Európskej komisie a tlačové správy Európskej centrálnej banky vo forme textových dát. Nad týmito neštruktúrovanými údajmi je aplikovaný veľký jazykový model na analýzu sentimentu. Praktická použiteľnosť riešenia je demonštrovaná na verejných financiách krajín V4 a Rakúska za posledné desaťročie.

Výstupom práce je automatizovaný dátový proces a interaktívny Power BI dashboard, ktorý umožňuje analyzovať fiškálne ukazovatele, zahraničný obchod, štruktúru daní a verejné výdavky, pričom je doplnený o sentiment európskych inštitúcií z dlhodobého aj krátkodobého hľadiska. Výsledky ukazujú, že navrhnuté riešenie poskytuje reprodukovateľný, rozšíriteľný a prakticky použiteľný rámec pre komplexnú dátovú analytiku v cloudovom prostredí. Zároveň však samotné webové zdroje a charakter veľkých jazykových modelov môžu predstavovať určité obmedzenia.

Kľúčové slová: business intelligence, self-service BI, Microsoft Power BI, cloud computing, Microsoft Azure, Azure Data Factory, analýza sentimentu, verejné financie, medailónová architektúra

Abstract

Every year a large amount of data is generated, which leads to increased demands on information systems. In academic literature, however, the individual components of data solutions are usually addressed in isolation. This thesis focuses on the design and implementation of an integrated cloud-based data solution that connects data collection, data processing, analytical methods, and the visualization layer into a single comprehensive analytical framework.

The proposed solution is implemented in the Microsoft Azure environment using Azure Data Factory (ADF), Azure Storage Account, and Azure Batch. The data layer is based on a medallion architecture, which makes it possible to separate raw, cleaned, and analytics-ready data. The solution is specifically focused on structured fiscal and macroeconomic data from the Eurostat web portal, while also integrating economic forecasts from the European Commission and press releases from the European Central Bank in the form of textual data. A large language model is applied to these unstructured data for sentiment analysis. The practical applicability of the solution is demonstrated using the public finances of the V4 countries and Austria over the past decade.

The output of the thesis is an automated data process and an interactive Power BI dashboard that enables the analysis of fiscal indicators, foreign trade, tax structure, and public expenditure, supplemented by the sentiment of European institutions from both long-term and short-term perspectives. The results show that the proposed solution provides a reproducible, extensible, and practically applicable framework for comprehensive data analytics in a cloud environment. At the same time, however, the web sources themselves and the nature of large language models may present certain limitations.

Keywords: business intelligence, self-service BI, Microsoft Power BI, cloud computing, Microsoft Azure, Azure Data Factory, sentiment analysis, public finances, medallion architecture

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Každoročne pribúda veľké množstvo dát, čo viedlo organizácie si uvedomovať ich hodnotu [1]. Tento trend bol ešte výraznejšie urýchlený príchodom pandémie a s ňou spojenou digitalizáciou súkromného aj verejného sektora [2, 3, 4]. To malo za dôsledok zvýšené nároky na informačné systémy, spracovanie dátových zdrojov a ich prezentáciu [5, 6].

V akademickej ale aj praktickej sfére sú vysoko rozšírené nástroje typu self-service business intelligence (SSBI), ktoré umožňujú tvorbu interaktívnych analytických výstupov aj netechnických používateľom [7]. Napriek ich rastúcej popularite však existujú viaceré obmedzenia súvisiace s izolovaným využívaním vizualizačnej vrstvy, obmedzenou integráciou dátových procesov a nízkou mierou automatizácie. Zároveň možno v akademickej literatúre pozorovať, že cloudové dátové riešenia a pokročilé analytické metódy, ako napríklad analýza sentimentu, sú často spracované oddelene, prevažne na konceptuálnej úrovni, bez praktickej implementácie end-to-end riešení.

Cieľom tejto práce je preto navrhnúť a implementovať integrované dátové riešenie v cloudovom prostredí Microsoft Azure, ktoré prepojí zber dát, ich spracovanie, analytické metódy a vizualizačnú vrstvu do jednotného analytického celku. Navrhované riešenie je demonštrované na empirickom prípade analýzy verejných financií, pričom fiškálne a makroekonomické dáta slúžia ako aplikačný use-case na overenie funkčnosti, škálovateľnosti a rozšíriteľnosti riešenia.

Práca sa zameriava nielen na návrh technickej architektúry a implementáciu dátových tokov, ale aj na integráciu textových dát a analýzy sentimentu ako doplnkovej analytickej vrstvy, ktorá môže rozšíriť tradičnú kvantitatívnu analýzu verejných financií. Cieľom nie je samotné ekonomické hodnotenie fiškálnej politiky, ale demonštrácia praktickej aplikácie komplexného dátového riešenia, ktoré je možné adaptovať aj na iné analytické domény.

1 Výskumný problém a ciele v oblasti cloudových analytických riešení

Aj napriek tomu, že pokročilé analytické a dátové technológie sú široko dostupné a v súkromnom sektore vo veľkej miere adoptované, ich aplikácia v akademickej sfére a vo verejných financiách je stále pomerne zriedkavá a len málo zjednotená. Existujúce štúdie a praktické aplikácie na Business Intelligence sa primárne zameriavajú na analýzu založenú na statických alebo periodicky aktualizovaných dátach. V mnohých z týchto prípadov sa príprava dát a ich súvisiaca infraštruktúra spomínajú len okrajovo. Hlavným dôvodom je práve ten, že tieto štúdie bývajú zamerané hlavne na analýzu výsledkov a nie na samotný dátový proces.

Na podobnej úrovni sa nachádzajú aj samotné cloudové dátové riešenia, keďže akademická literatúra ostáva prevažne na konceptuálnej a architektonickej úrovni. Praktická realizácia týchto cloudových riešení, ktorá zahŕňa zber a integráciu dát, ich spracovanie, vizualizáciu a pokročilé analýzy, zostáva obmedzená. To následne vedie k nedostatku konkrétnych usmernení o tom, ako možno takéto riešenia implementovať v praxi.

Popri izolovanej praktickej realizácii Business Intelligence a cloudových riešení možno podobný problém pozorovať aj v oblasti analýzy sentimentu. Napriek tomu, že sentimentová analýza je aj vďaka rozvoju umelej inteligencie a LLM modelov v súčasnosti čoraz populárnejšia, akademické štúdie sa prevažne zameriavajú na konkrétne prípady použitia, porovnávanie modelov alebo jej teoretické aspekty. Len zriedkavo sú tieto metódy zakomponované do komplexných end-to-end analytických riešení alebo kombinované s kvantitatívnymi dátami, napríklad v kontexte verejných financií.

Ako už bolo naznačené v predchádzajúcom texte, táto fragmentácia poukazuje na významnú medzeru v akademickej literatúre, keďže existujúce výskumy sa sústreďujú prevažne na izolované riešenia alebo teoretické koncepty.

Cieľom tejto práce je z týchto dôvodov návrh a implementácia cloudového dátového riešenia realizovaného v prostredí Microsoft Azure, ktoré prepája pravidelné spracovanie dát, Business Intelligence nástroje a analýzu sentimentu v rámci jedného analytického rámca. Konkrétne ciele práce sú teda nasledovné:

- Navrhnúť a implementovať dátový tok na spracovanie a prípravu štruktúrovaných dát verejných financií v cloudovom prostredí Microsoft Azure.
- Vytvoriť interaktívny BI dashboard v nástroji Power BI, ktorý podporuje exploratívnu analýzu fiškálnych dát naprieč vybranými indikátormi a krajinami.
- Zakomponovať analýzu sentimentu textových dát ako doplnkovú analytickú dimenziu ku kvantitatívnym ukazovateľom.
- Demonštrovať použiteľnosť navrhnutého analytického rámca na prípadovej štúdii verejných financií.

2 Nedostatky v aplikácii nástrojov na podnikové spravodajstvo

V tejto kapitole sa zameriame na nedostatky v aplikácii nástrojov podnikovej inteligencie v akademickej aj aplikačnej praxi, so zameraním na tri kľúčové oblasti, ktoré sú v literatúre často riešené izolovane.

2.1 Izolovaný charakter nástrojov podnikovej inteligencie

V posledných rokoch zaznamenávajú výrazný nárast popularity nástroje typu self-service business intelligence (SSBI) [7], ktoré umožňujú aj netechnickým používateľom, či už zamestnancom z netechnickej oblasti a manažmentu, využívať podnikové spravodajstvo samostatným spôsobom, bez potreby technických odborníkov. [8].

V aplikačnej aj akademickej praxi sú tieto nástroje využívané predovšetkým na tvorbu dashboardov a realizáciu deskriptívnej analytiky, čo vyplýva z dostupných komparatívnych analýz a prípadových štúdií zameraných na nástroje business intelligence, ako sú PowerBI, Tableau a Qlik Sense [9, 10, 11, 12].

Hoci takéto zameranie podporuje rýchlu interpretáciu dát a zvyšuje dostupnosť analytických výstupov, praktická aplikácia SSBI nástrojov naráža na viaceré obmedzenia. Tieto limitácie sa týkajú najmä prípravy a vyhľadávania dát, práce s rôznorodými dátovými zdrojmi a hlavne tendencie k využívaniu jedného izolovaného nástroja na všetko, čo vedie k vzniku analytických riešení s obmedzenou integráciou a limitovanými možnosťami škálovania [7, 8].

2.2 Konceptuálny prístup ku cloudovým dátovým riešeniam

Identifikované obmedzenia v SSBI nástrojoch naznačujú, že problém nespočíva výlučne v samotnej vizualizačnej vrstve, ale v chýbajúcich dátových procesov a infraštruktúre, ktoré by podporovali integráciu rôznorodých dátových zdrojov a posilnilo by to škálovanie analytických riešení.

Toto sa ukázalo najmä v poslednom desaťročí, kedy došlo k výraznému nárastu objemu dostupných dát, ktorý bol ešte viac urýchlený pandémiou COVID-19 a s ňou spojenou akceleráciou digitalizačných procesov v súkromnej aj verejnej sfére [2, 3, 4]. Dôsledkom týchto zmien sa tradičné analytické prístupy, založené prevažne na manuálnych pracovných postupoch a nástrojoch podnikovej inteligencie orientovaných na reporting, začali ukazovať ako nepostačujúce, najmä v oblasti prípravy a integrácie dát [13, 6].

Tento vývoj viedol k zvýšeniu komplexnosti analytických požiadaviek a k potrebe spracovávať dáta v kratších časových intervaloch. Zároveň tak zosilnel dopyt po dátovo orientovaných riešeniach podporujúcich agilné a škálovateľné spracovanie dát, nízkolatenčné analytické procesy a automatizovanú orchestráciu dátových tokov [13, 6]. V reakcii na tieto požiadavky

sa mnohé organizácie začali orientovať na cloudovo založené prostredia, ktoré umožňujú efektívne spracovanie a analýzu narastajúceho množstva dát [5].

V dôsledku sa stali významnou témou aj v akademickej sfére, čo sa odráža v narastajúcom počte vedeckých článkov a štúdií zameraných na túto oblasť. Existujúca literatúra sa však prevažne sústreďuje na porovnávanie cloudových poskytovateľov alebo na návrh architektonických konceptov pre využitie cloudových služieb v dátovej analytike.

Tento prístup je ilustrovaný napríklad v práci Li et al. [14] a v štúdii od Borra [15], ktoré sa zameriavajú na porovnanie popredných cloudových platforiem, ako sú Microsoft Azure, Amazon Web Services a Google Cloud Platform, najmä z pohľadu dostupných služieb a ich technických charakteristík.

Medzi praktickejšie orientované práce v oblasti cloudových riešení pre podnikové spravodajstvo patrí štúdia Talakola [16], ktorá demonštruje kombináciu platformy Google Cloud Platform (GCP) s nástrojom Microsoft Power BI na troch aplikačných prípadoch. Hoci táto práca prezentuje konkrétne scenáre využitia cloudových služieb, jej hlavný dôraz je kladený na opis jednotlivých nástrojov, ktoré poskytuje GCP a architektonického návrhu prepojenia s Microsoft PowerBI. Implementačné aspekty end-to-end analytického riešenia, vrátane detailného spracovania dátových tokov, nie sú v práci rozpracované.

Podobný prístup, kombinujúci praktické aplikačné scenáre s prevažne architektonickým opisom riešenia, môžeme pozorovať aj v ďalších prácach zameraných na cloudové BI riešenia. Jednou z nich je riešenie, ktoré znázornil Chillara [17] vo svojej štúdii z roku 2025, v ktorej navrhuje cloudovo založenú BI architektúru pre analytické spracovanie a vizualizáciu dát v oblasti zdravotníctva. Navrhované riešenie zahŕňa integráciu približne dvadsiatich dátových zdrojov prostredníctvom ETL procesov, využitie dimenzionálneho modelovania v cloudovom dátovom sklade a aplikáciu pokročilých analytických metód vrátane strojového učenia (ML) a spracovania prirodzeného jazyka (NLP). Napriek tomuto širokému spektru použitých technológií zostáva prezentované riešenie spracované predovšetkým na architektonickej úrovni. Štúdia explicitne nešpecifikuje použitú implementáciu dátového skladu ani to, či ide o riešenie poskytované konkrétnym cloudovým providerom alebo o vlastnú implementáciu. Rovnako absentuje detailný opis implementačných krokov potrebných na realizáciu uceleného end-to-end analytického riešenia v reálnom prostredí.

2.3 Izolácia analytických úloh: Analýza sentimentu

Nedostatky identifikované pri využívaní SSBI nástrojov, najmä ich izolované postavenie v analytickom procese, ako aj prevažne konceptuálny charakter cloudových architektonických návrhov v akademickej literatúre, sa následne premietajú aj do spôsobu, akým sú riešené pokročilé analytické úlohy, vrátane spracovania textových dát a sentimentovej analýzy. Výskumné práce sa v tomto kontexte prevažne zameriavajú na porovnávanie výkonnosti jednotlivých moderných modelov a algoritmov na štandardizovaných dátasetoch, čo ilustruje

napríklad štúdia Krugmanna a Hartmanna [18], ktorá prezentuje benchmarking sentimentovej analýzy v dobe generatívnej AI, alebo porovnávacia analýza od Annan et al. [19], hodnotiaca účinnosť a výkonnosť veľkých jazykových modelov (LLM) v úlohách analýzy sentimentu a identifikácie váhavosti voči očkovaniu na dátach zo sociálnych médií. Takéto štúdie často poskytujú detailné porovnanie modelových architektúr a tréningových stratégií, avšak sentimentnú analýzu riešia prevažne ako samostatnú analytickú úlohu.

Okrem modelovo orientovaných porovnávacích štúdií sa v literatúre vyskytujú aj aplikačné prípadové štúdie zamerané na využitie sentimentovej analýzy v konkrétnych doménach [20, 21, 22]. Hoci tieto práce demonštrujú praktické použitie sentimentových modelov v reálnych scenároch, otázky ich systematickej integrácie do komplexných end-to-end dátových analytických riešení, zostávajú v dostupných štúdiách často mimo ich rozsahu.

2.4 Identifikované nedostatky v súčasnej aplikácii BI nástrojov

Aj napriek rozsiahlemu množstvu literatúry zameranej na nástroje typu SSBI, cloudové dátové služby a prístupy v oblasti analýzy sentimentu, poukazuje analýza existujúcich vedeckých prác a aplikačnej praxe na vzorec fragmentácie analytických riešení a nedostatok prakticky aplikovateľných postupov umožňujúcich ich replikáciu.

Nástroje podnikového spravodajstva sú často využívané ako izolovaná prezentačná vrstva bez hlbšej integrácie dátových procesov. Súvisí to predovšetkým s ich orientáciou na netechnických používateľov a obmedzenými možnosťami škálovania pri samostatnom nasadení. Cloudové dátové platformy sú v akademických prácach prevažne spracované na úrovni architektonických konceptov, pričom detailné implementačné aspekty end-to-end riešení zostávajú často mimo rozsahu skúmania.

Podobný prístup je možné pozorovať aj v oblasti analýzy sentimentu, kde sa výskum sústreďuje najmä na porovnávanie modelov, zatiaľ čo otázky ich systematickej integrácie do dátových systémov a prepojenia s BI nástrojmi sú spravidla riešené len okrajovo. Z týchto vyššie uvedených dôvodov, považujem za nedostatok absenciu uceleného pohľadu na návrh a realizáciu komplexných dátových analytických riešení, ktoré by prepájali infraštruktúru, analytické metódy a vizualizačnú vrstvu do jedného konzistentného celku.

3 Teoretické východiská

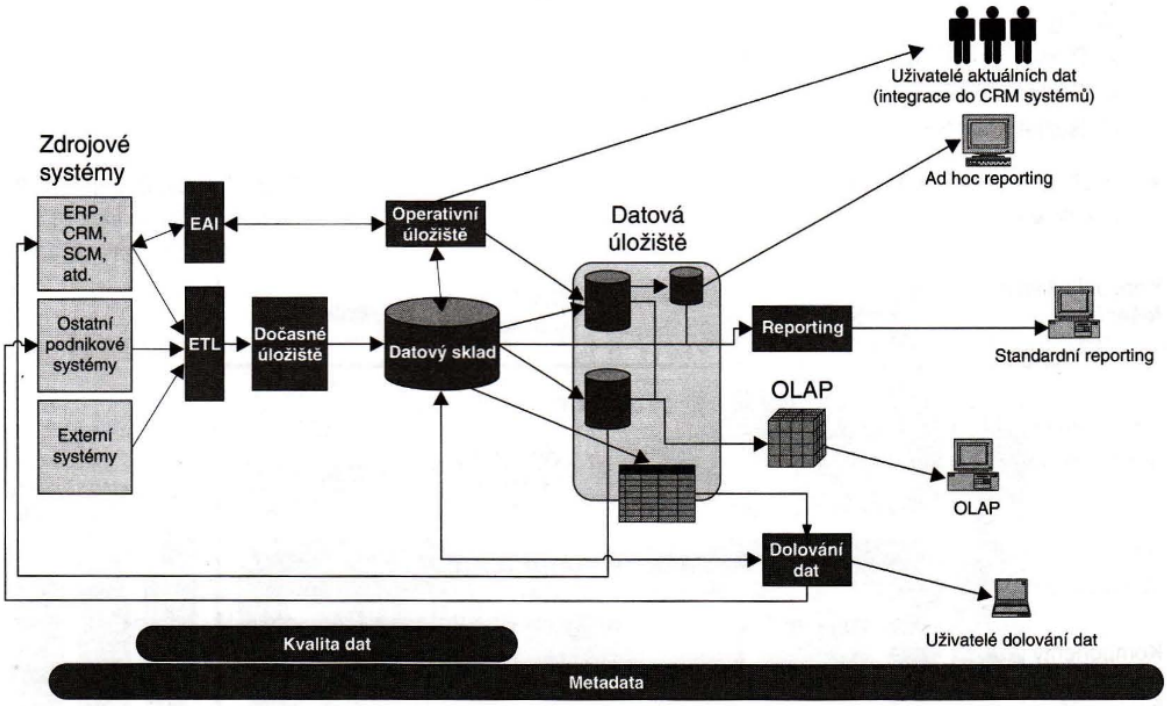
V tejto kapitole si objasníme základné pojmy a teoretické poznatky súvisiace s návrhom end-to-end riešenia. V úvode predstavíme pojem podnikovej inteligencie a self-service podnikovej inteligencie (SSBI). Ďalej sa zameriame na rôzne druhy dátových úložísk, ich charakteristiky a hlavné oblasti využitia. Nadviažeme opisom ETL a ELT dátových tokov, ako aj problematikou ich orchestrácie. V závere kapitoly sa budeme venovať cloud computingu, jeho porovnaniu s on-premise riešeniami a zodpovednostným modelom medzi používateľom a poskytovateľom služby. Kapitolu uzavrieme predstavením virtualizácie, ktorá je dôležitou súčasťou cloud computingu.

3.1 Podnikové spravodajstvo a self-service BI

Pojem podnikové spravodajstvo (business intelligence) je v odbornej literatúre definovaný rôznymi spôsobmi. Vo svojej podstate však predstavuje súbor procesov, technológií, konceptov a aplikácií informačného systému, ktorých cieľom je podporiť a uľahčiť rozhodovanie v podniku [7, 8, 23, 24]. Ako už bolo naznačené, tak v business intelligence nejde len o tvorbu dashboardov, ale o komplexný analytický proces, ktorý zahŕňa zber dát, ich integráciu z rôznych zdrojov, transformáciu, uloženie, analytické spracovanie a následne prezentácia výsledkov koncovým užívateľom. Túto skutočnosť znázorňuje aj obrázok 1 z knihy od Ota Novotného, et al. [24], na ktorom je možné vidieť prepojenie zdrojových systémov, ETL procesov, dátového skladu, dátových úložísk a analytických nástrojov, ako sú reporting, OLAP alebo dolovanie dát. V praxi spoločnosti využívajú systémy business intelligence najmä na získanie konkurenčnej výhody a zvýšenie svojej výkonnosti a ich význam narastá najmä pri rastúcom množstve údajov [25, 7, 24].

Pod vyššie uvedenou definíciou rozumieme najmä tradičné systémy podnikového spravodajstva. Tento prístup umožňuje vysokú mieru kontroly nad dátami, avšak v súčasnosti BI špecialisti často narážajú na úzke miesto pri spracovaní narastajúceho počtu požiadaviek od používateľov. Z tohto dôvodu vznikol koncept self-service business intelligence (SSBI), ktorý umožňuje aj netechnickým používateľom samostatne vytvárať vlastné reporty, interaktívne dashboardy a analytické pohľady nad dátami [8].

Passlick et al. [26] opísal podstatu self-service BI vo svojej štúdii v postupnom zvyšovaní miery samostatnosti používateľov pri práci s dátami, ako môžeme vidieť na obrázku 2. Na základnej úrovni používatelia najmä pristupujú k existujúcim reportom a využívajú pripravené informácie. Na vyššej úrovni už dokážu vytvárať vlastné reporty, využívať analytické funkcie a podieľať sa tak na tvorbe informácií. Najvyššia úroveň self-service BI zahŕňa aj prácu s novými dátovými zdrojmi, ich kombinovanie a prípravu vlastných analytických výstupov. Tento prístup tak znižuje závislosť používateľov od IT alebo BI oddelenia a podporuje dátovo orientované rozhodovanie v organizácii.



Obr. 1: Architektúra tradičného systému business intelligence [24]

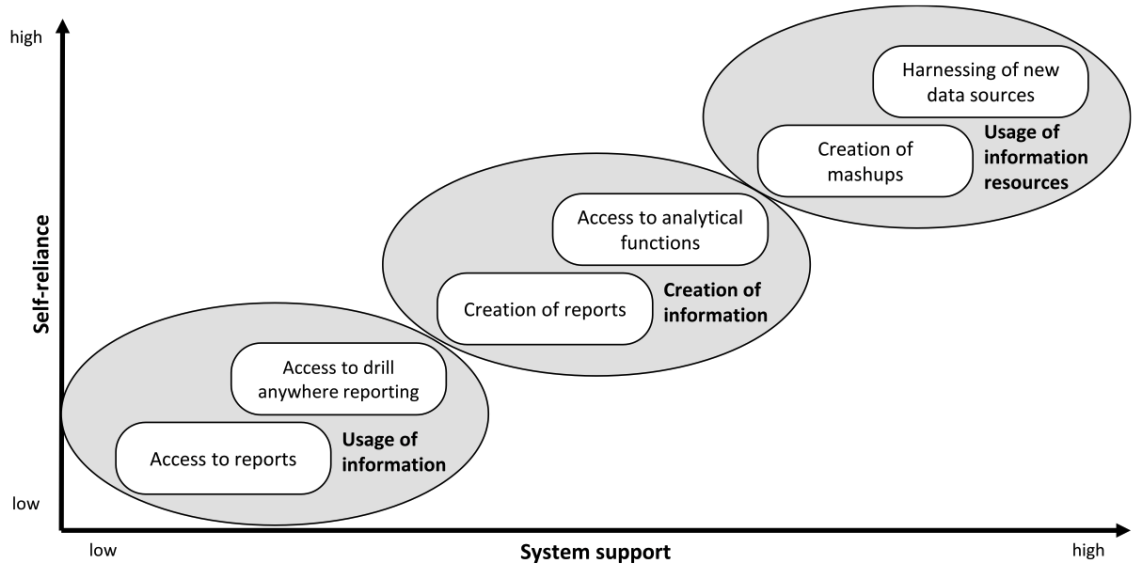
3.2 Dátové úložiská a spracovanie dát

Moderné analytické riešenia sa neopierajú iba o samotnú vizualizačnú vrstvu, ale vyžadujú aj vhodne navrhnutú dátovú infraštruktúru. Tá zabezpečuje ukladanie, organizáciu, transformáciu a sprístupnenie dát pre ďalšie analytické využitie. V závislosti od charakteru spracovávaných dát a požiadaviek analytického riešenia sa v praxi využívajú rôzne typy dátových úložísk, medzi ktoré patria najmä dátové sklady, dátové jazerá a novšie koncepty dátového lakehouse [27, 28]. S týmito úložiskami úzko súvisia aj procesy ETL a ELT, ktoré opisujú spôsob presunu a transformácie dát zo zdrojových systémov do špecifických dátových štruktúr [24]. Pri komplexnejších riešeniach je zároveň potrebné tieto procesy riadiť a automatizovať prostredníctvom orchestrácie dátových tokov [29].

3.2.1 Dátový sklad, dátové jazero a dátový lakehouse

Dátový sklad, označovaný ako data warehouse, predstavuje centralizované úložisko určené predovšetkým na spracovanie štruktúrovaných dát [27]. V literatúre sa dátové sklady často spájajú s prvou generáciou analytických dátových platforiem, ktoré boli navrhnuté najmä na podporu podnikového reportingu a rozhodovania [28]. Údaje v dátovom sklade sú ukladané podľa vopred definovanej schémy a spravidla nie sú ručne vkladané, ale získavané, integrované a transformované zo zdrojových systémov. Dátový sklad je zároveň charakteristický tým, že je subjektovo orientovaný, integrovaný, časovo premenlivý a nemenný, čo umožňuje jeho využitie na podnikové reportovanie a analytické spracovanie [27, 24].

Na rozdiel od dátového skladu je dátové jazero navrhnuté ako flexibilnejšie úložisko,



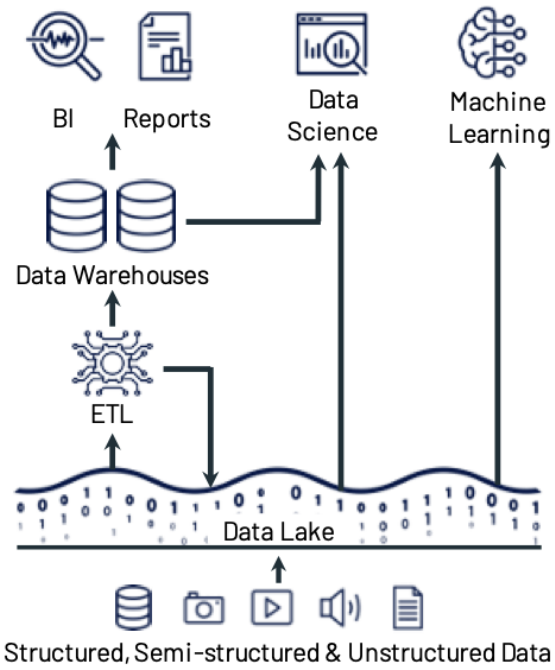
Obr. 2: SSBI úrovne [26]

ktoré okrem štruktúrovaných dát podporuje aj iné formáty údajov [27]. Môže ísť o polo-štruktúrované súbory, ako napríklad JSON alebo XML, ale aj neštruktúrované dáta, vo forme textov, dokumentov, alebo obrázkov. Dátové jazero v praxi často slúži ako centrálne miesto na ukladanie surových dát prichádzajúcich z rôznych zdrojov, čím podporuje centralizáciu dát a ich následné analytické využitie [27]. V moderných dátových architektúrach sa preto často využíva kombinácia dátového jazera a dátového skladu, ako je znázornené aj na obrázku 3 podľa štúdie Armbrust et al. [28]. V takomto prístupe sú dáta zo zdrojových systémov najskôr ukladané do dátového jazera a následne transformované a sprístupnené v dátovom sklade pre analytické a reportovacie účely.

Koncept data lakehouse vznikol ako snaha prepojiť vlastnosti dátového jazera a dátového skladu v rámci jednotnej architektúry. Jeho cieľom je zachovať flexibilitu dátového jazera pri ukladaní rôznorodých dát a zároveň doplniť vlastnosti typické pre dátové sklady, ako je podpora dátových schém, riadenie kvality dát a ACID transakcie [28]. Lakehouse architektúra tak umožňuje pracovať s rôznymi typmi dát v jednom úložisku a zároveň ich efektívne sprístupňovať pre analytické spracovanie, reporting alebo pokročilé analytické metódy. V praxi sa tento prístup často implementuje pomocou špeciálnych formátov, ako je Parquet, Delta Lake alebo Apache Iceberg [28].

3.2.2 ETL vs ELT procesy

Procesy ETL a ELT úzko súvisia s ukladaním a prípravou dát v dátových úložiskách, ako sú dátové sklady, dátové jazerá alebo data lakehouse riešenia. ETL a ELT predstavujú dva základné prístupy k presunu dát zo zdrojových systémov do cieľového prostredia [30]. Hoci sú si tieto prístupy svojou podstatou podobné, líšia sa najmä poradím vykonávania jednotlivých krokov a miestom, kde dochádza k transformácii dát.



Obr. 3: Architektúra s DL a DWH [28]

Skratky ETL a ELT pozostávajú z troch základných krokov. Extract predstavuje extrakciu, teda získavanie údajov zo zdrojových systémov. Týmito zdrojmi môžu byť napríklad informačné systémy typu ERP, CRM, prípadne jednoduché súbory. Transform označuje transformáciu údajov, ktorá zahŕňa čistenie, normalizáciu, agregáciu a obohacovanie. a Posledný krok Load, predstavuje načítanie alebo uloženie dát do cieľového prostredia [30].

Ako môžeme pozorovať, základný rozdiel medzi ETL a ELT spočíva v poradí týchto krokov. Pri ETL sú dáta najskôr extrahované zo zdrojových systémov, následne transformované mimo cieľového úložiska a až potom načítané do dátového skladu alebo iného analytického systému. Naopak pri procese ELT sú dáta po extrakcii najskôr načítané do cieľového úložiska a transformácie sa vykonávajú až následne priamo v tomto prostredí [30].

Štúdia od D. Seenivasan [30] uvádza porovnanie ETL a ELT procesov z hľadiska výkonnosti, škálovateľnosti a nákladov, ktoré je zhrnuté v tabuľke 1.

Z porovnania môžeme sledovať, že ETL prístup môže byť pomalší, keďže transformácie prebiehajú ešte pred načítaním dát do cieľového úložiska. Jeho škálovateľnosť je zároveň do väčšej miery závislá od kapacity transformačného nástroja. Naopak, ELT umožňuje rýchlejšie načítanie surových dát a následné vykonanie transformácií priamo v cieľovej dátovej platforme. Tento prístup preto lepšie využíva výpočtové možnosti moderných dátových skladov, dátových jazier alebo lakehouse riešení. Z pohľadu nákladov môže mať ELT nižšie počiatočné náklady, avšak v závislosti od objemu spracovania a využitia výpočtových zdrojov môžu byť jeho prevádzkové náklady vyššie.

Tabuľka 1: Porovnanie ETL a ELT procesov [30]

Kritérium	**ETL**	**ELT**
Výkonnosť	Pomalší prístup z dôvodu vykonávania transformácií pred načítaním dát.	Rýchlejší prístup vďaka priamemu načítaniu surových dát do cieľového úložiska.
Škálovateľnosť	Obmedzená kapacitou transformačného nástroja alebo transformačnej vrstvy.	Škáluje sa podľa možností cieľovej dátovej platformy alebo dátového skladu.
Náklady	Vyššie počiatočné náklady na nastavenie a následnú údržbu.	Nižšie počiatočné náklady, avšak potenciálne vyššie prevádzkové náklady.

3.2.3 Orchestrácia dátových tokov

Orchestrácia dátových tokov úzko súvisí aj s procesmi ETL a ELT, ktoré predstavujú konkrétne formy dátových pipeline. Využívanie takýchto tokov zvyšuje efektívnosť dátových analýz, ale zároveň vyžadujú podporu pre dizajn a manažment [29]. V súčasnosti je orchestrácia dátových tokov dôležitou súčasťou moderných dátových riešení [31], keďže umožňuje riadiť tok informácií medzi jednotlivými systémami a úlohami.

Pod pojmom orchestrácia dátových tokov rozumieme automatizované spúšťanie, časovanie a koordináciu úloh v rámci dátového procesu [31]. Kľúčovou vlastnosťou orchestrácie je modelovanie procesov vo forme DAG (Directed Acyclic Graph), teda orientovaného acyklického grafu, kde sú jednotlivé úlohy reprezentované ako vrcholy a závislosti medzi nimi ako hrany [31, 32]. Takýto model umožňuje definovať počiatočný krok dátového toku, následnosť jednotlivých úloh a podmienky jeho ukončenia.

V praxi orchestrácia dátových tokov často zahŕňa automatizáciu zberu dát, ich transformácie, validácie, monitorovania priebehu spracovania a ošetrenia prípadných chýb. Automatizácia týchto činností znižuje potrebu manuálnych zásahov, obmedzuje riziko ľudskej chyby a zvyšuje opakovateľnosť celého dátového procesu [31].

3.3 Cloud computing

V tejto sekcii sa zameriame na pojem cloud computing a porovnáme ho s lokálnym, teda on-premise riešením. Následne vysvetlíme jednotlivé modely zodpovednosti medzi používateľom a poskytovateľom cloudových služieb. V závere sekcie predstavíme pojem virtualizácie, ktorá patrí medzi základné technologické princípy cloud computingu.

3.3.1 Cloud vs On-premise riešenie

Pod pojmom cloud computing rozumieme model poskytovania výpočtových zdrojov a služieb prostredníctvom internetu [33]. Cloud computing používateľom sprístupňuje výpočtové zdroje, ako sú servery, úložiská, databázy, aplikácie alebo sieťové služby, a to na požiadanie

podľa ich aktuálnych potrieb [34]. Charakteristickou vlastnosťou cloudového prostredia je, že používatelia nemusia vlastniť ani priamo spravovať fyzickú infraštruktúru, na ktorej sú tieto služby prevádzkované. Namiesto toho využívajú zdroje poskytovateľa cloudových služieb, ktoré môžu byť dynamicky prideľované, rozširované alebo uvoľňované [33].

Cloud computing tak predstavuje alternatívu k tradičnému on-premise prístupu, pri ktorom organizácia vlastní a prevádzkuje potrebnú infraštruktúru vo vlastnom prostredí. Kým v prípade on-premise riešenia je organizácia zodpovedná za obstaranie, údržbu, aktualizácie a rozširovanie hardvéru aj softvéru, pri cloudovom riešení je značná časť tejto zodpovednosti presunutá na poskytovateľa cloudových služieb. Rozdiel medzi týmito dvoma prístupmi preto nespočíva iba v umiestnení infraštruktúry, ale aj v spôsobe jej správy, financovania a škálovania.

3.3.2 Zodpovednostné modely cloud computingu

Vzhľadom na mieru zodpovednosti, ktorá je presunutá z používateľa na poskytovateľa cloudových služieb, rozdeľujeme cloudové služby do troch základných modelov, ktoré možno pozorovať aj v tabuľke 2.

Tabuľka 2: Zodpovednosť poskytovateľa pri jednotlivých modeloch nasadenia a cloudových služieb

Vrstva / služba	**On-premise**	**IaaS**	**PaaS**	**SaaS**
Fyzická infraštruktúra		✓	✓	✓
Sieť a úložisko		✓	✓	✓
Virtualizácia		✓	✓	✓
Operačný systém			✓	✓
Runtime prostredie			✓	✓
Middleware			✓	✓
Aplikácia				✓
Dáta				

Pozn.: Symbol ✓ označuje zodpovednosť poskytovateľa služby.

Prvým modelom je **infraštruktúra ako služba** (Infrastructure as a Service – IaaS). V porovnaní s lokálnym, tzv. on-premise riešením zbavuje používateľa zodpovednosti za fyzickú infraštruktúru, ako sú dátové centrá, servery, sieťové prvky či úložiská. Poskytovateľ tak používateľovi sprístupňuje virtuálnu infraštruktúru v cloudovom prostredí. Používateľ alebo vývojár však naďalej zodpovedá za inštaláciu a správu operačných systémov, podporného softvéru, aplikácií a dát [35, 33]. Príkladmi tohto modelu sú Azure Virtual Machines, Amazon EC2 a Google Compute Engine.

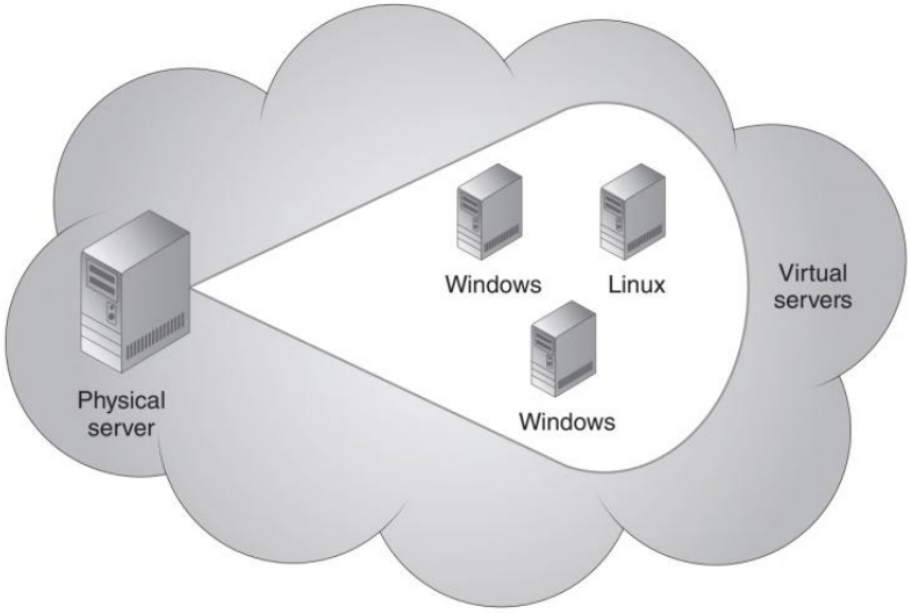
Druhým modelom je **platforma ako služba** (Platform as a Service – PaaS). Tento model okrem fyzickej infraštruktúry odbremeňuje používateľa aj od správy operačného systému, runtime prostredia, middleware a ďalších platformových komponentov. Vývojárom tak umožňuje sústrediť sa predovšetkým na vývoj a nasadzovanie vlastných aplikácií bez potreby

zaoberať sa správou hardvéru alebo aktualizáciami operačného systému. Používateľ v tomto modeli zodpovedá najmä za svoju aplikáciu, dáta a prístupové práva [35, 33]. Príkladom služby typu PaaS je Azure App Service.

Posledným modelom je **softvér ako služba** (Software as a Service – SaaS). Tento model poskytuje používateľovi hotovú softvérovú aplikáciu dostupnú prostredníctvom internetu. Používateľovi zvyčajne stačí webový prehliadač alebo klientska aplikácia, prostredníctvom ktorej k službe pristupuje. V tomto modeli používateľ zodpovedá najmä za konfiguráciu aplikácie, správu používateľských účtov a prístupových práv. Nevýhodou môže byť menšia miera kontroly nad samotnou aplikáciou a jej technickým prostredím [35, 33]. Príkladmi služieb typu SaaS sú Microsoft 365 a Gmail.

3.3.3 Virtualizácia

Dôležitým aspektom cloud computingu je práve virtualizácia, keďže cloud computing sa na ňu vo veľkej miere spolieha [34]. Pod pojmom virtualizácia možno rozumieť využitie hardvéru alebo softvéru na vytvorenie abstraktnej, virtuálnej reprezentácie fyzických výpočtových zdrojov. Inými slovami, fyzická infraštruktúra môže byť rozdelená na viacero samostatných virtuálnych prostredí. Jamsa [33] uvádza príklad servera, ktorý má procesor (CPU) schopný spúšťať konkrétny operačný systém, napríklad Windows alebo Linux. Pomocou špeciálneho softvéru však môže byť tento server nakonfigurovaný tak, aby sa používateľovi javil ako systém s viacerými procesormi, na ktorých môžu bežať rovnaké alebo rozdielne operačné systémy. Tento príklad je taktiež znázornený na obrázku 4 nižšie. Takýto princíp umožňuje efektívnejšie využívanie fyzických zdrojov a ich poskytovanie viacerým používateľom alebo aplikáciám v izolovaných virtuálnych prostrediach.



Obr. 4: Virtualizácia - príklad [33]

4 Vstupné dátové zdroje

Keďže praktická časť tejto práce sa nezameriava výlučne na návrh a implementáciu dátového riešenia v cloudovom prostredí, ale aj na jeho aplikáciu na konkrétnom aplikačnom prípade, táto časť poskytuje prehľad dát, ktoré predstavujú vstup do navrhovaného systému, spolu s opisom postupov použitých v úvodnej fáze riešenia.

4.1 Fiškálne dáta a ich spracovanie

Počas úvodnej fázy tejto práce sme nezačali priamo implementáciou infraštruktúry v cloude, keďže sme ešte nevedeli čo ideme spracovávať a čo má byť výstupom. Najskôr sme preto vykonali výber a predbežnú analýzu dostupných dátových zdrojov s cieľom overiť ich vhodnosť a dostatočnosť pre plánovanú empirickú štúdiu.

Empirická analýza vykonaná v tejto práci sa zameriava predovšetkým na štruktúrované makroekonomické a fiškálne data, ktoré sme obdržali prostredníctvom webového portálu Eurostat. Eurostat ponúka harmonizované a vysoko kvalitné štatistické dáta pre všetky členské štáty Európskej Únie, čo z neho robí vhodného kandidáta na porovnávanie verejných financií a makroekonomickej analýzy medzi krajinami. Ďalšou výhodou využitia dát z portálu Eurostat je ich konzistentnosť v porovnaní s národnými dátovými zdrojmi. Národné inštitúcie často aktualizujú svoje databázy v kratších časových intervaloch, čo môže viesť k častejším revíziám a zmenám historických hodnôt. Použitie harmonizovaných dát Eurostatu sa tak ukázali byť najvhodnejšie a umožňujú tak stabilnejší základ pre analytické spracovanie a porovnanie výsledkov.

4.1.1 Vybrané dátasety pre fiškálnu analýzu

Štruktúrované makroekonomické a fiškálne dáta sme extrahovali prostredníctvom webového rozhrania portálu Eurostat, ktoré umožňuje export dátových súborov na základe aktuálne aplikovaných filtrov. Pred samotným stiahnutím dátového súboru boli definované potrebné filtre zahŕňajúce dimenzie ako krajina, časové obdobie, položky národných účtov a sezónne úpravy, ktoré môžeme bližšie vidieť na obrázku 4 v prílohe. Po nastavení požadovaných filtrov boli dáta exportované vo formáte TSV (Tab-Separated Values). Uvedený postup nám platil pre všetky dátové súbory až na výnimku datasetu *ds-045409*, pri ktorom by manuálne filtrovanie jednotlivých HS kódov predstavovalo neefektívny a časovo náročný proces. HS kódy reprezentujú medzinárodne štandardizovaný číselný systém na klasifikáciu tovarov. Z tohto dôvodu sme vytvorili jednoduchý skript v jazyku Python, ktorý prostredníctvom prispôsobenej URL adresy rozhrania API získal, načítal a uložil požadované dáta do formátu TSV.

Všetky vybrané dátové súbory získané z databázy Eurostat sú nasledovné:

- `gov_10q_ggnfa` – Štvrťročné nefinančné účty sektora verejnej správy

- `gov_10dd_edpt1` – Údaje o deficite a dlhu verejnej správy
- `gov_10q_ggdebt` – Štvrťročné štatistiky dlhu verejnej správy
- `gov_10a_exp` – Ročné výdavky verejnej správy
- `gov_10a_main` – Hlavné agregáty financovania verejnej správy
- `gov_10a_taxag` – Ročné agregáty daní a sociálnych príspevkov
- `nama_10_gdp` – Hrubý domáci produkt a agregáty národných účtov
- `prc_hicp_manr` – Harmonizovaný index spotrebiteľských cien (HICP)
- `une_rt_q` – Štvrťročná miera nezamestnanosti
- `ds-045409` – Údaje o medzinárodnom obchode (dovoz a vývoz)

Zvolené dátasety nám tak pokrývajú kľúčové dimenzie verejných financií, makroekonomickej výkonnosti, cenového vývoja a podmienok na trhu práce.

4.1.2 Aplikácia filtrov

Ako už bolo uvedené v predchádzajúcej časti (pozri sekciu 4.1.1), dátové súbory sme pred ich uložím z portálu Eurostat museli vyfiltrovať, aby sa predišlo zbytočnému načítavaniu veľkého objemu dát. Zároveň sme tým vopred definovali dátové úpravy, ktoré budú aplikované aj pri návrhu dátového toku v cloudovej infraštruktúre.

Jedným z hlavných aplikovaných filtrov, ktorý sme na väčšiu polovicu súborov naniesli, je výber inštitucionálneho sektora. Na výber sme mali s rôznych podskupín, ako napríklad podsektor ústrednej štátnej správy (S1311) a podsektor štátnej správy na úrovni regiónov (S1312). Keďže cieľom analýzy nebolo ísť do detailnej dekompozície jednotlivých úrovní verejnej správy pri porovnávaní štátov, zvolili sme agregovaný sektor verejnej správy (S13), ktorý zahŕňa uvedené a aj ďalšie podsektory.

Druhým dôležitým filtrom boli položky národných účtov (*national accounts items*), ktoré inými slovami reprezentovali jednotlivé makroekonomické a fiškálne ukazovatele. Tento filter bol aplikovaný v závislosti od konkrétneho dátového súboru a spravidla určoval presný ukazovateľ, ktorý nás z daného dátasetu zaujímal v rámci analýzy. Jednalo sa napríklad o príjmy, výdavky, dlhy alebo iné súvisiace veličiny.

Na uľahčenie porovnateľnosti medzi jednotlivými krajinami sme ukazovatele vybrali v štandardizovanej jednotke percenta hrubého domáceho produktu (HDP). Tento prístup nám umožnil efektívnejšie porovnávanie ekonomík s rozdielnou veľkosťou, úrovňou príjmov a výdavkov, čím zároveň eliminoval skreslenie spôsobené absolútnymi hodnotami makroekonomických ukazovateľov.

Ďalším z aplikovaných filtrov bola sezónna úprava dát. Táto bola však k dispozícii len pre 4 dátové súbory. V prípade miery nezamestnanosti sme vybrali sezónne očistené údaje (*seasonally adjusted, SA*), keďže tento ukazovateľ vykazoval sezónne výkyvy, ktoré by mohli skresľovať krátkodobé porovnania. Naopak, pri fiškálnych ukazovateľoch, konkrétne pri deficite, príjmoch, výdavkoch a dlhu verejnej správy, sme použili neočistené údaje (*non-seasonally adjusted, NSA*). Tento výber sme zvolili z dôvodu, že použitie neočistených údajov nám umožňuje konzistentné porovnanie fiškálnych ukazovateľov v čase medzi jednotlivými krajinami.

Okrem filtrovania na úrovni jednotlivých ukazovateľov, ako bolo uvedené v predchádzajúcej časti, bolo počas úvodnej fázy zberu dát možné definovať aj periódu a krajiny, pre ktoré sme dáta chceli brať. Keďže cieľom počiatočnej fázy bola exploračná analýza, dátové súbory sme extrahovali so všetkými dostupnými časovými periódami a krajinami.

V samotnej analytickej časti práce sa však pozornosť sústreďuje predovšetkým na vybrané skupiny krajín, konkrétne na štáty V4 a Rakúsko, a to za obdobie rokov 2015 až 2024, resp. do druhého štvrťroka roku 2025 (2025Q2), keďže novšie údaje neboli v čase spracovania dostupné.

Zaujímavým aspektom analyzovaných dát bola prítomnosť metadátových označení pri niektorých hodnotách ukazovateľov. Tieto príznaky indikovali napríklad predbežné hodnoty (*p*) alebo chýbajúce údaje, resp. situácie, v ktorých hodnota nemôže existovať (*m*). Zachovanie týchto označení nám umožní zohľadniť kvalitu dát pri interpretácii výsledkov a identifikovať možné obmedzenia v časovej porovnateľnosti jednotlivých ukazovateľov.

4.2 Textové dáta pre analýzu sentimentu

Ešte pred samotným návrhom dátového riešenia bolo potrebné identifikovať vhodné texty, ktoré by dokázali tvoriť základ pre našu analýzu sentimentu. Súčasne bolo cieľom určiť, v akej forme budú tieto texty dostupné a aké budú ich hlavné charakteristiky z pohľadu ďalšieho spracovania.

4.2.1 Ekonomické prognózy

Ako prvý zdroj textov pre našu analýzu sentimentu sme zvolili ekonomické prognózy Európskej komisie (EC), ktoré sú verejne dostupné na oficiálnej stránke inštitúcie pod nasledujúcim odkazom:

`https://economy-finance.ec.europa.eu/economic-forecast-and-surveys/economic-forecasts_en`

Tieto prognózy vyjadrujú očakávaný vývoj kľúčových makroekonomických ukazovateľov, ako sú rast HDP, inflácia, nezamestnanosť a verejné financie, a poskytujú tak ideálny doplnok pre našu analýzu. Dokumenty sú typicky publikované vo forme štruktúrovaných

textov, najčastejšie vo formáte HTML, v niektorých prípadoch aj ako doplnok vo formáte PDF, a vyznačujú sa vyššou mierou odbornosti a formalizovaného jazyka.

Z pohľadu analýzy sentimentu predstavujú tieto prognózy vhodný zdroj najmä preto, že obsahujú kvalitatívne hodnotenia budúceho ekonomického vývoja a formulácie o rizikách, neistotách a očakávaných trendoch. Práve tieto jazykové konštrukcie umožňujú identifikovať pozitívne, neutrálne aj negatívne komunikačné rámce, čím explicitne odrážajú ekonomickú náladu a očakávania odborných inštitúcií v strednodobom až dlhodobom horizonte.

4.2.2 Tlačové správy

Druhým identifikovaným zdrojom sú tlačové správy Európskej centrálnej banky (ECB), ktoré obsahujú najmä oficiálne vyhlásenia o menovej politike, rozhodnutiach o úrokových sadzbách, inflácii, ekonomickom výhľade a ďalších témach súvisiacich s vývojom v eurozóne. Na rozdiel od ekonomických prognóz Európskej komisie majú tlačové správy ECB kratší rozsah a sú publikované častejšie, čo umožňuje zachytiť aktuálny vývoj a reakcie na meniace sa ekonomické podmienky.

Z pohľadu analýzy sentimentu sú tieto texty cenné najmä preto, že zachytávajú postoj centrálnej banky k aktuálnej ekonomickej situácií a vyjadrujú tak krátkodobý pohľad na ekonomický sentiment v porovnaní s prognózami z EC.

5 Návrh a implementácia cloudovej dátovej architektúry

V tejto kapitole si opíšeme postup, ako sme riešili skúmanú problematiku. Opisujeme proces návrhu cloudovej dátovej architektúry, výber technologických nástrojov a postup implementácie riešenia v prostredí MS Azure.

5.1 Metodologické východiská návrhu cloudového dátového riešenia

5.1.1 Komplexnosť návrhu cloudovej dátovej architektúry

Návrh a voľba vhodného cloudového dátového riešenia si vyžaduje systematický metodologický prístup, ktorý zohľadňuje viacero technologických, architektonických a prevádzkových aspektov. Pri návrhu takejto architektúry je nevyhnutné aby sme analyzovali charakter spracovávaných dát, a to najmä s ohľadom na ich štruktúru, periodicitu aktualizácie a spôsob využitia. Rozlišovať môžeme medzi viacerými typmi dát, a to napríklad jednorazovými dátovými sadami, pravidelne aktualizovanými údajmi, udalostnými (event-based) dátami, či pomaly sa meniacimi dimenzionálnymi údajmi. Tieto charakteristiky dát zásadne ovplyvňujú spôsob a prístup, akým navrhujeme dátové toky a architektúru úložných vrstiev.

Rovnako významnú úlohu zohrávajú prevádzkové a bezpečnostné požiadavky systému, ako je ochrana dát, riadenie prístupových práv, nastavenie retenčných politík, archivácia dát či zabezpečenie kontinuity prevádzky v prípade zlyhania alebo výpadku systému. Dôležitým faktorom je tiež plánované využitie dát, či už na analytické účely, reporting, alebo ako vstup pre modely umelej inteligencie.

Z vyššie uvedeného môžeme tušiť, že návrh cloudovej dátovej architektúry nie je jednoznačný proces, ale vyžaduje si zohľadnenie rôznorodých požiadaviek.

5.1.2 Odôvodnenie voľby cloudového riešenia

Základné možnosti tvorby dátového riešenia spočívajú v realizácii buď v cloudovom, alebo v lokálnom (on-premise) prostredí, ktoré sme bližšie opísali v časti 3.3.1. Výber vhodného prostredia súvisí s viacerými technologickými, prevádzkovými a ekonomickými aspektami, ktoré boli čiastočne naznačené už v predchádzajúcej časti.

Hoci môže lokálne riešenie pôsobiť na prvý pohľad atraktívne, najmä z dôvodu jeho nezávislosti od externého poskytovateľa a na prvý pohľad nižších priebežných nákladov, je potrebné sa zamyslieť, čo sa snažíme vybudovať. Lokálne riešenie so sebou prináša viacero významných výziev. Jednou z nich je potreba zabezpečenia dostatočne výkonného hardvérového vybavenia, ktoré by dokázalo pokryť nároky spojené s transformáciou dát, ich agregáciou a prípadným výpočtovým spracovaním modelov sentimentovej analýzy.

V prípade pravidelnej aktualizácie dát je zároveň nevyhnutné, aby bolo zariadenie nepretržite dostupné a prevádzkyschopné. To znamená nielen počiatočnú investíciu do infraštruktúry,

ale aj zabezpečenie jej kontinuálnej prevádzky, údržby a správy. Pri lokálnom riešení je navyše potrebné vo vyššej miere zohľadňovať fyzickú bezpečnosť zariadenia, ochranu pred kybernetickými hrozbami, ako aj zálohovanie a obnovu dát v prípade zlyhania systému.

Naopak, cloudové riešenie umožňuje flexibilne prispôsobovať výpočtové zdroje aktuálnym potrebám bez nutnosti počiatočnej kapitálovej investície do hardvéru. Zároveň poskytuje natívne mechanizmy pre automatizáciu procesov, škálovanie výkonu a zabezpečenie dostupnosti služby. V kontexte nášho riešenia, ktoré zahŕňa pravidelné spracovanie viacerých dátových zdrojov, integráciu štruktúrovaných aj neštruktúrovaných dát a ich následné analytické využitie, sa preto cloudové prostredie javilo ako vhodnejšia a udržateľnejšia alternatíva.

Dôležitým faktorom bola aj potreba minimalizovať manuálne zásahy do procesu spracovania dát a zabezpečiť reprodukovateľnosť analytických výstupov. Cloudové prostredie nám tak umožňuje implementovať automatizované dátové toky, centrálne riadiť spracovanie dát a efektívne oddeliť transformačnú logiku od prezentačnej vrstvy v nástroji Microsoft Power BI.

5.1.3 Zvolenie platformy MS Azure

Pri výbere cloudovej platformy sme zvažovali troch hlavných poskytovateľov, a to Microsoft Azure, Amazon Web Service (AWS) a Google Cloud Platform (GCP). Každá z uvedených platforiem disponuje špecifickými silnými stránkami. AWS je napríklad známe širokým portfóliom služieb a vysokou mierou flexibility pri budovaní komplexných infraštruktúrnych riešení. GCP je často preferovaný v oblasti pokročilej analytiky a spracovania veľkých dátových objemov, najmä v kontexte nástrojov orientovaných na dátovú vedu a strojové učenie.

Pri rozhodovaní sme však nezohľadňovali iba šírku ponúkaných služieb, ale najmä ich integrácie, kompatibilitu s existujúcimi nástrojmi a schopnosť vytvoriť ucelené dátové prostredie podporujúce automatizované spracovanie dát. V tomto kontexte sa ako najvhodnejšia platforma javila Microsoft Azure.

Významným faktorom bola najmä jeho prirodzená integrácia s nástrojom Microsoft Power BI, ktorý bol zvolený ako prezentačná a vizualizačná vrstva riešenia. Azure zároveň poskytuje jednotné prostredie pre návrh, implementáciu a orchestráciu dátových pipeline procesov, čo nám umožnilo efektívne oddeliť transformačnú logiku od analytickej a vizualizačnej vrstvy.

Ďalším rozhodujúcim aspektom bola dostupnosť natívnych služieb pre ukladanie dát, dávkové spracovanie a riadenie dátových tokov v rámci jedného ekosystému. Takéto integrované prostredie zjednodušuje správu riešenia, zvyšuje jeho transparentnosť a umožňuje jeho budúce rozšírenie bez potreby zásadnej zmeny technologickej platformy. Z týchto vyššie uvedených dôvodov, považujem za najvhodnejšiu platformu pre naše použitie práve Microsoft Azure.

5.1.4 Výber hlavných cloudových služieb MS Azure

Po zvolení platformy Microsoft Azure bolo potrebné určiť konkrétne cloudové služby, ktoré budú tvoriť technologický základ navrhovaného riešenia. Výber služieb vychádzal najmä z požiadaviek na ukladanie dát, orchestráciu dátových tokov, dávkové spracovanie vlastných skriptov a následné sprístupnenie dát analytickej vrstve.

Azure Storage Account

Ako prvé sa pozrieme na službu Azure Storage Account, ktorú sme zvolili ako centrálnu službu pre ukladanie dát. Azure Storage Account ponúka viacero možností ukladania dát, ako napríklad Blob Storage, File Storage alebo Queue Storage. V rámci nášho riešenia sme zvolili Azure Data Lake Storage Gen2 (ADLS), ktorý je podobný službe Blob Storage, ale na rozdiel od nej podporuje hierarchickú štruktúru adresárov a je optimalizovaný pre analytické scenáre. Táto vlastnosť je v našom prípade užitočná, pretože umožňuje logické aj fyzické oddelenie dát v rámci jednotlivých vrstiev medailónovej architektúry (pozri sekciu 5.2), a zároveň nám umožňuje vytvárať priečinky, ako napríklad *bronze/* alebo *silver/*. Týmto spôsobom sa zjednodušuje správa dát, ich organizácia a spätná dohľadateľnosť spracovania. Navyše je služba optimalizovaná pre prácu s veľkým objemom súborov a paralelné spracovanie dát, čo je v kontexte pravidelnej aktualizácie fiškálnych a textových dát kľúčové.

Azure Data Factory

Druhou významnou službou v našom navrhnutom riešení je Azure Data Factory (ADF). Táto služba predstavuje integračnú a orchestračnú vrstvu, ktorá umožňuje vytvárať dátové pipeline procesy, prepájať rôzne dátové zdroje a riadiť jednotlivé fázy spracovania dát.

Hlavnou výhodou ADF je, že umožňuje automatizovať zber dát zo zdrojových systémov, ich transformáciu a presun do cieľového úložiska. V kontexte nášho riešenia je táto vlastnosť dôležitá najmä preto, že pracujeme s viacerými dátovými zdrojmi, ktoré je potrebné pravidelne spracovávať a aktualizovať.

Ďalšou významnou výhodou služby je možnosť parametrizácie procesov v dátových tokoch, plánovania ich spúšťania a implementácie riadenia závislostí medzi jednotlivými krokmi. Týmto prístupom je možné minimalizovať potrebu manuálnych zásahov do procesov spracovania dát a zvýšiť mieru automatizácie.

ADF taktiež poskytuje mechanizmy monitorovania a správy spustených procesov, čím zvyšuje transparentnosť riešenia a uľahčuje identifikáciu prípadných chýb. V kontexte práce s externými dátovými zdrojmi je takáto funkcionalita dôležitá najmä preto, že dostupnosť alebo štruktúra poskytovaných dát sa môže v čase meniť.

Azure Batch

Treťou kľúčovou službou, ktorá je súčasťou navrhnutého riešenia, je Azure Batch. Túto službu sme zvolili ako výpočtový systém pre časti riešenia, ktoré si vyžadujú vlastnú spracovateľskú

logiku a ktoré nie je vhodné realizovať výlučne prostredníctvom natívnych transformačných nástrojov Azure Data Factory.

Azure Batch umožňuje efektívne spúšťať a škálovať dávkové úlohy vo forme individuálnych Python skriptov v prostredí cloudu. Táto vlastnosť je dôležitá najmä pri spracovaní neštruktúrovaných textových dát, kde je potrebné flexibilne reagovať na rozdielnu štruktúru zdrojových webových stránok a implementovať vlastné postupy extrakcie, čistenia a normalizácie textu.

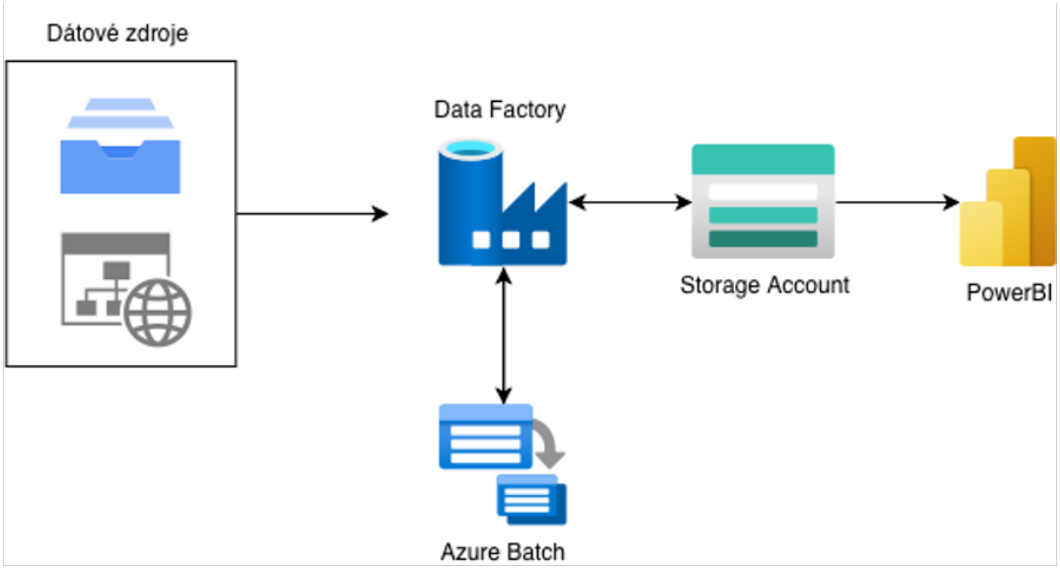
Výhodou tohto prístupu je oddelenie výpočtovej logiky od riadenia dátového toku. Azure Batch tak dopĺňa možnosti Azure Data Factory a umožňuje implementovať časti riešenia, ktoré by boli v samotnom ADF prostredí zložité, neprehľadné alebo menej efektívne.

5.2 Architektúra navrhnutého riešenia

Návrh riešenia sme vzhľadom na požiadavky rozdelili na viacero logických a technologických vrstiev, ktoré v konečnom dôsledku tvoria ucelený a navzájom prepojený systém.

5.2.1 Infraštruktúrny návrh riešenia

Celkový infraštruktúrny návrh riešenia je znázornený na diagrame 5, ktorý zohľadňuje nástroje použité pri jeho implementácii.



Obr. 5: Infraštruktúrny návrh cloudového dátového riešenia

Na začiatku architektúry sa nachádzajú vstupné zdrojové systémy, z ktorých čerpáme vstupné dáta. Ide najmä o štatistické a textové dáta poskytované inštitúciami ako Eurostat, Európska centrálna banka a Európska komisia, ktoré boli taktiež popísané v kapitole 4. Dáta sú získavané rôznymi mechanizmami, predovšetkým prostredníctvom API volaní, spracovaním webového obsahu alebo odoberaním RSS feedu.

Centrálnym riadiacim komponentom navrhnutej infraštruktúry je služba Azure Data Factory. Tá v riešení zabezpečuje orchestráciu dátových tokov a koordinuje komunikáciu medzi zdrojovými systémami, dátovým úložiskom a výpočtovou službou Azure Batch. Prostredníctvom tejto služby sú riadené procesy extrakcie dát, transformačné kroky a následné ukladanie výsledkov do dátovej vrstvy.

Služba Azure Batch zapájame do riešenia ako výpočtovú vrstvu pre úlohy, ktoré nie je vhodné alebo efektívne realizovať iba prostredníctvom natívnych nástrojov Azure Data Factory. Ide najmä o spracovanie kvalitatívnych textových dát, kde je potrebné vykonať vlastnú logiku extrakcie, čistenia, normalizácie a hodnotenia textov. Azure Batch tak dopĺňa Azure Data Factory tým, že zabezpečuje vykonanie vlastných Python skriptov, zatiaľ čo Azure Data Factory riadi ich spúšťanie a začlenenie do celkového dátového toku.

Centrálnu dátovú vrstvu riešenia predstavuje služba Azure Storage Account, konkrétne úložisko Azure Data Lake Storage Gen2. Do tejto vrstvy sú ukladané vstupné aj transformované dáta v jednotlivých fázach spracovania. Úložisko zároveň slúži ako spoločný dátový priestor pre Azure Data Factory, Azure Batch a Microsoft Power BI.

Poslednú vrstvu navrhnutého riešenia predstavuje nástroj Microsoft Power BI, ktorý slúži ako prezentačná a vizualizačná vrstva systému. Prostredníctvom tohto nástroja sú spracované dáta sprístupnené používateľovi vo forme interaktívnych reportov, dashboardov a analytických výstupov.

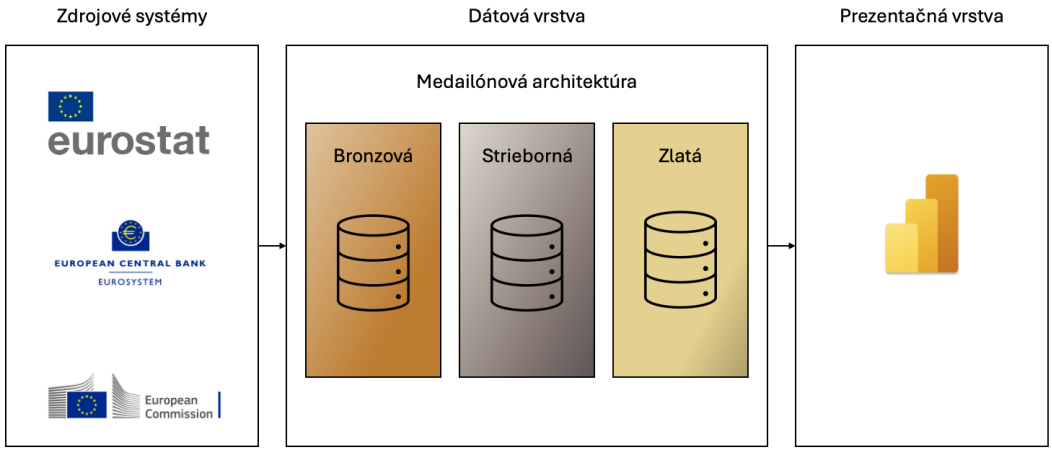
Zaradenie nástroja Microsoft Power BI ako samostatnej vrstvy zároveň reflektuje jeden z kľúčových princípov navrhnutej architektúry, ktorým je oddelenie transformačnej a dátovej vrstvy od vrstvy prezentačnej. Transformácie dát, čistenie údajov a aplikačná logika sú tak realizované v cloudovom prostredí, zatiaľ čo Power BI plní výlučne úlohu analytického a vizualizačného nástroja.

5.2.2 Medailónová architektúra dátovej vrstvy

Na základe našich požiadaviek vytvoriť robustné, reprodukovateľné, automatizované a škálovateľné riešenie sme pre dátovú vrstvu zvolili vrstvený architektonický model založený na medailónovom prístupe, ktorý môžeme vidieť znázornený na diagrame 6 Tento model nám umožňuje systematicky oddeliť jednotlivé fázy spracovania dát a postupne zvyšovať ich kvalitu a štruktúru v rámci troch logických vrstiev.

V rámci bronzovej vrstvy sú dáta ukladané v ich surovej podobe, čím zabezpečujeme zachovanie pôvodného stavu zdrojových údajov. V literatúre sa táto vrstva často označuje aj ako *source of truth*, teda primárny zdroj pravdy. Strieborná vrstva predstavuje priestor pre aplikáciu pravidiel dátovej kvality, čistenie, filtrovanie a štandardizáciu dát. Zlatá vrstva následne obsahuje transformované a agregované dátové súbory pripravené na načítanie do prezentačnej vrstvy.

Takto navrhnutá architektúra zároveň umožňuje uchovávanie historických dát a zabezpečuje spätnú dohľadateľnosť transformačných krokov. V prípade potreby je možné vrátiť sa k



Obr. 6: Medailónová architektúra

predchádzajúcej vrstve spracovania a opätovne realizovať transformačné operácie bez straty pôvodných dát. Týmto prístupom zvyšujeme transparentnosť riešenia, jeho auditovateľnosť a dlhodobú udržateľnosť.

5.3 Návrh a implementácia dátových tokov pre štruktúrované fiškálne dáta (Eurostat)

V tejto časti si predstavíme, akým spôsobom sme navrhli a následne implementovali spracovanie štruktúrovaných dát z platformy Eurostat, ktoré sme si bližšie opísali v kapitole 4.

5.3.1 Identifikácia a modelovanie zdrojov dát

Ako prvé sme sa v rámci návrhu museli zamyslieť nad tým, akým spôsobom chceme dáta z platformy získavať. Keďže sme chceli vytvoriť automatizované riešenie, manuálne sťahovanie TSV súborov, ktoré sme využívali v úvodnej fáze projektu, neprichádzalo do úvahy. Pri skúmaní možností sme narazili na dokumentáciu k možnostiam extrakcie dát pomocou REST API, kde bolo na výber z viacerých kategórií, konkrétne Statistics API, SDMX 2.1 a SDMX 3.0. Keďže sme sa rozhodli pracovať s dátami vo forme TSV, prvú možnosť, Statistics API, sme vylúčili, pretože poskytuje dáta výlučne vo formáte JSON-stat 2.0. Zvyšné dve možnosti ponúkajú štandardizovaný prístup k štatistickým údajom založený na modeli SDMX a umožňujú získavať rovnaké dátové sady prostredníctvom REST služieb. Hlavný rozdiel spočíva v tom, že SDMX 3.0 prináša modernejší návrh rozhrania s flexibilnejším filtrovaním a aktualizovaným spôsobom definovania výstupných formátov, zatiaľ čo SDMX 2.1 používa staršie URL štruktúry a parametre a zachováva sa najmä kvôli spätnej kompatibilite existujúcich riešení. Z týchto dôvodov sme sa rozhodli pre SDMX 3.0.

V ďalšom kroku sme sa snažili určiť základ URL, ktorý bude obsahovať nami požadované tabuľky a zároveň bude kompatibilný so SDMX 3.0. V dokumentácii sme sa dočítali, že dáta

Eurostatu sa nachádzajú pod základom **ec.europa.eu/eurostat/api/dissemination**.

Keď už sme mali zvolený základ URL a typ API, ktorý chceme použiť, vyzeral výsledný základ URL následovne:

`https://ec.europa.eu/eurostat/api/dissemination/sdmx/3.0/data/`

Následne bolo potrebné identifikovať jednotlivé endpointy pre konkrétne tabuľky. Na tento účel sme využili používateľské rozhranie dostupné na platforme Eurostat, v ktorom je možné vopred vyfiltrovať tabuľky v požadovanom formáte a následne zvoliť spôsob získania dát, pričom jednou z možností je práve API URL. Na každú z našich desiatich Eurostat tabuliek sme aplikovali základné filtre, ktoré sme uviedli v tabuľke 4 a vysvetlili v kapitole 4, a následne sme vygenerovali príslušné URL adresy.

Pri testovaní extrakcie dát pomocou týchto API URL sme narazili na jednu zaujímavú situáciu. Pre tabuľku DS-045409 sme boli v testovacom Jupyter Notebooku schopní extrahovať dáta so základom URL uvedeným vyššie. Pri implementácii tohto riešenia v Azure Data Factory však pri rovnakom URL dochádzalo k chybe. V dokumentácii sme následne zistili, že tabuľky začínajúce na DS sú špeciálne tabuľky pochádzajúce z databázy Comext a sú dostupné len prostredníctvom následujúcej URL adresy:

`https://ec.europa.eu/eurostat/api/comext/dissemination/`

kde môžeme sledovať, že oproti predchádzajúcej základnej URL pribudol prvok *comext*.

Po aplikovaní tejto zmeny sme už boli schopní dané dáta úspešne získavať aj pomocou Azure Data Factory.

Keď už boli identifikované všetky endpointy pre jednotlivé tabuľky, bolo potrebné rozhodnúť sa, akým spôsobom ich efektívne zapracovať do ADF tak, aby sa zároveň zachovala informácia o názve tabuľky a aby bolo možné endpoint v prípade zmien jednoducho upraviť. Z uvedených dôvodov boli URL odkazy uložené v JSON formáte v nasledujúcom tvare:

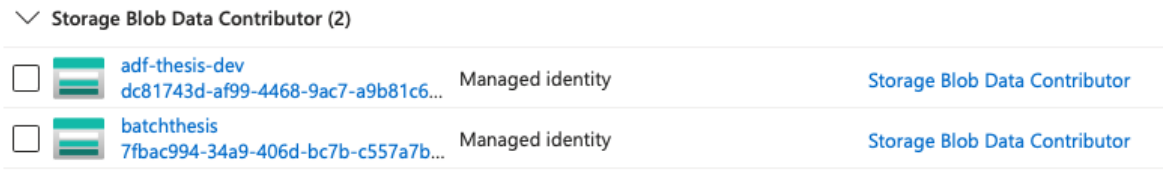
```
{
    "tableId": "TABLE_NAME",
    "relativeUrl": "RELATIVE_URL",
    "frequency": "FREQUENCY"
}
```

Ako je možné vidieť, jeden prvok JSON štruktúry obsahuje informácie o konkrétnej tabuľke, spôsobe prístupu k nej a periodicite poskytovaných dát. Tieto údaje sú dôležité nielen z pohľadu informovania používateľa, ale aj pre samotné dátové pipeline, keďže napríklad na základe názvu tabuľky je možné automaticky vytvoriť dedikovaný priečinok pre jej spracovanie pomocou parametrov dostupných v službe ADF.

5.3.2 Príprava integračného prostredia

Konfiguračný JSON súbor opísaný v predchádzajúcej časti sme následne nahrali do úložiska ADLS, aby bol dostupný priamo v rámci cloudového prostredia a mohol byť využívaný v dátových pipeline procesoch.

Pred samotným načítaním dát do ADF je potrebné nastaviť niekoľko predpokladov. Platforma Azure kladie silný dôraz na bezpečnosť, a preto neumožňuje jednej službe pristupovať k druhej, pokiaľ jej to explicitne nepovolíme. Aby sme teda mohli z Azure Data Factory pristupovať k REST endpointom a k úložisku v Azure Data Lake Storage, musíme priamo v službe Storage Account priradiť spravovanej identite našej inštancie príslušné IAM oprávnenia. Konkrétne sme inštancii Azure Data Factory pridelili rolu Storage Blob Data Contributor, ktorá jej umožňuje čítať, zapisovať a odstraňovať údaje v našom úložisku. Tieto IAM oprávnenia môžeme vidieť na obrázku 7 nižšie.



Obr. 7: IAM oprávnenia

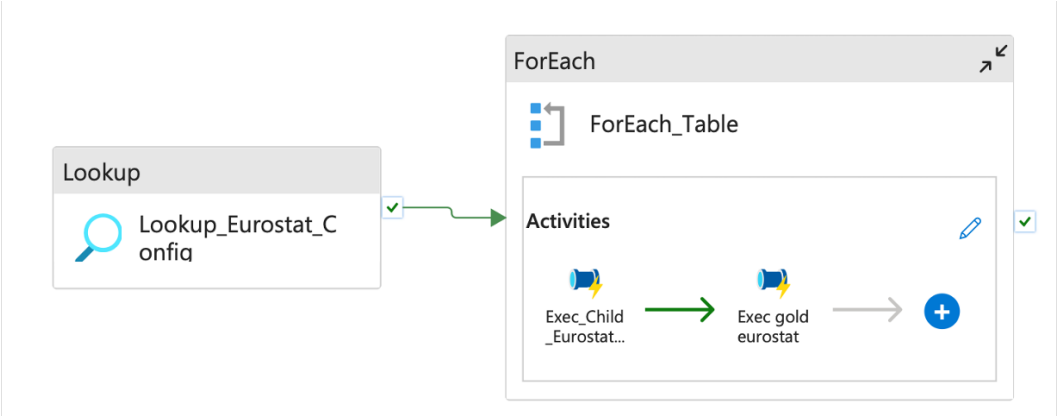
Druhým predpokladom je vytvorenie spojenia medzi ADF a jednotlivými zdrojmi dát. To realizujeme pomocou konektora s názvom linked service, ktorý je dostupný v ADF. V rámci riešenia sme vytvorili samostatné konektory pre ADLS a pre REST endpointy Eurostatu, pričom v prípade REST pripojenia bolo potrebné definovať základnú (base) URL adresu. Vzhľadom na to, že DS-045409 tabuľka využíva odlišnú základnú URL, vytvorili sme samostatné linked services pre oba typy základných endpointov.

Následne, keď už môže ADF komunikovať s ADLS, vytvoríme pre JSON súbor s endpointami dataset. Dataset v ADF predstavuje logickú definíciu dátovej štruktúry a umiestnenia zdroja alebo cieľa dát a slúži ako abstraktná vrstva medzi aktivitami v ADF a samotným úložiskom. Vďaka datasetu je možné jednoducho meniť zdroj dát bez úpravy celého dátového toku. Pri vytváraní datasetu je potrebné zadať typ úložiska, v ktorom sa nachádza (Azure Data Lake Storage Gen2), a formát súboru (JSON), zvoliť príslušný linked service a špecifikovať presnú cestu k súboru. V našom prípade ide o cestu *datalake/config/eurostat_tables.json*. Datasety sme v bronzovej vrstve obdobným spôsobom vytvorili aj pre oba typy základu URL a pre cieľový súbor typu TSV.

5.3.3 Riadiaci dátový tok

Riadiaci dátový tok, resp. master pipeline, predstavuje nadradený riadiaci mechanizmus, ktorý zabezpečuje dynamické spracovanie viacerých dátových tabuliek bez potreby manuálneho

zásahu alebo vytvárania desiatich rôznych variantov dátových tokov. Ako môžeme pozorovať z obrázka 8 nižšie, proces iniciuje aktivita typu Lookup, ktorá načíta konfiguračný JSON súbor obsahujúci definície jednotlivých tabuliek, ktoré sme predstavili v časti 5.3.1 vyššie.



Obr. 8: Eurostat - Riadiaci dátový tok

Výstup tejto aktivity slúži ako vstup pre aktivitu typu ForEach, ktorá iteratívne spúšťa podriadené dátové toky pre každú tabuľku samostatne. Aby tento proces mohol efektívne fungovať, museli sme pre každý podriadený dátový tok nastaviť parametre. V našom prípade išlo o parametre ako názov tabuľky, relatívnu URL adresu a frekvenciu aktualizácie. Zároveň treba pripomenúť, že na realizáciu úvodnej Lookup aktivity bolo potrebné špecifikovať dataset pre spomínanú JSON konfiguráciu.

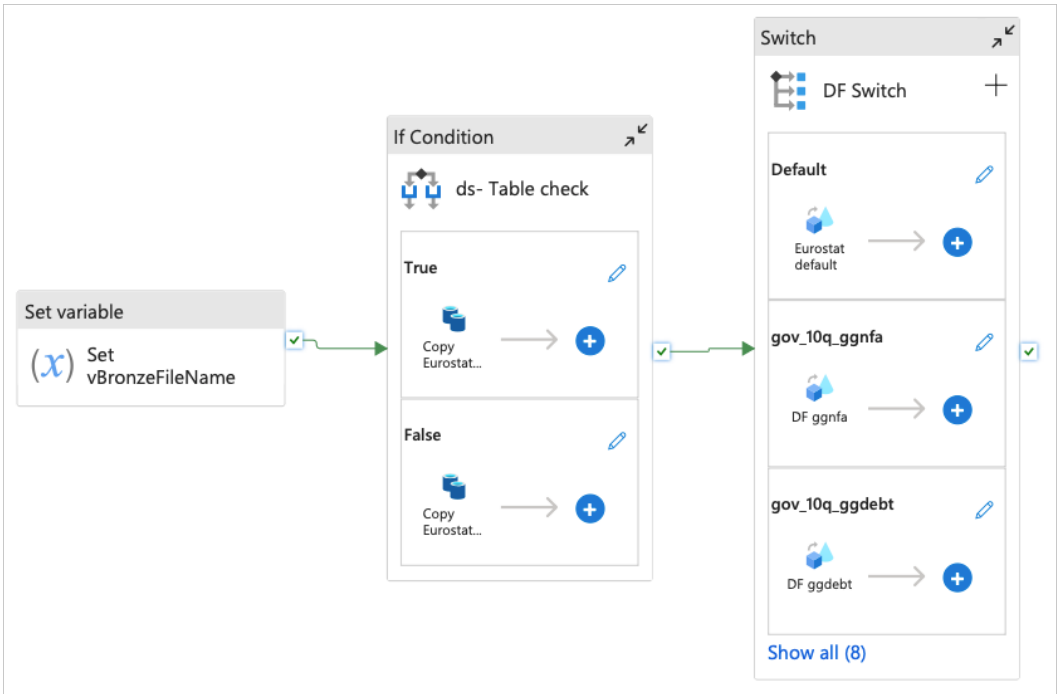
5.3.4 Implementácia extrakcie dát

V tejto časti opisujeme prvý podriadený dátový tok, ktorý je spúšťaný aktivitou typu *ForEach* v nadradenom (master) pipeline procese. Schéma tohto podriadeného pipeline procesu je znázornená na obrázku 9.

V úvodnej fáze nastavujeme premennú **vBronzeFileName**, ktorej priraďujeme názov tabuľky získaný z nadradeného toku. Tento krok umožňuje dynamickú parametrizáciu následných aktivít v rámci pipeline. V nasledujúcom kroku využívame podmienkovú aktivitu, prostredníctvom ktorej kontrolujeme, ktorá tabuľka je aktuálne spracovávaná. Táto podmienka je nevyhnutná z dôvodu existencie tabuľky DS-045409, ktorá využíva odlišnú základnú URL adresu. Pre túto tabuľku je preto potrebné použiť samostatnú *Copy* aktivitu s odlišným konektorom linked service.

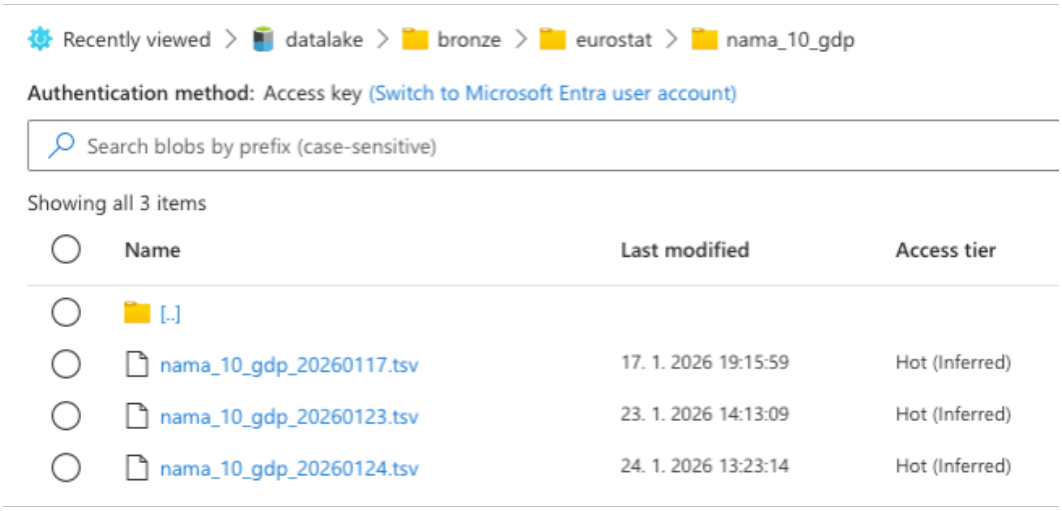
Aktivita typu *Copy* predstavuje hlavný komponent našej bronzovej vrstvy, keďže zabezpečuje samotnú extrakciu dát z definovanej URL adresy prostredníctvom príslušného datasetu. Následne sú dáta uložené vo formáte TSV do úložiska Azure Data Lake Storage pod nasledujúcou cestou:

```
bronze/eurostat/NAZOV_TABULKY/NAZOV_TABULKY_DATUM.tsv
```



Obr. 9: Eurostat - Bronze a Silver dátový tok

Môžeme pozorovať, že názov súboru okrem parametrizovaného názvu tabuľky obsahuje aj dátum. Tento dátum reprezentuje moment extrakcie dát zo zdrojového systému a slúži predovšetkým na odlíšenie jednotlivých súborov a zabezpečenie historizácie bronzovej vrstvy. Ukážku finálneho uloženia dát v bronzovej vrstve uvádzame na obrázku 10.



Obr. 10: Eurostat - Bronze ukážka

5.3.5 Transformácia dát

Po úspešnom uložení extrahovaných dát do bronzovej vrstvy nasleduje fáza transformácie dát. V tomto kroku, ako je znázornené na obrázku 9, využívame aktivitu typu *Switch*, ktorá

kontroluje názov aktuálne spracovávanej tabuľky. Hoci transformácie jednotlivých tabuliek vo svojej podstate pozostávajú z podobnej sekvencie operácií, na rozdiel od extrakcie dát prostredníctvom aktivity *Copy* nie je možné využiť jednotnú parametrizovanú logiku pre všetky tabuľky. Dôvodom je skutočnosť, že každá tabuľka obsahuje špecifickú štruktúru stĺpcov, ktorá si vyžaduje individuálny transformačný prístup.

V jednom prípade však došlo k situácii, kde tri tabuľky disponovali identickou štruktúrou stĺpcov a rovnakou periodicitou, a preto sme pre ne implementovali spoločnú transformačnú logiku. Konkrétne sa jednalo o tabuľky *gov_10dd_edpt1*, *gov_10a_main* a *gov_10_taxag*.

Pre uvedené tabuľky sa transformačný proces skladá z deviatich krokov, ktoré sú znázornené na obrázku 11.

Obr. 11: Eurostat - Silver DataFlow

V prvom kroku načítavame dáta z bronzovej vrstvy prostredníctvom definovaného datasetu. Po načítaní dát je možné pozorovať, že napriek uloženiu vo formáte TSV sú hlavné identifikačné hodnoty agregované v jednom stĺpci a oddelené čiarkou. Túto skutočnosť môžeme pozorovať na obrázku 12. Z tohto dôvodu sme v nasledujúcom kroku aplikovali transformáciu typu *Split*, prostredníctvom ktorej sme daný stĺpec rozdelili na päť samostatných stĺpcov. Pôvodný spojený stĺpec sme následne odstránili, keďže Data Flow ho po rozdelení automaticky nezruší. Data Flow v službe Azure Data Factory predstavuje vizuálny nástroj na transformáciu dát, ktorý umožňuje ich čistenie, spájanie a úpravu bez potreby manuálneho písania kódu.

freq,unit,sector,na_item,geo\TIME_PERIOD	1995	1996
A,MIO EUR,S13,B9,AT	-11332.3	-8504.5
A,MIO EUR,S13,B9,BE	-9929.3	-8803.1

Obr. 12: Eurostat - Spojené stĺpce

V ďalšej fáze sme aplikovali operáciu *Unpivot*, keďže jednotlivé roky boli reprezentované ako samostatné stĺpce. Touto transformáciou sme dosiahli normalizovanú štruktúru dát, kde rok predstavuje samostatný atribút.Následne sme implementovali základné pravidlá dátovej kvality a odstránili záznamy s nulovými hodnotami v atribútoch krajina, položka a rok. V ďalšom kroku sme upravili stĺpec *values*, keďže v niektorých prípadoch obsahoval aj textové označenia (napr. „p“), ktoré indikujú predbežné (provisional) hodnoty. Tieto znaky boli oddelené od numerickej hodnoty. Ďalej sme nahradili špecifické znaky reprezentujúce chýbajúce údaje (napr. dvojbodku) prázdnou hodnotou. V predposlednom kroku sme nastavili

správne dátové typy jednotlivých stĺpcov a v závere transformačného procesu sme výslednú tabuľku uložili vo formáte Parquet do striebornej vrstvy. Finálny výstup striebornej vrstvy je znázornený na obrázku 13.

freq	unit	sector	na_item	geo	year	value	flag
A	MIO EUR	S13	B9	AT	1995	-11332.30	
A	MIO EUR	S13	B9	AT	1996	-8504.50	
A	MIO EUR	S13	B9	AT	1997	-4901.60	

Obr. 13: Eurostat - Výstup striebornej vrstvy

Pre použitie formátu Parquet sme sa rozhodli z dôvodu jeho technologických výhod. Ide o stĺpcovo orientovaný a komprimovaný formát, ktorý dokáže výrazne znížiť objem uložených dát a zároveň zrýchľuje analytické operácie nad nimi. Táto vlastnosť je žiaduca najmä pri efektívnom načítavaní dát do nástroja Power BI, ako aj pri ďalšom spracovaní v rámci Azure Data Factory. Významnou výhodou je aj podpora schémy dát, ktorá umožňuje zachovať definované dátové typy a zabezpečiť ich konzistentný prenos do prezentačnej vrstvy.

Vyššie opísanou transformáciou sme dosiahli redukciu počtu stĺpcov z 31 na 8, zároveň sme aplikovali základné pravidlá dátovej kvality a zabezpečili normalizáciu dátovej štruktúry. Data Flow procesy pre zvyšné tabuľky sú implementované obdobným spôsobom, pričom sa líšia najmä v dodatočných transformačných krokoch reflektujúcich špecifickú štruktúru jednotlivých datasetov.

Je potrebné zároveň uviesť, že v striebornej vrstve na rozdiel od bronzovej neuchovávame historické verzie dát, ale priečinok príslušnej tabuľky pri každom spracovaní prepisujeme. Keďže vstupné dáta z Eurostatu pri každom načítaní obsahujú kompletnú historickú časovú radu, implementácia inkrementálnej stratégie by v tomto prípade neprinášala výraznú pridanú hodnotu. Úplná historizácia je zabezpečená už na úrovni bronzovej vrstvy, čo považujeme za dostatočné.

Vo vyššie opísanom dátovom toku, ktorý je znázornený na obrázku 9, je možné pozorovať význam parametrizácie viacerých prvkov, najmä názvu tabuľky, cieľovej cesty uloženia a relatívnej URL adresy. Bez využitia parametrizácie by bolo potrebné vytvárať samostatné *Copy* aktivity pre každú tabuľku, čo by výrazne znižovalo škálovateľnosť, udržateľnosť a prehľadnosť celého riešenia.

5.3.6 Finálne úpravy tabuliek

Zlatú vrstvu sme implementovali ako samostatný, dedikovaný dátový tok, ktorý je taktiež volaný v rámci hlavného pipeline procesu, ako je znázornené na obrázku 8. Samostatný dátový tok pre zlatú vrstvu bol zvolený najmä z dôvodu lepšej prehľadnosti riešenia a jednoduchšieho monitorovania logov.

Ako je možné vidieť na obrázku 14, zlatá vrstva využíva podobný mechanizmus riadenia transformácií ako strieborná vrstva. Nachádza sa tu aktivita typu *Switch*, ktorá na základe

názvu tabuľky priraďuje príslušnú transformačnú logiku implementovanú prostredníctvom aktivity *Data Flow*.

Obr. 14: Eurostat - Gold Pipeline

Hlavný rozdiel medzi striebornou a zlatou vrstvou spočíva v ich účele. Kým v striebornej vrstve sa zameriavame na čistenie dát, aplikáciu pravidiel dátovej kvality a normalizáciu štruktúry, v zlatej vrstve aplikujeme konkrétnu biznis logiku. Údaje sú tu upravované tak, aby boli priamo pripravené pre analytické spracovanie v nástroji Power BI. V prípade potreby viacerých analytických scenárov je zároveň možné vychádzať z plne normalizovaných tabuliek striebornej vrstvy.

Pre ilustráciu uvádzame príklad jednej z transformácií zlatej vrstvy, ktorá je znázornená na obrázku 15.

Obr. 15: Eurostat - Gold DataFlow

V prvom kroku načítavame dáta zo striebornej vrstvy vo formáte Parquet. Následne aplikujeme filter na atribút rok, pričom ponechávame iba údaje za obdobie od roku 2015.

Toto obmedzenie vychádza z analytického rozsahu, ktoré sme definovali pre vizualizácie v Power BI.

V ďalšom kroku odstraňujeme nulové hodnoty, ktoré by mohli negatívne ovplyvniť výsledky analýzy. Následne aplikujeme validačné podmienky zabezpečujúce, že frekvencia údajov zodpovedá požadovanej hodnote (v tomto prípade ročnej frekvencii označenej ako „A“) a že sektor je rovný hodnote „S13“. V prípade nesplnenia týchto podmienok je vyhlásená chyba, čím sa zabezpečuje konzistentnosť dát.

V predposlednom kroku zabezpečujeme konzistentné pomenovanie stĺpcov podľa očakávanej schémy a v závere transformačného procesu sú dáta opätovne ukladané vo formáte Parquet do zlatej vrstvy.

Významnejšia zmena oproti striebornej vrstve nastáva pri tabuľke *DS-045409*, ktorú v zlatej vrstve rozdeľujeme na dve samostatné tabuľky, keďže obsahuje údaje o importe aj exporte.

5.3.7 Zhrnutie návrhu dátového toku pre štruktúrované dáta

Navrhnutý dátový tok pre spracovanie štruktúrovaných fiškálnych dát z Eurostatu je založený na medailónovej architektúre, ktorá umožňuje systematické oddelenie extrakcie, čistenia a aplikácie biznis logiky. Prostredníctvom parametrizácie zdrojových URL, názvov tabuliek a cieľových ciest sme dosiahli vysokú mieru modularity a škálovateľnosti riešenia.

Spojením bronzovej a striebornej vrstvy do jedného podriadeného pipeline procesu sme zabezpečili procesnú konzistentnosť spracovania jednotlivých tabuliek, zatiaľ čo oddelenie zlatej vrstvy umožnilo aplikovať špecifickú analytickú logiku nezávisle od transformačnej časti. Z týchto uvedených dôvodov poskytuje navrhnuté riešenie robustný a reprodukovateľný základ pre následnú analýzu v prostredí Power BI.

5.4 Metodika sentimentovej analýzy

V tejto časti sa zameriavame na metodiku analýzy sentimentu, postupy, ktoré jej predchádzali, a modely zvolené v rámci navrhnutého riešenia.

5.4.1 Prístup k analýze sentimentu

Hlavným cieľom analýzy sentimentu v našom riešení je identifikovať, či komunikácia analyzovaných inštitúcií vykazuje skôr pozitívny, negatívny alebo neutrálny charakter. Za účelom dosiahnutia tohto cieľa sa snažíme v rámci nášho navrhnutého riešenia kvantifikovať tón textových dát vybraných európskych inštitúcií získaných z oficiálnych publikácií alebo predpovedí. Sentiment v našom kontexte nepredstavuje subjektívne emocionálne hodnotenie, ale aproximáciu obsahového vyznenia textu, ktorá umožňuje jeho následnú agregáciu, napríklad na mesiace a roky, a porovnanie v čase.

Pri návrhu sentimentovej analýzy sme sa rozhodli nevyužiť tradičné lexikónové metódy, ktoré určujú sentiment na základe vopred definovaných slovníkov, ale zvolili sme prístup založený na využití jazykových modelov typu Large language model (LLM), konkrétne *gpt-4o-mini* a *gpt-5.5*.

Dôvodom tejto voľby je charakter analyzovaných textov, ktoré pochádzajú z formálnej inštitucionálnej komunikácie. Takéto texty sú často štylizované neutrálne a využívajú odbornú terminológiu. Lexikónové metódy by v takomto prípade nemuseli poskytovať dostatočne presné výsledky, keďže nedokážu zachytiť širší význam viet. Na druhej strane LLM modely umožňujú interpretovať text komplexnejšie, pričom model zohľadňuje kontext, štruktúru viet aj význam jednotlivých výrazov v rámci celého textu. Lexikónové metódy by mohli byť vhodným kandidátom, pokiaľ by sme sa zamerali napríklad na sociálne siete, čo však nie je predmetom tejto práce.

Napriek výhodám zvoleného prístupu pomocou LLM modelu je potrebné zohľadniť aj jeho obmedzenia. Analýza sentimentu založená na LLM je závislá nielen od kvality vstupných dát, ale aj formulácie inštrukcií a príkladov, pričom rôzne nastavenia môžu viesť k odlišným výsledkom.

Ďalším obmedzením je samotný charakter analyzovaných textov. Inštitucionálna komunikácia je často formulovaná neutrálne a opatrne, čo môže viesť k nižšej variabilite sentimentu a sťažovať jeho jednoznačnú interpretáciu. Z tohto dôvodu môžeme vo výsledkoch uvedených v nasledujúcej kapitole očakávať skôr neutrálny charakter sentimentu.

A posledným obmedzením, ktoré treba spomenúť, je charakter, akým LLM modely operujú. Tieto modely nepracujú na princípe skutočného porozumenia obsahu v ľudskom zmysle slova, pretože ich výstup je výsledkom pravdepodobnostného spracovania textu na základe vzorov, ktoré sa daný model z rozsiahlych trénovacich dát naučil. Z tohto dôvodu môže byť výsledné sentimentové hodnotenie ovplyvnené spôsobom formulácie inštrukcií, štruktúrou vstupného textu alebo kontextom, ktorý model implicitne priraďuje analyzovanému obsahu. Pri použití rovnakého modelu tak nemožno úplne vylúčiť situácie, v ktorých by odlišná formulácia vstupu alebo jemná zmena inštrukcie viedla k mierne odlišnému výsledku.

Obmedzenie predstavuje aj rozdelenie dlhších textov na menšie časti. Hoci tento krok umožňuje efektívne spracovanie rozsiahlejších dokumentov, zároveň môže viesť k čiastočnej strate širšieho významového rámca textu. Jednotlivé časti dokumentu môžu niesť odlišný význam, ktorý je plne pochopiteľný až v kontexte celého dokumentu. Agregácia čiastkových výsledkov preto predstavuje pragmatický kompromis medzi technickou realizovateľnosťou a obsahovou presnosťou.

5.4.2 Očakávaný výstup LLM modelu

Ako výstup sentimentovej analýzy očakávame štruktúrované dáta vo formáte JSON, ktoré obsahujú identifikáciu zdroja, základné metadáta a samotné hodnotenie sentimentu vrátane doplňujúcich atribútov. Príklad takéhoto výstupu je uvedený nižšie:

```
{
 "url": URL,
 "year": 2020,
 "chunk_id": 0,
 "sentiment": "negative",
 "score": -1,
 "uncertainty": "medium",
 "confidence": 0.75,
 "rationale": "The forecast describes a severe shock to
    economic activity in the first half of 2020 followed by
    initial recovery. However, the resurgence of the pandemic
     suggests ongoing challenges and uncertainty for growth,
    inflation, and employment, indicating a cautious and
    somewhat pessimistic outlook.",
 "model": "gpt-4o-mini",
 "scored_on": "2026-02-16"
}
```

Výstup predstavuje hodnotenie sentimentu na úrovni jednotlivých textových segmentov (chunkov). Kľúčovým atribútom pre ďalšie spracovanie je numerické skóre (v príklade označené ako `score`), ktoré vyjadruje sentiment na škále od -2 do +2, pričom hodnota -2 reprezentuje silne negatívny sentiment, +2 silne pozitívny tón a 0 neutrálny tón.

Okrem toho model vracia aj atribút `confidence`, ktorý reprezentuje mieru istoty modelu pri danom hodnotení. Tento atribút zohráva dôležitú úlohu pri následnej agregácii výsledkov, keďže umožňuje zohľadniť spoľahlivosť jednotlivých hodnotení.

Ostatné atribúty, ako napríklad `uncertainty` alebo `rationale`, slúžia najmä na interpretáciu výsledkov a lepšie pochopenie rozhodnutia modelu.

5.4.3 Agregácia sentimentového skóre

Na úrovni jednotlivých textových segmentov získavame čiastkové sentimentové skóre, ktoré je následne potrebné agregovať na vyššie úrovne, ako sú dokumenty (URL) a časové obdobia.

Agregácia prebieha v dvoch krokoch. V prvom kroku agregujeme skóre na úrovni jednotlivých dokumentov ako vážený priemer, pričom váhou je hodnota `confidence`:

$$S_{\text{url}} = \frac{1}{n} \sum_{i=1}^{n} s_i \cdot c_i$$

kde s_i predstavuje sentimentové skóre jednotlivého segmentu a c_i jeho príslušnú mieru dôvery. Hodnota n označuje počet segmentov daného dokumentu.

Okrem váženého skóre evidujeme vo výstupe agregácie aj jednoduchý aritmetický priemer:

$$S_{\text{raw}} = \frac{1}{n} \sum_{i=1}^{n} s_i$$

Tento údaj nám slúži najmä ako referenčná hodnota a umožňuje tak porovnanie vplyvu váženia podľa confidence.

V druhom kroku agregujeme výsledné skóre na úrovni časových období (napr. rok alebo štvrťrok), kde využívame aritmetický priemer skóre jednotlivých dokumentov:

$$S_{\text{period}} = \frac{1}{m} \sum_{j=1}^{m} S_{\text{url},j}$$

kde m predstavuje počet dokumentov v danom období.

Pri tejto agregácii používame primárne vážené skóre S_{url}, ktoré zohľadňuje spoľahlivosť jednotlivých hodnotení na úrovni segmentov. Tento prístup sme zvolili najmä preto, aby sme redukovali vplyv menej istých predikcií modelu a zároveň zachovali interpretovateľnosť výsledkov.

Takto agregované skóre následne využívame na analýzu vývoja sentimentu v čase a jeho porovnanie medzi jednotlivými inštitúciami.

5.5 Návrh a implementácia dátových tokov pre kvalitatívne textové dáta (EK, ECB)

V tejto kapitole si predstavíme, ako sme navrhli dátové toky na spracovanie textových dát zo stránky Európskej komisie a Európskej centrálnej banky, na automatizovanú extrakciu textu, očistenie a úpravu. Výslednú formu týchto textov sme následne poskytli vybranému LLM modelu, ktorý nám následne poskytol pre jednotlivé texty rôzne parametre podľa inštrukcií ako napríklad skóre a istota. Výsledkom tohto procesu je konzistentne štruktúrovaný dataset vhodný na sentimentovú analýzu.

5.5.1 Architektonický návrh riešenia pre textové dáta

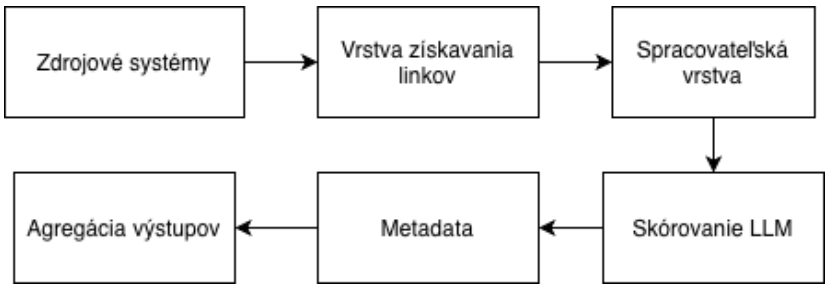
Pri návrhu architektúry riešenia sme museli zohľadniť špecifiká a charakter kvalitatívnych textových dát, ako aj požiadavky vyplývajúce z ich následného analytického využitia. Ako už bolo spomenuté v kapitole 4, vstupné dáta sú získavané z verejne dostupných zdrojov EC a ECB, pričom ide o textové dokumenty publikované v rôznych formátoch, v rôznych časových intervaloch a s odlišnou štruktúrou. Nami navrhnuté riešenie preto muselo podporovať nielen samotné získavanie týchto textov, ale aj ich čistenie, transformáciu, skórovanie a uloženie do jednotnej dátovej podoby.

Z funkčného hľadiska bolo potrebné zabezpečiť najmä automatizované získavanie textov z externých webových zdrojov, spracovanie neštruktúrovaného alebo pološtruktúrovaného textu, pravidelné spúšťanie dátových tokov, integráciu vlastných spracovateľských skriptov a vytvorenie konzistentného výstupu vhodného na ďalšiu analýzu.

Z vyššie uvedených dôvodov som za kľúčový prvok architektúry zvolil opäť službu Azure Data Factory, ktorá v navrhnutom riešení plnila úlohu orchestrácie dátového toku. Prostredníctvom tejto služby bolo možné definovať logiku tokov, riadiť poradie jednotlivých krokov spracovania aj mimo samotných skriptov, zabezpečiť ich plánované spúšťanie a monitorovať priebeh spracovania. Azure Data Factory tak nebola využitá na samotné detailné textové spracovanie, ale predovšetkým ako riadiaca vrstva, ktorá prepája jednotlivé časti riešenia do jedného konzistentného procesu.

Na implementáciu spracovateľskej logiky bola využitá služba Azure Batch v kombinácii s vlastnými Python skriptami. Tento prístup nám umožnil oddeliť riadenie dátových tokov od vykonávania výpočtovo a logicky náročnejších úloh, ako sú extrakcia textu z rôznych formátov, čistenie obsahu a transformácia textu do menších častí pre LLM. Výhodou tohto prístupu bola najmä flexibilita, keďže vlastné skripty umožňovali reagovať na heterogenitu zdrojových dokumentov a implementovať spracovanie, ktoré by bolo v čisto low-code prostredí zložité alebo neefektívne.

Z logického hľadiska možno navrhnutý proces spracovania textov rozdeliť do viacerých vrstiev, ktoré sú znázornené na obrázku 16. Jednotlivé kroky v rámci týchto vrstiev boli implementované prostredníctvom vlastných Python skriptov, ktoré realizovali konkrétne operácie nad textovými dátami a následne boli zoskupené do dátových tokov v rámci medailónovej architektúry. Prvú vrstvu tvorili zdrojové systémy reprezentované webovými stránkami Európskej komisie a Európskej centrálnej banky. Tieto zdroje poskytovali textové dokumenty vo forme HTML stránok, archívnych článkov alebo PDF súborov. Druhú vrstvu predstavovala vrstva získavania odkazov, ktorej úlohou bolo identifikovať relevantné odkazy, získať ich a zabezpečiť extrakciu ich obsahu. Na túto vrstvu nadväzovala spracovateľská vrstva, v rámci ktorej prebiehala extrakcia textu, odstránenie technických a netextových prvkov, normalizácia obsahu a transformácia dokumentov do jednotnej dátovej reprezentácie.



Obr. 16: Proces spracovania textov

Po spracovaní a očistení textov nasledovala vrstva analytického skórovania. V tejto fáze boli pripravené textové dáta poskytované vybranému LLM modelu, ktorý na základe vopred

definovaných inštrukcií generoval doplňujúce analytické atribúty, ako napríklad sentimentové skóre alebo mieru istoty hodnotenia. Výstupy z tejto fázy boli následne pripojené k jednotlivým odkazom, z ktorých pochádzal pôvodný text, a zároveň sme k nim doplnili aj metadáta. Tento výstup bližšie popisujeme v časti 5.4.2. V záverečnej fáze boli tieto texty agregované podľa rokov a mesiacov v závislosti od zdroja.

Celkový tok dát bol teda navrhnutý tak, aby jednotlivé kroky tvorili ucelený a opakovateľný proces. Azure Data Factory zabezpečovala riadenie a orchestráciu dátových tokov, Azure Batch vykonávala konkrétne spracovateľské úlohy a Python skripty implementovali samotnú logiku extrakcie, čistenia a transformácie dát. Takto navrhnutá architektúra umožnila oddeliť riadiacu vrstvu od implementačnej logiky, čím sa zvýšila modularita, prehľadnosť a rozšíriteľnosť riešenia.

Na základe uvedenej architektúry boli následne navrhnuté samostatné dátové toky pre jednotlivé zdrojové systémy. Hoci texty Európskej komisie a Európskej centrálnej banky sledovali rovnaký analytický cieľ, ich forma publikovania a technické charakteristiky sa líšili. Z tohto dôvodu bolo potrebné pristupovať k návrhu dátového toku pre každý zdroj samostatne, pričom v ďalších krokoch boli tieto texty transformované do jednotnej analytickej podoby.

5.5.2 Implementácia dátového toku pre EC texty

Po identifikovaní dátových zdrojov a definovaní základného postupu spracovania textov sme sa zamerali na analýzu štruktúry dát publikovaných Európskou komisiou. Odkazy na ekonomické prognózy Európskej komisie boli v čase spracovania dostupné prostredníctvom oficiálnej stránky uvedenej v kapitole 4, ktorá obsahovala odkazy na prognózy od roku 2015. Táto skutočnosť bola pre naše riešenie výhodná, keďže naša kvantitatívna analýza na dátach z Eurostatu bola takisto vykonaná na údajoch od roku 2015.

Bronzová vrstva

Keďže nám oficiálna stránka Európskej komisie poskytovala prehľad všetkých relevantných odkazov na jednom mieste, v prvom kroku sme ich pomocou Python skriptu zozbierali a uložili do formátu JSON na účely ďalšieho spracovania, uchovania a historizácie. Príklad takéhoto výstupu je uvedený nižšie:

```
{
  "url": "https://economy-finance.ec.europa.eu/economic-
     forecast-and-surveys/economic-forecasts/autumn-2015-
     economic-forecast_en",
  "first_seen": "2026-02-16",
  "last_seen": "2026-02-16",
  "year": 2015,
  "text_extracted": true,
```

```
    "text_extracted_on": "2026-02-16",
    "text_extracted_error": null
  }
```

Ako možno vidieť z uvedeného výstupu, okrem samotných odkazov sme extrahovali aj rok, pre ktorý je daná prognóza, a zároveň sme daný element doplnili o metadáta. Metadáta typu *first_seen* a *last_seen* boli pridané už v tomto kroku. Zvyšné metadáta sme doplnili až v kroku extrakcie obsahu, aby sme text nespracovávali viac ako raz.

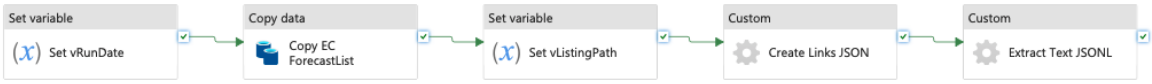
V nasledujúcej fáze sme analyzovali štruktúru jednotlivých prognóz s cieľom identifikovať spoločný vzor, na základe ktorého by sme vedeli efektívne a generalizovateľne extrahovať texty zo všetkých dostupných predpovedí. Počas tejto analýzy sme však zistili, že štruktúra jednotlivých prognóz nebola v priebehu sledovaného obdobia jednotná. Zmeny sa týkali tak obsahovej, ako aj štruktúrnej stránky. Niektoré prognózy obsahovali stručné textové zhrnutie priamo na webovej stránke vo forme samostatnej sekcie, iné boli doplnené o PDF dokumenty, obrázky alebo ďalšie sprievodné prvky. Rozsah a forma dostupného textu sa tak medzi jednotlivými publikáciami výrazne líšili.

Jedným z dôležitých rozhodnutí v rámci návrhu dátového toku bolo aj určenie primárneho textového zdroja. Ako hlavný vstup pre ďalšie spracovanie sme zvolili text dostupný priamo na webovej stránke, a nie plné PDF dokumenty priložené k prognózam. Dôvodom bola najmä ich rozsahová a obsahová komplexnosť. Zatiaľ čo text publikovaný priamo na stránke spravidla predstavoval zhrnutie danej prognózy, priložené PDF dokumenty často obsahovali rozsiahlu analytickú správu s detailným rozborom situácie. Takto rozsiahle texty by boli z pohľadu následného spracovania jazykovým modelom výrazne náročnejšie a menej vhodné na konzistentné skórovanie.

Z uvedených dôvodov sme sa rozhodli, že najefektívnejším spôsobom bude extrahovať všetky odseky z daného odkazu, ktoré obsahovali text týkajúci sa predpovede, pričom extrakcia bola ukončená v momente, keď sa na stránke začali objavovať sekcie nesúvisiace s hlavným obsahom, ako napríklad videá, médiá, súvisiace odkazy alebo prílohy. Tento prístup nám umožnil zachytiť hlavnú textovú časť prognózy v čo najkonzistentnejšej forme naprieč celým sledovaným obdobím. Po vykonaní extrakcie sme získaný text spolu s príslušnými metadátami (URL, rok a nadpis) uložili do súboru vo formáte JSONL.

Vyššie uvedené kroky boli následne implementované do bronzovej dátovej pipeline, ktorá tieto procesy zjednocuje a definuje ich postup. Tieto kroky sú znázornené na obrázku 17. Môžeme si všimnúť, že dátový tok navyše obsahuje kroky, v ktorých nastavujeme premenné **vRunDate** a **vListingPath**. Tie slúžia na zachytenie dátumu a názvu súboru pre konkrétny deň extrakcie.

V prípade striebornej a zlatej vrstvy neuvádzame samostatnú vizualizáciu, keďže tieto časti pozostávajú len z jednej hlavnej transformačnej aktivity.



Obr. 17: EC - Bronze dátový tok

Strieborná vrstva

V ďalšej fáze sme pristúpili k transformácii a predspracovaniu získaných textov. Jednotlivé záznamy sme načítali z JSONL súboru a vykonali sme ich základnú normalizáciu, ktorá zahŕňala najmä odstránenie nadbytočných medzier, prázdnych riadkov a ďalších formálnych nepravidelností. Následne sme pri dlhších textoch overovali, či je možné ich rozdeliť na menšie logické celky. V prípadoch, keď text presahoval stanovený rozsah, sme využili jednoduchý chunking založený na rozdelení podľa logických odstavcov identifikovaných pomocou oddeľovača `\n\n`. Výsledné textové segmenty sme opätovne uložili do nového JSONL súboru, pričom sme doplnili aj identifikáciu príslušného zdrojového odkazu a identifikátor konkrétneho chunku. Zároveň sme v pôvodnom súbore evidovali, ktoré dokumenty už boli spracované, aby sme ich nespracovávali opakovane.

Zlatá vrstva

V poslednej fáze dátového toku sme jednotlivé textové chunky poskytli ako vstup pre vybraný veľký jazykový model. Každý vstup bol doplnený o sadu systémových inštrukcií a o požadovanú štruktúru výstupu, ktorú mal model vrátiť. Príklad očakávaného výstupu a jeho popis môžeme vidieť v časti 5.4 tejto kapitoly. Cieľom tejto fázy bolo získať pre jednotlivé textové úseky konzistentné hodnotenie vo forme definovaných analytických atribútov. Výsledné skóre a ďalšie súvisiace informácie sme z dôvodu spätnej dohľadateľnosti ukladali do JSONL súborov aj po segmentoch, pričom následne prebehla ich agregácia na úrovni jednotlivých dokumentov, resp. odkazov a rokov. Presná metodika agregácie je taktiež popísaná v sekcii 5.4.3.

Vyššie uvedené dátové toky sú následne prepojené a spustené hlavným dátovým tokom, ktorý je zobrazený na obrázku 18.

Obr. 18: EC - Master pipeline

5.5.3 Implementácia dátového toku pre ECB texty

Pri analýze štruktúry webovej stránky Európskej centrálnej banky sme úspešne identifikovali centrálne miesto pre vyhľadávanie noviniek a publikácií. Pri skúmaní zdrojového kódu a počiatočnej extrakcii HTML obsahu sme však zistili, že extrahovaný obsah neobsahuje očakávaný text a vracia prázdny obsah elementu \body.

Podrobnejšou analýzou sme následne odhalili, že obsah stránky je dynamicky načítavaný prostredníctvom JavaScriptu. To znamená, že relevantné údaje nie sú dostupné cez pôvodnú HTML odpoveď servera, ale sú generované až počas načítavania stránky v klientskom prostredí.

Keďže tento spôsob publikovania obsahu komplikuje priamu extrakciu údajov, hľadali sme alternatívny prístup. Ako vhodné riešenie sme identifikovali RSS Feed, ktorý ECB poskytuje ako odporúčaný mechanizmus na získavanie noviniek a publikácií. RSS, celým názvom Really Simple Syndication, predstavuje štandardizovaný XML formát určený na distribúciu aktualizovaného obsahu, ako sú rôzne novinky a správy. RSS tak umožňuje efektívnu a inkrementálnu extrakciu údajov z webových zdrojov.

Nevýhodou tohto riešenia je obmedzená dostupnosť historických záznamov, keďže najstaršie dostupné dáta siahali len približne do polovice februára. Z tohto dôvodu sme implementovali vlastný Python skript na pravidelné odoberanie RSS feedu v mesačných intervaloch, čím sme začali postupne budovať vlastnú databázu tlačových správ.

RSS mechanizmus nám tak poskytol odkazy na jednotlivé tlačové správy spolu s dátumom ich vydania, ktorý sme taktiež uložili do JSON súboru spolu s príslušnými odkazmi. Tento dátum sme na konci procesu využili na agregáciu odkazov podľa mesiacov a rokov.

Pri preskúmaní obsahu viacerých tlačových správ sme identifikovali jednotný vzor, a to, že na začiatku správy sa nachádzalo niekoľko odrážok, ktoré stručne a prehľadne sumarizovali obsah textu a opisovali danú udalosť. Na základe tohto zistenia sme usúdili, že najvhodnejším prístupom bude extrahovať práve toto zhrnutie, čím zároveň zamedzíme duplicite informácií a spracovaniu nadmerného množstva obsahu.

Túto extrakciu sme realizovali taktiež pomocou vlastného Python skriptu, ktorý využíva aj spoločné funkcie s EC, ako napríklad čítanie a zapisovanie súborov z Azure Data Lake Storage Gen2. Týmto prístupom zabezpečujeme vyššiu efektivitu a zároveň centralizovanejšiu správu riešenia v prípade budúcich zmien.

Po úspešnej extrakcii textov z odsekov sme ich aj v tomto prípade podrobili normalizácii a očisteniu od rôznych neviditeľných znakov a aj prípadný chunking, pokiaľ je extrahovaný text odsekov príliš dlhý. Následne keď už bol text normalizovaný, spustili sme naň rovnaký LLM model ale s odlišnými inštrukciami pre lepšiu a presnejšiu odpoveď. Ako posledné sme výstup agregovali a zabezpečili sme rovnakú formu výstupu ako pri EC.

5.6 Bezpečnostné a prevádzkové aspekty riešenia

Pri navrhovaní dátového riešenia treba zohľadniť okrem samotnej implementácie dátových tokov, spracovania dát a účelu aj bezpečnosť a prevádzkové aspekty riešenia. Nami navrhnuté riešenie je implementované s dôrazom na jednoduchosť, transparentnosť a dostatočnú mieru zabezpečenia vzhľadom na charakter spracovaných dát. Keďže riešenie využíva prevažne cloudové služby platformy Microsoft Azure, bolo možné využiť vstavané mechanizmy na riadenie prístupov, monitorovanie, správu výpočtových zdrojov a fakturácie.

5.6.1 Správa prístupov a autentifikácia

V rámci riešenia sú využívané najmä služby Azure Data Factory, Azure Batch a Azure Data Lake Storage Gen2, ktoré sú aj bližšie opísané v časti 5.1.4 vyššie. Prístup k týmto zdrojom je riadený prostredníctvom mechanizmu riadenia prístupov, ktorý sa v skratke označuje ako IAM (Identity and Access Management), pričom k väčšine zdrojov má prístup výhradne len autor riešenia.

Výnimku predstavuje úložisko ADLS, ku ktorému bol poskytnutý prístup aj školiteľovi. Tento prístup bol obmedzený na čítanie prostredníctvom rolí *Reader* a *Storage Blob Data Reader*, čím bola zabezpečená možnosť kontroly dát bez možnosti ich modifikácie.

Pre zabezpečenie komunikácie medzi jednotlivými komponentmi riešenia sme využili spravované identity (Managed Identity). Konkrétne služby Azure Data Factory a Azure Batch boli nakonfigurované tak, aby pristupovali k úložisku prostredníctvom pridelených rolí typu *Storage Blob Data Contributor*. Tento prístup eliminuje potrebu správy prihlasovacích údajov a zvyšuje celkovú bezpečnosť riešenia.

5.6.2 Ochrana dát a ich správa

Z pohľadu zabezpečenia dát bolo riešenie navrhnuté s ohľadom na charakter spracovávaných informácií. Všetky spracovávané dáta pochádzajú z verejne dostupných zdrojov (Európska komisia, Európska centrálna banka, Eurostat), a preto neobsahujú citlivé ani osobné údaje.

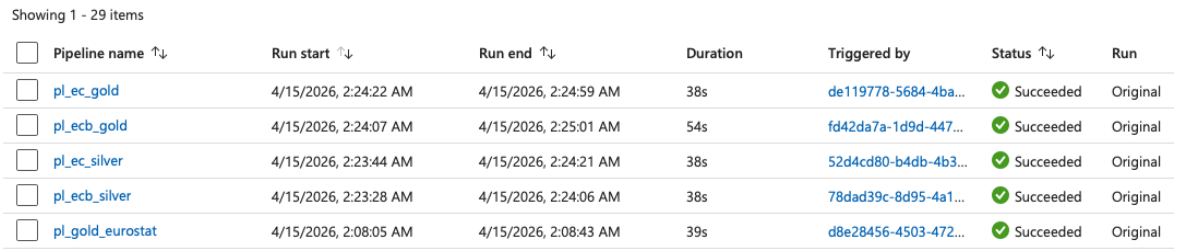
Z tohto dôvodu nebolo potrebné implementovať pokročilé mechanizmy ochrany dát nad rámec štandardných nastavení poskytovaných službou Azure Data Lake Storage Gen2. Tieto nastavenia zahŕňajú najmä kontrolu prístupov na úrovni kontajnerov a objektov, ako aj zabezpečenie integrity a dostupnosti dát.

Napriek tomu boli dáta organizované do logických štruktúr tak, aby bolo možné jednoducho identifikovať jednotlivé vrstvy spracovania a zabezpečiť ich prehľadnosť a konzistentnosť.

5.6.3 Prevádzková spoľahlivosť a spracovanie chýb

Keďže naše riešenie pracuje s externými dátovými zdrojmi, je potrebné zohľadniť aj riziko ich zmeny alebo dočasnej nedostupnosti. V našom riešení pokiaľ dôjde k zlyhaniu jednotlivých

krokov dátového toku, sú k dispozícii podrobné logy v prostredí Azure Data Factory a Azure Batch, ktoré umožňujú identifikáciu príčiny problému a jeho následnú nápravu. Na obrázku 19 je znázornená ukážka úspešného vykonania a zaznamenania dátového toku v ADF.


Showing 1 - 29 items

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run
pl_ec_gold	4/15/2026, 2:24:22 AM	4/15/2026, 2:24:59 AM	38s	de119778-5684-4ba...	Succeeded	Original
pl_ecb_gold	4/15/2026, 2:24:07 AM	4/15/2026, 2:25:01 AM	54s	fd42da7a-1d9d-447...	Succeeded	Original
pl_ec_silver	4/15/2026, 2:23:44 AM	4/15/2026, 2:24:21 AM	38s	52d4cd80-b4db-4b3...	Succeeded	Original
pl_ecb_silver	4/15/2026, 2:23:28 AM	4/15/2026, 2:24:06 AM	38s	78dad39c-8d95-4a1...	Succeeded	Original
pl_gold_eurostat	4/15/2026, 2:08:05 AM	4/15/2026, 2:08:43 AM	39s	d8e28456-4503-472...	Succeeded	Original

Obr. 19: ADF logy

Príkladom takéhoto problému bola zmena názvu stĺpca v jednej z tabuliek poskytovaných prostredníctvom API služby Eurostat, na ktorú sme počas nášho testovania narazili. Táto zmena spôsobila nefunkčnosť existujúceho API dopytu, čo si vyžiadalo jeho úpravu a aktualizáciu. Takéto situácie poukazujú na potrebu priebežnej údržby riešenia pri práci s externými zdrojmi a predstavujú obmedzenie pri plnej automatizácii.

5.6.4 Škálovateľnosť a prevádzková flexibilita

Škálovateľnosť v našom riešení závisí predovšetkým od charakteru vstupných dát. V prípade štruktúrovaných dát získavaných zo služby Eurostat je možné nové dátové zdroje integrovať relatívne jednoducho, a to prostredníctvom pridania príslušného API dopytu do konfiguračného JSON súboru. Ak majú tieto dáta rovnakú štruktúru ako existujúce dátové súbory, nie je potrebné implementovať dodatočnú logiku spracovania.

Naopak, pri spracovaní neštruktúrovaných textových dát je potrebné pristupovať individuálne ku každému zdroju dát, keďže jednotlivé webové stránky sa líšia štruktúrou aj spôsobom publikovania obsahu. Z tohto dôvodu je potrebné implementovať špecifickú extrakčnú logiku pre každý nový zdroj.

Na podporu rozšíriteľnosti riešenia sme vytvorili súbor spoločných Python funkcií, ktoré pokrývajú najčastejšie operácie, ako je čítanie a zapisovanie dát, spracovanie textu alebo komunikácia s úložiskom. Tento prístup zjednodušuje implementáciu nových komponentov a znižuje množstvo duplicitného kódu.

5.7 Limity navrhnutého riešenia

Aj napriek tomu, že naše navrhnuté a implementované riešenie spĺňa stanovené ciele, bolo možné v priebehu jeho tvorby identifikovať viaceré obmedzenia.

Prvým pozorovaným obmedzením je historická dostupnosť dát, resp. ich dostupnosť v strojovo spracovateľnej podobe. Toto obmedzenie sa prejavilo pri tlačových správach Európskej centrálnej banky, ktorá prostredníctvom RSS feedu poskytuje len najnovšie publikované

správy. Takýto spôsob sprístupnenia obsahu obmedzuje možnosti spätnej a dlhodobej analýzy. Z uvedeného dôvodu sme pristúpili k budovaniu vlastnej databázy týchto dát. Získanie dostatočne rozsiahlej vzorky vhodnej pre dlhodobú analýzu je však časovo náročný proces, ktorý nebolo možné v rámci časového horizontu tejto práce plnohodnotne realizovať.

Druhé obmedzenie súvisí s adaptáciou nových textových zdrojov. Na rozdiel od tabuliek zo stránky Eurostat, kde vo všeobecnosti postačuje získať správny API dopyt, overiť dátovú schému a prípadne doplniť nové stĺpce, textové dáta si vyžadujú podstatne vyššiu mieru individuálnej úpravy. Je to spôsobené najmä rôznorodosťou webových stránok, ich štruktúry a spôsobu publikovania obsahu.

Okrem vyššie uvedených obmedzení je potrebné zohľadniť aj limity vyplývajúce zo samotnej metodiky sentimentovej analýzy, ktoré sme podrobnejšie popísali v sekcii 5.4. Výsledky analýzy založenej na LLM modeloch sú ovplyvnené kvalitou vstupných dát a formuláciou inštrukcií, pričom rôzne nastavenia môžu viesť k odlišným výsledkom. Zároveň charakter analyzovaných textov, ktoré sú často formulované neutrálne a opatrne, znižuje variabilitu sentimentu a sťažuje jeho interpretáciu. Obmedzenie predstavuje aj rozdelenie textov na menšie časti, ktoré môže viesť k čiastočnej strate kontextu, pričom agregácia výsledkov predstavuje kompromis medzi presnosťou a technickou realizovateľnosťou.

6 Dátová analýza fiškálnych ukazovateľov

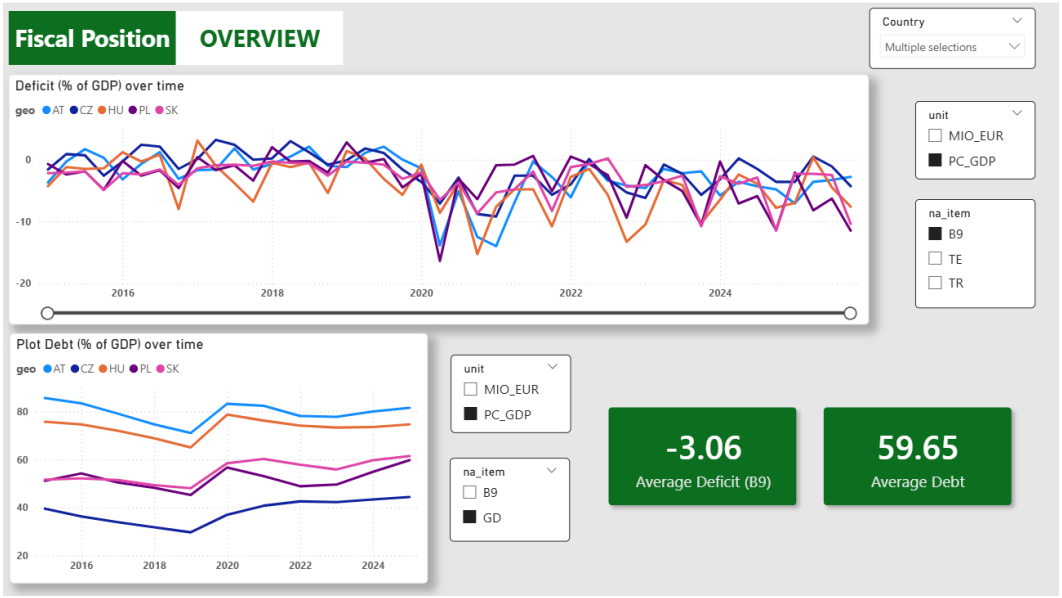
V tejto kapitole si predstavíme výsledky dátovej analýzy fiškálnych ukazovateľov, ktorá slúži ako empirická prípadová štúdia aplikácie navrhovaného dátového riešenia.

Po predstavení vstupných dát v kapitole 4 a opise návrhu a implementácie dátového riešenia v kapitole 5 nadväzujeme exploračnou dátovou analýzou. Jej cieľom je lepšie porozumieť nami spracovaným dátam, preskúmať ich štruktúru, vývoj v čase a základné vzťahy medzi sledovanými ukazovateľmi. Na analýzu našich dát sme využili SSBI nástroj Power BI, v ktorom sme vytvorili viacstranový interaktívny dashboard.

V nasledovných sekciách si postupne predstavíme a popíšeme jednotlivé časti vytvoreného dashboardu.

6.1 Prehľad fiškálnej pozície

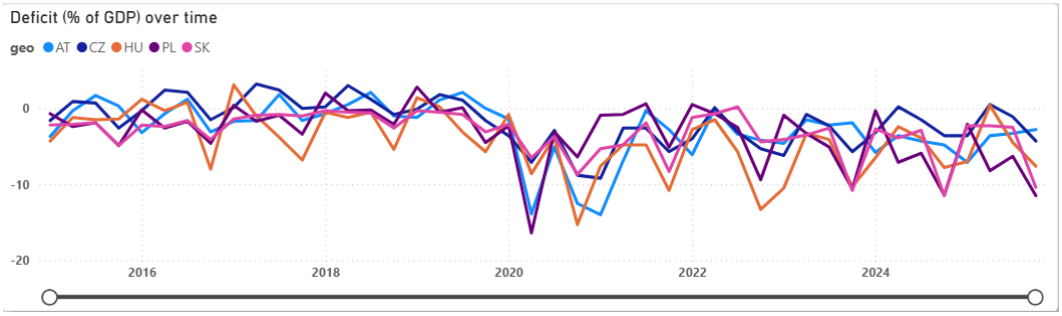
Aby sme získali základný prehľad o fiškálnej situácii, tak sme najprv pripravili deskriptívny pohľad, ktorý môžeme vidieť na obrázku 20 nižšie. Vizualizácia zachytáva vývoj deficitu verejných financií a štátneho dlhu v čase, ako aj ich priemerne hodnoty pre krajiny skupiny V4, konkrétne Slovenska (SK), Českej republiky (CZ), Maďarska (HU), Poľska (PL), a doplnené Rakúsko (AT). Ako sme už spomínali v kapitole 4, analyzované ukazovatele sú vyjadrené v percentách HDP, čo nám umožňuje jednoduchšie a efektívnejšie porovnávať krajín medzi sebou.



Obr. 20: Dashboard fiškálnej pozície

Prvý graf na obrázku 21 nám znázorňuje vývoj deficitu verejných financií vyjadreného v percentách HDP za obdobie od roku 2015 do konca roka 2025 na kvartálnej báze, pričom každý dátový bod reprezentuje hodnotu za príslušný kvartál, resp. štvrťrok. Tento graf zároveň

dopĺňame tabuľkou znázornenou na obrázku 22, v ktorej sumarizujeme minimálne, maximálne a priemerné hodnoty deficitu za tri vybrané obdobia. Vďaka tomu môžeme presnejšie identifikovať kritické fázy vývoja verejných financií.



Obr. 21: Deficit V4 a Rakúska

	2015-2019			2020-2022			2023-2025		
Krajina	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
AT	-3.7	2.1	-0.4	-14	-0.1	-5.89167	-7.1	-1.5	-3.79167
CZ	-2.6	3.2	0.53	-9.2	0.1	-4.56667	-6.2	0.5	-2.63333
HU	-8	3.1	-1.9	-15.3	-0.8	-6.65	-10.5	0.4	-5.625
PL	-4.9	2.8	-1.365	-16.4	0.6	-3.875	-11.5	-0.3	-6.05833
SK	-4.9	-0.3	-1.71	-8.7	0.2	-3.95	-11.5	-2.3	-4.95833

Obr. 22: Deficit tabuľka V4 a Rakúska

Z časového radu môžeme pozorovať, že v období rokov 2015 až 2019 sa analyzované krajiny pohybovali prevažne v miernom deficite, pričom vývoj bol v porovnaní s nasledujúcimi obdobiami relatívne stabilný. Z tabuľky môžeme vidieť, že najpriaznivejšiu priemernú hodnotu dosahovala Česká republika, ktorá ako jediná vykázala priemerný prebytok na úrovni 0,53% HDP. Naopak, najhoršiu priemernú bilanciu dosahovalo Maďarsko, ktorého priemerný deficit predstavoval -1,9% HDP. Z tabuľky zároveň vyplýva, že práve Maďarsko malo v tomto období najväčšiu volatilitu a zaznamenalo aj najvýraznejší pokles, keď jeho minimum dosiahlo -8% HDP. Slovensko malo v predpandemickom období pomerne stabilný vývoj, avšak jeho hodnoty sa pohybovali výlučne v deficite, keďže maximum dosiahlo len -0,3% HDP a priemerný deficit predstavoval -1,71% HDP.

Tento trend bol ale narušený začiatkom roka 2020 v dôsledku pandémie COVID-19. Z grafu je zrejmé, že najvýraznejší nárast deficitu zaznamenalo Poľsko a Rakúsko, neskoršie aj Maďarsko. Potvrdzuje to aj tabuľka, podľa ktorej v období rokov 2020 až 2022 dosiahlo najnižšiu hodnotu Poľsko s minimom -16,4% HDP, nasledované Maďarskom s minimom -15,3% HDP a Rakúskom s hodnotou -14% HDP. Česká republika a Slovensko dosiahli v tomto období miernejšie minimá, konkrétne -9,2% HDP v prípade Česka a -8,7% HDP v prípade Slovenska. Z grafu však môžeme vidieť, že Slovensko vykázalo relatívne silnú rezistenciu najmä začiatkom roka 2020, keď v porovnaní s ostatnými krajinami skupiny V4 a Rakúskom dosahovalo jeden z najnižších deficitov. Potvrdzuje to aj priemerná hodnota

za toto obdobie, keď Slovensko dosiahlo deficit -3,95% HDP. Najhoršie priemerné výsledky dosiahlo Maďarsko s hodnotou -6,65% HDP, nasledované Rakúskom s priemerným deficitom -5,89% HDP.

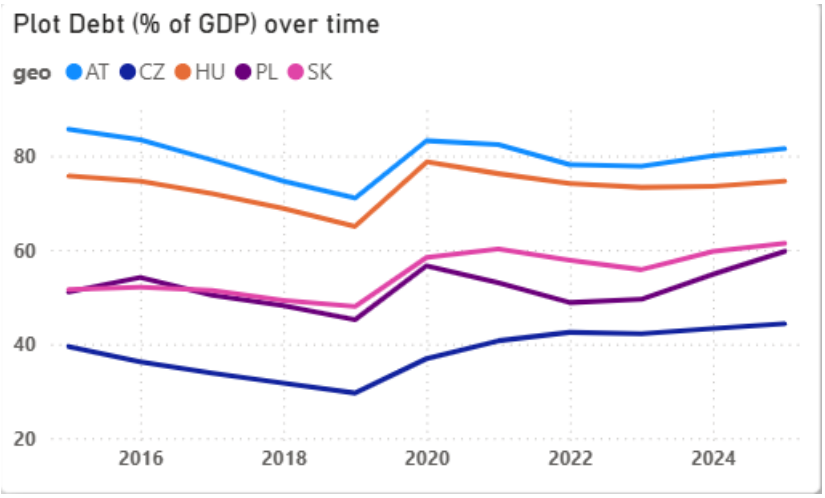
V nasledujúcom období po roku 2022 môžeme pozorovať relatívne rýchle zotavenie Poľska a Rakúska, pričom vo všetkých analyzovaných krajinách zároveň začala prevažovať zvýšená fluktuácia deficitu, čo naznačuje pretrvávajúce fiškálne tlaky. Táto volatilita je pozorovateľná až do konca analyzovaného obdobia, čo potvrdzujú aj minimálne a maximálne hodnoty uvedené v tabuľke. V období rokov 2023 až 2025 môžeme zároveň sledovať, že Slovensko ako jediné dosiahlo v určitom bode horší deficit ako počas pandemického obdobia. Z tabuľky tiež vyplýva, že priemerné hodnoty Poľska a Slovenska boli horšie ako v predchádzajúcom období, konkrétne -6,06% HDP v prípade Poľska a -4,96% HDP v prípade Slovenska. Ostatné krajiny si z hľadiska priemernej hodnoty mierne polepšili.

Zaujímavým zistením je, že hoci niektoré krajiny v jednotlivých kvartáloch obdobia 2023 až 2025 dosiahli krátkodobé zlepšenie, ich priemerné hodnoty zostali výrazne záporné. Česká republika a Maďarsko sa v tomto období dokázali v určitých kvartáloch dostať do mierneho prebytku alebo blízko vyrovnaného hospodárenia, keď maximálne hodnoty dosiahli 0,5% HDP v prípade Česka a 0,4% HDP v prípade Maďarska. Išlo o prvý prípad od začiatku pandémie COVID-19, keď sa fiškálna bilancia Maďarska opäť dostala na úroveň prebytku. Naopak, Slovensko už v období 2023 až 2025 prebytok nedosiahlo a jeho najvyššia hodnota bola stále záporná, konkrétne -2,3% HDP. Rakúsko a Poľsko v rovnakom kvartáli dosahovali vyšší deficit v porovnaní so Slovenskom, čo naznačuje rozdielny priebeh fiškálnej konsolidácie medzi jednotlivými krajinami. Celkovo tak môžeme konštatovať, že slovenská fiškálna bilancia sa po pandémii síce čiastočne stabilizovala, no v porovnaní s predpandemickým obdobím zostala na horšej úrovni.

V závere sledovaného obdobia však môžeme pozorovať opätovné zhoršenie fiškálnej bilancie vo väčšine analyzovaných krajín. Výraznejší pokles je viditeľný najmä v prípade Slovenska a Maďarska, ktoré sa po predchádzajúcom zlepšení opäť dostávajú do hlbšieho deficitu. Rakúsko sa v tomto období odlišuje od vývoja krajín V4, keďže jeho fiškálna bilancia sa naopak zlepšuje a v poslednom dostupnom období vykazuje najpriaznivejšiu hodnotu spomedzi sledovaných krajín. Celkovo možno konštatovať, že zatiaľ čo obdobie 2015–2019 bolo charakteristické relatívnou stabilitou a miernymi deficitmi, obdobie 2020–2022 predstavovalo najvýraznejší šok pre verejné financie. Obdobie 2023–2025 následne ukazuje, že fiškálna konsolidácia bola medzi krajinami nerovnomerná a vo viacerých prípadoch zostali deficity výrazne vyššie než pred pandémiou.

Na grafe zadĺženosti štátov, zobrazenom na grafe 23, pozorujeme vývoj verejného dlhu za obdobie rokov 2015 až 2025, pričom jeden dátový bod reprezentuje ročnú hodnotu v percentách HDP. Už na prvý pohľad si môžeme všimnúť, že aj v tomto prípade dochádza k šoku v roku 2020, podobne ako pri vývoji deficitu verejných financií. Zároveň však vidíme,

že miera nárastu zadlženosti je v porovnaní s deficitom relatívne podobná naprieč všetkými analyzovanými štátmi. Od roku 2020 vidíme, že krajiny si približne zachovávajú rovnakú úroveň dlhu vzhľadom na HDP. ktorú nadobudli počas pandémie, pričom už len mierne fluktuujú.



Obr. 23: Zadĺženosť V4 a Rakúska

Najnižšiu úroveň verejného dlhu spomedzi analyzovaných krajín dlhodobo sledujeme pri Českej republike, s priemernou hodnotou približne 39% HDP. Tento vývoj taktiež podporuje konzistenciu s výsledkami analýzy deficitu verejných financií, kde Česká republika v sledovanom období patrila medzi krajiny s relatívne nižšími a stabilnejšími deficitmi, čo sa postupne premietalo aj teda do priaznivejšej dynamiky verejného dlhu.

Slovensko sa v porovnaní s ostatnými krajinami nachádza približne na priemernej úrovni zadlženosti, avšak z hľadiska vývoja dlhu v čase je možné pozorovať dôležitý medzník. V roku 2024 dosiahla úroveň verejného dlhu Slovenska po druhýkrát v sledovanom období hranicu 60% HDP, pričom prvýkrát k prekročeniu došlo v roku 2021.

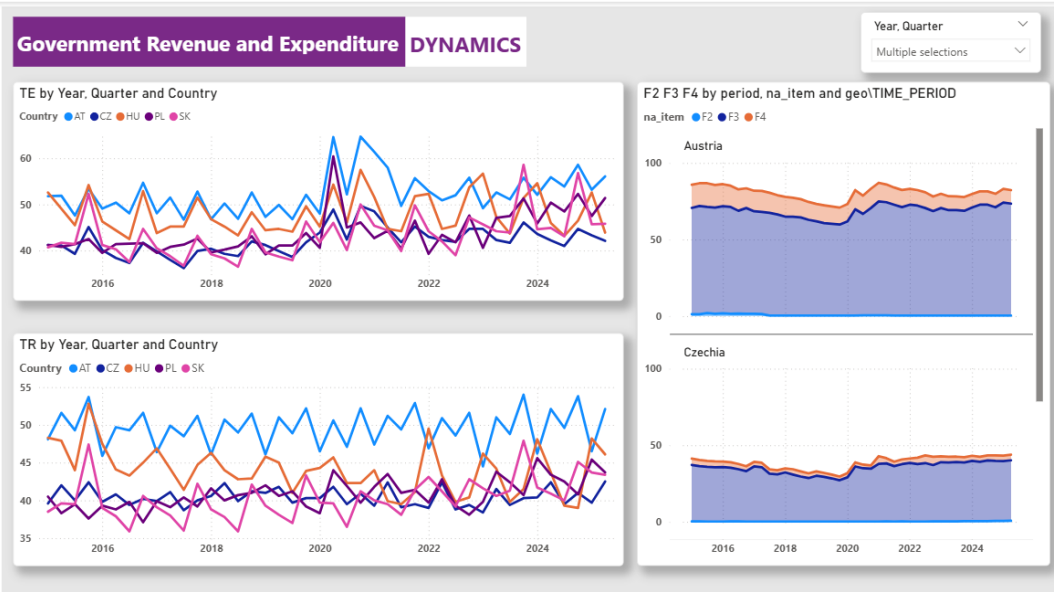
Táto hranica je často označovaná ako jeden z tzv. Maastrichtských konvergenčných kritérií, ktoré definujú základné makroekonomické podmienky pre vstup krajiny do Európskej únie a zároveň slúžia ako referenčný rámec pre hodnotenie dlhodobej udržateľnosti verejných financií [36, 37]. Dosiahnutie tejto úrovne dlhu preto predstavuje významný indikátor z hľadiska fiškálnej stability, aj keď samotná interpretácia jej implikácií presahuje rámec tejto práce.

Tieto grafy ďalej komplementujú priemerné hodnoty deficitu a dlhu. Vidíme, že priemerná hodnota deficitu je na úrovni -3.06 (B9), čo reprezentuje mierny deficit. Pre dlh je priemerná hodnota 59.65, okolo ktorej je aj Slovensko.

6.2 Príjmy a výdavky verejných financií

Aby sme lepšie porozumeli vzťahom zobrazeným na prvom dashboarde popísanom v predchádzajúcej sekcií 6.1, zameriame sa na ďalšiu stránku dashboardu, znázornenú na obrázku

24. Táto stránka zobrazuje časové rady príjmov a výdavkov verejných financií v percentách HDP, pričom rozdiel medzi ich hodnotami určuje výslednú hodnotu ukazovateľa **B9**, teda saldo verejných financií (deficit/prebytok). Zároveň je na tejto stránke zobrazená aj štruktúra verejného dlhu, ktorá poskytuje doplňujúci pohľad na zloženie zadlženosti jednotlivých krajín.



Obr. 24: Dashboard príjmov a výdavkov verejných financií

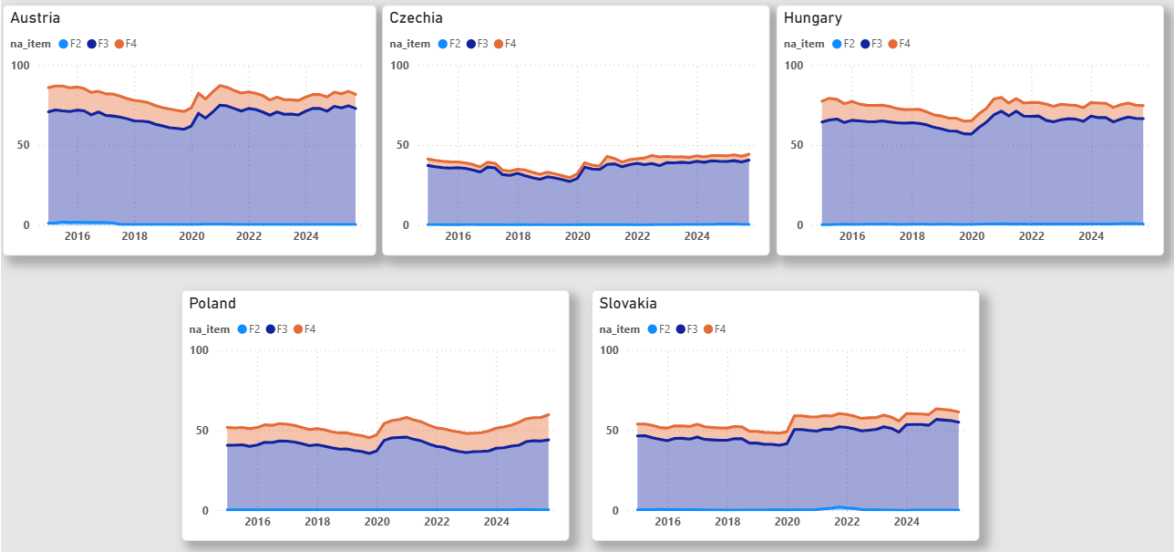
Na grafe výdavkov môžeme opätovne pozorovať výraznejší nárast v roku 2020, zatiaľ čo príjmy verejných financií v tomto období zostávajú približne na rovnakej úrovni. Tento vývoj nám naznačuje, že nárast deficitu v štátoch skupiny V4 a v Rakúsku počas pandémie bol primárne spôsobený zvýšenými verejnými výdavkami, zatiaľ čo príjmy to výrazne neovplyvnilo. Zároveň možno na grafe výdavkov pozorovať pravidelné zvýšenie výdavkov ku koncu jednotlivých rokov, čo môže súvisieť s čerpaním rozpočtových prostriedkov v závere rozpočtového obdobia. Od roku 2020 sa tiež mierne zvyšuje fluktuácia výdavkov, pričom vo všetkých analyzovaných krajinách môžeme pozorovať aj mierny nárast priemernej úrovne výdavkov v porovnaní so začiatkom sledovaného obdobia.

Na grafe príjmov môžeme sledovať dlhodobo stabilné hodnoty v prípade Rakúska, kde sa príjmy aj napriek menšiemu rozptylu pohybujú v priemere okolo úrovne 50 % HDP. Poľsko a Česká republika vykazujú pomerne stabilné a vzájomne porovnateľné hodnoty príjmov vyjadrených v percentách HDP, pričom od roku 2020 dochádza k ich miernemu kolísaniu.

V prípade Maďarska pozorujeme počas sledovaného obdobia pozorovateľný postupný pokles príjmov s občasnými výkyvmi. Naopak, pri Slovensku je možné zaznamenať nárast príjmov v priebehu sledovaného obdobia, ktorý je však sprevádzaný aj paralelným nárastom verejných výdavkov v podobnom tempe.

Posledný graf na tomto dashboarde reprezentuje, ako bolo uvedené vyššie, štruktúru finančných záväzkov jednotlivých štátov. Hlavnými sledovanými kategóriami sú mena a

vklady (F2), dlhové cenné papiere (F3) a úvery (F4). Tento graf môžeme vidieť taktiež znázornený na obrázku 25.



Obr. 25: Štruktúra finančných záväzkov

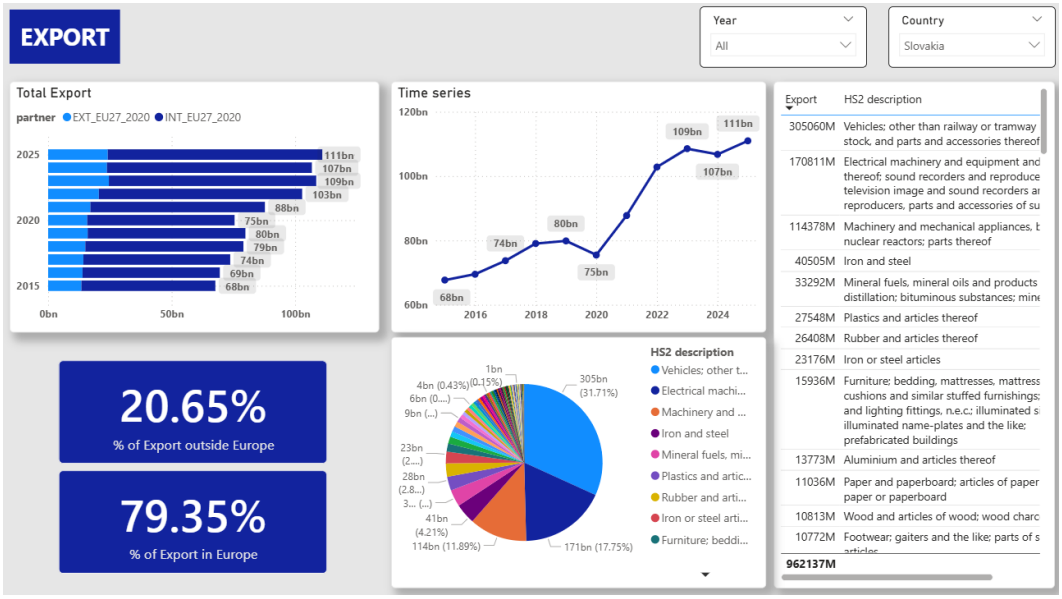
Vo všetkých analyzovaných krajinách predstavujú mena a vklady len zanedbateľnú časť celkových záväzkov, spravidla pod úrovňou 1% HDP. Najväčší podiel finančných záväzkov vo všetkých prípadoch tvoria dlhové cenné papiere, po ktorých nasledujú úvery, ktoré ale tiež tvoria menšiu čiastku. V prípade niektorých krajín, najmä Rakúska, predstavujú úvery relatívne vyšší podiel v porovnaní s ostatnými analyzovanými štátmi. Z týchto pozorovaní nám vyplýva, že všetky pozorované krajiny sa predovšetkým spoliehajú na trhové financovanie dlhu prostredníctvom štátnych dlhopisov, a nie na bankové úvery alebo depozitných nástrojov.

6.3 Zahraničný obchod: Export a Import

V tejto časti si predstavíme a analyzujeme zahraničný obchod, ktorý nám dopĺňa našu fiškálnu a makroekonomickú analýzu prezentovanú v predchádzajúcich sekciách. Zameriavame sa predovšetkým na Slovensko v období od rokov 2015 až 2025. Vytvorený dashboard, ktorý je znázornený na obrázkoch 26 a 27, kombinuje časové rady zahraničného obchodu, geografickú orientáciu obchodu, a produktovú štruktúru exportu a importu.

Na obrázku 26 môžeme pozorovať dashboard znázorňujúci Export. Výsledky tejto stránky jasne ukazujú, že objem medzinárodného obchodu v sledovanom období postupne rástol, s viditeľným poklesov v roku 2020, po ktorom nasledovalo prudké zotavenie a zrýchlenie tempa rastu. Tento trend pretrvával až do roku 2024, v ktorom po prvýkrát od pandémie pozorujeme opätovný pokles objemu exportu. V roku 2025 však opäť vzrástol.

Pri zohľadnení geografickej orientácie exportu vidíme, že prevažná väčšina slovenského vývozu sa uskutočňuje na území Európskej únie. Obchod v rámci 27 členských krajín EU tvorí v priemere 79.35% celkového vývozu, pričom medziročne pozorujeme nárast predov-



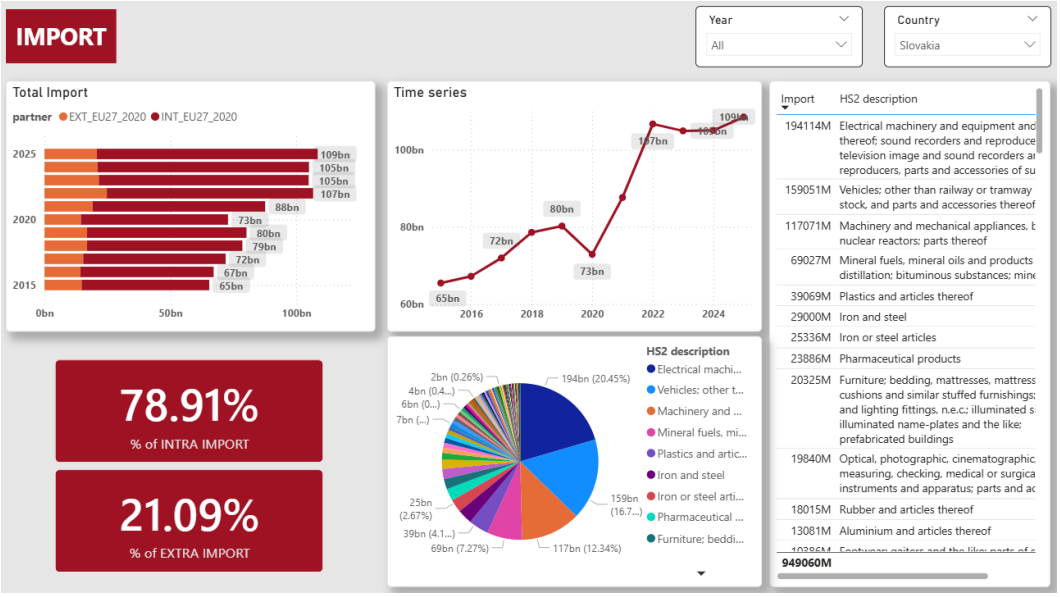
Obr. 26: Export Dashboard

šetkým v tejto časti. Export mimo krajín Európskej únie predstavuje v priemere 20.65% so stabilným rastom v sledovanom období. Z tohto pozorovania nám vyplýva výrazná orientácia slovenského vývozu na trh Európskej únie, ako aj jeho vysoká závislosť od vnútroeurópskych obchodných vzťahov.

Pri analýze produktovej štruktúry vývozu sledujeme, že najvýznamnejšou položkou sú práve automobily, ktoré počas zaznamenaného obdobia reprezentujú takmer 1,8 násobok druhej najväčšej položky. Zároveň vidíme, že automobily tvoria až 31.71% celkového exportu, čo naznačuje silnú orientáciu Slovenska na automobilové odvetvie. Druhou najvýznamnejšou kategóriou je elektrotechnika, ktorá sa na celkovom exporte podieľa približne 18%. Spolu tak tieto dve produktové skupiny tvoria takmer polovicu celkového objemu exportu.

Ak sa zameriame na dovoz znázornený na obrázku 27, môžeme pozorovať, že aj v tomto prípade došlo k výraznému poklesu v roku 2020. Tento pokles ovplyvnil dovoz tak z krajín Európskej únie, ako aj z krajín mimo EÚ, čo je zrejmé z grafu zobrazujúceho geografickú orientáciu dovozu. Avšak v nasledujúcom období je možné pozorovať postupné zotavenie a zrýchlenie tempa rastu dovozu, podobne ako pri vývoze. Geografická štruktúra dovozu je veľmi podobná štruktúre exportu. Dovoz z krajín Európskej únie tvorí v priemere 78,91% celkového importu, zatiaľ čo dovoz z krajín mimo EÚ predstavuje približne 21,09%.

Pri analýze produktovej štruktúry dovozu však už pozorujeme odlišnosti v porovnaní s exportom. Najvýznamnejšou dovážanou položkou je elektronika, ktorá sa na celkovom dovoze podieľa približne 20,45% za sledované obdobie. Druhou najvýznamnejšou kategóriou sú automobily s podielom okolo 17%, pričom na treťom mieste sa nachádzajú stroje a mechanické zariadenia. Tieto kategórie môžu poukazovať nielen na význam technologicky náročných produktov, ale aj na skutočnosť, že viaceré mechanické a elektronické komponenty potrebné



Obr. 27: Import Dashboard

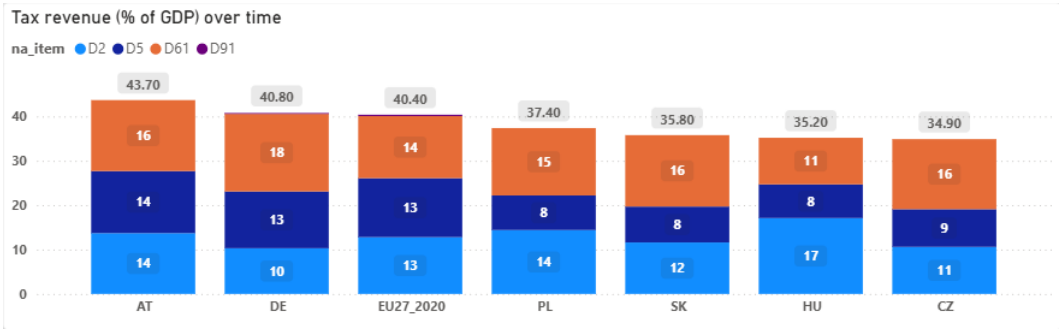
pre automobilový priemysel sú dovážané zo zahraničia.

6.4 Štruktúra daní

V tejto sekcii sa zameriavame na štruktúru daní a analyzujeme postavenie Slovenska v porovnaní s krajinami V4, ako aj s Rakúskom a Nemeckom. Okrem toho sa na údaje pozeráme aj zo širšieho pohľadu a porovnávame ich s európskym priemerom, resp. agregátom EU27_2020. Stĺpce v nižšie uvedenom grafe 28 sú rozdelené podľa typu dane (v tabuľke označené ako na_item). Konkrétne sledujeme dane z produkcie a dovozu (D2), dane z príjmov a majetku (D5), sociálne príspevky (D61) a kapitálové dane (D91).

Graf 28 teda zobrazuje príjmy z daní vyjadrené ako percento HDP. Údaje sú zobrazené pre rok 2024, ktorý bol v čase analýzy najaktuálnejší. Na samotnom grafe si môžeme ako prvé všimnúť celkové daňové zaťaženie krajín. Najvyššie hodnoty dosahujú Rakúsko a Nemecko, konkrétne 43,7% HDP a 40,8% HDP. Európsky priemer (EU27_2020) sa nachádza na úrovni 40,4% HDP, teda len mierne pod úrovňou Nemecka. Slovensko sa nachádza tesne nad Maďarskom s hodnotou 35,8% HDP, pričom nižšiu hodnotu spomedzi sledovaných krajín dosahuje už len Česko (34,9% HDP).

Ako môžeme vidieť, tento rozdiel je do značnej miery spôsobený nižším podielom daní z príjmov a majetku v krajinách V4 v porovnaní s Rakúskom, Nemeckom a európskym priemerom. Zároveň pozorujeme, že Maďarsko vykazuje najvyšší podiel daní z produkcie a dovozu a zároveň najnižší podiel sociálnych príspevkov spomedzi analyzovaných krajín. Toto rozloženie daňových príjmov môže naznačovať väčší dôraz krajín V4 na nepriame dane, čo môže mať implikácie napríklad pre rôzne príjmové skupiny obyvateľstva.



Obr. 28: Štruktúra daní

6.5 Štruktúra výdavkov

V predposlednej časti si predstavíme štruktúru verejných výdavkov podľa klasifikácie COFOG a analyzujeme, ako sú jednotlivé výdavkové kategórie zastúpené v rámci vybraných krajín. Aj v tomto prípade sa venujeme porovnaniu Slovenska s krajinami V4, ako aj s Rakúskom, Nemeckom a agregátom EU27_2020. Jednotlivý význam skratiek funkcií verejných výdavkov COFOG môžeme vidieť v tabuľke 3 nižšie.

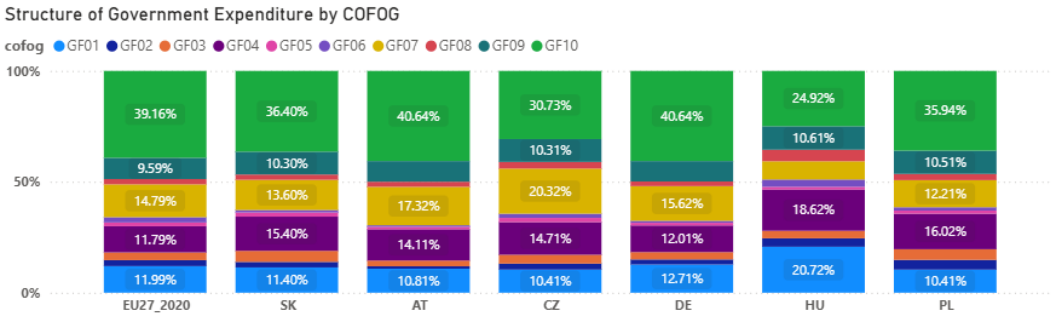
Tabuľka 3: Klasifikácia COFOG – prehľad funkcií verejných výdavkov

Kód	Názov
GF01	Všeobecné verejné služby
GF02	Obrana
GF03	Verejný poriadok a bezpečnosť
GF04	Ekonomické záležitosti
GF05	Ochrana životného prostredia
GF06	Bývanie a občianska vybavenosť
GF07	Zdravotníctvo
GF08	Rekreácia, kultúra a náboženstvo
GF09	Vzdelávanie
GF10	Sociálna ochrana

Graf 29 zobrazuje štruktúru verejných výdavkov za rok 2023, ktorý je pre tento typ dát najaktuálnejší. Na osi X sú znázornené jednotlivé krajiny a na osi Y percentuálne zastúpenie výdavkových kategórií, pričom hodnoty sú vyjadrené ako podiel na celkových výdavkoch (PC_TOT). Vo všetkých analyzovaných prípadoch môžeme pozorovať, že najvýznamnejšiu položku predstavuje sociálna ochrana (GF10), čo je typické pre vyspelé ekonomiky.

Najvyšší podiel sociálnej ochrany dosahujú Nemecko a Rakúsko (obe približne 40,6%), čo naznačuje silný dôraz na sociálny štát. Európsky priemer (EU27_2020) sa pohybuje na úrovni približne 39,2%, pričom Slovensko zaostáva s podielom 36,4%. Ešte nižší podiel pozorujeme v Česku (30,7%) a najnižší v Maďarsku (24,9%), čo môže naznačovať odlišný prístup k sociálnym politikám.

Druhou významnou kategóriou sú výdavky na zdravotníctvo (GF07). Najvyšší podiel



Obr. 29: Štruktúra výdavkov

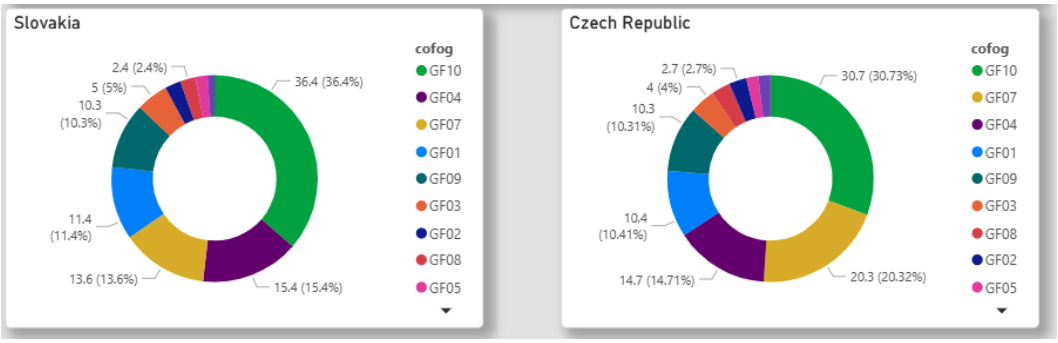
dosahuje Česko (20,3%), zatiaľ čo ostatné krajiny sa pohybujú približne v rozmedzí 12 – 17%. Slovensko alokuje na zdravotníctvo približne 13,6%, čo je mierne pod úrovňou niektorých porovnávaných krajín.

Významnú úlohu zohrávajú aj ekonomické záležitosti (GF04). Najvyšší podiel zaznamenávame v Maďarsku (18,6%) a Poľsku (16,0%), čo môže indikovať väčší dôraz na podporu ekonomiky, infraštruktúry alebo investičných stimulov. Slovensko sa v tejto kategórii nachádza na úrovni 15,4%, čo je porovnateľné s ostatnými krajinami V4.

V oblasti vzdelávania (GF09) sú rozdiely medzi krajinami relatívne malé, pričom väčšina sa pohybuje okolo 10%. Slovensko dosahuje približne 10,3%, čo je blízke európskemu priemeru.

Zaujímavé rozdiely možno pozorovať aj v kategórii všeobecných verejných služieb (GF01). Maďarsko výrazne vyčnieva s podielom 20,7%, čo je takmer dvojnásobok oproti ostatným krajinám. Tento jav môže súvisieť napríklad s vyššími nákladmi na správu štátu alebo obsluhu verejného dlhu.

Pri detailnejšom pohľade na štruktúru verejných výdavkov Slovenska a Českej republiky, zobrazenom na grafe 30, môžeme identifikovať viacero významných rozdielov v prioritách štátov.



Obr. 30: Štruktúra výdavkov SK vs CZ

Najvýraznejší rozdiel pozorujeme v kategórii sociálnej ochrany (GF10), kde Slovensko dosahuje podiel 36,4%, zatiaľ čo Česko 30,7%. Tento rozdiel naznačuje, že Slovensko alokuje väčšiu časť verejných zdrojov na sociálne transfery, ako sú dôchodky či sociálne dávky.

Naopak, Česko výrazne prevyšuje Slovensko vo výdavkoch na zdravotníctvo (GF07), kde dosahuje až 20,3%, kým Slovensko len 13,6%. Tento rozdiel môže indikovať vyššiu prioritu zdravotného systému v Českej republike alebo odlišnú štruktúru jeho financovania.

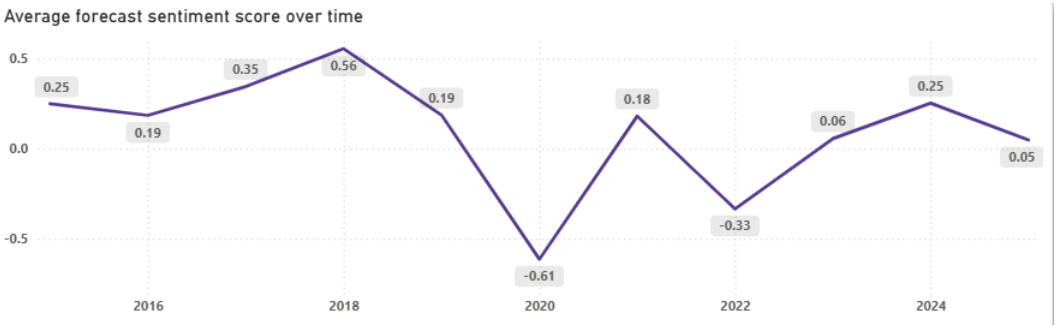
V ostatných kategóriách, ako napríklad vzdelávanie (GF09) alebo obrana (GF02), dosahujú obe krajiny relatívne podobné hodnoty.

6.6 Analýza sentimentu

V tejto sekcii sa zameriame na analýzu sentimentu dátových súborov Európskej komisie a Európskej centrálnej banky, ktorú sme spracovali v rámci kapitoly 5.

6.6.1 Sentiment ekonomických prognóz

Na grafe 31 môžeme pozorovať vývoj priemerného sentimentu ekonomických prognóz za posledné desaťročie. Z časového priebehu môžeme identifikovať niekoľko výrazných fáz.



Obr. 31: Vývoj sentimentu - EC

V období rokov 2015 až 2018 môžeme pozorovať postupný rast sentimentu, ktorý nadobúda maximum v roku 2018 na hodnote 0,56. Toto obdobie naznačuje fázu ekonomického optimizmu, pravdepodobne súvisiacu s priaznivým makroekonomickým vývojom v Európe, ktorý sme mohli pozorovať aj pri fiškálnych dátach.

Po tomto vrchole si však môžeme všimnúť výrazný pokles sentimentu, ktorý prechádza z pozitívnych hodnôt v rokoch 2018 a 2019 do výrazne negatívnych v roku 2020, keď dosahuje hodnotu -0,61, čo zároveň predstavuje najnižší bod celého sledovaného obdobia. Tento prudký prepad silno indikuje práve vplyv pandémie Covid-19 a s ňou súvisiacou ekonomickou neistotou.

V roku 2021 dochádza k výraznému oživeniu sentimentu na hodnotu približne 0,18, čo naznačuje čiastočné zotavenie ekonomických očakávaní. Tento trend však nie je stabilný, keďže v roku 2022 sentiment opäť klesá do negatívnych hodnôt (-0,33), pravdepodobne v dôsledku geopolitickej neistoty a energetickej krízy.

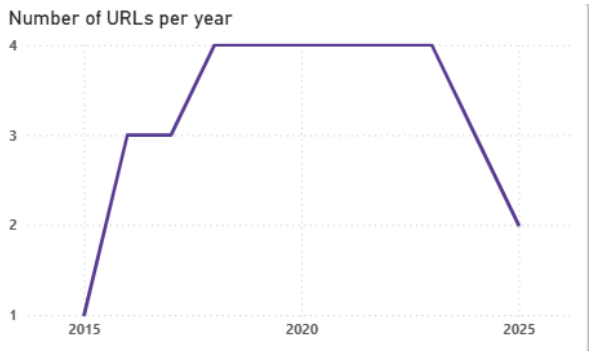
V nasledujúcich rokoch (2023 – 2025) môžeme pozorovať postupné zlepšovanie sentimentu, pričom hodnoty sa stabilizujú v mierne pozitívnom pásme (okolo 0,05 – 0,25). To naznačuje opatrný optimizmus, avšak bez návratu na úroveň predpandemického obdobia.

Analyzovaný vývoj sentimentu naznačuje, že ekonomické očakávania reagujú na vonkajšie vplyvy skôr než samotné fiškálne ukazovatele. Napríklad prudký pokles sentimentu v roku 2020 predchádzal zhoršeniu verejných financií v dôsledku pandémie COVID-19. Rovnako aj pokles v roku 2022 odráža rastúcu neistotu, ktorá sa následne premietla do fiškálneho vývoja. Na základe týchto zistení možno konštatovať, že sentiment môže slúžiť ako vhodný indikátor pri analýze fiškálnej situácie.

Celkové priemerné sentimentové skóre za pozorované obdobie dosahuje hodnotu 0,1. Táto hodnota sa nachádza blízko neutrálnej úrovne, čo naznačuje, že analyzované ekonomické prognózy Európskej komisie majú z dlhodobého hľadiska prevažne neutrálny charakter. Napriek výrazným medziročným výkyvom tak celkový priemer nepoukazuje na jednoznačne pozitívne ani negatívne ladenie obsahu.

Na druhej strane je potrebné zdôrazniť, že sentiment vykazuje vyššiu volatilitu a môže byť ovplyvnený krátkodobými faktormi a udalosťami. Z tohto dôvodu je jeho využitie vhodné najmä ako doplnkový indikátor, ktorý rozširuje interpretáciu fiškálnych dát.

Zároveň, ako môžeme vidieť na grafe 32, jednotlivé roky obsahujú rozdielny počet prognóz, čo môže ovplyvniť výsledný sentiment. Tento aspekt, spolu s charakterom použitých LLM modelov, ako bolo uvedené v kapitole 5, predstavuje jedno z obmedzení tejto analýzy.

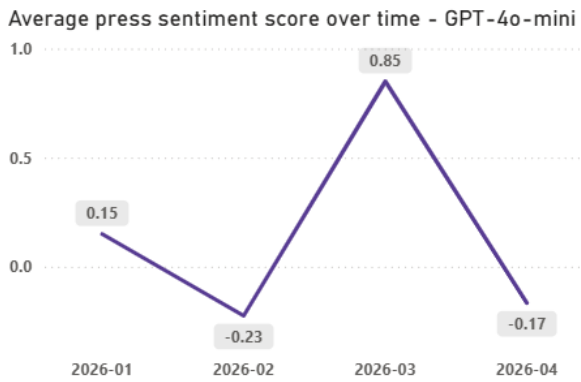


Obr. 32: Počet URL - EC

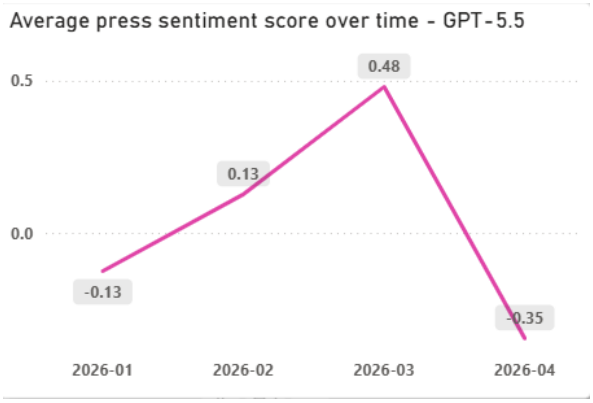
6.6.2 Sentiment tlačových správ

V tejto časti sa zameriame na vývoj sentimentu tlačových správ Európskej centrálnej banky, pričom sentiment sme analyzovali pomocou dvoch jazykových modelov, konkrétne *GPT-4o-mini* a *GPT-5.5*. Výsledky sú zobrazené na grafoch 33 a 34 nižšie.

Z grafov môžeme pozorovať, že analyzované obdobie zachytáva vývoj od januára do apríla 2026, pričom sentiment zobrazujeme na mesačnej granularite. Dôvod kratšieho časového horizontu sme bližšie popísali v kapitole 4. Keďže ide o krátkodobý vývoj, mesačná granularita nám umožňuje lepšie zachytiť zmeny v komunikačnom tóne ECB. Pri interpretácii výsledkov je však potrebné zohľadniť aj počet dostupných textov, ktorý predstavoval v priemere 6,5 tlačovej správy za mesiac.



Obr. 33: Vývoj sentimentu - ECB - GPT-4o-mini



Obr. 34: Vývoj sentimentu - ECB - GPT-5.5

Zo samotného vývoja priemerného sentimentového skóre môžeme v oboch prípadoch sledovať výrazné zmeny v komunikačnom tóne. Pri modeli GPT-4o-mini dosahovalo priemerné sentimentové skóre v januári hodnotu 0,15, čo naznačuje mierne pozitívny tón tlačových správ. Vo februári však sentiment klesol na hodnotu -0,23, čím sa dostal do mierne negatívneho pásma. Výraznú zmenu môžeme sledovať v marci, keď sentiment prudko vzrástol na hodnotu 0,85. Tento nárast poukazuje na podstatne pozitívnejšiu formuláciu obsahu tlačových správ v danom mesiaci. V apríli sa sentiment opäť znížil na hodnotu -0,17, čo naznačuje návrat k mierne negatívnemu tónu.

Pri modeli GPT-5.5 môžeme sledovať mierne odlišnú výšku hodnôt, ako aj čiastočne odlišný priebeh sentimentu. V januári dosahoval sentiment hodnotu -0,13, zatiaľ čo vo februári vzrástol na 0,13. Najvyššiu hodnotu dosiahol rovnako ako pri modeli GPT-4o-mini v marci, a to na úrovni 0,48. V apríli následne sentiment klesol na hodnotu -0,35, čo predstavuje najnižšiu hodnotu v celom sledovanom období podľa modelu GPT-5.5.

Porovnanie oboch modelov ukazuje, že napriek rozdielnym absolútnym hodnotám zachytávajú v marci a apríli podobný smer vývoja sentimentu. Naopak, rozdiel môžeme sledovať vo februári, keď pri modeli GPT-4o-mini sentiment klesol, zatiaľ čo pri modeli GPT-5.5 vzrástol a pokračoval v raste až do marca. V oboch prípadoch však môžeme sledovať maximum práve v marci, čo naznačuje pozitívnejší tón tlačových správ v tomto mesiaci. V apríli následne vi-

díme prudký pokles, pričom sa hodnoty oboch modelov dostávajú do záporného pásma. Tento pokles môže súvisieť s prebiehajúcimi konfliktmi vo svete, konkrétne s napätím spojeným s Hormuzským prielivom.

Výsledky preto poukazujú nielen na zmeny v komunikačnom tóne tlačových správ ECB, ale aj na dôležitosť voľby LLM modelu pri sentimentovej analýze. Zároveň môžeme sledovať krátkodobejší charakter a vyššiu intenzitu zmien pri tlačových správach v porovnaní s ekonomickými prognózami.

6.7 Zhrnutie analýzy

V tejto kapitole sme si prostredníctvom dátovej analýzy analyzovali vybrané fiškálne a makroekonomické ukazovatele, ako aj štruktúru zahraničného obchodu, s cieľom demonštrovať praktickú aplikáciu navrhovaného dátového riešenia. Analýzu sme realizovali prostredníctvom interaktívnych dashboardov v nástroji Power BI, ktoré nám poslúžili ako validačná a analytická vrstva nad spracovanými dátami.

Samotné výsledky fiškálnej analýzy poukázali na výrazný vplyv pandémie COVID-19 na vývoj deficitu a verejného dlhu vo všetkých analyzovaných krajinách. Zatiaľ čo nárast deficitu bol v pandemickom období primárne ovplyvnený zvýšenými verejnými výdavkami, vývoj verejného dlhu vykazoval naprieč krajinami podobnejšiu dynamiku a po roku 2020 sa stabilizoval na zvýšených úrovniach.

Analýza zahraničného obchodu Slovenska doplnila fiškálny kontext o pohľad na externé ekonomické vzťahy. Výsledky potvrdili silnú orientáciu slovenského exportu aj importu na trh Európskej únie, ako aj význam automobilového a elektrotechnického priemyslu v produktovej štruktúre zahraničného obchodu. Pandemické obdobie sa prejavilo dočasným poklesom obchodnej aktivity, po ktorom nasledovalo výrazné zotavenie a následná stabilizácia obchodných tokov.

Výsledky analýzy štruktúry daní ukázali, že Slovensko dosahuje v porovnaní s Rakúskom, Nemeckom a priemerom EÚ nižšie celkové daňové zaťaženie. Tento rozdiel je spôsobený najmä nižším podielom daní z príjmov a majetku, čo je charakteristické aj pre ostatné krajiny V4. Zároveň sa ukázalo, že jednotlivé krajiny sa líšia nielen celkovou úrovňou daňových príjmov, ale aj ich vnútornou štruktúrou, čo môže odrážať rozdielne nastavenie daňovej politiky a odlišný dôraz na priame a nepriame dane.

Analýza štruktúry verejných výdavkov podľa klasifikácie COFOG poskytla detailnejší pohľad na výdavkové priority sledovaných krajín. Vo všetkých analyzovaných prípadoch predstavovala najväčšiu výdavkovú položku sociálna ochrana, čo potvrdzuje jej kľúčové postavenie vo verejných financiách európskych krajín. Zároveň sa však ukázali rozdiely medzi jednotlivými krajinami, najmä pri porovnaní Slovenska, Českej republiky a Maďarska. Slovensko alokuje vyšší podiel výdavkov na sociálnu ochranu, zatiaľ čo Česko výraznejšie smeruje verejné zdroje do zdravotníctva a Maďarsko do všeobecných verejných služieb. Tieto

rozdiely tak poukazujú na odlišné výdavkové priority aj medzi susednými krajinami.

Doplnkovú analytickú vrstvu predstavovala sentimentová analýza textových dát Európskej komisie a Európskej centrálnej banky. V prípade ekonomických prognóz Európskej komisie sa ukázalo, že sentiment v sledovanom období reagoval citlivo na významné ekonomické a geopolitické udalosti. Najvýraznejší pokles bol zaznamenaný v roku 2020, čo korešponduje s obdobím pandémie COVID-19, zatiaľ čo v nasledujúcich rokoch dochádzalo k postupnému zlepšovaniu sentimentu. Celkové priemerné skóre na úrovni 0,1 však naznačuje prevažne neutrálny charakter analyzovaných textov z dlhodobého hľadiska.

Pri tlačových správach Európskej centrálnej banky sme pracovali s kratším časovým horizontom, čo obmedzuje možnosti dlhodobej interpretácie. Napriek tomu výsledky poukázali na zmenu komunikačného tónu v priebehu prvých mesiacov roka 2026. Pri oboch použitých modeloch, GPT-4o-mini a GPT-5.5, sme mohli pozorovať najvyššie hodnoty sentimentu v marci, po ktorých nasledoval pokles v apríli do záporného pásma. Rozdiely medzi modelmi sa prejavili najmä v absolútnej výške sentimentového skóre a v priebehu hodnôt na začiatku sledovaného obdobia, čo poukazuje na význam voľby LLM modelu pri interpretácii sentimentovej analýzy. Tento vývoj zároveň naznačuje, že textové dáta môžu vhodne dopĺňať kvantitatívne ukazovatele a rozširovať analytický pohľad o kvalitatívny rozmer komunikácie európskych inštitúcií.

Diskusia

V tejto kapitole zhrnieme hlavné kroky realizované v rámci práce, zhodnotíme naplnenie stanovených cieľov a poukážeme na limity navrhnutého riešenia, ako aj na možnosti jeho ďalšieho rozvoja.

V úvode práce sme začali prehľadom existujúcej literatúry zameranej na oblasť nástrojov typu SSBI, cloudové služby a prístupy k analýze sentimentu. Pri tejto analýze dostupnej literatúry sme narazili na viacero nedostatkov, ktoré následne formovali ciele našej práce. Konkrétne boli identifikované najmä tieto problematické oblasti: izolované využívanie vizualizačnej vrstvy pri SSBI nástrojoch bez hlbšej integrácie dátových procesov, prevažne konceptuálny charakter akademických prác zameraných na cloudové služby s minimálnym dôrazom na praktickú implementáciu end-to-end riešení, ako aj nedostatočné riešenie integrácie sentimentových modelov do komplexných dátových analytických systémov.

Z týchto dôvodov sme si stanovili ako hlavný cieľ tejto práce navrhnúť a implementovať cloudové dátové riešenie, ktoré prepája zber dát, ich spracovanie, analytické metódy a vizualizačnú vrstvu do jedného uceleného analytického rámca.

Po definovaní cieľa sme následne identifikovali vhodné dátové zdroje. Konkrétne sme identifikovali 10 relačných dátových súborov dostupných prostredníctvom webovej platformy Eurostat. Ich vhodnosť sme otestovali úvodným stiahnutím vo formáte TSV a vytvorením vizualizácií v Power BI. Následne sme identifikovali textové dáta, konkrétne ekonomické prognózy z Európskej komisie a tlačové správy z Európskej centrálnej banky.

V ďalšom kroku sme navrhli a následne implementovali end-to-end cloudové riešenie v prostredí Microsoft Azure. Navrhnuté riešenie pokrýva celý proces od získavania vstupných dát cez ich uloženie, historizáciu, čistenie, transformáciu, obohatenie a monitoring, až po ich sprístupnenie vo vizualizačnej vrstve v nástroji Power BI. Z technologického hľadiska sme využili najmä služby Azure Data Factory, Azure Data Lake Storage Gen2 a Azure Batch. Ich kombinácia umožnila riadenie dátových tokov, spúšťanie vlastných Python skriptov pri zložitejších transformáciách a integráciu LLM modelu do analytického procesu. Zároveň sme tým vytvorili riešenie, ktoré je modularizované, prehľadné a ďalej rozšíriteľné.

V rámci implementácie dátových tokov sa ukázalo, že štruktúrované dáta z Eurostatu sú vhodné na automatizované spracovanie, keďže sú dostupné prostredníctvom API rozhrania a majú naprieč tabuľkami relatívne konzistentnú štruktúru. Vďaka parametrizácii dátových tokov sme vytvorili riešenie, ktoré umožňuje spracovanie viacerých dátových súborov bez potreby manuálnej tvorby samostatných procesov pre každú tabuľku. Tento prístup zároveň podporuje budúce rozšírenie riešenia o ďalšie dátové zdroje s podobnou štruktúrou.

Pri práci s textovými dátami sa naopak prejavila vyššia technická aj metodická náročnosť. Webové zdroje Európskej komisie a Európskej centrálnej banky sa líšili štruktúrou, spôsobom publikovania obsahu aj dostupnosťou historických záznamov. Pri ekonomických prognózach Európskej komisie bolo možné využiť extrakciu odkazov a textov priamo z webovej stránky,

zatiaľ čo pri tlačových správach ECB sme z dôvodu dynamického načítavania obsahu zvolili odoberanie RSS feedu. Tento rozdiel ukázal, že spracovanie textových dát si pri každom novom zdroji vyžaduje individuálny prístup.

Empirická analýza fiškálnej situácie slúžila predovšetkým ako prípadová štúdia na overenie použiteľnosti navrhnutého riešenia. Výsledky ukázali, že vytvorené dátové toky a Power BI dashboardy umožňujú analyzovať fiškálne a makroekonomické ukazovatele naprieč krajinami a časom. V rámci analýzy sme identifikovali najmä vplyv pandémie COVID-19 na vývoj deficitu, verejného dlhu a verejných výdavkov, ako aj rozdiely medzi krajinami V4 a Rakúskom. Analýza zahraničného obchodu zároveň doplnila fiškálny pohľad o širší makroekonomický kontext Slovenska a poukázala na jeho silnú orientáciu na európsky trh a automobilový priemysel. Sentimentová analýza textových dát ukázala, že kvalitatívne dáta môžu vhodne dopĺňať kvantitatívne ukazovatele, avšak najmä ako doplnkový indikátor. Zároveň sa ukázalo, že ekonomické prognózy Európskej komisie mali z dlhodobého hľadiska prevažne neutrálny charakter, keďže priemerné sentimentové skóre za analyzované obdobie dosiahlo hodnotu 0,1.

Napriek splneniu stanovených cieľov sme počas tvorby riešenia identifikovali niekoľko limitácií. Prvým limitom je obmedzená historická dostupnosť pri tlačových správach ECB. Keďže RSS feed poskytoval iba obmedzené množstvo historických záznamov, bolo možné sledovať obdobie za posledných pár mesiacov. Tento limit je možné do budúcnosti zmierniť pravidelným odoberaním RSS feedu a postupným budovaním vlastnej historickej databázy, alebo preskúmaním alternatívnych extrakčných možností.

Druhým limitom je závislosť od dostupnosti a stability externých webových zdrojov. V našom prípade sme sa stretli so zmenou logiky pri tabuľke DS-045409, ktorá znefunkčnila naše API volanie a bolo treba ho poupraviť. Zároveň sme narazili aj na občasné odstávky webovej stránky alebo oneskorené odpovede API rozhrania.

Tretím limitom sú metodické obmedzenia sentimentovej analýzy. Použitý LLM model umožňuje lepšie zachytiť kontext formálnych textov než jednoduché lexikónové metódy, avšak výsledky môžu byť ovplyvnené formuláciou inštrukcií, štruktúrou vstupného textu, jeho rozdelením na menšie segmenty a spôsobom agregácie skóre, ako aj výberom konkrétneho LLM modelu. Sentimentové skóre preto nemožno chápať ako objektívne meranie ekonomickej nálady, ale skôr ako aproximáciu tónu analyzovaných textov.

Ďalším obmedzením je samotný charakter inštitucionálnych textov. Dokumenty Európskej komisie a tlačové správy ECB sú formulované odborným, formálnym a často neutrálnym jazykom. To prirodzene znižuje variabilitu sentimentového skóre a môže spôsobovať, že aj významné obsahové rozdiely sa v numerickom skóre prejavia len mierne. Z tohto dôvodu je sentimentová analýza v tejto práci chápaná ako doplnková analytická vrstva, nie ako hlavný nástroj hodnotenia fiškálnej alebo makroekonomickej situácie.

Na základe dosiahnutých výsledkov môžeme konštatovať, že ciele práce boli splnené. Podarilo sa nám navrhnúť a implementovať end-to-end cloudové riešenie, ktoré pokrýva celý

proces od zberu dát až po ich vizualizáciu. Zároveň sa podarilo prepojiť štruktúrované dáta z Eurostatu s textovými dátami európskych inštitúcií, čím sa rozšíril analytický pohľad na skúmanú problematiku. Integrácia LLM modelu do dátového toku ukázala, že sentimentová analýza môže byť súčasťou širšieho dátového riešenia a nemusí predstavovať izolovanú analytickú úlohu. Navrhnuté riešenie je zároveň flexibilné a umožňuje jeho ďalšie rozšírenie o nové dátové zdroje, transformačné postupy alebo pokročilejšie analytické metódy.

Z hľadiska ďalšieho rozvoja by bolo možné riešenie rozšíriť o ďalšie dátové zdroje, napríklad správy ministerstiev financií, národných bánk alebo medzinárodných organizácií. V oblasti analytických metód by bolo možné porovnať účinnosť rôznych LLM modelov. Z technologického hľadiska by ďalším krokom mohlo byť rozšírenie monitorovania dátových tokov, zavedenie pokročilejších validačných pravidiel alebo implementácia automatických notifikácií pri zlyhaní spracovania.

Použitá literatúra

1. TANG, Jinxuan; MA, Tianjun; LUO, Qianwen. Trends prediction of big data: a case study based on fusion data. *Procedia Computer Science*. 2020, roč. 174, s. 181–190.
2. CHI, Tin Wong; MAHMUD, Imran. Business intelligence system adoption: A systematic literature review of two decades. *International Journal of Industrial Management*. 2020, roč. 6, s. 1–8.
3. BASILE, Luigi Jesus; CARBONARA, Nunzia; PELLEGRINO, Roberta; PANNIELLO, Umberto. Business intelligence in the healthcare industry: The utilization of a data-driven approach to support clinical decision making. *Technovation*. 2023, roč. 120, s. 102482.
4. TIRNO, Rabbir Rashedin. Effect of business intelligence on organizational competitiveness-exploring the mediation of technology anxiety. *Computers in Human Behavior Reports*. 2024, roč. 16, s. 100536.
5. OGEAWUCHI, JEFFREY CHIDERA; UZOKA, ABEL CHUKWUEMEKE; ABAYOMI, AA; AGBOOLA, OA; GBENLE, TOLUWASE PETER; AJAYI, OLANREWAJU OLUWASEUN. Innovations in Data Modeling and Transformation for Scalable Business Intelligence on Modern Cloud Platforms. *Iconic Res. Eng. J.* 2021, roč. 5, č. 5, s. 406–415.
6. VO, Quoc Duy; THOMAS, Jaya; CHO, Shinyoung; DE, Pradipta; CHOI, Bong Jun; SAEL, Lee. Next generation business intelligence and analytics: a survey. *arXiv preprint arXiv:1704.03402*. 2017.
7. PAŁYS, Marcin; PAŁYS, Andrzej. Benefits and challenges of self-service business intelligence implementation. *Procedia Computer Science*. 2023, roč. 225, s. 795–803.
8. LENNERHOLT, Christian; VAN LAERE, Joeri; SÖDERSTRÖM, Eva. User-related challenges of self-service business intelligence. *Information Systems Management*. 2021, roč. 38, č. 4, s. 309–323.
9. TOIC, Andrea; POSCIC, Patrizia; JAKSIC, Danijela. Analysis of selected business intelligence data visualization tools. In: *Central European Conference on Information and Intelligent Systems*. Faculty of Organization a Informatics Varazdin, 2022, s. 25–32.
10. MISHRA, Adya. The Role of Data Visualization Tools in Real-Time Reporting: Comparing Tableau, Power BI, and Qlik Sense. *IJSAT-International Journal on Science and Technology*. 2020, roč. 11, č. 3.
11. DOKO, Fisnik; MISKOVSKI, Igor. Advanced analytics of big data using power BI: credit registry use case. In: *17th International Conference on Informatics and Information Technologies, Skopje, North Macedonia, May*. 2020, zv. 8.

12. BANY MOHAMMAD, Ashraf; AL-OKAILY, Manaf; AL-MAJALI, Mohammad; MASA'DEH, Ra'ed. Business intelligence and analytics (BIA) usage in the banking industry sector: an application of the TOE framework. *Journal of Open Innovation: Technology, Market, and Complexity*. 2022, roč. 8, č. 4, s. 189.

13. OGEAWUCHI, JC; UZOKA, AC; ALOZIE, CE; AGBOOLA, OA; OWOADE, S; AKPE, OEE. Next-generation data pipeline automation for enhancing efficiency and scalability in business intelligence systems. *International Journal of Social Science Exceptional Research*. 2022, roč. 1, č. 1, s. 277–282.

14. LI, Ang; YANG, Xiaowei; KANDULA, Srikanth; ZHANG, Ming. CloudCmp: comparing public cloud providers. In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010, s. 1–14.

15. BORRA, Praveen. Comparison and analysis of leading cloud service providers (AWS, Azure and GCP). *International Journal of Advanced Research in Engineering and Technology (IJARET) Volume*. 2024, roč. 15, s. 266–278.

16. TALAKOLA, Swetha. Analytics and reporting with Google Cloud platform and Microsoft Power BI. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*. 2022, roč. 3, č. 2, s. 43–52.

17. CHILLARA, Srimurali Krishna. Leveraging cloud-based BI architecture for scalable healthcare analytics: A technical framework for transformation. *World Journal of Advanced Engineering Technology and Sciences*. 2025, roč. 15, č. 01, s. 2291–2300. Dostupné z DOI: 10.30574/wjaets.2025.15.1.0489.

18. KRUGMANN, Jan Ole; HARTMANN, Jochen. Sentiment analysis in the age of generative AI. *Customer Needs and Solutions*. 2024, roč. 11, č. 1, s. 3.

19. ANNAN, Augustine; EIDEN, Amanda L; WANG, Dong; DU, Jingcheng; RASTEGAR-MOJARAD, Majid; NOMULA, Varun Kumar; WANG, Xiaoyan. Evaluating large language models for sentiment analysis and hesitancy analysis on vaccine posts from social media: Qualitative study. *JMIR Formative Research*. 2025, roč. 9, e64723.

20. MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*. 2014, roč. 5, č. 4, s. 1093–1113.

21. KOCHUIEVA, Zoia; BORYSOVA, Natalia; MELNYK, Karina; HULIIEVA, Dina. Usage of Sentiment Analysis to Tracking Public Opinion. In: *COLINS*. 2021, s. 272–285.

22. MORENO-MARCOS, Pedro Manuel; ALARIO-HOYOS, Carlos; MUÑOZ-MERINO, Pedro J; ESTÉVEZ-AYRES, Iria; KLOOS, Carlos Delgado. Sentiment analysis in MOOCs: A case study. In: *2018 IEEE global engineering education conference (EDUCON)*. IEEE, 2018, s. 1489–1496.

23. GROSSMANN, Wilfried; RINDERLE-MA, Stefanie. *Fundamentals of business intelligence*. Springer, 2015.

24. NOVOTNÝ, Ota; POUR, Jan; SLÁNSKÝ, David. *Business Intelligence: Jak využít bohatství ve vašich datech*. Praha: Grada Publishing, 2005. ISBN 80-247-1094-3.

25. MAJID, M Ehsanul; MARINOVA, Dora; HOSSAIN, Amzad; CHOWDHURY, ME; RUMMANI, Farah. Use of conventional business intelligence (bi) systems as the future of big data analysis. *American Journal of Information Systems*. 2024, roč. 9, č. 1, s. 1–10.

26. PASSLICK, Jens; GRÜTZNER, Lukas; SCHULZ, Michael; BREITNER, Michael H. Self-service business intelligence and analytics application scenarios: A taxonomy for differentiation. *Information Systems and e-Business Management*. 2023, roč. 21, č. 1, s. 159–191.

27. NAMBIAR, Athira; MUNDRA, Divyansh. An overview of data warehouse and data lake in modern enterprise data management. *Big data and cognitive computing*. 2022, roč. 6, č. 4, s. 132.

28. ARMBRUST, Michael; GHODSI, Ali; XIN, Reynold; ZAHARIA, Matei et al. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: *Proceedings of CIDR*. sn, 2021, zv. 8, s. 28. Č. 1.

29. MATSKIN, Mihhail; TAHMASEBI, Shirin; LAYEGH, Amirhossein; PAYBERAH, Amir Hossein; THOMAS, Aleena; NIKOLOV, Nikolay; ROMAN, Dumitru. A Survey of Big Data Pipeline Orchestration Tools from the Perspective of the DataCloud Project. *DAMDID/RCDL (Supplementary Proceedings)*. 2021, roč. 3036, s. 63–78.

30. SEENIVASAN, Dhamotharan. ETL vs ELT: Choosing the right approach for your data warehouse. *International Journal for Research Trends and Innovation (IJRTI)*. 2022, roč. 7, č. 2, s. 110–122.

31. MUVVA, Sainath. Data Pipeline Orchestration and Automation: Enhancing Efficiency and Reliability in Big Data Environments. *Int. J. Core Eng. Mgmt*. 2025, roč. 6, č. 11.

32. FOIDL, Harald; GOLENDUKHINA, Valentina; RAMLER, Rudolf; FELDERER, Michael. Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. *Journal of Systems and Software*. 2024, roč. 207, s. 111855.

33. JAMSA, Kris. *Cloud computing*. Jones & Bartlett Learning, 2022.

34. RUPARELIA, Nayan B. *Cloud computing*. Mit Press, 2016.

35. SATYANARAYANA, S. Cloud computing: SAAS. *Computer Sciences and Telecommunications*. 2012, č. 4, s. 76–79.

36. INSEE. *Convergence criteria (Maastricht Treaty)*. 2021. Dostupné tiež z: https://www.insee.fr/en/metadonnees/definition/c1348. Accessed: 2026-01-11.

37. POLASEK Wolfgang, Amplatz Christian. The Maastricht Criteria and the Euro: Has the Convergence Continued? *Journal of Economic Integration*. 2003, roč. 18, č. 4, s. 661–688. Dostupné tiež z: https://www.e-jei.org/upload/YQU9GXKL59VJ07KV.pdf.

Prílohy

1. Detailná špecifikácia dátových súborov z Eurostatu a konkrétny výber filtrov

Tabuľka 4: Detailná špecifikácia dátových súborov z Eurostatu a konkrétny výber filtrov

Dataset	Freq.	Period	Sector	NA Item / Indicator	Unit	Seasonal Adj.	Classification / Demography	Geography	Flags	Notes
gov_10q_ggnfa	Q	2000Q1–2025Q2	S13	B9, TR, TE	EUR; % GDP	NSA	–	All EU	P	Non-financial accounts
gov_10dd_edpt1	Q	2000Q1–2025Q2	S13	GD, B9	EUR; % GDP	NSA	–	All EU	–	EDP reporting
gov_10q_ggdebt	Q	2000–2024	S13	F2–F4	% GDP	NSA	–	All EU	P, M	Debt instruments
gov_10a_exp	A	2000–2024	S13	TE	% of total	–	COFOG GF01–GF10	All EU	–	Expenditure by function
gov_10a_main	A	2015–2024	S13	B9, TE, TR	% GDP	–	–	All EU	–	Main aggregates
gov_10a_taxag	A	2015–2024	S13	D2–D5, D61	% GDP	–	–	All EU	–	Tax revenues
nama_10_gdp	A	2000–2024	–	B1GQ, P6, P7	% GDP	–	–	All EU	–	National accounts
prc_hicp_manr	M	2000–2025	–	CP00	RCH_A	–	–	All EU	–	Inflation (HICP)
une_rt_q	Q	2003Q1–2025Q2	–	–	PC_ACT	SA	Y15–74; Total	All EU	–	Unemployment rate
ds-045409	A	2015–2024	–	Value in EUR	EUR	–	HS2/HS4	AT, CZ, HU, SK, PL, EU27	–	Trade data (API)