PROFINIT
NÁSKOK DÍKY ZNALOSTEM

Azure cvičení

Pipeline pro stažení a práci s veřejnými datasety Covid-19

Obsah

1 Popis projektu

1.1 Cíl projektu

Cílem projektu je za pomocí Azure komponent vytvořit v Azure Data Factory pipeline, která po spuštění stáhne aktuální data o testování na Covid-19 z portálu ministerstva zdravotnictví, následně provede transformace datasetu (přidání názvů krajů + okresů, odstranění sloupců...) a výsledný dataset předá Azure Databricks notebooku, kde se bude s datasetem dále pracovat. Výsledný notebook bude obsahovat jednoduchou analýzu dat a vizualizace.

1.2 Použité technologie

V projektu se používá zejména **Azure Data Factory**, **Azure Databricks** a **Azure Storage account**. Cílem je osahat si základy těchto technologií a vyzkoušet si je propojit prostřednictvím pipeline v Azure Data Factory.

2 Založení účtu

Před začátkem: doporučuji celý projekt dělat v AJ, tzn. přepnout si výchozí jazyk v nastavení.

1. Vytvořte si účet přes tento odkaz: https://azure.microsoft.com/en-us/free/students/. **Použijte svůj školní mail!**
2. Po vytvoření účtu se vám otevře Education Hub.
3. Zvolte možnost, že chcete získat přístup ke studenstkým výhodám a dokončete registraci.
4. Na stránce Subskripcí si ověřte, že se vám vytvořila *Azure for Students* subskripce.

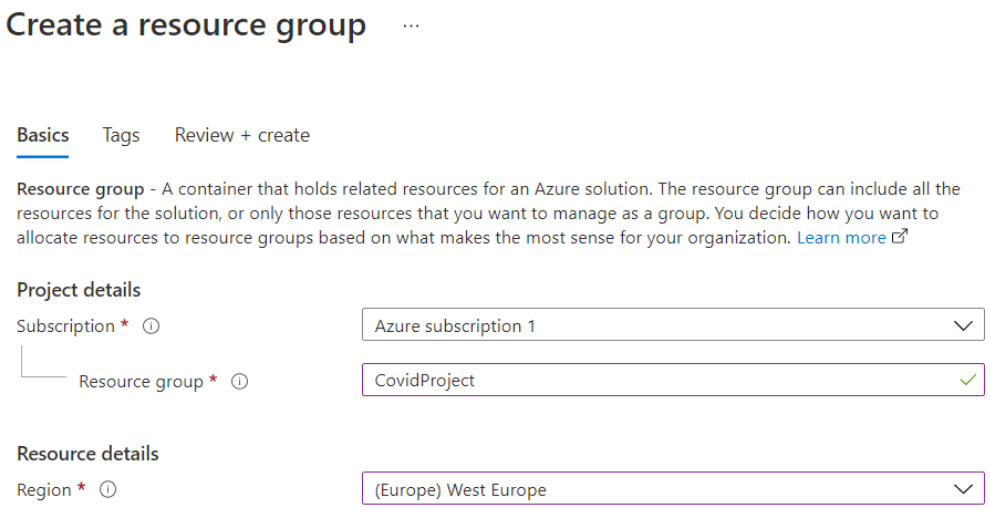
3 Příprava prostředků

`Resource` je označení, které se používá pro instance jednotlivých service, komponent, které Azure nabízí, jako např. VMs, Webová aplikace, Databricks, Storage account, Azure Data Factory, Azure SQL Databáze ...

3.1 Vytvoření Resource group

`Resource groups` se používají pro logické seskupování `resource`. Dobrá organizace `resource` následně usnadňuje práci; např. při nastavování přístupu a různých vlastností společných pro všechny `resource` v rámci jedné `resource group`. Pro tento projekt vytvoříme jednu `resource group`, do které umístíme všechny používané `resource`.

1. Na Azure portálu rozklikněte boční menu a vyberte záložku *resource groups*.
2. Vytvořte novou `resource group`.
3. Zvolte vhodné jméno a blízký region, potvrďte založení.

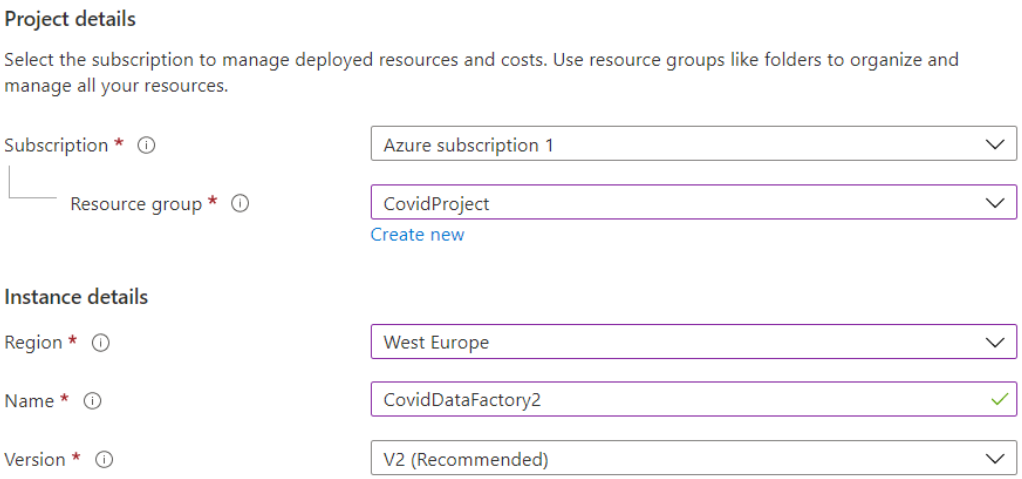


Obrázek 1: Zakládání resource group

3.2 Vytvoření Data Factory

1. V menu vyberte možnost *Create resource* (první možnost).
2. Vyhledejte Data Factory `resource` a klikněte na *Create*.
3. Vyplňte základní informace, jako verzi zvolte V2. Jako `resource group` zvolte tu vytvořenou v předchozím kroku.

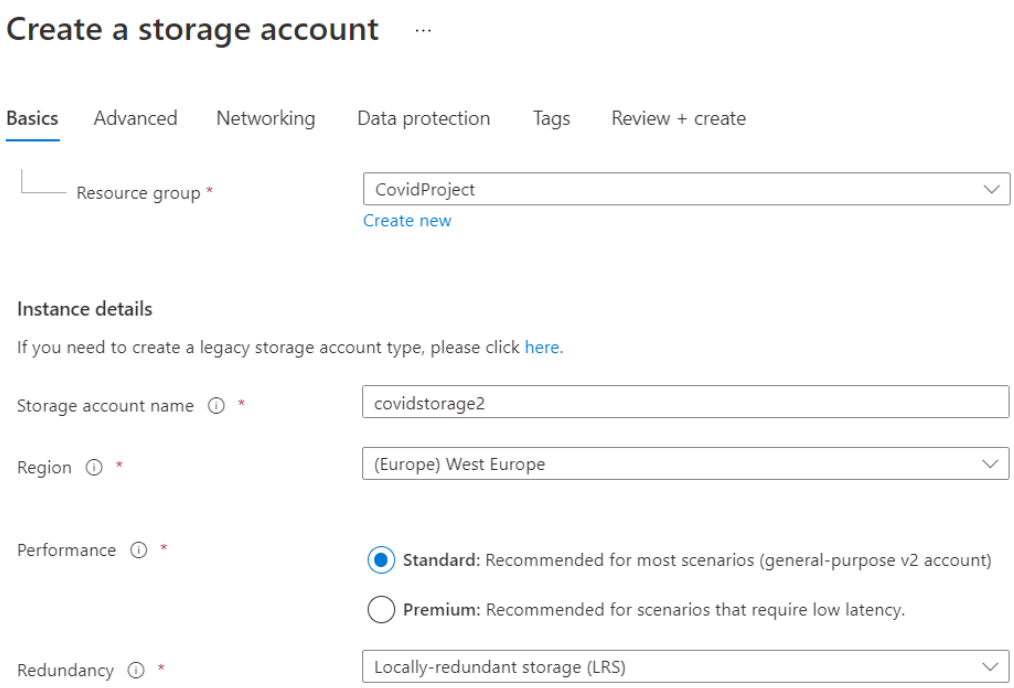
4. V záložce Git configuration zaškrtněte možnost *Configure Git later*.
5. **Review + Create**.



Obrázek 2: Založení Data Factory

3.3 Vytvoření Storage Account

1. Obdobným způsobem jako v předchozím kroku vytvořte novou `resource` - Storage Account.
2. Zvolte *Standard* performance a *LRS* redundancy. (Pro naše potřeby bohatě stačí.)
3. **Review + Create**.



Obrázek 3: Zakládání Storage Account

3.4 Vytvoření Azure Databricks

1. Stejným způsobem jako v předchozích krocích založte Azure Databricks `resource`. Opět není potřeba měnit **skoro** žádná nastavení - **Pricing tier nastavte jako Trial**.

2. Ujistěte se, že jste zvolili správnou `resource group`.

3.5 Otevření Azure Data Factory Studia

V ADF Studiu budeme tvořit většinu logiky projektu.

Proces otevření ADF Studia:

1. Na Azure Portalu si najděte vámi vytvořený Data Factory `resource`.

2. Klikněte na *Open Azure Data Factory Studio*. (Obrázek 4)

Obrázek 4: Otevření ADF Studia

4 Pipeline

Pipeline práce s daty zahrnuje práci s Azure Data Factory Pipeline logické seskupení aktivit, které dohromady pracují na nějakém úkolu. V tomto úkolu si zkusíme velmi zjednodušenou pipeline. **V našem případě to budou následující aktivity:**

1. **Download data**
   - Vytvoření Containeru uvnitř Account Storage storage.
   - Stažení covid dat.
   - Nahrání do Containeru.
2. **Data Flow**
   - Slouží k transformaci dat pomocí jednoduchého GUI.
   - Všechny optimalizace se dějou automaticky na pozadí.
   - Podporuje široké množství operací (např. join, sort, filter, select...).
3. **Databricks Notebook**
   - Aktivita sloužící ke spuštění konkrétního Databricks notebooku.

4.1 Vytvoření datasetů v Storage

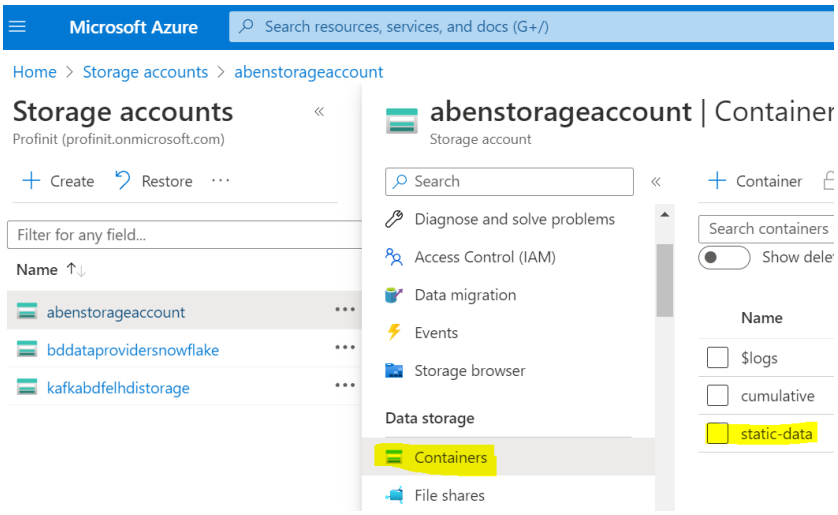
Dataset na covid portálu neobsahuje názvy krajů a okresů, pouze jejich identifikátory. Účelem našeho data flow bude spojení datasetů podle identifikátorů tak, abysme pro každý záznam měli i název kraje a okresu. Následně odstraníme zbytečné sloupce.

Stažení souborů s údaji:

- Stáhněte si připravená data o populaci (czech_population.csv) v krajích ČR a jejich NUTS/LAU kódech (CZ AREA CODES.xlsx).

Nahrání souborů do storage:

1. Na Azure Portálu přejděte na svůj Storage Account a přes možnost **Data storage** → **Containers** vytvořte nový container *static-data*.
2. Access level ponechte jako *private*.



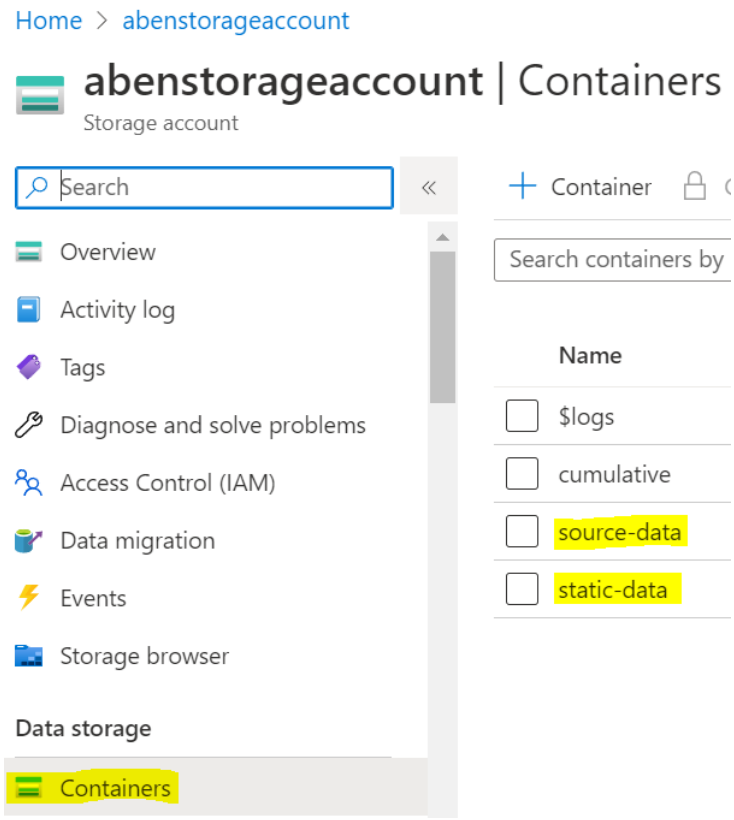
Obrázek 5: Static Container

3. Nahrajte do containeru stáhlé soubory (csv populace + excel soubor s kraji).

Na Covid portálu se nachází mnoho datasetů. V tomto cvičení budeme pracovat s datasetem **COVID-19: Celkový (kumulativní) počet provedených testů podle krajů a okresů ČR**.

1. Stáhněte si k sobě data z **https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/kraj-okres-testy.csv**.

2. Vytvořte v vašem Blob Container, například **source-data**, 6.
3. Pomocí Upload nahrajte stažené CSV.



Obrázek 6: Otevření ADF Studia

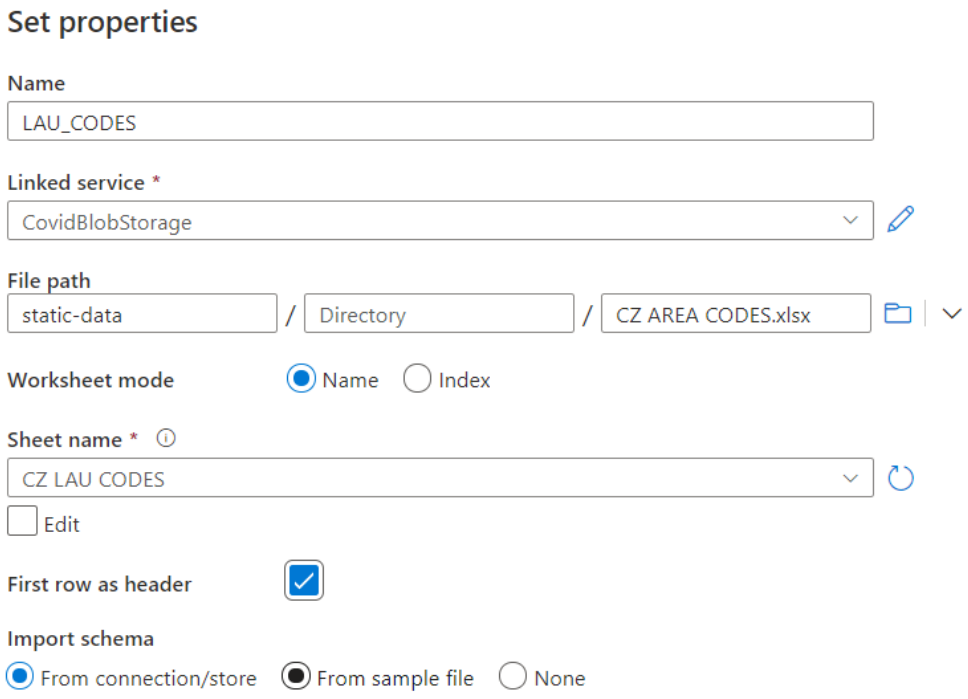
Excel soubor s názvy krajů/okresů budeme používat v Data Flow. Je tedy potřeba vytvořit v ADF Studiu nové Datasety, viz 4.2.

4.2 Data Flow

Jak již bylo řečeno, data flows slouží k transformaci a práci s daty.

1. Přes **Author** → **Datasets** → **New dataset** založte nový dataset.
2. Zvolte Blob Storage, typ datasetu Excel a specifikujte cestu k souboru.
3. Zbylé nastavení lze vidět na Obrázku 7. Všimněte si položky **Sheet name**, kterou specifikujete, jaký Sheet použít pro tento dataset - jeden dataset může být max. 1 excel stránka.

4. Založte **dva datasety** - jeden pro okresy (zkratka LAU), druhý pro kraje (NUTS). Dejte si pozor, ať vyberete správný sheet.



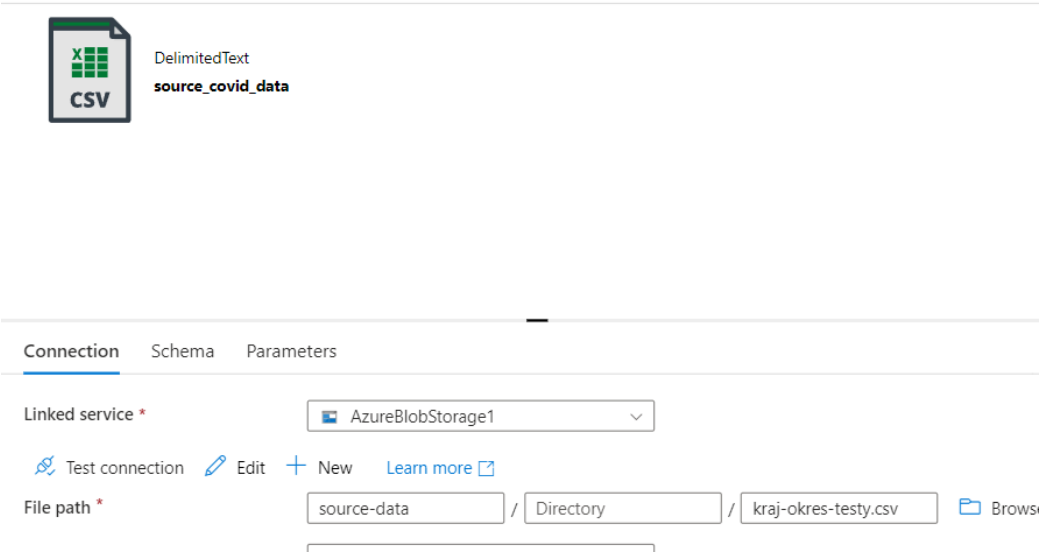
Obrázek 7: Vytvoření datasetu z Excel sheetu

5. Schema importujte **From connection**.

6. Než budete pokračovat, zkontrolujte, že máte dva datasety - jeden pro každý sheet.

7. Založte další **Dataset** jako Azure Blob Storage typu CSV.

8. Cestu zvolte k **source-dataset** Containeru vytvořeném dříve s kumulativními daty, viz 8.



Obrázek 8: Vytvoření datasetu z csv

4.2.1 Nahrání schematu

Ještě než budeme používat dataset (o testech na Covid-19) v Data Flow, musíme se ujistit, že ADF zná jeho schema.

1. Přejděte na Dataset (v ADF Studiu), který představuje soubor v Blob Storage.
2. V záložce schema vyberte možnost **Import schema**.
3. Pokud zvolíte **From connection/store**, ADF se pokusí naimportovat schema z cesty, která je specifikovaná v datasetu. Náš dataset má parametrizované jméno souboru a ADF se pokusí najít soubor s názvem, který jste specifikovali v default value. Pokud jste takový soubor manuálně do své storage nenahráli, tak se to nepovede. Mělo by ale vše proběhnout automaticky. Je více možností, pokud nastane nějaká chyba:
   - nastavit výchozí hodnotu na jméno souboru, které už ve storage máte - např. soubor který jste stáhli v kroku .
   - manuálně nahrát soubor do storage s názvem, který jste nastavili jako default value

- nenastavit default value a podobně jako u první možnosti zadat jméno souboru, který už máte stáhlý, ale ne jako default value, ale až budete vyzvání po kliknutí na import

4. Zkontrolujte, že se vám podařilo nahrát schema. Na typy sloupců teď koukat nemusíte - důležité je, aby bylo vidět jaké a kolik jich je.

4.2.2 Vytvoření Data Flow a přidání zdrojů

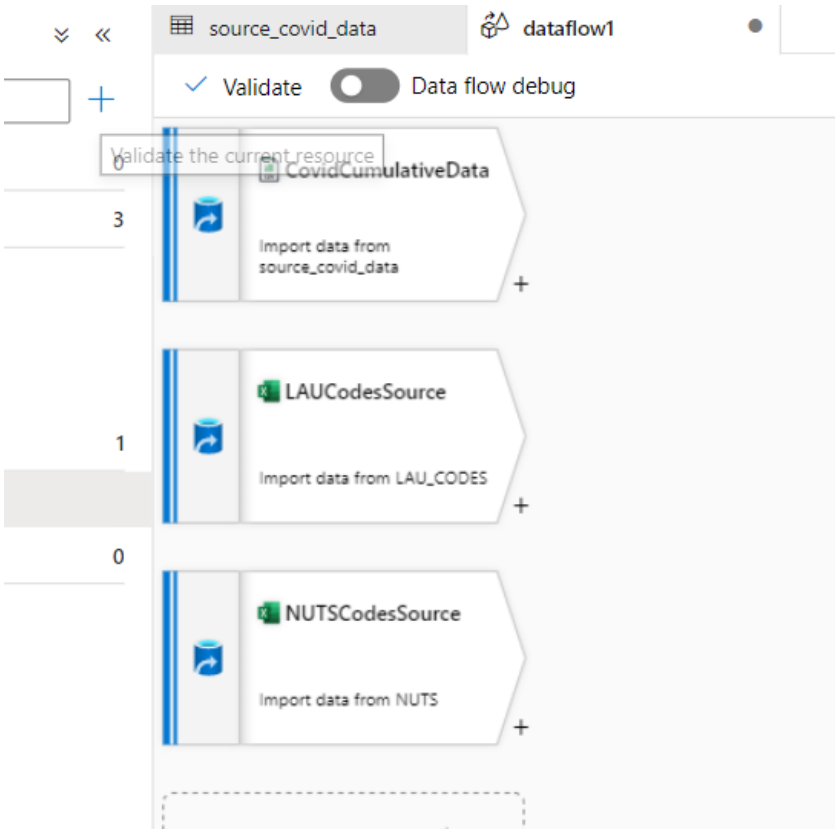
Ve chvíli kdy máme připravené zdrojové datasety, můžeme přejít k vytvoření Data Flow.

1. Možností **Author** → **Data flows** → **New data flow** vytvořte novou Data flow.

2. Vytvořte tři zdrojové datasety - LAU, NUTS a stažená data o testech.

3. Stačí kliknout na **Add source** a vybrat dataset. Výsledek by měl vypadat podobně jako na obrázku 9.

4. Ujistěte se, že u každého zdroje vidíte počet sloupců. Pokud vám to ukazuje 0 sloupců, znamená to, že dataset nemá definované schéma. V takovém případě schema nahrajte (viz 4.2.1).



Obrázek 9: Data Flow zdrojové datasety

4.2.3 Přidání transformací

Kliknutím na + vedle zdrojového datasetu lze přidat novou transformaci a ty lze tímto způsobem dále řetězit. Klikněte na + a prohlédněte si, jaké operace lze s daty provádět.

V našem Data flow budeme potřebovat hlavně **join** a **select**. **Select** umožňuje sloupce přejmenovat a smazat. Pro smazání sloupce zaškrtněte políčko vedle názvu a klikněte na ikonu koše. Pro obnovení nastavení zmáčkněte **Reset**.

1. Pomocí operace `join` spojte dataset s Covid portálu s Excel datasety tak, aby u každého záznamu bylo jméno kraje a okresu. `Join` bude potřeba použít dvakrát. (Pro připomenutí: LAU odpovídá NUTS4, kód kraje je NUTS3, občas zkrácený jen na NUTS).

2. Následné operace nejsou nutné, ale urychlí proces JOINování. Pomocí operace `select` smažte duplikátní (např. `kraj_nuts_kod` obsahuje stej-

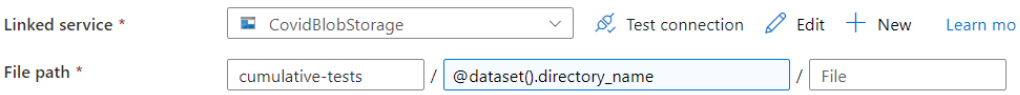
nou informaci jako sloupec `CZ-NUTS3`) a nepotřebné sloupce. Z datasetu `LAUCODES` (NUTS4) nepotřebujeme `kod_okresu` ani `nazev_kraje`, sloupce můžete smazat ještě před joinem. Sloupce `CZ-NUTS3` a `CZ-NUTS4` ponechejte.

3. Ukončující položkou Data flow je vždy operace `Sink` dataset - destinace, kam se mají transformovaná data uložit, na Obrázku 11 pojmenováno jako `SaveJoinedAndCleaned`.

   (a) Založte nový dataset, obdobným způsobem jako u datasetu, kam ukládáme stažená data z Covid portálu. Opět využijeme Blob Storage a dataset budeme ukládat jako CSV

   (b) **Pozor**, dataset, který používáme jako Sink nemůže mít specifikované jméno souboru. Directory nastavte jako dynamic value `@dataset().directory_name`. Data mohou být rozdělena do několik partition, pokud bysme vyžadovali, aby výsledek Sink operace byl v jednom souboru, museli bychom to explicitně nastavit v Data flow.

   (c) To ale není třeba, data budeme třídit do složek, které v názvu budou mít čas spuštění pipeliny. **Vytvořte parametr** *directory_name* a ten pak použijte v poli *directory* (na defaultní hodnotě nezáleží - dataset vždy budeme používat s konkrétní hodnotou, kterou mu předáme z pipeliny). Pole *filename* ponechte prázdné. Viz Obrázek 10.

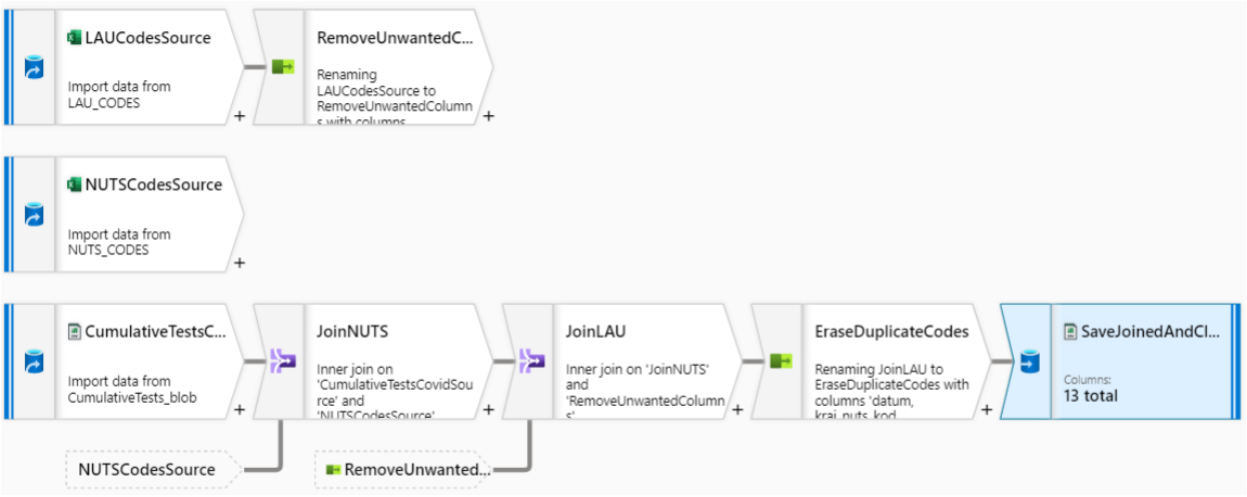


Obrázek 10: Nastavení cesty u Sink datasetu použitého v Data flow.

   (d) Kontejner buď můžete založit nový, nebo použít stejný kam ukládáte stažené datasety z Covid portálu.

4. Nově vytvořený dataset použijte jako Sink dataset ve vaší data flow.

5. Výsledná data flow může vypadat např. jako Obrázek 11.



Obrázek 11: Takto může vypadat výsledné Data flow.

6. Změny uložte kliknutím na **Publish all** → **Publish**.

4.3 Přidání Data flow aktivity do pipeline

1. Vytvořte vaší pipeline a vložte do ní Data flow aktivitu (v kategorii Move & Transform) a vhodně ji pojmenujte.

2. V záložce **Settings** u možnosti **Data flow** zvolte vámi vytvořenou Data flow v předchozím kroku.

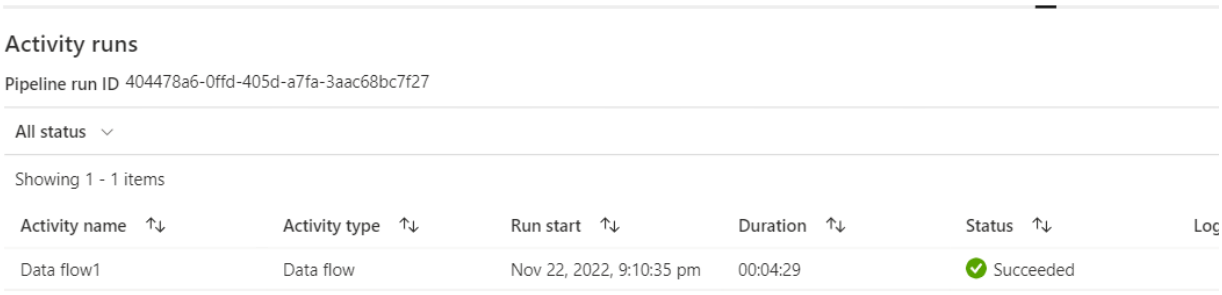
3. Po zvolení konkrétní data flow v **Settings** přibydou pole pro parametry jejích datasetů. V našem případě jde o: *directory_name* (jméno adresáře kam uložit výsledek).

   (a) **directory_name**: Sestavte název z vhodného jména a času spuštění pipeliny, např.

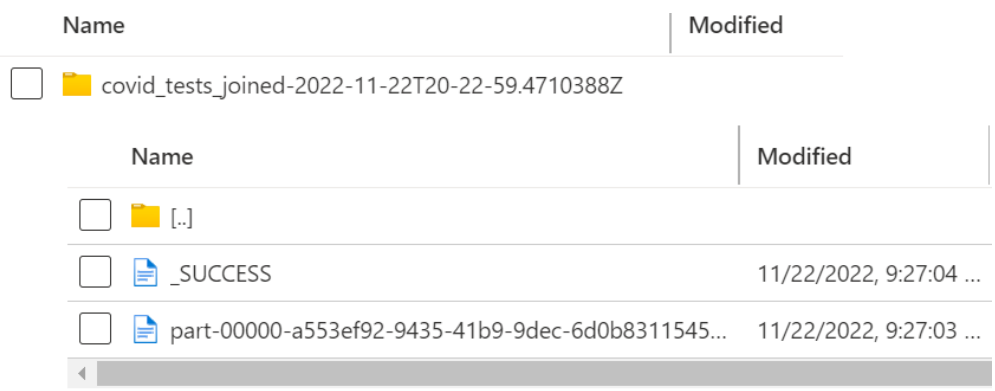
   ```
   {@concat('covid_tests_joined',
           '-',replace(pipeline().TriggerTime,':','-'))}.
   ```

   - **pokud kopírujete kód přímo z dokumentu, dejte si pozor jestli se zkopíroval správně (tj. uvozovky, podtržítka, bez newline, bez mezer navíc apod.)**

4. Změny uložte kliknutím na **Publish all**. Pro otestování zkuste pipelinu spustit pomocí **Trigger now**. Až pipeline doběhne, viz 12, podívejte se do vaší storage, jestli obsahuje adresář s najoinovanými daty, viz 13.



Obrázek 12: Takto může vypadat úspěšný běh dataflow.



Obrázek 13: Najoinovaná data v cumulative containeru.

5 Databricks

V poslední části projektu využijeme Databrick notebook - jednoduché webové rozhraní dokumentu, který obsahuje spustitelný kód, případně vizualizace a doplňující komentáře, více zde.

Nejdříve vytvoříme nový notebook a cluster, poté propojíme naše Databricks a Data Factory `resource` a následně přidáme parametrizovanou Databricks notebook aktivitu do pipeline. Pak už bude většina práce přímo v notebooku.

5.1 Otevření pracovního prostředí

- Najděte si na Azure Portálu vaší Databricks `resource` a následně otevřte pracovní prostředí kliknutím na **Launch Workspace**.

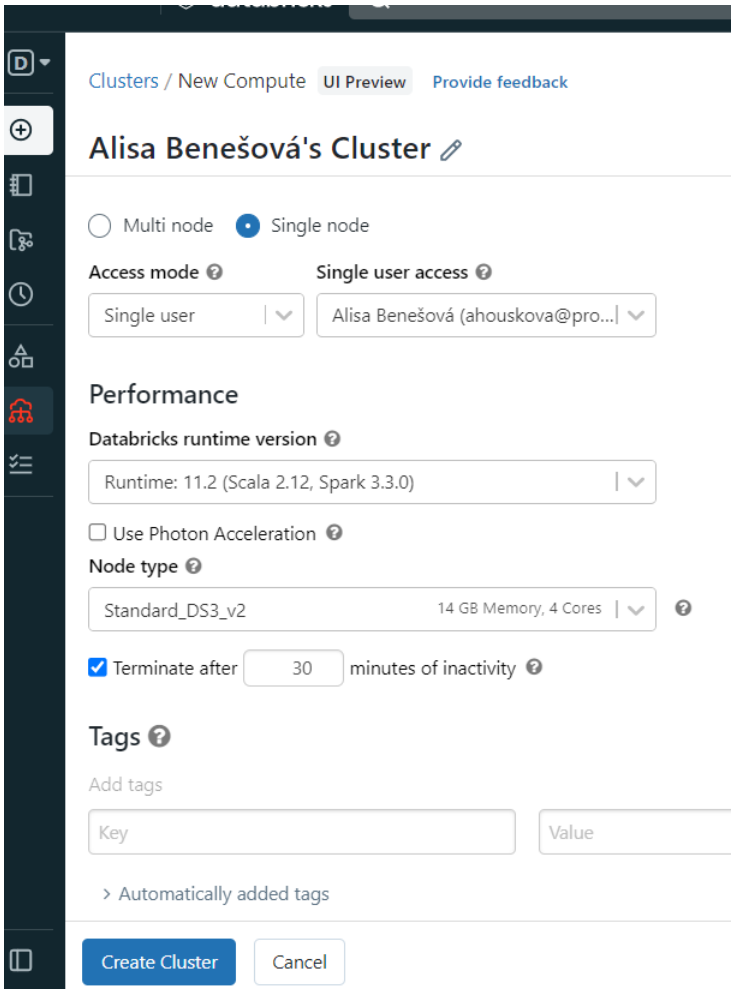
5.2 Vytvoření clusteru

Aby bylo možné provádět práci, výpočty (např. příkazy z Databricks notebooku), je potřeba zajistit výpočetní prostředky - **Cluster**. Azure Databricks rozlišuje 2 typy clusterů:

- **all-purpose cluster**: vytvoří se jednorázově, vypíná a zapíná se na pokyn, nebo když uplyne nastavená doba
- **job cluster**: vytvoří se automaticky dle specifikovaných parametrů až když je potřeba, např. když přijde trigger

Pro naše účely, jelikož budeme chtít notebook testovat během vytváření, použijeme **all-purpose cluster**. Více o clusterech.

1. Navigujte do **Compute** → + **Create cluster**.
2. Vytvořte cluster s nastavením na Obrázku 14



Obrázek 14: Nastavení clusteru

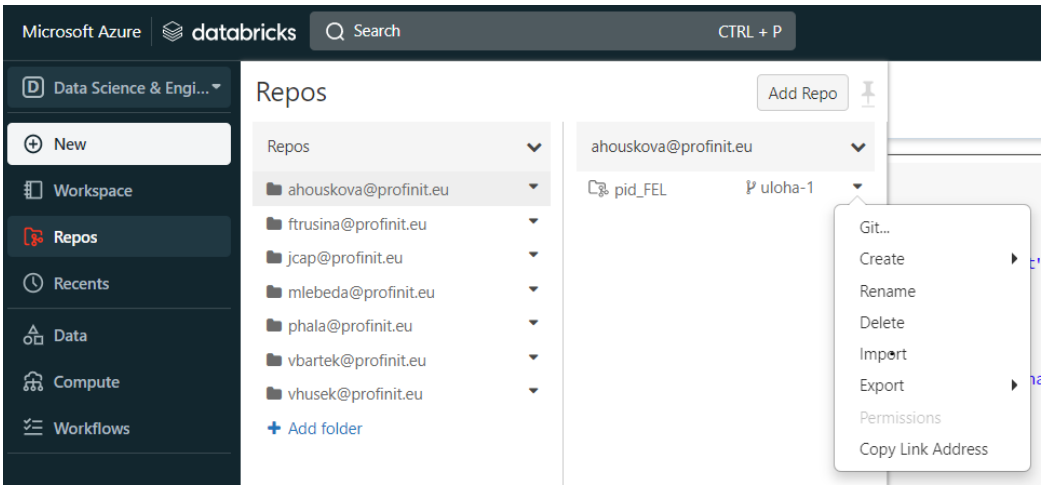
5.3 Import notebooku

1. Z menu vyberte **Workspace** → **Users** → **Váš účet** → **Import** → **URL** a naimportujte si notebook z URL: https://github.com/mjanec/azure-practicetask-covid/blob/master/Covid%20Practice%20t

2. Vlevo nahoře v dropdown menu můžete zvolit, jaký cluster bude vykonávat příkazy z notebooku. Vyberte váš cluster - měl by běžet. Pokud neběží, spusťte ho.

3. Not Mandatory step. You can clone the repository to your git provider and via personal access token connect to Databricks. You must use

menu Repos. How to generate tokens, see github, gitlab. Choose **Add Repos** and provide your URL, pull.



Obrázek 15: Repos download

5.4 Přidání Azure Databricks do linked services

Aby bylo možné spouštět notebooky v konkrétním ADB `resource` prostřednictvím Azure Data Factory, je nejdříve potřeba tyto komponenty propojit. V ADB vytvoříme *access token*, který pak využijeme při vytváření nové linked service v ADF.

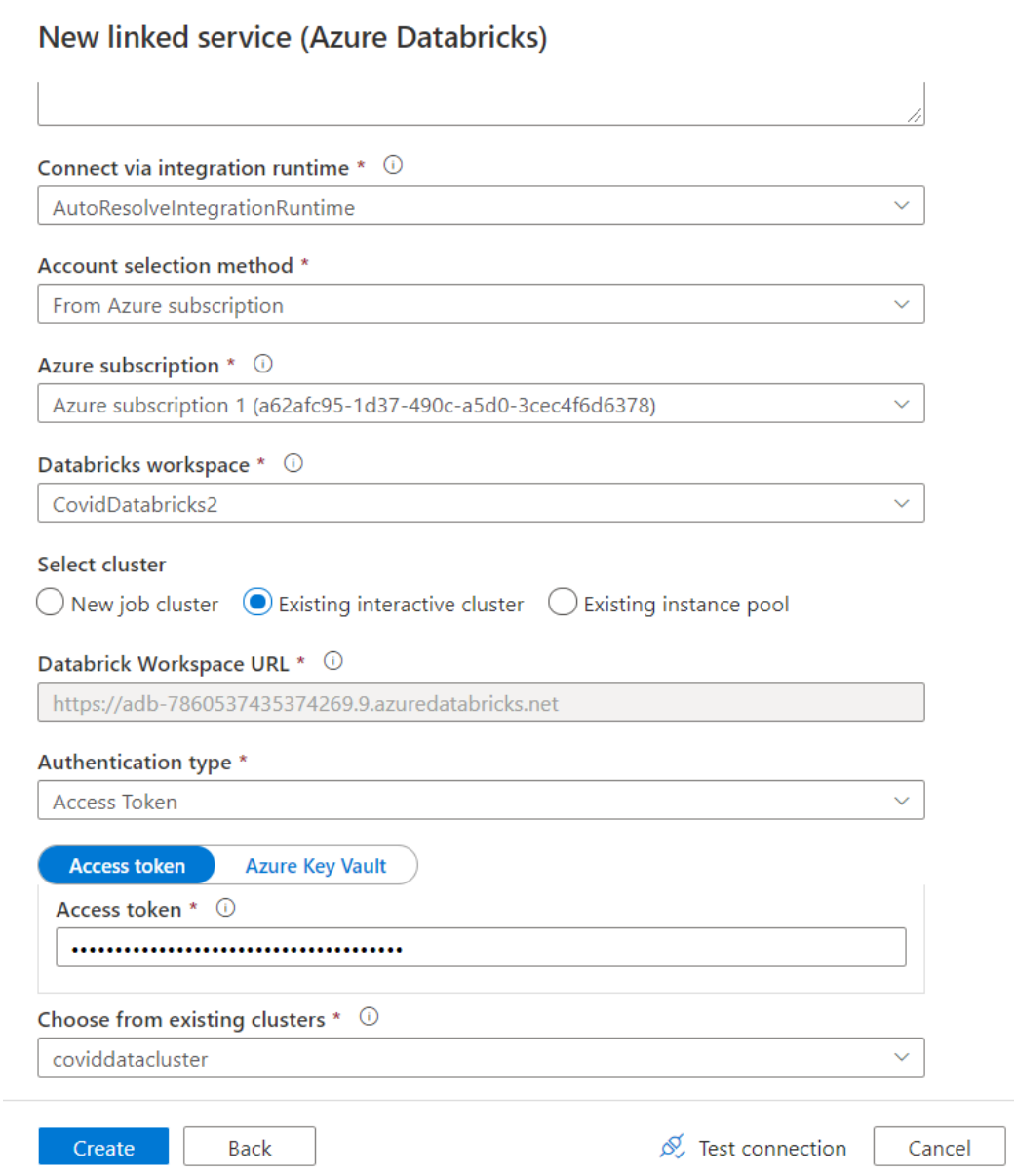
Vygenerování access tokenu:

1. V Databricks postraním menu navigujte do **Settings** → **User settings** → **Generate New Token**.
2. Vyplňte účel tokenu a jeho platnost - můžete ponechat defaultní hodnotu.
3. Token si zkopírujte do schránky nebo uložte do souboru. Nepůjde ho znovu zobrazit.

S tokenem ve schránce se vraťte do **Azure Data Factory Studia** a vytvořte novou linked service:

1. **Manage** → **Linked service** → **New**.
2. V záložce **Compute** vyberte Azure Databricks.

3. Vyplňte název a základní informace, přes možnost **From Azure subscription** vyberte váš Databricks workspace.
4. Typ autentizace zvolte token a vložte váš token ze schránky.
5. Vyberte možnost **Existing interactive cluster** a vyberte váš cluster.
6. Obrázek 16.



Obrázek 16: Příklad nastavení ADB linked service

5.5 Přidání Azure Databricks notebook aktivity

Po vytvoření notebooku jej můžeme přidat jako aktivitu do Data Factory pipeline.

1. Přidejte do pipeline novou **Databricks** → **Notebook** aktivitu.

2. Zajistěte (pomocí zelené šipky), aby se aktivita spustila až po úspěšném doběhnutí předchozí aktivity (data flow).

3. V nastavení aktivity vyberte vaší Databricks *linked service*.

4. V záložce **Settings** vaší aktivity vyberte cestu k notebooku, který má aktivita spustit.

5. V záložce **Settings** → **Base parameters** nastavte parametry, které budou notebooku předány při spuštění:

   - *storage_account_name*: Jméno vašeho Storage account.
   - *container*: Název kontejneru, kam ukládáte zprocesovaná data (pomocí data flow).
   - *directory*: Název adresáře, kde se nachází data, se kterými budete dále pracovat - hodnota tohoto parametru musí být stejná, jako hodnota Sink parametru předchozí aktivity (předáváme adresu, kam se uložila data po skončení data flow), viz 3a.
   - *storage_account_access_key*: klíč, který najdete na Azure portálu v nastavení vašeho Storage accountu (**Security + networking** → **Access keys**) - použijte **key1**
     - Šlo by použít i **key2**, proč jsou klíče dva si můžete přečíst zde.

5.6 Zobrazení detailů běhu notebooku

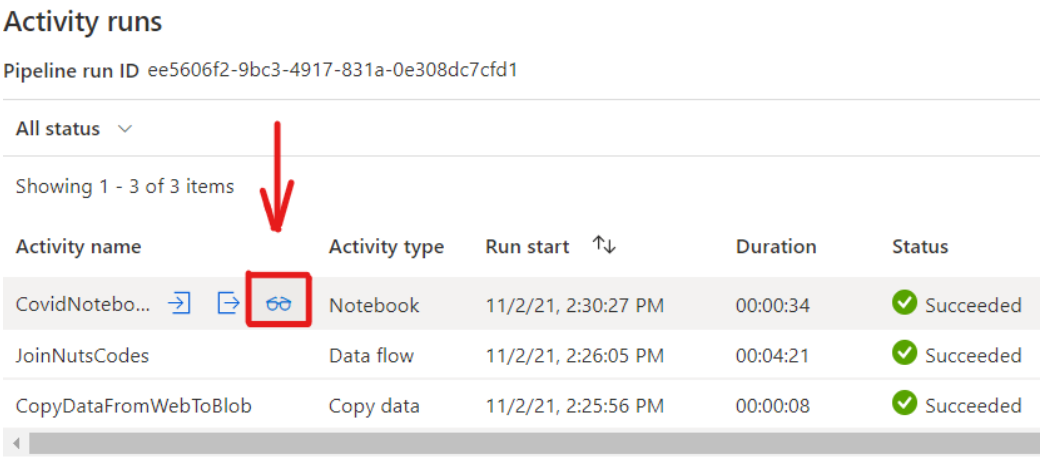
Po tom co přidáte notebook aktivitu do vaší pipeline, zkuste ji spustit (přes **Trigger now**).

1. Po spuštění běžte do **Monitor** → **Pipeline runs** a vyberte aktuální běh.

2. Kliknutím na tlačítko **Details** (viz Obrázek 17) se vám zobrazí **Run page url**.

3. Přejděte na **run page** - stránka obsahující konkrétní běh notebooku. Můžete se podívat na výstupy buněk či zkontrolovat (nahoře na stránce) jestli byly notebooku předány správné parametry.

4. Neděste se toho, že běh notebook aktivity nebude úspěšný - notebook čeká na vypracování.



Obrázek 17: Zobrazení běhu notebooku

Pokud se vám notebook úspěšně spouští se správnými parametry, přejděte do vašeho Azure Databricks workspace a dále se řiďte pokyny v notebooku. Hodně štěstí.