

# Azure cvičení

Pipeline pro stažení a práci s veřejnými datasety  
Covid-19



## Obsah

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Popis projektu</b>                                 | <b>1</b>  |
| 1.1      | Cíl projektu . . . . .                                | 1         |
| 1.2      | Použité technologie . . . . .                         | 1         |
| <b>2</b> | <b>Založení účtu</b>                                  | <b>1</b>  |
| <b>3</b> | <b>Příprava prostředků</b>                            | <b>1</b>  |
| 3.1      | Vytvoření Resource group . . . . .                    | 2         |
| 3.2      | Vytvoření Data Factory . . . . .                      | 2         |
| 3.3      | Vytvoření Storage Account . . . . .                   | 3         |
| 3.4      | Vytvoření Azure Databricks . . . . .                  | 4         |
| <b>4</b> | <b>Stažení datasetů</b>                               | <b>4</b>  |
| 4.1      | Otevření Azure Data Factory Studia . . . . .          | 4         |
| 4.2      | Založení HTTP linked service . . . . .                | 5         |
| 4.3      | Vytvoření datasetu . . . . .                          | 6         |
| <b>5</b> | <b>Pipeline</b>                                       | <b>7</b>  |
| 5.1      | Vytvoření pipeline . . . . .                          | 8         |
| 5.2      | Přidání Copy data aktivity . . . . .                  | 8         |
| 5.3      | Spuštění pipeline . . . . .                           | 10        |
| 5.4      | Data Flow . . . . .                                   | 11        |
| 5.4.1    | Nahrání schématu . . . . .                            | 12        |
| 5.4.2    | Vytvoření Data Flow a přidání zdrojů . . . . .        | 13        |
| 5.4.3    | Přidání transformací . . . . .                        | 14        |
| 5.5      | Přidání Data flow aktivity do pipeline . . . . .      | 16        |
| <b>6</b> | <b>Databricks</b>                                     | <b>17</b> |
| 6.1      | Otevření pracovního prostředí . . . . .               | 17        |
| 6.2      | Vytvoření clusteru . . . . .                          | 17        |
| 6.3      | Import notebooku . . . . .                            | 18        |
| 6.4      | Přidání Azure Databricks do linked services . . . . . | 19        |
| 6.5      | Přidání Azure Databricks notebook aktivity . . . . .  | 20        |
| 6.6      | Zobrazení detailů běhu notebooku . . . . .            | 21        |

# 1 Popis projektu

## 1.1 Cíl projektu

Cílem projektu je za pomoci 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 studentským výhodám a dokončete registraci.
4. **Na stránce Subskripce** 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

### 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.

#### Create a resource group ...

**Basics** Tags Review + create

**Resource group** - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization. [Learn more](#)

**Project details**

Subscription \* ⓘ Azure subscription 1 ▼

Resource group \* ⓘ CovidProject ✓

**Resource details**

Region \* ⓘ (Europe) West Europe ▼

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.

### 3.3 Vytvoření Storage Account

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

#### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

|                    |   |
|--------------------|---|
| Subscription * ⓘ   | <div>Azure subscription 1</div> ▼                       |
| Resource group * ⓘ | <div>CovidProject</div> ▼<br><a href="#">Create new</a> |

#### Instance details

|             |                                |
|-------------|--------------------------------|
| Region * ⓘ  | <div>West Europe</div> ▼       |
| Name * ⓘ    | <div>CovidDataFactory2</div> ✓ |
| Version * ⓘ | <div>V2 (Recommended)</div> ▼  |

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.**

### 3.4 Vytvoření Azure Databricks

#### Create a storage account ...

**Basics**   Advanced   Networking   Data protection   Tags   Review + create

Resource group \*  [Create new](#)

**Instance details**  
If you need to create a legacy storage account type, please click [here](#).

Storage account name ⓘ \*

Region ⓘ \*

Performance ⓘ \*  
☒ **Standard:** Recommended for most scenarios (general-purpose v2 account)  
☐ **Premium:** Recommended for scenarios that require low latency.

Redundancy ⓘ \*

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**.

## 4 Stažení datasetů

Prvním krokem naší pipeline bude stažení **aktuálního** datasetu z **portálu ministerstva zdravotnictví**.

### 4.1 Otevření Azure Data Factory Studio

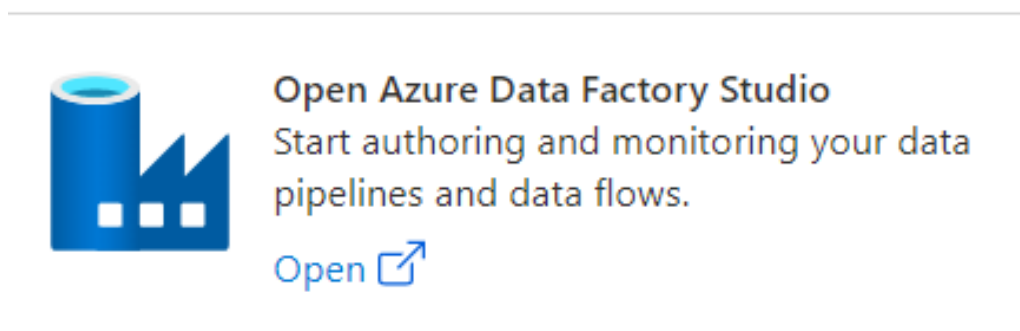
V ADF Studiu budeme tvořit většinu logiky projektu.

**Proces otevření ADF Studio:**

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

## 4.2 Založení HTTP linked service

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



Obrázek 4: Otevření ADF Studia

## 4.2 Založení HTTP linked service

1. Otevřete si ADF Studio.
2. Z postraního levého menu vyberte **Manage** → **Linked Services** → **New**.
3. Vyberte typ úložiště, ve kterém se nachází datasety, se kterými budeme pracovat. V našem případě se jedná o **HTTP**.
4. Jako *Base URL* nastavte adresu Covid portálu, *Authentication Type* nastavte na Anonymous.
5. Potvrďte vytvoření a v ADF Studiu zveřejněte a validujte změny kliknutím na **Publish all** → **Publish**.

## 4.3 Vytvoření datasetu

### New linked service (HTTP)

**Name \***

CovidPortal

**Description**

Portál o Covid-19 MZ ČR

**Connect via integration runtime \* ⓘ**

AutoResolveIntegrationRuntime

**Base URL \***

https://onemocneni-aktualne.mzcr.cz

**Server Certificate Validation ⓘ**

☒ Enable ☐ Disable

**Authentication type \***

Anonymous

Obrázek 5: Vytvoření linked HTTP Server service

## 4.3 Vytvoření datasetu

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**. Je potřeba tento dataset vytvořit v ADF Studiu, využijeme k tomu i v předchozím kroku vytvořenou HTTP service. Tímto způsobem zajistíme, že při každém spuštění naší *pipeline* budeme pracovat s aktuální verzí datasetu (tou která se v době spuštění nachází na Covid portálu).

1. Z postraního levého menu vyberte **Author** → **Dataset** → **New dataset**.
2. Vyberte data store, kde se dataset nachází - opět **HTTP**.
3. Zvolte typ datasetu - **csv**.
4. Jako *Linked service* zvolte HTTP Service vytvořenou v předchozím kroku.



5. Do *Relative URL* zadejte adresu datasetu na Covid portálu. Např. `/api/v2/covid-19/kraj-okres-testy.csv`.
6. Zaškrtněte *First row as a header*. Schema není potřeba importovat. (Obrázek 6).

### New linked service (HTTP)

**Name \***

CovidPortal

**Description**

Portál o Covid-19 MZ ČR

**Connect via integration runtime \* ⓘ**

AutoResolveIntegrationRuntime

**Base URL \***

https://onemocneni-aktualne.mzcr.cz

**Server Certificate Validation ⓘ**

☒ Enable ☐ Disable

**Authentication type \***

Anonymous

Obrázek 6: Vytvoření datasetu

## 5 Pipeline

Pipeline v Azure Data Factory je **logické seskupení aktivit**, které dohromady pracují na nějakém úkolu.

**V našem případě to budou 3 aktivity:**

### 1. Copy data

- Slouží k překopírování dat z jedné *data store* do jiné.

## 5.1 Vytvoření pipeline

- Zdroj i destinace jsou datasety (nedefinované v ADF Studiu). Ne všechny datasety mohou být použity jako *Sink* (destinace, kam chci data nakopírovat). [Přehled lze nalézt zde](#).

### 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.

## 5.1 Vytvoření pipeline

1. **Author** → **Pipeline** → **New pipeline**
2. Zvolte vhodné pojmenování.

Po vytvoření pipeline byste měli vidět menu aktivit. Nové aktivity lze do pipeline přidat přetáhnutím do pipeline okna. Projděte si, jaké druhy aktivit ADF nabízí.

## 5.2 Přidání Copy data aktivity

Jelikož Data Flow, se kterou budeme později pracovat, neumí jako zdrojový dataset přijímat datasety využívající *HttpServer*, bude potřeba si dataset nejdříve stáhnout a umístit ho do Azure Storage (nebo do jiného podporovaného úložiště).

1. Projděte do vaší pipeline, z nabídky aktivit vyberte **Move & Transform** → **Copy data** aktivitu a přetáhněte ji do pipeline.
2. Rozklikněte si aktivitu, jako Source dataset zvolte ten vytvořený v předchozích krocích (csv na Covid portálu).
3. Sink dataset jsme zatím nevytvořili. Dal by se použít způsob jako u předchozího datasetu, tzn. nejdříve vytvořit novou linked service a dále nový dataset za použití postraní menu. Využijeme ale možnost dataset vytvořit přímo v aktivitě.

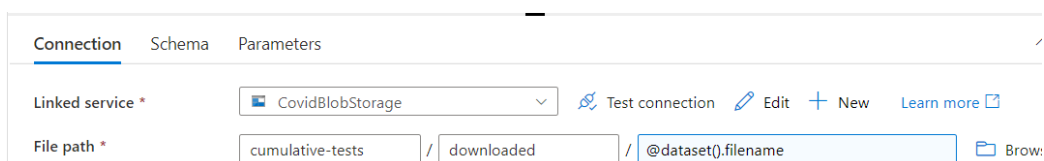
- (a) V záložce **Sink** vyberte možnost **New**.
- (b) Jako typ úložiště zvolte **Azure Blob Storage**, typ datasetu opět **csv**.
- (c) Nyní je třeba nalinkovat naší Blob Storage, která je součástí Storage Accountu, který jsme vytvořili v kroce 3.3.
  - i. V dropdown menu *Linked service* vyberte možnost **New**.
  - ii. Nastavení ponechte jak jsou. Z nabídky vyberte vaší subskripci a následně váš storage account.
  - iii. Dole klikněte na *Test Connection*, abyste ověřili, zda jste storage nalinkovali správně a následně klikněte na **Create**.
- (d) Opět zaškrtněte možnost **First row as a header**, nastavení potvrďte.
- (e) Z nabídky (**Author** → **Dataset**) vyberte nově vytvořený dataset.
  - i. Jelikož chceme mít variabilní jméno datasetu - podle času stažení - přidejte v záložce **Parameters** nový parametr a pojmenujte jej *filename*. Na defaultní hodnotě příliš nezáleží, protože ji nebudeme používat.
  - ii. Zvolte vhodnou cestu pro uložení datasetu, např. podle obrázku 7. Pod políčkem *File* (poslední pole) klikněte na *Add dynamic content* (musíte kliknout do pole) a poté zvolte vámi vytvořený parametr (Obrázek 7).
- (f) Přejděte zpět do **Copy data** aktivity do záložky **Sink**. Pokud jste správně nakonfigurovali dataset pro uložení, měli byste vidět políčko pro parametr *filename*. Pomocí *Add dynamic content* vytvořte název souboru tak, aby obsahoval čas spuštění pipeline.
  - Jelikož dvojtečky v názvech souborů dělají problémy, je potřeba je za něco nahradit, můžete využít funkci `replace`.
  - např. `@concat('covid_cumulative_tests', replace(pipeline().TriggerTime, ':', '-'))`
    - pokud kopírujete kód přímo z dokumentu, dejte si pozor jestli se zkopíroval správně (tj. bez newline, bez mezer navíc apod.)
  - Pokud byste se chtěli vyhnout hard-coded názvu souboru, můžete si udělat parametr na úrovni pipeline a použít pak ten:
 

```
@concat(pipeline().parameters.filename_cumulative_test_data, '-', replace(pipeline().TriggerTime, ':', '-'))
```

### 5.3 Spuštění pipeline

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

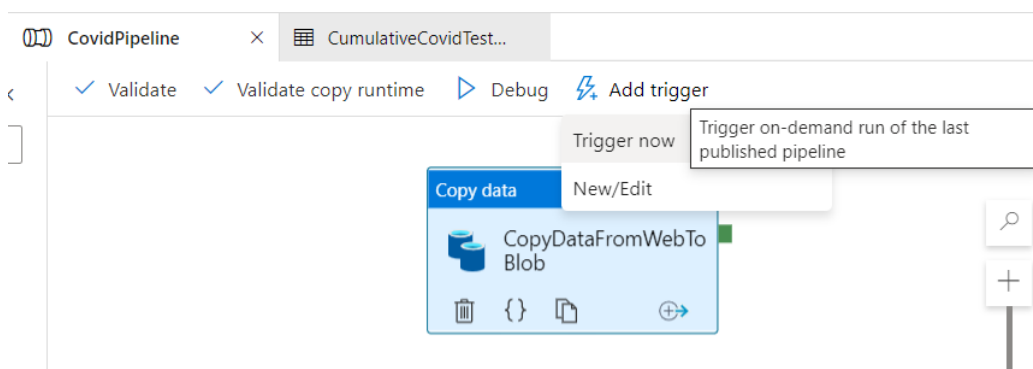
4. Změny uložte kliknutím na **Publish all** → **Publish**.



Obrázek 7: Konfigurace datasetu, do nějž ukládáme stažená data

### 5.3 Spuštění pipeline

V tuto chvíli máme již funkční pipeline o jedné aktivitě, jejíž funkce je stažení souboru z Covid portálu a uložení v Blob Storage. Jestli pipeline funguje zjistíme tak, že ji spustíme. Podle obrázku 8 manuálně triggerněte pipeline. V menu **Monitor** → **Pipeline runs** se můžete podívat na aktuální run a historii.



Obrázek 8: Triggernutí pipeline

Po doběhnutí pipeline si na Azure Portálu najděte váš Storage Account a ujistěte se, že se soubor stáhl. Měli byste vidět nově vytvořený kontejner (pokud jste jej nevytvářeli manuálně přes Azure Portal) a v něm adresář obsahující jeden soubor - stažené csv. Zkuste si pustit pipeline ještě jednou.

## 5.4 Data Flow

Jak již bylo řečeno, data flows slouží k transformaci a práci s daty. 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*.
3. Nahrajte do containeru stáhlé soubory (csv populace + excel soubor s kraji).

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

1. Opět 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 9. Všimněte si položky **Sheet name**, kterou specifikujete, jaký Sheet použít pro tento dataset - jeden dataset může být max. 1 excel stránka.
4. Založte **dva datasetsy** - jeden pro okresy (zkratka LAU nebo NUTS4), druhý pro kraje (NUTS3 nebo jen NUTS). Dejte si pozor, ať vyberete správný sheet.
5. Schema importujte **From connection**.
6. Než budete pokračovat, zkontrolujte, že máte dva datasetsy - jeden pro každý sheet.

## Set properties

**Name**  
LAU\_CODES

**Linked service \***  
CovidBlobStorage

**File path**  
static-data / Directory / CZ AREA CODES.xlsx

**Worksheet mode**  
☒ Name ☐ Index

**Sheet name \***  
CZ LAU CODES

☐ Edit

**First row as header**  
☒

**Import schema**  
☒ From connection/store ☒ From sample file ☐ None

Obrázek 9: Vytvoření datasetu z Excel sheetu

### 5.4.1 Nahrání schematu

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

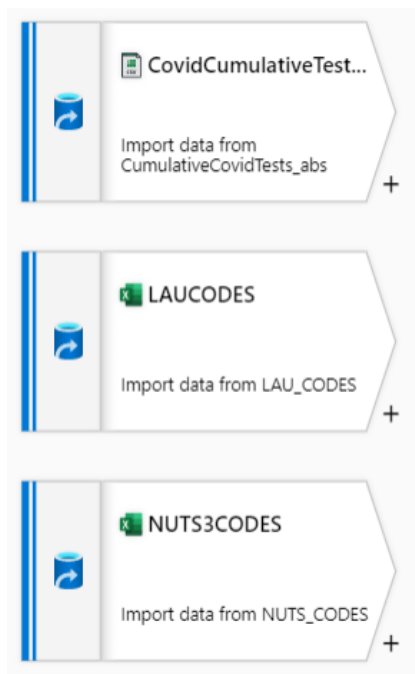
1. Přejděte na Dataset (v ADF Studiu), který představuje soubor v Blob Storage.
2. V záložce schema vyberte možnost **Import schema**.
3. Pokud zvolíte **From connection/store**, ADF se pokusí naimportovat schema z cesty, která je specifikovaná v datasetu. Náš dataset má parametrizované jméno souboru 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. Je více možností:
  - nastavit výchozí hodnotu na jméno souboru, které už ve storage máte - např. soubor který jste stáhli v kroku 5.3.

- 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 vyzváni po kliknutí na import
4. Pro možnost **From sample file** bude soubor pravděpodobně moc velký, museli byste jej zkrátit.
  5. Zkontrolujte, že se vám podařilo nahrát schema. Na typy sloupců teď koukat nemusíte - důležité je, aby bylo vidět jaké a kolik jich je.

#### 5.4.2 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 10.
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 5.4.1).



Obrázek 10: Data Flow zdrojové datasety

### 5.4.3 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. Následně pomocí operace **select** smažte duplikátní (např. `kraj_nuts_kod` obsahuje stejnou 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.



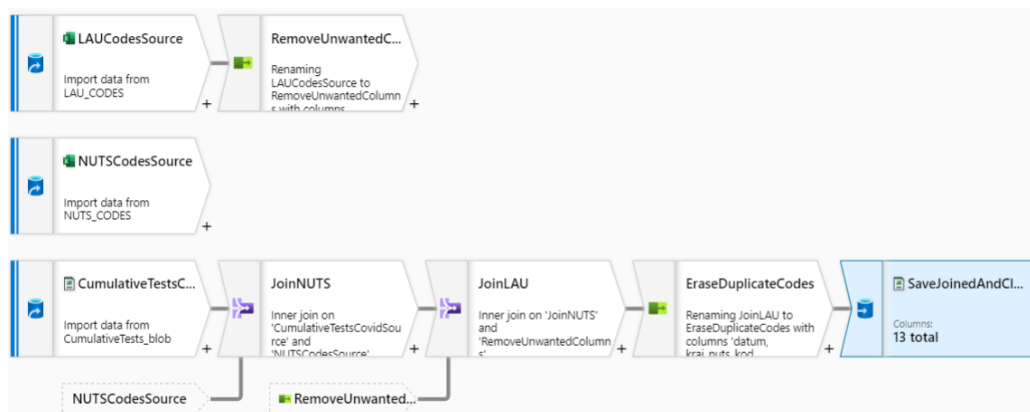
3. Ukončující položkou Data flow je vždy Sink dataset - destinace, kam se mají transformovaná data uložit.
  - (a) Založte nový dataset, obdobným způsobem jako u datasetu, kam ukládáme stažená data z Covid portálu. Opět využijeme Blob Storage a dataset budeme ukládat jako CSV
  - (b) **Pozor**, dataset, který používáme jako Sink nemůže mít specifikované jméno souboru. 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 opět budou mít čas spuštění pipeline. **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 pipeline). Pole `filename` ponechte prázdné. Viz Obrázek 11.
  - (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 12.
6. Změny uložte kliknutím na **Publish all** → **Publish**.

Linked service \* CovidBlobStorage Test connection Edit + New Learn more

File path \* cumulative-tests / @dataset().directory\_name / File

Obrázek 11: Nastavení cesty u Sink datasetu použitého v Data flow.

## 5.5 Přidání Data flow aktivity do pipeline



Obrázek 12: Takto může vypadat výsledné Data flow.

## 5.5 Přidání Data flow aktivity do pipeline

1. Přejděte zpět do 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: *filename* (jméno zdrojového datasetu) a *directory\_name* (jméno adresáře kam uložit výsledek).
  - (a) **filename**: Jelikož chceme, aby vstupem do Data flow byl dataset stažený předchozí **Copy data** aktivitou, musí být obsah *filename* stejný, jako u předchozí aktivity u Sink datasetu (viz krok 3f).
  - (b) **directory\_name**: Sestavte název z vhodného jména a času spuštění pipeline, např. `@concat('covid_tests_joined', '-' ,replace(pipeline().TriggerTime,':','-' ))`.
    - pokud kopírujete kód přímo z dokumentu, dejte si pozor jestli se zkopíroval správně (tj. bez newline, bez mezer navíc apod.)
4. Kliknutím na zelený obdélníček u **Copy data** aktivity přetáhněte šipku vedoucí do Data flow aktivity. Tím zajistíte, že se dataset nejdříve stáhne a až pak se spustí Data flow.

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

## 6 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.

### 6.1 Otevření pracovního prostředí

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

### 6.2 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 [13](#)

# New Cluster

Cancel
Create Cluster
DB

Cluster Name

coviddatacluster

Cluster Mode ?

Single Node | v

Databricks Runtime Version ? [Learn more](#)

Runtime: 8.3 (Scala 2.12, Spark 3.1.1) | v

**Note** Databricks Runtime 8.x and later use Delta Lake as the default table format.

Autopilot Options

☒ Terminate after 

60

 minutes of inactivity ?

Node Type ?

Standard\_DS3\_v2 14 GB Memory, 4 Cores | v ?

DBU / hour: 0.75 ?

Standard\_DS3\_v2

▶ Advanced Options

Obrázek 13: Nastavení clusteru

## 6.3 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, spustte ho.

## 6.4 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 14.

## 6.5 Přidání Azure Databricks notebook aktivity

**New linked service (Azure Databricks)**

Connect via integration runtime \* ⓘ  
 AutoResolveIntegrationRuntime

Account selection method \*  
 From Azure subscription

Azure subscription \* ⓘ  
 Azure subscription 1 (a62afc95-1d37-490c-a5d0-3cec4f6d6378)

Databricks workspace \* ⓘ  
 CovidDatabricks2

Select cluster  
☐ New job cluster ☒ Existing interactive cluster ☐ Existing instance pool

Databrick Workspace URL \* ⓘ  
 https://adb-7860537435374269.9.azuredatabricks.net

Authentication type \*  
 Access Token

☒ Access token ☐ Azure Key Vault

Access token \* ⓘ  
 .....

Choose from existing clusters \* ⓘ  
 coviddatacluster

Obrázek 14: Příklad nastavení ADB linked service

## 6.5 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.

## 6.6 Zobrazení detailů běhu notebooku

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 [3b](#).
  - *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](#)**.

## 6.6 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 [15](#)) se vám zobrazí **Run page url**.
3. Přejděte na **run page** - stránka obsahující konkrétní běh notebooku. Můžete se podívat na výstupy buněk či zkontrolovat (nahore na stránce) jestli byly notebooku předány správné parametry.
4. Neděste se toho, že běh notebook aktivity nebude úspěšný - notebook čeká na vypracování.




## 6.6 Zobrazení detailů běhu notebooku

### Activity runs

Pipeline run ID ee5606f2-9bc3-4917-831a-0e308dc7cfd1

All status ▾

Showing 1 - 3 of 3 items

| Activity name  | Activity type | Run start ↑↓        | Duration | Status      |
|--|---------------|---------------------|----------|-------------|
| CovidNotebo...    | Notebook      | 11/2/21, 2:30:27 PM | 00:00:34 | ✓ Succeeded |
| JoinNutsCodes  | Data flow     | 11/2/21, 2:26:05 PM | 00:04:21 | ✓ Succeeded |
| CopyDataFromWebToBlob  | Copy data     | 11/2/21, 2:25:56 PM | 00:00:08 | ✓ Succeeded |

Obrázek 15: 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.