* Heb je de usecase van je bedrijf (datalake etc) beter in kaart kunnen brengen?
  + Data van Microsoft 365 Customer Engagement of Power Platform wordt geëxporteerd naar CSV bestanden en in Azure Data Lake geplaatst (1 grote pool die opgesplitst moet worden per groep per jaar)
  + Minstens 1 keer per dag:
    - Opkuisen
    - Aanvullen
    - Opsplitsen
    - Ten slotte: Doorsturen naar de klant via SFTP of email (CSV bestanden of parquet files)
* Heb je al een idee van concrete vergelijkingscriteria? (stel een lijst op van meetbare/aantoonbare zaken)
  + Meetbaar:
    - Kostprijs
    - Performantie
    - Mogelijkheid tot debuggen (kan aangetoond worden met een voorbeeld?)
    - Verschil in implementatietijd? (kan verschillen door ervaring?)
  + Minder gemakkelijk meetbaar:
    - Moeilijkheidsgraad in opzet
      * Alles via source control
      * Azure Pipelines om te deployen
      * Moeilijker om te deployen 🡪 minder goeie score
    - Mogelijkheden van de tool (wat is de feature set?)
    - Onderhoudbaarheid
    - Testbaarheid
* Wat is je doelstelling tegen volgende meeting? (wat wil je over 2 weken bereikt hebben)

Er bestaan on-premises tools maar er worden ook cloud etl tools aangeboden, hierdoor is er geen pc nodig

Literatuurstudie

* Wat zijn ETL’s of ELT’s?
* Wat is het verschil tussen ETL en ELT?
* Welke verschillende soorten ETL tools bestaan er?
  + On-premises
  + Open-source
  + Cloud-based
  + Hybrid
* ??? Kort per soort ETL tool de meest populairste opties opnoemen
* ??? Uitleggen waarom bij Net IT er gekozen zal worden voor implementatiemogelijkheden binnen Azure
* ??? Verschillen uitleggen tussen de implementatiemogelijkheden binnen Azure

Vragen:

* Kan documentatie gebruikt worden als bron? Hoe kan dit gebruikt worden?
* Moeten/mogen er tussentitels in de literatuurstudie/stand van zaken zitten?
* Kan een online artikel zonder auteur gebruikt worden? Vaak is dit een bedrijf/organisatie die het artikel publiceerde.

Vragen voor Net IT:

* Waarom wordt de data naar de klant doorgestuurd?

Vragen 15/03:

# Feedback 01/03

zoals terraform bekijken

wat moet er in literatuurstudie?

data factory moet beschreven worden in de literatuurstudie

cli, pipelines etc moeten beschreven worden

elk onderwerp dat ik aanraak moet "gedocumenteerd" worden

specifieker gaan -> uitleggen wat hjet is (uitleggen wat een datalake is)

kan ik connecteren met andere datalakes ook?

auteur: microsoft, geraadpleegd op voor datum

er mogen tussentitels, moet leuk te lezen zijn

Welke IDEs?

Monitoring -> Azure monitoring

Workspace losgekoppeld van persoon

Verschillende soort computes uitzoeken

19/04 feedback

nog niet tastbaar genoeg, niet iedereen kent het (je kan queries doen maar wat wordt er mee bedoeld?)

Titel aanpassen, waar is ELT?

Goeie voorbeeldjes voorzien zodat de lezer goed mee is met wat er gebeurd

Automatisatiemogelijkheden, infrastructure as code?

elke afbeelding moet een voetnoot hebben

screenshots zijn te klein

Volledige implementaties in bijlage toevoegen, belangrijkste in bachelorproef

oefenpresentatie vragen !!!

Feedback Koen:

Details vergelijken met elkaar, ik heb de ene transformatie hier en die voer ik zo uit (in adf) en bij databricks doe ik het zo, die in detail gaan vergelijken

# Creation

## Databricks

Managed Resource group source: https://learn.microsoft.com/en-us/answers/questions/1332779/azure-databricks

# Source control

## ADF

* Data Factory service geeft geen repository voor het opslaan van JSON entiteiten voor wijzigigen, de enigste manier op wijzigingen op te slaan is via de **Publish All** knop en alle wijzigingen worden direct gepubliceerd naar data factory service
* Data Factory service is niet geoptimaliseerd voor samenwerking en version control
* Azure Resource Manager (ARM) template voor het deployen van Data Factory is niet inbegrepen
* Azure Data Factory laat het toe om een Git repository te configureren via Azure Repos of Github

Voordelen :

* Mogelijkheid om wijzigen bij te houden/controleren
* Mogelijkheid om wijzigingen die bugs hebben geïntroduceerd ongedaan te maken
* **Gedeeltelijk opslaan:** wijzigingen kunnen niet als concept opgeslaan worden en alle publicaties moeten datafactory-validatie doorstaan
* **Samenwerking en controle:** elke contributor kan bepaalde rechten krijgen, sommige team members zullen bijvoorbeeld geen changes via git kunnen maken terwijl andere team members die wel zullen kunnen
* **Betere CI/CD:** er kunnen release pipelines gebruikt worden
* **Betere performance:** een factory met git kan 10 keer sneller laden, dit doordat resources via Git gedownload worden

Elke Azure Repos Git repository dat is geassocieerd met data factory heeft een collaboration branch (default: main). Daarnaast kunnen gebruikers ookfeature branches aanmaken. Met behulp van ene pull request kunnen deze feature branches dan gemerged worden met de collaboration branch. Er kan enkel gepubliceerd worden naar Data Factory service vanaf de collaboration branch.

Standaard genereert data factory Resource Manager templates van de gepubliceerde factory naar een branch met de naam *adf\_publish*. (kan gewijzigd worden)

Data factory slaat geen secrets op in git. Aanpassingen in Linked Services waarin secrets zoals passwords zitten worden direct gepubliceerd naar Azure Data Factory service. Het gebruik van een Key Vault of MSI authentication maakt dus continuous integration en deployment makkelijker doordat er geen rekening meer gehouden moet worden met deze secrets tijdens Resource Manager deployment

Source: <https://learn.microsoft.com/en-us/azure/data-factory/source-control>

Azure Data Factory access geven aan andere gebruikers -> **Data Factory Contributor** role geven bij de **Resource Group** van de Data Factory

* Create, edit & delete data factories en child resources including datasets, linked services, pipelines, triggers, en integration runtimes
* Deploy Resource Manager templates
* Manage App Insight alerts voor data factory
* Aanmaken van support tickets

Permissions bij Azure Repos en GitHub :

* Afhankelijk van Data Factory permissions
* Gebruiker die enkel member is van de Reader role kan enkel Data Factory child resources wijzigen en commits uitvoeren op de repo maar niet publishen

Custom roles kan ook maar die skippen we ff

Source: <https://learn.microsoft.com/en-us/azure/data-factory/concepts-roles-permissions>

# Costs

## Azure Data Factory

Berekenen met vCore-hour, controleren in subscription

Vragen promotor

* Mag mijn onderzoek aangepast worden naar een vergelijkende studie tussen Azure Data Factory en Azure Databricks
  + Mag long list en short list weg gelaten worden?
  + Optimaliseren van gegevensverwerkinging in Microsoft Azure: Een vergelijkende analyse tussen Azure Data Factory en Azure Databricks voor het implementeren van Extractie, Transformatie en Laden (ETL)  
      
    OF  
      
    Optimaliseren van gegevensverwerkinging in Microsoft Azure: Een vergelijkende analyse van implementatiemethoden voor Extractie, Transformatie en Laden (ETL)
* In de methodologie heb ik zelf nu veel uitgewerkt, moet ik hier aparte hoofdstukken van maken en dan in de methodologie uitleggen waarover deze hoofdstukken zullen gaan?
* Er gaat een stuk over de publish of collaboration branch van Azure Data Factory. Is dit iets wat verder uitgelegd moet worden in de literatuurstudie? Of moet dit verder uitgewerkt worden in de methodologie zelf?

TODO:

* Textbf tussentitels aanpassen qua stijl

Hoofdstukken

* Proof-of-concepts
  + Azure Data Factory (ADF)
    - Opzet van resources
    - Collaboration en source control
      * Roles en permissions
      * Source control
    - Ophalen van data uit Azure Data Lake
    - Belangrijkste transformaties
      * Determinatie van welke groepen de premie in hun bestand krijgen
      * Bepalen van de kolom “Gr” voor een premie
      * Bepalen van de kolommen “AntheaNumber” en “MembershipNumber” voor een premie
  + Azure Databricks
    - Opzet van resources
    - Collaboration en source control
      * Management en permissions
      * Source control
    - Ophalen van data uit Azure Data Lake
    - Belangrijkste transformaties
      * Determinatie van welke groepen de premie in hun bestand krijgen
      * Bepalen van de kolom “Gr” voor een premie
      * Bepalen van de kolommen “AntheaNumber” en “MembershipNumber” voor een premie

Nieuw:

* Methodologie: uitleggen wat er gebeurd in de volgende hoofdstukken
* Long list / short list: is dit wel nodig doordat enkel adf en databricks vergeleken worden?
  + *selectie van geschikte tools (bij vergelijkende studie, "short-list")*
* Opzet PoC voor zowel ADF als Databricks
  + Opzet resources
  + Collaboration en source control
  + Ophalen data uit Data Lake
  + Belangrijkste transformaties (zodat er een verschil tussen ADF en databricks getoond wordt)
* Uitvoeren testen en verzamelen resultaten
  + Kostprijs
  + Performantie
* Analyse van resultaten

# Testen

## Azure Data Factory

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pipeline | Azure IR | Core count | TTL (Time to live) | Region |
| Small Compute Size | integrationRuntimeSmall | 4 Cores + 4 Driver cores | 0 minutes | West Europe |
| Medium Compute Size | integrationRuntimeMedium | 8 Cores + 8 Driver cores | 0 minutes | West Europe |

## Azure Databricks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Databricks name | Managed Resource Group name | Workflow name | Compute name | Core count |
| databricks-loeka-standard | databricks-loeka-standard-mrg | ComputeSizeSmall | ComputeSizeSmall\_cluster | 4 Cores + 4 Driver cores |
| databricks-loeka-standard-2 | databricks-loeka-standard-2-mrg | ComputeSizeMedium | ComputeSizeMedium\_cluster | 8 Cores + 8 Driver cores |

# IDE

## ADF

<https://github.com/Huachao/azure-content/blob/master/articles/data-factory/data-factory-build-your-first-pipeline-using-vs.md>

## Databricks

<https://learn.microsoft.com/en-us/azure/databricks/dev-tools/vscode-ext/>

<https://learn.microsoft.com/en-us/azure/databricks/dev-tools/visual-studio-code>

## Feedback 03/05

kort de andere dingen bespreken

long en short list niet benoemen, in literatuurstudie benoemen dat er alternatieven zijn maar dat we over azure

als iemand lastig doet -> verdedigen dat er gekeken is naar alternatieven maar dat short/long niet relevant is

Niet bij zijn -> meeting gecanceld (niet langer dan 10 minuten wachten)

nog steeds recht op feedback -> zelf mail sturen naar mevrouw Lena De Mol, in CC: mevrouw liesbeth lewylie (team coordinator)

Ook voor presentatie oefenen er zelf op staan om dit te kunnen oefenen

laatste week van examenperiode komt datum presentatie

# Aparte feedback (datum onbekend)

Welke IDEs?

Monitoring -> Azure monitoring

Workspace losgekoppeld van persoon

Verschillende soort computes uitzoeken

# Costs

## 05/05

ADF totaal: 0.72

Small: 1.1259 x 0,251 = **0,282601 (10m) (3m 11s cluster)**

Medium: 1.7344 x 0,251 = **0,435334 (7m 53s) (3m 6s cluster)**

0,282601 + 0,435334 = 0,717935 = 0,72

Databricks

Databricks-loeka-standard: 0,10

Databricks-loeka-standard-mrg: 0,06 + 0,06

= 0,10 + 0,06 + 0,06 = **0,22 (10m 44s)**

Databricks-loeka-standard-2: 0,12

Databricks-loeka-standard-2mrg: 0,10 + 0,10 + 0,05 + 0,05

= 0,12 + 0,10 + 0,10 + 0,05 + 0,05 = **0,42 (7m 17s)**

## 06/05

ADF total: 0,76

Small: 1.1513 x 0,251 = **0.288976 (10m 48s) (2m 57s cluster)**

Medium: 1.8602 x 0,251 = **0.46691 (8m 6s) (3m 24s cluster)**

0.288976 + 0.46691 = 0.755887 = 0.76

Databricks

Databricks-loeka-standard: 0.09

Databricks-loeka-standard-mrg: 0.06 + 0.06

= 0.09 + 0.06 + 0.06 = **0.21 (10m 16s)**

Databricks-loeka-standard-2: 0.12

Databricks-loeka-standard-2-mrg: 0.10 + 0.10 + 0.05 + 0.05

= 0.12 + 0.10 + 0.10 + 0.05 + 0.05 = **0.42 (7m 43s)**

## 07/05

ADF total: 0.73

Small: 1.1702 x 0.251 = **0.2937202 (10m 2s) (3m 9s)**

Medium: 1.7266 x 0.251 = **0.4333766 (7m 56s) (2m 42s)**

Databricks

Databricks-loeka-standard: 0.09

Databricks-loeka-stanard-mrg: 0.6 + 0.6

= 0.9 + 0.6 + 0.6 = **0.21 (10m 20s)**

Databricks-loeka-standard-2: 0.12

Databricks-loeka-standard-2-mrg: 0.12 + 0.12 + 0.05 + 0.05

= 0.12 + 0.12 + 0.12 + 0.05 + 0.05 = **0.46 (10m 30s)**

**ADF SMALL**

|  |  |  |
| --- | --- | --- |
| Date | Costs | Time |
| 05/05 | 0.28 | 10m (3m 11s) |
| 06/05 | 0.29 | 10m 48s (2m 57s) |
| 07/05 | 0.29 | 10m 2s (3m 9s) |

**ADF MEDIUM**

|  |  |  |
| --- | --- | --- |
| Date | Costs | Time |
| 05/05 | 0.44 | 7m 53s (3m 6s) |
| 06/05 | 0.47 | 8m 6s (3m 24s) |
| 07/05 | 0.43 | 7m 56s (2m 42s) |

**DATABRICKS SMALL**

|  |  |  |
| --- | --- | --- |
| Date | Costs | Time |
| 05/05 | 0,22 | 10m 44s (3m 48s) |
| 06/05 | 0,21 | 10m 16s (3m 45s) |
| 07/05 | 0.21 | 10m 20s (4m 8s) |

**DATABRICKS MEDIUM**

|  |  |  |
| --- | --- | --- |
| Date | Costs | Time |
| 05/05 | 0,42 | 7m 17s (3m 22s) |
| 06/05 | 0,42 | 7m 43s (3m 44s) |
| 07/05 | 0.46 | 10m 30s (6m 27s) |

# Uitleggen literatuurstudie

* **Azure**
  + Subscription
  + Resource group
* **Azure Data Factory (ADF)**
  + Verschillende transformaties
  + Verschillende source types
* **Azure Databricks**
  + Managed Resource Group
  + Cluster
  + Databricks Connect
  + Databricks Asset Bundles (DABs)
* Source control
  + Azure Repos
* IntelliSense
* Schema-validatie
* **Azure Resource Manager (ARM) templates**
* **Continuous Integration (CI) en Continuous Deployment (CD)**
* **Azure Data Lake Storage Gen2**
* **Key Vault**
* **App registration**
* **Microsoft Power BI**
  + Power query
* CSV bestanden
* Wat zijn worker cores en driver cores
* **Microsoft Cost Management / Subscription costing bekijken**
* **Infrastructure as Code (IaC)**

Small:

* ADF iets duurder
* ADF en databricks consistent in prijs
* Performantieverschillen moeilijk te zien, lopen vrij gelijk

Medium:

* ADF waarschijnlijk iets duurder
* Beide redelijk consistent in prijs
* Databricks duidelijk sneller met uitzondering van 1 hoge cluster startup tijd

Cluster startup tijd sws hoger bij databricks

Schrijven naar CSV bestand nog doen!!