

Dataplatform – informatie uit onderzoek samengevat

+++++

Omschrijving opdracht / probleem

De provincie Utrecht beschikt momenteel over twee operationele platforms. Beide platforms hebben hun eigen uitdagingen, de provincie Utrecht heeft om advies gevraagd bij het beoordelen en verbeteren van deze platforms om aan de huidige eisen te voldoen en tegelijkertijd flexibel te blijven voor toekomstige uitbreidingen.

+++++

Aanbevelingen

- Integreer Microsoft Fabric met bestaande netwerkinfrastructuur, waarbij directe verbindingen met on-premises systemen worden overwogen en afhankelijkheid van het publieke internet wordt geminimaliseerd.
- Neem security en monitoring in acht, inclusief het instellen van alert rules en het toepassen van naming conventions en tagging binnen Fabric Workspaces.
- Pas gebruikers- en groepenbeheer toe binnen Fabric via Microsoft Entra ID (voorheen Azure AD) en sluit aan op het IAM-systeem van de provincie.
- Gebruik geïntegreerde Fabric-componenten, zoals Lakehouse, Dataflows, Notebooks en Power BI, voor een uniforme dataverwerking en -analyse.
- Ondersteun zowel gestructureerde als ongestructureerde data binnen het OneLake-concept van Microsoft Fabric.
- Zorg voor een hybride benadering van BI en AI, waarbij het platform niet alleen wordt ingezet voor rapportage, maar ook voor geavanceerde analyses en machine learning-modellen.
- Implementeer een DTAP-strategie binnen Fabric, waarbij test- en acceptatieomgevingen in Workspaces worden ingericht met duidelijke versiebeheer- en releaseprocessen.
- Zorg voor actief beheer en lifecyclemanagement van Fabric Workspaces, inclusief datagovernance en kostenbeheer via Microsoft Purview en Cost Management.
- Implementeer een metadatamodel in Fabric en zorg voor strikte governance en documentatie van datastromen en datasets.

+++++

Samenvatting

De provincie Utrecht, bestaande uit 26 gemeenten met een gezamenlijke bevolking van 1,36 miljoen, is de oudste en kleinste provincie van Nederland en staat bekend om haar hoge bevolkingsdichtheid.

Analyse en advies

De provincie beheert een divers scala aan gegevens, variërend van publieke en private data tot zeer gevoelige Persoonlijk Identificeerbare Informatie (PII), in gestructureerde, semi-gestructureerde en ongestructureerde formaten. De huidige infrastructuur van het dataplatform kent uitdagingen op het gebied van datadeling, engineering, bruikbaarheid en beveiliging. Er is behoefte aan een meer gecentraliseerde aanpak, maar de route daarnaartoe is nog onduidelijk.

Het advies is om het huidige platform te migreren naar een nieuw platform gebaseerd op een centraal data lakehouse, gefaciliteerd door Microsoft Fabric. Dit biedt een moderne en flexibele architectuur die de huidige uitdagingen van de provincie Utrecht beter kan adresseren.

Migratieplan voor de implementatie

1. Bepalen welke datastromen (ADF of Synapse) migratie vereisen naar het nieuwe platform.
2. Templates verkrijgen voor de geïdentificeerde datastromen. Omdat Fabric zowel Synapse als ADF ondersteunt, kan de bestaande code eenvoudig worden hergebruikt.
3. Vaststellen van de gegevensoverdrachtseisen van de bestaande naar de nieuwe platformen. Indien nodig een eenmalige bulkoverdracht uitvoeren naar een uniforme data lake.
4. Analyse van data-afhankelijkheden: Een grondige analyse uitvoeren om afhankelijkheden tussen verschillende datastromen, applicaties en diensten te identificeren. Dit helpt bij het prioriteren van migratietaken en waarborgt een soepele overgang.
5. Datakwaliteit en transformatie: Voor de migratie de bestaande data beoordelen en waar nodig opschonen of transformeren, zodat alleen relevante en schone data wordt overgezet.

6. Uitvoeren van uitgebreide tests en controles om te verzekeren dat het nieuwe dataplatform klaar is voor productie.
7. Beveiliging en compliance: Controleren of het nieuwe platform voldoet aan beveiligings- en nalevingsvereisten. Dit omvat toegangscontrole, encryptie en compliance monitoring om gevoelige data te beschermen.
8. Na succesvolle implementatie het oude dataplatform gefaseerd buiten gebruik stellen.

Fabric en DTAP

Om meerdere omgevingen (Dev, Test, Prod) binnen Microsoft Fabric in te richten, zijn de volgende activiteiten van belang:

1. **Workspace-organisatie:** Gebruik aparte Fabric Workspaces per omgeving (Dev, Test, Acceptatie, Productie).
2. **OneLake Data Lake beheer:** Gebruik gescheiden lakehouses binnen the OneLake voor de verschillende omgevingen.
3. **Consistente naamgeving:** Hanteer een duidelijke en consequente naamgeving van Workspaces, Lakehouses, Pipelines en Dataflows.
4. **Omgevingsspecifieke configuraties:** Stem instellingen zoals toegangsrechten, netwerkinstellingen en dataretentie af per omgeving.
5. **Geautomatiseerde inrichting:** Maak gebruik van Infrastructure as Code(IaC), CI/CD pipelines en Deployment Pipelines voor het overzetten van assets tussen omgevingen en het managen van capaciteit.
6. **Releaseworkflow:** Gebruik versiebeheer binnen Fabric Workspaces en implementeer een gestandaardiseerd deploymentproces.
7. **Dataseparatie en beveiliging:** Pas technieken zoals row-level security (RLS), column-level security (CLS) en datamaskering toe voor het beschermen van gevoelige gegevens.
8. **Monitoring en logging:** Maak gebruik tools om van Microsoft Fabric te monitoren.
9. **Beleidsafdwinging:** Definieer governance policies via Microsoft Purview en zorg voor compliance met beveiligingsrichtlijnen.
10. **Continue optimalisatie:** Evalueer en verbeter het gebruik van Microsoft Fabric op basis van monitoringgegevens en gebruikersfeedback.

Conclusie

De adoptie van Microsoft Fabric biedt de provincie Utrecht een robuuste en toekomstbestendige oplossing voor een gecentraliseerd dataplatform. Dit platform ondersteunt een effectieve datagovernance, samenwerking en innovatie. Met een doordachte implementatiestrategie en continue ondersteuning kan de provincie haar data optimaal benutten voor betere besluitvorming, efficiëntere operaties en het behalen van strategische doelstellingen.

+++++