

Evolution d'un Entrepot de Donnees

Projet Data Engineering E6 – Region Hauts-de-France

Hamza Elbrek

BTS SIO – Option SLAM

Fevrier 2026

Competences visees

C16 – Exploiter des donnees

C17 – Gerer le patrimoine de donnees

Contexte

Azure SQL · Python · Terraform

5 schemas · 7 tables de faits

- 1 Contexte et objectifs
- 2 Architecture de l'entrepot
- 3 Maintenance et journalisation
- 4 Strategie de sauvegarde
- 5 Securite des acces (RBAC + RLS)
- 6 Variations de dimensions (SCD)
- 7 Bilan et competences

- 1 Contexte et objectifs
- 2 Architecture de l'entrepot
- 3 Maintenance et journalisation
- 4 Strategie de sauvegarde
- 5 Securite des acces (RBAC + RLS)
- 6 Variations de dimensions (SCD)
- 7 Bilan et competences

Projet E5 – Acquis

- Entrepot Azure SQL deploye via Terraform
- Schema en etoile : 6 dimensions, 5 faits sur 7
- Pipelines ETL Python
- 3 roles RBAC initiaux

Limites identifiees

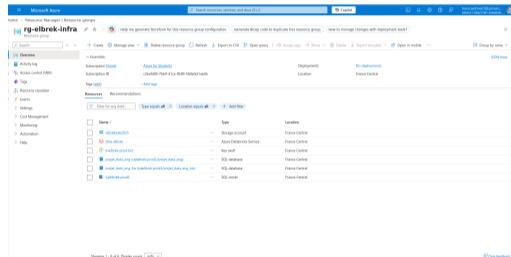
- Pas de journalisation structuree
- Pas de strategie de sauvegarde
- 2 tables de faits manquantes
- Aucune restriction geographique des acces

Objectif E6

Faire evoluer l'entrepot vers un **systeme effectif en conditions reelles** :
maintenance, securite, historisation, nouvelles sources.

Ressources provisionnees (Terraform)

- **Azure SQL Database – Data Warehouse**
- **ADLS Gen2 – Data Lake (raw / curated)**
- **Azure Databricks – Traitement Spark**
- **Azure Data Factory – Orchestration**
- **Key Vault – Gestion des secrets**
- **Log Analytics Workspace – Auditing SQL**



- 1 Contexte et objectifs
- 2 Architecture de l'entrepot**
- 3 Maintenance et journalisation
- 4 Strategie de sauvegarde
- 5 Securite des acces (RBAC + RLS)
- 6 Variations de dimensions (SCD)
- 7 Bilan et competences

5 schemas fonctionnels

- stg – Zone de staging (donnees brutes CSV)
- dwh – Dimensions + Faits (cœur du DWH)
- dm – Datamarts analytiques (vues materialisees)
- analytics – Tableaux de bord + monitoring
- **security** – RBAC, RLS, agences, employes [E6]

Nouveautés E6

- fait_emploi – Chomage / emploi
- fait_menages – Composition menages
- log_etl – Journalisation ETL
- log_erreurs – Erreurs pipeline
- Schema **security** complet



- 1 Contexte et objectifs
- 2 Architecture de l'entrepot
- 3 Maintenance et journalisation**
- 4 Strategie de sauvegarde
- 5 Securite des acces (RBAC + RLS)
- 6 Variations de dimensions (SCD)
- 7 Bilan et competences

Workflow en 8 étapes :

- ① Reception mail → creation ticket
- ② Prise en charge (correctif ou escalade)
- ③ Priorisation manager si necessaire
- ④ Accuse de reception expéditeur
- ⑤ Diagnostic via vues de monitoring
- ⑥ Resolution et tests
- ⑦ Notification de cloture
- ⑧ Archivage backlog

Matrice de priorisation

P1 Critique	Pipeline ETL en erreur
P2 Haute	Donnees incorrectes
P3 Normale	Demande evolution
P4 Basse	Documentation

Regle cle

Aucune demande ne reste sans ticket ouvert.

Journalisation – Double niveau de tracabilite

Niveau 1 – Logs Python

Module logging standard :
INFO, WARNING, ERROR
→ Console + fichier local

Niveau 2 – Table SQL

dwh.log_etl enregistre :
etape, table, statut, duree, nb lignes

Notifications email

etl_notifier.py envoie un rapport HTML
a chaque fin de pipeline (succes ou erreur)

log_id	data_execution	etape	table_cible	statut	nb_lignes
31	2024-02-17 15:54:31.000000	BACKUP_BACFAC	DATABASE	SUCCESS	0
30	2024-02-17 15:29:07.000000	BACKUP_BACFAC	DATABASE	DEBUT	0
29	2024-02-17 15:25:37.000000	load_facts	ALL	SUCCESS	0
28	2024-02-17 15:25:41.000000	load_facts	ALL	DEBUT	0
27	2024-02-17 15:24:58.000000	load_dimensions	ALL	SUCCESS	0
26	2024-02-17 15:24:37.000000	load_dimensions	ALL	DEBUT	0
25	2024-02-17 15:04:40.000000	BACKUP_BACFAC	DATABASE	SUCCESS	0
24	2024-02-17 15:01:59.570000	BACKUP_BACFAC	DATABASE	DEBUT	0

✗ ETL DWH – ERREUR Faits (dwh.fait_entreprises) – 11:51:32

✗ ETL Data Warehouse – Erreur de chargement

Date : 20240218 11:51:34
Projet : Data Engineering – Région Hauts-de-France

Etape	Publ
Table	dwh.fait_entreprises
Heure	11:51:32
Erreur	"Column(s) 'obs_value' do not exist"

Tables traitées avant l'erreur

Table	Statut	Lignes	Heure	Duree	Detail
dwh.fait_regulation	SKIP	0	11:51:31	0.3s	
dwh.fait_entreprises_demo	SKIP	0	11:51:31	0.3s	
dwh.fait_revenus	SKIP	0	11:51:32	0.3s	
dwh.fait_entreprises	SKIP	0	11:51:33	0.3s	
dwh.fait_messages	SKIP	0	11:51:33	0.3s	

- 1 Contexte et objectifs
- 2 Architecture de l'entrepot
- 3 Maintenance et journalisation
- 4 Strategie de sauvegarde**
- 5 Securite des acces (RBAC + RLS)
- 6 Variations de dimensions (SCD)
- 7 Bilan et competences

Strategie de sauvegarde – 2 niveaux

Niveau 1 – PITR Azure (automatique)

- Sauvegarde complete + differentielle toutes les 12h
- Retention court terme : **14 jours**
- Retention long terme : 4 sem / 12 mois / 3 ans
- Restauration a la minute pres (RTO \approx 1h)

Niveau 2 – Export BACPAC (pipeline ETL)

- Etape 5 du pipeline ETL
- Export .bacpac horodate vers ADLS Gen2

The image displays two screenshots from the Microsoft Azure portal. The top screenshot shows the configuration page for a SQL database named 'projet_data_eng' (sqldebprod2/projet_data_eng). It details various settings including Service tier (Standard), Max storage (250 GB), Availability (High), Backup (Differential backup frequency: 12 hours, Retention: 14 days), and Security (Microsoft Defender for SQL: Disabled). The bottom screenshot shows a file explorer view for a container named 'raw'. It lists three backup files: 'projet_data_eng_20240717_170100.bacpac', 'projet_data_eng_20240717_170100.bacpac', and 'projet_data_eng_20240717_170100.bacpac', all with a size of 6.45 MB and an 'Available' state.

- 1 Contexte et objectifs
- 2 Architecture de l'entrepot
- 3 Maintenance et journalisation
- 4 Strategie de sauvegarde
- 5 Securite des acces (RBAC + RLS)**
- 6 Variations de dimensions (SCD)
- 7 Bilan et competences

4 roles metier (E6)

role_admin	Controle total
role_etl_process	R/W stg + dwh
role_analyst	Lecture dwh/dm/analytics
role_consultant	Lecture filtree par RLS

Changement cle

Suppression de role_dwh_admin et introduction de role_consultant avec filtrage géographique.

Query 1: `SELECT * FROM sys.database_principals`

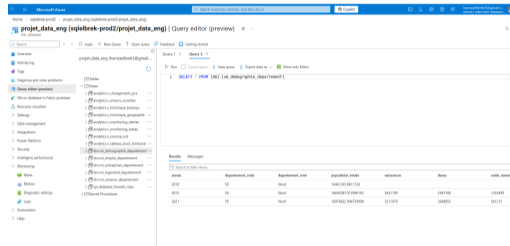
id	name	type	schema_name	database_name	creation_date	modification_date
1	dbo	5	SQL_USER	dbo	2023-04-08T09:19:42.2875000	2023-04-01-01
2	guest	5	SQL_USER	guest	2023-04-08T09:19:42.2875000	2023-04-01-01
3	INFORMATION_SCHEMA	5	SQL_USER		2023-04-08T09:19:42.2875000	2023-04-01-01
4	sys	5	SQL_USER		2023-04-08T09:19:42.2875000	2023-04-01-01
5	role_etl_process	8	DATABASE_ROLE		2023-02-02T11:17:52.9875000	2023-02-01-01
6	role_analyst	8	DATABASE_ROLE		2023-02-02T11:17:52.9875000	2023-02-01-01
7	jean.dupont	5	SQL_USER	dbo	2023-02-02T11:18:00.8000000	2023-02-01-01
8	sophia.martin	5	SQL_USER	dbo	2023-02-02T11:18:00.8000000	2023-02-01-01
9	role_admin	8	DATABASE_ROLE		2023-02-02T11:18:00.8000000	2023-02-01-01
10	role_consultant	8	DATABASE_ROLE		2023-02-02T11:18:00.8000000	2023-02-01-01

Referentiel employes

519 employes fictifs generes
(contrainte RGPD – pas d'import de donnees bancaires reelles sur Azure)

Architecture RLS

- Predicat applique sur `dwh.dim_geographie`
- Propagation automatique a toutes les vues `dm.*` et `analytics.*`
- Fonction `security.fn_rls_geographie`



Hierarchie – 4 niveaux

Directeur Regional	1
Directeurs Departement	5
Directeurs Agence	101
Collaborateurs	412

101 agences

Communes Hauts-de-France
> 10 000 habitants
(02, 59, 60, 62, 80)

- 1 Contexte et objectifs
- 2 Architecture de l'entrepot
- 3 Maintenance et journalisation
- 4 Strategie de sauvegarde
- 5 Securite des acces (RBAC + RLS)
- 6 Variations de dimensions (SCD)**
- 7 Bilan et competences

Slowly Changing Dimensions – 3 types implementes

SCD Type 1

Ecrasement

Correction de libelle NAF.
L'historique est perdu.

→ dim_activite

SCD Type 2

Historisation

Fusion de communes.
Colonnes est_actif,
version, date_debut.

→ dim_geographie

SCD Type 3

Colonne precedente

Renommage categorie PCS.
1 seul niveau d'historique
conserve.

→ dim_demographie

Microsoft Azure | Search resources, services, and docs (Ctrl+K)

projct_data_eng (sqlbirek-prod2/projct_data_eng) | Query editor (preview)

Query 1 X Query 2 X

1 SELECT TOP (1000) FROM [dw].[dim_geographie]

Results Messages

date_creation	date_modification	date_fin_valide	est_actif	version	date_debut_valide
2026-01-23T13:47:29.8566667	2026-01-23T13:47:29.8566667		True	1	2010-01-01T00:00:00.0000000

- 1 Contexte et objectifs
- 2 Architecture de l'entrepot
- 3 Maintenance et journalisation
- 4 Strategie de sauvegarde
- 5 Securite des acces (RBAC + RLS)
- 6 Variations de dimensions (SCD)
- 7 Bilan et competences**

Maintenance & monitoring

- ✓ Processus incidents (8 étapes)
- ✓ Table `dwh.log_etl + log_erreurs`
- ✓ Notifications email automatiques
- ✓ Vues analytics de supervision

Sauvegarde

- ✓ PITR Azure 14 jours + retention LT
- ✓ Export BACPAC vers ADLS Gen2
- ✓ Nettoyage automatique

Sécurité

- ✓ 4 rôles RBAC métier
- ✓ RLS sur `dwh.dim_geographie`
- ✓ 519 employés / 101 agences chargées
- ✓ Schéma `security` dédié

Évolution du modèle

- ✓ 2 nouvelles tables de faits
- ✓ SCD Type 1, 2 et 3 implémentés
- ✓ 7 tables de faits complétées

C16 – Exploiter des données

- Intégration de nouvelles sources CSV
- Requetage SQL analytique (datamarts)
- Vues de monitoring et supervision
- Row-Level Security sur les consultations

C17 – Gérer le patrimoine de données

- Politique de sauvegarde et restauration
- Journalisation et traçabilité des ETL
- Gestion des rôles et des accès (RBAC)
- Historisation SCD Type 1/2/3
- Procédures de scalabilité documentées

Valeur ajoutée

Un entrepôt de données passe de l'**état initial E5** à un **système opérationnel complet** :
monitore, sécurise, sauvegarde et prêt à évoluer.

Merci pour votre attention

Auteur	Hamza Elbrek
Projet	E6 – Evolution d'un Entrepot de Donnees
Region	Hauts-de-France
Stack	Azure SQL · Python · Terraform · ADLS Gen2
Date	Fevrier 2026

Questions ?