RAPPORT DE PROJET

Projet d'Intégration de Données et Business Intelligence

realiser par : Aymane ELMKADMI

PostgreSQL → Snowflake → Pentaho → Power BI

Analyse des Données Étudiantes

Date : Janvier 2026

Table des Matières

Table des Matières 2
1. Introduction 4
1.1 Contexte du Projet 4
1.2 Objectifs 4
2. Architecture du Projet 5
2.1 Vue d'Ensemble de l'Architecture 5
2.2 Flux de Données 5
3. Implémentation Technique 6
3.1 Phase 1 : PostgreSQL et pgAdmin 6
3.1.1 Configuration de la Base de Données 6
3.2 Phase 2 : Migration vers Snowflake (RAW) 6
3.2.1 Configuration Snowflake 6
3.2.2 Avantages de Snowflake 6
3.3 Phase 3 : Transformations avec Pentaho 7
3.3.1 Processus de Transformation 7
3.3.2 Types de Transformations Appliquées 7
3.3.3 Résultats des Transformations 7
3.4 Phase 4 : Chargement dans Snowflake (ANALYTICS) 7
3.4.1 Structure du Schéma ANALYTICS 7
3.5 Phase 5 : Visualisation avec Power BI 8
3.5.1 Connexion Power BI - Snowflake 8
3.5.2 Visualisations Créées 8
4. Résultats et Insights 9
4.1 Données Analysées 9
4.2 Insights Clés 9
4.2.1 Distribution Démographique 9
4.2.2 Performance Académique 9
4.2.3 Tendances Temporelles 9
5. Technologies Utilisées 10
6. Défis et Solutions 11
6.1 Défis Rencontrés 11
6.1.1 Migration PostgreSQL vers Snowflake 11
6.1.2 Transformations Pentaho 11

6.1.3 Performance des Requêtes..... 11
6.2 Bonnes Pratiques Appliquées 11
7. Conclusion 12
7.1 Réalisations 12
7.2 Compétences Acquises 12
7.3 Perspectives d'Amélioration 12
7.4 Conclusion Générale 12
Annexes 13
Annexe A : Structure de la Base de Données 13
Annexe B : Champs de la Table STUDENTS_POSITION 13
Annexe C : Outils et Versions 13

1. Introduction

Ce rapport présente un projet complet d'intégration de données et de Business Intelligence, mettant en œuvre une architecture moderne de gestion et d'analyse de données étudiantes. Le projet démontre l'utilisation de technologies de pointe pour transformer des données brutes en informations exploitables à travers des visualisations interactives.

1.1 Contexte du Projet

Dans le cadre de l'amélioration de la gestion des données académiques, ce projet vise à créer une chaîne complète d'intégration et d'analyse de données, permettant aux décideurs d'obtenir des insights pertinents sur les performances et les caractéristiques des étudiants.

1.2 Objectifs

- Mettre en place uneinfrastructure d'intégration de données robuste et scalable
- Transformer et nettoyer les données brutes pour l'analyse
- Créer un entrepôt dedonnées cloud sur Snowflake
- Développer des visualisations interactives avec Power BI
- Fournir des insightsactionnables pour la prise de décision

2. Architecture du Projet

Le projet suit une architecture en plusieurs couches, allant de la source de données jusqu'à la visualisation finale. Cette approche garantit la séparation des responsabilités et facilite la maintenance.

2.1 Vue d'Ensemble de l'Architecture

Couche	Technologie	Rôle
Source	PostgreSQL (pgAdmin)	Base de données source contenant les données brutes
Staging	Snowflake (Schema: RAW)	Zone de réception des données brutes
Transformation	Pentaho Data Integration	ETL - Transformation et nettoyage des données
Analytique	Snowflake (Schema: ANALYTICS)	Entrepôt de données transformées et optimisées
Visualisation	Microsoft Power BI	Tableaux de bord interactifs et analyses visuelles

2.2 Flux de Données

Le flux de donnéessuitunparcours structuré en cinq étapes principales, garantissant l'intégrité et la qualitédesdonnées à chaque niveau de traitement.

1. ***Extraction:***Lesdonnées sont extraites depuis PostgreSQL via pgAdmin
2. ***ChargementInitial:***Les données brutes sont chargées dans Snowflake (base de donnéesSTUDENT_DB, schéma RAW)
3. ***Transformation:***Pentaho Data Integration applique les transformations nécessaires(nettoyage, agrégation, enrichissement)
4. ***ChargementAnalytique :*** Les données transformées sont enregistrées dans le schéma ANALYTICSde Snowflake
5. ***Visualisation:***PowerBI se connecte à Snowflake pour créer des tableaux de bord interactifs

3. Implémentation Technique

3.1 Phase 1 : PostgreSQL et pgAdmin

La première phase du projet consiste à préparer les données sources dans PostgreSQL. Cette étape fondamentale établit la base de données opérationnelle qui alimentera l'ensemble du pipeline.

3.1.1 Configuration de la Base de Données

- Installation et configuration de PostgreSQL
- Utilisation de pgAdmin pour l'administration
- Création des tables pour les données étudiantes
- Chargement des données initiales

3.2 Phase 2 : Migration vers Snowflake (RAW)

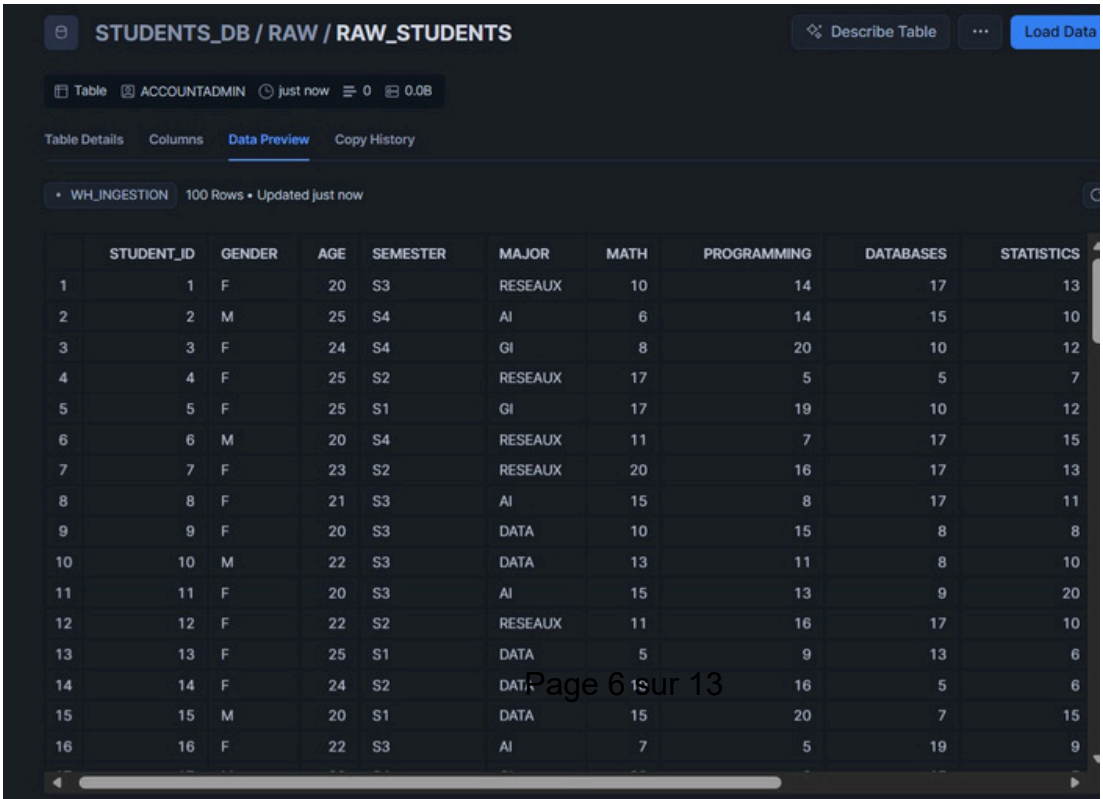
Cette phase marque la transition vers le cloud avec Snowflake, offrant scalabilité et performance pour le traitement des données.

3.2.1 Configuration Snowflake

- ***Base de données :*** STUDENT_DB
- ***Schéma :*** RAW (pour les données brutes)
- Migration complète des données depuis PostgreSQL vers Snowflake

3.2.2 Avantages de Snowflake

- Séparation du stockage et du calcul pour une performance optimale
- Scalabilité automatique selon les besoins
- Support natif du semi-structuré (JSON, Parquet, etc.)
- Partage de donnéessécurisé entre différentes applications



	STUDENT_ID	GENDER	AGE	SEMESTER	MAJOR	MATH	PROGRAMMING	DATABASES	STATISTICS
1	1	F	20	S3	RESEAUX	10	14	17	13
2	2	M	25	S4	AI	6	14	15	10
3	3	F	24	S4	GI	8	20	10	12
4	4	F	25	S2	RESEAUX	17	5	5	7
5	5	F	25	S1	GI	17	19	10	12
6	6	M	20	S4	RESEAUX	11	7	17	15
7	7	F	23	S2	RESEAUX	20	16	17	13
8	8	F	21	S3	AI	15	8	17	11
9	9	F	20	S3	DATA	10	15	8	8
10	10	M	22	S3	DATA	13	11	8	10
11	11	F	20	S3	AI	15	13	9	20
12	12	F	22	S2	RESEAUX	11	16	17	10
13	13	F	25	S1	DATA	5	9	13	6
14	14	F	24	S2	DATA	15	16	5	6
15	15	M	20	S1	DATA	15	20	7	15
16	16	F	22	S3	AI	7	5	19	9

3.3 Phase 3 : Transformations avec Pentaho

Pentaho Data Integration (PDI) est utilisé pour effectuer des transformations complexes sur les données, garantissant leur qualité et leur pertinence pour l'analyse.

3.3.1 Processus de Transformation

Les transformations Pentaho ont été conçues pour préparer les données pour l'analyse. Ces transformations incluent le nettoyage, l'enrichissement et l'agrégation des données étudiantes.

3.3.2 Types de Transformations Appliquées

- ***Nettoyage des Données :*** Suppression des doublons, traitement des valeurs manquantes, normalisation des formats
- ***Enrichissement :*** Calcul de métriques dérivées (groupes d'âge, niveaux de performance)
- ***Agrégation :*** Création de vues agrégées par semestre, spécialité, et caractéristiques démographiques
- ***Validation :*** Vérification de l'intégrité des données et des règles métier

3.3.3 Résultats des Transformations

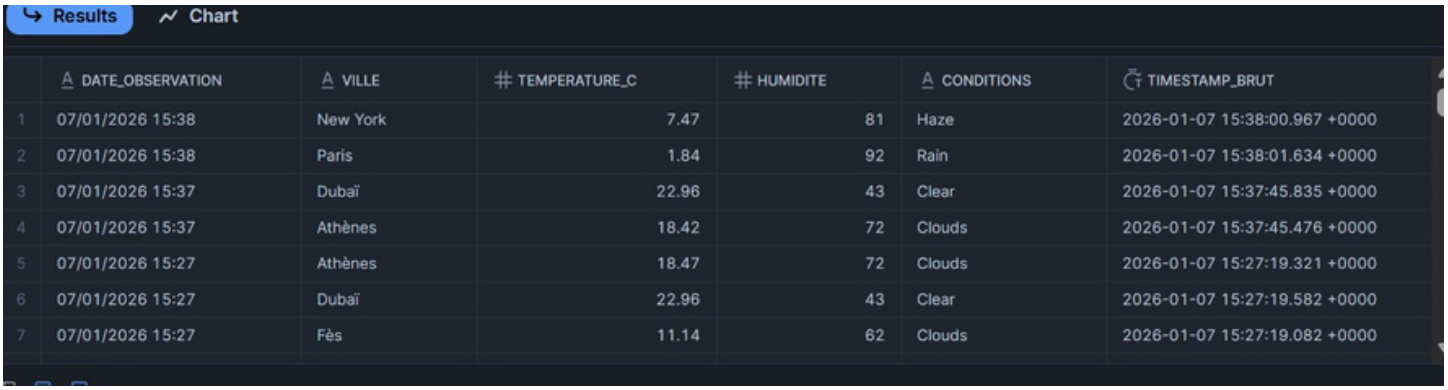
Les transformations Pentaho ont produit des données propres et structurées, prêtes pour l'analyse. Les captures d'écran des résultats ont été documentées pour validation et traçabilité du processus.

3.4 Phase 4 : Chargement dans Snowflake (ANALYTICS)

Après transformation, les données ont été chargées dans le schéma ANALYTICS de Snowflake, créant ainsi un entrepôt de données optimisé pour l'analyse et le reporting.

3.4.1 Structure du Schéma ANALYTICS

- Table principale:STUDENTS_POSITION
- Données enrichiesavec attributs calculés
- Optimisation desindex pour des requêtes performantes



Results | Chart

	DATE_OBSERVATION	VILLE	TEMPERATURE_C	HUMIDITE	CONDITIONS	TIMESTAMP_BRUT
1	07/01/2026 15:38	New York	7.47	81	Haze	2026-01-07 15:38:00.967 +0000
2	07/01/2026 15:38	Paris	1.84	92	Rain	2026-01-07 15:38:01.634 +0000
3	07/01/2026 15:37	Dubaï	22.96	43	Clear	2026-01-07 15:37:45.835 +0000
4	07/01/2026 15:37	Athènes	18.42	72	Clouds	2026-01-07 15:37:45.476 +0000
5	07/01/2026 15:27	Athènes	18.47	72	Clouds	2026-01-07 15:27:19.321 +0000
6	07/01/2026 15:27	Dubaï	22.96	43	Clear	2026-01-07 15:27:19.582 +0000
7	07/01/2026 15:27	Fès	11.14	62	Clouds	2026-01-07 15:27:19.082 +0000

3.5 Phase 5 : Visualisation avec Power BI

La dernière phase du projet consiste à créer des tableaux de bord interactifs dans Power BI, connectés directement au schéma ANALYTICS de Snowflake.

3.5.1 Connexion Power BI - Snowflake

Power BI a été connecté à Snowflake en utilisant le connecteur natif, permettant des rafraîchissements automatiques et des performances optimales grâce au mode DirectQuery.

3.5.2 Visualisations Créées

Le tableau de bord Power BI comprend 8 visualisations principales, offrant une vue complète des données étudiantes :

Visual	Type	Description
1	Graphique Circulaire	Distribution des étudiants par groupe d'âge
2	Graphique Combiné	Analyse des spécialités (MAJOR) avec tendances
3	Graphique Linéaire	Évolution par spécialité dans le temps
4	Graphique Combiné	Performance par semestre avec comparaisons
5	Graphique en Anneau	Répartition par genre et spécialité
6	Nuage de Points	Analyse de corrélation par spécialité
7	Arbre de Décomposition	Analyse hiérarchique : spécialité → âge → performance
8	Graphique à Barres	Comparaison des semestres

4. Résultats et Insights

4.1 Données Analysées

Le tableau de bord Power BI analyse les données de la table STUDENTS_POSITION du schéma ANALYTICS, incluant les dimensions suivantes :

- ***Démographie :*** Âge, genre, groupe d'âge
- ***Académique :*** Spécialité (MAJOR), semestre, niveau de performance
- ***Métriques :*** Effectifs, moyennes, tendances temporelles

4.2 Insights Clés

4.2.1 Distribution Démographique

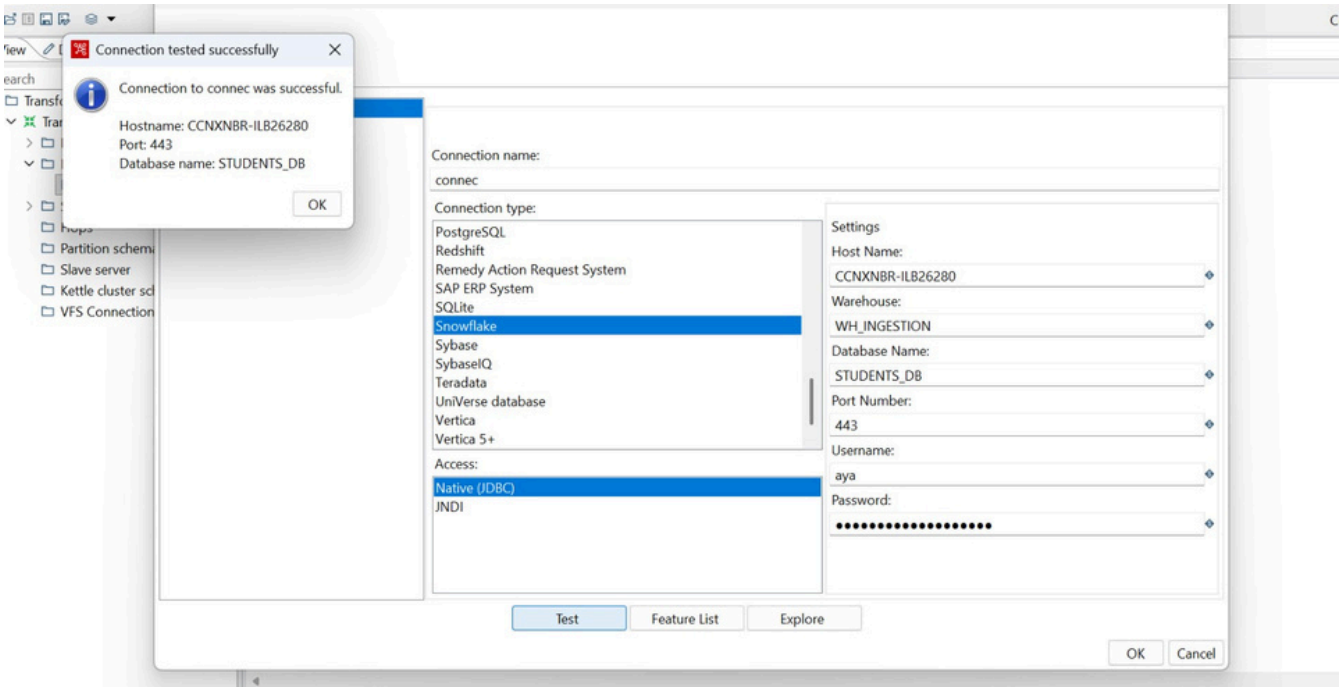
- Identification des groupes d'âge prédominants
- Analyse de la répartition hommes/femmes par spécialité

4.2.2 Performance Académique

- Comparaison des performances entre semestres
- Identification des spécialités avec les meilleurs résultats
- Corrélations entre âge, spécialité et niveau de performance

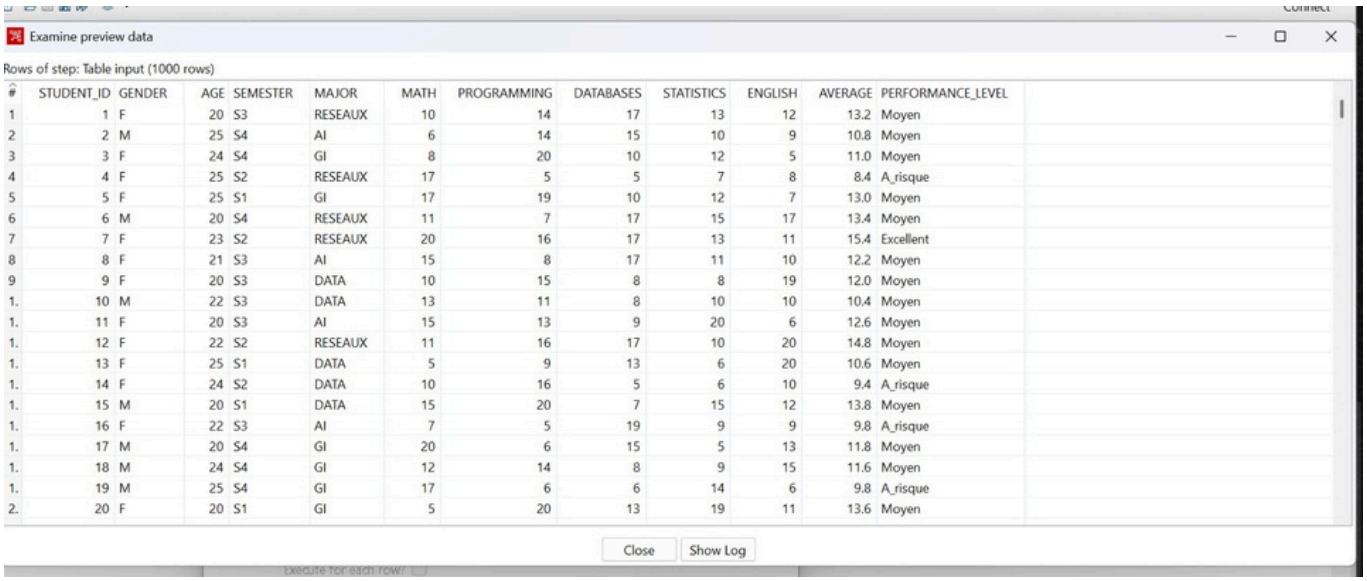
4.2.3 Tendances Temporelles

- Évolution des effectifs par spécialité
- Analyse des variations semestrielles



5. Technologies Utilisées

Technologie	Version/Type	Utilisation
PostgreSQL	SGBD Relationnel	Base de données source
pgAdmin	Outil d'administration	Gestion et administration PostgreSQL
Snowflake	Cloud Data Warehouse	Entrepôt de données (RAW + ANALYTICS)
Pentaho PDI	ETL Open Source	Transformations et intégration de données
Power BI	Plateforme BI Microsoft	Visualisation et tableaux de bord



#	STUDENT_ID	GENDER	AGE	SEMESTER	MAJOR	MATH	PROGRAMMING	DATABASES	STATISTICS	ENGLISH	AVERAGE	PERFORMANCE_LEVEL
1	1	F	20	S3	RESEAUX	10	14	17	13	12	13.2	Moyen
2	2	M	25	S4	AI	6	14	15	10	9	10.8	Moyen
3	3	F	24	S4	GI	8	20	10	12	5	11.0	Moyen
4	4	F	25	S2	RESEAUX	17	5	5	7	8	8.4	A_risque
5	5	F	25	S1	GI	17	19	10	12	7	13.0	Moyen
6	6	M	20	S4	RESEAUX	11	7	17	15	17	13.4	Moyen
7	7	F	23	S2	RESEAUX	20	16	17	13	11	15.4	Excellent
8	8	F	21	S3	AI	15	8	17	11	10	12.2	Moyen
9	9	F	20	S3	DATA	10	15	8	8	19	12.0	Moyen
1.	10	M	22	S3	DATA	13	11	8	10	10	10.4	Moyen
1.	11	F	20	S3	AI	15	13	9	20	6	12.6	Moyen
1.	12	F	22	S2	RESEAUX	11	16	17	10	20	14.8	Moyen
1.	13	F	25	S1	DATA	5	9	13	6	20	10.6	Moyen
1.	14	F	24	S2	DATA	10	16	5	6	10	9.4	A_risque
1.	15	M	20	S1	DATA	15	20	7	15	12	13.8	Moyen
1.	16	F	22	S3	AI	7	5	19	9	9	9.8	A_risque
1.	17	M	20	S4	GI	20	6	15	5	13	11.8	Moyen
1.	18	M	24	S4	GI	12	14	8	9	15	11.6	Moyen
1.	19	M	25	S4	GI	17	6	6	14	6	9.8	A_risque
2.	20	F	20	S1	GI	5	20	13	19	11	13.6	Moyen

6. Défis et Solutions

6.1 Défis Rencontrés

6.1.1 Migration PostgreSQL vers Snowflake

- ***Défi :*** Compatibilité des types de données entre PostgreSQL et Snowflake
- ***Solution :*** Mapping systématique des types de données et validation des schémas

6.1.2 Transformations Pentaho

- ***Défi :*** Gestion des données manquantes et incohérentes
- ***Solution :*** Mise en place de règles de validation et de nettoyage automatisées

6.1.3 Performance des Requêtes

- ***Défi :*** Optimisation des temps de réponse pour les visualisations Power BI
- ***Solution :*** Utilisation du mode DirectQuery et optimisation des requêtes Snowflake

6.2 Bonnes Pratiques Appliquées

- Séparation claire entre données brutes (RAW) et données analytiques (ANALYTICS)
- Documentation complète des transformations Pentaho
- Validation systématique à chaque étape du pipeline
- Tests de performanceet optimisation continue

7. Conclusion

7.1 Réalisations

Ce projet a permis de mettre en place une infrastructure complète d'intégration et d'analyse de données, démontrant la maîtrise de technologies modernes de Data Engineering et Business Intelligence. Les principales réalisations incluent :

- Migration réussie de PostgreSQL vers Snowflake avec préservation de l'intégrité des données
- Mise en œuvre de transformations ETL complexes avec Pentaho
- Création d'un entrepôt de données cloud scalable et performant
- Développement de visualisations interactives fournissant des insights actionnables

7.2 Compétences Acquises

- ***Bases de données :*** PostgreSQL, Snowflake, architecture cloud
- ***ETL :*** Pentaho Data Integration, transformations complexes
- ***Business Intelligence :*** Power BI, conception de tableaux de bord
- ***Architecture :*** Design de pipelines de données end-to-end

7.3 Perspectives d'Amélioration

Pour améliorer davantage ce projet, les axes suivants pourraient être explorés :

- ***Automatisation :*** Mise en place de pipelines automatisés avec orchestration (Airflow, etc.)
- ***Machine Learning :*** Intégration de modèles prédictifs pour anticiper les performances étudiantes
- ***Temps Réel :*** Implémentation de streaming pour des analyses en temps réel
- ***Alertes :*** Système d'alertes automatiques pour les anomalies ou tendances critiques

7.4 Conclusion Générale

Ce projet démontre une maîtrise complète du cycle de vie des données, de la source à la visualisation. L'architecture mise en place est scalable, maintenable et suit les meilleures pratiques de l'industrie. Les tableaux de bord Power BI créés fournissent des insights précieux qui peuvent guider la prise de décision dans le contexte académique.

L'utilisation combinée de PostgreSQL, Snowflake, Pentaho et Power BI illustre la capacité à orchestrer différentes technologies pour créer une solution d'analyse de données robuste et performante.

Annexes

Annexe A : Structure de la Base de Données

Base de données : STUDENT_DB

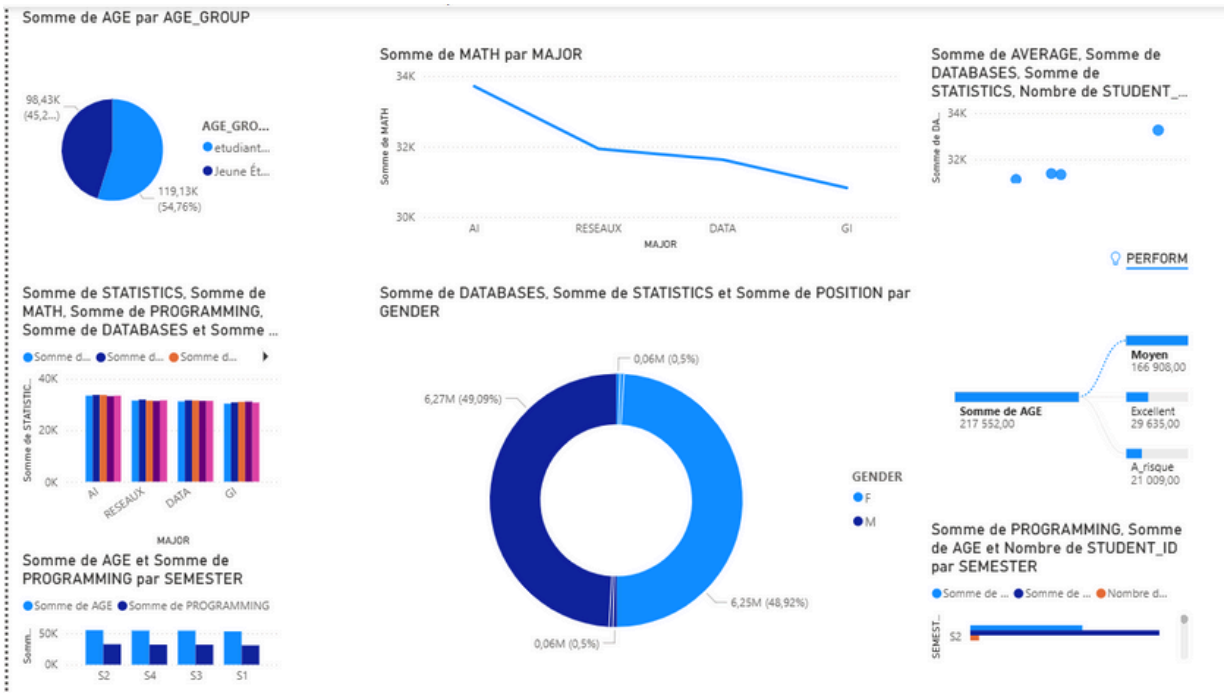
- ***Schéma RAW :*** Contient les données brutes importées de PostgreSQL
- ***Schéma ANALYTICS :*** Contient la table STUDENTS_POSITION avec données transformées

Annexe B : Champs de la Table STUDENTS_POSITION

- AGE_GROUP : Groupe d'âge des étudiants
- GENDER : Genre (M/F)
- MAJOR : Spécialité d'études
- SEMESTER : Semestre académique
- PERFORMANCE_LEVEL : Niveau de performance académique

Annexe C : Outils et Versions

- PostgreSQL : Version compatible avec pgAdmin
- Snowflake : Cloud Data Warehouse
- Pentaho Data Integration : Version Community ou Enterprise
- Microsoft Power BI : Version Desktop et/ou Service



--- Fin du Rapport ---