PROSIT : TRANSFORMERS

FODIL Nel | MARCELLI Enzo | GOUADFEL Rayan
08 avril 2024

Table des matières

I. Introduction 2
II. Analyse du contexte 2
III. Objectifs 3
IV. Problématique 3
V. Plan d'Action 4
VI. Définitions 5
ELT (Extract, Load, Transform) 10
Modèle en étoile 10
Modèle en flocon 10
Talend Open Studio pour Big Data 11
Amazon S3 11
Azure Data Lake Storage 11
ER/Studio 12
IBM Data Architect 12
SQL 13
Business Intelligence 13
PgAdmin 4 13
Conclusion 14
Webographie 15

I. Introduction

Dans le Prosit "Transformers", nous nous concentrons sur la mise en place d'une architecture de data warehouse qui intègre des données variées. Nous explorons la création d'un référentiel intermédiaire qui simplifie l'analyse des données et respecte leur confidentialité. Face au défi de modéliser et d'intégrer ces données en respectant les contraintes de temps du projet, l'accent est mis sur l'utilisation efficace de l'ETL pour préparer le premier livrable crucial.

II. Analyse du contexte

Le contexte du projet "Transformers" nous situe à l'intersection de l'analyse de données avancée et de la gestion stratégique des informations dans un environnement Big Data. Nous sommes confrontés à la tâche complexe d'harmoniser des données de provenances et de formats divers, ce qui nécessite une infrastructure capable de les traiter sans égard à leur hétérogénéité initiale. L'architecture envisagée par Archie promet un stockage unifié, levant les barrières structurelles et permettant une analyse flexible et sécurisée.

Cependant, cette unification des données passe par une série de transformations, où la place et le processus de l'ETL sont questionnés. Faut-il transformer avant ou après le chargement dans le référentiel ? Cette interrogation technique masque un enjeu plus vaste : assurer la cohérence et la qualité des données dans le cadre d'analyses décisionnelles pertinentes.

Dans un calendrier projet serré, ces considérations théoriques doivent rapidement laisser place à des décisions pratiques et à l'action pour produire un livrable conforme aux attentes des utilisateurs et aux standards du Big Data. La pression est d'autant plus forte que l'architecture à réaliser doit être non seulement fonctionnelle mais aussi évolutive, pour s'adapter aux changements rapides et imprévisibles des besoins en données.

III. Objectifs

L'objectif principal du Prosit "Transformers" est de concevoir un système d'entreposage de données capable de traiter et d'analyser des sources d'informations variées. Les objectifs spécifiques comprennent :

- **Intégrer des Données Hétérogènes** : Fusionner des données structurées, semi-structurées et non structurées dans un référentiel unique pour faciliter l'accès et l'analyse.
- **Assurer la Confidentialité** : Protéger les données tout au long du processus d'intégration et d'analyse.
- **Modélisation des Données Cibles** : Développer un modèle de données multidimensionnel qui s'aligne avec les besoins d'analyse des utilisateurs.
- **Optimiser le Processus ETL :** Déterminer le placement le plus efficace pour les transformations de données - avant ou après le chargement dans l'entrepôt.
- **Respecter les Délais** : Produire le livrable dans un délai serré tout en assurant la qualité et la performance de l'entrepôt de données.
- **Gérer les Incohérences** : Identifier et corriger les erreurs, doublons et enregistrements vides dans les données source.
- **Faciliter les Analyses Décisionnelles** : S'assurer que l'entrepôt de données final est bien adapté pour soutenir des décisions commerciales informées.

IV. Problématique

Quelle est la meilleure stratégie pour modéliser et intégrer rapidement des données hétérogènes dans un entrepôt de données tout en facilitant les analyses décisionnelles complexes ?

V. Plan d’Action

Pour répondre aux objectifs du Prosit "Transformers", le plan d'action suivant est proposé :

- **Établissement d'une Plateforme d'Intégration** : Utiliser Talend Open Studio pour Big Data comme outil principal pour l'extraction, la transformation et le chargement (ETL) des données. Cela facilitera l'importation de données hétérogènes et leur préparation pour l'analyse.
- **Construction d'un Data Lake** : Créer un data lake comme référentiel intermédiaire pour stocker toutes les données collectées, quelle que soit leur structure. Cela peut se faire en utilisant des technologies telles que Hadoop ou une solution cloud comme Amazon S3 ou Azure Data Lake Storage.
- **Modélisation Multidimensionnelle** : Développer un modèle de données en étoile ou en flocon en se servant de logiciels de modélisation de données comme ER/Studio ou IBM Data Architect pour capturer les dimensions et mesures nécessaires aux analyses.
- **Normalisation des Données** : Assurer que les formats de données, en particulier les dates et les identifiants uniques, soient normalisés pour permettre une comparaison et une analyse cohérentes. Des outils tels que Talend ou des fonctions SQL personnalisées peuvent être utilisés pour ce traitement.
- **Assurer la Qualité des Données** : Implémenter des contrôles de qualité de données pour détecter et corriger les erreurs et les doublons, en utilisant des composants dédiés dans Talend ou des scripts Python pour le nettoyage des données.
- **Automatisation des Processus** : Mettre en place des jobs Talend automatisés pour l'extraction et la transformation des données, garantissant une intégration régulière et fiable dans l'entrepôt de données.
- **Conception de Tableaux de Bord** : Utiliser des outils de business intelligence comme Power BI pour créer des visualisations interactives et des tableaux de bord qui permettront d'exploiter les données de l'entrepôt pour des insights actionnables.

VI. Dimensions et Tables

Décès

- Table Deces

Etablissement

- Table Etablissement_Sante
    - Table Salle (+ clé étrangère consultation)
    - Table Laboratoire

Hospitalisation

- Table Hospitalisation (+ clé étrangère patient + clé étrangère etablissement_sante)
- Table Consultation
    - Table Mutuelle
    - Table Diagnostic
    - Table Prescription (+ clé étrangère Consultation)
        - Table Médicaments
    - Table Patients
    - Table Professionnels de Sante (rajouter colonnes communes de l'excel)
        - Table Activités_Professionnel_Sante
        - Table Specialites

Satisfaction_2014

- Table DPA_SSR _table_es_2014
- Table DPA_SSR _table_lexique_2014
- Table DPA_SSR _table_participant_2014
- Table RCP_MCO _table_es_2014
- Table RCP_MCO_table_lexique_2014
- Table RCP_MCO_table_participant_2014

Satisfaction_2015

- Table hpp_mco _lexique_2015
- Table hpp_mco_tables_es_2015
- Table idm_mco _tables_es_2015

Satisfaction_2016 :

- Table dan_mco donnees_2016
- Table dan_mco _lexique_2016
- Table dpa_had donnees_2016
- Table dpa_had lexique_2016

Satisfaction_2017_2018 :

- Table dpa-ssr- donnees_2017_2018
- Table dpa-ssr- lexique_2017_2018
- Table ete-ortho-ipaqss-donnees_2017_2018
- Table ete-ortho-ipaqss-lexique_2017_2018
- Table rcp-mco -donnees_2017_2018
- Table rcp-mco -lexique_2017_2018
- Table ESATIS48H_MCO _donnees_2017_2018
- Table ESATIS48H_MCO_ lexique_2017_2018

Satisfaction_2019 :

- Table lexique-esatis48h
- Table lexique-esatisca
- Table lexique-iqss
- Table resultats-esatis48h
- Table resultats-esatisca
- Table resultats-iqss

Satisfaction_2020 :

- Table lexique-esatis48h_2020
- Table lexique-esatisca_2020
- Table lexique-iqss_2020
- Table resultats-esatis48h_2020
- Table resultats-esatisca _2020
- Table resultats-iqss_2020

Listes fichiers dossiers Satisfaction :

Fichiers 2014 :

- DPA_SSR _table_es
- DPA_SSR _table_lexique
- DPA_SSR _table_participant
- RCP_MCO _table_es
- RCP_MCO_table_lexique
- RCP_MCO_table_participant

Fichiers 2015 :

- hpp_mco _lexique
- hpp_mco_tables_es
- idm_mco _tables_es

Fichiers 2016 :

- dan_mco donnees
- dan_mco _lexique
- dpa_had donnees
- dpa_had lexique

Fichiers 2017/2018 :

- dpa-ssr- donnees
- dpa-ssr- lexique
- ete-ortho-ipaqss-donnees
- ete-ortho-ipaqss-lexique
- rcp-mco -donnees
- rcp-mco -lexique
- ESATIS48H_MCO _donnees
- ESATIS48H_MCO_ lexique

Fichiers 2019 :

- lexique-esatis48h
- lexique-esatisca
- lexique-iqss
- resultats-esatis48h
- resultats-esatisca
- resultats-iqss

Fichiers 2020 :

- lexique-esatis48h
- lexique-esatisca
- lexique-iqss
- resultats-esatis48h
- resultats-esatisca
- resultats-iqss

ESATIS -> 2019=2020

ESATIS48 (résultats)-> 2017=2019=2020

VII. Définitions

ELT (Extract, Load, Transform)

ELT est une approche de traitement des données où les données sont d'abord extraites de leurs sources originales et chargées directement dans une cible de stockage, comme un data lake ou un data warehouse, sans transformation préalable. La transformation des données s'effectue après leur chargement dans la cible de stockage. Cette méthode est souvent privilégiée dans les environnements Big Data en raison de la capacité de stockage et de traitement massifs offerts par des technologies telles que Hadoop et les plateformes de cloud computing. ELT permet une plus grande flexibilité dans le traitement des données et est bien adapté pour gérer de grands volumes de données en réduisant le temps nécessaire à leur préparation avant analyse.

Modèle en étoile

Le modèle en étoile est une architecture de base de données utilisée pour les entrepôts de données et les applications de Business Intelligence. Il est ainsi nommé en raison de sa structure qui ressemble à une étoile avec une table de faits au centre, entourée de tables de dimensions. La table de faits stocke les données transactionnelles ou événementielles quantitatives, telles que les ventes ou les transactions, avec des clés étrangères référençant les tables de dimensions. Les tables de dimensions contiennent des données descriptives sur les aspects des transactions, comme les clients, les produits, le temps, et d'autres catégories d'analyse. Ce modèle est conçu pour optimiser les requêtes analytiques en simplifiant les relations entre les données, rendant les requêtes plus rapides et plus intuitives pour l'utilisateur.

Modèle en flocon

Le modèle en flocon, ou schéma en flocon de neige, est une variation du modèle en étoile pour la modélisation des entrepôts de données. Dans ce modèle, les tables de dimensions sont normalisées, c’est-à-dire décomposées en sous-tables pour réduire la redondance des données. Cette normalisation se traduit par une structure qui ressemble à un flocon de neige, où la table de faits est entourée par des dimensions qui peuvent elles-mêmes être reliées à d'autres tables de dimensions de niveau inférieur. Bien que le modèle en flocon puisse conduire à des structures de base de données plus complexes et à des requêtes potentiellement plus lentes en raison du nombre accru de jointures, il offre des avantages en termes d'économie d'espace de stockage et de maintien de l'intégrité des données grâce à une meilleure normalisation.

Talend Open Studio pour Big Data

Talend Open Studio pour Big Data est un outil d'intégration de données open-source qui permet aux utilisateurs de concevoir, de développer et de déployer des tâches d'intégration de données et des pipelines ETL (Extract, Transform, Load) pour les environnements Big Data. Il offre une interface graphique qui simplifie la création de jobs d'intégration de données, en permettant aux utilisateurs de glisser-déposer des composants et de configurer leurs propriétés sans nécessiter une programmation complexe. Talend supporte une large variété de sources de données, y compris les bases de données relationnelles, les fichiers plats, les systèmes de fichiers distribués comme HDFS, et les plateformes de données cloud. Il est particulièrement apprécié pour sa capacité à gérer de grands volumes de données de manière efficace, en facilitant le travail avec des technologies de Big Data comme Hadoop, Spark, et Apache Hive.

Amazon S3

Amazon Simple Storage Service (S3) est un service de stockage d'objets offert par Amazon Web Services (AWS) conçu pour stocker et récupérer n'importe quelle quantité de données, à tout moment et de n'importe où sur le web. Il fournit une interface de service web simple qui peut être utilisée pour stocker et récupérer n'importe quelle quantité de données, à tout moment, à partir de n'importe où sur le web. Amazon S3 est largement utilisé pour la sauvegarde et la récupération, l'archivage, les applications d'entreprise, les sites web IoT, et les applications mobiles, et est particulièrement populaire pour le stockage de grands volumes de données non structurées, tels que des vidéos, des photos et des fichiers de log. Il est conçu pour offrir une durabilité de 99.999999999% et stocke les données pour des millions d'applications pour des entreprises du monde entier.

Azure Data Lake Storage

Azure Data Lake Storage est une solution de stockage de données massivement scalable et sécurisée sur le cloud de Microsoft Azure. Conçu spécifiquement pour les besoins du Big Data analytics, Azure Data Lake Storage combine les capacités de stockage d'objets de grande capacité avec les fonctionnalités de fichiers hiérarchiques, rendant ainsi le stockage et l'analyse de données de différents formats et tailles plus efficaces. Il est optimisé pour les charges de travail analytiques, telles que le traitement parallèle massif, l'analyse de données big data et l'apprentissage automatique, offrant une intégration transparente avec les services d'analyse Azure tels que Azure HDInsight, Azure Databricks et Azure Synapse Analytics. Azure Data Lake Storage assure une sécurité des données de haut niveau, supporte les politiques de rétention des données, et fournit une scalabilité sans limite pour s'adapter aux exigences croissantes de stockage et de traitement des données.

ER/Studio

ER/Studio est un outil puissant pour la modélisation des données qui aide les organisations à gérer et à optimiser leur architecture de données. Il permet aux architectes de données, aux concepteurs et aux parties prenantes de l'entreprise de collaborer efficacement en créant des modèles de données logiques et physiques. ER/Studio supporte une large variété de plateformes de bases de données, y compris SQL Server, Oracle, MySQL, et PostgreSQL, facilitant ainsi la conception, la visualisation, et la documentation des structures de données complexes.

Cet outil propose des fonctionnalités avancées telles que la modélisation des données d'entreprise, la gestion des dictionnaires de données, la génération de schémas de bases de données, et l'analyse d'impact, permettant aux entreprises de mieux comprendre leurs actifs de données et de garantir la cohérence et la qualité des données à travers l'organisation. ER/Studio se distingue par sa capacité à aligner les modèles de données avec les objectifs d'affaires, en fournissant une vue globale des données qui facilite la prise de décision basée sur des informations précises et à jour.

IBM Data Architect

IBM Data Architect est une solution avancée de modélisation de données qui permet aux utilisateurs de concevoir, de développer et de gérer des architectures de données sophistiquées pour leurs environnements d'entreprise. Cet outil fait partie de la famille de produits IBM Data Studio et offre une plateforme intégrée pour la modélisation relationnelle, dimensionnelle, et NoSQL, supportant une variété de bases de données comme IBM DB2, Oracle, Teradata, et d'autres.

IBM Data Architect aide les organisations à comprendre et à visualiser la structure de leurs données, les relations entre les données, et à créer des modèles logiques et physiques qui peuvent être déployés directement dans les systèmes de gestion de bases de données. Il propose des fonctionnalités telles que la comparaison et la synchronisation de modèles, la génération de DDL (Data Definition Language), et la documentation automatique de bases de données, facilitant ainsi la gestion du cycle de vie des données et l'alignement des initiatives de données avec les besoins métier.

En mettant l'accent sur la collaboration et la gouvernance des données, IBM Data Architect permet aux équipes multidisciplinaires de travailler ensemble de manière efficace pour concevoir des solutions de données qui soutiennent les objectifs stratégiques et opérationnels de l'entreprise.

SQL

SQL (Structured Query Language) est un langage de programmation utilisé pour communiquer avec les bases de données relationnelles. Il permet de manipuler les données en effectuant des requêtes pour récupérer, insérer, mettre à jour ou supprimer des données dans une base de données.

Business Intelligence

La Business Intelligence (BI) désigne l'ensemble des technologies, des processus et des outils permettant de collecter, d'analyser et de présenter des données afin d'aider les entreprises à prendre des décisions éclairées. La BI implique souvent l'utilisation de logiciels et de systèmes pour transformer les données brutes en informations exploitables.

PgAdmin 4

PgAdmin 4 est une interface graphique open source utilisée pour administrer et gérer les bases de données PostgreSQL. Il permet aux administrateurs de bases de données de gérer les utilisateurs, les schémas, les tables, les requêtes SQL, ainsi que d'autres aspects de la base de données via une interface conviviale.

Conclusion

En conclusion, le Prosit "Transformers" nous guide à travers le processus complexe de création d'un système d'entreposage de données robuste et évolutif, capable de gérer des données hétérogènes dans un environnement Big Data. L'utilisation stratégique d'outils comme Talend Open Studio pour Big Data et la mise en place d'un data lake offrent une fondation solide pour l'intégration, la transformation, et l'analyse des données. En adoptant une approche méthodique pour la modélisation multidimensionnelle et en mettant l'accent sur la qualité des données, nous pouvons surmonter les défis posés par les sources d'informations diversifiées. Ce projet illustre l'importance de la flexibilité, de la précision, et de la sécurité dans le domaine de l'analyse de données, tout en soulignant la nécessité d'une planification rigoureuse et d'une exécution efficace pour atteindre les objectifs dans des délais serrés. Les enseignements tirés et les solutions mises en œuvre ici serviront de référence pour des initiatives futures dans l'espace dynamique du Big Data.

Webographie

https://www.talend.com/fr/resources/elt-vs-etl/

https://www.ediservices.com/fr/etl-integration/

https://www.talend.com/fr/resources/guide-data-lake/

https://www.cartelis.com/blog/data-lake-vs-data-warehouse/

https://openclassrooms.com/fr/courses/4467481-creez-votre-data-lake/4467488-identifiez-les-besoins-de-votre-data-lake

https://www.data-transitionnumerique.com/cube-olap-decisionnel-big-data/

https://www.data-transitionnumerique.com/cube-olap-decisionnel-big-data/

https://moodle.cesi.fr/pluginfile.php/24914/mod_resource/content/4/res/Informatique_Desicionnelle.pdf

https://www.edureka.co/blog/videos/etl-using-big-data-talend/

https://www.edureka.co/blog/talend-big-data-tutorial/