CONSTRUCTION D'UN PIPELINE DE TRAITEMENT DE DONNÉES POUR LES DONNÉES DE DIFFÉRENTS FORMATS (FARKLI FORMATTAKİ VERİLER İÇİN BİR VERİ İŞLEME BORU HATTI OLUŞTURMA)

Doğa Yağmur Yılmaz, Dr. Sultan N. Turhan

Département de Génie Informatique
Faculté d'Ingénierie et de Technologie, Université Galatasaray

SUJET ET OBJECTIF

L'objectif du projet est de réaliser l'intégration des données en créant un pipeline de traitement des données pour des données de différents formats qui proviennent de différentes sources.

Il est créé une base de données orientée colonnes d'académiciens, sur laquelle des requêtes peuvent être exécutées, en extrayant et en intégrant les données des académiciens à partir de deux sources différentes.

Les informations "nom, prénom, ORCID, ID de chercheur, titre, université, position de cadre, domaine principal, domaine scientifique, spécialités et e-mail" des académiciens sont collectées à partir de la plateforme "Yök Académique". Ensuite, les articles et spécialités précis des académiciens sont extraits par "Google Scholar" pour réduire leurs domaines d'expertise à des titres plus spécifiques.

VALEUR ORIGINALE

Le but donc est de faciliter l'accès à des données précises et fiables sur les académiciens en Turquie, ce qui pourrait être utile pour les recherches et les analyses. Le traitement par lots permettra également de traiter un grand nombre de données rapidement et efficacement. Ce pipeline facilite donc la collecte, le stockage et la gestion de données académiques en Turquie.

METHODES ET TECHNOLOGIES

Un code de scrape a été écrit pour interagir avec la page dynamique et les données de «Yök Académique» ont été extraites sous forme de tableau html. Les données sont été converties et écrites dans un seul fichier de «csv».

À l'aide ded librairies scrapy et scholarly, les données ont été extraites de «Google Scholar» et collectées dans un fichier de «json».
Le code qui intègre des données sous différents formats a été écrit.

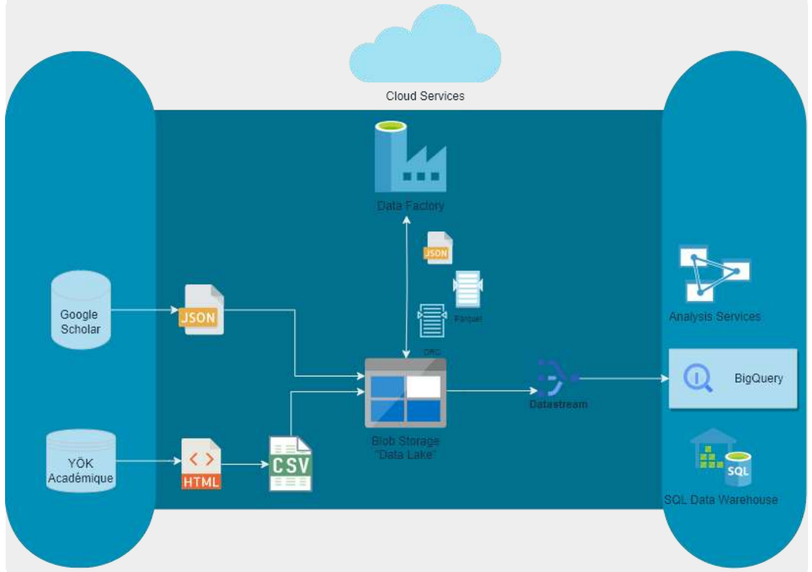
Une zone de stockage où des données de différents formats provenant de différentes sources sont stockées ensemble dans leur forme originale.

Compte tenu de l'évolutivité, du coût et de l'accessibilité, l'architecture en nuage a été préférée.

Une grande quantité de données est traitée en blocs et en une seule fois. flux de travail composé d'une série de tâches interconnectées.

Google BigQuery; un entrepôt de données qui aide à gérer et à analyser des données,
peut lire des données provenant de sources externes,
conçu pour les plus grands ensembles de données,
permet d'éxecuter des requêtes rapide, de type SQL.

ARCHITECTURE DU MODELE



Google BigQuery / Requêtes sur Data

Query results

SAVE RESULTS | EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH | PREVIEW

Row	Temel Alan	Kadro Yeri	Ad Soyad	E-Posta
54	Sosyal-Beşeri ve İdari Bilimler T...	GALATASARAY ÜNİVERSİTESİ/İLETİŞİM FAKÜLTESİ/İLETİŞİM BÖLÜMÜ/İLETİŞİM BİLİMLERİ ANABİLİM DALI/	AYŞE TOY PAR	atoy[at]gsu.edu.tr
55	Sosyal-Beşeri ve İdari Bilimler T...	GALATASARAY ÜNİVERSİTESİ/FEN-EDEBİYAT FAKÜLTESİ/FELSEFE	SELAMİ ATAKAN ALTINÖRS	aaltinors[at]gsu.edu.tr

Row	num_citations	publications.bib.citation	pu... pub_year	publications.bib.title	publications.source
54	9		2009	El kapılarında yeşilçam: 1970-1990 arası Türkiye'de dış göç-sinema ilişkisi	AUTHOR_PUBLICATION_ENTRY
55	38	Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi 1 (28), 389-401, 2010	2010	Düşünce ile dil arasındaki ilişki...	AUTHOR_PUBLICATION_ENTRY

Problemes Rencontrés & Les Solutions

1. Lors de la Scrape, Google Scholar bloque l'adresse IP qui demande à plusieurs reprises l'accès à la page. Pour cette raison, les données de google scholar ont été retirées par scraping avec plusieurs machines différentes à des intervalles de temps.
2. Étant donné que le projet a été mené sur des comptes d'étudiants en nuage, toutes les données retirées n'ont pas pu être exploitées. Les données des académiciens de l'université Galatasaray ont donc été conservées séparément et l'étude a été réalisée sur ces données.
3. Comme la connexion Spark ne pouvait pas être realisée par Azure, d'autres solutions cloud capables d'extraire des données en format ORC d'un stockage blob et de fournir une connexion Spark ont été recherchées. Dans cette direction, la solution cloud de google a été mise en place pour travailler avec le modèle de flux.
4. Comme la connexion Spark ne pouvait pas être realisé, la mise en correspondance des données a été assurée par un script de python et la conversion de format des données a été assurée à l'aide d'Azure data factory dans un seul fichier.
5. Étant donné que les données au format ORC écrites dans le blob peuvent être extraites directement vers le nuage Google et utilisées dans les opérations BigQuery, il n'est pas nécessaire de disposer d'une connexion spark en cas d'utilisation de BigQuery.
6. La corruption et la perte de données s'étant produites lors de la conversion du format de données en parquet via data factory, ORC, qui est un autre format en colonnes sur lequel des requêtes peuvent être exécutées, a été préféré.

CONCLUSION

Deux ensembles de données sous différents formats provenant de différentes sources sont stockés dans le stockage Blob qui est une architecture de type "lac de données" . Puis elles ont été intégrées par des infos «université et prénom-nom» à l'aide d'un code de scrape et converties au format ORC (colonné) via Data Factory avec succès.

Les données ont été requêtées en les coulant au Google BigQuery. La requête a permis d'obtenir des informations sur les académiciens et leurs articles sous forme de tableaux.

En raison de contraintes liées au coût des systèmes en nuage, seule la partie de l'université Galatasaray des données extraites a été utilisée.

A la fin d'étude, le base de données contient 137 académiciens et leurs articles.