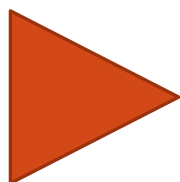
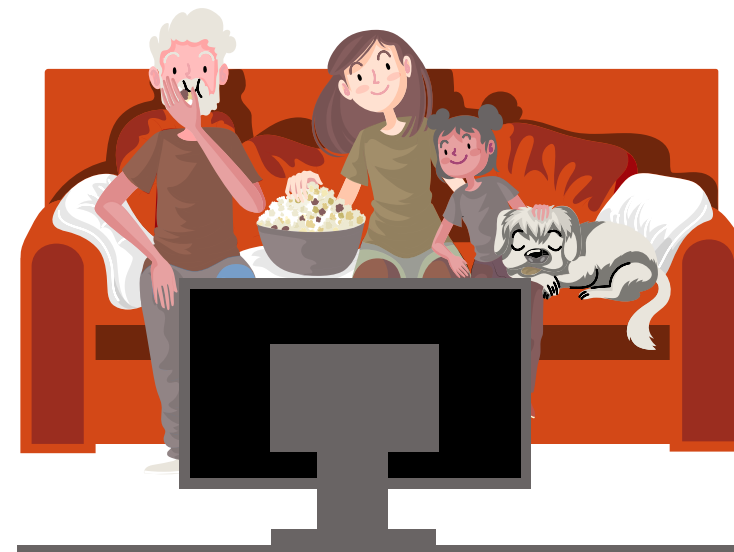


Prototype de système de recommandation de films



PAUL DESCHLIDRE, VANICK DJOFANG DJAMEN,
MOHAMED EL AMINE MEGUANANI, EMMA OLLIVIER
KENIA PINEDA MESA, BAPTISTE VIERA

Sommaire

- Introduction
- Présentation des données et du modèle de données
- Exploitation des données et tableau de bord
- Entrepôt de données
- Interface utilisateur et recommandations
- Architecture du prototype
- Perspectives
- Sources et références
- Démonstration



Les systèmes de recommandations

- Largement répandus dans ce domaine et indispensables pour répondre aux besoins d'affaire
- Font face à de nombreux défis (performance, éthique) (Schedl *et al.*) (Milano *et al.*)
- Basés fréquemment sur l'intelligence artificielle (Roy et Dutta)
- Nécessite une grande quantité de données (Roh *et al.*)



Les sources des données

MovieLens

The Movie
Database

Blockbuster
Database

Rounak Banik, The movie Dataset, Kaggle www.kaggle.com/datasets/rounakbanik/the-movies-dataset

CrowsFlower, Blockbuster Database, Data world, <https://data.world/crowdfower/blockbuster-database>

Les informations fournis sur les données

Informations générales sur les films
(titre, poster, date de sortie, pays, genres)



Acteurs et équipe de tournage



Mots clefs associés



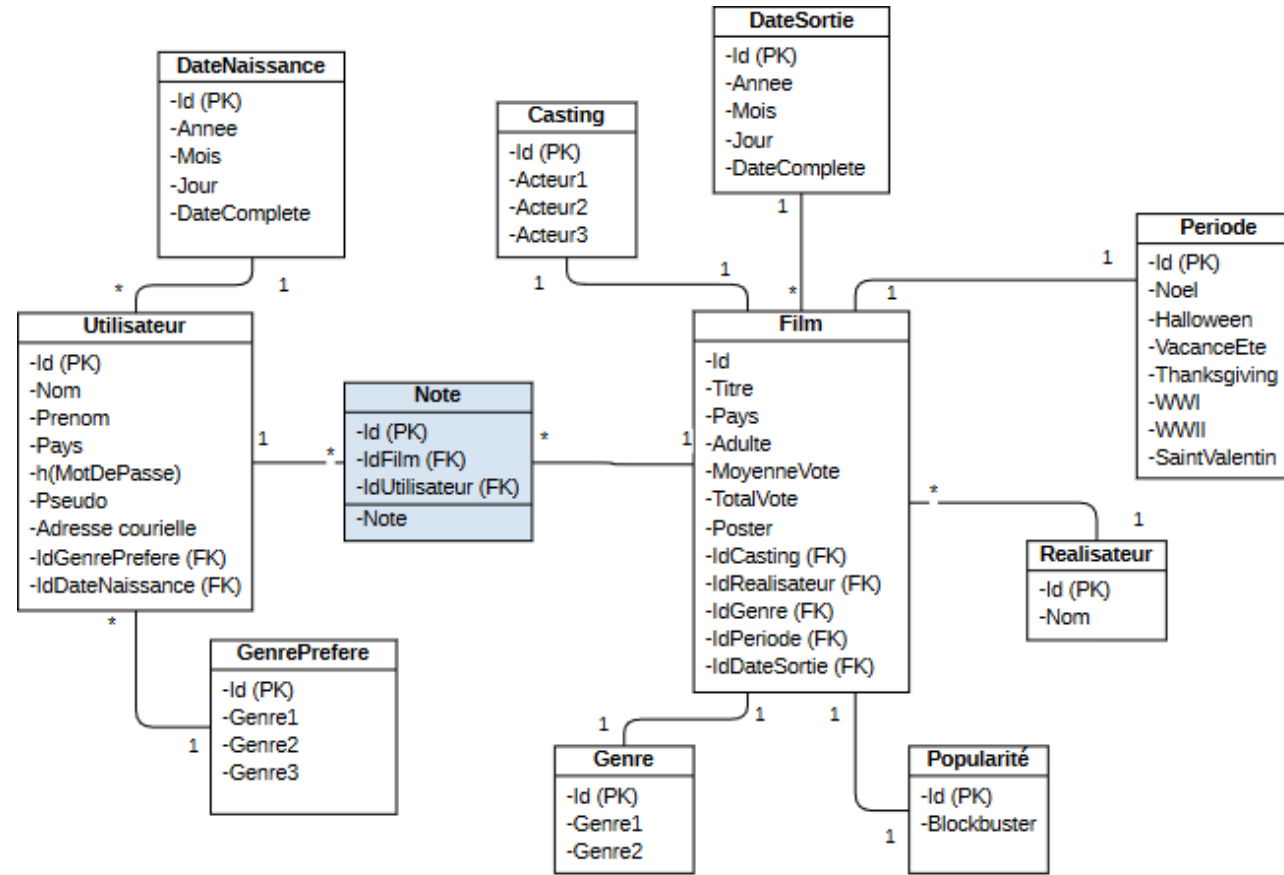
Notes



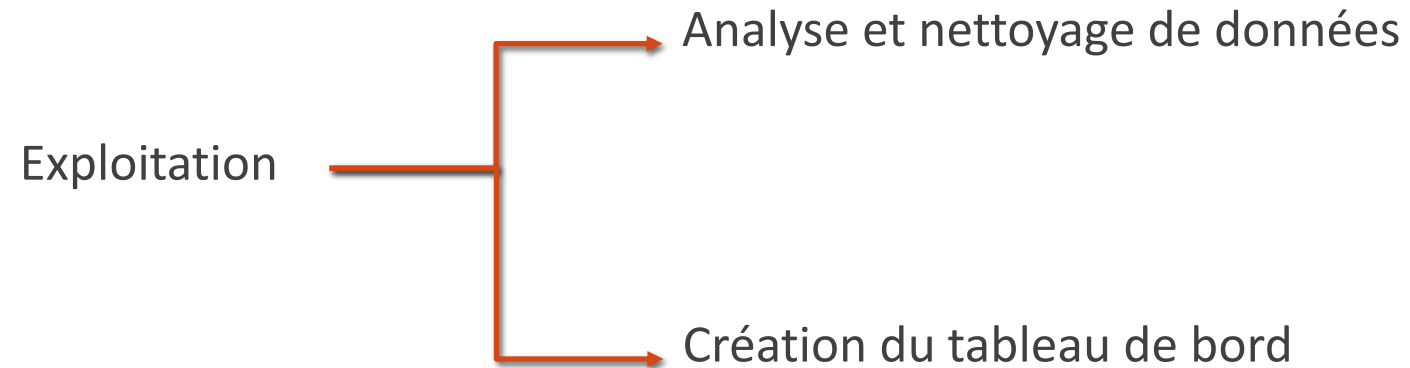
Popularité



Le modèle de données



Exploitation de données



Analyse et nettoyage de données

Analyse et détection des anomalies

`'read_csv'`

`"info()"`

`"dtype"`

`"duplicated()"`

`"groupby()"`

`"sort_values()"`

`"describe()"`

`"isnull()"`

`"unique()"`

`"count() "`



Nettoyage des données

`drop_duplicates()`

`drop_duplicates(subset=['Title', 'Pays'], keep='last')`

`drop_duplicates(subset=['Id'])`

`dropna(subset=['Id'])`

`astype('int64')`



Création du tableau de bord

Transformation

projet - Éditeur Power Query

Fichier Accueil Transformer Ajouter une colonne Affichage Outils Aide

Fermer & appliquer Nouvelle source Sources récentes Entrer des données Paramètres de la source de données Gérer les paramètres Actualiser l'aperçu Propriétés Éditeur avancé Gérer Choisir les colonnes Supprimer les colonnes Conserver les lignes Supprimer les lignes Fractionner la colonne Regrouper par Type de données : Nombre entier Utiliser la première ligne pour les en-têtes Remplacer les valeurs Transformer

Requêtes [8] film_cl casting_cl date_cl genre_cl periode_cl popularite_cl realisateur_cl note_cl

= Table.TransformColumnTypes(#"Valeur remplacée",{{"MoyenneVote", type number}})

	A ⁰ c Pays	A ⁰ c Adult	1.2 MoyenneVote	1.2 TotalVote	A ⁰ c Poster_left
1	volution	FAUX	4,3	2	/hfkqKt23CvBxHSNsCeLM6b
2	camp	US	FAUX	3,4	23 /hcNZF3UzShQX8xFjNjBTE9
3		FAUX	7	1	/yokrKMyTx81PRkh8tnHUBz
4		US	FAUX	3,4	53 /48zZhMQ7HABuHrgQL1dD
5	black	IT	FAUX	6	2 /eoe3t99Djfh5xsgECerDju5c
6	ngo	IT	FAUX	5	1 /6vC9ttnGacblzzvkUWgd4t9
7		US	FAUX	6	24 /lwbZwoy0571LualFyTNHyG
8	girl: a love story	US	FAUX	6,3	11 /i4AxBeJDzqvaRhWsP8guiOI
9		AU	FAUX	6	28 /g3wmvVxwISW2BcbZh94ks
10		US	FAUX	5,5	9 /8aA51sgy3q8ocnlCU9Zc55C
11		FAUX	0	0	/2qe7F8gcJWJGNia70t7pi1T
12		GB	FAUX	6,7	414 /b8dmfG84peFdouN2N8wO
13		IN	FAUX	5,8	3 /IDL3RikOCwbH6qEUtVHr7F
14		US	FAUX	6,5	2 /1cDs39BQM8DGe5YIISr5F4
15	... and the boys		FAUX	6,3	3 /81cHqZpcxvWB3ZODTKYml
16	ona skies	US	FAUX	3,7	9 /rfxURSqmXFjikmedWMWl
17		US	FAUX	5,8	6 /8bpgDDGx9RK6dDBBEcBET
18		FR	FAUX	5,3	3 /pyEIYMsxPlp0tb0nxFWnjzl
19	us part	JP	FAUX	6,8	6 /lu6GKNol4Q0kqitWbEArt1I
20	for love		FAUX	6	11 /rcFF0JMG7mAYKHelO3EbL
21		US	FAUX	5,1	4 /4cbQ1TwPRgpUHmykrVOY
22	before christmas	US	FAUX	5,9	11 /i5qPFBSeKjNbFPI1CseOUSN
23	ummer	US	FAUX	7,2	2993 /55jtNPD1bb182vzQccvEU
24					

14 COLONNES, 999+ LIGNES Profilage de la colonne en fonction des 1000 premières lignes

APERÇU TÉLÉCHARGÉ À 15:21

Paramètres d'une requête

PROPRIÉTÉS

Nom

film_cl

Toutes les propriétés

ÉTAPES APPLIQUÉES

Source

En-têtes promus

Type modifié

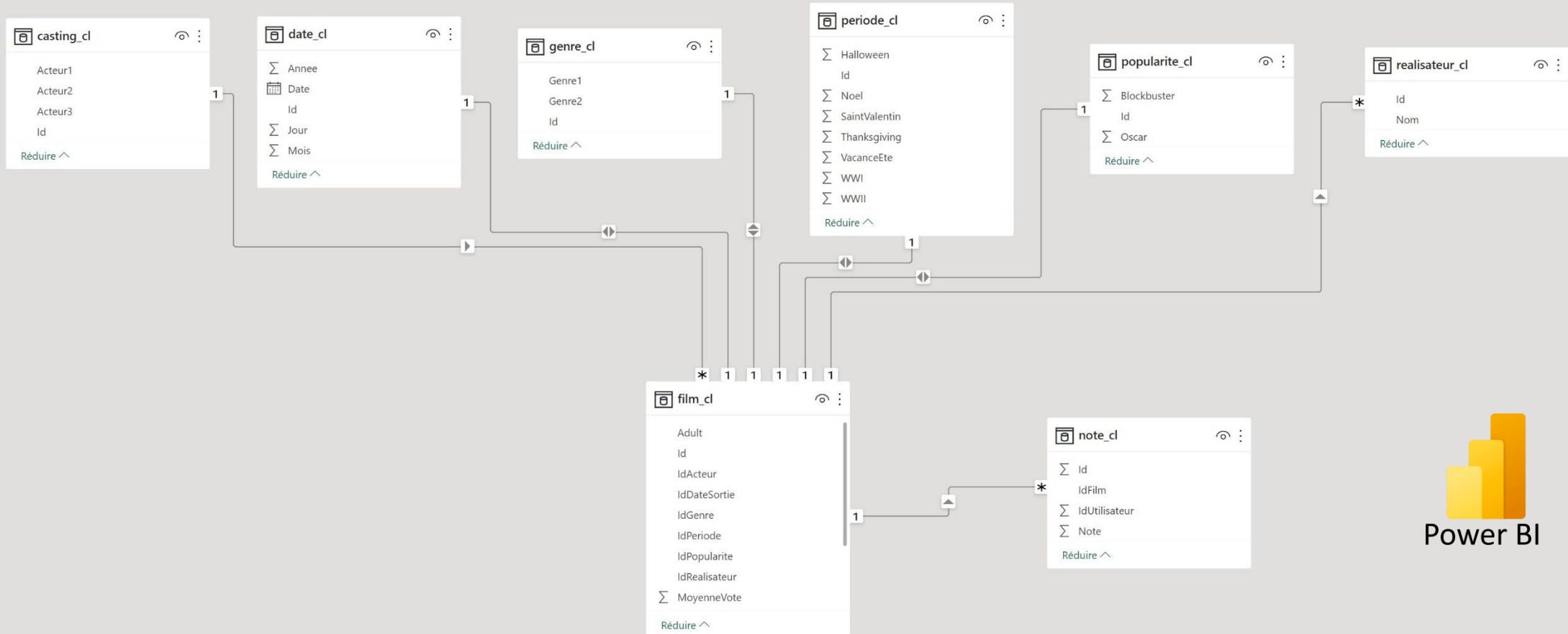
Valeur remplacée

Type modifié1



Création du tableau de bord

Modèle



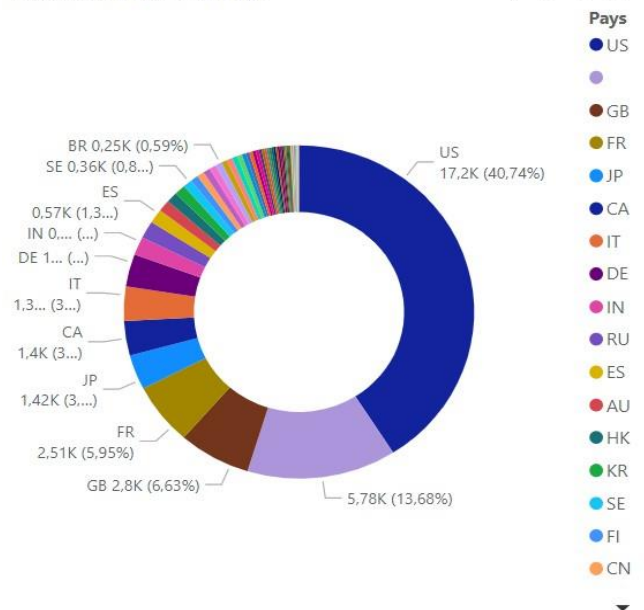
Création du tableau de bord

Visualisation

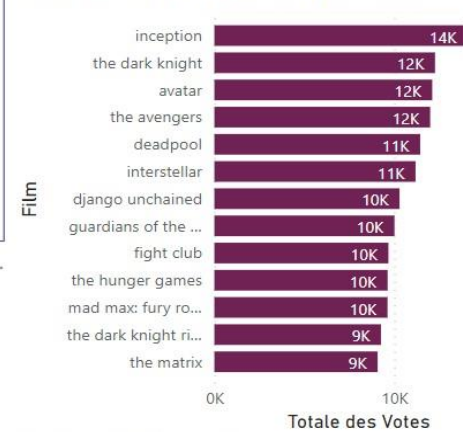


42227	4902160	5,63
Nombre de Film	Total Vote	Moyenne de Vote
439	257	828
Halloween	Noel	Saint Valentin
24	127	394
Thanks giving	Vacance	WWII

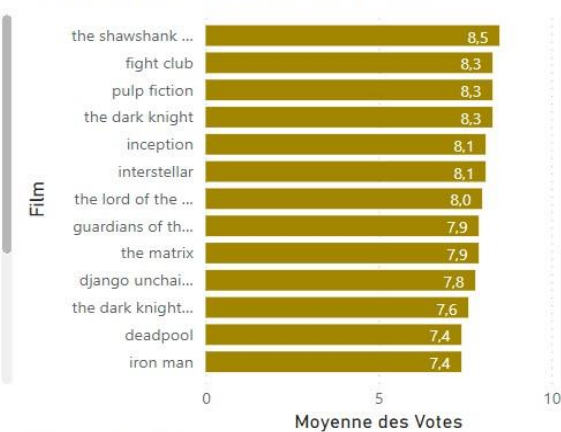
Nombre de Film par Pays



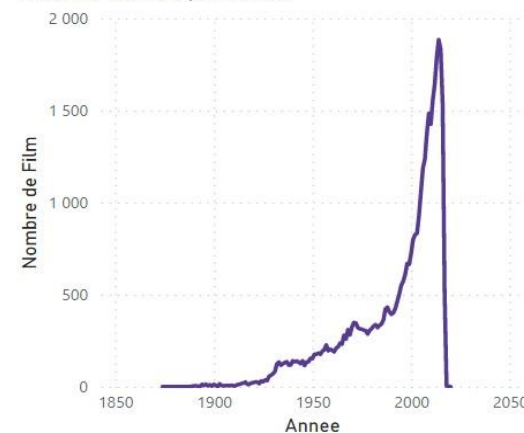
Classement des 20 films les plus notes



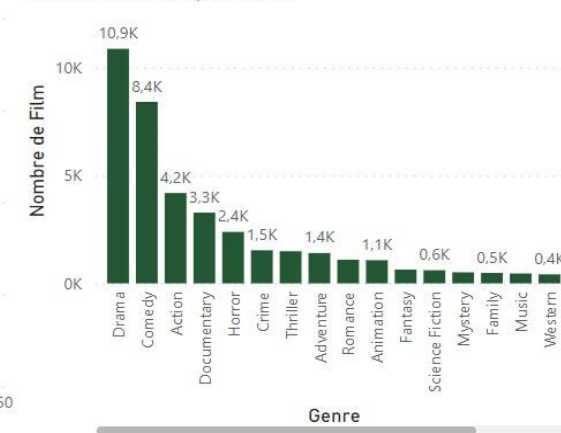
Classement des 20 films les mieux notes



Nombre de Film par Année



Nombre de Film par Genre



Création du tableau de bord

Partage

Power BI

Mon espace de travail

Essai : 59 jours

Accueil

Créer

Parcourir

Centre de données

Métriques

Applications

Pipelines de déploiement

Apprenez

Espaces de travail

Mon espace de travail

Mon espace de travail

+ Nouveau

Charger

Tout

Contenu

Jeux de données + flux de données

Nom	Type	Propriétaire	Actualisé	Prochaine actualisation	Promotion	Confidentialité
projet	Rapport	Mohamed El Amine Meguenani	02/04/23 23:27:45	—	—	—
projet	Jeu de données	Mohamed El Amine Meguenani	02/04/23 23:27:45	Non applicable	—	—

[Lien vers la page du dashboard](#)



Les logiciels utilisés

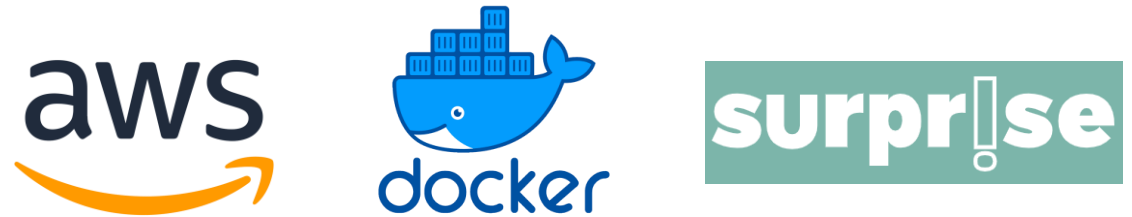
L'entrepôt de données



L'interface utilisateur



Système de recommandation



Environnement de programmation



L'entrepôt de données



BV Baptiste Viera ACCOUNTADMIN

Search

Tables

- CASTING
- DATE_NAISSANCE
- DATE_SORTIE
- FILM**
- GENRE
- GENRE_PREFERE
- MACTH_MOVIES_ID
- MOVIES
- MY_RATINGS_TRAIN_DATA
- MY_USER_MOVIE_RECOM...
- NOTE
- NO_RATINGS
- PERIODE
- POPULARITE
- RATINGS
- RATINGS_TRAIN_DATA
- RATINGS_TRAIN_DATA_LO...
- REALISATEUR

MOVIELENS / PUBLIC / FILM

Table ACCOUNTADMIN 3 days ago 42.2K 3.8MB

Table Details Columns Data Preview Copy History

Table definition

```
1 create or replace TABLE
2 MOVIELENS.PUBLIC.FILM (
3   ID NUMBER(38,0) NOT NULL,
4   TITLE VARCHAR(255),
5   PAYS VARCHAR(50),
6   ADULT VARCHAR(50),
7   MOYENNEVOTE FLOAT,
8   TOTALVOTE FLOAT,
9   POSTER_LEFT VARCHAR(255),
10  IDACTEUR NUMBER(38,0),
11  IDREALISATEUR NUMBER(38,0),
12  IDGENRE NUMBER(38,0),
13  IDPOPULARITE NUMBER(38,0),
14  IDPERIODE NUMBER(38,0),
15  IDDATESORTIE NUMBER(38,0),
16  POSTER_RIGHT VARCHAR(255),
17  primary key (ID),
```

MOVIELENS / PUBLIC / FILM

Table ACCOUNTADMIN 3 days ago 42.2K 3.8MB

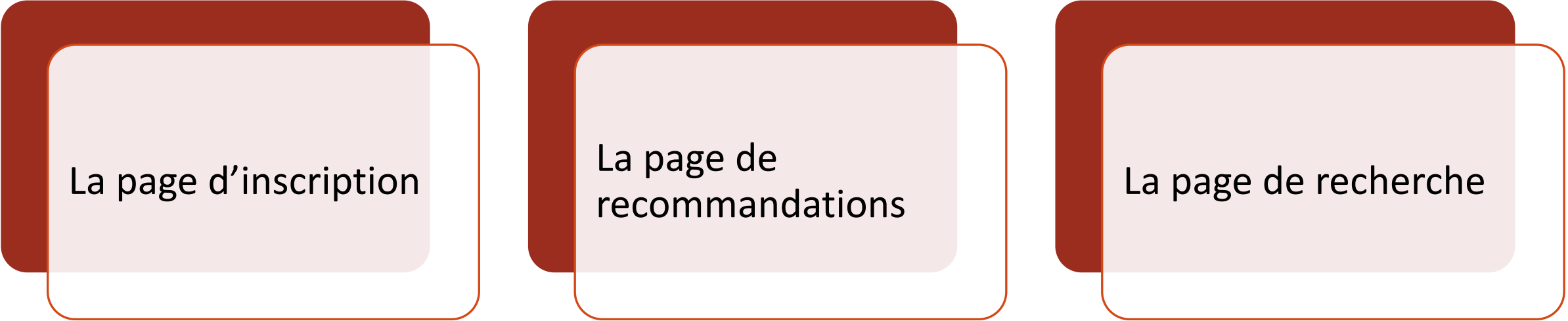
Table Details Columns **Data Preview** Copy History

• COMPUTE_WH 100 of 42.2K Rows • Updated just now

	↑	ID	TITLE	PAYS
1		1,652	00 schneider - jagd auf nihil baxter	DE
2		2,357	10 items or less	US
3		4,204	\$5 a day	US
4		4,951	10 things i hate about you	US
5		7,840	10,000 bc	US
6		8,209	1. mai – helden bei der arbeit	DE
7		8,420	...and god created woman	FR
8		9,051	10	US
9		10,035	100 girls	US
10		10,884	10 questions for the dalai lama	US
11		14,062	.45	US

L'interface utilisateur

Composé de 3 parties principales :



La page d'inscription

La page de
recommandations

La page de recherche

Inscris-toi !

Prenom
Pierre

Nom
Eloite

Nouveau nom d'utilisateur
pasdinspiration

Nouveau mot de passe
•

Adresse courriel
pierre@gmail.com

Pays
Canada

Anniversaire
2000/04/27

Selectionne tes 3 genres préférés
Romance x Family x Aventure x

S'inscrire

La page d'inscription

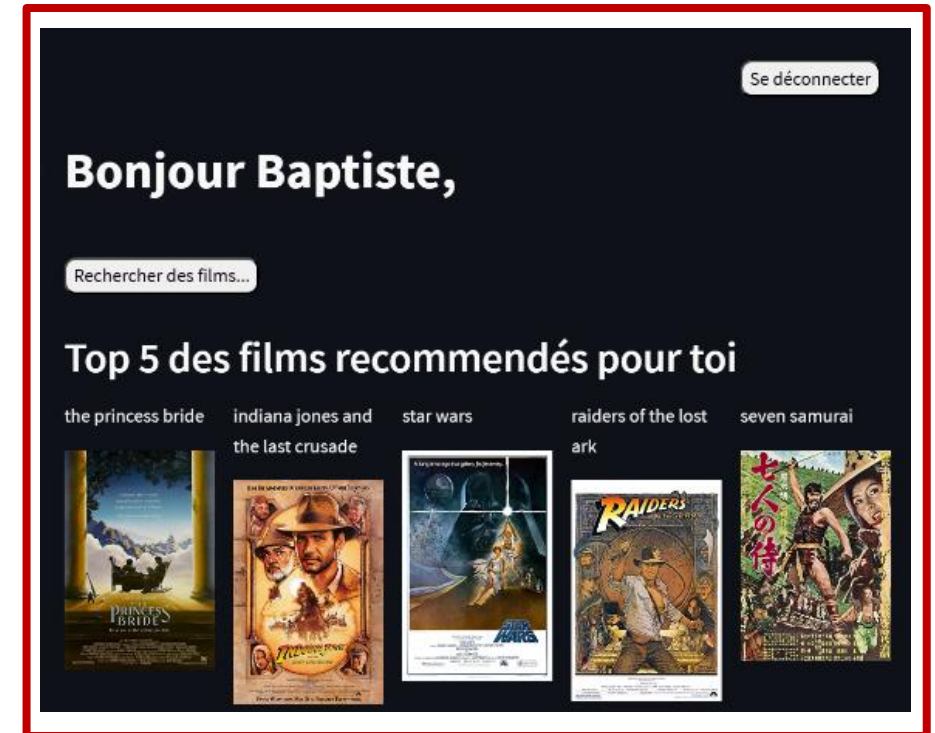
Permet à l'utilisateur de :

- Rentrer ses informations qui lui permettront par la suite de s'authentifier et d'accéder à son profil
- Donner au système ses préférences afin qu'il puisse dès de début lui proposer des films pertinents

La page des recommandations

Recommandations personnalisées pour un utilisateur en fonction :

- Des notes données par d'autres utilisateurs (filtrage collaboratif)
- De ses genres préférés
- De la période de l'année
- Du succès rencontré par les films
- Du pays de résidence de l'utilisateur
- De la date de naissance de l'utilisateur



La page des recommandations

Des films Adventure pour toi

the hidden fortress



man from deep river



storks



dream city



roar



Films pour Noel

girlfriends of christmas past



snow bride



the polar express



once upon a holiday



merry matrimony



Ca vient de chez vous...

precious



repentance



day of the fight



escape from new york



the end of the tour



Rechercher un film

Recherche par titre de films

star wars

☐ Filtrage Avancé

MOY : 8.1 - NB : 6778.0 MOY : 6.4 - NB : 4526.0 MOY : 6.4 - NB : 4074.0 MOY : 7.1 - NB : 4200.0 MOY : 5.8 - NB : 434.0

US-1994 US-1999 US-2001 US-2005 SG-2008

star wars star wars: episode i - the phantom menace star wars: episode ii - attack of the clones star wars: episode iii - revenge of the sith star wars: the clone wars



Note Estim

"3.517396674278
79"



Note Estim



Note Estim



Note Estim



Note Estim

Choisissez votre genre de Film

Action

Choisissez le réalisateur

James Cameron

Choisissez un intervalle de score

0.00 9.24

Choisissez une période

1959 2018

Appliquer

Rechercher un film

Recherche

☒ Filtrage Avancé

MOY : 6.8 - NB : 1138.0 MOY : 7.7 - NB : 4274.0 MOY : 7.4 - NB : 4208.0 MOY : 6.8 - NB : 19.0 MOY : 7.2 - NB : 12114.0

US-1994 FR-1991 GB-1984 US-2005 US-2009

true lies terminator 2: judgment day the terminator aliens of the deep avatar

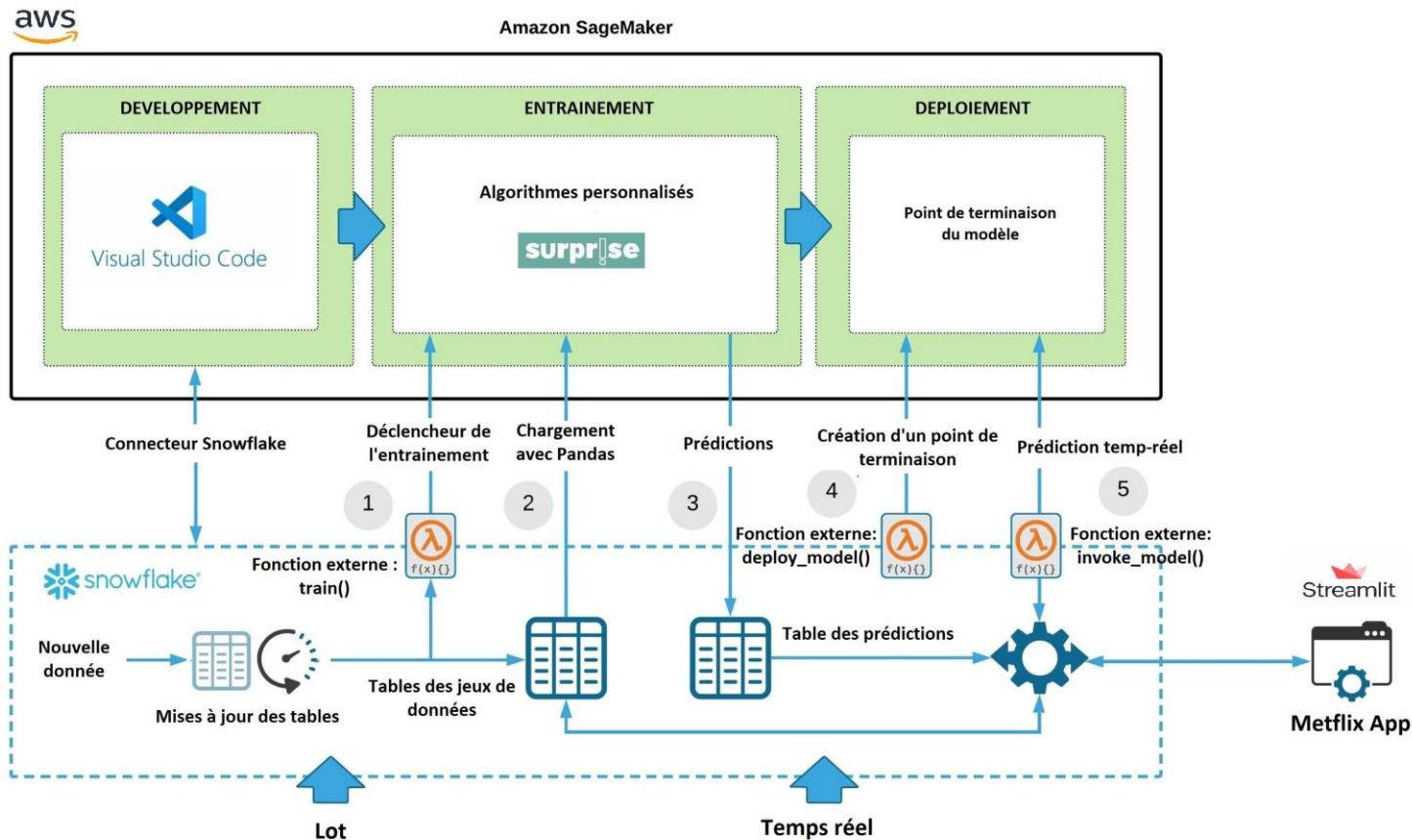


La page de recherche de films

Permet à l'utilisateur de :

- chercher des films à partir de leur titre
- les filtrer par rapport à leur réalisateur, leur date de sortie ou encore leur genre
- voir les notes estimées qu'il aurait pu mettre à ceux-ci grâce au modèle d'apprentissage machine

Architecture du système de recommandation



- **Etape 1 :** Charger l'ensemble des données relatives aux films dans Snowflake
- **Etape 2 :** Créer une image docker de l'entraînement et d'inférence personnalisée pour SageMaker
- **Etape 3 :** Créer une application sans serveur pour connecter Snowflake et SageMaker en déploiement des fonctions AWS Lambda et API Gateway.
- **Etape 4 :** Entraînement, déploiement mais aussi inférence du modèle à l'aide des fonctions externes de Snowflake

Création et déploiement du modèle sous format d'image Docker dans AWS

Training.py

```
data = Dataset.load_from_df(df[['USERID', 'MOVIEID', 'RATING']], reader)

# SVD algorithm.
algo = SVD()

# 5-fold cross-validation
cross_validate(algo, data, measures=['RMSE', 'MAE'], cv=5, verbose=True)

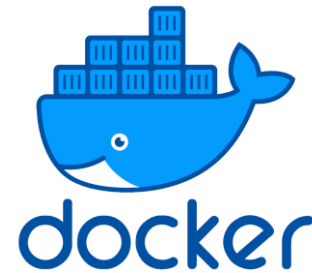
trainset = data.build_full_trainset()

# Predict ratings that are not in the training set.
testset = trainset.build_anti_testset()
predictions = algo.test(testset)

top_n = get_top_n(predictions, n=10)

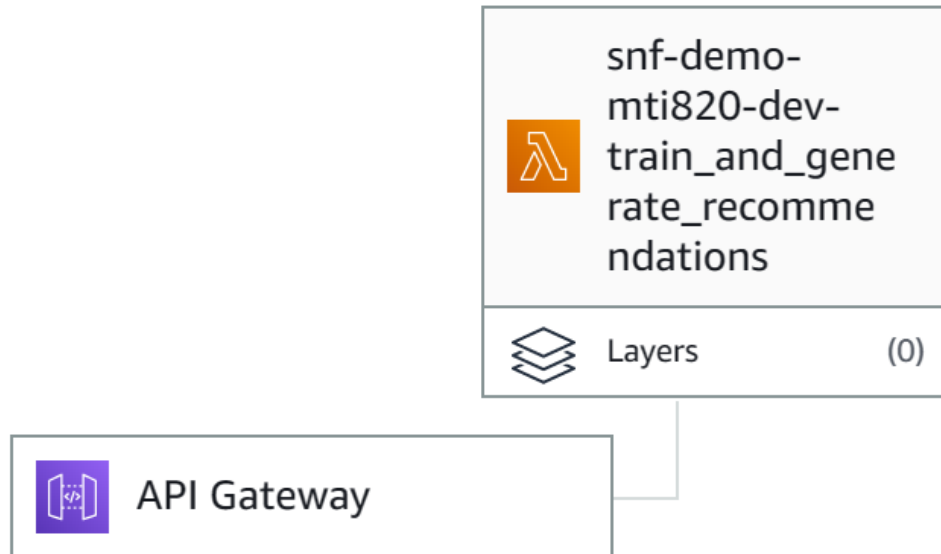
# save the top 10 recommended ratings into Snowflake
save_predictions_to_snowflake(top_n, cur, output_table_name)

# save the model
dump.dump(os.path.join(model_path, 'model.pkl'), algo=algo)
```



Amazon ECR
Elastic Container Registry

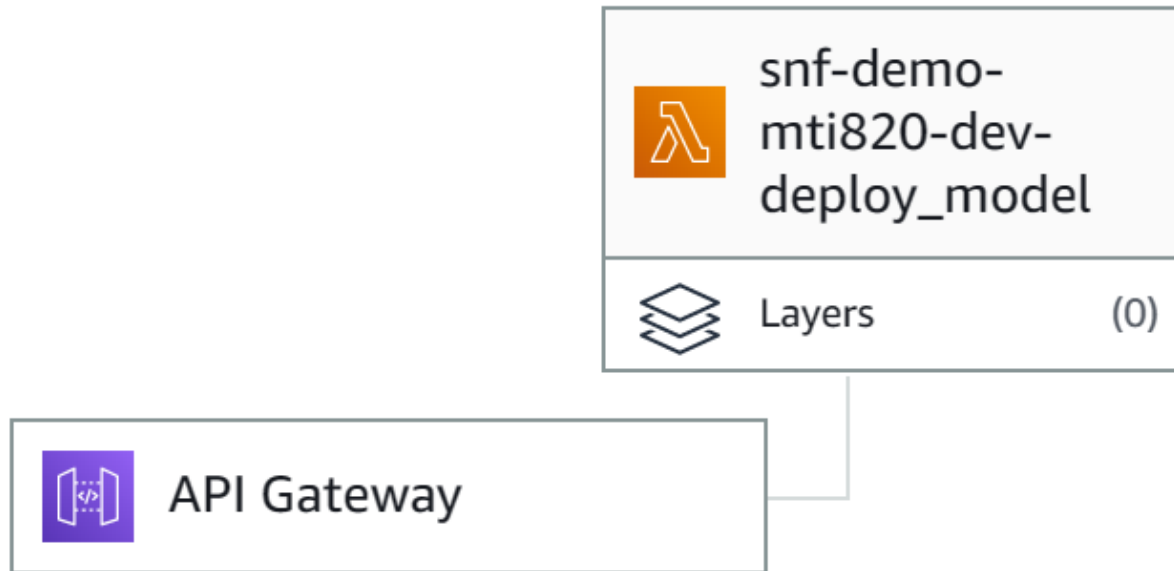
Fonction Lambda : Entraînement et recommandation



```
training_job_name = prefix
TRAINING_IMAGE_ECR_PATH = os.environ['training_image_ecr_path']
SAGEMAKER_ROLE_ARN = os.environ['sagemaker_role_arn']

response = client.create_training_job(
    TrainingJobName=training_job_name,
    HyperParameters=dict(input_table_name=input_table_name, output_table_name=output_table_name),
    AlgorithmSpecification={
        'TrainingImage': TRAINING_IMAGE_ECR_PATH,
        'TrainingInputMode': 'File'
    },
    RoleArn=SAGEMAKER_ROLE_ARN,
    OutputDataConfig={
        'S3OutputPath': s3_output_location
    },
    ResourceConfig={
        'InstanceType': 'ml.m5.xlarge',
        'InstanceCount': 1,
        'VolumeSizeInGB': 10
    },
    StoppingCondition={
        'MaxRuntimeInSeconds': 3600
    }
)
```

Fonction Lambda : Déploiement



```
# start the SageMaker training job
client = boto3.client('sagemaker')

ECR_PATH = os.environ['training_image_ecr_path']
SAGEMAKER_ROLE_ARN = os.environ['sagemaker_role_arn']

response = client.create_model(
    ModelName=model_name,
    PrimaryContainer={
        'Image': ECR_PATH,
        'ModelDataUrl': model_data_url
    },
    ExecutionRoleArn=SAGEMAKER_ROLE_ARN
)

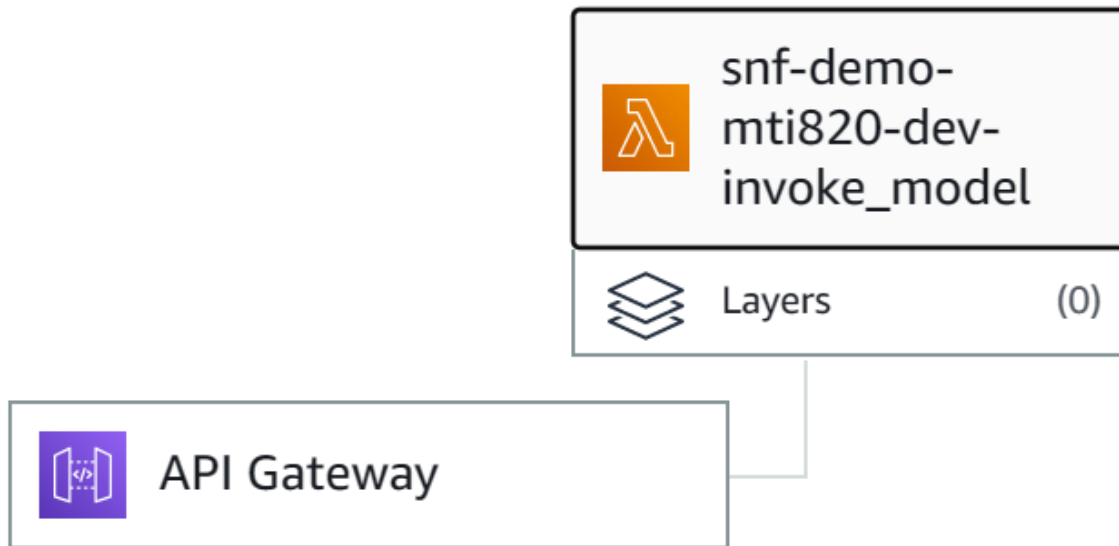
print(response)
print("now trying to create endpoint config...")

response = client.create_endpoint_config(
    EndpointConfigName=model_name,
    ProductionVariants=[
        {
            'VariantName': 'variant-1',
            'ModelName': model_name,
            'InitialInstanceCount': 1,
            'InstanceType': 'ml.t2.medium'
        }
    ]
)

print(response)
print("now trying to create the endpoint...")

response = client.create_endpoint(
    EndpointName=model_name,
    EndpointConfigName=model_name
)
```

Fonction Lambda : Appel du model



```
# invoke the SageMaker endpoint
client = boto3.client('sagemaker-runtime')
response = client.invoke_endpoint(
    EndpointName=model_name,
    Body=body.encode('utf-8'),
    ContentType='text/csv'
)
```


Fonctions Externes dans Snowflake

- Fonction « **train_and_get_recommendations** » pour l'entraînement et les prédictions

```
create or replace external function train_and_get_recommendations
(table_entree varchar, table_sortie varchar)
returns variant
    api_integration = snf_recommender_api_integration
as '<API_Gateway_entrainement_point_termination_url>';
```

- Fonction « **deploy_model** » pour le déploiement du modèle
- Fonction « **invoke_model** » pour l'appel du modèle

Utilisation du modèle par lot et en temps réel

Par lot :

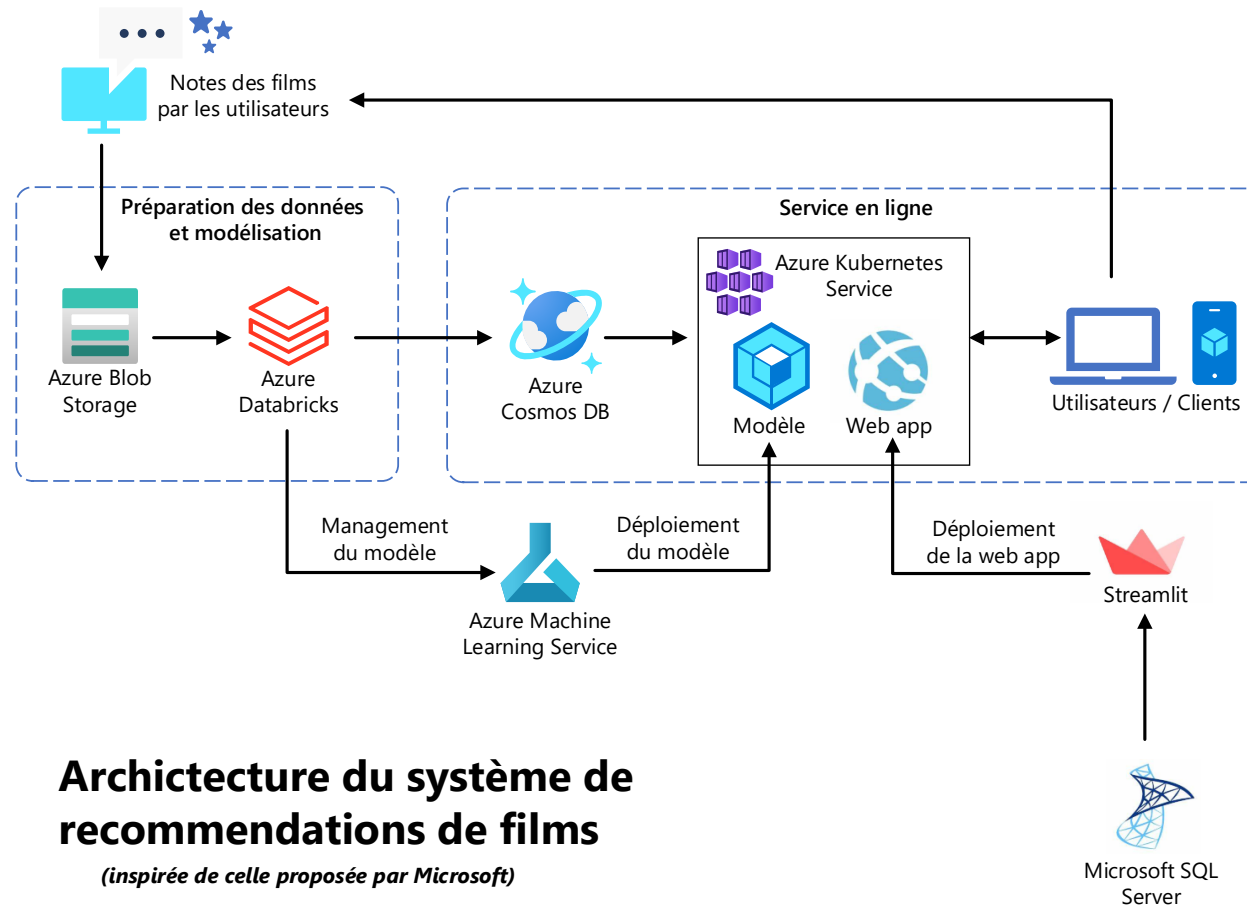
```
Select train_and_get_recommendations(" donnees_entrainement ", "recommandation_predite");
```

Temps-réel :

```
Select deploy_model('movielens-model-v1', '<s3_artefact_du_model>') ;
```

```
Select nn.utilisateur_id, nn.film_id, m.titre, invoke_model('movielens-model-v1', nn.utilisateur_id,  
nn.film_id) as note_predite from non_note nn, film f where nn.film_id = f.id;
```

Difficultés rencontrées : Microsoft Azure



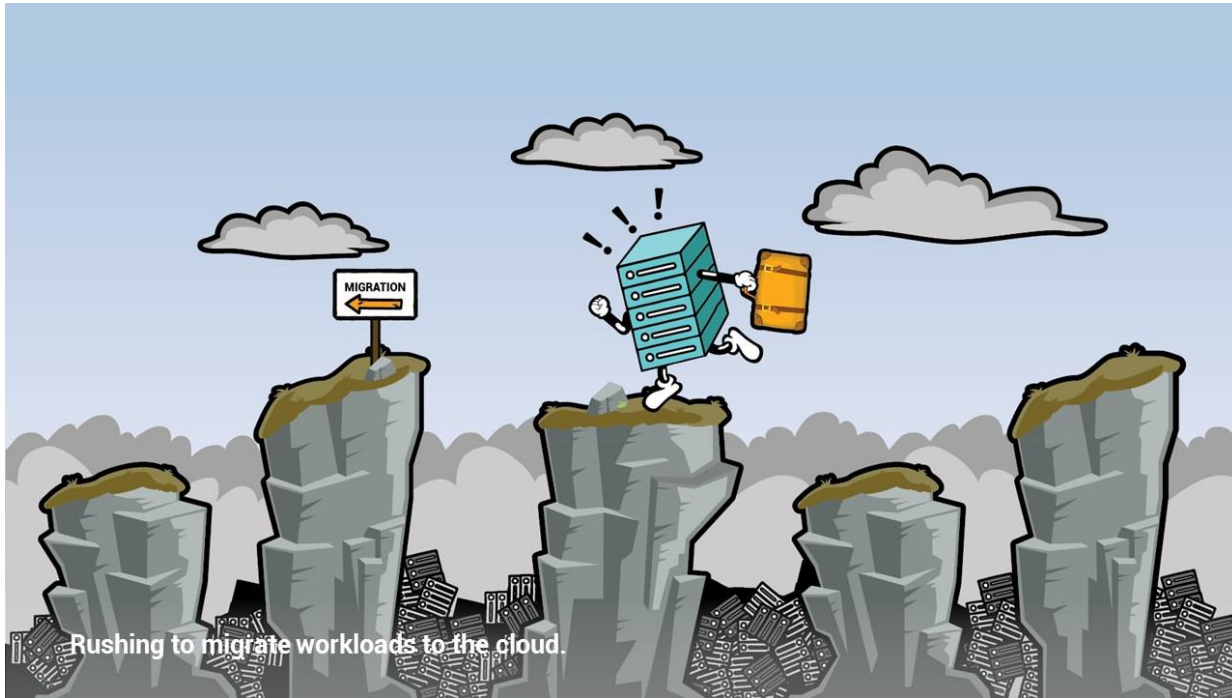
Architecture du système de recommandations de films

(inspirée de celle proposée par Microsoft)

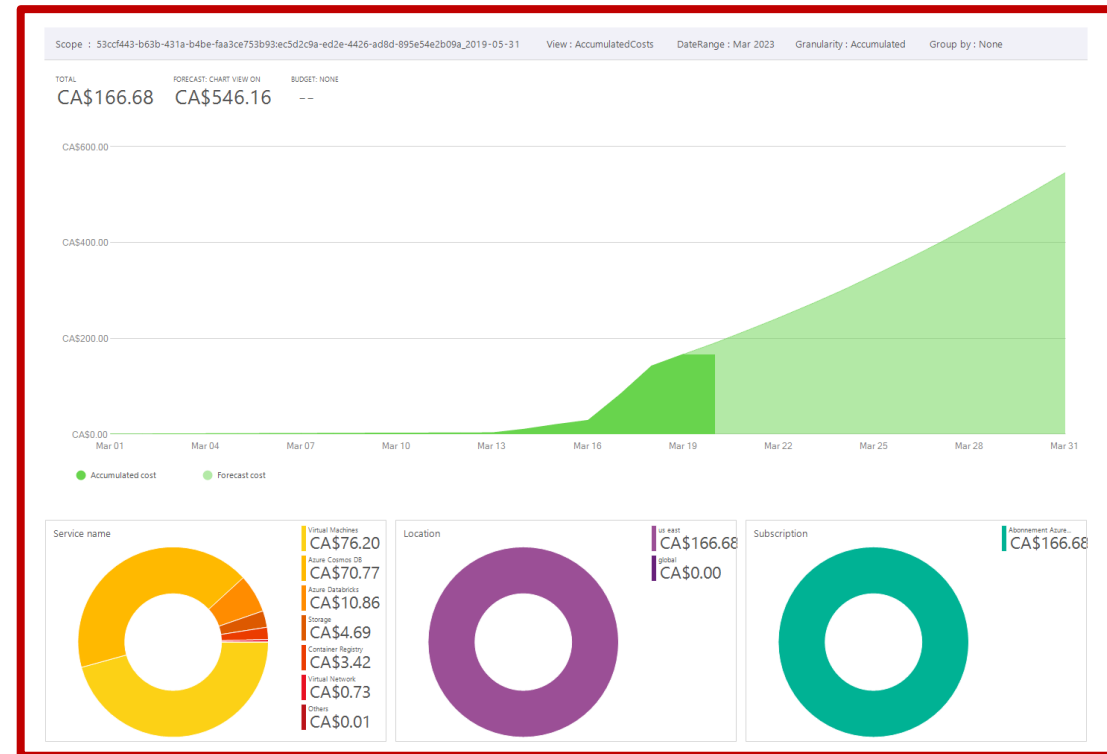
Architecture du système de recommandation initial basé sur les services de Microsoft Azure et non sur AWS et Snowflake

Difficultés rencontrées : Microsoft Azure

Migration de Spark 2.3 vers Spark ≥ 3.0



Problèmes financiers



Perspectives

- Possibilité des utilisateurs de noter des films et de mettre à jour la base de données
- Automatisation du pipeline d'apprentissage automatique pour le TOP 10 des recommandations
- Ajouter des algorithmes de filtrage par contenu ou des algorithmes hybrides basée sur du deep learning (apprentissage profond)
- Utilisation des données non supervisées
- Pouvoir choisir de ne pas avoir de recommandé certains types de film

Sources et références

- Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. J Big Data 9, 59 (2022). <https://doi.org/10.1186/s40537-022-00592-5>
- Schedl, M., Zamani, H., Chen, CW. et al. Current challenges and visions in music recommender systems research. Int J Multimed Info Retr 7, 95–116 (2018). <https://doi.org/10.1007/s13735-018-0154-2>
- Milano, S., Taddeo, M. & Floridi, L. Recommender systems and their ethical challenges. AI & Soc 35, 957–967 (2020). <https://doi.org/10.1007/s00146-020-00950-y>
- Seaver, N., “Captivating algorithms: Recommender systems as traps.” Journal of Material Culture, 24(4), 421–436, 2019, <https://doi.org/10.1177/1359183518820366>
- Gomez-Urbe, C-A., Hunt, N., The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Trans. Manage. Inf. Syst. 6, 4, Article 13 (January 2016), <https://doi.org/10.1145/2843948>
- Roh, Y., Heo, G., Euijong Whang S., A Survey on Data Collection for Machine Learning: a Big Data -- AI Integration Perspective, 2019, 1811.03402, arXiv, <https://doi.org/10.48550/arXiv.1811.03402>
- Rounak Banik, The movie Dataset, Kaggle, www.kaggle.com/datasets/rounakbanik/the-movies-dataset
- CrowsFlower, Blockbuster Database, Data world, <https://data.world/crowdflower/blockbuster-database>
- Documentation de Streamlit, <https://docs.streamlit.io/>



DEMO