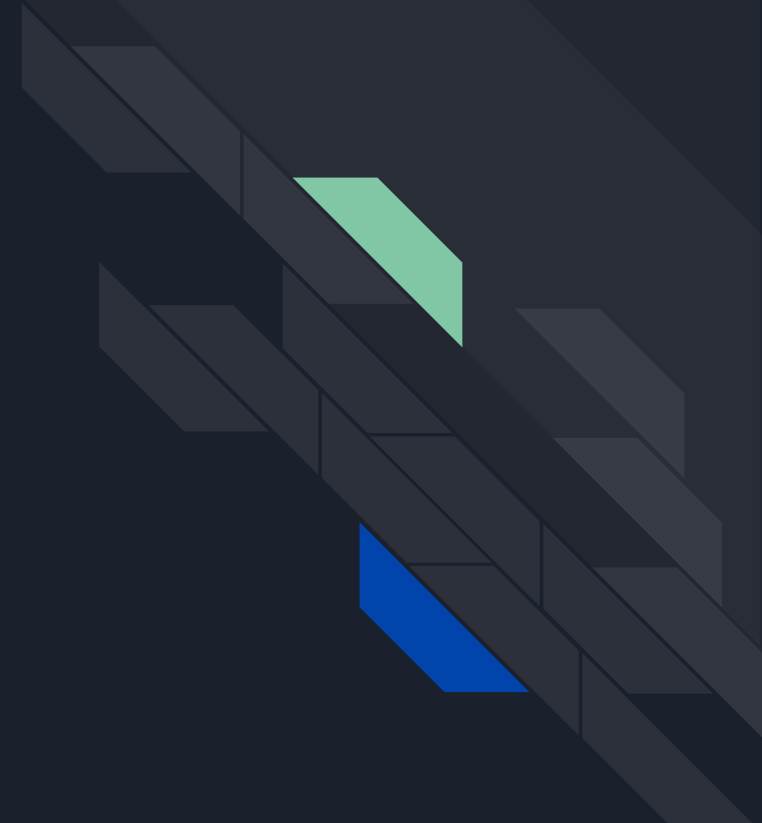




Construction de données publiques

Alexandre Marie
Thomas Capodano
Tony Lim
Salomé Rivoallanou-Drevet
Ludovic Hamel
Nicolas Bourneuf

Description critique des caractéristiques du projet





Les sources des données

Nom	Format	Lien
(API) IGDB	JSON	https://api-docs.igdb.com/#about
The Game Awards	CSV	https://www.kaggle.com/unanimad/the-game-awards
Video Games Sales	CSV	https://www.kaggle.com/gregorut/video-gamesales

Ingestion des données

Les données sont ingérées à l'aide d'un script C# et de la librairie LINQ.

Les données sont extraites depuis les fichiers CSV téléchargés et depuis les retours JSON généré par l'API IGDB.

Voici les différentes colonnes utilisées pour lier les jeux de données distincts.

Jeu de don	Colonnes	Type	Description	Exemple
Ventes	Rank	int	Identifiant (propre à ce jeu de données!!)	2
	Name	string	Nom du jeu	Wii Sports
	Platform	string	Console / plateforme (PS4, wii, etc)	Wii
	Year	int	Année de sortie	2006
	Genre	string	Genre du jeu	Sports
	Publisher	string	Éditeur	Nintendo
	NA_Sales	double	Ventes NA	01.08
	EU_Sales	double	Vente EU	05.02
	JP_Sales	double	Ventes JP	4.025
	Other_Sales	double	Autres ventes	16.25
	Global_Sales	double	Ventes globales	12.05
Game Awards	Year	int	année de la nomination à la récompense	2006
	Category of nominee	string	nom de la récompense pour laquelle le jeu est nommé	The Year (GOTY)
	name	string	nom du jeu ou de l'entreprise nominé(e)	Born in the Year of the Dragon
	company	string	développeurs (pas éditeur) du jeu nominé	BioWare
	won	bool	obtention ou non de la récompense	1
IGDB	voters	string	qui peut voter pour les jeux nominés	fan, jury
	Name	string	Nom du jeu	Terra Nova: Strik
	Platform	string	Console / plateforme (PS4, wii, etc)	PlayStation 2
	Player perspectives	string	Point de vue personnage (1er / 3eme personne)	First person
	Aggregated rating	float	Note du jeu sur 100	72.0
	Release date	string	Date de sortie	oct. 02, 2018
	Summary	string	Description du jeu	Lead an elite fig
	Theme	string	Type du jeu (Thriller, Science fiction ...)	Thriller
	company	string	Nom studio de création	Tomolo Games
	Game mode	string	Type de mode de jeu (solo, multi ...)	Single player
	Genre	string	genre du jeu	Role-playing (RPG)
	Keywords	string	mot pour décrire le jeu	exo zombies
	Similar game	string	liste de nom similaire à ce jeu	CyberPunk 2077
	Franchise	string	Nom de la franchise du jeu	Cyberpunk
	Game engine	string	Moteur du jeu	Flutter
	Web site	string	Url vers le site du jeu	www._____.com



Format produit

En sortie nous générons un fichier json.

A l'aide de nos jeux de données nous pouvons aboutir à un fichier qui a les colonnes suivantes.

Colonnes	Type	Description	Exemple
Rank	int	Identifiant propre à ce jeu de données	2
Name	string	Nom du jeu	Wii Sports
Platform	string	Console / plateforme (PS4, wii, etc)	Wii
Year	int	Année de sortie	2006
Genre	string	Genre du jeu	Sports
Publisher	string	Editeur	Nintendo
NA_Sales	double	Ventes NA	01.08
EU_Sales	double	Vente EU	05.02
JP_Sales	double	Ventes JP	4.025
Other_Sales	double	Autres ventes	16.25
Global_Sales	double	Ventes globales	12.05
RatingAPI	double	Moyenne des note attribuées par IGDB sur 100	82.05
GamesModes	array<string>	Type de jeu	["Single player", "Multiplayer"]
Company	array<string>	Entreprise de publication du jeu	["Nintendo", "Nintendo EAD"]
PlayerPerspective	array<string>	Perspective de la vision du personnage	["Side view"]



Pérennité de la stratégie automatique

Le processus d'ingestion des données puis de fusion pour la construction de la ressource est favorable à l'automatisation car la lecture de nos données porte sur des structures fiables et pérennes.

Un souci probable serait d'obtenir une erreur de synchronisation avec l'API IGDB, qui est une source de complexité pour notre automatisation.



Questions liées quantité / fiabilité

Notre ensemble de données généré regroupe les 500 jeux vidéos les plus vendus.

Il a été difficile de trouver un code identificateur permettant d'associer un même jeu vidéo d'une source de données à une autre. Nous nous sommes basés sur les noms des jeux vidéos.

Ceci a rendu l'ensemble de données produit moins fiable pour certains jeux.


Discussion adaptabilité
principes FAIR





Objectifs

- Générer un ensemble de données.
- Depuis 3 jeux de données ouverts.
- Centraliser des informations relatives aux jeux vidéos.



Plan des gestions de données ou comment être FAIR ?

- Edition d'une documentation pour décrire chaque colonne du modèle de données. Il faudrait ajouter un identifiant persistant autre que le rang du jeu (qui peut changer) ou le nom qui peut contenir des doublons ou différés selon les langues pour que la donnée soit plus "Findable".
- Avoir des sources référencées pour être "Accesible".
- Garder la même syntaxe pour chaque version pour être "Interoperable".
- Dépendre de plus d'API comme IGDB qui peuvent mettre à jour leurs données et les faire évoluer pour être plus "Reusable".



Discussions faisabilité

- Mise à jour périodique des données
 - Les données en provenance de l'API IGDB peuvent être facilement mises à jour. Cependant pour les données en provenance des fichiers CSV devront être mis à jour avec les nouvelles versions de ces fichiers mis à disposition en ligne.
- Mise à jour du schéma
 - Le schéma a une structure peu flexible à la modification / suppression. Cependant, il est possible d'ajouter aisément de nouvelles propriétés.
- Invalidation / correction des données
 - Un système de mesure de qualité ou de données incomplètes notamment pour les jeux qui manquent d'entreprises référencées. En effet, ils n'ont pas pu être associés dans tous les jeux de données.
- Contribution par un tiers
 - La structure du fichier permet une bonne lisibilité par l'être humain.