Rapport de projet

Big Data Warehouse

Mastère 1 Dev Manager Full Stack

Groupe composé de : Prashath Sivayanama, … Corentin, … Brice, Ye Jack

Professeur : Alexandre Bergere

Introduction

Pour garantir une exécution optimale des traitements et des fonctionnalités comme l'utilisation du format Iceberg, il est recommandé d'utiliser un cluster configuré avec la version 12.2 de Databricks. Lors de nos tests, nous avons rencontré des limitations avec la version 11, rendant difficile l'intégration et l'utilisation fluide d'Iceberg.

De plus, il est essentiel d'ajouter la bibliothèque suivante au cluster pour permettre la prise en charge d'Iceberg :

org.apache.iceberg:iceberg-spark-runtime-3.3_2.12:1.3.1

Cette bibliothèque peut être intégrée en utilisant Maven dans la configuration du cluster.

Valorant est un jeu de tir tactique en équipe, développé par Riot Games, qui combine des éléments de compétition et de stratégie. Sorti en 2020, il met en scène deux équipes de cinq joueurs s'affrontant dans des modes de jeu où chaque joueur incarne un "agent", disposant de compétences uniques. Ces agents sont classés en différents rôles, tels que Duelist, Sentinel, Initiator et Controller, qui influencent les stratégies d'équipe. Le jeu propose plusieurs cartes distinctes, chacune possédant des caractéristiques uniques, qui obligent les joueurs à adapter leurs choix d'agents et leurs tactiques. Avec une scène compétitive en plein essor, Valorant est devenu une référence dans l'univers de l'e-sport.

Objectif

Ce projet a pour objectif principal d'analyser les données relatives au jeu compétitif Valorant, en mettant un accent particulier sur les performances des agents en fonction des cartes jouées par des équipes professionnelles. Cette analyse vise à répondre à plusieurs questions fondamentales qui permettent de mieux comprendre les stratégies adoptées dans le cadre des matchs compétitifs. Ces questions incluent notamment : quels sont les agents les plus utilisés sur chaque carte ? Quels agents offrent les meilleures probabilités de victoire ? Et enfin, quelle carte maximise les performances des agents pour des équipes spécifiques ? Ces informations sont essentielles pour permettre aux équipes de perfectionner leurs stratégies et de mieux anticiper les choix de leurs adversaires.

Données utilisées

Ce projet s'appuie essentiellement sur des données collectées sur Kaggle, plateforme populaire regorgeant de jeux de données variés et complets, à l'heure actuelle, cette plateforme donne accès uniquement à des données structurées et prêtes à analyser, ce qui la rend tout à fait propice à l'exercice. De ce fait, en plus d'une couverture exhaustive des agents joués, elle met en évidence les cartes jouées et l'historique des scores réalisés des équipes, elle met aussi à disposition l'ensemble des performances que ces dernières peuvent effectuer.

https://www.kaggle.com/datasets/aliibrahim10/valorant-stats

https://www.kaggle.com/datasets/qualidea1217/valorant-pro-matches-since-april-2021

https://www.kaggle.com/datasets/evangower/valorant-esports-top-earnings?select=Players.csv

Architecture du système

Le schéma d'architecture proposé ici est par nature modulaire et scalable, et organisé en plusieurs couches : Bronze, Silver et Gold, pour une meilleure gestion et traitement des données. Les principales étapes se déroulent comme suit.

Collecte des données : Les données sont extraites des bases ouvertes des plateformes spécialisées.

Stockage initial (Bronze Layer) : les données brutes sont déposées dans un lac de données sur Azure Data Lake centrant à la fois les fichiers d'origine intacts.

Nettoyage et Transformation (Silver Layer) : Les données sont nettoyées, enrichies et organisées via Databricks et Spark. Cette étape inclut la suppression des doublons, la gestion des valeurs manquantes et la structuration des données pour travailler efficacement sur l'analyse.

Structuration avancée (Gold Layer) : Les données traitées sont organisées dans un modèle dimensionnel sous forme de tables de faits et de dimensions pour assurer une analyse plus poussée et utiliser les outils de visualisation.

Modélisation et Visualisation : Un notebook spécifique a été conçu dans le but de centraliser les visualisations des données transformées. Ce notebook, exécuté dans un Databricks, permet de générer des graphiques et de contrôler la cohérence des données.

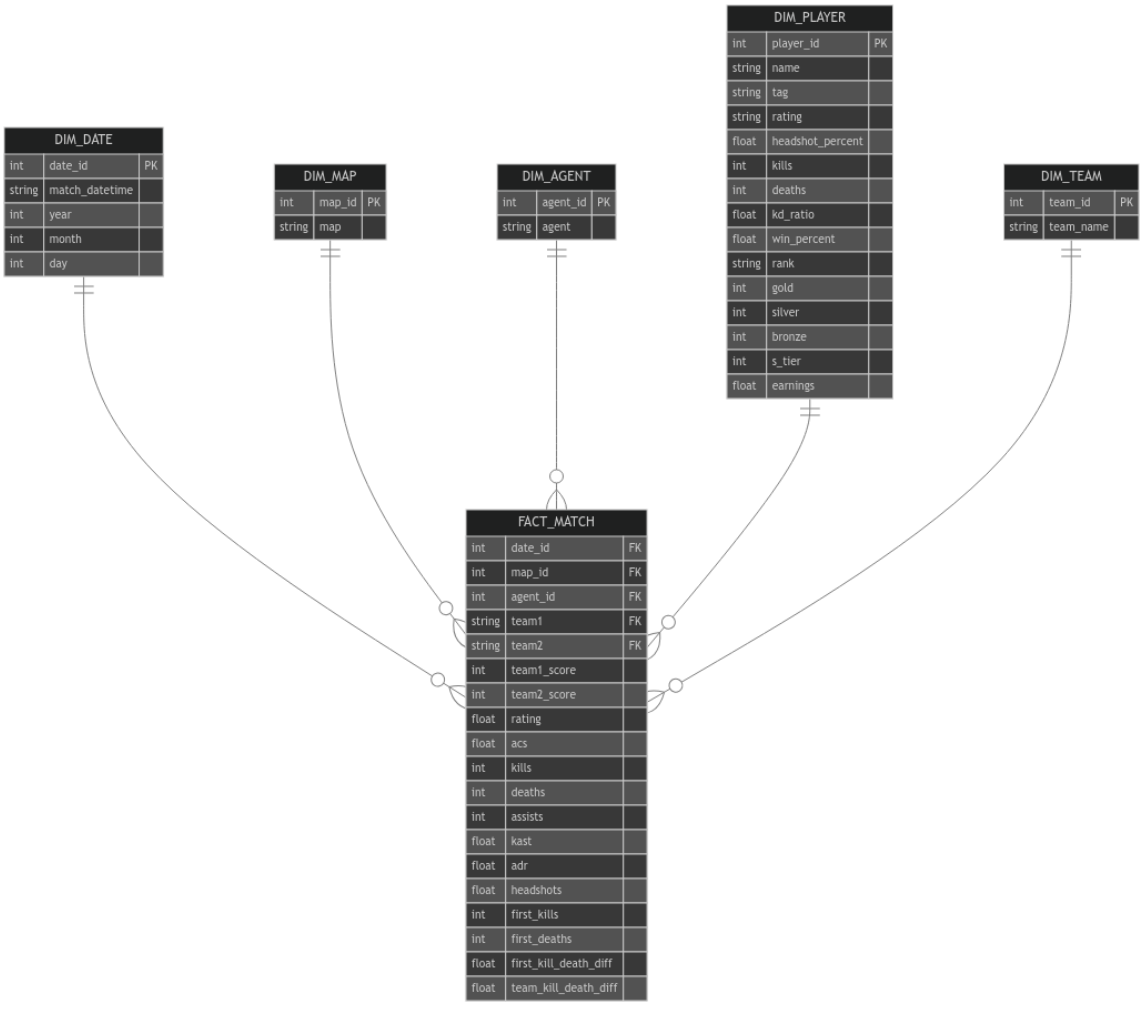
projet_data

Share | Create

Name	Type	Owner	Created at
ds-bronze	Notebook	prashath sivayanama	11/28/2024, 05:40:46 PM
ds-gold	Notebook	prashath sivayanama	01/02/2025, 01:57:48 PM
ds-silver	Notebook	prashath sivayanama	12/10/2024, 05:03:45 PM
Notebook visualization	Notebook	prashath sivayanama	12/29/2024, 10:13:36 PM

Modèle dimensionnel

La modélisation adoptée est orientée sur un modèle dimensionnel en étoile, qui regroupe les données en une table centrale entourée de plusieurs tables de dimensions. Ce modèle rend possible les analyses rapides et précises.



Traitement de données

Bronze Layer : Création et stockage des données brutes

1. Création de la table data_since_april

Cette table data_since_april, structure les données brutes en colonnes appropriées pour faciliter leur traitement.

Table

	id	match_datetime	patch	map	team1	team2	team1_score	team2_score	player_name	player_team	
1	0	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	HolyM0Ly	Impulse GW	Ki
2	1	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	Dolfo	Impulse GW	Br
3	2	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	zePiCzz	Impulse GW	As
4	3	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	s2newb	Impulse GW	Ne
5	4	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	Minikid	Impulse GW	Sc
6	5	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	Phatt	EGN Esports	Br
7	6	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	Dext	EGN Esports	Cy
8	7	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	fainz	EGN Esports	Je
9	8	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	al1en	EGN Esports	Or
10	9	2023/4/16 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	silenttt	EGN Esports	Sc
11	10	2023/4/16 7:15	6.05	Ascent	ZETA DIVISION	DetonatioN FocusMe	13	9	SugarZ3ro	ZETA DIVISION	Or
12	11	2023/4/16 7:15	6.05	Ascent	ZETA DIVISION	DetonatioN FocusMe	13	9	Laz	ZETA DIVISION	Sc
13	12	2023/4/16 7:15	6.05	Ascent	ZETA DIVISION	DetonatioN FocusMe	13	9	crow	ZETA DIVISION	Ka
14	13	2023/4/16 7:15	6.05	Ascent	ZETA DIVISION	DetonatioN FocusMe	13	9	TENNN	ZETA DIVISION	Ki

8,623+ rows | Truncated data

2. Création de la table player_bronze

Cette table player_bronze organise les données des joueurs, notamment leur classement, leurs performances et leurs gains financiers.

Table

	rank	player	gold	silver	bronze	s_tier	earnings
1	1	zombs	13	5	1	1	$113,550
2	2	ShahZaM	13	4	1	1	$113,450
3	3	dapr	14	3	1	1	$112,870
4	4	SicK	12	3	1	1	$112,650
5	5	cNed	12	4	2	1	$107,735
6	6	starxo	12	4	1	1	$104,394
7	7	Kiles	6	4	2	1	$104,093
8	8	nAts	14	6	0	1	$103,829
9	9	Chronicle	13	6	0	1	$103,574
10	10	d3ffo	12	4	2	1	$102,773
11	11	Sheydos	10	4	0	1	$102,626
12	12	BONECOLD	5	1	1	1	$101,918
13	13	Redgar	9	4	1	1	$101,115
14	14	zeek	13	4	1	1	$94,754
15	15	TenZ	8	4	3	1	$92,750

500 rows

3. Création de la table val_stats_bronze

Cette table val_stats_bronze structure les données statistiques détaillées des joueurs, incluant des informations sur leurs performances globales, leurs armes utilisées, ainsi que leurs résultats associés aux agents joués. Ces données sont essentielles pour effectuer des analyses approfondies sur les choix d'agents et leur impact sur les victoires.

Création de colonnes calculées pour fournir des indicateurs clés tels que le taux de victoire et le taux de sélection des agents.

Table

	region	name	tag	rating	damage_round	headshots	headshot_percent	aces	clutches	flawless	first_bloc
1	NA	ShimmyXD	#NA1	Radiant	135.8	null	24.9	null	null	null	
2	NA	XSET Cryo	#cells	Radiant	170.3	879	28.3	2	122	94	
3	NA	PuRelittleone	#yoruW	Radiant	147.5	720	24	3	117	59	
4	NA	Boba	#0068	Radiant	178.2	856	37.3	3	83	49	
5	NA	i love mina	#kelly	Radiant	149.8	534	24.4	2	71	38	
6	NA	Decay	#GODK	Radiant	134.1	null	26	2	162	94	
7	NA	Osias	#1212	Radiant	163.4	null	25.2	7	186	92	
8	NA	Knights RIKU	#KRN	Radiant	153.3	510	17.5	2	112	64	
9	NA	RaijuACE	#3131	Radiant	153.7	null	24.6	2	189	132	
10	NA	dawn	#24k	Radiant	153.6	339	20.8	1	56	44	
11	NA	100T Derrek	#100	Radiant	144.2	795	35.3	1	91	46	
12	NA	acts017	#ttv	Radiant	137.8	868	29.2	0	146	80	
13	NA	sam wow	#linda	Radiant	142.2	null	32.3	1	157	84	
14	NA	kipp	#0002	Radiant	131.1	null	20.1	2	175	98	

10,000+ rows | Truncated data

Silver Layer : Nettoyage et Transformation des Données

Dans la couche Silver, nous avons effectué un nettoyage approfondi des données brutes en utilisant Apache Spark via Databricks. Cette étape cruciale a impliqué plusieurs processus de transformation :

1. Configuration de l'Environnement

- Mise en place d'une connexion sécurisée avec Azure Data Lake Storage Gen2
- Configuration du catalogue Spark pour utiliser Iceberg comme format de table
- Création de deux bases de données distinctes :
  - spark_catalog.current : pour les données nettoyées
  - spark_catalog.trash : pour les données rejetées

2. Traitement des Données de Matches (player_april)

- Création d'une structure de table normalisée avec 45 colonnes incluant :
  - Informations de match (ID, date, patch, map)
  - Données d'équipe (noms, scores)
  - Statistiques de joueur (rating, kills, deaths, assists)
  - Métriques de performance (ADR, KAST, headshots)
- Nettoyage des données en :
  - Supprimant les valeurs NULL
  - Éliminant les doublons
  - Validant l'intégrité des données numériques

3. Traitement des Données de Joueurs (player)

- Nettoyage des données de classement et de récompenses
- Standardisation des formats de données
- Vérification de la cohérence des métriques (gold, silver, bronze, s_tier)

4. Traitement des Statistiques Valorant (val-stat)

- Normalisation des données avec :
  - Conversion en minuscules et suppression des espaces superflus
  - Arrondissement des valeurs décimales à 2 chiffres
  - Standardisation des noms d'agents et d'armes
- Optimisation de la structure en :
  - Supprimant la colonne 'region' jugée non pertinente
  - Validant l'intégrité des statistiques de tir (headshots, body shots, leg shots)

5. Gestion de la Qualité des Données

- Mise en place d'un système de séparation des données :
  - Les données valides sont stockées dans les tables "current"
  - Les données problématiques sont isolées dans les tables "trash"
- Traçabilité complète des données rejetées pour analyse ultérieure

Cette couche Silver assure ainsi la qualité et la cohérence des données qui seront utilisées dans la couche Gold pour les analyses plus poussées.

Gold Layer : Modélisation Dimensionnelle et Création du Data Warehouse

Dans la couche Gold, nous avons implémenté un modèle en étoile pour optimiser les analyses des données Valorant. Cette structuration permet une analyse plus efficace et plus pertinente des performances de jeu.

1. Structure du Modèle en Étoile

Tables de Dimensions :

1. **Dimension Agent (dim_agent)**
    - Identifiant unique pour chaque agent
    - Nom de l'agent
    - Permet l'analyse des performances par agent
2. **Dimension Map (dim_map)**
    - Identifiant unique pour chaque carte
    - Nom de la carte
    - Facilite l'analyse des performances par carte
3. **Dimension Team (dim_team)**
    - Identifiant unique pour chaque équipe
    - Nom de l'équipe
    - Permet le suivi des performances par équipe
4. **Dimension Player (dim_player)**
    - Identifiant unique pour chaque joueur
    - Informations détaillées :
        - Tag et nom du joueur
        - Statistiques globales (rating, headshot%, K/D ratio)
        - Palmarès (médailles or, argent, bronze)
        - Gains en tournois
5. **Dimension Date (dim_date)**
    - Identifiant unique pour chaque date
    - Décomposition temporelle (année, mois, jour)
    - Permet l'analyse des tendances temporelles

Table de Faits (fact_match) :

- Métriques clés de performance :
    - Scores des équipes
    - Statistiques individuelles (rating, ACS, kills, deaths, assists)
    - Indicateurs tactiques (KAST, ADR, headshots)
    - Performances spécifiques (first kills, first deaths)

2. Implémentation Technique

- Utilisation du format Iceberg pour le stockage des tables
- Configuration optimisée pour les requêtes analytiques
- Mise en place de clés étrangères pour assurer l'intégrité référentielle
- Stockage dans Azure Data Lake avec une structure organisée

3. Avantages du Modèle

1. **Performance**
    - Optimisation des requêtes analytiques
    - Réduction de la redondance des données
    - Facilité de maintenance
2. **Flexibilité**
    - Possibilité d'ajouter de nouvelles dimensions
    - Adaptation facile aux évolutions du jeu
    - Support de différents types d'analyses
3. **Qualité des Données**
    - Intégrité référentielle assurée
    - Traçabilité des modifications
    - Cohérence des données garantie

Cette structuration en Gold Layer permet de répondre efficacement aux besoins d'analyse des performances dans Valorant, en facilitant la création de tableaux de bord et de rapports détaillés.

Visualisation et Analyse des Données

Notre analyse des données Valorant s'est concentrée sur quatre axes principaux, chacun apportant des insights spécifiques sur différents aspects du jeu compétitif.

1. Performance des Agents par Carte

Analyses Réalisées :

- **Heatmap des performances** : Visualisation croisée des ratings moyens des agents sur chaque carte
- **Top 3 agents par carte** : Identification des agents les plus performants sur chaque map

```
=== Analyse des performances des agents ===
Top 3 agents par map:

Map: Ascent
+----+-------+----------+------+------------+
|rank|agent  |avg_rating|avg_kd|games_played|
+----+-------+----------+------+------------+
|1   |Sova   |1.04      |1.12  |5059740     |
|2   |Chamber|1.03      |1.2   |752640      |
|2   |Sage   |1.03      |1.17  |887880      |
+----+-------+----------+------+------------+
only showing top 3 rows


Map: Bind
+----+-------+----------+------+------------+
|rank|agent  |avg_rating|avg_kd|games_played|
+----+-------+----------+------+------------+
|1   |Jett   |1.07      |1.26  |1040760     |
|1   |Reyna  |1.07      |1.22  |201096      |
|2   |Chamber|1.06      |1.24  |971964      |
```

- **Statistiques globales** : Analyse des performances générales incluant :
  - Rating moyen
  - Ratio K/D
  - Pourcentage de headshots
  - Dégâts moyens par round (ADR)

Insights Clés :

- Identification des agents meta pour chaque carte
- Patterns de performance spécifiques à certaines cartes
- Agents polyvalents vs. spécialistes de certaines cartes

2. Analyse des Performances des Équipes

Visualisations :

- **Graphique des taux de victoire** : Comparaison des performances des 50 meilleures équipes

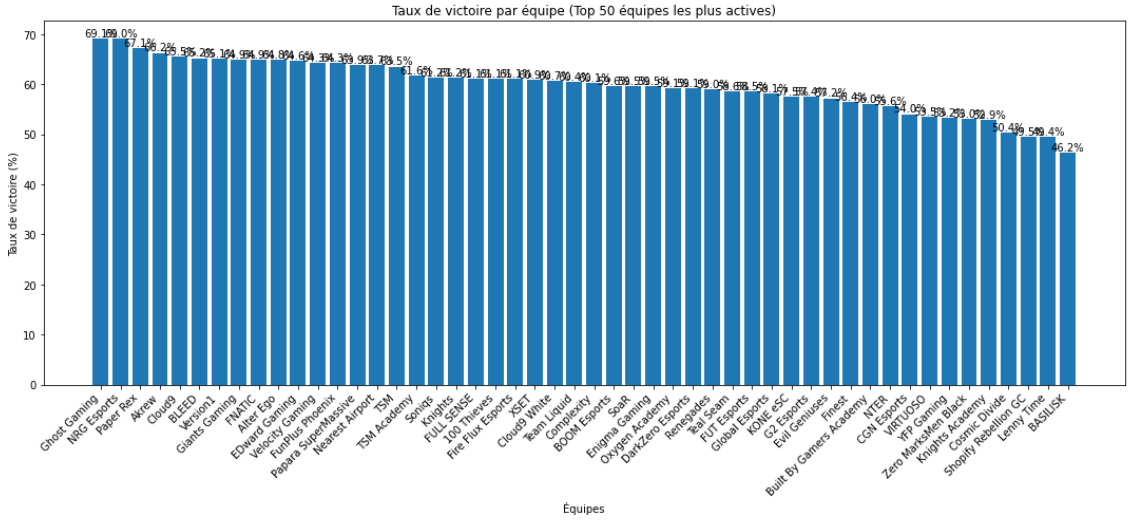
```
=== Analyse des performances des équipes ===

Statistiques des équipes:
+----------------------+-------------+---------+--------+
|team_name             |total_matches|avg_score|win_rate|
+----------------------+-------------+---------+--------+
|Ghost Gaming          |28488        |11.65    |69.14   |
|NRG Esports           |18936        |11.54    |69.0    |
|Paper Rex             |26160        |11.54    |67.1    |
|Akrew                 |22176        |11.46    |66.23   |
|Cloud9                |18348        |11.69    |65.51   |
|BLEED                 |24552        |11.26    |65.19   |
|Version1              |25008        |11.55    |65.07   |
|Giants Gaming         |18936        |11.64    |64.91   |
|FNATIC                |21792        |11.88    |64.88   |
|Alter Ego             |18312        |11.51    |64.81   |
|EDward Gaming         |22032        |11.67    |64.63   |
|Velocity Gaming       |29220        |11.32    |64.28   |
|FunPlus Phoenix       |18744        |11.36    |64.28   |
|Papara SuperMassive   |19212        |11.45    |63.88   |
```

- **Scatter plot matches/victoires** : Corrélation entre l'expérience et le succès
- **Graphique des scores moyens** : Distribution des performances offensives

Métriques Analysées :

- Taux de victoire global
- Nombre total de matches joués
- Score moyen par match
- Consistance des performances



3. Évolution Temporelle des Performances

Analyses Temporelles :

- **Courbes d'évolution** : Suivi des performances des top 5 équipes
- **Métriques suivies** :
  - Taux de victoire mensuel
  - Score moyen par période
  - ADR (Average Damage per Round)

Observations :

- Tendances de performance sur la durée
- Périodes de domination
- Impact des changements de meta

4. Analyse des Performances Individuelles

Visualisations :

- **Distribution des ratings** : Répartition des niveaux de performance
- **Corrélations statistiques** :
  - Headshots vs. Win Rate
  - Rating vs. Performance globale

Métriques Individuelles :

- Pourcentage de headshots
- Taux de victoire individuel
- Rating par niveau de jeu

Conclusions Principales

1. **Méta du Jeu** :
    - Identification des agents dominants par carte
    - Évolution des stratégies gagnantes
2. **Performance des Équipes** :
    - Corrélation entre expérience et succès
    - Impact de la consistance sur les résultats
3. **Évolution Temporelle** :
    - Adaptabilité des équipes aux changements
    - Stabilité des performances sur la durée
4. **Performances Individuelles** :
    - Impact du skill mécanique sur les victoires
    - Importance de la polyvalence des joueurs

Ces visualisations permettent une compréhension approfondie des différents aspects du jeu compétitif Valorant, offrant des insights précieux pour les équipes et les analystes.