PROJET DATA WAREHOUSE & POWER BI

PRESENTED BY

Kabil boufares

Firas ben Ali

Nourhenne ben Abdelghaffar

Contexte du projet

🎯 Objectif global:

L’entreprise souhaite renforcer sa capacité à prendre des décisions stratégiques en exploitant efficacement ses données internes.

 Rôle du data specialist

En tant que spécialiste des données, vous êtes chargé de :

· Collecter et structurer les données pertinentes

· Mettre en place un entrepôt de données (Data Warehouse)

· Créer des dashboards interactifs pour les utilisateurs métiers à l’aide de Power B

 Pourquoi un entrepôt de données ?

· Centralisation des données hétérogènes

· Amélioration de la qualité, cohérence et traçabilité des informations

· Facilitation de l’analyse multidimensionnelle

Objectifs du projet:

Choisir un thème stratégique

→ Thème sélectionné : Cybersécurité

1

Collecter des données pertinentes

→ Datasets structurés et complémentaires issus de Kaggle

2

Créer un entrepôt de données relationnel

→ Modélisation en étoile avec table de faits et tables de dimensions

3

Développer un processus ETL avec SSIS

→ Extraction, nettoyage, transformation et chargement des données dans un Data Warehouse

4

Construire des tableaux de bord interactifs avec Power BI

→ Visualisation claire des tendances, des risques et des indicateurs clés

5

Choix du thème : Cybersécurité

Pourquoi la cybersécurité ?

- Sujet d’actualité critique face à la montée des cyberattaques mondiales
- Risques croissants : phishing, ransomwares, DDoS, fuites de données
- Fortes implications économiques, juridiques et organisationnelles

Intérêt analytique

- Données riches et multidimensionnelles
- Permet des analyses croisées : type d’attaque, secteur touché, localisation, pertes financières
- Sujet idéal pour démontrer la puissance d’un projet de Business Intelligence

infographie sur la cybersécurité mondiale.

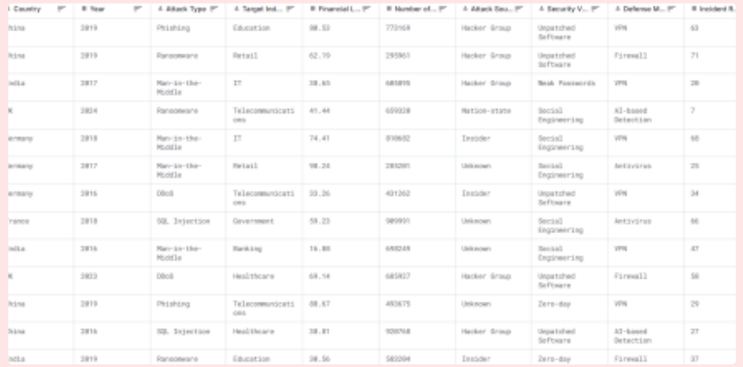
Jeux de données utilisés

Source principale

- **Plateforme :** Kaggle
- **Nom du dataset :** Global Cybersecurity Threats (2015–2024)
- **Format :** Fichier CSV structuré et prêt à intégrer dans Power BI / SSIS

Contenu du dataset

- **Pays :** localisation des incidents
- **Année :** période de survenue
- **Type d’attaque :** phishing, malware, ransomware, etc.
- **Secteur ciblé :** finance, santé, industrie...
- **Pertes financières estimées (en millions $)**
- Utilisateurs affectés, origine de l’attaque, durée de résolution



Country	Year	Attack Type	Target Ind...	Financial L...	Number of...	Attack Sou...	Security V...	Defense M...	Incident R...
hina	2019	Phishing	Education	80.53	773169	Hacker Group	Unpatched Software	VPN	63
hina	2019	Ransomware	Retail	62.19	295969	Hacker Group	Unpatched Software	Firewall	71
nita	2017	Man-in-the-Middle	IT	38.65	605895	Hacker Group	Weak Passwords	VPN	28
K	2024	Ransomware	Telecommunicati ons	41.44	609338	Nation-state	Social Engineering	AI-based Detection	7
rmany	2018	Man-in-the-Middle	IT	74.41	816652	Insider	Social Engineering	VPN	68
rmany	2017	Man-in-the-Middle	Retail	98.24	283201	Unknown	Social Engineering	Antivirus	25
rmany	2016	DDoS	Telecommunicati ons	33.26	431262	Insider	Unpatched Software	VPN	34
rance	2018	SQL Injection	Government	59.23	989891	Unknown	Social Engineering	Antivirus	66
nita	2016	Man-in-the-Middle	Banking	16.88	698248	Unknown	Social Engineering	VPN	47
K	2023	DDoS	Healthcare	68.14	685827	Hacker Group	Unpatched Software	Firewall	58
hina	2019	Phishing	Telecommunicati ons	88.67	491675	Unknown	Zero-day	VPN	29
hina	2016	SQL Injection	Healthcare	38.81	926768	Hacker Group	Unpatched Software	AI-based Detection	27
nita	2019	Ransomware	Education	38.56	581004	Insider	Zero-day	Firewall	37

data set

✅ **Critères de sélection**

- Dataset structuré et cohérent
- Richesse analytique
- Pertinent pour un entrepôt de données et l’analyse stratégique

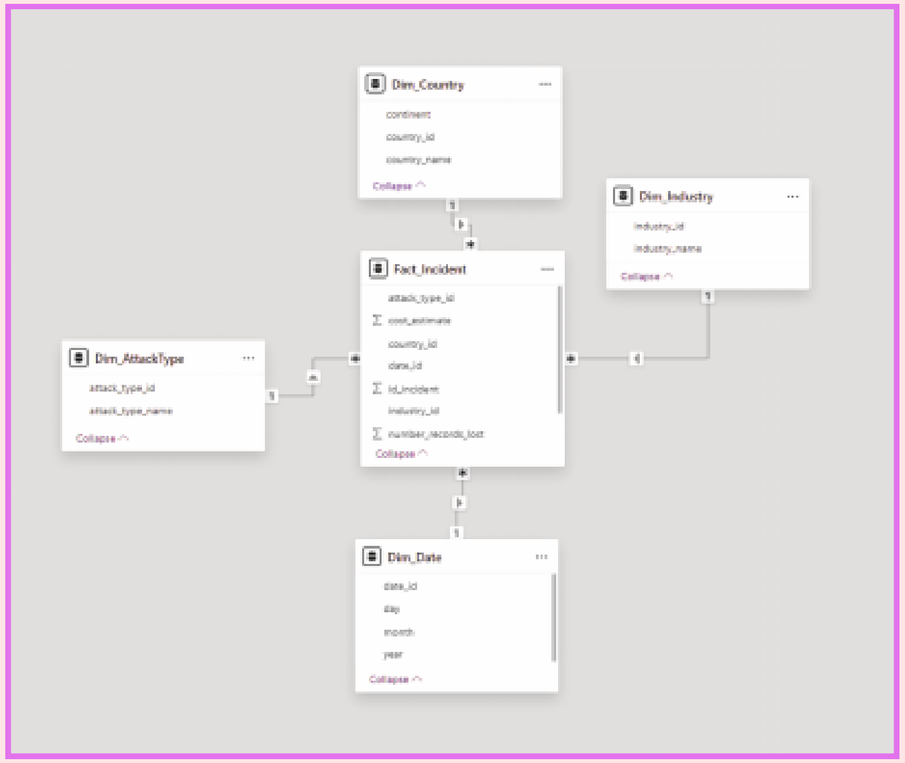
Nettoyage et préparation des données

Étapes de prétraitement appliquées

1. Suppression des doublons
→ Pour éviter les biais statistiques
2. Traitement des valeurs manquantes
→ Suppression ou imputation selon le contexte
3. Normalisation des formats
→ Uniformisation des noms (majuscules, minuscules, accents...)
4. Conversion des types de données
→ Montants financiers → float, années → int
5. Création de colonnes dérivées
→ ID pour pays, année, secteur, type d'attaque, etc.

Objectif : garantir la qualité, la cohérence et la compatibilité des données avec le Data Warehouse.

Architecture du Data Warehouse



modèle en étoile avec les noms des tables et leurs liens

Modèle de données adopté : Modèle en étoile

Le modèle en étoile a été choisi pour sa simplicité, sa lisibilité et son efficacité en analyse décisionnelle.

Structure principale

· Table de faits : Fact_Incident

→ Contient les mesures quantitatives (coûts, utilisateurs impactés)

· Tables de dimensions :

→ Dim_Country : Pays

→ Dim_Date : Année

→ Dim_Industry : Secteur

→ Dim_AttackType : Type d'attaque

Relations

· Clés primaires dans les dimensions

· Clés étrangères dans la table de faits

· Relations de type 1:N

Processus ETL avec SSIS

ETL : Extraction, Transformation, Chargement

Le processus ETL a été développé avec SQL Server Integration Services (SSIS), dans Visual Studio.

Étapes réalisées :

· Extraction

→ Lecture des fichiers CSV via composant Flat File Source

· Transformation

→ Nettoyage : doublons, conversion des types

→ Colonnes dérivées : identifiants, formats normalisés

→ Lookup vers les dimensions (pays, secteur, attaque...)

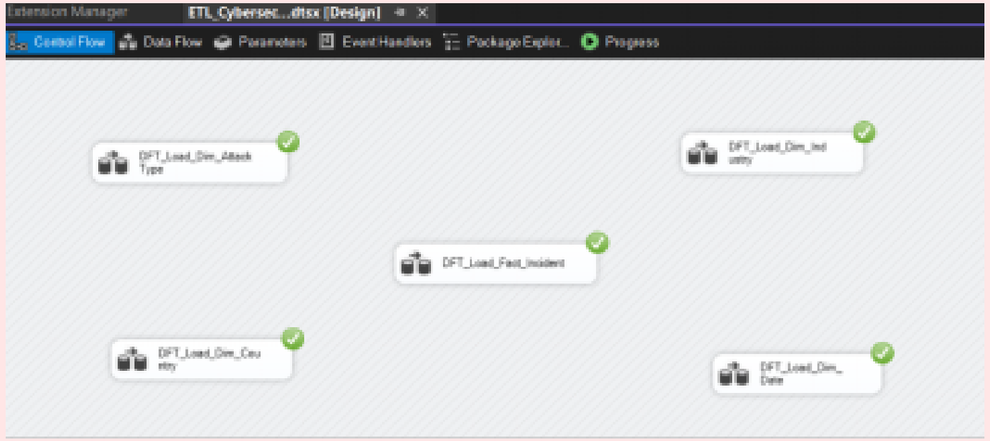
· Chargement

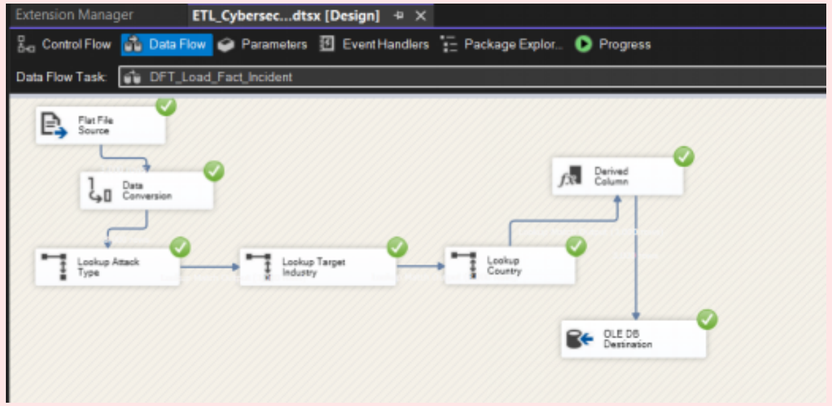
→ Envoi des données nettoyées vers les tables SQL Server via OLE DB Destination

Organisation du projet SSIS

Chaque table a un Data Flow Task dédié :

- DFT_Load_Dim_Country
- DFT_Load_Dim_AttackType
- DFT_Load_Dim_Date
- DFT_Load_Dim_Industry
- DFT_Load_Fact_Incident





flux ETL dans SSIS (avec les transformations, Lookups, etc.)

Création des tables SQL

Implémentation dans SQL Server (SSMS)

Les tables du modèle en étoile ont été créées à l'aide de scripts SQL exécutés dans SQL Server Management Studio.

Tables de dimensions créées :

Dim_Country (country_id, country_name)
· Dim_AttackType (attack_type_id, attack_type_name)
· Dim_Date (date_id, year)
· Dim_Industry (industry_id, industry_name)

Table de faits : Fact_Incident

Clés étrangères vers chaque dimension
· Colonnes de mesure : financial_loss, affected_users

Respect de l'intégrité référentielle

· Clés primaires auto-incrémentées
· Contraintes de clés étrangères dans la table de faits
· Vérification par des requêtes SELECT et jointures

Connexion à Power BI & Préparation

Connexion au Data Warehouse

- Utilisation de Power BI Desktop
- Connexion directe à SQL Server
- Chargement des tables du schéma (faits + dimensions)

01

Préparation des données dans Power BI

- Renommage des colonnes pour plus de lisibilité
- Création de mesures DAX :
  - o SUM(financial_loss)
  - o COUNT(incident_id)
- Vérification des relations (liens entre ID de dimensions)

02

Objectif : rendre les données prêtes à l'analyse visuelle tout en assurant la cohérence et la performance

Dashboards Power BI : Vue générale

Objectif des dashboards

- Offrir une vision stratégique claire des cybermenaces
- Permettre une exploration interactive des données par les décideurs
- Mettre en lumière les tendances, zones critiques, et secteurs vulnérables

Principaux éléments visuels créés :

- KPI (coûts totaux, nombre d’incidents…)
- Histogrammes par type d’attaque
- Courbes d’évolution annuelle
- Cartes géographiques interactives
- Graphiques par secteur d’activité

Interactivité assurée via :

- Filtres (pays, année, industrie…)
- Info-bulles dynamiques
- Signets pour navigation fluide entre vues

Répartition géographique des attaques

Objectif de la visualisation

- Montrer où se concentrent les cyberattaques dans le monde
- Identifier les zones géographiques à risque
- Comparer les niveaux d’exposition entre régions

Insights clés

- Les pays les plus touchés sont généralement les plus numérisés :
  - États-Unis, Inde, pays européens...
- Les pays émergents présentent un risque de sous-détection
- Corrélations possibles avec les contextes géopolitiques et économiques

Filtrage dynamique

- Possibilité de filtrer par type d’attaque, année, ou secteur ciblé

Types d’attaques les plus fréquents

Objectif de la visualisation

- Identifier les vecteurs d’attaque les plus courants
- Quantifier leur fréquence relative
- Comparer l’impact des différents types d’attaques

Observations clés

- Phishing : attaque la plus répandue sur la période analysée
- Suivi par : Ransomware, DDoS, Intrusions internes
- Certaines attaques moins fréquentes ont un coût unitaire très élevé

Utilité métier

- Permet de prioriser les ressources de défense
- Aide à orienter la formation et la sensibilisation interne

Secteurs les plus impactés

Objectif de la visualisation

- Identifier les industries les plus ciblées par les cybermenaces
- Estimer les pertes financières sectorielles
- Mettre en lumière les secteurs à forte criticité

Secteurs les plus touchés

- Santé : très ciblée, pertes critiques, données sensibles
- Finance : attaques fréquentes, enjeux de réputation
- Technologie & Industrie : souvent visées pour sabotage ou espionnage
- Autres secteurs à surveiller : gouvernement, éducation, commerce

Visualisation croisée

- Nombre d’incidents vs. Coûts totaux
- Possibilité de filtrer par pays ou type d’attaque

Analyse des résultats

Répartition par type d’attaque

- Phishing domine largement les incidents signalés
- Les ransomwares provoquent les pertes les plus lourdes
- Certaines attaques peu fréquentes sont très destructrices (ex : menaces internes)

Évolution dans le temps

- Augmentation constante entre 2015 et 2023
- Pics d’incidents liés à des événements mondiaux (COVID-19, tensions géopolitiques)

Répartition géographique

- Pays les plus touchés : États-Unis, Inde, pays européens
- Les zones émergentes souffrent d’un manque de détection ou de protection

Impact par industrie

- Santé et finance : pertes financières très élevées
- Industrie et technologie : fortement exposées aux cyberattaques

Conclusion stratégique

- Nécessité d’une veille active en cybersécurité
- Importance d’une analyse en temps réel des incidents

Conclusion du projet

Résumé de la démarche

- Création d’un Data Warehouse sur les menaces de cybersécurité
- Utilisation de SSIS pour l’ETL : extraction, transformation, chargement
- Construction de dashboards Power BI interactifs et dynamiques
- Analyse stratégique des types d’attaques, des pays ciblés et des secteurs vulnérables

✅ Résultats atteints

- Données centralisées, nettoyées et historisées
- Visualisations claires et utiles à la prise de décision
- Plateforme évolutive, prête à intégrer d’autres sources de données

Objectif atteint :

Fournir aux décideurs une vision globale et exploitable de la menace cyber pour orienter leurs choix de sécurité.

Recommandations & pistes d’amélioration

Perspectives d’évolution du projet

1. Renforcer l’analyse prédictive

→ Intégration d’algorithmes de Machine Learning pour anticiper les menaces

2. Ajouter de nouvelles sources de données

→ CERTs nationaux, bases de vulnérabilités (CVE), flux RSS spécialisés

3. Mettre en place des alertes automatisées

→ Notifications via Power BI ou services connectés en cas de pics d’incidents

4. Automatiser l’ETL

→ Utilisation de SQL Server Agent pour les mises à jour régulières

5. Former les utilisateurs métier

→ Sensibilisation à la lecture de dashboards et aux enjeux de cybersécurité

Vision long terme

Faire évoluer le projet vers une plateforme complète de cybersurveillance connectée en temps réel aux données internes et externes.

Merci pour votre attention