

# Women's Rights

**Clémence AMIARD - Jimmy LETTE VOUETO - Gina PALESCH**

M1 Dev Manager Full Stack

XDEV703 – BIG DATA / SPARK

06/01/2025

## Table of contents

<b>1</b>	<b>Introduction .....</b>	<b>3</b>
1.1	Membre du groupe .....	3
1.2	Présentation du projet.....	3
1.3	Objectif .....	3
<b>2</b>	<b>Données utilisées.....</b>	<b>4</b>
2.1	Description des sources de données .....	4
2.2	Type de données et leur format .....	6
<b>3</b>	<b>Architecture .....</b>	<b>8</b>
3.1	Schéma d'architecture .....	8
3.2	Explication du schéma .....	8
<b>4</b>	<b>Modèle dimensionnel.....</b>	<b>9</b>
4.1	Schéma en étoile.....	9
4.2	Explication du schéma .....	10
<b>5</b>	<b>Modèle Conceptuel des Données .....</b>	<b>11</b>
5.1	Schéma .....	11
<b>6</b>	<b>Implémentation.....</b>	<b>12</b>
6.1	Azure.....	12
6.2	Databricks.....	13
6.3	Bronze.....	14
6.4	Silver .....	15
6.5	Gold.....	18
6.6	Visualisation .....	21
<b>7</b>	<b>Conclusion.....</b>	<b>22</b>

## 1 Introduction

### 1.1 Membre du groupe

Clémence Amiard
Jimmy Lette Voueto
Gina Palesch

### 1.2 Présentation du projet

Dans le cadre de notre analyse des inégalités de genre à l'échelle mondiale, nous avons entrepris un projet visant à explorer et comparer les indicateurs clés liés aux droits des femmes et à leur représentation socio-économique, à la fois à l'échelle des continents et au sein des pays individuels, sur différentes périodes.

Pour cela, nous avons intégré et enrichi les données existantes en ajoutant une colonne supplémentaire indiquant le **continent** pour chaque entrée. Cette amélioration permet de :

- Comparer les indicateurs entre les continents pour observer les disparités régionales.
- Analyser les tendances au sein d'un continent donné ou d'un pays spécifique sur plusieurs années.

### 1.3 Objectif

L'objectif principal de ce projet est d'analyser et de visualiser les inégalités entre les sexes en s'appuyant sur des indicateurs économiques et sociaux. Plus précisément, nous cherchons à :

- **Comparer les disparités de genre entre les continents** afin d'identifier les régions les plus touchées par les inégalités.
- **Analyser les tendances à l'intérieur des continents et des pays**, permettant une compréhension plus fine des dynamiques locales.
- **Mettre en lumière les progrès ou les reculs dans le temps**, en étudiant les données sur plusieurs années pour identifier les facteurs de changement.

## 2 Données utilisées

### 2.1 Description des sources de données

Les données utilisées proviennent du jeu de données intitulé "*Women's Rights*" disponible sur [Kaggle](#). Ce jeu de données couvre divers aspects des inégalités entre les sexes à travers plusieurs fichiers décrits comme suit :

#### 1. Données sur le rapport femmes-hommes dans le travail non rémunéré

- **Nom du fichier** : female-to-male-ratio-of-time-devoted-to-unpaid-care-work.csv
- **Description** : Ce fichier met en évidence le temps consacré par les femmes par rapport aux hommes aux activités non rémunérées, telles que les soins aux enfants, les tâches domestiques et d'autres formes de travail informel. Les données sont basées sur une étude menée par l'OCDE en 2014.
- **Interprétation des valeurs** :
  - Un **rapport élevé** indique que les femmes consacrent significativement plus de temps que les hommes au travail non rémunéré, soulignant une répartition inégale des tâches.
  - Un **rapport faible** suggère une répartition plus équilibrée du travail non rémunéré entre les sexes.

#### 2. Données sur la part des femmes dans les percentiles de revenus

- **Nom du fichier** : 2- share-of-women-in-top-income-groups-atkinson-casarico-voitchovsky-2018.csv
- **Description** : Ce fichier montre la proportion de femmes dans les différents percentiles de revenus (top 0,1 %, top 1 %, top 10 %, etc.). Ces données révèlent la représentation des femmes dans les couches les plus riches de la population, ce qui reflète les inégalités d'accès aux opportunités économiques.
- **Interprétation des valeurs** :
  - Un **pourcentage élevé** dans un percentile donné indique une meilleure représentation des femmes dans cette tranche de revenus.
  - Un **pourcentage faible** montre que les femmes sont sous-représentées dans les revenus les plus élevés, ce qui peut refléter des inégalités systémiques.

### 3. Données sur le taux de participation au marché du travail des femmes par rapport aux hommes

- **Nom du fichier** : 3- ratio-of-female-to-male-labor-force-participation-rates-ilo-wdi.csv
- **Description** : Ce fichier présente le ratio de participation des femmes par rapport aux hommes sur le marché du travail, en pourcentage. Ces données permettent de comprendre les différences dans la participation économique et d'évaluer les progrès en matière d'égalité des sexes dans le domaine de l'emploi.
- **Interprétation des valeurs** :
  - Un **pourcentage élevé** signifie que la participation des femmes est proche ou égale à celle des hommes, ce qui reflète une égalité plus grande dans l'accès au marché du travail.
  - Un **pourcentage faible** indique que la participation des femmes est bien inférieure à celle des hommes, mettant en évidence des obstacles économiques ou culturels pour les femmes.

### 4. Données sur l'écart salarial entre les sexes

- **Nom du fichier** : 6- gender-gap-in-average-wages-ilo.csv
- **Description** : Ce fichier documente l'écart salarial en pourcentage entre les hommes et les femmes dans différents pays et sur plusieurs années.
- **Interprétation des valeurs** :
  - Un **pourcentage positif** indique que les hommes gagnent davantage que les femmes. Plus la valeur est élevée, plus l'écart est significatif.
  - Un **pourcentage négatif** montre que les femmes gagnent, en moyenne, plus que les hommes dans un contexte donné.

## 2.2 Type de données et leur format

### 1. Données sur le rapport femmes-hommes dans le travail non rémunéré

- **Description** : Montre le rapport entre le temps consacré par les femmes et les hommes au travail non rémunéré (soins, tâches domestiques).
- **Colonnes** :
  - Entity : Nom du pays, Texte
  - Code : Code ISO-3 du pays, Texte
  - Year : Année, Entier
  - Female to male ratio : Rapport, Nombre décimal
- **Format** : CSV

### 2. Données sur la part des femmes dans les percentiles de revenus

- **Description** : Montre la proportion de femmes dans les différents percentiles de revenus (top 0,1 %, top 1 %, top 10 %, etc.).
- **Colonnes** :
  - Entity : Nom du pays, Texte
  - Code : Code ISO-3 du pays, Texte
  - Year : Année, Entier
  - Share of women in top X% : Pourcentage de femmes dans les percentiles, Nombre décimal ou Null
- **Format** : CSV

### 3. Données sur le taux de participation au marché du travail des femmes par rapport aux hommes

- **Description** : Indique le ratio femmes-hommes pour la participation au marché du travail, exprimé en pourcentage.
- **Colonnes** :
  - Entity : Nom du pays, Texte
  - Code : Code ISO-3 du pays, Texte
  - Year : Année, Entier
  - Ratio of female to male labor force participation rate (%) : Ratio femmes-hommes, Nombre décimal
- **Format** : CSV

### 4. Données sur l'écart salarial entre les sexes

- **Description** : Documente l'écart salarial entre les hommes et les femmes, exprimé en pourcentage.
- **Colonnes** :
  - Entity : Nom du pays, Texte
  - Code : Code ISO-3 du pays, Texte
  - Year : Année, Entier
  - Gender wage gap (%) : Écart salarial, Nombre décimal
- **Format** : CSV

## 3 Architecture

### 3.1 Schéma d'architecture



### 3.2 Explication du schéma

#### 1. Source de données (Kaggle)

- Les données brutes sont téléchargées depuis **Kaggle**. Il s'agit de **fichiers CSV** contenant des informations sur des sujets liés aux femmes, comme le ratio femme-homme pour le travail non rémunéré.

#### 2. Stockage dans Azure Data Lake Storage (Couche Bronze)

- Les fichiers CSV sont chargés dans **Azure Data Lake Storage**. Ils se trouvent dans la **couche Bronze**, où les données sont **non structurées** et doivent être nettoyées et transformées avant d'être analysées.

#### 3. Stockage dans DBFS

- Les **fichiers CSV** de la couche Bronze sont chargés dans Databricks, où un **mount** est effectué pour connecter Databricks à Azure Data Lake Storage, permettant ainsi l'accès aux données via **DBFS**.

#### 4. Traitement des données dans Databricks (Apache Spark et Iceberg)

- Apache Spark est utilisé pour effectuer les transformations et le nettoyage des données. Ensuite, les **données nettoyées** sont stockées dans la **couche Silver**, où elles sont structurées et prêtes à être analysées.
- Les **données agrégées** et **analysées** se trouvent dans la **couche Gold**, prêtes à être utilisées pour des rapports ou des visualisations.
- Les données sont stockées sous forme de **tables Iceberg**, garantissant une gestion efficace des versions et des métadonnées. Ces tables Iceberg sont ensuite sauvegardées au format Parquet dans Azure Data Lake Storage.

#### 5. Visualisation et Analyse

- Les visualisations dans **Databricks** permettent de tirer des insights des données et de prendre des décisions éclairées.



## 4 Modèle dimensionnel

### 4.1 Schéma en étoile

CONTINENT
ContinentId
ContinentName

INDICATOR
IndicatorId
IndicatorName
Unit
Source

GenderStats
Id
IdEntity
IdYear
IdContinent
IdIndicator
Value

ENTITY
EntityId
EntityName

YEAR
YearId
YearValue

## 4.2 Explication du schéma

Le **Schéma en étoile** est une structure utilisée dans les **Data Warehouses**. Il se compose d'une **table de faits** centrale et de plusieurs **tables de dimensions**.

### Table de faits (FactGenderStats) :

Cette table contient les données quantitatives utilisées pour l'analyse. Dans notre cas, elle comprend divers **indicateurs** liés à l'égalité des sexes, tels que :

- Rapport entre le temps consacré par les femmes et les hommes au travail non rémunéré
- Proportions de femmes dans les tranches de revenus les plus élevées
- Écart salarial entre les sexes
- Rapport de la participation des femmes par rapport aux hommes sur le marché du travail

Cette table de faits, qui contient les principales mesures du projet, n'est pas au format Iceberg car l'auto-incrémentation n'est pas supportée. Pour plus de facilité, nous avons exceptionnellement créé cette dernière table au format Delta.

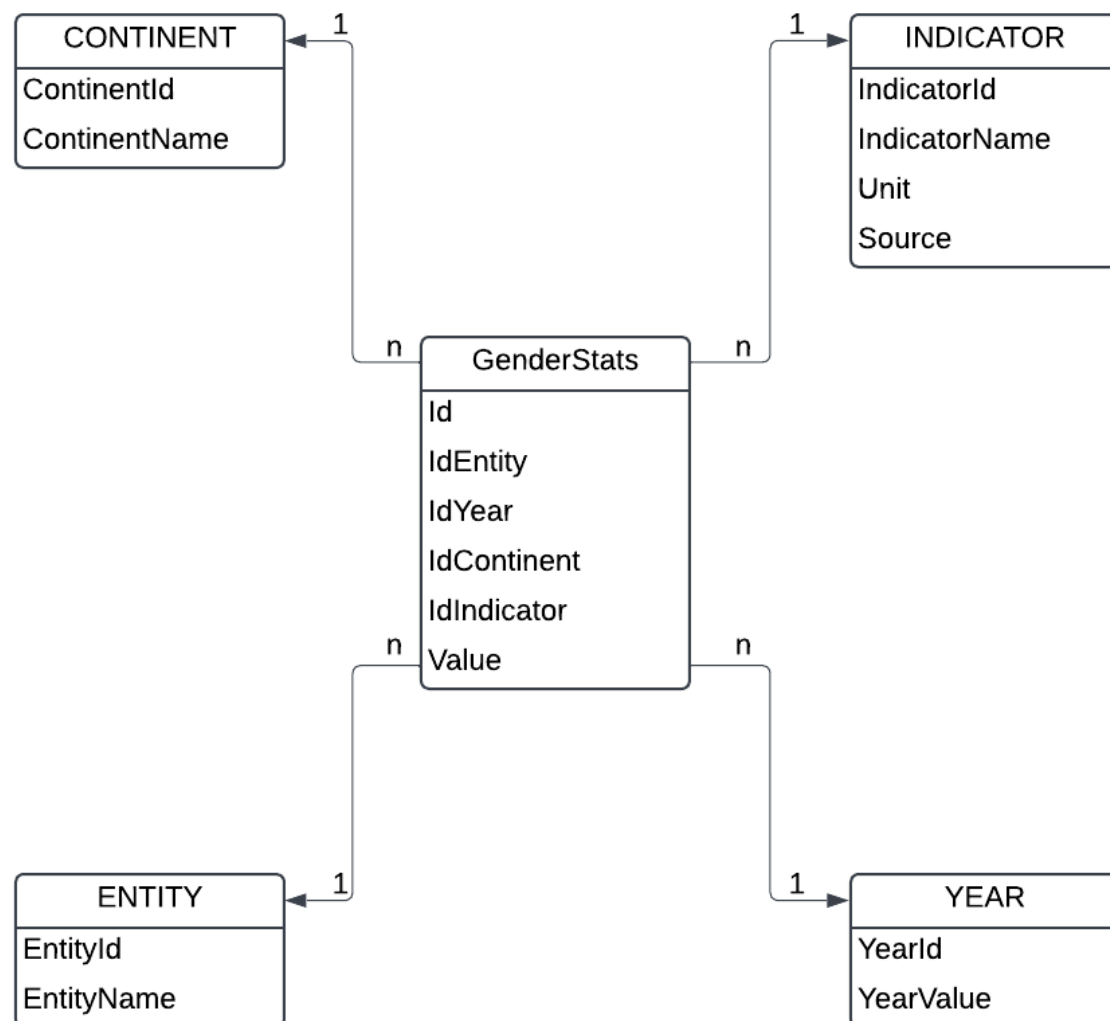
### Tables de dimensions :

Les tables de dimensions contiennent des informations descriptives qui contextualisent les données de la table des faits. Dans notre projet, nous avons plusieurs tables de dimensions :

- **DimensionEntity** : Cette table contient des informations sur les **pays** présentes dans les données.
- **DimensionYear** : Cette table contient les **années** pertinentes pour chaque observation.
- **DimensionContinent** : Cette table contient la **classification des continents**, permettant de grouper et d'analyser les données par régions géographiques.
- **DimensionIndicator** : Cette table contient les **indicateurs sociaux** spécifiques utilisés dans les données.

## 5 Modèle Conceptuel des Données

### 5.1 Schéma



## 6 Implémentation

### 6.1 Azure

Tout d'abord, il était nécessaire de créer un **Azure Storage Account** pour stocker les données dans le cloud Azure. Ce compte de stockage sert de base pour le téléchargement et la gestion des fichiers et des dossiers. À l'intérieur de ce compte de stockage, tu peux ensuite créer plusieurs **containers**, qui agissent comme des "dossiers" structurés pour organiser les données.

Une fois le compte de stockage configuré, des **containers** ont été créés pour différentes couches de données dans le modèle **Lakehouse**, à savoir **Bronze**, **Silver** et **Gold**. Ces containers représentent différentes étapes du traitement des données :

- **Bronze** : Données brutes, non structurées et non modifiées.
- **Silver** : Données nettoyées et transformées.
- **Gold** : Données agrégées et finales prêtes pour l'analyse.

The screenshot shows the Azure Storage Explorer interface for a storage account named 'dlkefreijlettevoueto'. The left sidebar shows the navigation pane with 'Conteneurs de blobs' selected. The main pane displays a list of blob containers. The table below represents the data shown in the interface.

Nom	Dernière modification	Niveau d'accès anonyme	État du bail
\$logs	26/11/2024 16:10:32	Privé	Disponible
ds-bronze	26/11/2024 16:12:05	Privé	Disponible
ds-gold	05/01/2025 22:20:25	Privé	Disponible
ds-silver	05/01/2025 20:57:48	Privé	Disponible

Les fichiers CSV ont été téléchargés dans le container **ds-bronze**, qui représente la **couche Bronze** pour stocker les données brutes.

The screenshot shows the Azure Storage Explorer interface for the 'ds-bronze' container. The left sidebar shows the navigation pane with 'ds-bronze' selected. The main pane displays a list of blob objects. The table below represents the data shown in the interface.

Nom	Dernière modification	Niveau d'accès	Type de blob	Taille	État du bail
[.]					
1- female-t...	01/12/2024 17:30:03	Élevé (déduit)	Objet blob de ...	1.59 KiB	Disponible
2- share-of-...	01/12/2024 17:30:03	Élevé (déduit)	Objet blob de ...	6.6 KiB	Disponible
3- ratio-of-f...	01/12/2024 17:30:03	Élevé (déduit)	Objet blob de ...	180.46 KiB	Disponible
6- gender-g...	01/12/2024 17:30:03	Élevé (déduit)	Objet blob de ...	9.43 KiB	Disponible

## 6.2 Databricks


Pour accéder à ces données dans **Azure Databricks**, les containers Blob Storage ont été **montés dans le Databricks File System (DBFS)**, permettant de les traiter comme des fichiers locaux. Le point de montage pour la couche Bronze était **/mnt/ds-bronze**.

```
dbutils.fs.unmount("/mnt/ds-bronze")
dbutils.fs.mount(
  source = f"wasbs://{container_name}@{storage_name}.blob.core.windows.net/",
  mount_point = mount_point_name,
  extra_configs = {
    f"fs.azure.account.key.{storage_name}.blob.core.windows.net": access_key
  }
)
```

```
fileCareWork = "dbfs:/mnt/ds-bronze/kaggle/current/1- female-to-male-ratio-of-time-devoted-to-unpaid-care-work.csv"
fileWomenInTop = "dbfs:/mnt/ds-bronze/kaggle/current/2- share-of-women-in-top-income-groups-atkinson-casarico-voitchovsky-2018.csv"
fileLaborForceParticipationRates = "dbfs:/mnt/ds-bronze/kaggle/current/3- ratio-of-female-to-male-labor-force-participation-rates-ilo-wdi.csv"
fileGenderGapWages = "dbfs:/mnt/ds-bronze/kaggle/current/6- gender-gap-in-average-wages-ilo.csv"
```

Le **cluster** sur Databricks a été configuré pour travailler avec **Apache Iceberg**.

Dans les **librairies**, a été ajouté (via Maven) :

<input type="checkbox"/>	Status	Name 	Type	Source
<input type="checkbox"/>	-	<a href="#">org.apache.iceberg:iceberg-spark-runtime-3.1_2.12:1.1.0</a>	Maven	-

Dans la **configuration Spark**, les paramètres suivants ont été ajoutés pour activer et configurer **Iceberg** :

### Spark config ⓘ

```
spark.sql.catalog.iceberg.warehouse dbfs:/mnt/ds-iceberg/
spark.sql.catalog.iceberg org.apache.iceberg.spark.SparkCatalog
spark.databricks.rocksDB.fileManager.useCommitService false
spark.sql.extensions
org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions
spark.sql.catalog.iceberg.type hadoop
```

## 6.3 Bronze

En utilisant **Spark**, les fichiers CSV téléchargés depuis le Blob Storage ont été chargés sous forme de **DataFrame** dans Databricks. Les fichiers ont été chargés avec les options `header="true"` et `inferSchema="true"` pour détecter automatiquement la structure des données.

```

▶ ✓ Yesterday (10s) 15

df = spark.read.format("csv").option("header", "true").option("inferSchema", "true").
load(FileGenderGapWages)

df.write.format("iceberg").mode("overwrite").saveAsTable("iceberg.bronze.
GenderGapWages")

▶ (3) Spark Jobs
▶ df: pyspark.sql.dataframe.DataFrame = [Entity: string, Code: string ... 2 more fields]

```

Après avoir chargé les données sous forme de DataFrame, elles ont été stockées dans des **tables Iceberg**. Ces tables ont été placées dans le schéma **Bronze** pour organiser les données brutes et les préparer pour un traitement ultérieur.

```

▶ ✓ Just now (1s) 17 SQL
%sql
SHOW TABLES IN iceberg.bronze;

```

▶ \_sqldf: pyspark.sql.dataframe.DataFrame = [database: string, tableName: string ... 1 more field]

Table ▾ + 🔍 🏠

	database	tableName	isTemporary
1	bronze	WomenInTop	false
2	bronze	CareWork	false
3	bronze	GenderGapWages	false
4	bronze	LaborForceParticipationRat...	false
5	bronze	MaleFemaleRatio	false









↓ 5 rows | 0.60 seconds runtime Refreshed now

*This result is stored as `_sqldf` and can be used in other Python cells.*

## 6.4 Silver















Les données de la couche Bronze ont été transférées vers la couche Silver, où des enregistrements non valides ont été nettoyés. Le nettoyage de chaque table (CareWork, WomenInTop, LaborForceParticipationRates, GenderGapWages) a été effectué dans des notebooks séparés, et un notebook distinct a été dédié à la gestion de la table de rejets (referential\_reject), le tout organisé dans le répertoire "silver\_steps" pour une gestion et traçabilité optimisées.

### silver\_steps

Name 	Type
 0.Create_table_reject	Notebook
 1.CareWork	Notebook
 2.WomenInTop	Notebook
 3.LaborForceParticipationRates	Notebook
 4.GenderGapWages	Notebook
 5.Create_continent_referential	Notebook
 6.Add_continent_to_table	Notebook

Afin d'organiser le processus de nettoyage, une table **referential\_reject** a été créée. Cette table documente les enregistrements problématiques des données brutes, en précisant :

- IdReject : Identifiant unique du rejet.
- targetTable : Table cible d'où provient le rejet.
- rejectCause : Raison du rejet.

Table  					  	
	  IdR...	 	  targetTable	  rejectCause		
1		3003	LaborForceParticipationRat...	income		
2		1001	CareWork	ratio nul		
3		2001	WomenInTop	données avant 2000		
4		2002	WomenInTop	données après 2012		
5		2003	WomenInTop	aucune données sur les parts des femm...		
6		3001	LaborForceParticipationRat...	ratio nul		
7		3002	LaborForceParticipationRat...	localisation inexploitable		
8		4001	GenderGapWages	pourcentage nul (0)		
 8 rows   2.41 seconds runtime					Refreshed 2 hours ago	

Dans les données nettoyées, des informations sur les continents ont été ajoutées pour les tables **CareWork**, **WomenInTop**, **LaborForceParticipationRates**, et **GenderGapWages**. Cela permet une analyse géographique plus détaillée.

▶ ✓ Yesterday (1s) 4

```
display(spark
  .table("iceberg.silver.CareWork"))
```

Table ▾ + 🔍 🔍 📄

	Entity	Code	Year	Ratio	IdContinent
1	Sweden	SWE	2014	1.49	3
2	Turkey	TUR	2014	6.22	2
3	Germany	DEU	2014	1.79	3
4	Cambodia	KHM	2014	4	2
5	France	FRA	2014	1.9	3
6	Algeria	DZA	2014	6.75	1
7	Argentina	ARG	2014	2.88	6

Une fois nettoyées, les données ont été sauvegardées sous forme de tables Iceberg.

▶ ✓ 12/24/2024 (2s) 5

```
schemaTable = spark.read.format("iceberg").load("iceberg.silver.GenderGapWages").schema

truncate_df = spark.createDataFrame([], schema=schemaTable)
truncate_df.write.format("iceberg").mode("overwrite").save("iceberg.silver.
GenderGapWages")
```

▶ 📄 truncate\_df: pyspark.sql.dataframe.DataFrame = [Entity: string, Code: string ... 2 more fields]

Les données nettoyées et transformées dans Databricks sont sauvegardées manuellement dans **Azure Data Lakehouse** au format **Parquet**. Elles sont stockées dans la couche **Silver** sous forme de **tables Iceberg** pour une gestion efficace des données.



Accueil > dlkefreijettevoueto

dlkefreijettevoueto | Navigateur de stockage

Compte de stockage

Rechercher

Vue d'ensemble

Journal d'activité

Étiquettes

Diagnostiquer et résoudre les problèmes

Contrôle d'accès (IAM)

Migration des données

Événements

**Navigateur de stockage**

Solutions de partenaire

Stockage des données

Conteneurs

Partages de fichiers

dlkefreijettevoueto

Favoris

Consultés récemment

Conteneurs de blobs

Slogs

ds-bronze

ds-gold

**ds-silver**

Tout afficher

Partages de fichiers

Files d'attente

Tables

Ajouter un répertoire

Charger

Actualiser

Supprimer

Copier

Coller

Renommer

Acquérir le bail

Conteneurs de blobs > ds-silver > kaggle

Méthode d'authentification : Clé d'accès (Basculer vers le compte d'utilisateur Microsoft Entra)

Rechercher les objets blobs par préfixe (respect de la casse)

Afficher uniquement les objets actifs

Affichage de tous les éléments 3

	Nom	Dernière modification	Niveau d'accès	Type de blob	Taille	État du bail
<input type="checkbox"/>	[.]					...
<input type="checkbox"/>	current	08/12/2024 13:23:20				...
<input type="checkbox"/>	data	10/12/2024 16:16:50				...
<input type="checkbox"/>	reject	10/12/2024 16:16:59				...

Accueil > dlkefreijettevoueto

dlkefreijettevoueto | Navigateur de stockage

Compte de stockage

Rechercher

Vue d'ensemble

Journal d'activité

Étiquettes

Diagnostiquer et résoudre les problèmes

Contrôle d'accès (IAM)

Migration des données

Événements

**Navigateur de stockage**

Solutions de partenaire

Stockage des données

Conteneurs

Partages de fichiers

dlkefreijettevoueto

Favoris

Consultés récemment

Conteneurs de blobs

Slogs

ds-bronze

ds-gold

**ds-silver**

Tout afficher

Partages de fichiers

Files d'attente

Tables

Ajouter un répertoire

Charger

Actualiser

Supprimer

Copier

Coller

Renommer

Acquérir le bail

Conteneurs de blobs > ds-silver > kaggle > data > CareWork

Méthode d'authentification : Clé d'accès (Basculer vers le compte d'utilisateur Microsoft Entra)

Rechercher les objets blobs par préfixe (respect de la casse)

Afficher uniquement les objets actifs

Affichage de tous les éléments 4

	Nom	Dernière modification	Niveau d'accès	Type de blob	Taille	État du bail
<input type="checkbox"/>	[.]					...
<input type="checkbox"/>	_SUCCESS	05/01/2025 21:42:04	Élevé (déduit)	Objet blob de ...	0	Disponible
<input type="checkbox"/>	_committed...	05/01/2025 21:42:00	Élevé (déduit)	Objet blob de ...	123 B	Disponible
<input type="checkbox"/>	_started_29...	05/01/2025 21:41:54	Élevé (déduit)	Objet blob de ...	0	Disponible
<input type="checkbox"/>	part-00000-...	05/01/2025 21:41:57	Élevé (déduit)	Objet blob de ...	3 KiB	Disponible

## 6.5 Gold

Dans la **couche Gold**, les **données nettoyées et transformées** de la couche **Silver** sont préparées et **optimisées** pour être utilisées dans des outils de **Business Intelligence** et des **rapports**. Les données ne sont pas agrégées, mais sont **structurées** et prêtes à être exploitées pour des analyses finales.

Dans la **couche Gold**, une **table des faits** est créée, contenant les **mesures détaillées** pour chaque pays, année et continent. Cette table inclure les champs suivants :

- Ratio femme-homme de la participation au marché du travail
- Écart salarial entre les sexes
- Ratio femme-homme du temps consacré au travail non rémunéré
- Part des femmes dans différents groupes de revenu (Top 0.1%, 0.25%, 0.5%, 1%, etc.)

```
%sql
CREATE TABLE IF NOT EXISTS iceberg.gold.dim_indicator (
  IndicatorID BIGINT NOT NULL,
  IndicatorName STRING NOT NULL,
  Unit STRING NOT NULL,
  Source STRING NOT NULL
);

INSERT INTO iceberg.gold.dim_indicator (IndicatorID, IndicatorName, Unit, Source)
VALUES
(1, 'Female to male ratio of time devoted to unpaid care work', 'proportion', 'OECD (2014)'),
(2, 'Share of women in top 0.1%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
(3, 'Share of women in top 0.25%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
(4, 'Share of women in top 0.5%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
(5, 'Share of women in top 1%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
(6, 'Share of women in top 5%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
(7, 'Share of women in top 10%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
(8, 'Ratio of female to male labor force participation rate', '%', 'ILO, WDI'),
(9, 'Gender wage gap', '%', 'ILO');
```

Les **tables de dimensions** dans la couche Gold permettent de **décrire** et de **lier** les **données factuelles** en fonction de différents **attributs**. Les dimensions sont les suivants :

- DimensionEntity : Cette dimension décrit les pays présents dans les faits.
- DimensionYear : Cette dimension contient les années associées aux mesures.
- DimensionContinent : Permet de regrouper les données par continents.
- DimensionIndicator : Cette dimension décrit les différents indicateurs (par exemple, participation au marché du travail, écart salarial) analysés dans les faits.

```
%sql
CREATE TABLE IF NOT EXISTS iceberg.gold.dim_indicator (
  IndicatorID BIGINT NOT NULL,
  IndicatorName STRING NOT NULL,
  Unit STRING NOT NULL,
  Source STRING NOT NULL
);

INSERT INTO iceberg.gold.dim_indicator (IndicatorID, IndicatorName, Unit, Source)
VALUES
  (1, 'Female to male ratio of time devoted to unpaid care work', 'proportion', 'OECD (2014)'),
  (2, 'Share of women in top 0.1%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
  (3, 'Share of women in top 0.25%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
  (4, 'Share of women in top 0.5%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
  (5, 'Share of women in top 1%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
  (6, 'Share of women in top 5%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
  (7, 'Share of women in top 10%', '%', 'Atkinson, Casarico and Voitchovsky (2018)'),
  (8, 'Ratio of female to male labor force participation rate', '%', 'ILO, WDI'),
  (9, 'Gender wage gap', '%', 'ILO');
```

Les données de la couche Silver sont manuellement enregistrées dans la couche **Gold** au format **Parquet** dans **Azure Data Lakehouse**. Elles sont optimisées pour des analyses finales et stockées sous forme de **tables Iceberg** ou **Delta** pour garantir une gestion robuste et des performances optimales.

The image displays two screenshots of the Microsoft Azure Storage Explorer interface, showing the 'ds-gold' container within the 'dlkefreijlettevoueto' storage account.

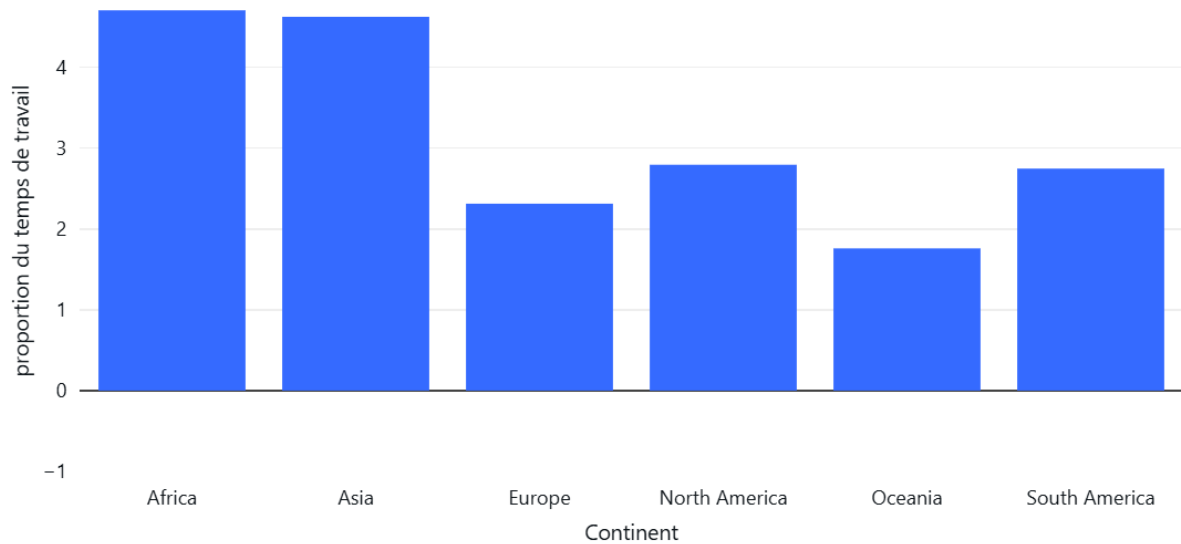
**Top Screenshot:** The interface shows the 'Conteneurs de blobs' view for the 'ds-gold' container. The breadcrumb path is 'Conteneurs de blobs > ds-gold > dimensions'. The table lists the following items:

Nom	Dernière modification	Niveau d'accès	Type de blob	Taille	État du bail
[.]					...
dim_contine...	05/01/2025 23:42:06				...
dim_entity	05/01/2025 23:41:42				...
dim_indicator	05/01/2025 23:41:17				...
dim_year	05/01/2025 23:40:52				...

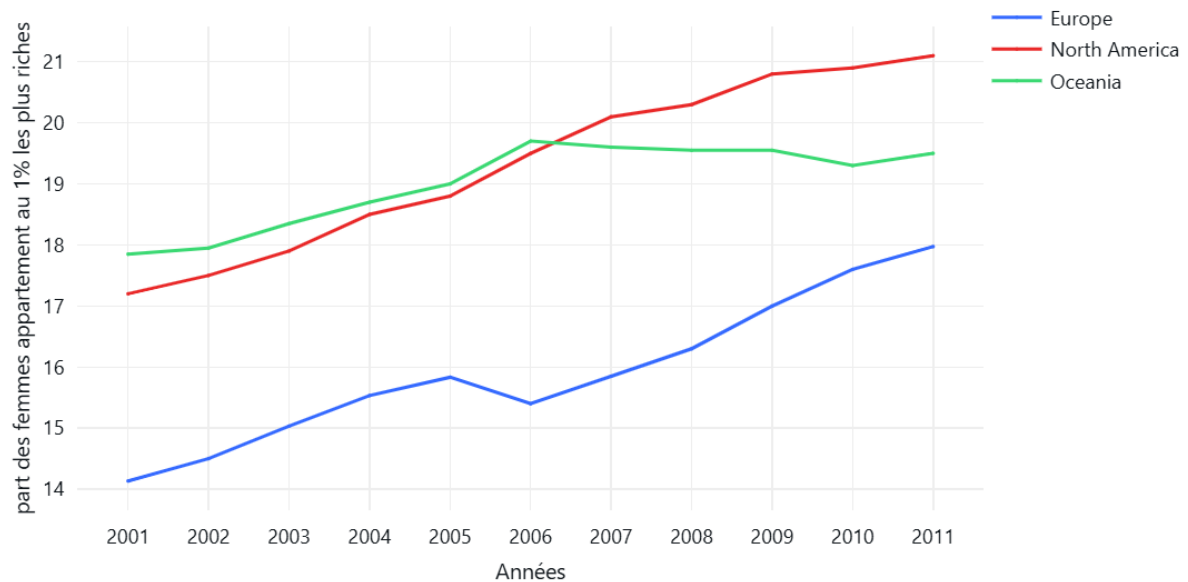
**Bottom Screenshot:** The interface shows the 'Conteneurs de blobs' view for the 'ds-gold' container. The breadcrumb path is 'Conteneurs de blobs > ds-gold > fact'. The table lists the following items:

Nom	Dernière modification	Niveau d'accès	Type de blob	Taille	État du bail
[.]					...
_SUCCESS	05/01/2025 23:36:37	Élevé (déduit)	Objet blob de ...	0	Disponible
_committed...	05/01/2025 23:36:33	Élevé (déduit)	Objet blob de ...	524 B	Disponible
_started_76...	05/01/2025 23:36:27	Élevé (déduit)	Objet blob de ...	0	Disponible
part-00000...	05/01/2025 23:36:29	Élevé (déduit)	Objet blob de ...	135.95 KiB	Disponible
part-00001...	05/01/2025 23:36:30	Élevé (déduit)	Objet blob de ...	8.88 KiB	Disponible
part-00002...	05/01/2025 23:36:30	Élevé (déduit)	Objet blob de ...	8.34 KiB	Disponible
part-00003...	05/01/2025 23:36:29	Élevé (déduit)	Objet blob de ...	6.4 KiB	Disponible
part-00004...	05/01/2025 23:36:29	Élevé (déduit)	Objet blob de ...	2.47 KiB	Disponible

## 6.6 Visualisation



Analyse: Proportion du temps de travail non rémunéré des femmes par rapport aux hommes (par exemple si la valeur est de 2, cela signifie que les femmes font en moyenne deux fois plus de soins non rémunérés par rapport aux hommes).



Analyse : On constate une augmentation du nombre de femmes qui appartenant aux 1% des

personnes ayant des revenus les plus élevés entre 2001 et 2011, surtout en Europe, même si on reste encore loin de la parité.

## 7 Conclusion

Nous pouvons constater que les femmes fournissent encore beaucoup de soins non rémunérés, surtout dans les pays d'Afrique et d'Asie, mais qu'il y a une légère augmentation des salaires les mieux payés, surtout dans les pays où les femmes fournissent un peu moins de soins non rémunérés.