

SATURDAY, AUGUST 12, 2023

TALEND

ETL

REPORT

**M. ABDERRAHIM
ELOUTMADI**

AYMANE SABRI

DATA DEVELOPER

Objectif du projet :

L'objectif principal de ce projet est de développer une compréhension approfondie des concepts fondamentaux de l'intégration de données en utilisant l'outil Talend. Nous nous sommes engagés à mettre en œuvre un scénario d'extraction, de transformation et de chargement (ETL)

Ce projet vise à démontrer notre capacité à concevoir et à mettre en œuvre un flux de travail ETL efficace, tout en traitant les défis courants tels que les nettoyages de données et l'optimisation des performances.

I. Introduction :

1. Contexte de projet :

L'explosion de la quantité de données disponibles, qu'elles soient structurées ou non, a mis en évidence la nécessité d'adopter des solutions d'intégration de données sophistiquées. Dans ce contexte, notre projet vise à explorer les fonctionnalités d'intégration de données en Talend en concevant un flux de travail **ETL** simple mais représentatif.

L'objectif principal de ce travail se résume à “ **ETL Talend**”

- Extraction Données .
- Transformation Des Données .
- Chargement Des Données .



2. Planification du plan de réalisation du brief :

Dans l'objectif de bien mener le **Brief**, j'ai commencé par établir le **planning** à suivre durant la période de brief .

Pour ce faire, j'ai d'abord décomposé mon projet en **phases**, où chaque phase est définie par un certain nombre de tâches. Ensuite, j'ai élaboré une planification de ces phases sur la durée du projet, à l'aide d'un diagramme de Gantt.

2.1. Etapes Suivie :

Les **étapes** suivies sont :

- **Comprendre** la nature et l'étendue du travail demandé.
- **Identifier** le type de recherche d'information demandé.
- Comprendre les **objectifs** d'apprentissage visés par le projet et les relier à la matière de l'étude.
- **Adopter** la démarche logique pour exécuter le travail.
- **Prêter** attention aux consignes et aux critères d'évaluation du projet (indiqués par écrit afin d'éviter toute erreur d'interprétation).
- Connaître les échéances et avoir l'intention de les respecter.

Suite à ces étapes j'ai identifié les besoins à satisfaire, défini l'aspect fonctionnel de projet et sa conception, réalisé le système et finalement je l'ai soumis à plusieurs **tests** pour s'assurer de son adaptation aux **besoins** exprimés précédemment.



2.2. Diagramme de Gant :

Ce diagramme représente la durée de chaque tâche effectué dans mon projet.

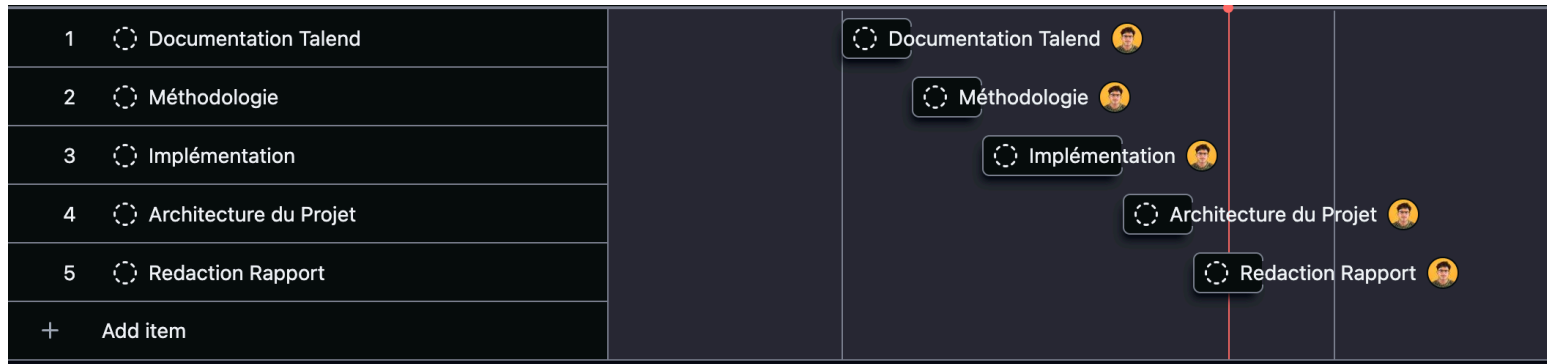


Figure : Diagramme de Gant

La portée du projet couvre les étapes fondamentales de l'ETL : de la configuration de la source de données à la conception des transformations, en passant par le chargement dans une base de données. Nous abordons également les considérations liées à la qualité des données et à la documentation du flux de travail ETL .

II. Methodologie :

1. Contexte de projet :

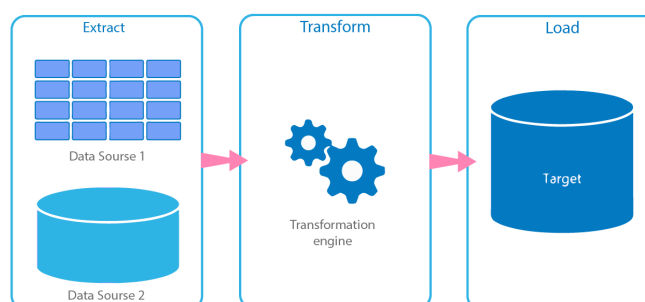
Dans cette section, nous détaillerons la méthodologie que j'ai suivie pour la réalisation de projet d'intégration de données avec Talend. Les différentes étapes, de la sélection de la source de données à la mise en place du processus [ETL](#), seront exposées en détail.

2. Conception de la Transformation des Données :

La conception de la transformation des données était une phase cruciale pour garantir que les données extraites soient adaptées à la destination souhaitée. Nous avons élaboré un plan détaillé pour chaque étape de transformation nécessaire, en prenant en compte les règles métier et les exigences spécifiques du projet. Les composants de transformation de Talend ont été utilisés pour nettoyer, enrichir et restructurer les données en fonction du modèle demandés .

3. Processus d'Extraction, de Transformation et de Chargement (ETL) :

Le cœur du projet résidait dans le processus [ETL](#), où nous avons mis en œuvre les étapes précédemment définies de manière cohérente et efficace. À l'aide des fonctionnalités de conception visuelle de Talend, nous avons créé des flux de travail [ETL](#) en reliant les différentes étapes, de la lecture initiale des données à la chargement final dans la base de données cible.



II. Architecture Du Projet :

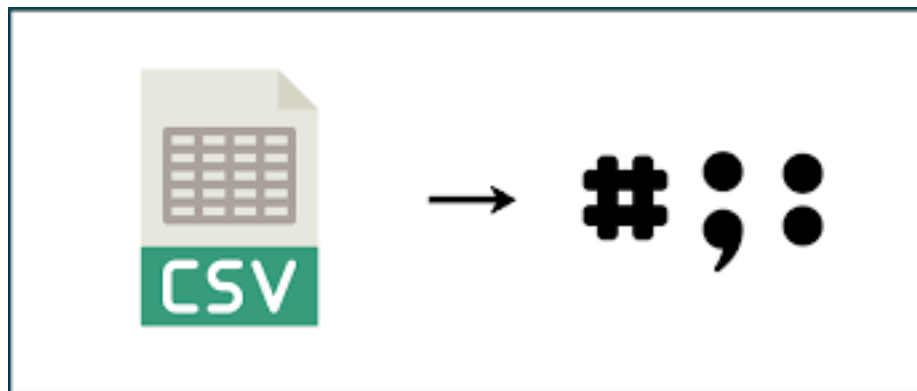
1. Vue d'Ensemble de l'Architecture ETL :

L'architecture **ETL** que nous avons adoptée pour ce projet suit un modèle classique de flux de données. Les données sont **extraites** de la source, subissent des **transformations** appropriées, puis sont chargées dans la base de données cible. Nous avons organisé notre flux de travail en différentes étapes, chacune étant gérée par des composants spécifiques de Talend. Cela a permis une séparation claire des responsabilités et une gestion plus efficace des données tout au long du processus **ETL**.

2. Description des Composants Talend Utilisés :

Nous avons utilisé une variété de composants Talend pour mettre en œuvre notre flux de travail **ETL**. Parmi les composants clés, citons :

- **tInputFileDelimited** : Ce composant est utilisé pour lire des fichiers délimités (comme CSV, TSV) à partir d'une source. Il permet d'extraire les données du fichier et de les transformer en enregistrements utilisables pour les étapes suivantes du flux de travail.



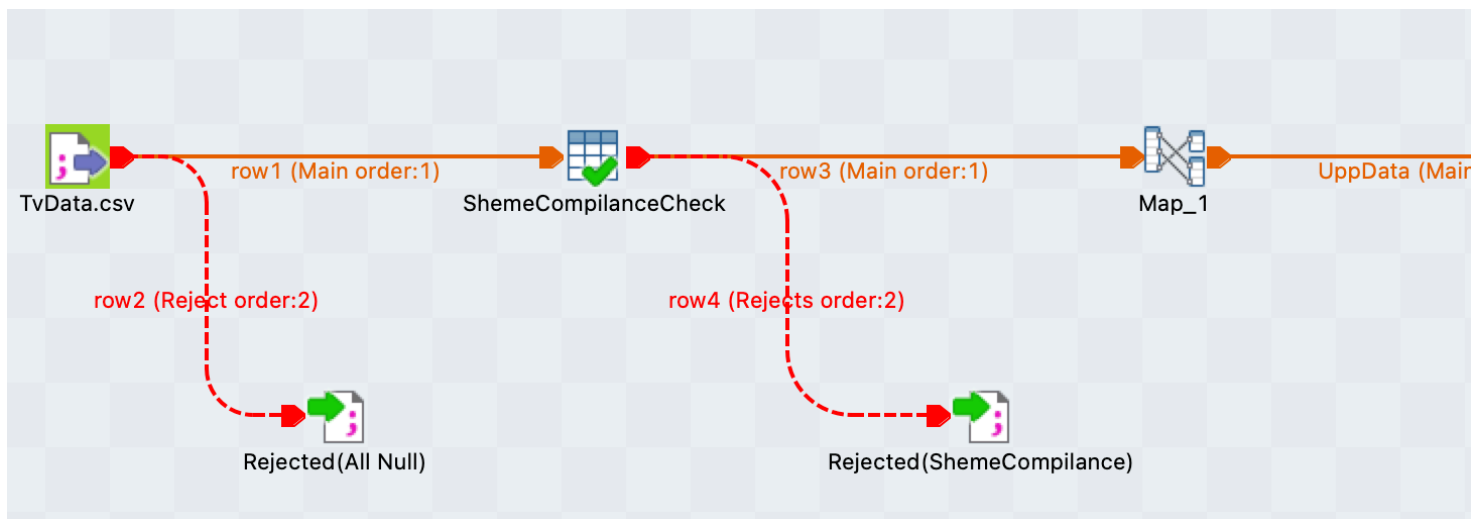
- **tMap** : Le composant tMap est au cœur de la transformation des données. Il permet de créer des règles de transformation visuelles pour nettoyer, filtrer, enrichir ou restructurer les données. Les règles sont appliquées à chaque ligne d'entrée, produisant ainsi une sortie modifiée en fonction des conditions et des opérations définies.
- **tAggregateRow** : Ce composant est utilisé pour agréger les données en fonction de certaines colonnes clés et d'opérations d'agrégation telles que la somme, la moyenne, le comptage, etc. Il vous permet de créer des résumés statistiques à partir de vos données
- **tReplicate** : Le composant tReplicate permet de diviser le flux de données en plusieurs copies identiques. Cela peut être utile lorsque vous souhaitez appliquer différentes transformations aux mêmes données ou les envoyer vers plusieurs destinations.
- **tDBOutput** : Ce composant est utilisé pour charger les données transformées dans une base de données cible. Il peut effectuer des opérations d'insertion, de mise à jour ou de suppression en fonction des actions définies. Il est souvent utilisé à la fin du flux de travail ETL pour persister les données transformées.

III. Implémentation :

Dans cette section, nous allons explorer en détail l'implémentation pratique de notre projet d'intégration de données en utilisant [Talend](#). Nous aborderons la configuration de la connexion à la source de données, la [transformation](#) des données à l'aide des composants [Talend](#) et le chargement des données transformées dans la base de données cible.

1. Configuration de la Connexion à la Source de Données :

Nous avons d'abord configuré la connexion à la source de données à l'aide du composant `tnputFileDelimited`. Nous avons spécifié le chemin du fichier source, ainsi que les délimiteurs appropriés (virgules) en fonction du format CSV. Cette étape a permis à Talend de lire les données brutes à partir de la source et de les préparer pour les transformations ultérieures.



File – Step 3 of 4

Add a Metadata File on repository
Define the setting of the parse job



File Settings

Encoding US-ASCII

Field Separator Comma Corresponding Character " "

Row Separator Standard EOL Corresponding Character "\r"

Escape Char Settings

☒ CSV ☐ Delimited

Escape Char Empty

Text Enclosure Empty

☐ Split row before field

Rows To Skip

If any rows must be ignored, specify the following parameters

Header ☒ 1

Footer ☐

☐ Skip empty row

Limit Of Rows

If the number of lines must be limited, specify this number

Limit ☐

Preview Output

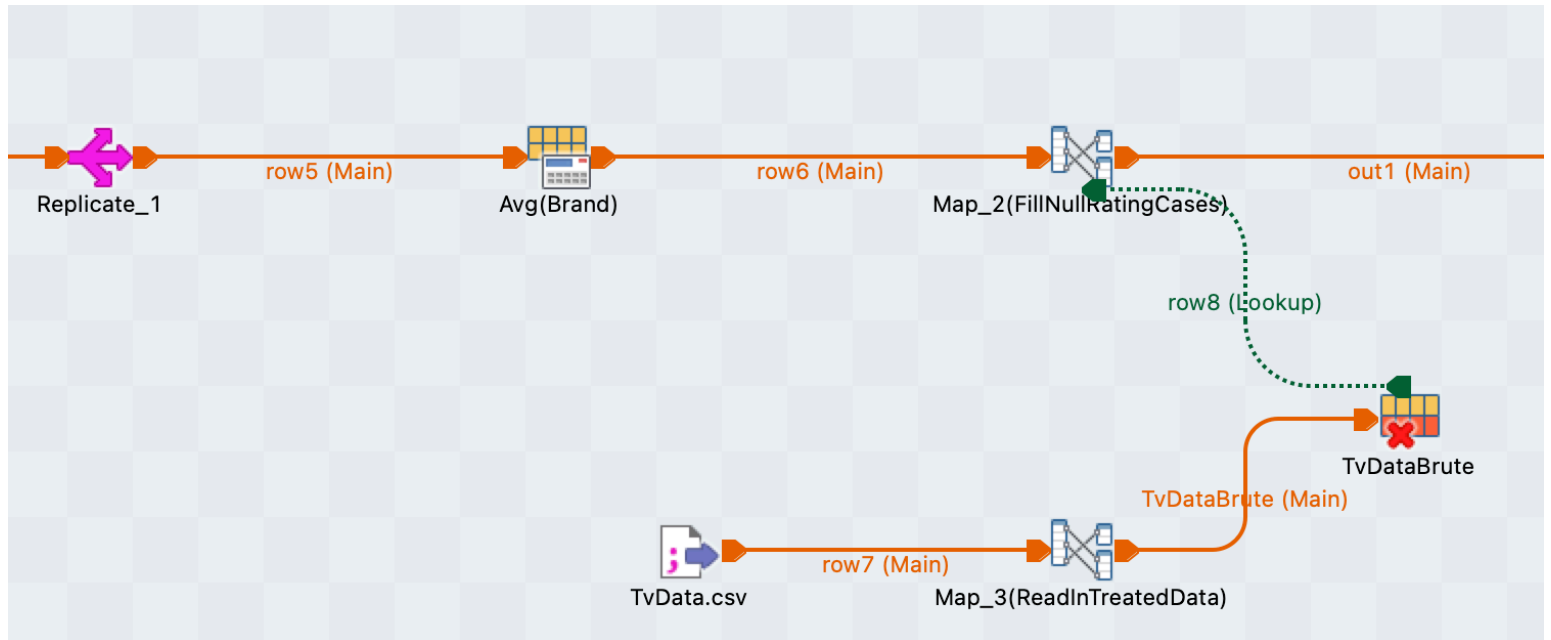
☒ Set heading row as column names Refresh Preview

| Brand | Resolution | Size | Selling Price | Original Price | Operating System | Rating |
|---------|---------------|------|---------------|----------------|------------------|--------|
| TOSHIBA | Ultra HD LED | 55 | 37999 | 54990 | VIDAA | 4.3 |
| TCL | QLED Ultra HD | 55 | 52999 | 129990 | Android | 4.4 |
| realme | HD LED | 32 | 13999 | 17999 | Android | 4.3 |
| Mi | HD LED | 32 | 14999 | 19999 | Android | 4.4 |
| realme | HD LED | 32 | 12999 | 21999 | Android | 4.3 |
| OnePlus | HD LED | 32 | 15999 | 19999 | Android | 4.3 |
| OnePlus | Full HD LED | 43 | 25999 | 29999 | Android | 4.3 |
| TCL | Ultra HD LED | 65 | 57999 | 119990 | Android | 4.2 |

Export as context Revert Context

2. Transformation des Données à l'Aide de Talend :

La transformation des données a été réalisée à l'aide des composant [tMap](#), [tAggregate](#) , [tUnique](#) ... où nous avons créé des règles de [transformation](#) pour nettoyer, normaliser les données. Par exemple, nous avons appliqué des filtres pour exclure les enregistrements inutiles(over 5 null values), effectué des opérations mathématiques pour calculer la moyenne (Rating). Cette étape a permis de préparer les données en vue de leur chargement dans la base de données cible.



| row6 | |
|--------|--|
| Column | |
| Brand | |
| Rating | |

| row8 | |
|-----------------|-------------|
| Property | Value |
| Lookup Model | Load once |
| Match Model | All matches |
| Join Model | Inner Join |
| Store temp data | false |

| Expr. key | Column |
|------------|------------------|
| row6.Brand | Brand |
| | Resolution |
| | Size |
| | Selling_Price |
| | Original_Price |
| | Operating_System |
| | Rating |

| row12 | |
|------------------|--|
| Column | |
| ID | |
| Brand | |
| Resolution | |
| Size | |
| Selling_Price | |
| Original_Price | |
| Operating_System | |
| Rating | |

| row15 | |
|-----------------|-----------------|
| Property | Value |
| Lookup Model | Load once |
| Match Model | Unique match |
| Join Model | Left Outer Join |
| Store temp data | false |

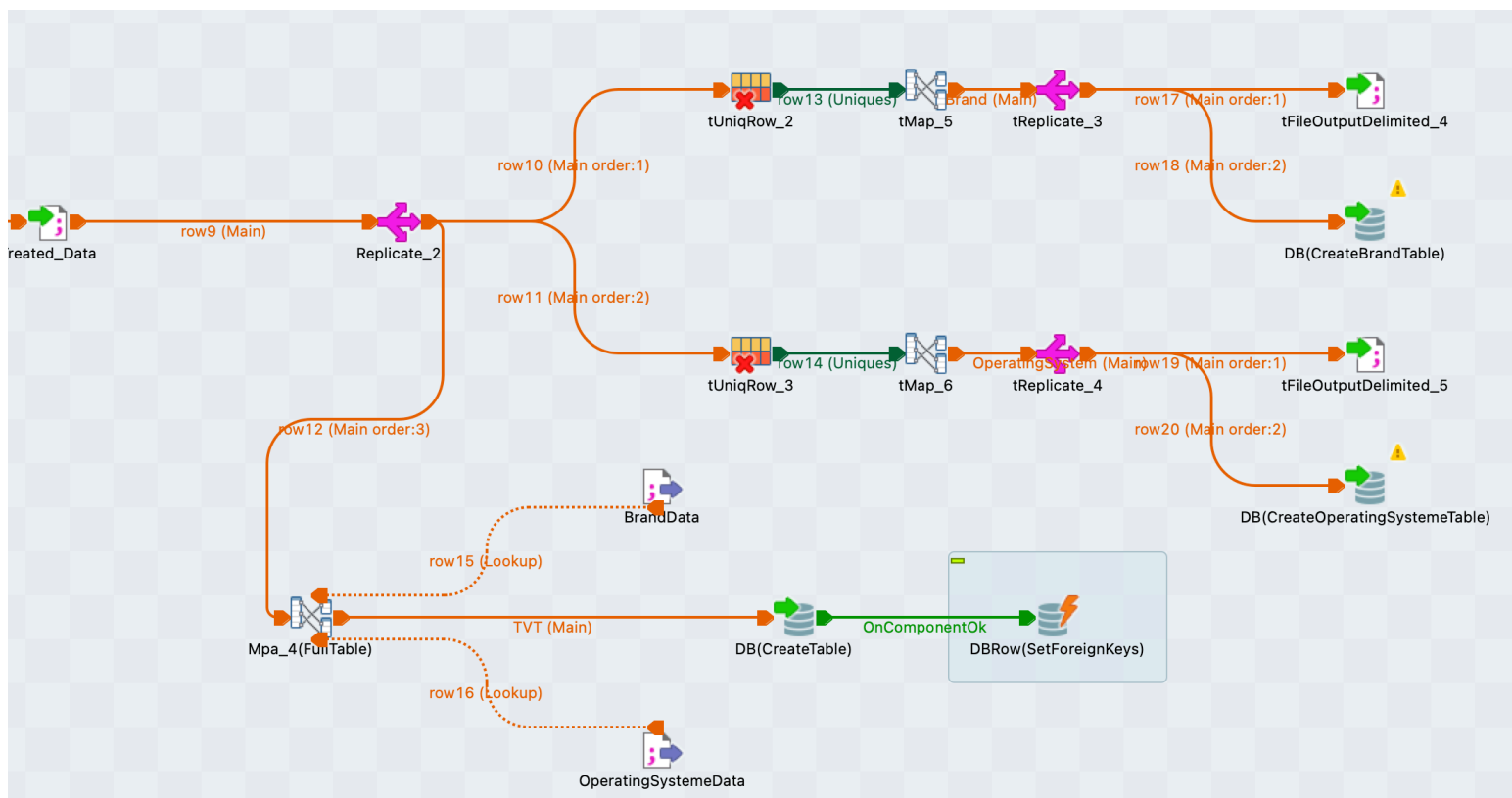
| Expr. key | Column |
|-------------|---------|
| row12.Brand | BrandID |
| | Brand |

| row16 | |
|------------------------|-------------------|
| Expr. key | Column |
| row12.Operating_System | OperatingSystemID |
| | Operating_System |

| TVT | |
|-------------------------|-------------------|
| Expression | Column |
| row12.ID | ID |
| row12.Resolution | Resolution |
| row12.Size | Size |
| row12.Selling_Price | Selling_Price |
| row12.Original_Price | Original_Price |
| row12.Rating | Rating |
| row15.BrandID | BrandID |
| row16.OperatingSystemID | OperatingSystemID |

3. Chargement des Données dans la Base de Données Cible:





























Une fois les données transformées, nous avons utilisé les composants **tDBOutput** , **tDbRow** ... pour charger les données dans la base de données cible. Nous avons configuré les paramètres de connexion à la base de données et indiqué les tables cibles où les données doivent être chargées. Selon les besoins, nous avons spécifié les actions à effectuer (insertion, mise à jour) et **mappé** les champs transformés aux colonnes de la base de données. Cette étape a permis de persister les données transformées dans un format adapté à l'analyse ultérieure.













































































4. Résultats et Performance

Une fois les données transformées et chargées dans la base de données cible, nous avons mené une analyse approfondie pour évaluer la qualité et la cohérence des données résultantes .



|  | | | | ID | Operating_System |
|---|--|--|--|----|------------------|
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 1 | Android |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 2 | Linux |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 3 | HomeOS |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 4 | Tizen |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 5 | FireTV OS |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 6 | WebOS |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 7 | Unknown |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 8 | VIDAA |
| <div><input type="checkbox"/> Check all</div> <div>With selected:  Edit  Copy</div> | | | | | |

| ←T→ | | | | | | ID | Brand |
|--------------------------|--|--|--|----|-----------------|----|-------|
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 1 | KODAK | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 2 | SHARP | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 3 | WESTON | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 4 | TCL | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 5 | AISEN | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 6 | SAMSUNG | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 7 | PANASONIC | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 8 | CROMA | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 9 | JVC | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 10 | THOMSON | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 11 | ONIDA | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 12 | LUMX | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 13 | IFFALCON BY TCL | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 14 | MI | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 15 | DYANORA | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 16 | LG | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 17 | CANDES | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 18 | NOKIA | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 19 | MOTOROLA | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 20 | IMPEX | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 21 | INTEX | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 22 | SANSUI | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 23 | PHILIPS | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 24 | HISENSE | | |
| <input type="checkbox"/> |  Edit |  Copy |  Delete | 25 | DEKTRON | | |

| ← T → | | | | ▼ ID | Resolution | Size | Selling_Price | Original_Price | Rating | BrandID | OperatingSystemID | |
|--------------------------|---|------|--|--|------------|--------------|---------------|----------------|--------|---------|-------------------|---|
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 1 | Ultra HD LED | 55 | 36999 | 47999 | 4.4 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 2 | HD LED | 32 | 12999 | 18499 | 4.4 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 3 | Full HD LED | 40 | 19499 | 20999 | 4.4 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 4 | Full HD LED | 42 | 20499 | 27999 | 4.4 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 5 | Ultra HD LED | 43 | 25999 | 37999 | 4.4 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 6 | Ultra HD LED | 55 | 34499 | 39990 | 4.3 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 7 | HD LED | 24 | 8499 | 10499 | 4 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 8 | Ultra HD LED | 50 | 32999 | 42999 | 4.4 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 9 | Ultra HD LED | 50 | 30999 | 33999 | 4.3 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 10 | HD LED | 32 | 11499 | 20990 | 4.3 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 11 | Full HD LED | 43 | 22499 | 24999 | 4.4 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 12 | HD LED | 32 | 11499 | 15999 | 4.1 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 13 | Ultra HD LED | 49 | 47990 | 50999 | 4.1 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 14 | Full HD LED | 40 | 15499 | 31990 | 4.3 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 15 | Full HD LED | 49 | 22999 | 31990 | 4.3 | 1 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 16 | Ultra HD LED | 50 | 37396 | 79990 | 3.35 | 2 | 2 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 17 | Full HD LED | 60 | 229900 | 229900 | 3.35 | 2 | 2 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 18 | Full HD LED | 46 | 151990 | 151990 | 3.35 | 2 | 2 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 19 | Full HD LED | 70 | 449900 | 449900 | 3.35 | 2 | 2 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 20 | Full HD LED | 52 | 199990 | 199990 | 3.35 | 2 | 2 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 21 | Full HD LED | 40 | 103990 | 103990 | 3.35 | 2 | 2 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 22 | HD LED | 32 | 15990 | 22500 | 3 | 2 | 2 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 23 | HD LED | 24 | 8220 | 14990 | 3.7 | 2 | 2 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 24 | Ultra HD LED | 55 | 59990 | 59990 | 4.2 | 3 | 1 |
| <input type="checkbox"/> |  | Edit |  Copy |  Delete | 25 | Full HD LED | 40 | 27990 | 27990 | 4.4 | 3 | 1 |