**Objetivo:**

Construir el proceso ETL utilizando Pentaho Data Integration (PDI) con el fin de obtener un archivo csv, realizando el proceso de extracción y transformación de los datos disponibles en la base de datos relacional BD Ames, en archivos csv y en MongoDB.

**Actividades:**

En grupos de 3 estudiantes, realizar las siguientes actividades:

* Entender las fuentes relacionales disponibles y su relación con el archivo esperado de salida del proceso ETL.

DB Ames . Postgres

* Entender las fuentes disponibles no relacionales y su relación con el archivo esperado de salida del proceso ETL.
* Realizar el proceso de extracción y transformación de datos de la base de datos relacional.
* Realizar el proceso de extracción y transformación de datos de las fuentes no relacionales.
* Obtener un archivo csv, con los datos definidos en [[detalle archivo](#_30j0zll)]. Cada fila del archivo es el detalle de una propiedad.

Un informe que incluya las actividades y conclusiones. Y el proyecto PDI.

**Fuentes de datos disponibles:**

A continuación, se definen las fuentes en las que se encuentran los datos:

|  |  |
| --- | --- |
| Fuentes | Descripción |
| Archivo [Ames Property.csv](https://drive.google.com/file/d/1tpiDI4KcBRH41rFjsBh3QmuqcOUKE6Bd/view?usp=drive_link) | Información general de la propiedad. (19 campos): 2930 registros  PID, Lot Frontage, Lot Area, Street, Alley, Lot Shape, Land Contour, Utilities, Lot Config, Land Slope, Neighborhood, Condition 1, Condition 2, Bldg Type, House Style, Overall Qual, Overall Cond, Year Built, Year Remod/Add |
| BD Relacional  [[Modelos de datos](https://docs.google.com/document/d/1ym-VP7RZ9SJ4_wEHsu1gMtjAosxwasTN8DLQjiPXYo4/edit?usp=sharing)]  [[Script DB](https://drive.google.com/file/d/156iclNsLGvf6XYmuzhp1MAjiPY26EfL6/view?usp=drive_link)]  [[ScriptDatos](https://drive.google.com/drive/folders/1bC-Ct8rAnUcJJzed9mELj156xJMKLPjZ?usp=sharing)] | **amesdbtemp** ( 35 columnas) : 2930 filas  Detalles relacionados con la construcción.  PID, MS SubClass, MS Zoning, Roof Style, Roof Matl, Exterior 1st, Exterior 2nd, Mas Vnr Type, Mas Vnr Area, Exter Qual, Exter Cond,Foundation,Heating, Heating QC,Central Air,Electrical, 1st Flr SF, 2nd Flr SF,Low Qual Fin SF, Kitchen AbvGr, Kitchen Qual, TotRms AbvGrd, Functional, Fireplaces, Fireplace Qu, Paved Drive, Wood Deck SF,Open Porch SF, Enclosed Porch, 3Ssn Porch, Screen Porch, Fence  **MSSubClass** ( 2 columnas) : 16 filas  Clase de construcción.  code, description  **MSZoning**( 3 columnas) : 8 filas  id, code, description  **TypeQuality** ( 3 columnas) : 5 filas  Tipos de calidad.  id, code, description  **SaleProperties** ( 5 columnas ) : 2930 filas  PID, Sale Date, Sale Type, Sale Condition, SalePrice  **FloorDetail** (5 columnas) : 4182 filas  Para cada propiedad por piso, se encuentra la cantidad de habitaciones y baños.  PID, Floor, Bedroom, Full Bath, Half Bath |
| MongoDB Atlas  [Datos conexión](https://docs.google.com/document/d/14NbPDg40BGuY0Y8KIH3CPpGZMAFUVUVd/edit?usp=drive_link&ouid=106441471743091147224&rtpof=true&sd=true) | Colecciones:   * **garage** ( 8 campos ) : 2773 documentos   PID, Garage Type, Garage Yr Blt, Garage Finish, Garage Cars, Garage Area, Garage Qual, Garage Cond   * **pool** (3 campos) : 13 documentos   PID, Pool Area, Pool QC   * **bsmt** ( 12 campos ) : 2851 documentos   PID, Bsmt Qual, Bsmt Cond, Bsmt Exposure, BsmtFin Type 1, BsmtFin SF 1, BsmtFin Type 2,BsmtFin SF 2, Bsmt Unf SF, Total Bsmt SF, Bsmt Full Bath, Bsmt Half Bath   * **misc** ( 3 campos) : 106 documentos   PID,Misc Feature, Misc Val  Nota: Al realizar el cargue de estos datos, si no se encuentran los datos, debe especificarse NA si es cualitativa y 0 si es numérica. |

**Transformaciones:**

Gr Liv Area: Pies cuadrados de superficie habitable sobre el nivel del suelo (suelo). Se calcula a partir de la suma de los campos: 1st Flr SF, 2nd Flr SF y Low Qual Fin SF.

FullBath, HalfBath, Bedroom: Se deben calcular teniendo en cuenta que una propiedad puede tener varios pisos. En la tabla FloorDetail se encuentran los datos de cantidad de habitaciones, y baños, por piso para cada propiedad. La transformación consiste en calcular el total de baños FullBath, HalfBath, Bedroom.

Garage, pool, bsmt, misc: Al realizar el cargue de estos datos (garage, pool, bsmt, misc), si no se encuentran los datos, debe especificarse NA si es cualitativa y 0 si es numérica.

Mo Sold,Yr Sold: En la tabla saleProperty se encuentra la fecha de venta y se requiere que en la salida esté solo mes (Mo Sold) y año de venta (Yr Sold).

Year Remod/Addcontiene la fecha de remodelación, si no tiene, en vez de dejarlo en Null colocar la misma fecha de construcción Year Built.

**Salida esperada** [[Ejemplo salida](https://drive.google.com/file/d/1v-SfD2MhFej7I8rFbScnu9Or8kpMG8qv/view?usp=drive_link)]:

1. **PID:** Property ID
2. **MSSubClass**: The building class
3. **MSZoning**: The general zoning classification
4. **LotFrontage**: Linear feet of street connected to property
5. **LotArea**: Lot size in square feet
6. **Street**: Type of road access
7. **Alley**: Type of alley access
8. **LotShape**: General shape of property
9. **LandContour**: Flatness of the property
10. **Utilities**: Type of utilities available
11. **LotConfig**: Lot configuration
12. **LandSlope**: Slope of property
13. **Neighborhood**: Physical locations within Ames city limits
14. **Condition1**: Proximity to main road or railroad
15. **Condition2**: Proximity to main road or railroad (if a second is present)
16. **BldgType**: Type of dwelling
17. **HouseStyle**: Style of dwelling
18. **OverallQual**: Overall material and finish quality
19. **OverallCond**: Overall condition rating
20. **YearBuilt**: Original construction date
21. **YearRemodAdd**: Remodel date
22. **RoofStyle**: Type of roof
23. **RoofMatl**: Roof material
24. **Exterior1st**: Exterior covering on house
25. **Exterior2nd**: Exterior covering on house (if more than one material)
26. **MasVnrType**: Masonry veneer type
27. **MasVnrArea**: Masonry veneer area in square feet
28. **ExterQual**: Exterior material quality
29. **ExterCond**: Present condition of the material on the exterior
30. **Foundation**: Type of foundation
31. **BsmtQual**: Height of the basement
32. **BsmtCond**: General condition of the basement
33. **BsmtExposure**: Walkout or garden level basement walls
34. **BsmtFinType1**: Quality of basement finished area
35. **BsmtFinSF1**: Type 1 finished square feet
36. **BsmtFinType2**: Quality of second finished area (if present)
37. **BsmtFinSF2**: Type 2 finished square feet
38. **BsmtUnfSF**: Unfinished square feet of basement area
39. **TotalBsmtSF**: Total square feet of basement area
40. **Heating**: Type of heating
41. **HeatingQC**: Heating quality and condition
42. **CentralAir**: Central air conditioning
43. **Electrical**: Electrical system
44. **1stFlrSF**: First Floor square feet
45. **2ndFlrSF**: Second floor square feet
46. **LowQualFinSF**: Low quality finished square feet (all floors)
47. **GrLivArea**: Above grade (ground) living area square feet
48. **BsmtFullBath**: Basement full bathrooms
49. **BsmtHalfBath**: Basement half bathrooms
50. **FullBath**: Full bathrooms above grade
51. **HalfBath**: Half baths above grade
52. **Bedroom**: Number of bedrooms above basement level
53. **Kitchen**: Number of kitchens
54. **KitchenQual**: Kitchen quality
55. **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
56. **Functional**: Home functionality rating
57. **Fireplaces**: Number of fireplaces
58. **FireplaceQu**: Fireplace quality
59. **GarageType**: Garage location
60. **GarageYrBlt**: Year garage was built
61. **GarageFinish**: Interior finish of the garage
62. **GarageCars**: Size of garage in car capacity
63. **GarageArea**: Size of garage in square feet
64. **GarageQual**: Garage quality
65. **GarageCond**: Garage condition
66. **PavedDrive**: Paved driveway
67. **WoodDeckSF**: Wood deck area in square feet
68. **OpenPorchSF**: Open porch area in square feet
69. **EnclosedPorch**: Enclosed porch area in square feet
70. **3SsnPorch**: Three season porch area in square feet
71. **ScreenPorch**: Screen porch area in square feet
72. **PoolArea**: Pool area in square feet
73. **PoolQC**: Pool quality
74. **Fence**: Fence quality
75. **MiscFeature**: Miscellaneous feature not covered in other categories
76. **MiscVal**: $Value of miscellaneous feature
77. **MoSold**: Month Sold
78. **YrSold**: Year Sold
79. **SaleType**: Type of sale
80. **SaleCondition**: Condition of sale
81. **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.

**Neighborhood:** la salida de este campo está organizado por el código y la descripción, de la siguiente forma.

|  |  |
| --- | --- |
| code | description |
| Blmngtn | Bloomington Heights |
| Blueste | Bluestem |
| BrDale | Briardale |
| BrkSide | Brookside |
| ClearCr | Clear Creek |
| CollgCr | College Creek |
| Crawfor | Crawford |
| Edwards | Edwards |
| Gilbert | Gilbert |
| IDOTRR | Iowa DOT and Rail Road |
| MeadowV | Meadow Village |
| Mitchel | Mitchell |
| Names | North Ames |
| NoRidge | Northridge |
| NPkVill | Northpark Villa |
| NridgHt | Northridge Heights |
| NWAmes | Northwest Ames |
| OldTown | Old Town |
| SWISU | South & West of Iowa State University |
| Sawyer | Sawyer |
| SawyerW | Sawyer West |
| Somerst | Somerset |
| StoneBr | Stone Brook |
| Timber | Timberland |
| Veenker | Veenker |

**Condition 1,** en este campo se maneja la siguiente información:

|  |  |
| --- | --- |
| code | description |
| Artery | Adjacent to arterial street |
| Feedr | Adjacent to feeder street |
| Norm | Normal |
| RRNn | Within 200' of North-South Railroad |
| RRAn | Adjacent to North-South Railroad |
| PosN | Near positive off-site feature--park greenbelt etc. |
| PosA | Adjacent to postive off-site feature |
| RRNe | Within 200' of East-West Railroad |
| RRAe | Adjacent to East-West Railroad |

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

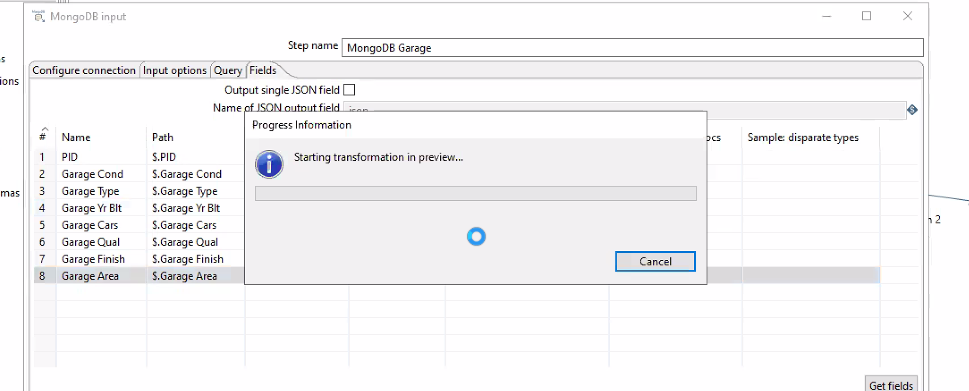
IR2 Moderately Irregular

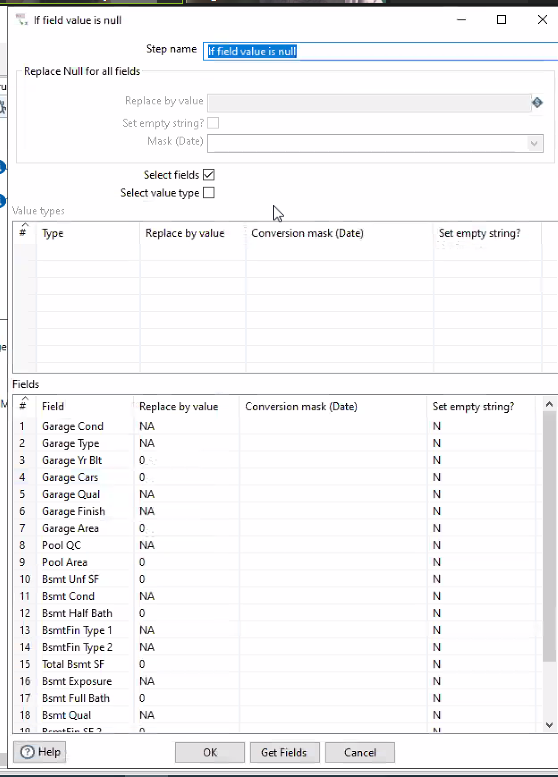
IR3 Irregular

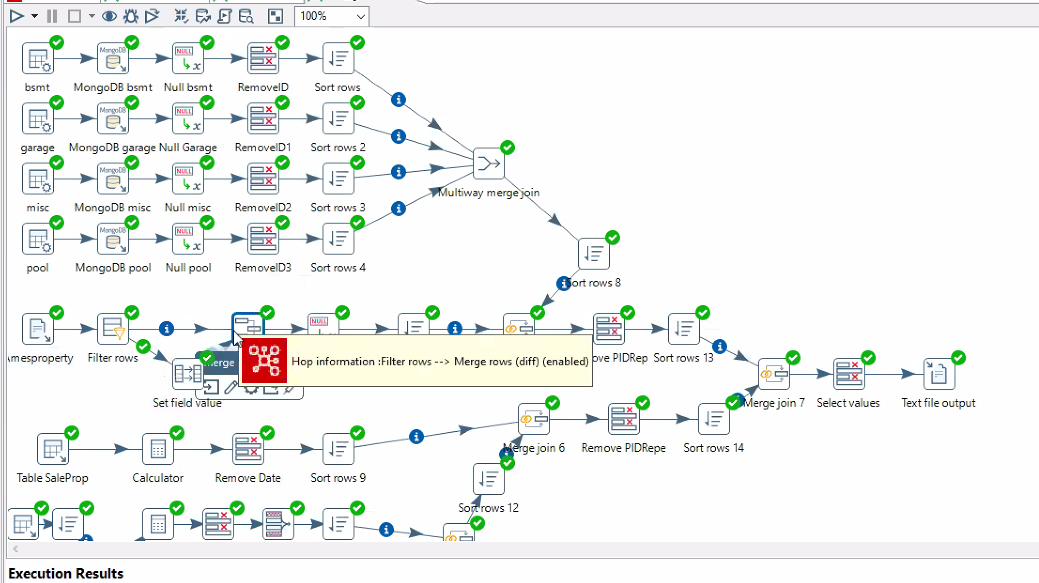
Referencias:

Canal de vídeos sobre uso de PDI

<https://www.youtube.com/watch?v=o7If1a-gkyI&list=PLPgjON4ZM0JBMHkH8JpNbdg0UjdZ8g1Bk>







Reporte 5 Paginas

Nombres,

Introduccion

Descripcion fuentes

ETLS por fuentes

Actividades realizadas

Investigaciones

Lecciones aprendidas

Referencias

Pentaho : https://www.hitachivantara.com/en-us/products/pentaho-plus-platform/data-integration-analytics/pentaho-tutorials.html#getting-started

Cheat sheet : <https://www.postgresqltutorial.com/wp-content/uploads/2018/03/PostgreSQL-Cheat-Sheet.pdf>

Docs: https://pentaho-public.atlassian.net/wiki/spaces/EAI/pages/376442127/If+field+value+is+null