CIS 4400

Protoy Saha

HW 1

04/31/2025

**Business Requirements**:

* Predict Real Estate Investment ROI: Use the NYC data to predict where it would be the best to invest in to get the most out of the investment.
* Predict Housing Prices: Create a model to predict property prices based on property related attributes such as zip code, square footage, etc.
* Find Market Trends: Analyze trends in specific regions that can help drive real estate investment.

**Functional Requirements**:

* Data Ingestion & Cleaning: Importing real estate data, cleaning missing values or NULL values, and standardizing the various data fields.
* EDA: Begin with the summary statistics, histograms, and visual trends in housing sales.
* ROI Prediction Model: Create and implement a machine learning algorithm to predict real estate investment ROI.
* Trend Analysis: Create reports on top performing zip codes, cities, and municipalities, and type of property in demand.
* Data Visualization: Build dashboards to allow users to interact with the data to their liking and task requirements to explore trends or filter the data.

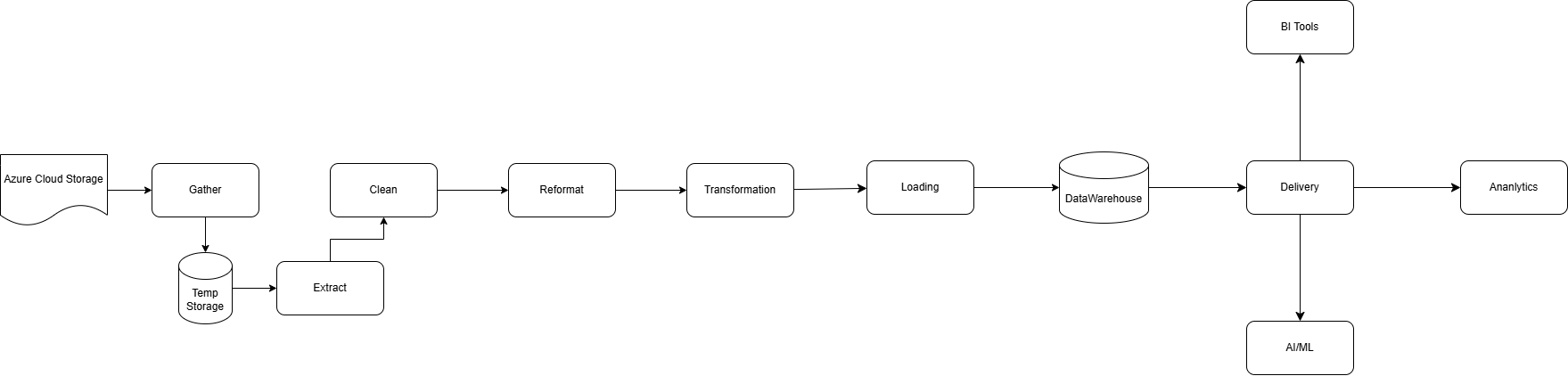
**Data Requirements**:

* NYC Real Estate Data
* Source: [DoltHub](https://www.dolthub.com/repositories/dolthub/us-housing-prices-v2/data/main)
* Meets all the necessary requirements of the project.

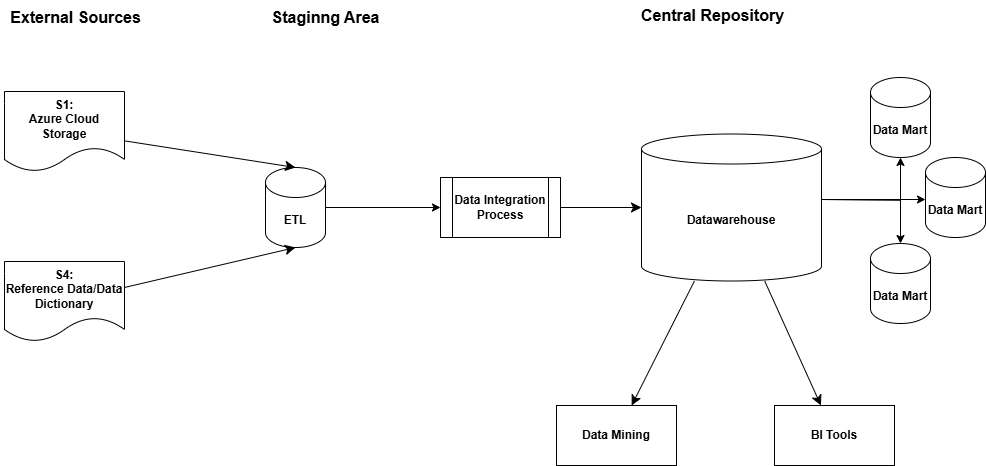
**Data Sourcing**:

* Connection string via Azure cloud storage.

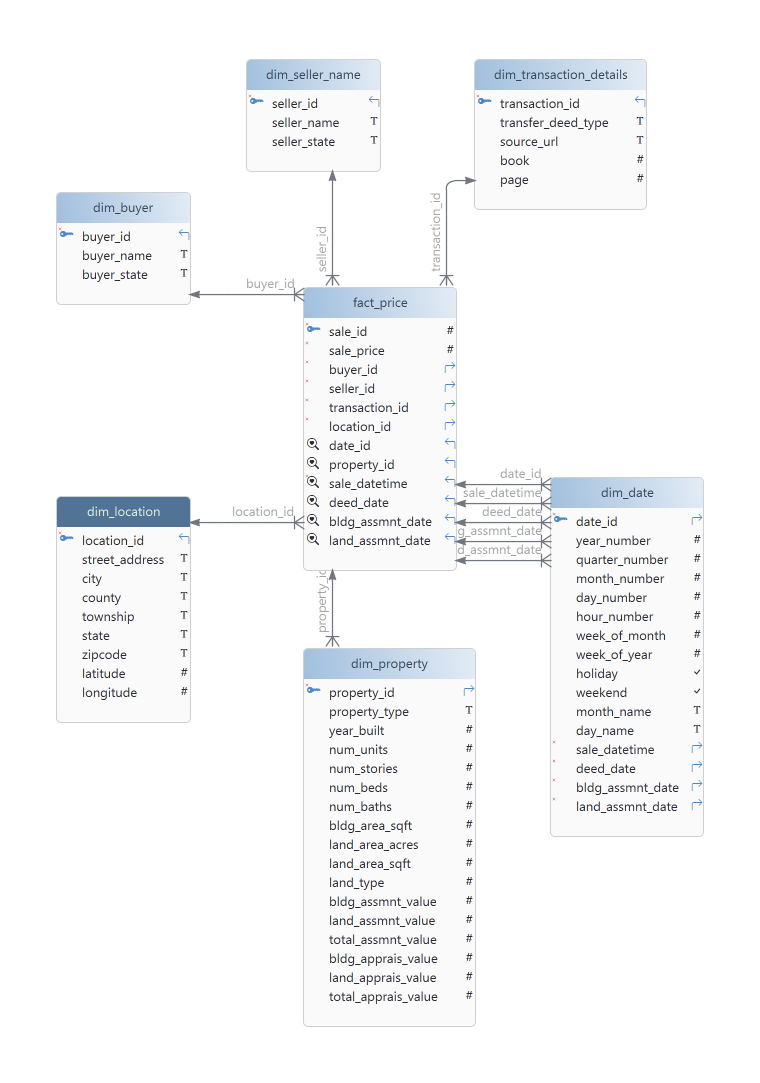
**Information Architecture**:

* 
* The data is streamed via a connection string from Azure cloud storage to my Azure cloud storage. My initial cloud storage will act as temporary storage. Thereafter, it will go through ELT process and put into the data warehouse. From there, it will go to the BI tool for data visualization.

**Data Architecture**:

* 
* The data arrives from the Azure cloud storage via a connection string to my Azure cloud storage. Additionally, a 2nd source is used as a data dictionary to provide metadata for the data. Later the data will go through a data integration process. The data will be stored in a top-down process. Where all of the data will be kept in a data warehouse and specific data as required for the business requirements will be stored in data marts.

**Dimensional Modeling**:

* 

**Data Dictionary**:

[CIS 4400 HW1 Data Dictionary](https://baruchmailcuny-my.sharepoint.com/:x:/g/personal/protoy_saha_baruchmail_cuny_edu/EeHVbPIDvjtHrhrJ_7V2QA8BPG-aFfa_wtxoTM37Cx8Suw?e=CiJX8d)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| buyer\_id | INT |  | PK | Surrogate key for the buyer dimension. | 1010 |  |
| buyer\_1\_name | VARCHAR | 1024 |  | Name of the primary buyer. | John Doe | buyer\_1\_name |
| buyer\_1\_state | ENUM |  |  | Two-letter state code of the primary buyer. | AK | buyer\_1\_state |
| buyer\_2\_name | VARCHAR | 1024 |  | Name of the secondary buyer. | John Smith | buyer\_2\_name |
| buyer\_2\_state | ENUM |  |  | Two-letter state code of the secondary buyer. | IL | buyer\_2\_state |
| seller\_id | INT |  | PK | Surrogate key for the seller dimension. | 1010 |  |
| seller\_1\_name | VARCHAR | 1024 |  | Name of the primary seller. | John Doe | seller\_1\_name |
| seller\_1\_state | ENUM |  |  | Two-letter state code of the primary seller. | NY | seller\_1\_state |
| seller\_2\_name | VARCHAR | 1024 |  | Name of the secondary seller. | John Smith | seller\_2\_name |
| seller\_2\_state | ENUM |  |  | Two-letter state code of the secondary seller. | John Smith | seller\_2\_state |
| transaction\_id | INT |  | PK | Surrogate key for the transactiion dimension. | 1020 |  |
| transfer\_deed\_type | VARCHAR | 100 |  | Type of deed used for the transfer. | Grant Deed | transfer\_deed\_type |
| source\_url | VARCHAR | 2048 |  | URL of the original record source. | [http://putnamcounty](http://putnamcounty/) | source\_url |
| book | INT |  |  | Book reference for the property deed record. |  | book |
| page | INT |  |  | Page reference for the property deed record. |  | page |
| location\_id | INT |  | PK | Surrogate key for the location dimension. | 1030 |  |
| street\_address | VARCHAR | 1024 |  | Street address of the property. | 123 Elm Street | property\_street\_address |
| city | VARCHAR | 255 |  | City of the property. | Paytown | property\_city |
| county | VARCHAR | 255 |  | County of the property. | Putnam County | property\_county |
| township | VARCHAR | 255 |  | Township of the property. | Rochdale Village | property\_township |
| state | CHAR | 2 |  | Two-letter abbreviation for the state where the property is located. | NY | state |
| zipcode | CHAR | 5 |  | 5-digit US ZIP code where the property is located. | 10010 | property\_zip5 |
| latitude | DOUBLE |  |  | Latitude coordinate of the property. | 12.2333 | property\_lat |
| longitude | DOUBLE |  |  | Longitude coordinate of the property. | 118.2223 | property\_lon |
| sale\_id | VARCHAR | 255 | PK | Unique identifier for the property sale. | 1040 | sale\_id |
| sale\_price | BIGINT |  |  | Sale price of the property. | 5050505 | sale\_price |
| property\_id | VARCHAR | 255 | PK | Unique identifier for the property. | 1050 | property\_id |
| property\_type | VARCHAR | 255 |  | Type of property. | Residential | property\_type |
| year\_built | INT |  |  | The year property was built. | 1995 | building\_year\_built |
| num\_units | INT |  |  | The number of units in the building. | 0 | building\_num\_units |
| num\_stories | INT |  |  | The number of stories in the building. | 2 | building\_num\_stories |
| num\_beds | INT |  |  | The number of bedrooms in the building. | 3 | building\_num\_beds |
| num\_baths | INT |  |  | The number of bathrooms in the building. | 4 | building\_num\_baths |
| bldg\_area\_sqft | INT |  |  | The square footage of the building. | 1250 | building\_area\_sqft |
| land\_area\_acres | INT |  |  | The land area in acres. | 0.005 | land\_area\_acres |
| land\_area\_sqft | INT |  |  | The land area in square footage. | 2000 | land\_area\_sqft |
| land\_type | VARCHAR | 500 |  | The description of the land type of the property. | Agricultural. | land\_type |
| bldg\_assmnt\_value | BIGINT |  |  | The assessed value of the building | 505050505 | building\_assessed\_value |
| land\_assmnt\_value | BIGINT |  |  | The assessed value of the land. | 505050 | land\_assessed\_value |
| total\_assmnt\_value | BIGINT |  |  | The total assessed value of the property. | 5050505 | total\_assessed\_value |
| bldg\_apprais\_value | BIGINT |  |  | The appraised value of the building. | 50505 | building\_appraised\_value |
| land\_apprais\_value | BIGINT |  |  | The appraised value of the land. | 50505 | land\_appraised\_value |
| total\_apprais\_value | BIGINT |  |  | The total appraised value of the property. | 505050 | total\_appraised\_value |
| date\_id | INT |  | PK | Surrogate key for the date dimension. | 1060 |  |
| year\_number | INT |  |  | The year of the date. | 2025 |  |
| quarter\_number | INT |  |  | The quarter based on the date | 3 |  |
| month\_number | INT |  |  | The month of the date. | 4 |  |
| day\_number | INT |  |  | The day of the date. | 5 |  |
| hour\_number | INT |  |  | The hour of the date. | 12 |  |
| week\_of\_month | INT |  |  | The week of the month based on the date. | 2 |  |
| week\_of\_year | INT |  |  | The week of the year based on the date. | 30 |  |
| holiday | BOOLEAN |  |  | Holiday identifier based on the date. | True/False |  |
| weekend | BOOLEAN |  |  | Weekend identifier based on the date. | True/False |  |
| month\_name | VARCHAR | 25 |  | The name of the month based on the date | April |  |
| day\_name | VARCHAR | 25 |  | The name of the day based on the date. | Friday |  |
| sale\_datetime | DATETIME |  |  | The datetime of the property sale. | 4/1/2024 | sale\_datetime |
| deed\_date | DATETIME |  |  | The deed datetime of the property. | 1/25/1940 | deed\_date |
| bldg\_assmnt\_date | DATETIME |  |  | The building assessment datetime of the property. | 1/30/2024 | building\_assessed\_date |
| land\_assmnt\_date | DATETIME |  |  | The land assessment datetime of the property. | 1/20/2024 | land\_assessed\_date |