# **Dataset Selection**

## Description

Dataset Selected – [Insurance Claims Fraud Data](https://www.kaggle.com/datasets/mastmustu/insurance-claims-fraud-data) (click on the text to view the original dataset)

Description –

The Insurance Claims Fraud Dataset contains data that can be used in the process of insurance fraud detection. The original dataset contains a one-year worth data, from 2020/06/01 till 2021/06/30 in three data files which contains the Insurance, Vendor, and Employee data tables.

The dataset chosen gives a multidimensional view of the insurance data from the perspectives of the Customer, Agent (Employee), and the Vendor (the Insurance Company).

The original data tables of the data set have been edited, configured, and rearranged to suit the requirements of the project. Hence 5 data tables have been identified:

1. Insurance – Contains the insurance transaction data related to the insurances.
2. Agent – This contains the data about insurance agents who are involved in the managing and handling of customer insurance.
3. Customer – Contains the details about the customer or the holders of an insurance policy.
4. Vendor – Contains the details of the insurance service provider or the insurance company.
5. PolicyClaim – Contains the details of policy claims made by customers.

Diagram

Description automatically generated

# **Preparation of the Data Sources**

Initially, the original three data files were in the csv format. Then after they were downloaded and broken down into five separate tables in five separate data files. These files were saved in different data formats.

Three types of data sources were utilized: cvs, txt, database.

1. .csv –

The Agent data were kept in the csv source type file (Agent.csv).

This contains the data about insurance agents who handle the client / customer insurance policies.

1. .txt –

The Policy Claims data were saved in a txt source type file (PolicyClaim.txt).

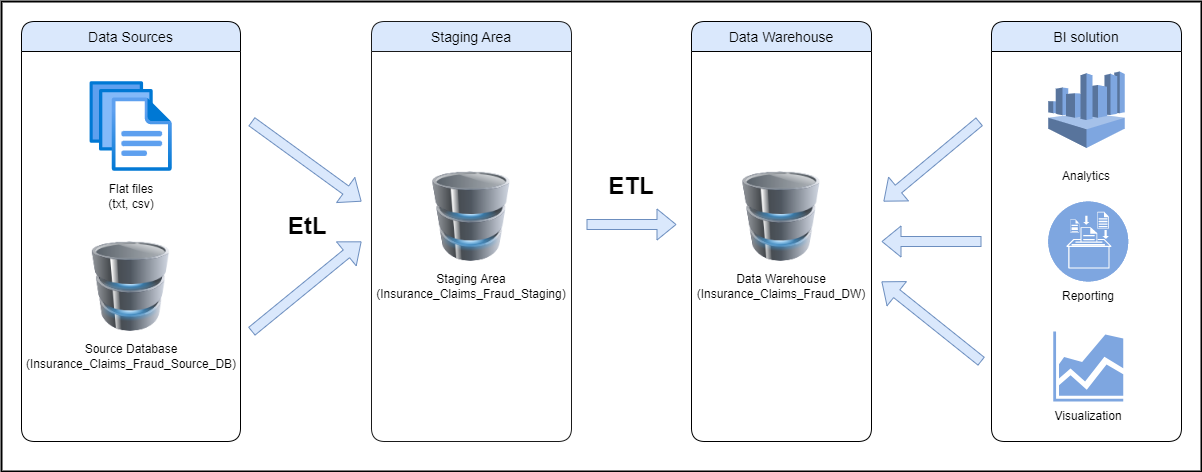
This contains all the details of the insurance policy claims that have been made by customers.

1. Database –

A source database (Insurance\_Claims\_Fraud\_Source\_DB) was created by importing the Insurance.csv, Customer.csv and Vendor.csv files.

* Insurance – This table contains all the transactional details related to an insurance transaction.
* Customer – This table contains all the details related to an insurance policy holder / owner of and insurance.
* Vendor – This tables contains the details of the insurance service provider of the company that offers the insurance cover to the customer.

# **Solution Architecture**

****

The above is a high-level Data Warehousing and Business Intelligence architectural solution for the chosen dataset and topic. We can identify for main layers here:

1. Data Sources –

The first stage in establishing a solid architecture is to collect data from many data sources, such as CRM, ERP, databases, files, or APIs, depending on the goals and resources available.

For the given scenario, there are 2 data sources, a source database, and flat files (csv and txt)

* Source database (Insurance\_Claims\_Fraud\_Source\_DB) which contains the customer, vendor, and insurance data tables.
* CSV (Agent.csv) file contains the agent details.
* TXT (PolicyClaim.txt) file contains the policy claims details.

1. Staging area –

A data staging area acts as a temporary storage facility between the data sources and the data warehouse. The staging area is primarily used to extract data quickly from its data sources while minimizing the effect of the sources.

In the given scenario a database named Insurance\_Claims\_Fraud\_Staging acts as the data staging area

1. Data warehouse –

A data warehouse is a large collection of business data that is used to enhance internal decision-making. This has a lot of historical information.

In this scenario a database file named Insurance\_Claims\_Fraud\_DW is used as the data warehouse. The said data warehouse comprises of 5 dimensional tables and 1 fact table.

1. BI solution (consumption) –

This employs technology and services to transform data into actionable insights that assist organizations in making better decisions.

In this scenario’s business solution, the data could be analyzed, visualized, and reported to understand the insurance frauds and to take actions necessary business decisions to mitigate them and take necessary informed decisions.

1. ETL –

ETL (Extract, transform, and load) is a data integration procedure that integrates data from several sources into a single, consistent data store that can subsequently be fed into a data warehouse or other destination system.

# **Diagram Description automatically generatedData Warehouse Design and Development**

The above is the dimensional model used for the given scenario. In summary the dimension model is designed with 5 dimensional tables (including the date dimension) and a single fact table.

## **The schema used – Snowflake Schema**

Snowflake schema has been utilized in the dimensional modelling to reduce redundancy through normalization. As visible the customer dimension table has been normalized.

* It was assumed that the address details would provide a greater benefit in categorizing and analyzing the data in various ways, hence the customer, vendor and agent contains hierarchical attributes describing the respective attributes.
  + - State > Postal code > City > AddressLine1 > Addressline2

## **Dimension and Fact Tables**

Five dimension tables and a fact table was created:

1. DimPolicyClaim – The policy claim dimension table contains the policy claim details. PolicyClaimsSK is the surrogate key.
2. DimCustomer – The customer dimension contains insurance policy holder / customer details. CustomerSK is the surrogate key.
3. DimAgent – Contains details of insurance agents who manages the customer insurances. AgentSK is the surrogate key.
4. DimVendor – Contains insurance provider details. VendorSK is the surrogate key.
5. DimDate – This is a common dimension. DateKey is the surrogate key. An SQL script was used to generate the date dimension.
6. FactInsurance – Contains all the transactional data. References dimension tables via foreign keys.

## **Slowly changing dimensions –**

It was assumed that the Customer and Agent could change certain details over time. Hence DimCustomer and DimAgent were considered as Slowly Changing Dimensions

* DimAgent – City and State are historical attributes.
* DimCustomer – City, State, PostalCode, AddressLine1, AddressLine2, EmploymentStatus, SocialClass, RiskSegmentation and MaritalStatus are historical attributes.

# **ETL Development**

## Load Data from Source to Staging

* Text

  Description automatically generatedOrder of Execution

Data is extracted from various data sources and is staged in an intermediate location until being loaded into the data warehouse. Individual extractions into the staging database happens as below images:

* Timeline

  Description automatically generated with low confidencePolicy Claim data are extracted from a flat file source and is loaded into staging area.
* Customer data are extracted from the customer data table of a source database.Timeline

  Description automatically generated with low confidence
* Agent data are extracted from another flat file source.Text

  Description automatically generated with low confidence
* A picture containing timeline

  Description automatically generatedVendor data are extracted from the customer data table of a source database.
* Insurance data are extracted from the Insurance data table of the source database.

Timeline

Description automatically generated

## Transforming and Loading Data into the Data Warehouse

* Text

  Description automatically generatedOrder of Execution

The staged data then undergoes some transformations and then finally loaded into the data warehouse as shown below:

* The following transformations are done to handle the NULL values of the city and addressline2 fields where they are found and replaced. Insertions of the modified and insert date are also assigned to obtain the system dates during insertion and modification.

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

* Following procedures are utilized in managing new insertions and updates of the DimVendor and DimPolicyClaim dimension tables.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

* A picture containing diagram

  Description automatically generatedLoading the transformed policy claims into the data warehouse.
* Loading the transformed customer data as a slowly changing dimension into the date warehouse.

Graphical user interface

Description automatically generated with low confidence

* Loading the transformed agent data as a slowly changing dimension into the date warehouse.

Graphical user interface

Description automatically generated with medium confidence

* Loading the transformed vendor data into the data warehouse.

Funnel chart

Description automatically generated

* Graphical user interface, text, application

  Description automatically generatedLoading the transactional data of this scenario into the data warehouse as a fact table (FactInsurance).

**Appendix**