**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 for of one year (2020/06/01 – 2021/06/30) that can be used in the process of insurance fraud detection. The original dataset contained three data files which contained 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 transactional data related to the insurance.
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 –

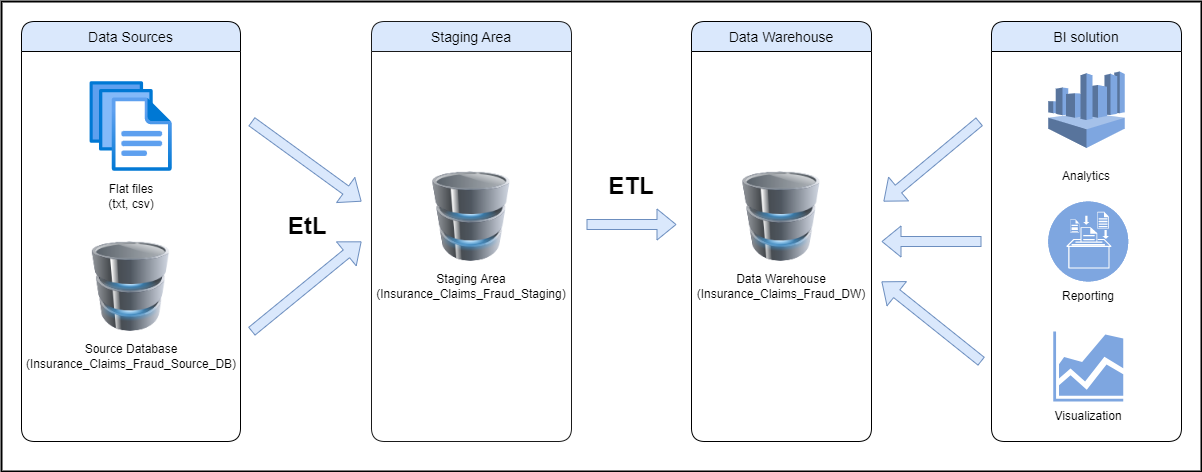
The Policy Claims data were saved in a txt source type file (PolicyClaim.txt).

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.