Data Mining with Maze2insights

by Mawunyo Adanu, consultant, Maze2insights LLC. June 6, 2026

"*Data is the new gold*" is now a familiar concept. Extracting this gold entails two endeavors:

(a) **Organizing/Engineering/ Preparing** real world **data** to be ready for analysis, by *integrating* multiple data sources into a *consistent* dataset. Real world data is akin to a **maze** that is overwhelming unless organized. For the best results, this effort should be scaled by building and maintaining data *hubs* and *pipelines*.

(b) **Deriving *actionable* insights** from the dataset and *sharing* with stakeholders (via *dashboards*, reports, etc.) and/or *encoding* them into data-dependent processes.

The primary hurdles to extracting actionable information from data include:

- access to computing facilities and analytical tools,
- ability to integrate data from multiple sources at scale and/or
- access to analytical skills.

There is a wide range of computing facilities and tools to choose from for curating data and producing and sharing analytics; from sophisticated proprietary software packed with pre-built features (bells-and-whistles) to open-source alternatives. The quality of both the data and analytics output is a function of the type of software used but also the skills and creativity of data engineers and analysts. Hence, there is potential *trade-off* between spending on computing facilities and tools *versus* skilled personnel. Small-to-medium sized companies with *limited* computing budgets can leverage skilled personnel to produce high quality databases and analytics nonetheless; basic software could be bolstered using creative and custom automation using **business**-savvy programming to expedite data and metrics production.

Benefits of Integrated Data Hubs/ Databases

***Standardization/*Consistency**

When different Business Units (BUs) in an organization make independent data pulls from source tables to satisfy *ad hoc* needs, the result is a ***litter*** of datasets with multiple versions of identical *measures*, often undocumented and creating confusion. This is especially true for industries that require specificity when describing measures. For example, in the insurance industry, **premiums** may be defined as *written*, *earned*, *unearned* and on a *calendar year*, *policy year* or *accident year* accounting basis. Having an integrated data hub that is based on *consensus definitions* eliminates confusion.

Reliability

Furthermore, such a hub serves as a reliable source of data and ready access to accurate information for all BUs.

Speed

It also facilitates seamless communication regarding business measures, leading to quicker execution and reliable decision-making.

AI and Automation of Workflows/Processes

Having a streamlined data process is essential for *automation* of data-dependent workflows and processes. Automation makes workflows/processes more dynamic and efficient and facilitates *scenario testing* of models used to determine business parameters. This allows analysts to focus on making sense of the output of their models and/or processes, not being bogged down by tedious data entry and reconciliation, fostering *comprehensive* analysis.

With the advent of AI, many companies are exploring its benefits (including workflow automation). However, a 2025 report by MIT found that 95% of AI pilots failed to advance to production. Data problems have been cited as one of the reasons for the failed AI efforts. For example, **Salesforce** encountered an apparent failure of an AI agent it launched. Investigations found that the problem was not the AI agent but defective underlying data that was feeding the AI agent. A robust data infrastructure will be essential to realizing benefits from AI initiatives.

Why Partner with Maze2insights For your Data Mining?

At **Maze2insights**, we combine understanding of your industry or business (via initial studies and consultation with client), coding and analytical skills to integrate all your data collection (transactional flat files, text files, Excel workbooks, Access tables, etc.) into a consistent database including all the **measures** employed by your business, and set up an automated refresh process to update the database at a desired frequency, in close consultation with your IT and business stakeholders. We can deploy the database in a DataMart of your choice (Tableau, Power BI, etc.) and build dashboards or other analytics (including models and/or data cubes for *models*), at the client's request. Furthermore, we will train your staff to maintain the process and be available to them for troubleshooting. Proprietary software is not required. We can work with both open-source software (R, Python) and proprietary software such as SAS.

Traditional in-house data projects require a *cross-functional team* of personnel from IT and other business units, who contribute their respective expertise to the project. Although such cooperative effort is laudable, there are challenges, notably:

- a learning curve for the team
- communication gaps and
- scheduling difficulties.

These can result in delayed execution. On the other hand, hybrid talent that combines technical knowledge of the client's business with coding skills can make decisions promptly, in consultation with IT and business stakeholders and be more agile. Furthermore, the ability to employ creative techniques in the data process/workflow means the client's existing resources would suffice for the project. This is what **Maze2insights** brings to the partnership – a high-quality data mining solution *customized* to the client's needs and *available* resources.

Extracting Alpha from Data

In today's age, every organization understands that there is actionable information (alpha) embedded in the data it collects during normal operations.

To distill the gold/alpha from the data requires synthesis of various sets of transactional data captured during operations into meaningful business or accounting measures. This synthesis no doubt requires IT resources (software and coding) as well as business knowledge to merge the transactional data consistently and to compute quantitative and qualitative measures that are *pertinent* to the business.

Levels (Tiers) of Data Synthesis

In practice, data synthesis occurs to varying degrees and is a function of available resources. At the highest level (designated here as ***Tier 1***), there is integration of internal transactional data and external data into a comprehensive business database, which feeds DataMart underlying reports and analyses, and the whole process refreshes periodically and automatically. **Large** organizations generally implement data synthesis at this level.

At the intermediate level (***Tier 2***), there is integration of *essential* internal transactional data and a *limited* amount of external data. Some compromises about *timing* relationships among data elements are acceptable, meaning accuracy of the data for Data Science applications may be reduced. The *marginal cost* of incorporating other internal data in the main process may be considered *unjustified*; only, when *necessary*, will such data be pulled on an *ad hoc* basis. This 80/20 solution is characteristic of **medium**-sized companies.

Finally, at the subsistent level (**Tier 3**), resources are limited so emphasis is placed on mission-critical functions. This is typical of **small** companies. The IT team would be responsible for capturing the transactional data but may be more focused on processes that deliver *imminent* data and information needs of staff to maintain operations. There are several *ad hoc* data pulls maintained by various Business Units often producing conflicting values for "identical" measures due to different definitions and/or assumptions underlying the queries used. Usually, there is a plan to build out a comprehensive data solution such as described above (Tier 1 or 2). However, such plans are based on *traditional* resource allocation strategies that require a dedicated cross-functional team from IT and Business stakeholders, where the IT team provides the infrastructure, including, a data-warehouse of primary (transactional) data and coding skills and the Business stakeholders provide the business data definitions needed for coding, to transform the transactional data into business measures. This *cooperation* tends to be rather challenging as there may be

repeated engagements to align each concept and its execution. Furthermore, in a small company environment, the resources of these teams are indeed “borrowed” rather than “dedicated. The result is protracted execution of the project, characterized by delays. A more efficient approach would be to employ talent with **bridged skills** – who are well-versed in the business fundamentals and methodologies but are also skilled coders and have a fair understanding of IT operations – such hybrid talent will consult with IT to locate the needed IT infrastructure and directly create the business measures required by stakeholders with minimal direction. They can also set up automated delivery of the output to facilitate efficient and optimal access for users in a business-friendly format. This will result in smoother and quicker execution.

The Case for Building an Integrated Database

Improvement in technology means businesses, more than ever, are operating in a fast-paced environment, having to respond quickly to changing events. The speed of response is a function of how quickly they can access information about driving factors – information that is systemic (public information) but also, ***more importantly***, specific information that is embedded in their ***internal*** data. Only a comprehensive data infrastructure that weaves all available information into a consistent granular database, which is directly connected to business processes used to analyze and monitor results, can ensure timely responses without interrupting the normal course of business. Furthermore, quick turnarounds improve customer service, help to take advantage of current windows of opportunity or to take immediate evasive action to stem impending losses, as well as to ***update*** business strategy.

Data Requirements in the Age of AI

Companies use data as *input* for determining business parameters (including *price*, *output*, *performance/profitability* indicators) and to improve processes. In all these endeavors, the models/frameworks/processes will become more *efficient* if the input data can *plug in directly* (from a data hub) without several manual interventions.

Recently, as companies have sought to leverage **AI** for productivity improvements, it has become apparent that the most *practical* application of AI is to automation of existing *defined* **workflows** (deployment of *autonomous* **AI agents** for Business applications is limited currently and is largely aspirational). Building a robust data infrastructure will be essential to realizing benefits from AI initiatives, especially automation of workflows and data-dependent processes, and having a streamlined data process is essential.

Automation makes workflows/processes more dynamic and efficient and facilitates ***scenario testing*** of analyses or models used to determine business parameters. This allows analysts to focus on making sense of the output of their analyses/models and processes, not being bogged down by tedious data *entry* and *reconciliation*, fostering ***comprehensive*** analyses.

Below we illustrate the construction of an ***integrated database*** using the example of an ***automobile insurance company***. We also include simplified examples of how the database can be utilized downstream in data analysis and other workflows.

Building an Integrated Database for an Automobile Insurance Company

In the simplest terms, insurance companies *organize* **risk pools** and *administer* them by determining the contributions (*premiums*) of members, making payments to members with valid claims (*losses*), paying the *expenses* incurred for the administration of the pool as well as the *cost of* capital used to run the risk pool.

Insurers strive to measure the risk level (loss potential) of pool members so that they can assign a fair premium, respectively to them. Accordingly, insurers are interested in premium and loss measures and their *attribution* to the *characteristics* of pool members.

Below we illustrate how premium and loss data can be obtained from transactions, organized, transformed into accounting measures, integrated with risk characteristics and deployed in a database for use in business analysis, reporting and process improvements. Some of the terminologies are specific to the industry and may not be familiar to every reader. However, the basic goal of the illustration should be apparent.

Terminologies

Premium Measures

The insurance business data flow begins with Premium (revenue). The Premium transactions are entered into a ledger system. Information on customers (insureds) and covered items may be stored in separate tables. To determine financial results, however, accounting measures of Premium must be calculated and associated with the natural groupings of insureds and covered items, by their attributes. Knowledge of insurance (premium) *accounting* is required to code this transformation.

Loss Measures

Booking Premiums gives rise to payment of Losses when covered events happen. Generally, Loss transactions are recorded in a ledger system, and other *occurrence* information about the loss event may be recorded in separate tables. To determine financial results, however, *accounting* measures of Losses must be calculated and associated with the natural groupings of insureds, covered items and occurrence details, by their attributes. Again, knowledge of insurance (loss) *accounting* is required to code this transformation.

Profit and Loss Measurement:

For an insurer, Premium represents *revenue,* and Loss represents *cost* (largest component). To measure financial results, Premium and Loss measures must be combined to determine appropriate *metrics* (e.g. *loss ratio = loss/premium**)***. Furthermore, including the factors giving rise to the two measures allows proper attribution of the metrics to factors, which enables understanding of business results and provides tools to manage the business successfully.

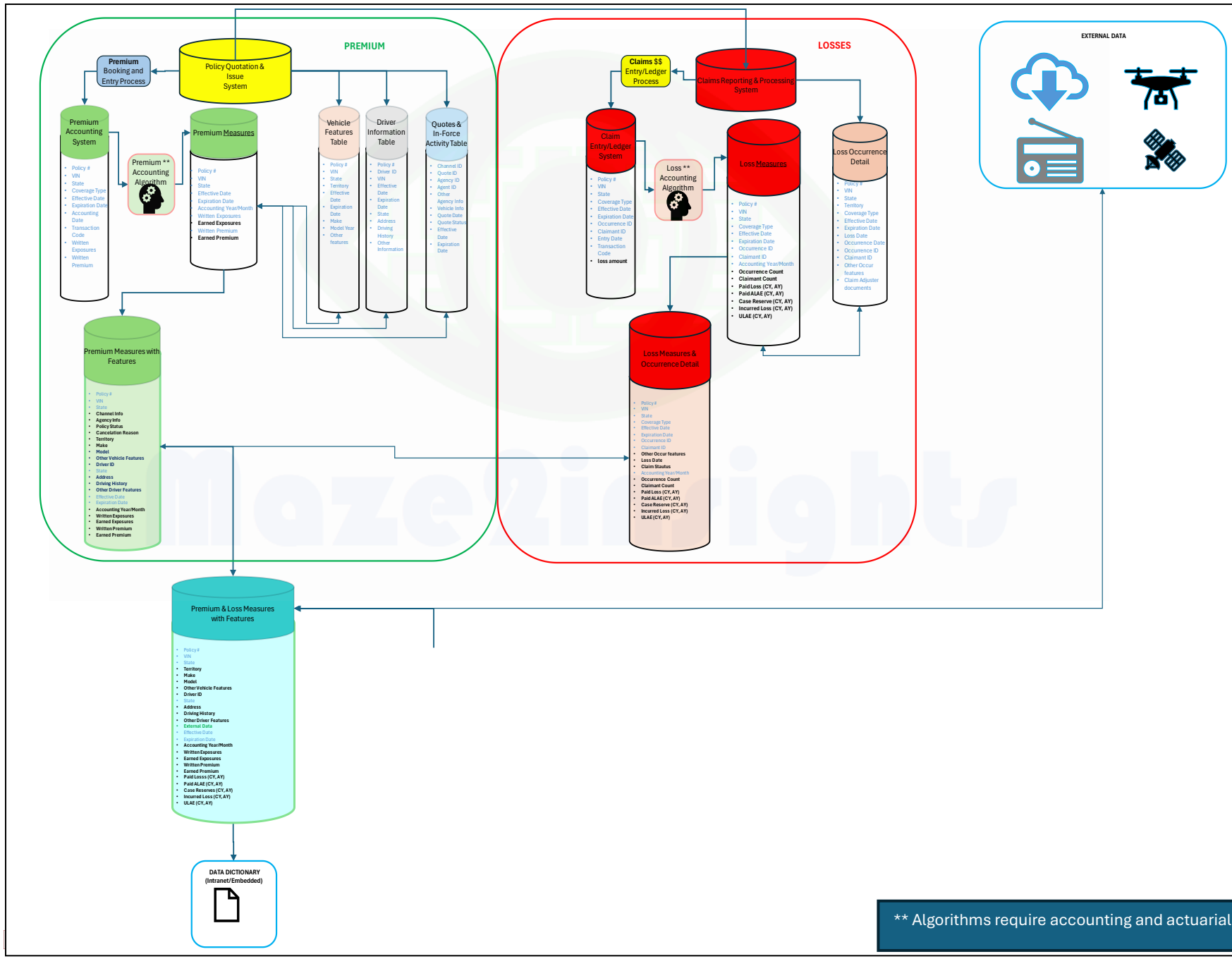
Flowchart 1 below illustrates the process of calculating *accounting* ***measures*** from premium and loss ***transactions***, respectively, and merging them with the relevant risk ***attributes***. The resulting premium and accounting measures are then *combined* by corresponding risks and attributes into *one* database.

External Data

Since external factors also impact the operation and results of any business, it is instructive to include pertinent external factors when combining Premiums and Losses. The outcome is a Premium and Loss database that **properly** attributes the measures to several driving (no pun intended) factors. Attribution *logic* is very important to capture effects correctly, especially where *correlation or causation* is implied.

Finally, it is good practice to create a *dictionary* for the combined database to provide documentation to users.

Flowchart 1 - Calculating Premium & Loss measures and merging to risk attributes



** Algorithms require accounting and actuarial knowledge

Data Consumption for *Alpha*

Congratulations on assembling a comprehensive Premium and Loss database. The next step is to extract as much *systematic* business information as possible to measure and monitor business performance, improve operational efficiency and achieve strategic goals.

Accurate summarization of measures by factors (attributes) and **effective** presentation and communication of such output is essential to the ability to leverage any "alpha" inherent in the data for improvement of margins through making processes more efficient and identifying profitable opportunities while minimizing costly mistakes.

Actionable information learned should be encoded into processes to boost efficiency and monitor performance, preferably by automatically feeding the relevant data elements into the processes.

As much as possible, financial analyses (Pricing, Reserving, etc., in the case of insurers) should be linked to input tools that are generated automatically from the database and should include enough *detail* that can be *modified on demand* to facilitate agile *scenario testing* and enhance fast and effective decision making. The latter (*not* rote manual updates to static models) is where analytical professionals shine. In practical terms, some "engine" components of analytical frameworks/models should be moved upstream, closer to data sources.

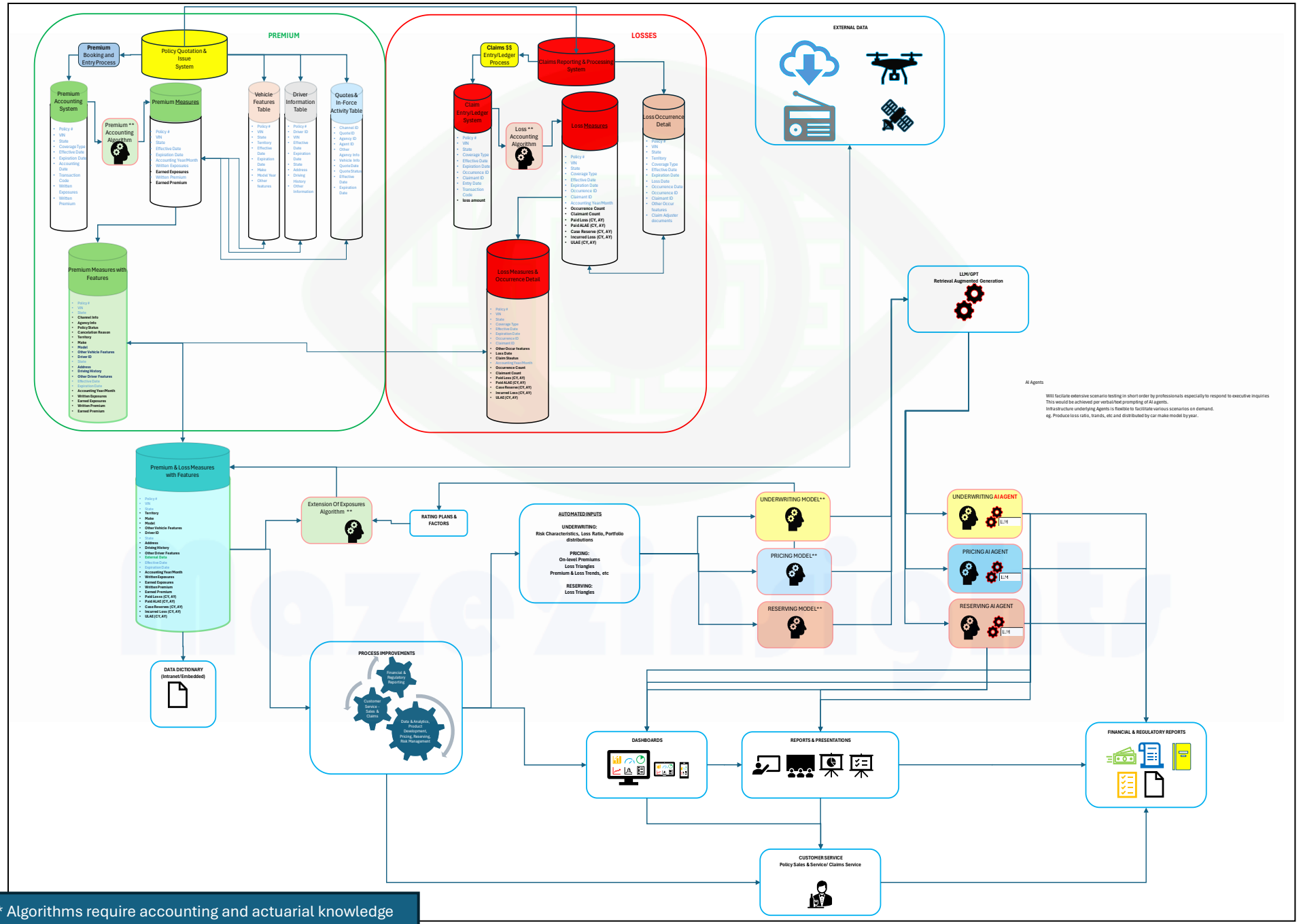
In fact, in the age of **generative AI**, analytical professionals could define their analyses/models with great flexibility and assign them to ***AI Agents*** that are fed by "dynamic" granular inputs and then issue commands (prompts) to the ***AI Agents*** to produce different scenarios.

Ultimately, a robust data pipeline (of internal and relevant external sources), accurate statistics of which are properly *connected* to the company's processes and analytical models will culminate in responsive actions from stakeholders to ensure customer satisfaction, regulatory compliance and sustainable financial success, even in the face of uncertainty.

Workflow 2, which extends workflow 1, shows how the database can be employed for data and financial analyses, process improvements and reporting.

In conclusion, putting all the components mentioned above together results in an integrated infrastructure that allows a company to operate in the matrix of its internal data as well as external data, leading to operational and financial success. The illustration here, albeit complex, is only a simplification of the real world and does not include many other contingencies that would emerge, but which can be overcome with determination.

Flowchart 2 – *Integrated* Database feeds Processes, Analyses and Reports



** Algorithms require accounting and actuarial knowledge

REFERENCE:

https://fortune.com/2025/08/21/an-mit-report-that-95-of-ai-pilots-fail-spooked-investors-but-the-reason-why-those-pilots-failed-is-what-should-make-the-c-suite-anxious/

https://fortune.com/2025/09/12/common-reasons-ai-products-fail-bad-data/

https://www.zdnet.com/article/50-ai-agents-get-their-first-annual-performance-review-6-lessons-learned/