Database Systems Project Part III. Machine Learning Model Creation

Name: Chenyi Lyu NID: N15220231
Name: Zoe Tan NID: N15948322

Optimization

Here, we assume the reference architecture that is most suitable for an insurance company to use in order to leverage hybrid data will depend on the specific needs and requirements of the company. In general, a hybrid data architecture combines both on-premises and cloud-based data storage and processing solutions, allowing the company to take advantage of the benefits of both. The ability to scale and adapt to changing business needs is one key benefit of a hybrid data architecture. As the company grows and its data needs evolve, the architecture can be easily expanded to accommodate the increased volume of data. This allows the company to continue to gain insights and make data-driven decisions, even as its data requirements change over time.

The Architecture

Some key elements of a hybrid data architecture for an insurance company might include:

- Data sources: These are the various systems and databases where the insurance company's data come from. For example, internal applications or external data sources such as public datasets.
- A cloud relational database management system (RDBMS): this would be used to store structured data, such as policy information, customer details, and transaction records. This data would be organized into tables with well-defined schema, making it easy to query and manipulate using SQL.
- A data lake: A central repository for storing structured and unstructured data from a variety of sources, including on-premises and cloud-based systems. A data lake would be used to store large volumes of historical data, such as claims data and customer interactions. This data would be extracted from the operational systems and transformed into a format that is optimized for analysis, data analytics, machine learning, and reporting.
- Data pipelines: These are the processes and tools used to extract, transform, and load data from various sources into the data lake. Data pipelines can be designed to support real-time or batch data processing, depending on the needs of the company.
- Data warehousing: A data warehouse is a specialized type of database designed for storing and analyzing large amounts of data. In a hybrid data architecture, the data warehouse can be located on-premises or in the cloud, depending on the specific requirements of the insurance company.
- Analytics and visualization tools: These are the tools and technologies used to analyze and make sense of the data stored in the data lake and data warehouse. This might include tools for data mining, machine learning, and visualization, such as dashboards and reports. Advanced analytics tools and machine learning algorithms would be used to gain insights from the data and make predictions.

    - Specifically, big data platforms, such as Hadoop and Spark, would be used to process and analyze unstructured and semi-structured data, such as social media posts and sensor data. These platforms would enable the insurance company to perform complex analyses on large volumes of data in near real-time.
    - For example, a machine learning model could be trained to predict the likelihood of a customer filing a claim, or to identify patterns in claims data that could indicate fraudulent activity.
- In addition to these core components, a hybrid data architecture for an insurance company might also include additional elements such as data governance and security measures, as well as ongoing maintenance and support to ensure that the architecture continues to meet the changing needs of the business.

Overall, the hybrid data architecture would allow the insurance company to store and manage a wide range of data, and to use this data to make better business decisions and improve customer service.

Components and data flow

The design would leverage Azure cloud platform services. Corresponding to the components we mentioned above, the services we aim to deploy are as follows:

- **Azure Database for MySQL**
    - stores structured data such as customer's data in a transactional database hosted.
    - The data from Azure Database for MySQL can be processed using **Azure Databricks** and stored in the analytics platform
    - Further, SQL Server Machine Learning Services can be used for computation.
- **Azure Blob storage**
    - This service can be a good candidate for a data lake as a massively scalable object storage for any type of unstructured data-images, videos, audio, documents.
- **SQL Server Integration Services** and **SQL Server Agent** can be used to automate these solutions.
- **Azure Data Explorer**
    - Fast, fully managed and highly scalable data analytics service for real-time analysis on large volumes of data streaming from applications, websites, IoT devices, and more.
    - Raw structured, semi-structured, and unstructured data such as any type of logs, business events, and user activities can be ingested into Azure Data Explorer.
- **Azure Synapse Analytics**
    - According to the documentation, Azure Synapse Analytics is the fast, flexible, and trusted cloud data warehouse that allows us to scale, compute, and store elastically and independently, with a massively parallel processing architecture. Analytics service that brings together enterprise data warehousing and Big Data analytics.
    - Use Azure Synapse Analytics to build a modern data warehouse and combine it with the Azure Data Explorer data to generate BI reports on curated and aggregated data models.
- **Azure Data Explorer Dashboards**
    - We can leverage the dashboards to natively export Kusto queries that were explored in the Web UI to optimized dashboards.
- **Power BI**
    - Finally, power BI helps drive better decision making with data visualization. Visualizations help gain deeper data insight. Use Power BI to interpret this data and generate new visualizations and insights.

- During the process, we can use **Synapse Pipelines Documentation** to create, schedule and orchestrate workflows.