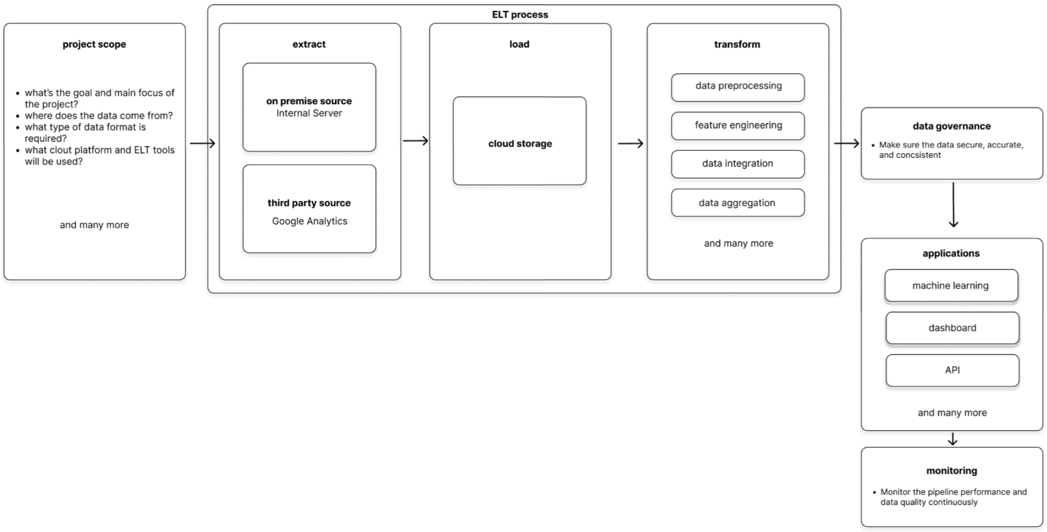
Question:
How would you create a data platform end-to-end system? The data might have internal data or external data, but the end data would be stored into cloud platforms like Google Cloud Platform or Azure Platform or AWS Platform. Please give details step by step, including data preparation, model evaluation, etc.

Answer:



My step if I will ever create a data platform end-to-end system can be seen on the diagram on top of this. The explanation will be explained below.

- Before creating a data platform end-to-end system, the first process that needs to be done is determine the project scope. The project scope is important because it will be the first step to ensure the data platform is aligned with the business objective and the resources to make sure the result is achieved. Defining project scope can help to identify the data sources, data tools and platforms to be used. It also can prevent unnecessary data activities.
- After determining the project scope, the next step will be ELT or Extract-Transform-Load. On this end-to-end system, I prefer to use ELT rather than ETL or Extract-Transform-Load because the data will be stored in cloud storage. With the data stored in it, the process of transforming the dataset can be done inside the cloud storage with less complexity, faster process, and flexible transformation, where the raw data is also saved in the cloud.
  In the extraction process, I will take the data from on premise source, such as internal server and third party services, such as Google Analytics. It can align with the business objective to choose the source datasets. Moreover, the raw data will be loaded to cloud storage. The data will be transformed

with the process includes data preprocessing, feature engineering, data integration, data aggregation, and many more based on the business needs.

- Furthermore, the data will be evaluated to make sure data is secure, accurate, and consistent. It also determines who will have access to the dataset, what permissions they have, and many more. This is also an important step due to the internal dataset is privacy and only certain people can gain access to it.
- Following that, the process of data pipeline did not end to make the final dataset, but also to make an analysis or modelling of it. Since the data is stored in cloud platforms like Google Cloud and AWS,it can be used for building dashboards and training machine learning models. Moreover, the final dataset can be used via an API, making it part of a complete end-to-end data system.
- Lastly, the final step of an end-to-end data system will be monitoring the process of the data pipeline to ensure data quality, performance, and reliability. This step identifies any problems in ETL processes and makes sure the dashboards, machine learning can continue to produce accurate and consistent results. The continuous monitoring also allows for feedback from users to create the system more aligned with the business objectives.