

To ensure that a data pipeline is kept up to date with accurate sales data, I would suggest the following tools and processes:

Source System Integration: First, I would recommend integrate the source system with the solution available to manage data pipelines in the company. This can be done through various methods, such as using APIs, ETL (Extract, Transform, Load) tools, or real-time data streaming, depending on the needs and available products.

Data Validation and Quality Control: Once the data is extracted from the source system(s), it is important to validate and clean the data to ensure that it is accurate and reliable. This can be achieved using automated data quality control tools and processes, such as data profiling, data cleaning, and data standardization. If using DBT, there are several data and schema tests available, built-in with DBT.

Data Transformation: Next, the data needs to be transformed into a format that is compatible with the target system(s) that will consume the data. This can involve mapping data fields, filtering, and aggregating data, and performing calculations on the data.

Data Loading: Once the data is transformed, it can be loaded into the target system(s). This can be done using tools such as SQL Server Integration Services, Azure Data Factory, dbt, databricks delta and more, depending on the solution.

Incremental Data Updates: Incremental updates can be used to ensure that the data is updated daily. This involves tracking changes in the source system(s) and only extracting and transforming the data that has changed since the last update. This can be achieved through various methods, such as using change data capture (CDC) or delta extraction.

Error Handling and Monitoring: Finally, it is important to have a robust error handling and monitoring process in place to ensure that any issues with the data pipeline are identified and resolved quickly. This can involve using automated monitoring tools and alerts, as well as having a dedicated team to monitor the pipeline and troubleshoot any issues that arise.

As a practical suggestion, based on my experience, I would recommend the creation of a medallion architecture on the data platform, following the logic and rules stated below. This allows data engineers, BI engineers and data champions to use a single source of information, with data governance and scheduled updates. This also makes it easier to change data in gold tables, as the origin tables are all in the system and everything is previously standardized.

