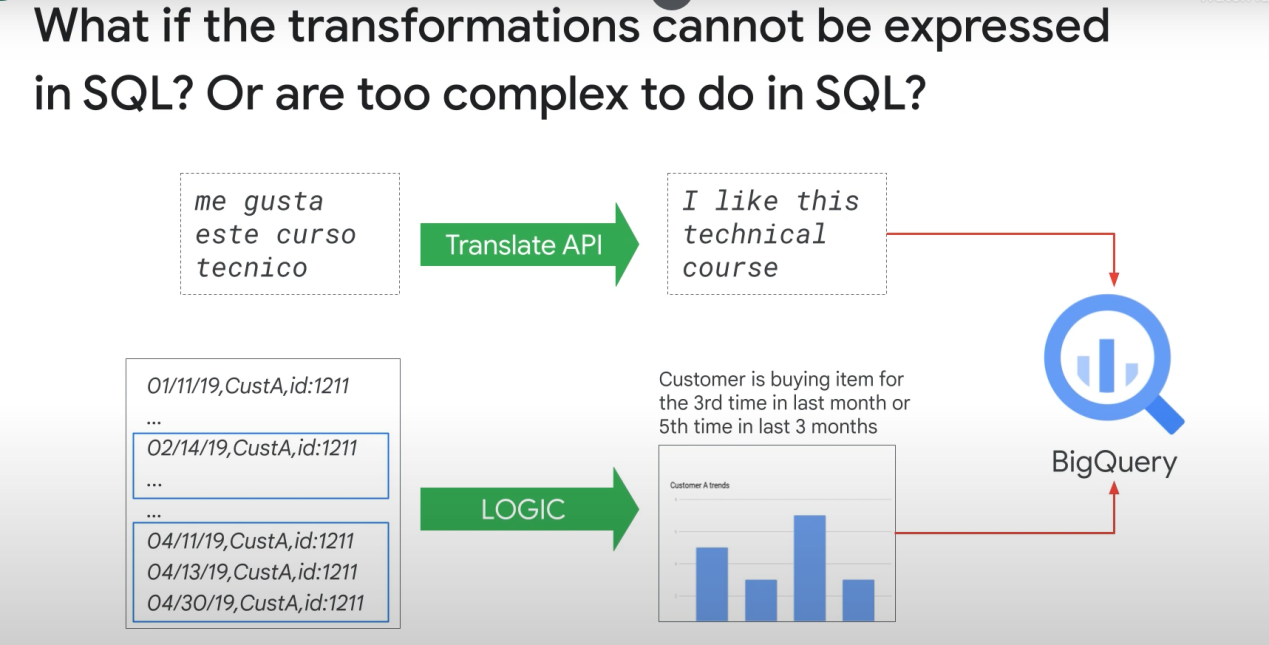


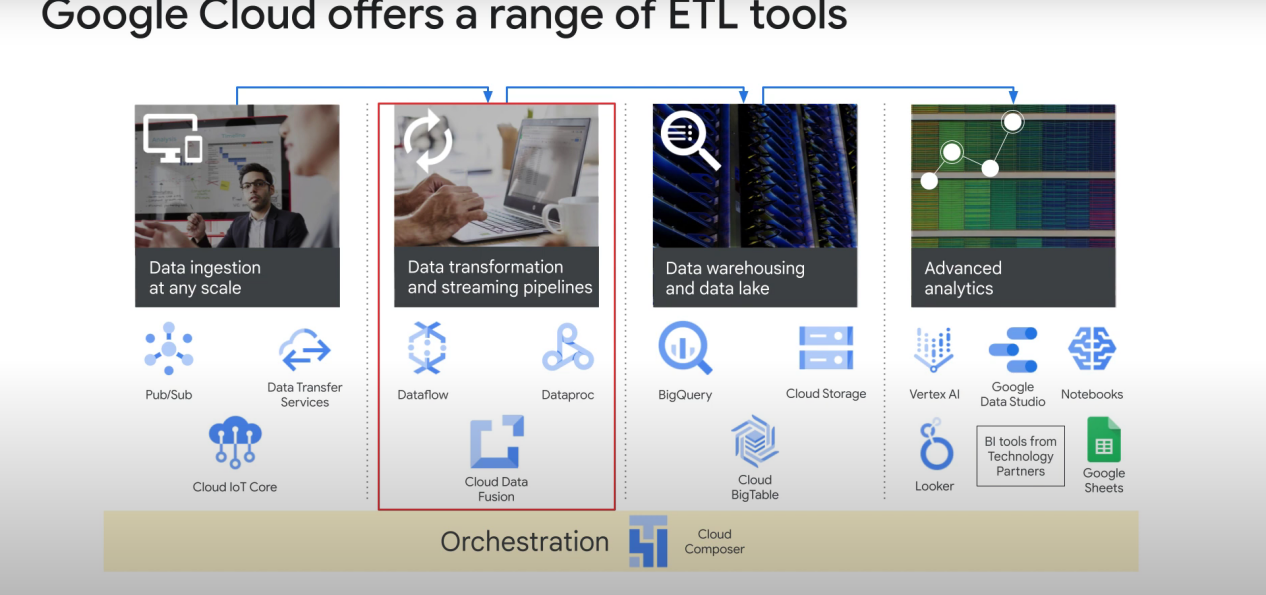
When to use ETL:



ETL can be used for this situation:



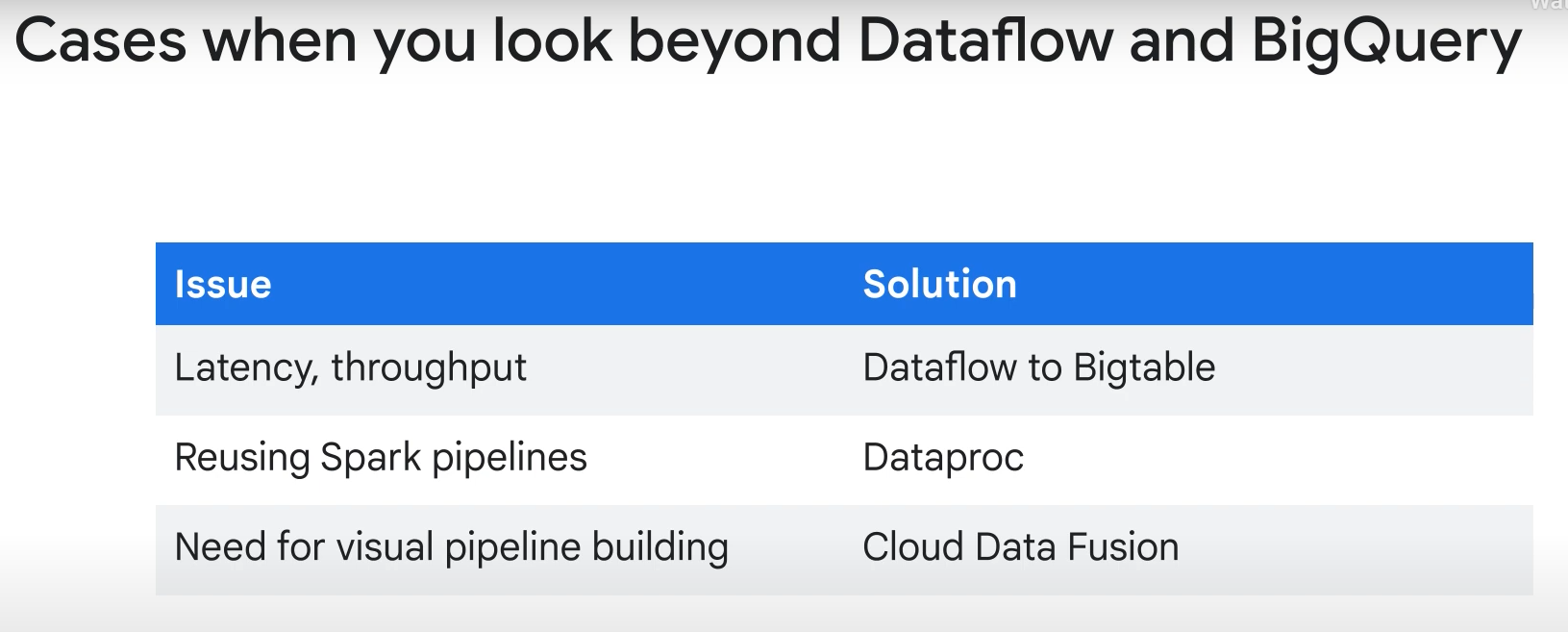
ETL TOOLS:

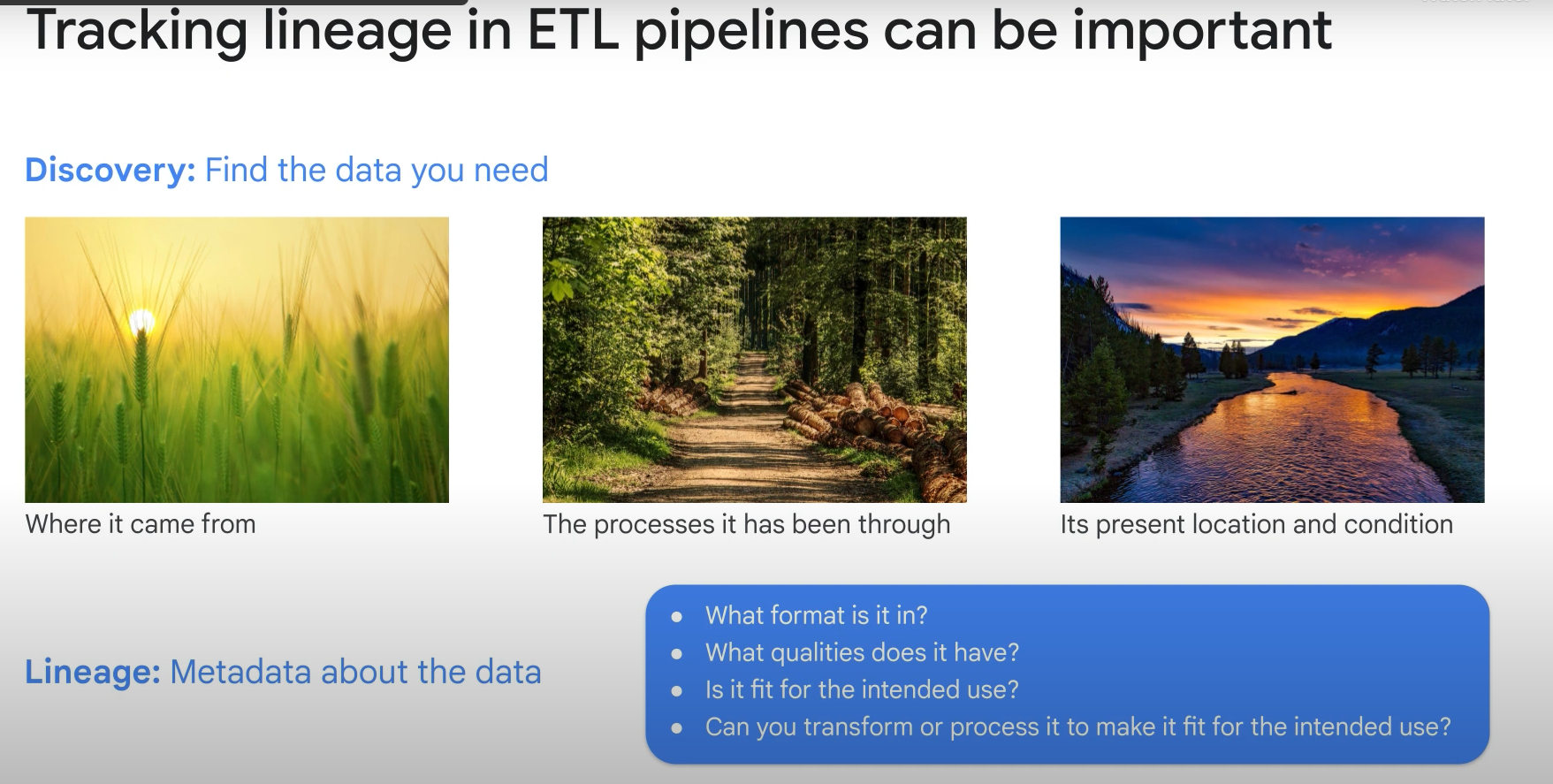


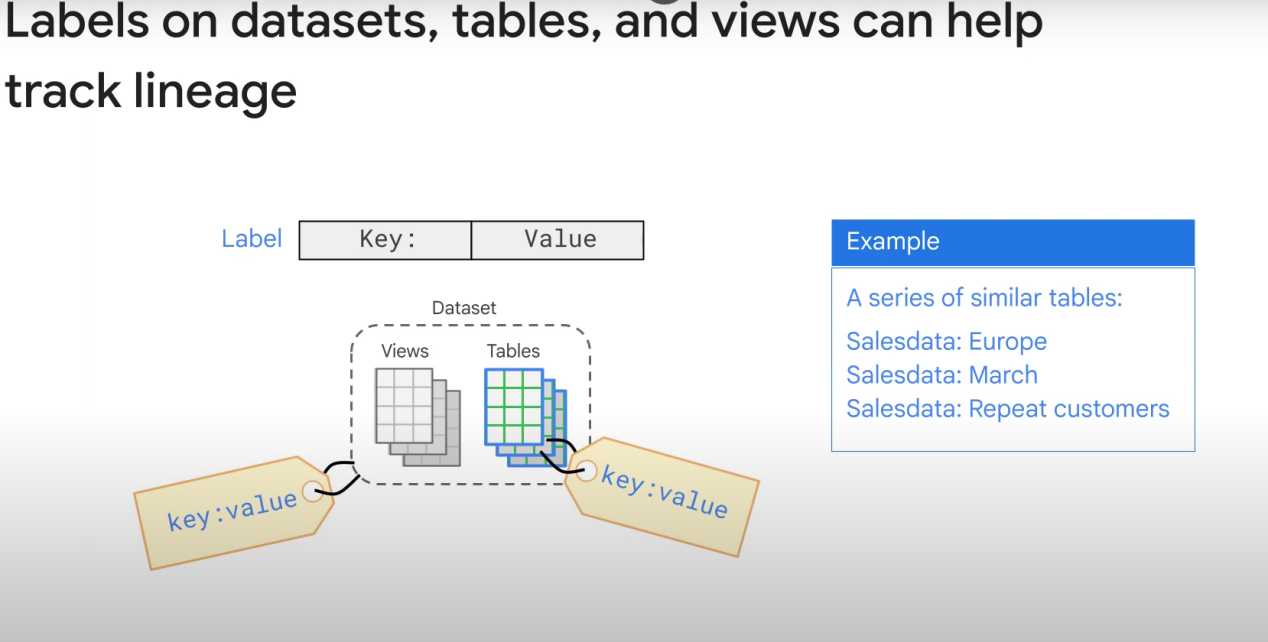
**Data proc:** requires significant hadoop expertise

**Datafusion**: provides simple UI to build ETL pipelines than can be deployed to data proc clusters

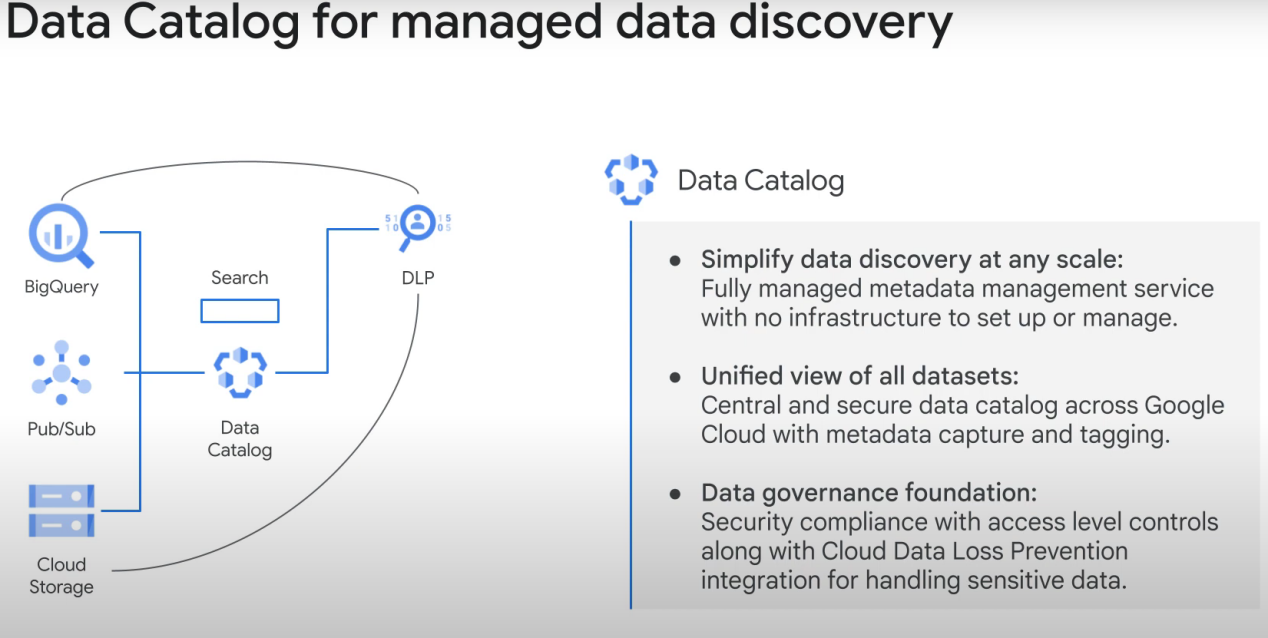
**Dataflow:** Batch + streaming data pipelines. Apache beans experience is good to have, but google also provides quick start templates for dataflow which makes it easy to use.



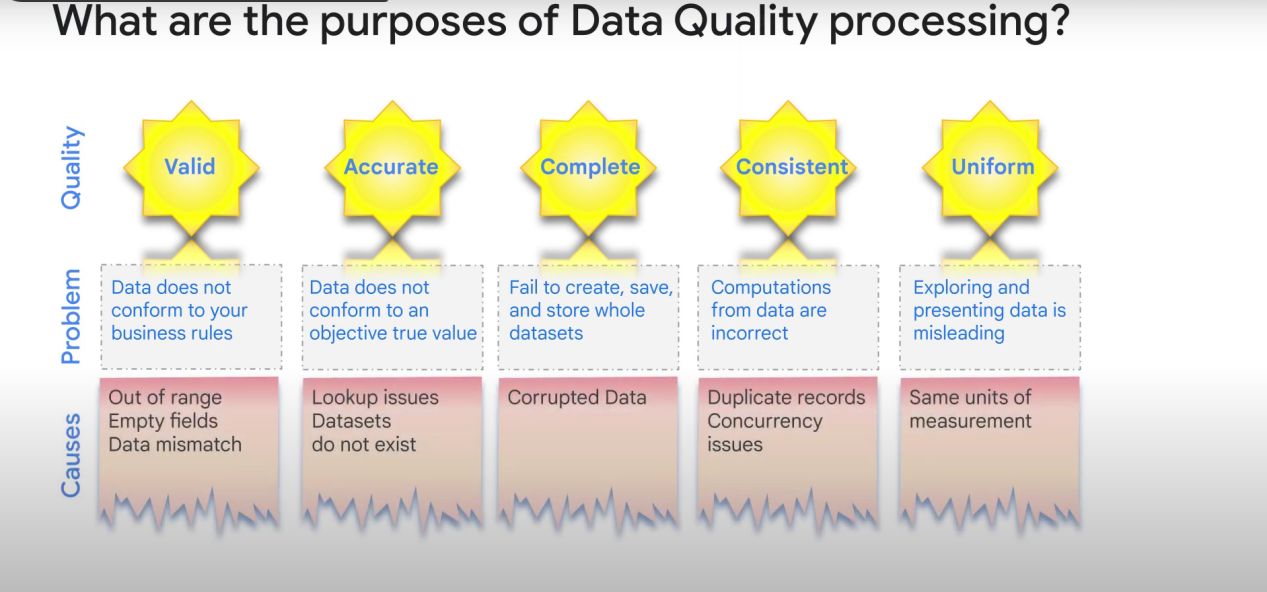


**Data catalog** allows you to see metadata (labels)/ lineage. It’s sort of a **metadata as a service thing**. It also allows to do data discovery.

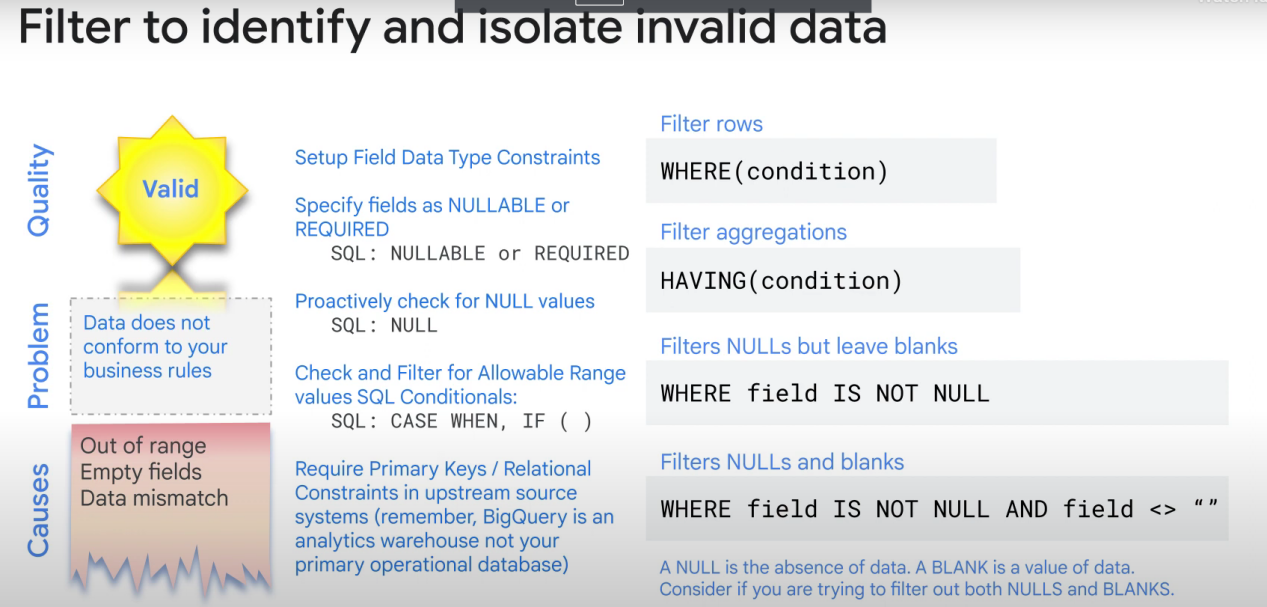
You can add labels to datasets, tables, and views

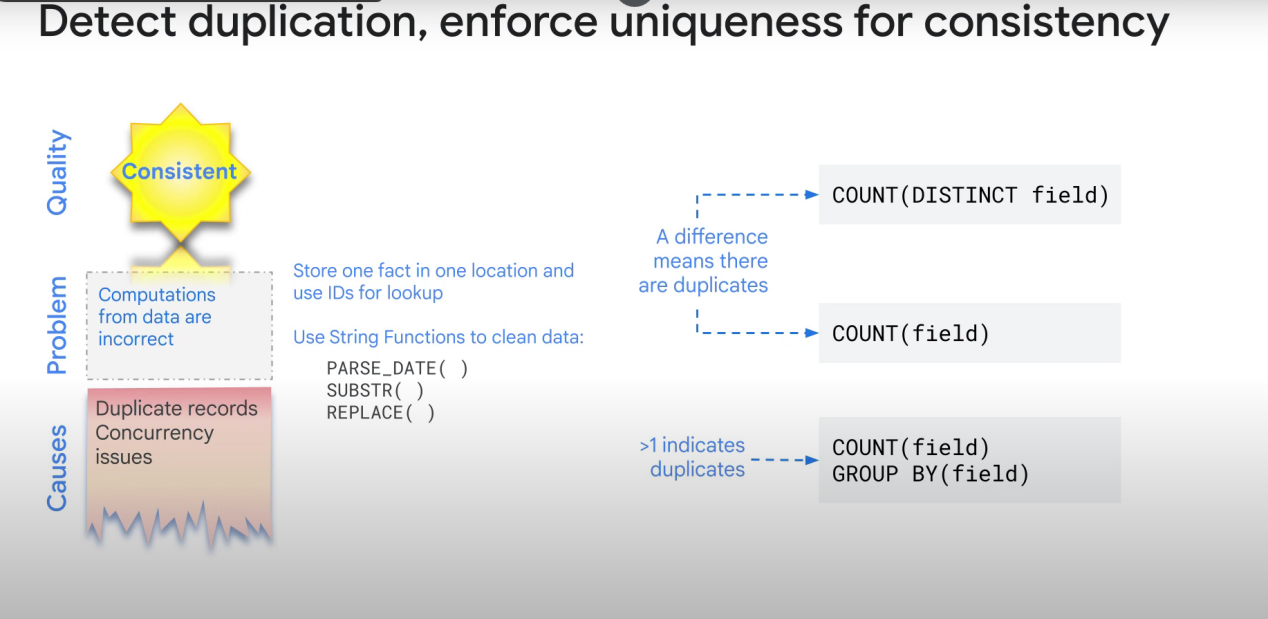


------------------------------------------------------------------------------------------------------------------------

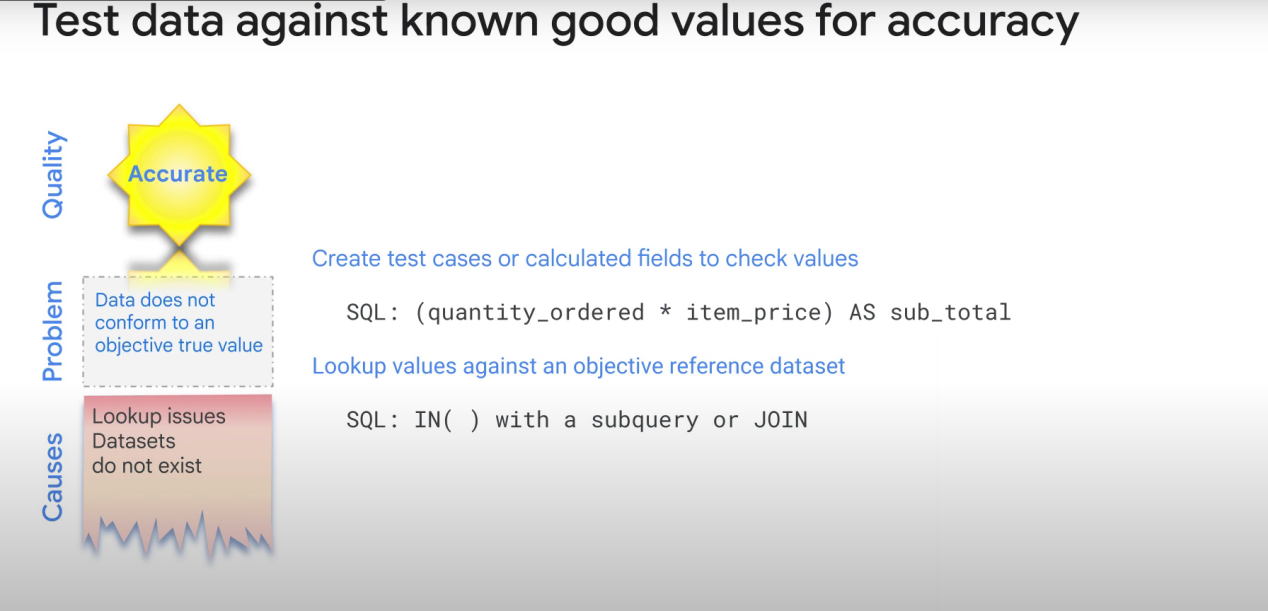


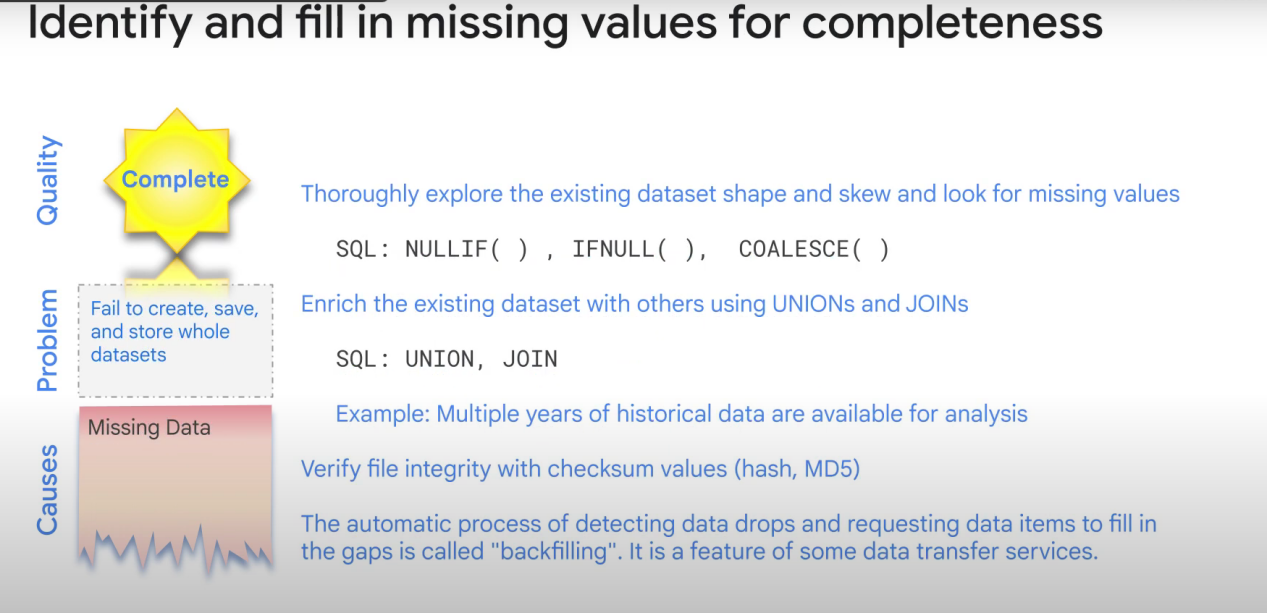
HOW TO FIX THEM?

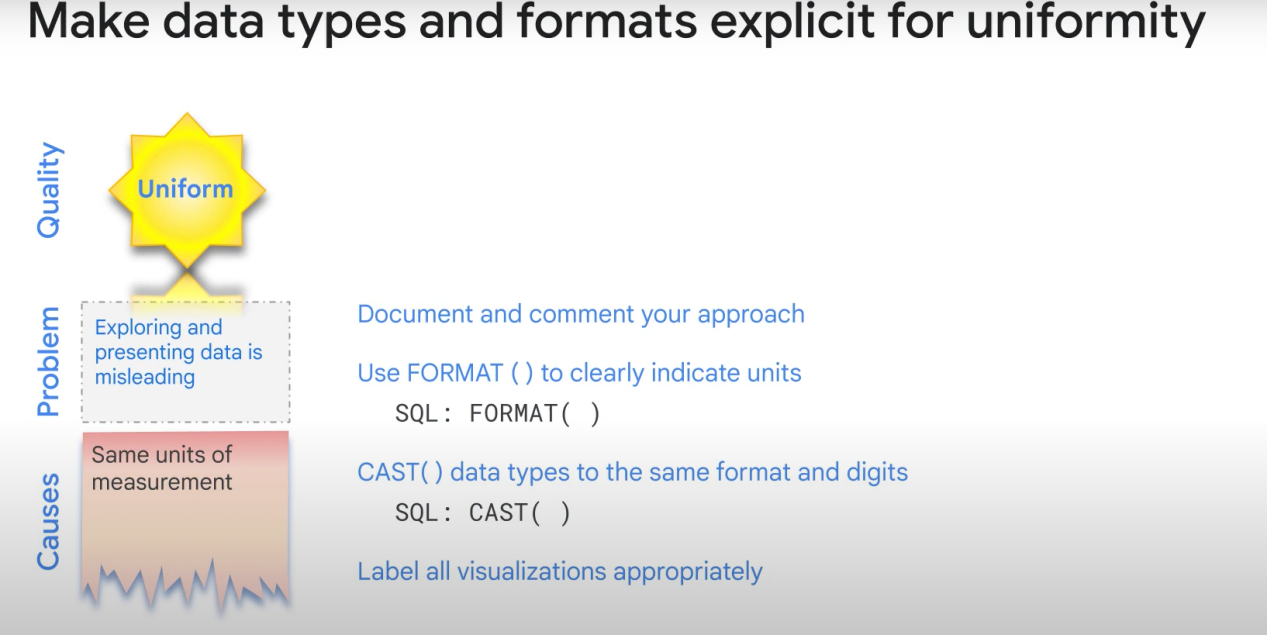




(duplicates, etc...)







(Not using the same units, etc...)