Data Architecture Design

Considering that streaming table is required, using delta table, specifically delta live table by databricks will serve the purpose.

We can use readStream function to read the parquet file, withWatermark method to specify the window of monitoring based on timestamp_detected column in the trans table. Along with dropDuplicationWithinWatermark method on detection_oid column.

Alternatively, write a writeback function against the table uising df.dropDuplicate(col("detection_oid")) and runs on microbatch or batch schedule.

Recommendation: given the scenario, recommend to use **watermark** to account of late or errors upstream ingestion for more control implementation.

Watermark feature is native to spark streaming engine thus not only specific to snowflake or databricks. However, would recommend this setup in the Azure Environment.