Handle CDC (Change Data capture) using python and Spark. work on it, come up with Architecture –

CDC is a process to identify changes to records in a source system, interpret the change(s) accurately, and then replicate the change to a target system with an intent to either replicate the source system or to record the changes for historical analysis. Or in other words, it can be said that change Data capture is an innovative mechanism for data integration. It is a technology for efficiently reading the changes made to a source database and applying those to a target database. It records the modifications that happen for one or more tables in a database. CDC records write, delete, and update events. It copies a selection of tables in their entirety from a source database into the target database.

There are two types of data changes :

1. Query-based - his approach regularly checks the production database for changes. This method can also slow production performance by consuming source CPU cycles. So many organisation don't track changes directly alternatively they use different CDC methods.
2. Log-based - The CDC process is a more non-intrusive approach and does not involve the execution of SQL statements at the source. this method involves reading log files of the source database to identify the data that is being created, modified, or deleted from the source into the target Data Warehouse.

Steps to Perform CDC

STEP 1: Extract: Raw data is extracted from an array of sources and sometimes placed in a Data Lake. This data could be formatted in JSON – Social media (Facebook, etc.), XML – Third-party sources, RDBMS

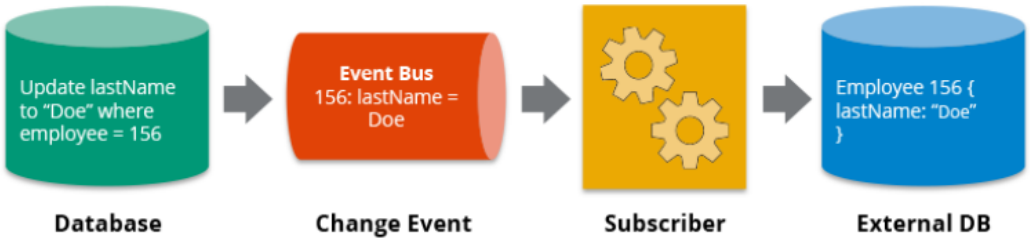
STEP 2: Transformation:

The transformation stage is where we apply any business rules and regulations to achieve.

Standardization, Deduplication, Verification, Sorting

STEP 3: Load:

To load this extracted transformed data into a new home by executing a task (job)



> The change data capture process via the publisher/subscriber method. Multiple databases and applications can subscribe to the change data.