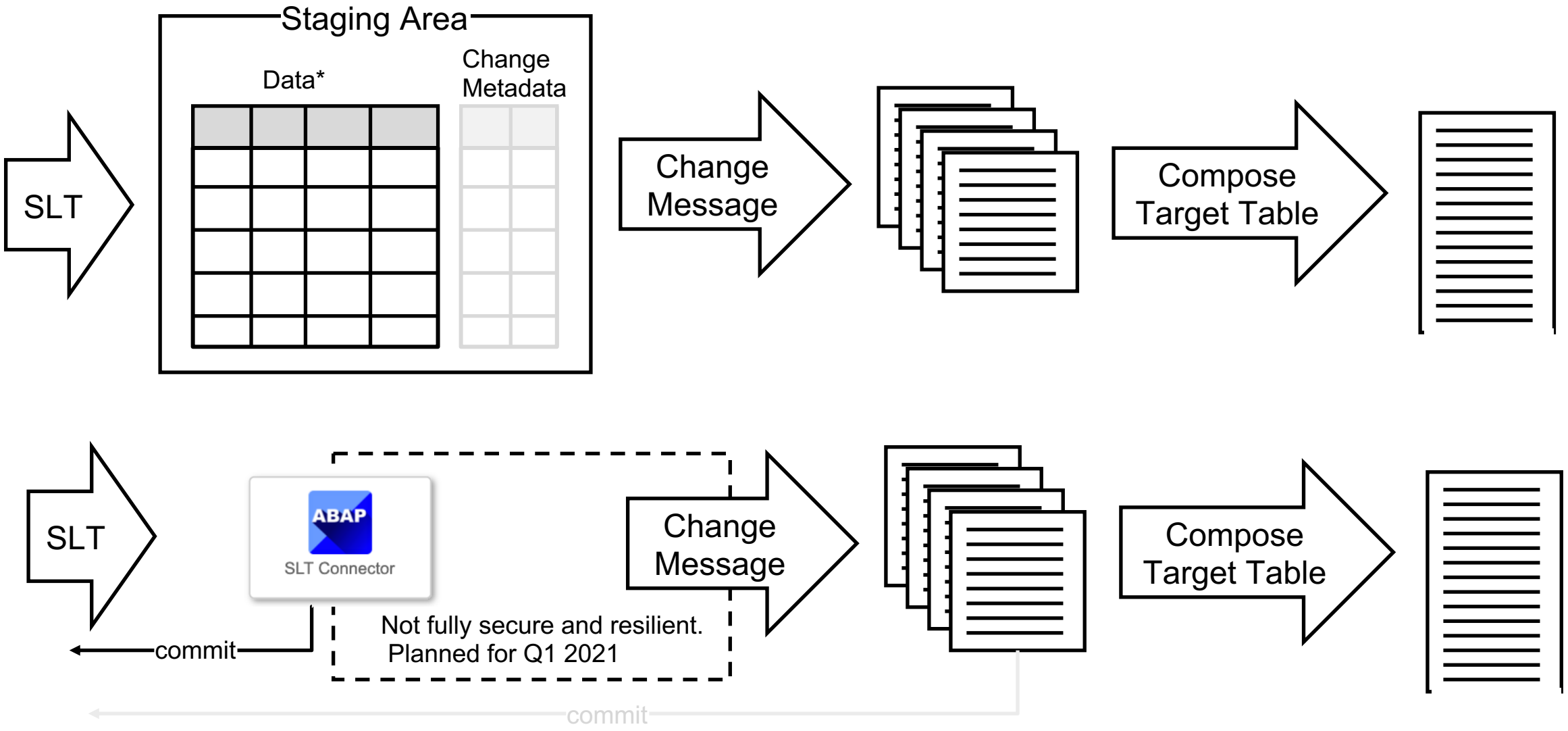
Replication with SAP Data Intelligence

Dr. Thorsten Hapke
Product Manager SAP Data Intelligence

August 2020

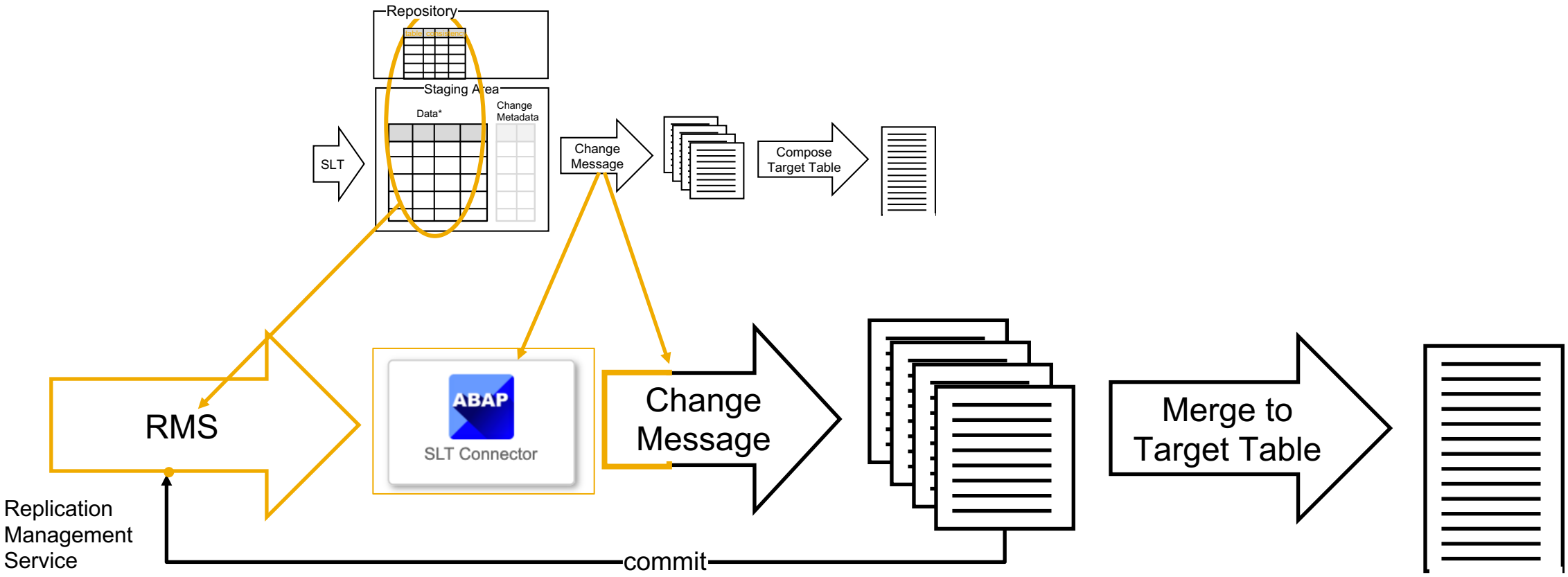
THE BEST RUN SAP

Design Options - SAP Data Intelligence 3.0

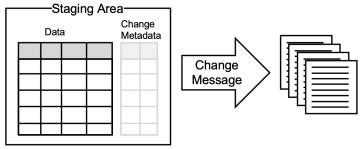


*Mirrored data or only Data Changes

Design Options – Planned SAP Data Intelligence 3.2

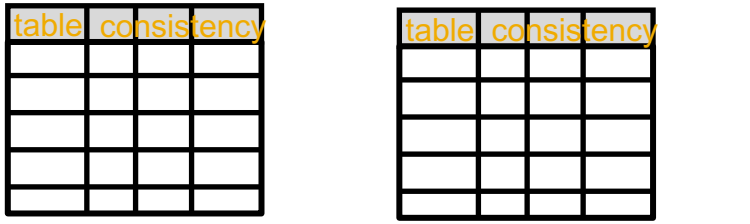


Staging Design Detail

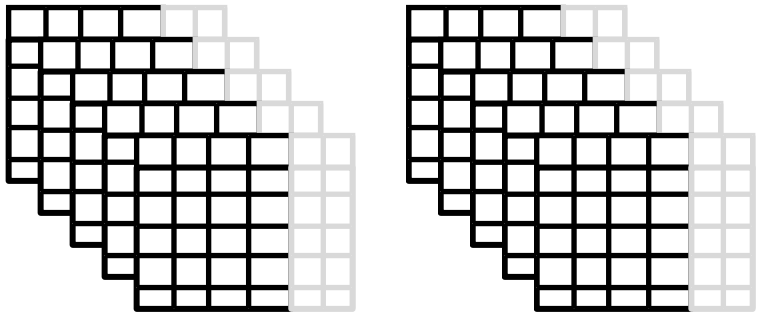


Repository Table

List of table names used as bundle for 1 replication pipeline with consistency check result.



Replication Tables



Staging HANA

SAP DI Replication Pipeline

Folder

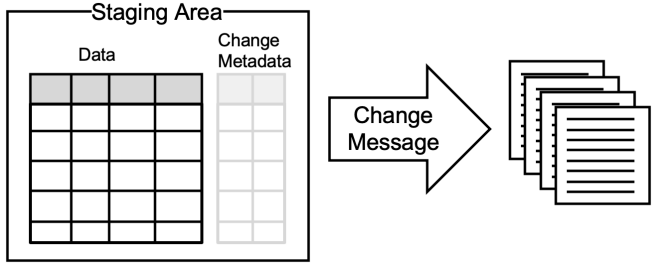
- Schema
  - Table
    - Files
      (change files, base file, primary key file, …)

Example:

- CUSTOMER
  - ADDRESS
    - ADDRESS_primary_keys.csv
    - ADDRESS.csv — Base File
    - 153531234_ADDRESS.csv — Change Files
    - 153531235 _ ADDRESS.csv — Change Files

Object Store

Design with ”HANA-Staging”



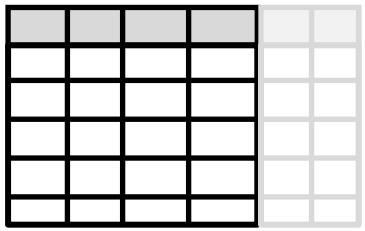
Preparation:

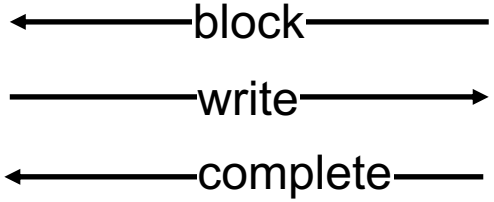
1. Adding “Change”-metadata columns to ”data”-table.

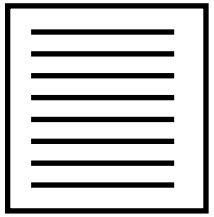
DIREP_TYPE	DIREP_UPDATED	DIREP_STATUS	DIREP_PID
Type of change: Insert/Update/Delete	Timestamp for having a change order	Status of the replication: Wait/Blocked/Completed	Provided by SAP DI as kind of transaction ID

2. Optional – When a “hard”-delete is required on SLT-message a table-trigger is needed for keeping the information of this change in a “delete-shadow”-table.

Process:







Appending Update and Insert-changes to change-files with timestamp/pid part of file name.

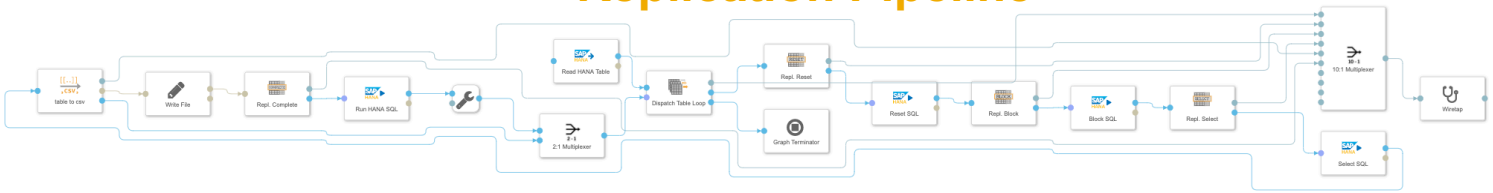
Replication of Changes

Replication Repository Table:

TABLE	…
Table name with Schema	…

- Permanent Loop or
- Scheduled

Replication Pipeline



Object Store

- Schema
  - Table
    - Files

(change files, base file, primary key file, …)

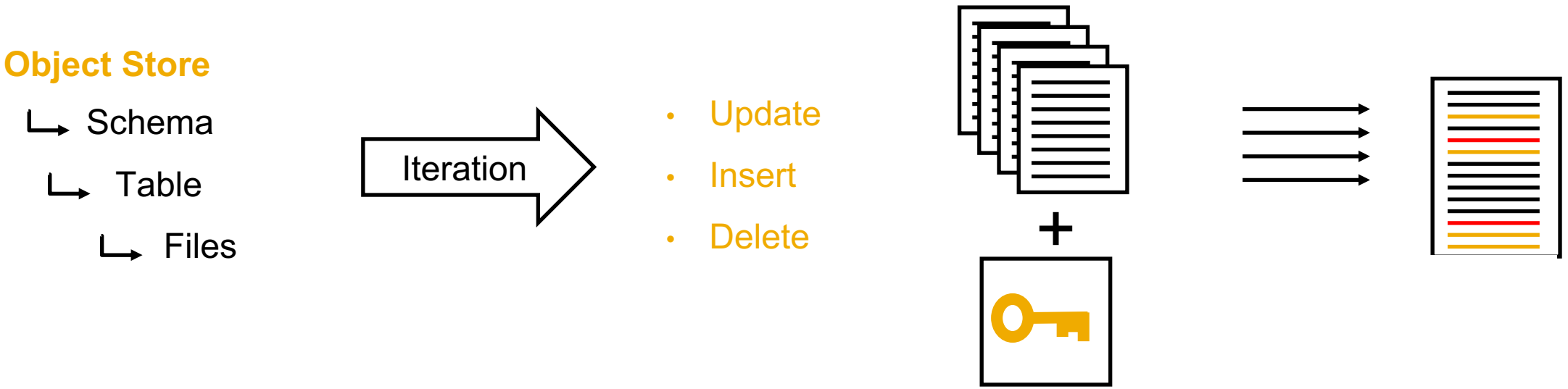
Compose Target Table

Preparation:

Run Pipeline for “primary-keys”. Stores primary keys of all tables to corresponding object store location.

Process:

Run Pipeline for “merging” all “change” tables to “target” table.



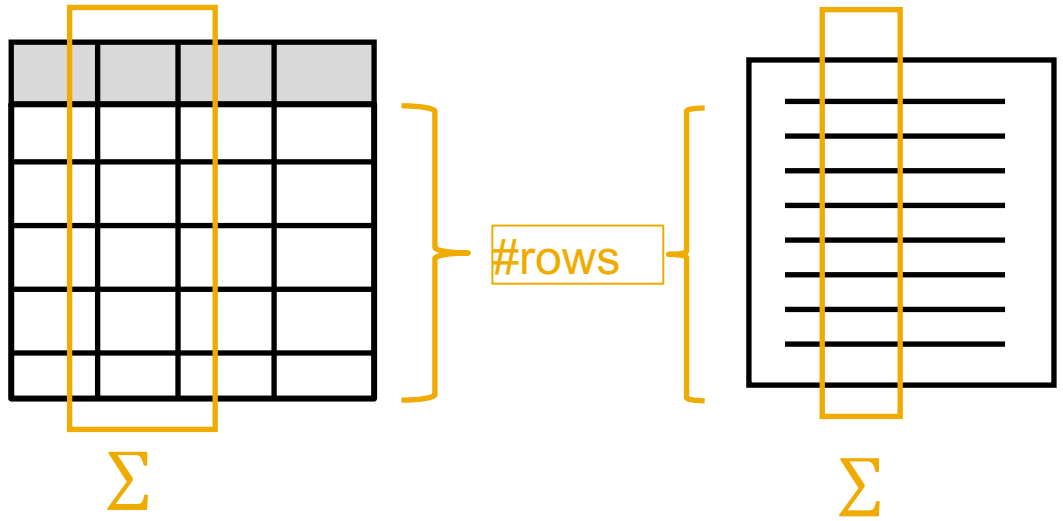
Check Consistency

TABLE	…	CHECKSUM_COLUMN	FILE_CHECKSUM	FILE_ROWS	FILE_UPDATED
Table name with Schema	…	Column of table to calculate check sum to verify consistency (Type: Int, Distinct or close to distinct)	Sum of the CHECKSUM_COLUMN	Number of rows	Timestamp of the last check

…

TABLE_CHECKSUM	TABLE_ROWS	TABLE_UPDATED
Sum of the CHECKSUM_COLUMN	Number of rows	Timestamp of the last check

- FILE-checksum calculated after each file merge.
- Table-checksum after running the pipeline TableProfile



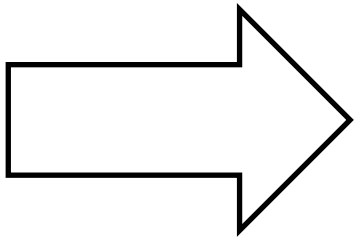
Latency to Real-Time Replication

Impact on Latency

replication process

Only changes loaded and send.

merging process



Entire file needs to be loaded for an update.

Recommendation

update and insert changes should run in separate replication pipelines and

Inserts directly been appended to target table.

No Conflicts, due to

The timestamp “DIREPL_UPDATED” and primary-key determines the merging result, not the sequence of pipeline processes

Partition Target for Performance Gain – Business Content Partitioning

Rationale:

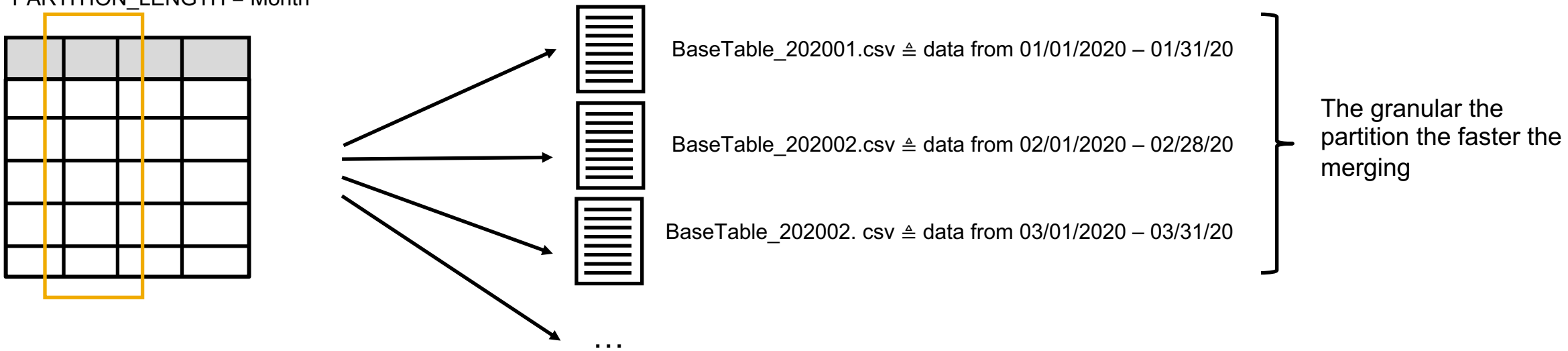
The “merging”-process is more time consuming than a table “copy”/initial load, because the whole target csv-file has to be loaded for updating the records. Therefore a partitioning of the target file can help by using a specified column that has a metric and is evenly increment over time.

Replication Repository Table:

TABLE	…	PARTITION_COLUMN	PARTITION_LENGTH
Table name with Schema	…		

Planned (open)

PARTITION_COLUMN,
e.g. Transaction date with
PARTITION_LENGTH = Month



Test Pipeline

Generating Test Tables

Config:

- Number of tables
- Number of rows for each table

Outcome:

- Test Tables in HANA
- Test Tables added to Replication Repository table

INDEX	NUMBER	DATE	DIREP_TYPE	DIREP_UPDATED	DIREP_STATUS	DIREP_PID
Row-Number	Integer	Hana date Type	Type of change: Insert/Update/Delete	Timestamp for having a change order	Status of the replication: Wait/Blocked/Completed	Provided by SAP DI as kind of transaction ID

Update Test Tables

Config:

- Modulo factor
  (number of records updated, e.g. 2 ≙ 50%
- Maximum random Number

Outcome:

- Updated test tables

Insert Test Tables

Config:

- Modulo factor
  (number of records updated, e.g. 2 ≙ 50%
- Maximum random Number

Outcome:

- test tables with additional records

Misc. Replication Management Pipelines

Prepare Object Store

Config:

- Table Repository
- Root folder on Object Store

Outcome:

- Creates folder structure
- Creates empty files with header

Get Primary Keys

Config:

- Table Repository
- Root folder on Object Store

Outcome:

- Fetches primary keys of all tables in Table Repository and stores them to corresponding folder location

Open

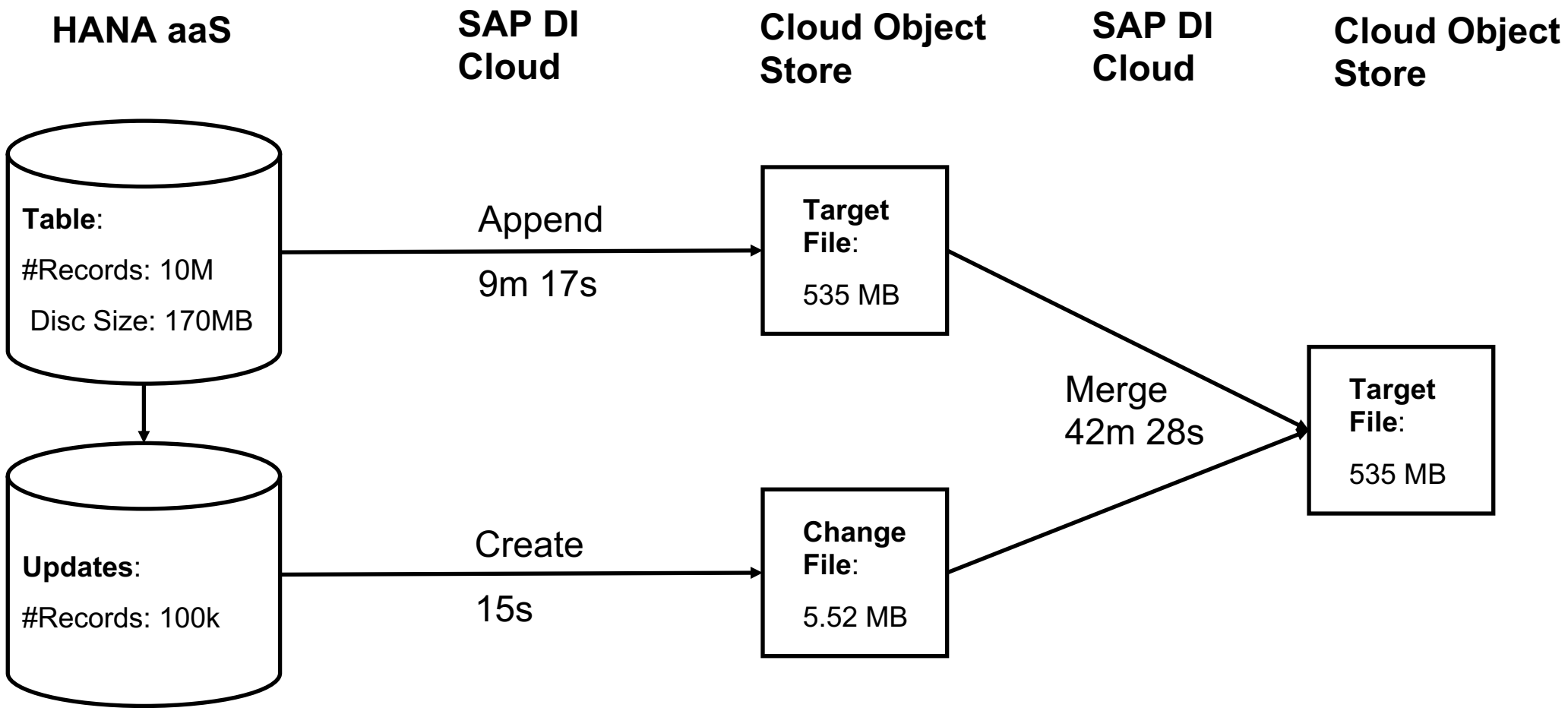
Release Blocked Records

Config:

Outcome:

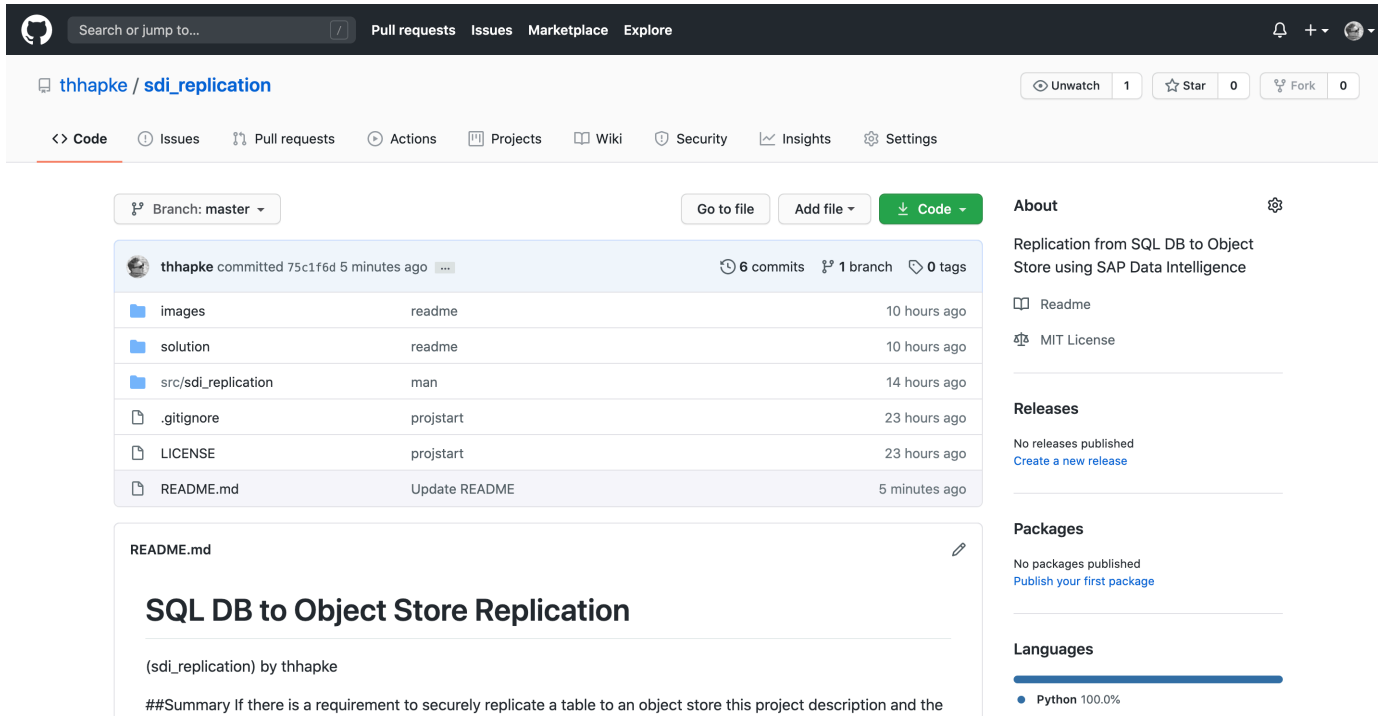
-

Performance Measurements



SAP DI Process Documentation and Download Solution

PUBLIC github



README

Source-Code of operators

Solutions for import into SAP DI

- Operators
- Pipelines