Для стандартизации систем источников заводим таблицу sys.sources

* id
* name
* short\_name
* description - произвольное короткое описание
* confluence - ссылка на confluence
* type (table/api/s3)
* db\_type (ms/postgres/mysql)

Таблица для сущностей источников sys.entities

* id
* sources\_id
* description (любое пояснение для себя и коллег)
* schema\_src (схема в источнике)
* name\_src  (имя в источнике)
* name\_trgt  (имя в DWH в нотации lower\_snake\_case) - Это именно универсальная часть именования в каждом слое
* filter (по сути условие where для отсечения не нужных данных) - Полей и значений может быть несколько. Пр: date > '2023-01-01' and type = 'type'
* dag\_name - название для DAGа airflow
* load\_type
  + none - не переносим, устарело, не понадобится
  + once - перенести вручную один раз. Сущность более не обновляется на источнике
  + full - полностью забрать все данные и полностью обновить (для маленьких справочников, менее 10.000 и не чаще чем раз в сутки)
  + inc - инкремент (нужно определить поле инкремента на источнике)
  + inc\_advanced - когда нет поля дата обновления, а есть только поле дата создания. Будут ситуации когда меняется строки или бесследно удаляется, а мы детектить не можем
* inc\_column (поле инкремента в источнике, если load\_type = inc )
* detect\_rule - потенциально правило, которое помогает отслеживать  изменения в источниках для  load\_type = inc\_advanced
* частота и время обновления таблицы - можно расписание строкой в формате cron сделать

Схемы в trgt не прописываем. Так как он идет по слоям - нескольким схемам.

Таблица для полей сущностей источников sys.entity\_columns  
Допущения

* не учитываем порядок полей
* не учитываем уникальность
* не учитываем PK, FK
* не учитываем значение по умолчанию, переносим что есть
* не учитываем not null, но есть правило потенциального преобразования пустых строк и null значений

======

* id
* entities\_id
* description (любое пояснение для себя и коллег)
* name\_src  (имя колонки в источнике)
* name\_trgt  (имя колонки в DWH в нотации lower\_snake\_case)
* data\_type\_src (тип данных на источнике. varchar, integer...)
* data\_type\_trgt (тип данных в dwh. varchar, integer...)
* data\_size\_src (длина / размер (для строк, чисел) на источнике. 128, 10 ...)
* data\_size\_trgt (длина / размер (для строк, чисел) в dwh. 128, 10 ...)
* data\_scale\_src (точность и масштаб (для чисел с плавающей точкой / decimal) на источнике. 10, 2...)
* data\_scale\_trgt (точность и масштаб (для чисел с плавающей точкой / decimal) в dwh. 10, 2...)
* data\_transform\_src (строка, приведение типа или другие минимальные трансформации на стороне источника)
* data\_transform\_trgt (строка, приведение типа или другие минимальные трансформации в dwh)
* null\_rule - правила трансформации при типе null, none...
* empty\_rule - правила трансформации при переносе пустой строке

**Data quality будет различаться для разных load\_type:**

**- full - сравниваем количество строк на ods слое(?) с источником**

**- inc - аналогично + проверяем кол-во строк в инкременте (если знаем примерный размер или хотя бы что он больше 0, то выдаем ошибку при нарушении условия)**

**- другие типы могут быть, если у нас есть инсайды по данным**

S2T можно вести единый файл json по каждому источнику. Но сдается мне что это будет не удобно из за потенциальной величины этого файла.

Поэтому предлагаю S2T сразу перевести в табличный вид.

[скрипты для заполнения метаданных](https://confluence.bistrodengi.ru/pages/viewpage.action?pageId=251200997)

sys.sources

INSERT INTO sys.sources (name, short\_name, description, type)  
VALUES ('tz\_r\_bpmonlinetz', 'tz\_bpm', 'реплика bpmonlinetz Турбозайм', 'ms\_db');

sys.entities

INSERT INTO sys.entities  
(sources\_id, description, schema\_src, name\_src, name\_trgt, "filter", dag\_name, load\_type, inc\_column, detect\_rule, update\_params)  
VALUES(1, 'Займ и заявка', 'dbo', 'Application', 'application', '', '', 'inc', 'modifiedon', 'modifiedon', 'daily');

sys.entity\_columns

INSERT INTO sys.entity\_columns (  
    entities\_id, description, name\_src, name\_trgt,   
    data\_type\_src, data\_type\_trgt, data\_size\_src,  
    data\_size\_trgt, data\_scale\_src, data\_scale\_trgt,   
    data\_transform\_src, data\_transform\_trgt, null\_rule, empty\_rule  
)  
VALUES   
    ('4', 'Id доп соглашения', 'Id', 'id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Дата создания доп соглашения', 'CreatedOn', 'created\_on', 'datetime2', '', 8, NULL, 7, NULL, '', '', '', ''),  
    ('4', 'Дата изменения', 'ModifiedOn', 'modified\_on', 'datetime2', '?', 8, NULL, 7, NULL, '', '', '', ''),  
    ('4', 'Тип доп соглашения', 'TypeId', 'type\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Дата начала доп соглашения', 'StartDate', 'start\_date', 'date', '', 3, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Дата окончания доп соглашения', 'EndDate', 'end\_date', 'date', '', 3, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Id статуса доп соглашения', 'StatusId', 'status\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Не используется', 'OwnerId', '', '', '', 16, NULL, NULL, null, '', '', '', ''),  
    ('4', 'Стоимость пролонгации', 'Amount', 'amount', 'decimal', '', 9, NULL, 2, NULL, '', '', '', ''),  
    ('4', 'Не используется', 'DocumentId', 'document\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Не используется', 'Comment', 'comment', 'nvarchar', '', 4000, NULL, NULL, null, '', '', '', ''),  
    ('4', 'Cвязь с займом', 'ApplicationId', 'application\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Cвязь с операциями', 'CashflowId', 'cashflow\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Порядковый номер доп соглашения', 'Number', 'number', 'int', '', 4, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Срок доп соглашения', 'Period', 'period', 'int', '', 4, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Дата завершения предыдущего доп соглашения по займу', 'PreviousEndDate', 'previous\_end\_date', 'date', '', 3, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Проценты за период', 'PlanInterestAmount', 'plan\_interest\_amount', 'decimal', '', 9, NULL, 2, NULL, '', '', '', ''),  
    ('4', 'id оферты на пролонгацию', 'OfferFileId', 'offer\_file\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Код АСП, которым подписано доп соглашение ', 'AspCode', 'asp\_code', 'nvarchar', '', 250, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'Id Типа платежного инструмента, которым оплачено доп соглашение', 'FinancingInstumentTypeId', 'financing\_instument\_type\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'id карты, с которой оплачено доп соглашение', 'FinancingInstrumentId', 'financing\_instrument\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', 'id cчета, которым оплачено доп соглашение', 'PaySystemOrderId', 'pay\_system\_order\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', '?', 'CanceledContactId', 'canceled\_contact\_id', 'uniqueidentifier', '', 16, NULL, NULL, NULL, '', '', '', ''),  
    ('4', '?', 'CanceledDate', 'canceled\_date', 'datetime2', '', 8, NULL, 7, NULL, '', '', '', '');