# **[JobGenerator 2.0](http://confluence.moscow.alfaintra.net/display/HADOOP/JobGenerator+2.0)**

Новая утилита генерации кода.

Основное отличие от предыдущей версии - генерация основана на реестрах объектов и полей (etl\doc\object\_registry.csv и etl\doc\object\_attributes\_registry.csv).

Для работы с генератором необходимо рядом с jar-файлом создать файл CONFIG.csv.

В CONFIG.csv необходимо прописать пути 4 файлам, на основе которых производится сборка кода:

1. **OBJECT\_REGISTRY\_FILE**- реестр объектов
2. **OBJECT\_ATTRIBUTES\_REGISTRY\_FILE**-реестр полей объектов
3. **HIVE\_SCHEMAS\_FILE**- параметры схем БД Hive
4. **SOURCES\_CONNECTION\_FILE**- параметры подключений к источникам

**OBJECT\_REGISTRY\_FILE. Требования к реестру объектов**

Должны быть заполнены первые 9 полей. Выделенные поля обязательны для заполнения

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Система источник** | **Имя сущности на источнике** | **БД Hive** | **Имя таблицы/view** | Партиционирование | Время запуска | Глубина | **Способ загрузки** | **Режим репликации ежедневно(последние x дней)/еженедельно(в какой день)/ежемесячно(в какой день/дни)** |
| **ANKSTANDBY** | **ALFAUSER\_AUTO.SLOLP\_APPLICATIONS** | **S\_SOURCEDATA** | **ANKSTANDBY\_APPLICATION\_AUTO** | filing\_date | 18:00 | c 01.05.2016 | **инкрементально по партициям** | **ежедневно** |

**OBJECT\_ATTRIBUTES\_REGISTRY\_FILE. Требования к реестру полей объекта**

Должны быть заполнены первые 4 поля.

Опционально можно заполнить поля (с 5 по 8 соответственно):

* Комментарий. Произвольный строковый комментарий
* Преобразование. Преобразование поля. Поддерживаются следующие значения:
  1. **format\_to\_date**- преобразование даты к формату YYYY-MM-DD
  2. **format\_to\_datetime**- преобразование даты к формату YYYY-MM-DD HH:MI:SS
  3. **format\_to\_milliseconds**- преобразование даты к формату YYYY-MM-DD HH:MI:SS.FFFFFF
  4. **format\_to\_string**- преобразование поля к строке. Используется для полей типа CLOB, BLOB, XML, итд
* Партиционирование. Указание на то, что поле используется для партиционирования таблицы. Поддерживаются следующие значения:
  1. **partition\_by\_year**- секции по году. В целевой таблице добавляется поле year\_part формата YYYY
  2. **partition\_by\_month**- секции по месяцу. В целевой таблице добавляется поле month\_part формата YYYYMM
  3. **partition\_by\_date**- секции по дню. В целевой таблице добавляется поле date\_part формата YYYYMMDD
  4. **partition\_by\_hour**- секции по часу. В целевой таблице добавляется поле hour\_part формата YYYYMMDDHH
* Дельта. Указание на то, что механизм перегрузки выделяет дельту по полю. Поддерживаются следующие значения:
  1. **delta\_column**- Указание на то, что поле используется для выделения дельты по интервалу даты
  2. **merge\_key** - отмечает ключевое поля для соединения дельты с предыдущими данными. Можно отметить несколько полей для составного ключа

Например:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **БД Hive** | **Имя таблицы/view** | **Имя атрибута** | **Тип данных** | Комментарий | Преобразование | Партиционирование | Дельта |
| **s\_dwh** | **vocstatus\_sdim** | **uk** | **double** | Унифицированный ключ измерения в DWH |  |  | merge\_key |
| **s\_dwh** | **vocstatus\_sdim** | **ccode** | **string** | Код статуса |  |  | merge\_key |
| **s\_dwh** | **vocstatus\_sdim** | **name** | **string** | Наименование статуса |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **description** | **string** | Описание |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **vocstatustype\_uk** | **double** | Тип статуса |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **src\_ccode** | **string** | Код типа льготы из СИ |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **module\_uk** | **double** | Система-источник |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **deleted\_flag** | **string** | Признак удаления записи |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **default\_flag** | **string** | Признак записи, введенной в хранилище для поддержания ссылочной целостности |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **dwscmix** | **double** | Уникальный код сочетания исходных систем |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **as\_of\_day** | **string** | Дата актуализации DWH | format\_to\_datetime |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **job\_insert** | **double** | Идентификатор  JOB`а, который вставил данную запись |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **job\_update** | **double** | Идентификатор  JOB`а, который обновил данную запись |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **dwh** | **double** | Тип статуса |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **test\_decimal** | **decimal(10,8)** |  |  |  |  |
| **s\_dwh** | **vocstatus\_sdim** | **evtime** | **string** | partitioned by hour (YYYYMMDDHH)  Дата получения информации | format\_to\_datetime | partition\_by\_hour | delta\_column |

**HIVE\_SCHEMAS\_FILE. Параметры схем БД Hive**

Данный файл описывает целевую hive-схему и соответствующие ей tmp-схему и hdfs-путь.

Разработчик должен дополнять файл новыми новыми схемами по мере необходимости.

Путь к данному файлу необходимо прописать в CONFIG.txt

**SOURCES\_CONNECTION\_FILE. Параметры подключений к источникам**

Данный файл описывает подключения к источникам.

Разработчик должен дополнять файл новыми новыми источниками по мере необходимости.

Поля:

* SOURCESYSTEM (поле 1) должно в точности соответствовать полю **Система источник**из **OBJECT\_REGISTRY\_FILE**
* TYPE (поле 2). Поддерживаются значения:
  1. **oracle**
  2. **sqlserver**
* CONNECTION (поле 3) - будет вставлено в соответсвующий man-поток
* USER (поле 4) - будет вставлено в соответсвующий man-поток
* PASSWORD (поле 5) - будет вставлено в соответсвующий man-поток

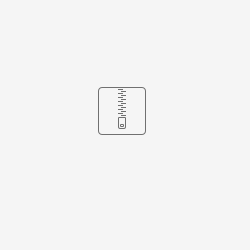
Например:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SOURCESYSTEM | TYPE | CONNECTION | USER | PASSWORD |
| DWH 2.0 Ядро DWH | oracle | jdbc:oracle:thin:@(DESCRIPTION = (ADDRESS = (PROTOCOL = TCP)(HOST = EXA2-SCAN)(PORT = 1521))(ADDRESS = (PROTOCOL = TCP)(HOST = EXA1-SCAN)(PORT = 1521))(CONNECT\_DATA =(SERVER = DEDICATED)(SERVICE\_NAME = DWSTPROD\_TAF))) | VSHILIN | REDACTED |

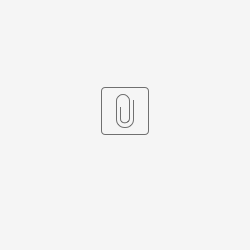
Путь к данному файлу необходимо прописать в CONFIG.txt

**Инструкция по генерации кода**

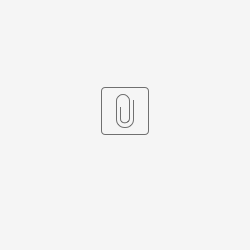
* Скачать утилиту job-generator. Положить ее в отдельный локальный каталог генератора.

[[](http://confluence.moscow.alfaintra.net/download/attachments/273422613/job_generator-2.0.jar?version=5&modificationDate=1562232007000&api=v2)job\_generator-2.0.jar](http://confluence.moscow.alfaintra.net/download/attachments/273422613/job_generator-2.0.jar?version=5&modificationDate=1562232007000&api=v2)

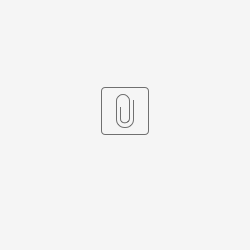
* Скачать файл с настройками HIVE. Положить его в локальный каталог генератора.

[[](http://confluence.moscow.alfaintra.net/download/attachments/273422613/HiveSchemas.csv?version=1&modificationDate=1556615630000&api=v2)HiveSchemas.csv](http://confluence.moscow.alfaintra.net/download/attachments/273422613/HiveSchemas.csv?version=1&modificationDate=1556615630000&api=v2)

* Скачать файл с настройками подключений к источникам. Заполнить его своей информацией для генерации объектов с разных источников. Положить его в локальный каталог генератора.

[[](http://confluence.moscow.alfaintra.net/download/attachments/273422613/Connections.csv?version=1&modificationDate=1556616411000&api=v2)Connections.csv](http://confluence.moscow.alfaintra.net/download/attachments/273422613/Connections.csv?version=1&modificationDate=1556616411000&api=v2)

* Получить актуальные файлы с реестрами объектов и полей. Например выкачать актуальную для текущей задачи ветку GIT в локальную рабочую копию. Найти в ней файлы реестров объектов и полей.
* Убедиться, что реестры объектов и полей заполнены в соответствии с описанными выше правилами.
* Скачать файл конфига. Настроить конфиг так, чтобы все его параметры смотрели на соответствующие файлы, настроенные ранее.

[[](http://confluence.moscow.alfaintra.net/download/attachments/273422613/CONFIG.csv?version=1&modificationDate=1556615466000&api=v2)CONFIG.csv](http://confluence.moscow.alfaintra.net/download/attachments/273422613/CONFIG.csv?version=1&modificationDate=1556615466000&api=v2)

* Запустить генерацию кода из командной строки.
  + При запуске нужно передать параметры запуска:
    1. БД Hive. Параметр должен соответствовать значению поля **БД Hive**из реестра объектов
    2. Имя таблицы/view. Параметр должен соответствовать значению пол **Имя таблицы/view**из реестра объектов
    3. Тип загрузки. Поддерживаются следующие типы загрузки:
       1. **simple\_nopart** - полная перегрузка без секционирования
       2. **simple\_part**- полная перегрузка с секционированием
       3. **delta\_nopart** - перегрузка по дельте без секционирования
    4. опциональный модификатор**drop\_tables**. Все hql-скрипты создания таблиц будут содержать пердварительную инструкцию drop для создаваемых объектов
    5. опциональный модификатор **ctl\_prototype**. Для объекта буден создан новый (или добавлен в существующий) прототип ctl-потока.
  + Например:
    1. java -jar job\_generator-2.0.jar S\_DWH vocstatus\_sdim simple\_part drop\_tables ctl\_prototype
    2. java -jar job\_generator-2.0.jar S\_DWH vocstatus\_sdim simple\_part "" ctl\_prototype
    3. java -jar job\_generator-2.0.jar S\_DWH vocstatus\_sdim simple\_part drop\_tables
    4. java -jar job\_generator-2.0.jar S\_DWH vocstatus\_sdim simple\_part