Основные требования:

1. Структура базы данных должна быть в PostgreSQL версии не ниже 12

1.1. Необходимо в одной базе данных создать две схемы. Первая схема ЕГРЮЛ вторая схема ЕГРИП

1.2. Кодировка базы данных должна быть UTF-8

1.3. Структура таблиц базы данных определены в xsd файлах аналогично названиям схем

1.4. Описание таблиц базы данных приведены в doc файлах аналогично названиям схем

1.5. Согласно пунктам 1.2. и 1.3. необходимо создать таблицы базы данных

1.6. Статические данные определены в таблице №1 (см в самом конце данного документа)

1.7. Перед загрузкой сведений регистр текстов приводится к одному ввиду

1.8. При загрузке статических данных (Таблица №1) информация должна проверяться на дубли

2. Вся среда разработки должна осуществляться на Unix подобных системах.

2.2. Предпочтительнее для парсера использовать язык программирования C++

2.3. Содержимое каталогов

В зависимости от вида предоставляемых сведений, загрузка производится из соответствующего каталога:

- EGRUL – в данный каталог хранятся сведения из ЕГРЮЛ
- EGRIP – в данный каталог хранятся сведения из ЕГРИП

В вышеуказанных каталогах содержатся следующие подкаталоги:

- Полные сведения из ЕГРЮЛ/ЕГРИП на начало года. В этом случае каталог имеет наименование вида 01.01.YYYY_FULL, где YYYY – год, по состоянию на начало которого выгружены сведения.

  Например, в каталоге \EGRUL\01.01.2016_FULL\ будут находиться сведения из ЕГРЮЛ по состоянию на 01.01.2016;
- Сведения о лицах, по которым в указанный день были внесены изменения. В этом случае каталог имеет наименование вида DD.MM.YYYY, где:
  - DD – день
  - MM – месяц
  - YYYYY – год, в котором были внесены изменения в ЕГРЮЛ/ЕГРИП.

  Например, в каталоге \EGRIP\12.01.2016\ будут находиться сведения о лицах, по которым 12 января 2016 были внесены изменения в ЕГРИП.

Архивы В каталоге, содержащем полные сведения или сведения о лицах, по которым в указанный день были внесены изменения в ЕГРЮЛ/ЕГРИП, содержатся архивы с файлами. Имя архива будет имеет вид:

- EGRUL_FULL_YYYY-MM-DD_N.zip – для архивов, содержащих полные сведения.
- EGRUL_YYY-MM-DD_N.zip – для архивов, содержащих сведения о лицах, по которым на указанную в имени файла дату были внесены изменения в ЕГРЮЛ/ЕГРИП.

  Где:
  - DD – день
  - MM – месяц
  - YYYYY – год
  - N – уникальный идентификатор архива, число

Файлы

В каждом архиве может быть до 100 xml-файлов, содержащих до 1000 сведений о лицах. Формат имени файла имеет следующий вид:

- EGRUL_FULL_YYYY-MM-DD_N.xml – для файлов, содержащих полные сведения.
- EGRUL_YYY-MM-DD_N.xml – для файлов, содержащих сведения о лицах, по которым на указанную в имени файла дату были внесены изменения в ЕГРЮЛ/ЕГРИП.

  Где:

- DD – день
- MM – месяц
- YYYYY – год
- N – уникальный идентификатор файла, число.

Таким образом, полный путь к файлу (включая архив) будет следующим:

Для ЕГРЮЛ:

- Полные сведения:

  \EGRUL\01.01.2016\EGRUL_FULL_2016-01-01.zip\EGRUL_FULL_2016-01-01_17648.xml
- Сведения о лицах, по которым на указанную в имени файла дату были внесены изменения:

  \EGRUL\12.01.2016\EGRUL_2016-01-12.zip\EGRUL_ 2016-01-12_17678.xml

Для ЕГРИП:

- Полные сведения: \EGRIP\01.01.2016\EGRIP_FULL_2016-01-01.zip\EGRIP_FULL_2016-01-01_18413.xml
- Сведения о лицах, по которым на указанную в имени файла дату были внесены изменения:

  \ EGRIP \12.01.2016\ EGRIP 2016-01-12.zip\EGRIP_ 2016-01-12_19010.xml

2.4. Формат файла

Каждый файл, содержащий полные сведения или сведения о лицах, по которым на указанную в имени файла дату были внесены изменения, представляет собой xml-файл в кодировке windows-1251, который имеет следующую структуру:

ЕГРЮЛ:

```
<?xml version="1.0" encoding="windows-1251"?>

<EGRUL ДатаВыг="2015-07-23">
<СвЮЛ ДатаВып="2015-07-23" …>…</СвЮЛ>
…
<СвЮЛ ДатаВып="2015-02-12" …>…</СвЮЛ>
</EGRUL
```

ЕГРИП:

```
<?xml version="1.0" encoding="windows-1251"?>

<EGRIP ДатаВыг="2015-07-23">
<СвИП ДатаВып="2015-07-23" …>…</СвИП >
…
< СвИП ДатаВып="2015-06-21" …>…</СвИП >
</ EGRIP
```

Каждый узел *СвЮЛ* или *СвИП* – соответствует выписке о юридическом лице или индивидуальном предпринимателе на указанную в атрибуте *ДатаВып* дату. Атрибут *ДатаВыг* содержит дату формирования файла.
Обработка ранее загруженных сведений

Сведения об одном и том же лице в течение года могут быть выгружены несколько раз. Таким образом, возникнет ситуация, когда в файле \EGRUL\12.01.2016\EGRUL_2016-01-12.zip\ EGRUL_2016-01-

12_17678.xml содержится информация о ЮЛ с ОГРН 1234567890123 по состоянию на 12 января 2016 года вида:
<?xml version="1.0" encoding="windows-1251"?>

<EGRUL ДатаВыг="2016-01-12">
<СвЮЛ ДатаВып="2016-01-12" ОГРН="1234567890123 ">…</СвЮЛ>
</EGRUL

в файле \EGRUL\09.07.2016\EGRUL_2016-07-09.zip\ EGRUL_2016-07-09_17899.xml содержится информация о ЮЛ с ОГРН 1234567890123 по состоянию на 9 сентября 2016 года вида:
<?xml version="1.0" encoding="windows-1251"?>

<EGRUL ДатаВыг="2016-07-09">
<СвЮЛ ДатаВып="2016-07-09" ОГРН="1234567890123 ">…</СвЮЛ>
</EGRUL

Примеры файлов во вложение с расширением xml

2.5. Парсер должен вести лог действий в текстовый файл следующего вида
2020-08-18 00:59:30,731 - Парсер - INFO - Запуск парсера.
2020-08-18 00:59:30,732 - EGRIP - INFO - Запуск парсера EGRIP.
2020-08-18 00:59:30,732 - EGRIP - INFO - Поиск zip-файлов.
2020-08-18 00:59:30,732 - EGRIP - INFO - Количество найденных zip-файлов - 1
2020-08-18 00:59:30,732 - EGRIP - INFO - Обработка найденых zip-файлов
2020-08-18 00:59:30,741 - ZipFile - INFO - Обработка файла: EGRIP_2018-08-01_1.zip
2020-08-18 00:59:30,741 - ZipFile - INFO - Начало обработки файла: EGRIP_2018-08-01_1.zip
2020-08-18 00:59:30,741 - ZipFile - INFO - Распаковка файлов...
2020-08-18 00:59:31,020 - ZipFile - INFO - Количество xml-файлов: 3
2020-08-18 00:59:31,025 - EGRIP-xml - INFO - Обработка файла - EGRIP_2018-08-01_344534.XML
2020-08-18 00:59:31,027 - EGRIP-xml - INFO - Обработка файла - EGRIP_2018-08-01_344535.XML
2020-08-18 00:59:31,028 - EGRIP-xml - INFO - Обработка файла - EGRIP_2018-08-01_344536.XML
2020-08-18 00:59:46,486 - EGRIP-xml - INFO - Конец обработки файла - EGRIP_2018-08-01_344534.XML
2020-08-18 01:00:13,112 - EGRIP-xml - INFO - Конец обработки файла - EGRIP_2018-08-01_344535.XML
2020-08-18 01:00:16,097 - EGRIP-xml - INFO - Конец обработки файла - EGRIP_2018-08-01_344536.XML
2020-08-18 01:02:50,464 - ZipFile - INFO - Конец обработки файла: EGRIP_2018-08-01_1.zip
2020-08-18 01:02:50,467 - EGRIP - INFO - Общее количество вставленных СвИП - 2466
2020-08-18 01:02:50,470 - EGRIP - INFO - Конец обработки zip-файлов.
2020-08-18 01:02:50,471 - Парсер - INFO - Парсер EGRIP завершил работу успешно.
2020-08-18 01:02:50,471 - Парсер - INFO - Парсер закончил работу.
Примерный код написанный на языке python во вложение название файла runner.py необходимо переписать в C++

2.5.1. Также необходимо в каждой из схем создать таблицу для технических сведений такие как дату загрузки, название обработанных папок, название архива, название файла, а также размер архива, файла
Примерный вид технической таблицы:

Дата загрузки	Название Папка	Название Архива	Название файла	Размер архива	Размер файла
18.08.2020	16.08.2020	EGRUL_2020-08-16_1.zip	EGRUL_2020-08-16_456296.XML	925 926 байт	11 034 624 байт
18.08.2020	16.08.2020	EGRUL_2020-08-16_1.zip	EGRUL_2020-08-16_456297.XML	925 926 байт	8 934 624 байт

2.5.2. Парсер должен записывать в таблицу п.2.5.1. сведения в соответствие с примером

2.5.2. Загрузка начинает с папки 01.01.2018_FULL если он ранее загружен уже то загружает следующую папку 02.01.2018 итд до 31.12.2018 затем грузит 01.01.2019_FULL и далее 02.01.2019 и каждый день смотрит если папки нет то идет дальше имеется ввиду что может быть так что папки с датой 02.01.2019 нет а есть папка 03.01.2019

2.6. Скрипт распаковывает архив во временную папку

2.7. Скрипт автоматический удаляет загруженные файлы во временной папке

2.8. При следующем запуске скрипта ранее отработанные папки, архивы, файлы, не должны обрабатываться за исключением если размер файлов, архивов изменился то тогда обрабатывать архивы файлы необходимо заново методом UPDATE

Таблица №1

До внесения в базу данных регистр привести в единый вид затем проверить записи на дубли в базе данных если идентичной записи нет, то тогда вставить запись если запись есть, то добавить только ссылку на id записи

ВидЗап
Город
НаселПункт
Район
Регион
СвНО
СвОКВЭД
СвОргПФ
СвОргФСС
СвРегОрг
СвСтатус
СвПрекращ