**ТЕХНИЧЕСКОЕ ЗАДАНИЕ**

Необходимо собрать витрины с помощью Spark/Scala (Spark версия 2.4.0, scala 2.11.12, java 1.8.131)

Подразумевается, что все таблицы-источники лежат в Hadoop в базе данных TEST\_DB

1. Необходимо реализовать алгоритм, описанный в файле, также предполагается, что функционал отрабатывает на ежедневной основе

Структура проекта должна соответствовать данному примеру на рисунке 1

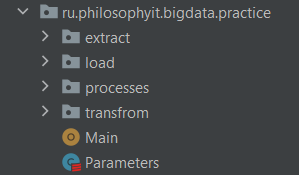


Рис.1 Структура проекта

-- extract - для считывания источников и применения к ним фильтров

-- transform - сборка витрин

-- load - запись витрин

-- processes – инкрементальная загрузка

файл Parameters – объявление (и парсинг) глобальных переменных и их определение с помощью аргументов, передающихся при запуске программы

1. должна быть предусмотрена тестовая сборка на local в IntelliJ IDEA для проверок, разработать собственную синтетику
2. Написать DAG-файл для запуска сборки витрин на ежедневной основе с помощью оркестратора Airflow
3. Необходимо приложить собранные эталоны витрин в формате .csv

**spark.sql для решения данного задания запрещен**

**ОПИСАНИЕ ТАБЛИЦ**

Таблица STUDENT\_DME\_44

|  |  |  |
| --- | --- | --- |
| Название колонки | Тип | Описание |
| stud\_id (PK) | Integer | Уникальный номер студента |
| spec\_id (FK) | String | Номер специальности на которой учится студент |
| last\_name | String | Фамилия |
| name | String | Имя |
| middle\_name | String | Отчество |
| birth\_date | String | Дата рождения |
| gender | String | Пол |

Таблица FORM\_DME\_44

|  |  |  |
| --- | --- | --- |
| Название колонки | Тип | Описание |
| form\_id (PK) | Integer | Номер договора (у одного stud\_id может быть несколько договоров) |
| stud\_id (FK) | Integer | Номер студента |
| form\_of\_education | String | Форма обучения (очная/заочная) |
| finance | String | Финансирование (бюджет/договор) |
| course | Integer | Курс поступления/изменения условий обучения/отчисления |
| education\_start | String | Дата начала обучения |
| date\_apply | String | Полная дата, когда было поступление/изменение условий обучения/отчисления |
| code | Integer | Учится(0) /Переведен из другого ВУЗа(1) /Сменил форму финансирования(2) /Отчислен(4) |

Таблица SPECIALIZATION\_DME\_44

|  |  |  |
| --- | --- | --- |
| Название колонки | Тип | Описание |
| spec\_id | String | Номер специальности |
| filial\_id | Integer | Идентификатор филиала |
| spec\_name | String | Название специальности |
| years | Integer | Годы обучения |
| cost | Integer | Стоимость за курс |
| quota | Integer | Количество человек |

Таблица FILIAL\_DME\_44

|  |  |  |
| --- | --- | --- |
| Название колонки | Тип | Описание |
| filial\_id | Integer | Идентификатор филиала |
| university | String | Название университета |
| filial\_name | String | Название филиала |
| filial\_address | String | Адрес филиала |
| filial\_number | String | Номер телефона филиала |

Необходимо собрать 5 витрин

1. Витрина PAYMENT\_DME\_44

Добавлять записи инкрементом

Должна содержать студентов, обучающихся только на платной основе, подсчеты на каком курсе и семестре находится сейчас студент и сколько он заплатил за обучение и учтены переводы на бюджет, отчисление и семестры

|  |  |  |
| --- | --- | --- |
| Исходная таблица | Название колонки | Описание |
| student | stud\_id | Уникальный код студента |
|  | fio | Фамилиям Имя Отчество |
| specialization | spec\_id | Номер специальности |
| specialization | spec\_name | Название специальности |
| filial | filial\_name | Название филиала |
| form | form\_id | Номер договора |
| form | finance | Форма обучения |
|  | total\_course | Текущий курс (учитывать кол-во годов обучения из таблицы specialization) |
|  | total\_semester | Текущий семестр |
| specialization | cost | Стоимость обучения в год |
|  | pay | Сколько студент заплатил за обучение |
|  | total\_pay\_semester | Количество семестров за которые платил студент |
|  | status | статус «отчислен» с указанием курса и семестра, на котором был отчислен, для обучающихся и переведенных статус «обучается» с такого-то курса с такого-то семестра |
| form | code | Код обучения |
|  | current\_date | текущая дата |

ПРИМЕР: Студент поступил на 1 курс 01.08.2022 – исходя из текущей даты понимаем, что он на 1 курсе 1 семестр  
Студент поступил на 1 курс 01.08.2020 – исходя из текущей даты можем вычислить, что студент на 3 курсе 5 семестр

1. Витрина BUDGET\_DME\_44

Добавлять записи инкрементом

Должна содержать в себе студентов, учащихся только на бюджетной основе, нужно учитывать тех студентов, которые перевелись на бюджет во время обучения

|  |  |  |
| --- | --- | --- |
| Исходная таблица | Название колонки | Описание |
| student | stud\_id | Уникальный код студента |
|  | fio | Фамилиям Имя Отчество |
| specialization | spec\_id | Номер специальности |
| specialization | spec\_name | Название специальности |
| filial | filial\_name | Название филиала |
| form | form\_id | Номер договора |
| form | finance | Финансирование |
|  | total\_course | Текущий курс (учитывать кол-во годов обучения из таблицы specialization) |
|  | total\_semester | Текущий семестр |
|  | status | статус «отчислен» с указанием курса и семестра, на котором был отчислен, для обучающихся и переведенных статус «обучается» с такого-то курса с такого-то семестра |
| form | code | Код обучения |
|  | current\_date | текущая дата |

1. Витрина STUDY\_DME\_44

Добавлять записи инкрементом

Должна содержать в себе только обучающихся студентов

|  |  |  |
| --- | --- | --- |
| Исходная таблица | Название колонки | Описание |
| student | stud\_id | Уникальный код студента |
|  | fio | Фамилиям Имя Отчество |
| specialization | spec\_id | Номер специальности |
| specialization | spec\_name | Название специальности |
| filial | filial\_name | Название филиала |
| form | form\_id | Номер договора |
| form | finance | Финансирование |
|  | total\_course | Текущий курс (учитывать кол-во годов обучения из таблицы specialization) |
|  | total\_semester | Текущий семестр |
|  | status | для обучающихся и переведенных статус «обучается» с такого-то курса с такого-то семестра |
| form | code | Код обучения |
| form | max\_date | Дата изменения договора |
|  | current\_date | текущая дата |

1. Витрина EXPELLED\_DME\_44

Добавлять записи инкрементом

Должна содержать только отчисленных студентов

|  |  |  |
| --- | --- | --- |
| Исходная таблица | Название колонки | Описание |
| student | stud\_id | Уникальный код студента |
|  | fio | Фамилиям Имя Отчество |
| specialization | spec\_id | Номер специальности |
| specialization | spec\_name | Название специальности |
| filial | filial\_name | Название филиала |
| form | form\_id | Номер договора |
| form | finance | Финансирование |
|  | total\_course | Текущий курс (учитывать кол-во годов обучения из таблицы specialization) |
|  | total\_semester | Текущий семестр |
|  | status | статус «отчислен» с указанием курса и семестра, на котором был отчислен |
| form | code | Код обучения |
|  | current\_date | текущая дата |

1. Витрина CALC\_DME\_44

Должна содержать подсчет студентов в каком филиале и на каком курсе обучаются

Перезаписывать при каждом запуске

Разбить на 4 стейджа

Stage 1 – подсчет сколько студентов на специальностях

|  |  |  |
| --- | --- | --- |
| Исходная таблица | Название колонки | Описание |
|  | stg | Содержит в себе литерал с названием стейджа (пример: select 'SPEC' stg) |
| specialization | spec\_id.as(“id”) | Номер специальности |
| specialization | spec\_name.as(“name”) | Название специальности |
|  | count | Подсчет студентов |

Stage 2 – подсчет студентов в филиалах

|  |  |  |
| --- | --- | --- |
| Исходная таблица | Название колонки | Описание |
|  | stg | Содержит в себе литерал с названием стейджа (пример: select 'FILIAL' stg) |
| filial | filial\_id.as(“id”) | Номер специальности |
| filial | filial\_name.as(“name”) | Название специальности |
|  | count | Подсчет студентов |

Stage 3 – подсчет студентов по курсам

|  |  |  |
| --- | --- | --- |
| Исходная таблица | Название колонки | Описание |
|  | stg | Содержит в себе литерал с названием стейджа (пример: select 'COURSE' stg) |
| form | filial\_id.as(“id”) | Номер специальности |
|  | name | select 'курс' name |
|  | count | Подсчет студентов |

Stage 4 – union stage 1, stage2, stage 3