## **Анализ вакансий hh.ru для заказчика.**

## **Работа с БД в PostgreSQL**

## **1. Введение. Знакомство с датасетом**

## **✍** Большинство крупных компаний хранят данные в реляционных базах, так как в таком виде они более структурированы и адаптированы для внешних систем, с помощью которых, как правило, данные в базу и заносятся. В системах с такими базами данных имеют значение именно отношения между сущностями. Если вам понадобилось, например, посчитать статистику по заказам, то здесь важны отношения: заказ, клиент, товар и т. д.

## → Мы переделали уже знакомый вам по предыдущему проекту датасет *hh.ru* в реляционный, чтобы вы могли попрактиковаться и в этой области.

## Представьте, что вы работаете в некотором дата-центре по трудоустройству, куда стекаются данные с различных платформ для найма. Вам необходимо познакомиться с данными, понять, с какими резюме вы имеете дело, а дальше собрать статистику для различных кадровых агентств и статистических центров.

## → Вы получите данные по возрасту кандидатов, по городам с наиболее активным рынком труда, поймёте специфику найма (где какие вакансии более активны), а также узнаете желаемые позиции для тех или иных кандидатов.

## Проект включает в себя несколько этапов:

## знакомство с датасетом;

## предварительный анализ данных;

## анализ кандидатов;

## глобальный анализ показателей.

## ⚡ В качестве итогового задания проекта вы подготовите текстовый Google-документ, в котором по каждой встречающейся вам в этом модуле задаче приведёте ваш ответ в формате:

## номер задания;

## код с пояснениями;

## результат запроса (в виде таблицы);

## общий вывод в конце по результатам анализа рынка труда.

## ЗНАКОМСТВО С ДАТАСЕТОМ

## Напоминаем, что датасет, с которым мы будем работать, уже вам знаком. Мы взяли те же данные *HeadHunter*, только разбили их по таблицам и перенесли в [БД SQL](http://sql.skillfactory.ru:3000/browse/2/schema/hh). Такой процесс называется нормализацией (вы можете почитать про неё дополнительно, если вам понадобятся знания по проектированию баз данных).

## 

## 

## 

## 

## 

## 

## 

## На диаграмме мы видим четыре таблицы:

## 

## **CANDIDATE**

## Таблица хранит в себе общие данные по кандидатам: *id*, пол, возраст, желаемая должность, город, вид занятости, текущая должность, дата обновления записи и зарплата.

## 

## 

## 

## 

## 

## 

## 

## **CITY**

## city — таблица-справочник для наших кандидатов — хранит код города и его название.

## 

## **CANDIDATE\_TIMETABLE\_TYPE**

## Это дополнительная таблица. Она существует для организации связи *многие-ко-многим*, так как у нас есть много кандидатов и у них может быть несколько подходящих типов рабочего графика.

## Примечание. Нам стоило бы вынести в дополнительную таблицу и колонку employment\_type из таблицы candidate, так как вид связи здесь одинаковый, но мы решили оставить всё как есть, так как это даст нам дополнительные возможности для постановки интересных задач и проведения изысканий.

## 

## **TIMETABLE\_TYPE**

## Это таблица-справочник вариантов рабочего графика, подходящего кандидату.

## 

## 

## **2. Предварительный анализ данных**

### 

### Задание 2.1

Рассчитайте максимальный возраст (max\_age) кандидата в таблице.

2.1.1 Код:

SELECT

max(age) /\*определяем максимальный возраст в колонке age \*/

FROM hh.candidate /\*из таблицы hh.candidate\*/

2.1.2 Результат:

# **100**

Вывод: максимальный возраст кандидата равен 100 годам, что является неадекватной информацией.

### Задание 2.2

Теперь давайте рассчитаем минимальный возраст (min\_age) кандидата в таблице.

Код:

SELECT

min(age) /\*определяем максимальный возраст в колонке age \*/

FROM hh.candidate /\*из таблицы hh.candidate\*/

Результат:

# **14**

Вывод: минимальный возраст кандидата равен 14 годам, что является адекватной информацией, но требует дополнительного анализа.

### Задание 2.3

Попробуем «почистить» данные. Напишите запрос, который позволит посчитать для каждого возраста (age) сколько (cnt) человек этого возраста у нас есть.

Отсортируйте результат по возрасту в обратном порядке.

Код:

SELECT

c.age /\*выбираем столбец age\*/,

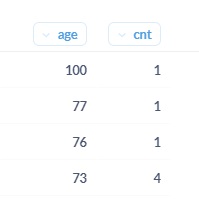
count(c.age) AS cnt /\*выбираем столбец age и создаем ему алиас cnt\*/

FROM hh.candidate c /\*из таблицы hh.candidate и создаем ей алиас c\*/

GROUP BY c.age /\*группируем по возрасту age\*/

ORDER BY c.age DESC /\*сортируем по возрасту age в обратном порядке (по убыванию)\*/

Результат:



Вывод: Большая часть кандидатов в возрасте от 16 до 75 лет

### Задание 2.4

По данным Росстата, средний возраст занятых в экономике России составляет 39.7 лет. Мы округлим это значение до 40. Найдите количество кандидатов, которые старше данного возраста. *Не забудьте отфильтровать «ошибочный» возраст 100.*

Код:

SELECT

COUNT(\*) /\*число кандидатов\*/

FROM hh.candidate /\*из таблицы hh.candidate\*/

WHERE

(age > 40 AND age != 100)/\*по условию, что возраст больше 100 лет и не равен 100\*/

Результат:

# **6,263**

Вывод: количество кандидатов в возрасте более 40 лет достаточно велико.

## **3. Глобальный анализ показателей**

### Задание 3.1

Для начала напишите запрос, который позволит узнать, сколько (cnt) у нас кандидатов из каждого города (city).

Отсортируйте результат по количеству в обратном порядке.

Код:

SELECT

cc.title city /\*столбец title с алиасом city из таблицы с алиасом cc\*/,

count(c.id) AS cnt /\*количество кандидатов в столбце id с алиасом cnt в таблице с алиасом c\*/

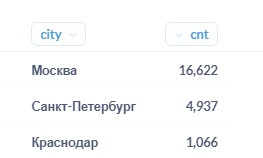
FROM hh.candidate c /\*таблица hh.candidate с алиасом c\*/

JOIN hh.city cc ON c.city\_id=cc.id /\*оператор соединения таблиц; таблица hh.city с алиасом cc; условие: city\_id таблицы c равен id таблицы cc\*/

GROUP BY cc.title /\*группируем по столбцу title из таблицы с алиасом cс\*/

ORDER BY cnt DESC /\*сортируем по количеству кандидатов в обратном порядке (по убыванию)\*/

Результат:



Вывод: наибольшее количество кандидатов из Москвы и Санкт-Петербурга

### Задание 3.2

Москва бросается в глаза как, пожалуй, самый активный рынок труда. Напишите запрос, который позволит понять, каких кандидатов из Москвы устроит «проектная работа».

**Формат выборки:** gender, age, desirable\_occupation, city, employment\_type.

Отсортируйте результат по *id* кандидата.

Код:

SELECT

c.gender /\*столбец gender из таблицы с алиасом c\*/,

c.age /\*столбец age из таблицы с алиасом c\*/,

c.desirable\_occupation /\*столбец desirable\_occupation из таблицы с алиасом c\*/,

cc.title as city/\*столбец title с алиасом city из таблицы с алиасом cс\*/,

c.employment\_type /\*столбец employment\_type из таблицы с алиасом cс\*/

FROM

hh.candidate c /\*таблица hh.candidate с алиасом c\*/

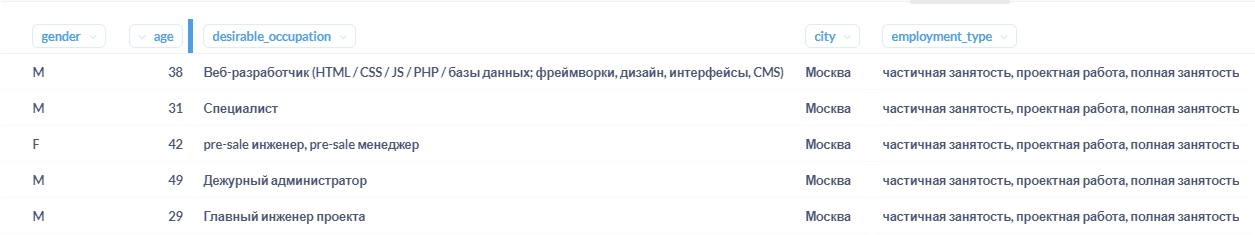
JOIN hh.city cc on c.city\_id = cc.id /\*оператор соединения таблиц; таблица hh.city с алиасом cc; условие: city\_id таблицы c равен id таблицы cc\*/

WHERE

cc.title = 'Москва' /\* условие: title таблицы cс равен 'Москва'\*/

AND c.employment\_type ilike '%проектная работа%' /\* условие: в employment\_type таблицы c встречается 'проектная работа' (без учета регистра)\*/

Результат:



Вывод: проектная работа является достаточно популярным у кандидатов.

### Задание 3.3

Данных оказалось многовато. Отфильтруйте только самые популярные *IT*-профессии — разработчик, аналитик, программист.

Обратите внимание, что данные названия могут быть написаны как с большой, так и с маленькой буквы.

Отсортируйте результат по *id* кандидата.

Код:

SELECT

c.gender /\*столбец gender из таблицы с алиасом c\*/,

c.age /\*столбец age из таблицы с алиасом c\*/,

c.desirable\_occupation /\*столбец desirable\_occupation из таблицы с алиасом c\*/,

cc.title as city /\*столбец title с алиасом city из таблицы с алиасом cс\*/,

c.employment\_type /\*столбец employment\_type из таблицы с алиасом c\*/

FROM

hh.candidate c /\*таблица hh.candidate с алиасом c\*/

JOIN hh.city cc on c.city\_id = cc.id /\*оператор соединения таблиц; таблица hh.city с алиасом cc; условие: city\_id таблицы c равен id таблицы cc\*/

WHERE

cc.title = 'Москва' /\* условие: title таблицы cс равен 'Москва'\*/

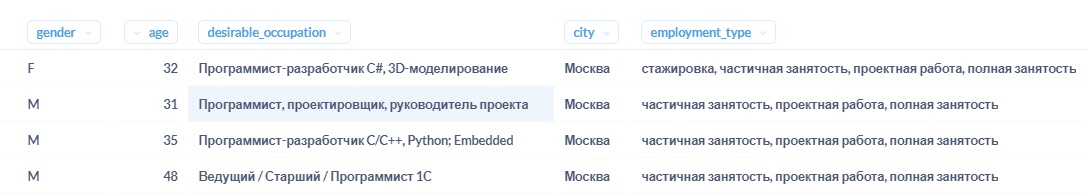
AND c.employment\_type ilike '%проектная работа%' /\* условие: в employment\_type таблицы c встречается 'проектная работа' (без учета регистра)\*/

AND (c.desirable\_occupation ilike '%разработчик%' /\* условие: в desirable\_occupation таблицы c встречается 'разработчик' (без учета регистра)\*/

OR c.desirable\_occupation ilike '%аналитик%' /\* условие: в desirable\_occupation таблицы c встречается ‘аналитик' (без учета регистра)\*/

OR c.desirable\_occupation ilike '%программист%') /\* условие: в desirable\_occupation таблицы c встречается ‘программист' (без учета регистра)\*/

Результат:



Вывод: у разработчиков, аналитиков и программистов проектная работа достаточно привлекательна.

### Задание 3.4

Для общей информации попробуйте выбрать номера и города кандидатов, у которых занимаемая должность совпадает с желаемой.

**Формат выборки:** id, city.

Отсортируйте результат по городу и *id* кандидата.

Код:

SELECT

c.id /\*столбец id из таблицы с алиасом c\*/,

cc.title as city /\*столбец title с алиасом city из таблицы с алиасом cс\*/

FROM

hh.candidate c /\*таблица hh.candidate с алиасом c\*/

JOIN hh.city cc on c.city\_id = cc.id /\*оператор соединения таблиц; таблица hh.city с алиасом cc; условие: city\_id таблицы c равен id таблицы c\*/

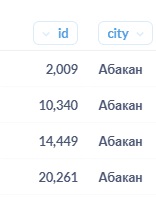
WHERE

(c.current\_occupation = c.desirable\_occupation

AND c.age= '100') /\* условие: содержимое current\_occupation равно desirable\_occupation таблицы c и age таблицы с алиасом c не равен 100\*/

ORDER BY cc.title, c.id /\*сортируем по title из таблицы с алиасом cс и id из таблицы с алиасом c\*/

Результат:



Вывод: кандидаты в основном ищут работу похожую занимаемой должности с более высокой заработной платой.

### Задание 3.5

Определите количество кандидатов пенсионного возраста.

*Пенсионный возраст для мужчин наступает в 65 лет, для женщин — в 60 лет.*

Код:

SELECT

COUNT (\*) /\*число кандидатов\*/

FROM hh.candidate c /\*таблица hh.candidate с алиасом c\*/

WHERE

c.age != 100

AND (c.gender = 'M' AND c.age>=65)

OR (c.gender = 'F' AND c.age>=60) /\* условие: age таблицы с алиасом c не равен 100\* и ( age больше или равно 65 при gender равном 'M' или age больше или равно 60 при gender равном 'F' таблицы с алиасом c )\*/

Результат:

# **75**

Вывод: количество кандидатов пенсионного возраста мало и при анализе может не учитываться.

## **4. Анализ кандидатов для заказчиков**

### Задание 4.1

1 point possible (graded)

Для добывающей компании нам необходимо подобрать кандидатов из Новосибирска, Омска, Томска и Тюмени, которые готовы работать вахтовым методом.

**Формат выборки:** gender, age, desirable\_occupation, city, employment\_type, timetable\_type.

Отсортируйте результат по городу и номеру кандидата.

Код:

SELECT

c.gender /\*столбец gender из таблицы с алиасом c\*/,

c.age /\*столбец age из таблицы с алиасом c\*/,

c.desirable\_occupation /\*столбец desirable\_occupation из таблицы с алиасом c\*/,

cc.title AS city /\*столбец title с алиасом city из таблицы с алиасом cс\*/,

c.employment\_type /\*столбец employment\_type из таблицы с алиасом c\*/,

tt.title AS timetable\_type /\*столбец title с алиасом timetable\_type из таблицы с алиасом tt\*/

FROM hh.candidate c /\*таблица hh.candidate с алиасом c\*/

JOIN hh.city cc ON c.city\_id=cc.id /\*оператор соединения таблиц; таблица hh.city с алиасом cc; условие: city\_id таблицы c равен id таблицы cc\*/

JOIN hh.candidate\_timetable\_type ctt on ctt.candidate\_id=c.id /\*оператор соединения таблиц; таблица hh.candidate\_timetable\_type с алиасом ctt; условие: candidate\_id таблицы ctt равен id таблицы c\*/

JOIN hh.timetable\_type tt on tt.id=ctt.timetable\_id /\*оператор соединения таблиц; таблица hh.timetable\_type с алиасом tt; условие: id таблицы tt равен timetable\_id таблицы ctt\*/

WHERE

( cc.title = 'Новосибирск'

OR cc.title = 'Омск'

OR cc.title = 'Томск'

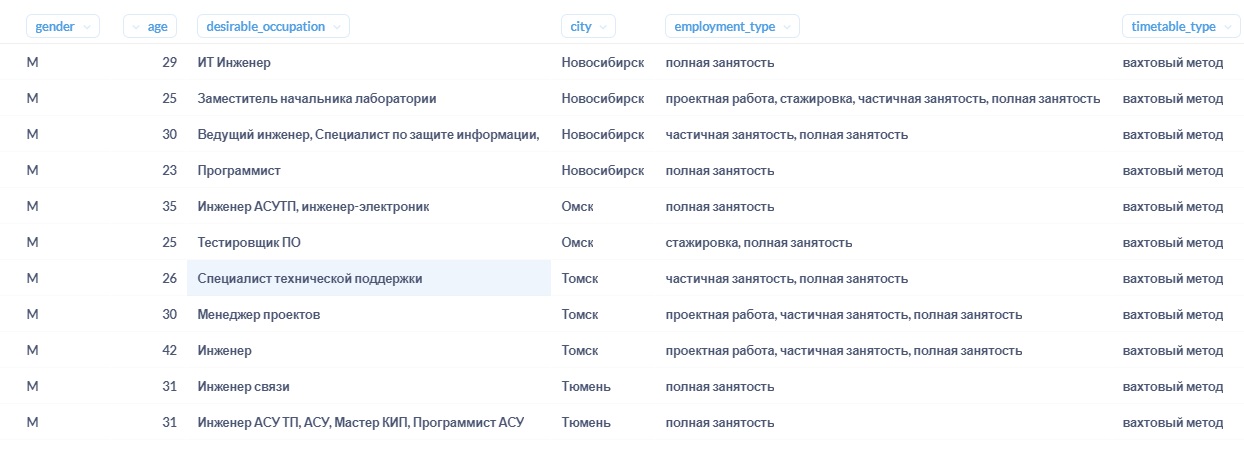
OR cc.title = 'Тюмень')

AND tt.title = 'вахтовый метод'/\* условие: title с алиасом таблицы tt равен 'вахтовый метод'\*/

AND c.age != 100 /\* условие: title таблицы с алиасом cс равен 'Новосибирск' или 'Омск' или 'Томск' или 'Тюмень' и age таблицы с алиасом c не равен 100\*/

ORDER BY cc.title, c.id /\*сортируем по title из таблицы с алиасом cс и id из таблицы с алиасом c\*/

Результат:



Вывод: кандидатов из Новосибирска, Омска, Томска и Тюмени согласных работать вахтовым методом всего 11, что говорит о непривлекательности такого вида занятости.

### Задание 4.2

Для заказчиков из Санкт-Петербурга нам необходимо собрать список из 10 желаемых профессий кандидатов из того же города от 16 до 21 года (в выборку включается 16 и 21, сортировка производится по возрасту) с указанием их возраста, а также добавить строку Total с общим количеством таких кандидатов. Напишите запрос, который позволит получить выборку вида:

Код:

(SELECT

c.desirable\_occupation /\*столбец desirable\_occupation из таблицы с алиасом c\*/,

c.age /\*столбец age из таблицы с алиасом c\*/

FROM hh.candidate c /\*таблица hh.candidate с алиасом c\*/

JOIN hh.city cc on cc.id = c.city\_id /\*оператор соединения таблиц; таблица hh.city с алиасом cc; условие: city\_id таблицы c равен id таблицы cc\*/

WHERE

cc.title = 'Санкт-Петербург' and c.age between 16 and 21 /\* условие: title таблицы с алиасом cс равен 'Санкт-Петербург' и age таблицы с алиасом c равен числу от 16 до 21\*/

ORDER BY c.age /\*сортируем по age из таблицы с алиасом c\*/

LIMIT 10 /\*выбираем первые 10 строк\*/

)

UNION ALL /\*оператор присоединения строк\*/

(SELECT

'Total '/\*значение ‘Total’ ячейки первого столбца\*/,

count(\*) /\*общее количество кандидатов выбранных профессий\*/

FROM hh.candidate c /\*таблица hh.candidate с алиасом c\*/

JOIN hh.city cc on cc.id = c.city\_id /\*оператор соединения таблиц; таблица hh.city с алиасом cc; условие: city\_id таблицы c равен id таблицы cc\*/

WHERE

cc.title = 'Санкт-Петербург' AND c.age BETWEEN 16 AND 21) /\* условие: title таблицы с алиасом cс равен 'Санкт-Петербург' и age таблицы с алиасом c равен числу от 16 до 21\*/

Результат:



Вывод: 10 наиболее привлекательных профессий в Санкт-Петербурге связаны с IT-сферой и возраст большинства кандидатов в выбранном диапазоне равен 18.

Общий вывод: Профессии связанные с информационными технологиями очень популярны среди кандидатов. Обусловлено это быстрыми темпами развития рынка труда в этой сфере, востребованностью профессионалов и, как следствие, высокими заработными платами. Основная масса кандидатов в возрасте до 65 лет. У кандидатов набирают популярность проектная работа, дистанционная работа.