Кейс-чемпионат «GPN Intelligence CUP»

направление «Разработка алгоритмов BIG DATA»

Погрузитесь в атмосферу нашего бизнеса, решите аналитическую задачу и станьте стажером нашей компании

ОГЛАВЛЕНИЕ

О КОМПАНИИ

1 **«Газпром нефть»** — вертикально-интегрированная нефтяная компания, основные виды деятельности которой — разведка и разработка месторождений нефти и газа, нефтепереработка, а также производство и сбыт нефтепродуктов.

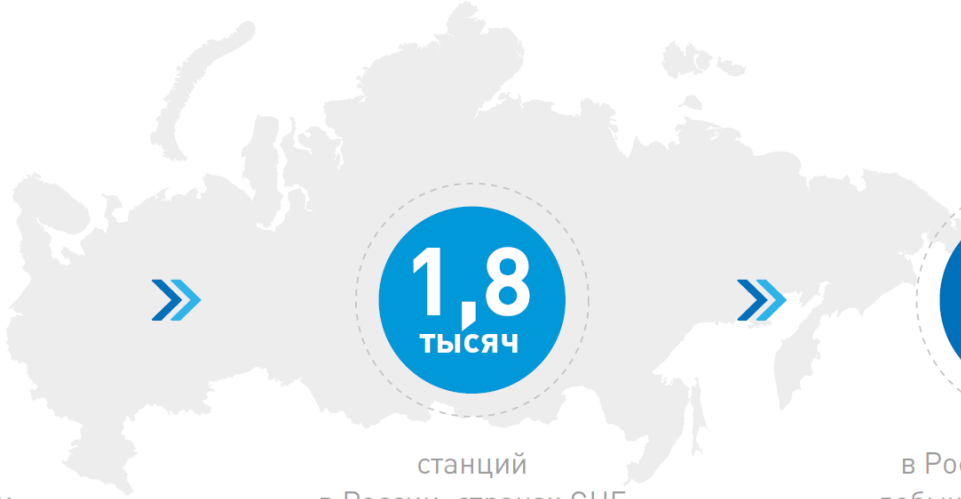
2 В 2018 году «Газпром нефть» достигла рекордных финансовых результатов, получив самую высокую прибыль в своей истории. Компания **делает ставку на технологическое развитие**, внедряет передовые решения для достижения стратегических целей.

3 **В структуру** «Газпром нефти» **входят более 70** нефтедобывающих, нефтеперерабатывающих и сбытовых предприятий в России, странах ближнего и дальнего зарубежья.

4 **По объему добычи** углеводородов «Газпром нефть» **входит в тройку крупнейших** компаний России. «Газпром нефть» стремится внедрять в своей работе передовые методики разведки, добычи и переработки нефти.

5 **Продукция «Газпром нефти»** экспортируется **более чем в 50 стран мира** и реализуется на всей территории России и за рубежом через разветвленную сеть собственных сбытовых предприятий. В настоящее время **сеть АЗС** компании насчитывает более **1,8 тыс. станций** в России, странах СНГ и Европы.

О КОМПАНИИ



1 место

в России
по объему прокачки
через **1 АЗС (20,7 т/сут)**

1,8 тысяч

станций
в России, странах СНГ
и Европы

3 место

в России по объему
добычи углеводородов
(92,9 млн тонн н.э.)

СБЫТОВОЙ БИЗНЕС КОМПАНИИ

Также в сбытовой блок моторных топлив входят дочерние общества, которые осуществляют хранение, перевалку и доставку топлива (ГПН-Транспорт, ГПН-Терминал).

СБЫТОВОЙ БЛОК МОТОРНЫХ ТОПЛИВ «ГАЗПРОМ НЕФТИ» ОБЪЕДИНЯЕТ ДОЧЕРНИЕ ОБЩЕСТВА И ПОДРАЗДЕЛЕНИЯ:

- **клиенты сети АЗС «Газпромнефть»;**
- **корпоративные клиенты** (ГПН-КП, заправляются по сервисным картам);
- **мелкооптовые клиенты** (ГПН-РП, предприятия и независимые АЗС).

СБЫТОВОЙ БИЗНЕС КОМПАНИИ

1 820 штук
Сеть АЗС

11 миллионов
лояльных клиентов*

> 54 000
корпоративных клиентов

ДИФФЕРЕНЦИАЦИЯ УСЛУГ:

магазины

кафе

мойки

ОБЪЕКТЫ УПРАВЛЕНИЯ:

бензовозы / газовозы

нефтебазы

Центр аналитических решений ДРП

1. Разработка аналитических решений

- BI-приложения
- Хранилища данных для аналитики
- Решения для «больших данных»

2.

- Управление качеством данных
- Единый каталог данных и бизнес-словарь
- База знаний и консультирование по данным

3. Центр компетенций Data Science

- Модели и прототипы на данных
- Проверка гипотез
- Участие в проектах с аналитической составляющей

4. Бизнес-партнерство в части новых проектов

- Оценка новых инициатив с точки зрения данных

5. Управление аналитической инфраструктурой

- Аналитические песочницы и инструменты
- Инжиниринг и окружение

6. Обучения по методам и инструментам анализа данных

- Python, R, SQL, Qlik, Power BI
- Методы анализа данных
- Машинное обучение

ОПИСАНИЕ КЕЙСА

Задание 1. Bash-скрипты

Необходимо написать параметризированный bash-скрипт «search_by_path.sh», основной задачей которого является получение информации о txt-файлах в указанной директории.

Входящие параметры bash-скрипта:
PATH - путь к директории, в которой находятся папки и файлы, обязательный
DATETIME – дата и время последнего изменения файла, обязательный
OUT_CSV_FILE_NAME – путь и название csv-файла с результатами обработки, обязательный

Функциональные требования к bash-скрипту:
В указанной директории (PATH) найти все txt-файлы, которые были изменены или созданы начиная с даты (DATETIME) включительно, и вывести их в консоль терминала в формате: полный путь к файлу (path, название, расширение)
Получить указанные ниже атрибуты для каждого txt-файла и записать их в результирующий файл OUT_FILE_NAME в формате CSV. Порядок атрибутов соответствует порядку столбцов в результирующем файле, название столбцов не указывать:

- дата сканирования;
- путь до файла;
- название файла;
- дата-время последнего изменения файла;
- дата-время последнего доступа к файлу;
- размер файла(Mb);
- число строк в файле;

Желательно выполнить задание с использованием пользовательских функций и дескрипторов ввод-вывод.

Требования к результату:
В качестве результата ожидается файл с названием <Фамилия Имя Отчество>_task01.txt, в котором будет содержаться полный текст скрипта и команда для его запуска с краткими комментариями.

ОПИСАНИЕ КЕЙСА

Задание 2- Регулярные выражения

На входе вашего ETL-процесса, есть ***txt-файл (case02_file.txt) (см. «материалы»)***, полученный от системы хранения метаинформации. Система имеет свой формат выгрузки данных, которые изменить невозможно.

Ваша задача – написать bash-скрипт/команду для получения из файла всех текстовых идентификаторов систем в виде очищенного списка. Текстовым идентификатором систем считается значение, заключенное в квадратные скобки, идущее после комбинации символов “ID: ”.
Пример: [ID: WAREHOUSE], значение WAREHOUSE.

Требования к результату:
В качестве результата ожидается файл <Фамилия Имя Отчество>_task02.txt, в котором будет содержаться полный текст bash-скрипта или команда с краткими комментариями.

ОПИСАНИЕ КЕЙСА

Задание 3 - Работа с Python

В качестве источника данных у вас есть созданный сотрудником компании ***excel-файл case03_input_file.xlsx (см. «материалы»)***, содержащий в себе коэффициент эффективности партнёрской сети в стандартной для всех партнеров объемов продаж.
К сожалению, данные в файле структурированы не самым оптимальным для загрузки в hdfs образом. См. лист data.

Описание файла:

1. Размер таблиц всегда одинаковый (9 строк, 9 столбцов).
2. В одном файле содержится только один лист с данными и в нем может быть более 10 000 таблиц.
3. Таблицы всегда разделены пустой строкой.
4. Значения «диапазон», всегда одинаковы: 0 – 10, 100 – 500, 500 - 1 000, 1 000 - 5 000,
5. Название региона, всегда содержится в первой строке – в первом столбце таблицы

Ваша задача - написать скрипт обработки данного файла с использованием языка Python (3.x). Результатом обработки должен быть csv файл, состоящий из следующих столбцов:

- File_Name
- Region
- Partner
- Range,Value

Имена столбцов должны содержаться в первой строке файла. Пример см. на листе «Result».

Требования к результату:
В качестве результата ожидается файл <Фамилия Имя Отчество>_task03.py, в котором будет содержаться код с краткими комментариями

ОПИСАНИЕ КЕЙСА

Задание 4 - SQL Запросы. Генерация данных.

Для решения заданий Вам потребуется скачать и установить любую бесплатную РСУБД (SQL Server Developer, PostgreSQL, MySQL, Oracle XE и т.п.)
В базе-источнике, есть три таблицы:

1. Таблица с данными о магазинах сети - «**stores**»:
- **store_id** – id магазина (целочисленный тип)
- **store_name** – Название магазина
- **store_region** – регион магазина
- **store_id** - primary key таблицы stores

2. Таблица с данными о продажах во всех магазинах сети - «**sales**»:
- **check_num** - Номер Чека (целочисленный тип)
- **sales_date** – год, месяц, день продажи, без времени.
- **store_id** – id магазина, foreign key stores.store_id
- **good_name** – Название товара
- **s_count** – количество товара в позиции (целочисленный тип)
- **s_sum** – сумма продажи (целочисленный тип, для упрощения)
- primary key таблицы sales – это набор полей [check_num, sales_date, store]

3. Таблица с данными о входе-выходе сотрудников из магазинах сети - «**store_acs**»:
- **store_id** - id магазина, foreign key stores.store_id
- **employee_id** сотрудник (целочисленный тип)
- **event_ts** – дата-время входа/выхода сотрудника из магазина.
- **event_type** – тип события (1 – вход сотрудника, -1 – выход сотрудника)

ОПИСАНИЕ КЕЙСА

Задание 4 - SQL Запросы. Генерация данных.

С помощью SQL-запросов, используя циклы и рекурсивные CTE, сгенерируйте данные в таблицах, удовлетворяющие следующим условиям:

1. «**stores**»: 1-3 магазина в 5 разных регионах. Форматы: Регион01,Регион02…Регион05, Магазин01, Магазин02…
2. «**sales**»: 20 уникальных товаров в формате: товар01, товар02… товар20.
3. «**sales**»: продажи за 3 месяца, продажи каждого из 20-ти товаров не менее 5 штук в день в каждом магазине.
4. Таблицу «**store_acs**» достаточно заполнить инструкцией INSERT за одну дату для двух магазинов. Ограничения: рабочий день с 9 до 21 – для всех магазинов, приход – уход не регламентирован, каждый сотрудник минимум один раз выходит из магазина.

Требования к результату:

В качестве результата ожидается файл <Фамилия Имя Отчество>_task04.sql, в котором будет содержаться код создания таблиц, генерации данных и краткие комментарии.

ОПИСАНИЕ КЕЙСА

Задание 5 - SQL Запросы. Продажи товар01 и товар02

Для решения задания Вам потребуются 4 таблицы, созданные в задании 4.

Напишите один запрос, который вернёт полную сумму чеков в разрезе по месяцам (строки) для следующих 4 –х условий (столбцы):

1. Одновременно продан товар01 и товар02.
2. Продан только товар01
3. Продан только товар02
4. Ни товар01, ни товар01 не присутствуют в чеке.

Для Вашего удобства, рекомендуется добавить тестовые записи в таблицу **sales**.

Требования к результату:
В качестве результата ожидается файл <Фамилия Имя Отчество>_task05.sql, в котором будет содержаться код и краткие комментарии.
йл <Фамилия Имя Отчество>_task04.sql, в котором будет содержаться код создания таблиц, генерации данных и краткие комментарии.

ОПИСАНИЕ КЕЙСА

Задание 6 - SQL Запросы. Учет рабочего времени.

Для решения задания Вам потребуется созданная в задании 4 таблица **store_acs**. Внимательно ознакомьтесь с требованиями к заполнению таблицы данными.

Напишите запрос, который предоставит информацию по кол-ву сотрудников, находящихся в магазине на каждый час рабочего дня, в каждом магазине за выбранную дату. В качестве упрощения, используйте только те магазины, для которых вы сгенерировали данные.

Требования к результату:
В качестве результата ожидается файл <Фамилия Имя Отчество>_task06.sql, в котором будет содержаться код и краткие комментарии.

ОПИСАНИЕ КЕЙСА

Задание 7 - SQL Запросы. Аналитика продаж.

Для решения задания Вам потребуются созданные в задании 4 таблицы stores и sales.

Напишите один запрос, который вернёт ТОП 3 товаров по сумме продаж, для каждого региона в разрезе месяцев, а также % от суммы продаж данного товара в данном регионе за месяц и % от суммы продаж данного товара во всей сети за месяц.

Требования к результату:
В качестве результата ожидается файл <Фамилия Имя Отчество>_task07.sql, в котором будет содержаться код и краткие комментарии.

.

ТРЕБОВАНИЯ К ОТПРАВКЕ РЕШЕНИЯ

1. Поместить файлы в указанном в задании формате в один zip-архив (Фамилия Имя Отчество.zip).
2. Выложить zip-архив на google или yandex – Диск.
3. Через личный кабинет прислать ссылку на загрузку zip-архива.