Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСиС»

ИНСТИТУТ ИНФОРМАЦИОННЫХ БИЗНЕС СИСТЕМ

ОТЧЕТ О ПРОЕКТНОЙ РАБОТЕ

ПО ПРЕДМЕТУ

«ПРОГРАММНЫЕ ПЛАТФОРМЫ И ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ»

**ИССЛЕДОВАНИЕ ДАННЫХ И МЕТОДЫ ВЫЯВЛЕНИЯ ЗАКОНОМЕРНОСТЕЙ В ЗАДАЧЕ «РУССКИЕ ШПИОНЫ»**

Программа подготовки магистра по направлению:

09.04.02 – Информационные системы и технологии

Студенты:

Войсят Ю.С.

Егорова В.А.

Злобин П.К.

Крючкова К.С.

Москва 2019

Оглавление

[Постановка задачи 2](#_Toc26750806)

[Алгоритм решения задачи 2](#_Toc26750807)

[Изучение исходных данных, аналитический обзор предметной области 2](#_Toc26750808)

[Анализ данных и обсуждение поиска выявлений аномалий 3](#_Toc26750809)

[Создание ER-диаграммы 5](#_Toc26750810)

[Приведение данных к единому формату 7](#_Toc26750811)

[Создание структуры БД 7](#_Toc26750812)

[Импорт данных в БД 7](#_Toc26750813)

[Приложение 7](#_Toc26750814)

# Постановка задачи

Необходимо разработать алгоритмическое и программное обеспечение для анализа специфических данных с целью выявления закономерностей и выбросов в этих данных для дальнейшего использования в качестве инструментов поиска.

Входными данными данной задачи являются полученные документы из разных источников, таких как сайты бронирования, форумы часто летающих пассажиров, агентства по продаже билетов и т. д. о рейсах нескольких подозрительных граждан России.

Программная реализация алгоритма должна включать в себя автоматизированные инструменты поиска аномалий и выбросов в некотором структурированном наборе данных, полученном из базы данных, структура и данные которой получены непосредственно по имеющимся входным данным.

# Алгоритм решения задачи

Для решения поставленной задачи данные были представлены в структурированном виде и нормализованы. Далее путем мозгового штурма были отобраны идеи выявления аномальных и подозрительных данных. Все идеи были разделены по характеру данных, а затем реализованы путем создания запросов и обработкой данных специальными инструментами для работы с данными.

1. Изучение исходных данных, аналитический обзор предметной области
2. Анализ данных и обсуждение поиска выявлений аномалий
3. Создание ER-диаграммы
4. Приведение данных к единому формату – csv, а также настройка полей по единому шаблону
5. Создание структуры БД
6. Импорт данных в БД
7. Создание плана реализации идей, разработанных экспертной группой
8. Реализация идей: создание скриптов
9. Представление результатов в наглядном виде (диаграммы, схемы)

# Изучение исходных данных, аналитический обзор предметной области

Предметная область проектной работы – авиаперевозки. Исходные данные представляют собой купленные билеты, информацию с форума о программах лояльности, личные данные пассажиров за 2017 год, а также расписания рейсов самолетов за 2017 и ноябрь 2018 – январь 2019 года.

Особенностями предметной области являются строго стандартизированные поля: аэропорт, номер рейса, код еды, и т.д.

Каждый аэропорт имеет уникальных по всему миру трехбуквенный код (кодировка ИАТА), например, DME – аэропорт Домодедово, EGO — аэропорт Белгород.

Номер рейса также кодируется с помощью ИАТА. Буквенная составляющая номера представляет перевозчика. Например, Delta обозначается как DL, American Airlines – как AA, United Airlines – как UA, а «Аэрофлот» – SU (Soviet Union). Номер рейса должен помещаться в диапазоне чисел от 1 до 9999. Как правило, номера рейсов на север и на восток — чётные, а на юг и запад — нечётные, но бывают и исключения. Авиакомпании обычно делают номер обратного рейса на одну цифру больше, чем номер исходящего рейса. Также, обычно, чем меньше номер рейса, тем важнее маршрут для авиакомпании. Рейс из аэропорта Джона Кеннеди в лондонский Хитроу — главный для Delta и обозначается как DL1. Чтобы не создавать путаницу с моделью самолёта Boeing-757, вылетам не присваивают номер 757. Кроме того, исключительно из суеверия, компании ни за что не будут выполнять рейсы 13 или 666. Только Finnair оказались выше своих страхов и несколько раз отправляли рейс 666 в пятницу, 13-го (последний раз в октябре 2017 года). Вылет из Копенгагена в Хельсинки (аэропорт обозначается как HEL, что созвучно с английским hell – «ад») прошёл более чем успешно.

# Анализ данных и обсуждение поиска выявлений аномалий

Анализ данных и выработка эффективных, а главное аргументированных методов обнаружения аномалий и нехарактерных отклонений является ключевой задачей проекта. Исследование имеющейся информации показало, что аномалии могут отражать несколько основных типов выбросов:

1. Несоответствие действий человека нормам, принятым в его социальной, этнической или религиозной группе.
2. Нестандартное или неестественное (нехарактерное) поведение человека
3. Нестандартная окружающая обстановка

Первый тип аномалий включает в себя выявление человеческих поступков, отклоняющихся от норм, установленных в социальной, этнической или религиозной группе, к которой эти лица предположительно относятся. Для этого были выявлены особенности, по которым можно определить людей к какому-либо классу, самая статистически вероятная из которых - частое посещение одного и того же города конкретной страны. Исходя из этого факта делается вывод, что пассажир имеет дом в этой стране и с высокой долей вероятности принадлежит к национальной и этнической группе, широко распространенной в этом государстве. На следующем этапе были определены отличительные черты поведения членов данных групп, однозначно отличающих их от остальных.

* Люди, относящиеся к мусульманской религии , полностью исключают из своего рациона питания блюда из свинины
* Люди, исповедующие индуизм не употребляют пищу, включающую в себя говядину
* Иудаизм характеризуется запретом использования различных технических и транспортных средств, в том числе самолетом в шаббат (субботу)
* Общество вегетарианцев (в том числе веганов) не принимают пищу, включающую в свой состав ингредиенты мясного происхождения

Несоблюдение этих догм может говорить о возможном притворстве и блефовании человека, что с высокой степенью уверенности выдает его как шпиона.

Второй тип аномалий отражает странное поведение человека, резко отличающееся от привычных действий окружающих. Принцип поиска таких нестандартных отклонений основывается на аналитической работе, подразумевающей под собой поиск и исключение привычных и обыденных поступков, тем самым выявляя остальные случаи. В ходе изучения информации и фильтра неподходящих данных были определены особенности неоднозначно трактуемых действий.

* Покупка билетов на один день (одно время), при заранее известной невозможности пребывания в забронированных самолетах может говорить о возможной попытке человека скрыть истинное направление своего движения. Данный фактор часто используется на практике работниками секретных подразделений.
* Полет одних и тех же людей на борту самолета несколько раз (при условии, что они не являются родственниками и не имеют общий номер брони) не выглядит как простое стечение обстоятельств и заставляет рассмотреть этих пассажиров более детально
* Чрезвычайно маленькое время нахождения человека в разных городах между перелетами дает сигнал об отсутствии туристического интереса и нахождении в этой точке родственных связей с кем-либо, что заставляет обратить особое внимание на деятельность данного пассажира
* Путешествия или командировки на длительное время, как правило сопровождаются, перевозкой багажа, а его систематическое отсутствие в течение длительного времени является нехарактерным фактором практически для любого человека
* Несомненно редким и требующим рассмотрения является факт систематического прилета человека в один город, а улет совершенно из другого. Данный случаи могут говорить о возможном опасении пассажира и его попыткой запутать возможное преследование
* Авиакомпании при частых полетах предоставляют существенные бонусы таким пассажирам, поэтому отказ вступления в программу лояльности такого лица может говорить о незаинтересованности в экономии средств, из чего можно сделать справедливый вывод, что деньги, на которые приобретен билет являются ни личными, ни предоставленными компанией, в случае, если это работник, отправившийся в командировку. Разведывательные службы используют государственные деньги и не заинтересованы во внесении имени своего сотрудника в дополнительные базы данных, которые могут его скомпрометировать.

Выявленные возможные случаи неординарного поведения маловероятно могут оказаться простым стечением обстоятельств, что говорит о необходимости детального рассмотрения данных лиц.

Третий тип выбросов заключает в себе нетривиальные совокупности факторов, вызывающих определенные сомнения в их случайности. Комплекс маловероятных и незапланированных обстоятельств может быть в интересах человека, который находится в эпицентре этих действий, таким образом, он может сам провоцировать создание группы этих событий.

* Вероятность регистрации только одного человека на рейс вне зависимости от вместительности самолета, практически однозначно убеждает в отсутствии случайной компоненты. С высокой долей уверенности можно утверждать, что пассажир преднамеренно подстроил данную ситуацию в личных целях, что однозначно говорит о нем, как о нетривиальном пассажире.
* О возникновении аналогичной ситуации может говорить лишь одна заказанная порция еды на борту или один багаж, сданный в грузовое отделение даже при покупке всех билетов.

В случае выявления аномалий в ходе анализа данных, базируясь на определении перечисленных событий можно практически со 100% уверенностью утверждать об их преднамеренном планировании и разработке, что может быть на руку только шпиону.

# Создание ER-диаграммы

При анализе данных были выявлены следующие стержневые сущности: Пассажир, Полет (рис. 1)

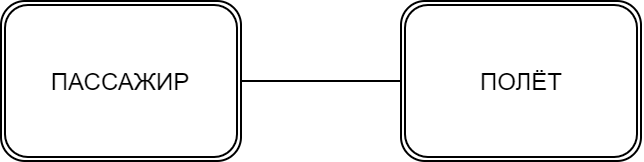


Рисунок - Стержневые сущности

Далее после нормализации были получены следующие таблицы:

1. Пассажир
2. Полет
3. Билет (4 НФ) – после уничтожения связи M:M
4. Рейс (4 НФ) – после уничтожения связи M:M
5. Самолет (3 НФ) – таблица-справочник
6. Аэропорт (3 НФ) – таблица-справочник
7. Тариф (3 НФ)
8. Агрегатор (3 НФ) – таблица-справочник
9. Еда (3НФ) – таблица-справочник
10. Бонусная программа (3 НФ)

Итоговая ER-диаграмма представлена на рис.2

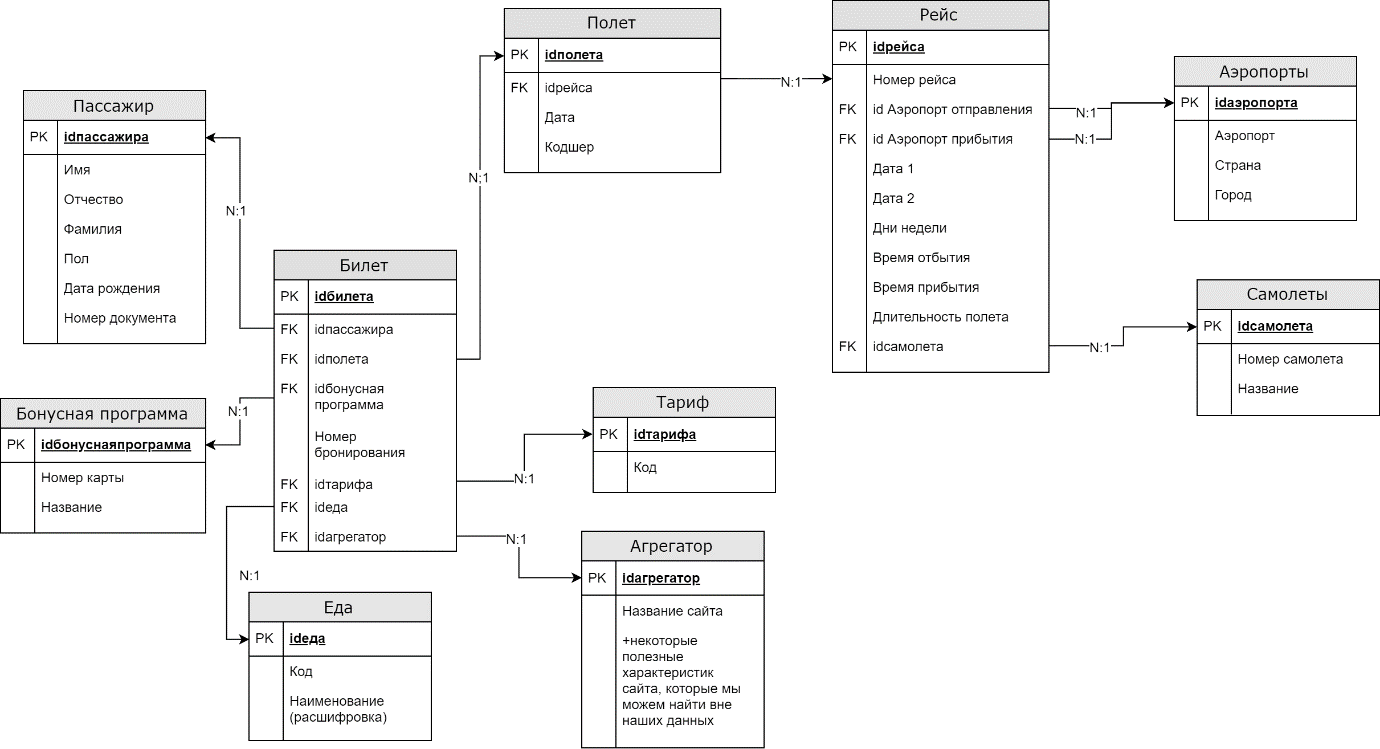


Рисунок - ER-диаграмма

# Приведение данных к единому формату

Все исходные данные изначально представлены в виде файлов разного формата: json, xml, pdf, csv, tab, yaml, xlsx. Для удобства импорта данных в базу данных было решено выбрать единый формат, предполагающий наличие разделителей – csv.

Таким образом, каждый файл был преобразован в csv формат. Для некоторых были написаны скрипты на языке Python, где при помощи регулярных выражений по шаблону вычленялись необходимые поля. Некоторые файлы, например, tab, был преобразован непосредственно в Excel – при помощи функций и с использованием масок.

В каждом файле дата была преобразована в формат “##.##.####”. Имя, фамилия и отчество разбиты на отдельные колонки (1 НФ).

Все использованные скрипты, а также файлы в формате csv расположены в репозитории (см. Приложение).

# Создание структуры БД

По представленной ER-диаграмме была создана база данных. Скрипт для базы написан на языке SQL. Развертывание базы выполнено в среде PostgreSQL 12 PgAdmin 4.

Cкрипт по созданию БД расположен в репозитории (см. Приложение).

# Импорт данных в БД

Импорт данных в БД был выполнен путем развертывания csv-файлов в среде PostgreSQL 12 PgAdmin 4. Дальнейшие манипуляции с данными по разбиению данных на таблицы выполнены при помощи SQL-запросов.

Backup-файл базы расположен в репозитории (см. Приложение).

# Приложение

<https://github.com/barbara11e/Bashirov>

Здесь актуальный отчет, презентация и скрипты.