«Национальный исследовательский ядерный университет «МИФИ»

Институт интеллектуальных кибернетических систем

Отчёт по командной работе

“DataScience”

**“Catch “The Russian Spies” Challenge”**

Студенты: Нурматов Салим

Шипилов Артем

Калинкин Артем

Чыонг Тхи Ан Хай

Группа: Б17-565

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Москва 2020

1. **Постановка задачи**

В данной работе команда поставила задачу анализа данных о рейсах с целью поиска потенциальных «русских шпионов» среди часто летающих пассажиров.

1. **Парсинг файлов и преобразование данных к формату csv с целю их последующего использования.**

Шипилов Артем – JSON, XML. Использовалось несколько файлов, парсеры залиты на GitHub. На языке Java для JSON использовалась библиотека gson; для XML – xstream. Была создана иерархия классов, соответствующая файлам.

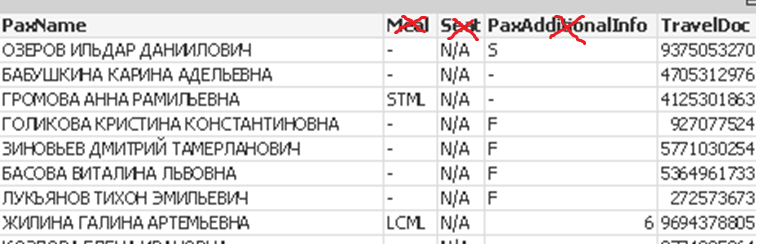
Нурматов Салим – PDF. Данный pdf файл был распарсен на языке Python.

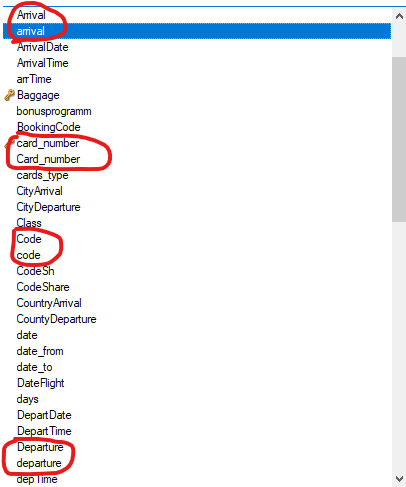
Калинкин Артем – YAML. На языке С++ был написан скрипт, с использованием библиотек <iostream>, <string>, <fstream>, <regex>, <vector>.

Чыонг Тхи Ан Хай – TAB. На языке С++ был написан скрипт, для преоразования файла .tab.

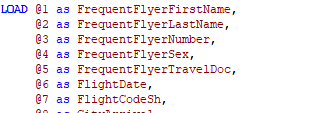
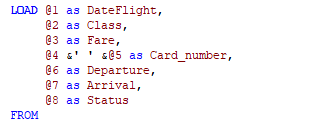
1. **Обработка данных (Нурматов, Калинкин). (Слияние, очистка, фильтрация)**

После преобразования файлов в формат csv, была произведена загрузка файлов в приложение “Qlickview”. При загрузке были решены все конфликты имен (было произведено слияние столбцов, отличающихся по названия, но имеющие одни и те же значения), были убраны не нужные данные.

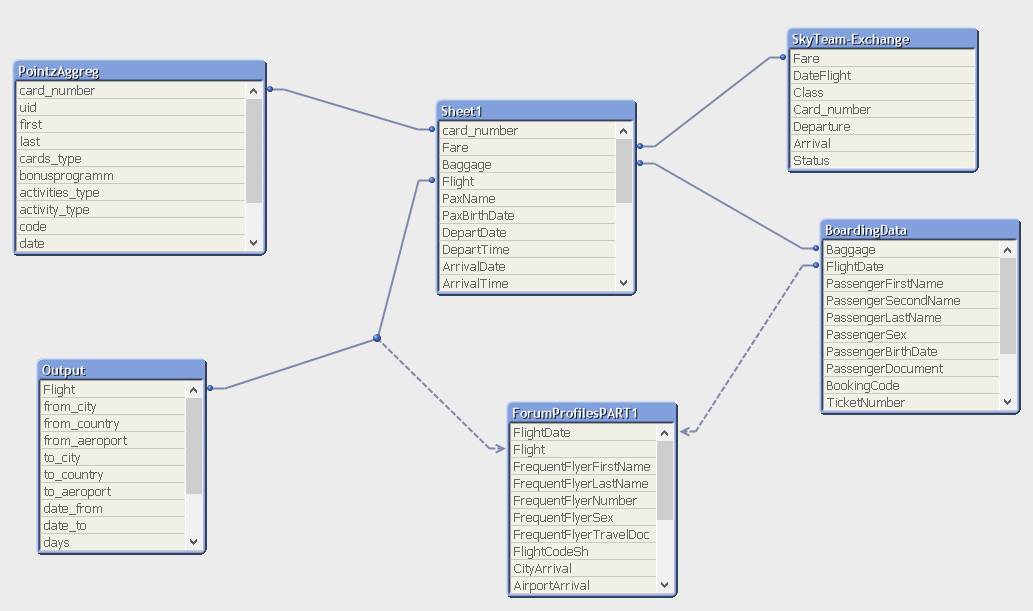




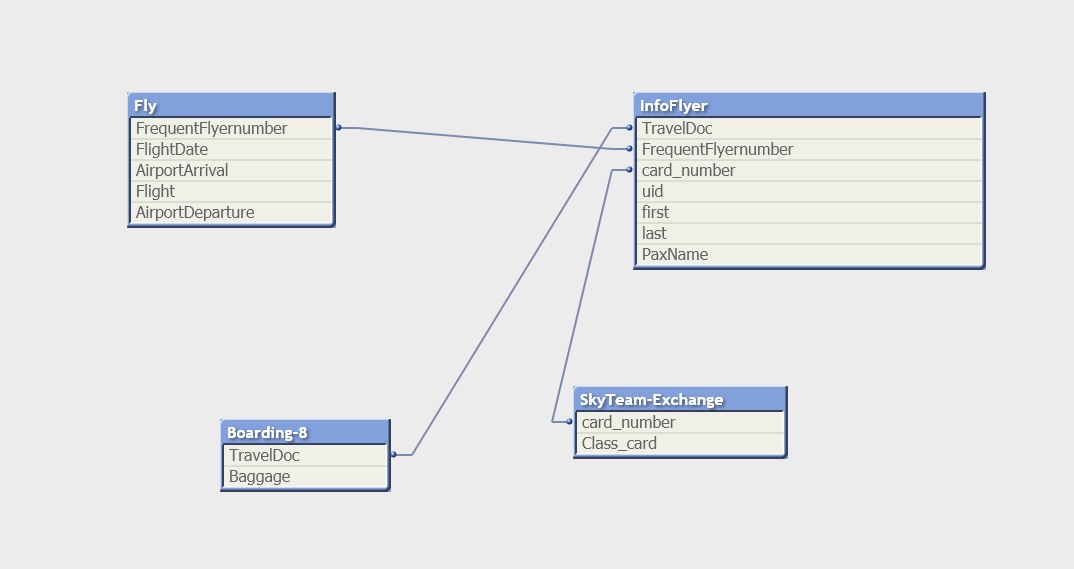
Кроме того, некоторые файлы не имели названия столбцов, для решения это проблемы столбцы были названы в соответствии с названиями остальных таблиц.



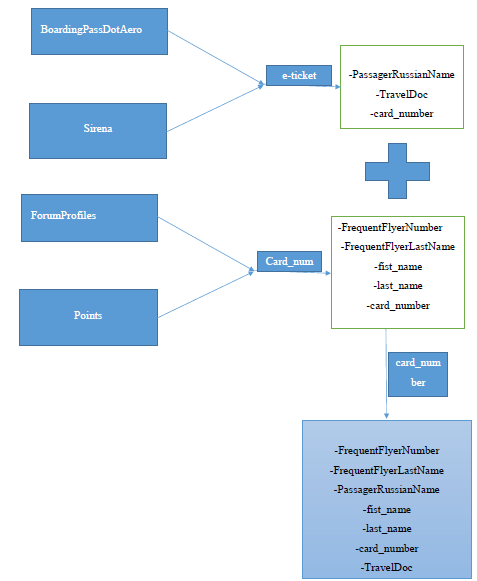
После переименования некоторых значений столбцов, была получена диаграмма сущностей, чтобы понять, что из себя представляет «База данных»

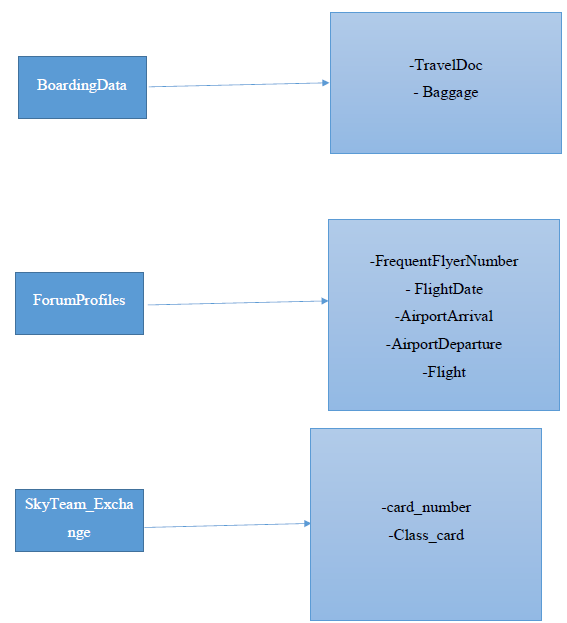


После очистки и слияния данных была получена следующая диаграмма сущностей



Процесс слияния данных проходил в соответствии схеме:



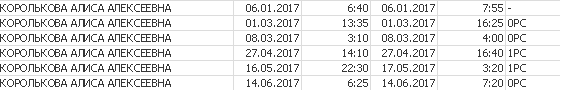


1. **Гипотезы и Алгоритмы поиска подозреваемых по определенным критериям (Нурматов, Калинкин, Шипилов, Ан Хай)**

В процессе анализа данных был составлен перечень критериев в поиске возможных «Русских шпионов».

Каждый из участников группы предлагал свои критерии по поиску подозреваемых, после чего совместно были проанализированы каждые из предложенных критериев и выбраны наиболее подходящие для определения шпиона.

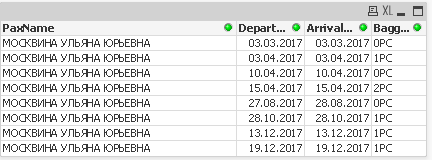
1. Подозреваемый вылетел на долгий срок без багажа. Для этого был осуществлен поиск, где полученные данные были отсортированы в первую очередь по дате вылета и дате прилета соответственно. Так же это было использованы для составления логической цепочки полетов из одного города в другой. Наши подозрения были выдвинуты в соответствии с тем фактом, что подозреваемый летал в тот или иной город на долгий срок без багажа, где 0PC означает без багажа, 1PC – 1 багаж, 2PC – 2 багажа.



1. Перелеты с целью сбить с толку или запутать следы. В ходе поиска также проходила сортировка по дате перелетов, создать логическую цепочку. Причиной наших подозрений является перелеты в пределах одного города.



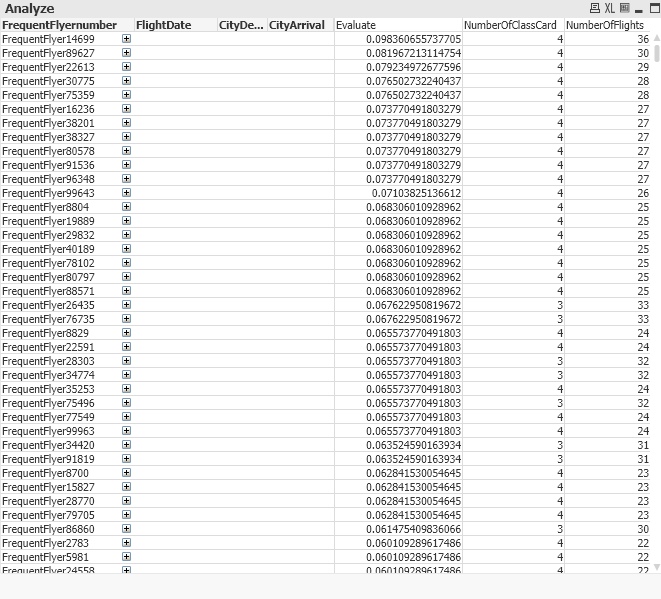
1. Подозреваемый улетел без багажа, прилетел с багажом. Поиск в данном критерии происходит также по сортировке даты отлета и прилета.



1. Подозреваемый вылетел из РФ и не вернулся. Стоит учесть тот факт, что обычно шпионы летают очень много и на долгие сроки – около года. В этом методе так же используется сортировка по датам перелетов а так же по ФИО. Этот метод является уникальным, так как он позволяет анализировать историю полетов и просматривать состояние багажа при каждом перелете, а также следить не пытается ли он запутать следы.
2. Кроме вышеописанных критериев были рассмотрены и другие, которые в теории оправдывают свои подозрения, но они были исключены так как мы не нашли подтверждения им. Примером тому является подсчет количество пересадок, совершенных подряд с учетов того, какие города были транзитными. В мире есть всего несколько крупных городов, которые используются в качестве транзита: Пекин, Стамбул, Москва и один из городов Европы. Следовательно, 2 -3 пересадки минимум сразу вызывают подозрения.

Для идентификации каждого пассажира были использованы паспортные данные.

1. **Общий список подозреваемых (Ан Хай)**



Общий список подозреваемых. FrequentFlyers отсортированы в порядке «уровня подозрительности». чем выше Evaluate, тем он «подозрительнее».

1. **Ссылка на GitHub**

После завершения работы, полный перечень документов был загружен на GitHub.

https://github.com/artemGBF/Russian\_Spies