**Кейс «Обнаружение аномалий в данных о потреблении интернет-трафика»**

**Задание**

Обнаружение нетипичного потребления интернет-трафика абонентами компании — оператора связи с целью определения взлома оборудования абонента.

Данное задание предполагает проведение анализа представленных данных о потреблении интернет-трафика абонентами. Нетипичное изменение расхода трафика могут сигнализировать о взломе оборудования абонента и установке на него вредоносного ПО — например для спам рассылок, ddos атак и т.д. Выявление таких фактов помогает временно заблокировать учетную запись абонента, до исправления ситуации. Для решения этой задачи необходимо использовать методы очистки данных, загрузки данных, расчет витрин данных и т. д. Помимо анализа данных, важно применять практики управления данными (data governance) для контроля качества данных на всех этапах анализа (data quality), определения происхождения данных (data lineage), описания глоссария данных.

Полученные результаты помогут в улучшения качества обслуживания клиентов оператора связи.

**Исходные данные**

Вам представлен набор синтетических данных: информации о клиентах, файлы периодической выгрузки с линейного оборудования связи с информацией о текущем трафике.

**Ожидаемый результат**

Необходимо на основе имеющихся данных построить набор выходных таблиц с интервалом расчета в 1 час входных данных и содержащих информацию о расходе трафика потребителями и признака подозрения на взлом. Схема входных данных и витрины, а также основания для детекции указанной аномалии представлены в детальном описании.

**Формат решения**

Результаты которые демонстрируются участниками хакатоны должны содержать:

1. список имен взломанных абонентов,
2. набор выходных таблиц, содержащих расчетные данные, построенные по данным выгрузки с телеком-оборудованием. Именно на основании этих расчетных данных и делается вывод о взломе отдельных абонентов. Поскольку период построения витрины соответствует 1 часу входящих оперативных данных, то количество витрин должно быть равно периоду (в часах) за который выгружались оперативные данные,
3. набор отчетов о контроле качества данных (data quality) на всех этапах обработки данных,
4. диаграмму происхождения данных (data lineage) для предоставляемых витрин,
5. глоссарий предметной области,
6. документ — презентация с демонстрацией хода решения.