Описание задачи.

Представленная задача создана на основе реального производственного процесса, решаемой аналитиками данных в одной из зарубежной компании — оператора связи. Компания, предоставляющая услуги связи, имеет различные тарифы на предоставляемые каналы связи интернет. Однако, несмотря на то, что каналы ограничены рамками тарифов, компания отслеживает динамику потребления интернет-трафика потребителями, и в случае фиксации нетипичного всплеска потребления, предполагает, что возможно компьютер абонента был взломан, и при помощи вредоносного ПО стал выполнять роль сервера спам-рассылки, элемента DDOS сети, сайта с запрещенным контентом и т. д. В этом случае, предоставление услуг интернет-связи временно блокируется, а с абонентом связываются для выяснения обстоятельств.

С технической точки зрения система контроля должна работать следующим образом. Данные из различных первичных систем оператора связи поступают в виде файлов различного формата (в зависимости от системы источника). Представленные данные можно подразделить на:

- справочные (абоненты, профили, тарифы и т. д.), представленные в виде комплекта файлов с текущим состоянием справочников,
- оперативные, выгружаемые из систем биллинга каждые 10 минут, которые содержат информацию о подключениях абонентов, объеме трафика и т. д. и представлены в виде текстовых файлов.

Оперативные данные представлены за период времени в несколько дней.

Необходимо пред обработать оперативные данные, и в совокупности с имеющимися справочными данными следует построить обновляемую таблицу (витрину данных) в соответствии с заданной схемой для прогнозирования взлома клиентов оператора связи для оперативного отключения клиентов, до выяснения обстоятельств. Каждый новый экземпляр витрины данных должен строится по исходным оперативным за каждый 1 час.

Кроме построения обновляемой витрины данных, необходимо, используя подходы Data Governance, настроить контроль качества данных (data quality) и происхождения данных (data lineage).