

Нужно создать витрину, содержащую в себе идентификаторы пользователей определенного веб-сайта.
 За основу берется Витрина А этого сайта, партиционированная по датам.
 Для решения задачи нужно использовать Spark DataFrame API (на Scala или PySpark).

- В результате нужно получить витрину, **максимально** содержащую все возможные идентификаторы пользователей (ИНН, хэш, телефон, id внутри сайта, различные куки, GAID или IDFA)
- Таблицы из источника предлагается заполнить самостоятельно. Сгенерировать синтетические данные тоже часть задания=)
- Алгоритм необходимо реализовать таким образом, чтобы его можно было запускать ежедневно для добавления новых данных и удаления дубликатов
- Т.е. в витрине идентификаторов нам неважно, что клиент приходит три дня подряд, нам важно лишь зафиксировать клиента однажды (и его новые устройства (куки) если он заходит с них)

Ваши источники

Витрина А

Название	Описание	Комментарий
raw_cookie	[{"key": "_sa_cookie_a", "value": "SA1.e519e40f-f599-4651-b168-40eefe821bf4.1592250469"}, {"key": "_fa_cookie_b", "value": "ads2.2020127030.1589493240"}, {"key": "_ym_cookie_c", "value": "15894932061040605701"}, {"key": "_fbp", "value": "fb.1.1589493214770.259895265"}, {"key": "org_uid", "value": "1234565"}, {"key": "user_uid", "value": "1234565"}, {"key": "user_phone", "value": "9 777 312 11 31"}, {"key": "user_mail", "value": " usermail@user.io "}]	колонку raw_cookie Вы должны заполнить такой сложной структурой, как представлен а выше
event_type	SUBMIT, REGISTER, SUBMIT_MD5	если SUBMIT_MD5, то в колонке data_value будет лежать хэш инн мд5 если SUBMIT, то в ней будет хэш ша256
event_action	pageview, event, login-check-otp	
data_value	e519e40ff5994651b16840eefe821bf4	хэши ИНН в ша256 и мд5, иногда ничего
geocountry	Mongolia	
city	Ulan-Batar	

user_os	Mac	
systemlanguage	RU, ENG	
geoaltitude	123,432	
meta_platform	WEB, MOBAIL	
screen size	1920x1080	
timestampcolumn	2023-01-01	

Витрина В

Название	Описание	Комментарий
id	Идентификатор системы В	
inn	ИНН	

Витрина С

Название	Описание	Комментарий
id	Идентификатор системы С	
cookie_a	Куки сайта	

Витрина D

Название	Описание	Комментарий
id	Идентификатор системы D	
inn	ИНН	

Витрина Е

Название	Описание	Комментарий
id	Идентификатор системы Е	
hash_phone_md5	Номер телефона хэш MD5	

Витрина F

Название	Описание	Комментарий
id	Идентификатор системы F	
hash_email_md5	Email хэш MD5	

Витрина G

Название	Описание	Комментарий
match_code	Тип идентификатора	match_code может быть GAID или IDFA, в зависимости от этого user_uid будет, соответственно, являться либо GAID (идентификатором андроид устройства), либо IDFA (идентификатор IOS устройства)
user_uid	Идентификатор пользователя	
id	Идентификатор системы F	