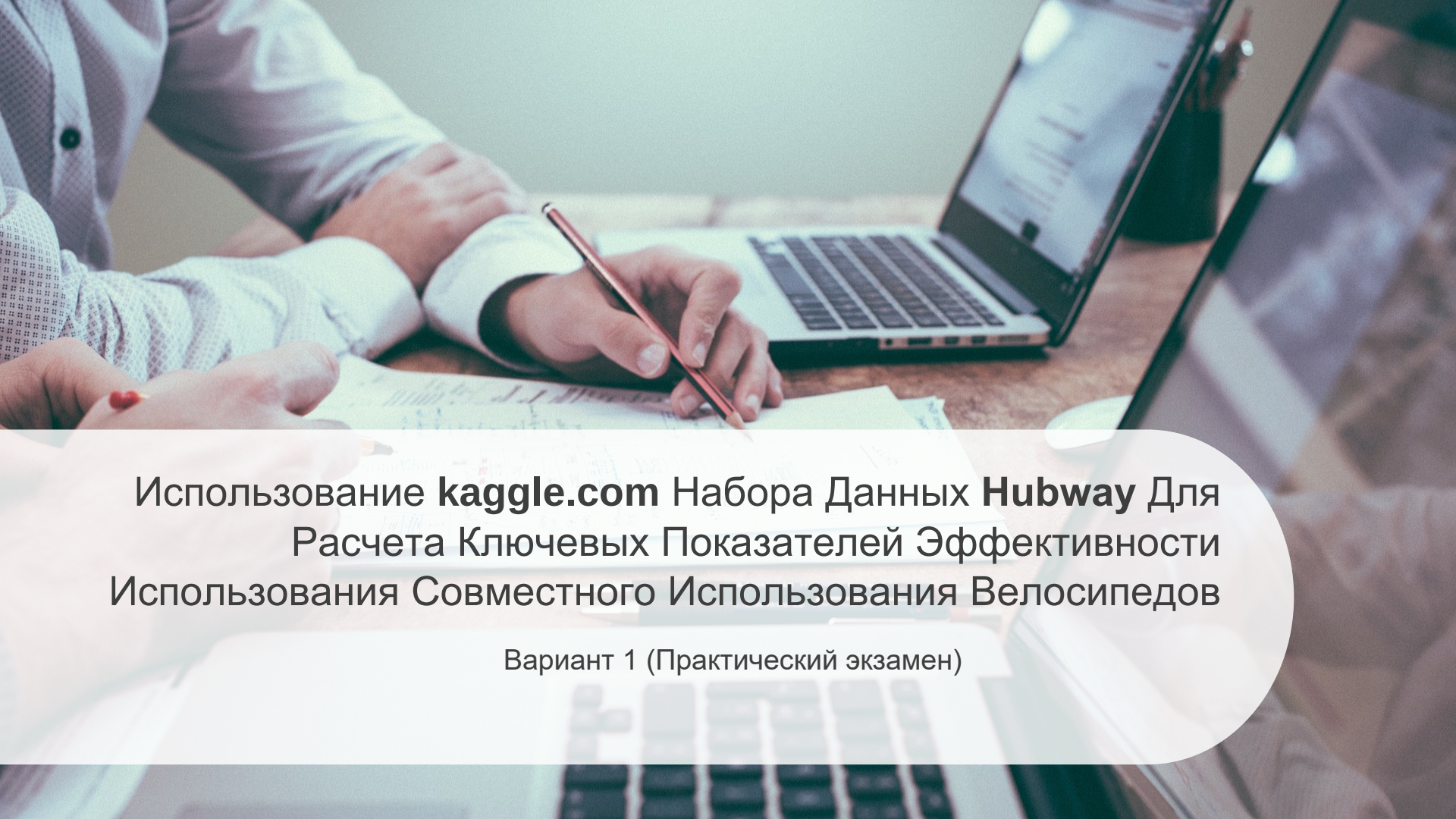

Использование **kaggle.com** Набора Данных **Hubway** Для Расчета Ключевых Показателей Эффективности Использования Совместного Использования Велосипедов

Вариант 1 (Практический экзамен)

Цель

kaggle.com обеспечивает ежемесячный экспорт записей о поездках на велосипедах Hubway 2011-2017:

-https://www.bluebikes.com/

Последние Полные Дампы: **https://www.kaggle.com/acmeyer/hubway-data**

tripduratio	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype	birth year	gende
125	01.10.2017 0:03	01.10.2017 0:05	107	Ames St at Main St	42.3625	-71.08822	117	Binney St / Sixth St	42.366095	-71.086388	720	Subscriber	1992	
1195	01.10.2017 0:04	01.10.2017 0:24	46	Christian Science Plaza - Massachusetts Ave at Westland Ave	42.3436658245146	-71.08582377433777	66	Allston Green District - Griggs St at Commonwealth Ave	42.34922469338298	-71.13275302578586	1320	Subscriber	1989	
513	01.10.2017 0:04	01.10.2017 0:13	227	Silber Way	42.34949599514002	-71.10057592391968	67	MIT at Mass Ave / Amherst St	42.3581	-71.093198	166	Subscriber	1996	
1609	01.10.2017 0:06	01.10.2017 0:33	60	Charles Circle - Charles St at Cambridge St	42.36075761041193	-71.07132909824031	17	Soldiers Field Park - 111 Western Ave	42.365064	-71.119233	438	Subscriber	1987	
246	01.10.2017 0:09	01.10.2017 0:13	22	South Station - 700 Atlantic Ave	42.352175	-71.055547	81	Boylston St at Washington St	42.352409	-71.062679	577	Subscriber	1982	
1857	01.10.2017 0:10	01.10.2017 0:41	90	Lechmere Station at Cambridge St / First St	42.370677	-71.076529	77	Somerville City Hall	42.386844	-71.09812	802	Subscriber	1995	
320	01.10.2017 0:10	01.10.2017 0:16	108	Harvard University / SEAS Cruft-Pierce Halls at 29 Oxford St	42.377945	-71.116865	74	Harvard Square at Mass Ave/ Dunster	42.373268	-71.118579	939	Subscriber	1998	
579	01.10.2017 0:11	01.10.2017 0:21	8	Union Square - Brighton Ave at Cambridge St	42.353334	-71.137313	174	Washington St at Brock St	42.3489528466951	-71.16031676530838	885	Subscriber	1990	
702	01.10.2017 0:12	01.10.2017 0:23	53	Beacon St at Massachusetts Ave	42.35082680669095	-71.0898108780384	120	Charles St and Beacon St	42.356052	-71.069849	710	Subscriber	1993	
1231	01.10.2017 0:14	01.10.2017 0:35	176	Lesley University	42.38674802045056	-71.11901879310608	74	Harvard Square at Mass Ave/ Dunster	42.373268	-71.118579	980	Subscriber	1982	
838	01.10.2017 0:14	01.10.2017 0:28	39	Washington St at Rutland St	42.338514601785995	-71.07404083013535	20	Aquarium T Stop - 200 Atlantic Ave	42.35991176110118	-71.0514298081398	50	Subscriber	1990	
687	01.10.2017 0:15	01.10.2017 0:26	91	One Kendall Square at Hampshire St / Portland St	42.366277	-71.09169	60	Charles Circle - Charles St at Cambridge St	42.36075761041193	-71.07132909824031	926	Customer	\N	
650	01.10.2017 0:15	01.10.2017 0:26	91	One Kendall Square at Hampshire St / Portland St	42.366277	-71.09169	60	Charles Circle - Charles St at Cambridge St	42.36075761041193	-71.07132909824031	339	Customer	\N	
174	01.10.2017 0:16	01.10.2017 0:19	21	Prudential Center - Belvedere St	42.345959	-71.082578	46	Christian Science Plaza - Massachusetts Ave at Westland Ave	42.3436658245146	-71.08582377433777	282	Subscriber	1946	

Цель

Мы хотим использовать эти данные для расчёта некоторых ключевых пока зателей эффективности использования сервиса

Этапы выполнения задания:

- Скачать данные с ресурса **htt**ps://**www.kaggle.co**m/**ac**m**eyer/hubway**-d**ata**.
- Сохранить необработанные данные (**CS**V-файлы) в **HD**F**S** (с разбивкой по YYYYMM).
- Оптимизировать, уменьшить и очистить исходные данные и сохранить их в конечном каталоге на **HD**F**S**.
- Рассчитайте ключевые показатели эффективности.
- экспортируйте их в файл Ex**cel**.

Дополнительное задание! Весь рабочий процесс обработки данных должен реализовать в инструменте ETL workflow (например, Pentaho Data Integration или Airflow) и выполняться автоматически.

ШАГ 1. Получение данных о совместном использовании велосипедов Hubway

Скачать с kaggle.com

https://www.kaggle.com/acmeyer/hubway-data

CSV CSV CSV

/user/hadoop/hubway/ / /*.csv
/user/hadoop/hubway/ / /*.csv
/user/hadoop/hubway/ / /*.csv
…

ШАГ 2. Предобработка данных

/user/hadoop/hubway/**raw**/**201606**/*.csv
/user/hadoop/hubway/**raw**/**201607**/*.csv
/user/hadoop/hubway/**raw**/**201608**/*.csv
…

- Переместить данные из исходного каталога в конечный каталог.
- При необходимости оптимизировать и уменьшить структуру данных для последующих запросов.
- При необходимости удалить дубликаты.

/user/hadoop/hubway/**final**/**201606**/*
/user/hadoop/hubway/**final**/**201607**/*
/user/hadoop/hubway/**final**/**201608**/*
…

ШАГ 3. Расчет И Экспорт Ключевых показателей эффективности

- Рассчитать ключевые показатели эффективности.
- Экспортируйте ключевые показатели в Excel.
- Используйте для достижения цели инструменты Hive, Spark или PySpark.

ШАГ 4. Ключевые Показатели Эффективности Для Расчета

Рассчитайте в месяц:

- Средняя продолжительность поездки (в минутах)
- Среднее расстояние поездки (в км)
- Доля использования в разбивке по полу (в процентах)
- Доля использования по возрасту (в процентах)
- Топ-10 самых подержанных велосипедов
- Топ-10 самых популярных стартовых станций
- Топ-10 самых конечных станций
- Доля использования на временной интервал (в процентах):
    - 00:00-06:00
    - 06:00-12:00
    - 12:00-18:00
    - 18:00-24:00