Использование данных **NYC_Taxi** для Расчета Ключевых Показателей Эффективности

Вариант 2 (Практический экзамен)

Цель

NYC.gov предоставляет ежемесячный экспорт записей о поездках на желтом такси в Нью-Йорке:

- https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Последние Полные Дампы:

- https://nyc-tlc.s3.amazonaws.com/trip+data/yellow_tripdata_2022-01.parquet
- https://nyc-tlc.s3.amazonaws.com/trip+data/yellow_tripdata_2022-02.parquet
- https://nyc-tlc.s3.amazonaws.com/trip+data/yellow_tripdata_2022-03.parquet

Цель

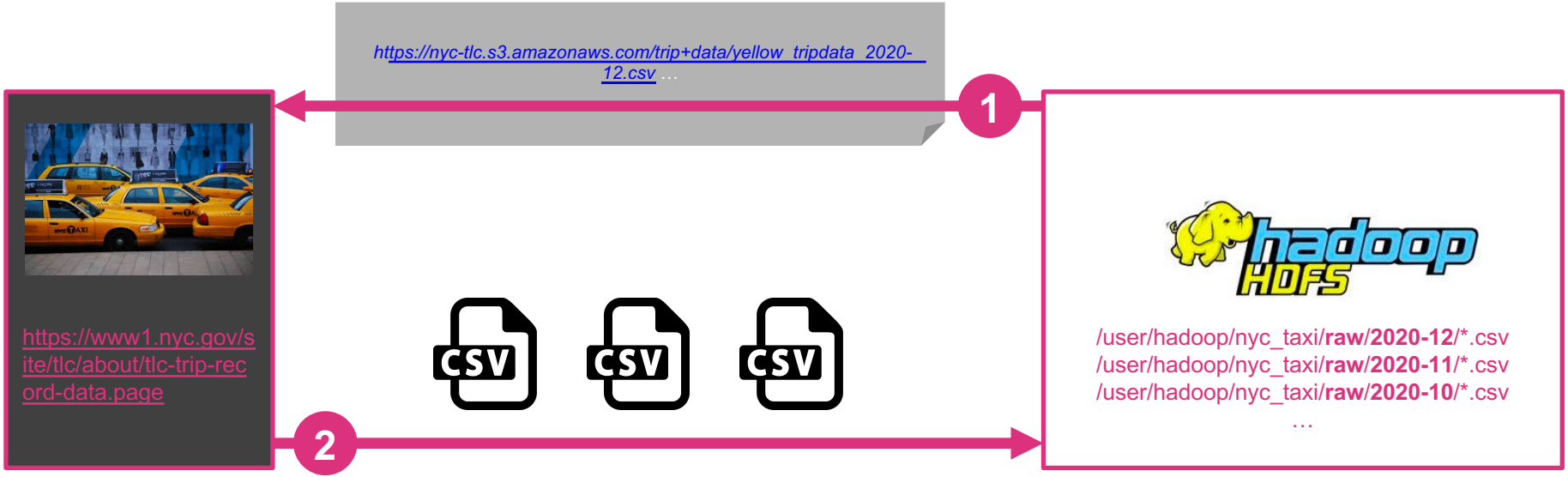
Мы хотим использовать эти данные для расчёта некоторых ключевых показателей эффективности использования сервиса

Этапы выполнения задания:

- Скачать данные с ресурса https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.
- 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. Получить Данные о такси **TLC NYC**



ШАГ 2. Предобработка данных

/user/hadoop/nyc_taxi/**raw**/**2020-12**/*.csv
/user/hadoop/nyc_taxi/**raw**/**2020-11**/*.csv
/user/hadoop/nyc_taxi/**raw**/**2020-10**/*.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/nyc_taxi/**final**/**2020-12**/*
/user/hadoop/nyc_taxi/**final**/**2020-11**/*
/user/hadoop/nyc_taxi/**final**/**2020-10**/*
…

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:
- Среднюю продолжительность поездки (в минутах)
- Среднее расстояние поездки (в милях)
- Среднюю общую сумму поездки (в долларах США)
- Среднюю сумму чаевых (в долларах США)
- Среднее количество пассажиров (в виде числа)
- Долю использования по типу оплаты
(кредитная карта, наличные и т.д. в процентах)