Анализ публикуемых новостей

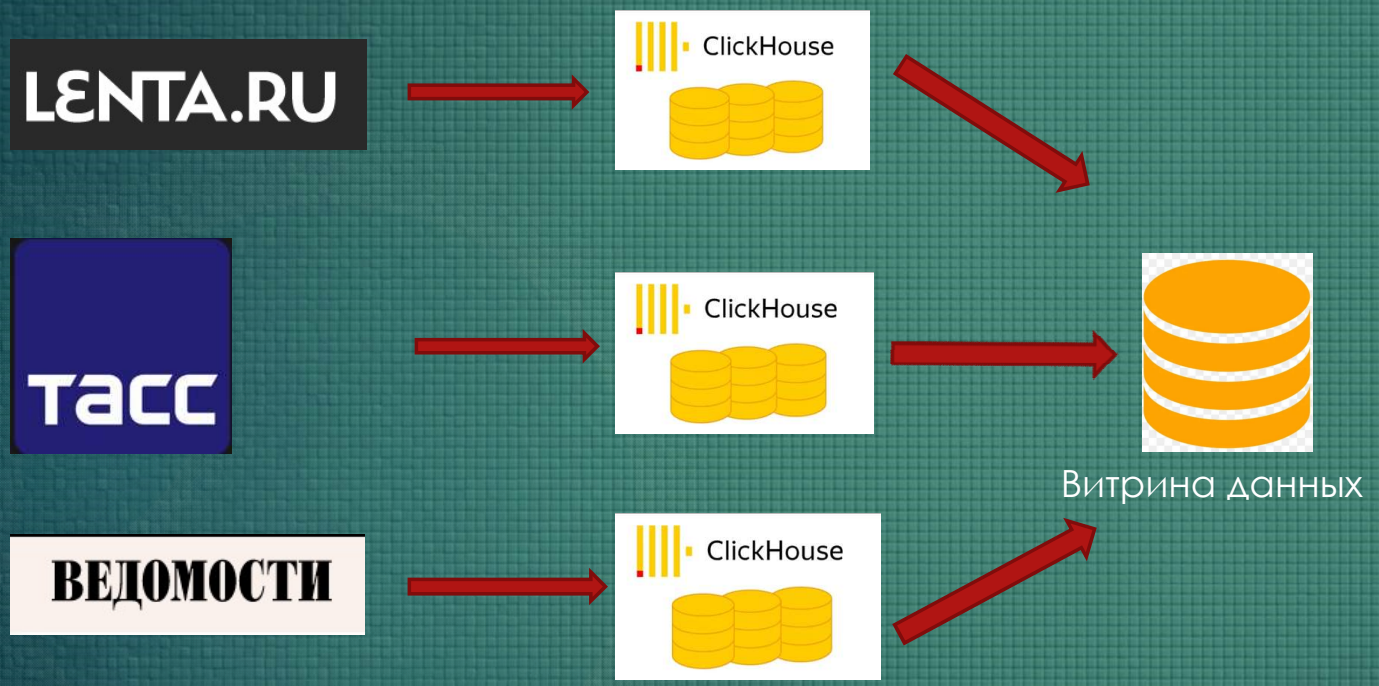
ВЫПОЛНИЛ: ГИНДУЛЛИН Э.Р.

Общая задача: создать ETL-процесс формирования витрин данных для анализа публикаций новостей.

- Подробное описание задачи:
- Разработать скрипты загрузки данных в 2-х режимах:
- o Инициализирующий – загрузка полного слепка данных источника
- o Инкрементальный – загрузка дельты данных за прошедшие сутки
- Организовать правильную структуру хранения данных
- o Сырой слой данных
- o Промежуточный слой
- o Слой витрин

Технологический стек и порядок обработки данных

Порядок обработки данных и отказоустойчивость



Данные собираются и Обрабатываются независимо друг от друга , по общим правилам, но в разные таблицы. Выход из строя одного из источников данных не приведет к выходу из строя всей системы. Ошибки можно исправить с помощью отладочных скриптов, а испорченные данные восстановить из сырых.
Объединение происходит только на конечном этапе сборки витрины данных.

Вывод и заключение:

- Использование Clickhouse в качестве БД значительно усложнило выполнение задания и потребовало больше времени на подготовку и выполнение. С другой стороны никаких нареканий к БД по производительности не было замечено, даже учитывая то, что БД была развернута в docker Образе с минимум ресурсов база данных откликалась мгновенно, а все задержки были на стороне python. Пока, конечно, делать окончательные выводы преждевременно, нужен объем БД ~1000000 строк, но уже есть уверенность, что Clickhouse справится легко.

- Использование Python так же оправдалось. Опубликованные на RSS каналах данные не всегда не содержат ошибок или идеально отформатированы. Были случаи, когда у записи отсутствовало поле о названии категории или данные не корректно парсились и в DataFrame попадали в одно поле название новости и дата публикации. Из-за этого происходил сбой в определении даты и времени последней публикации. При запаковке в jar было бы сложно восстановить работу программы.