Описание финальной работы

Обзор проекта

Цель
Запуск первого проекта по работе с большими данными, включающий сбор данных, исследовательский анализ и упаковку приложения.

Сценарий

Diego Cervo / Adobe Stock

Компания NYC Taxi стремится улучшить свои услуги. У компании есть данные обо всех поездках на такси с 2009 по 2016 год, и они наняли нашу команду для предоставления аналитики и стратегических рекомендаций.

Описание данных:
1,4 миллиарда записей о поездках на такси, собранных в период с 2009 по 2016 год, в формате 35GB snappy Parquet (~400GB в несжатом формате CSV), что представляет собой полноценную задачу по обработке больших данных.

Источники данных:
1. nyc.gov
2. academictorrents.com

Технические требования

Для заказчика важно, чтобы проект соответствовал требованиям:

- **Язык разработки:** код проекта должен быть написан на Python с использованием PySpark для обработки данных.
- **Развёртывание кластера:** кластер для обработки данных должен быть развёрнут либо локально, либо с помощью Yandex Data Proc для масштабируемости и производительности.
- **Хранилище исходных данных:** исходные данные должны быть загружены в Yandex Object Storage (S3-совместимое хранилище) для обеспечения надёжного доступа к данным и интеграции с Data Proc.
- **Моделирование:** промежуточные слои данных должны быть сформированы с использованием подходов Инмана, Кимбалла или Data Vault.
- **Оркестрация:** вся оркестрация процессов должна быть реализована с помощью Apache Airflow (AF), чтобы обеспечить автоматизацию ETL-пайплайнов.
  Вам понадобится инструкция по автоматизации Yandex Managed Service для Apache Airflow
- **Хранилище результатов:** результаты анализа и витрины данных должны быть загружены и отдельно выложены в ClickHouse для дальнейшей аналитики и визуализации.
  Вам понадобится инструкция по обмену данными с Yandex Managed Service для ClickHouse

Анализ данных

Набор данных состоит из двух основных DataFrame:

1. **Данные о поездках на такси** — включают детали, такие как временные метки, расстояния поездок, количество пассажиров, способы оплаты и многое другое.
2. **Описание такси-зон** — метаданные о каждой зоне для контекста посадок и высадок пассажиров.

Аналитические вопросы

1. Определить зоны с наибольшим количеством посадок и высадок.

2. Выяснить пиковые часы для поездок на такси.
3. Проанализировать распределение поездок и понять мотивы пассажиров.
4. Определить пиковые часы для коротких и длинных поездок.
5. Ранжировать топ-3 зон посадки и высадки для разных типов поездок.
6. Оценить способы оплаты для различных типов поездок.
7. Отследить эволюцию предпочтений в оплате со временем.
8. Изучить возможность организации райдшеринга для коротких поездок в близлежащих зонах.

Ключевые выводы и предложения

1. **Предложение 1.**
   Из-за значительного перекоса данных в сторону Манхэттена рекомендуется внедрить модель дифференцированной оплаты в зависимости от зоны посадки/высадки и спроса для управления нагрузкой и максимизации дохода.
2. **Предложение 2.**
   Наблюдаются пиковые часы с высоким спросом, что открывает возможности для динамического ценообразования и повышения прибыльности.
3. **Предложение 3.**
   Существует чёткое разделение между длинными и короткими поездками, причём короткие поездки преобладают в богатых районах (например, бары, рестораны). Специальные промоакции для коротких поездок можно ориентировать на такие зоны.
4. **Предложение 4.**
   Наличные платежи становятся менее популярными. Необходимо обеспечить бесперебойную работу процессора для карт 24/7, чтобы удовлетворить растущие потребности клиентов.
5. **Предложение 5.**
   Множество близких поездок можно объединить. Для этого предлагается стимулировать пассажиров использовать групповые поездки со скидкой. Это позволит:
   - Снизить затраты.
   - Стать более конкурентоспособными за счёт снижения цен.
   - Снизить выбросы, что может позволить претендовать на субсидии для этого проекта.

Пример экономического обоснования для предложения 5

Создана модель для оценки потенциального экономического эффекта от группировки поездок:

- 5% поездок можно объединить в группы.
- 30% пассажиров соглашаются на групповую поездку.
- Предоставляется скидка 5 долларов за групповую поездку и дополнительная плата 2 доллара за индивидуальную (за конфиденциальность и время).

Итоги

Проект «Большие данные для такси» включает пять предложений, из которых каждое имеет потенциал улучшить бизнес. Лучшее предложение может обеспечить экономический эффект в размере ~1 миллиарда долларов. Оценочная стоимость консультационного проекта — от 10 000 до 100 000 долларов.

Формат сдачи и критерии оценки

Формат сдачи

Отправьте на проверку куратору следующие материалы.

- **Код на Python с использованием Spark** — скрипты должны включать полную обработку данных и оптимизацию для работы с большими объемами информации (более 1,4 млрд записей).
- **Код Python с DAG AirFlow** — процессы ETL должны быть расписаны через DAG в AirFlow и настроены на автоматизацию этапов обработки данных.
- **Скриншоты из AirFlow** с успешно выполненными задачами — требуется приложить визуальные доказательства успешного завершения DAG и всех тасков.
- **SQL-скрипты и скриншоты проверки данных** — SQL-запросы для финальной выгрузки в ClickHouse и доказательства корректности результатов на скриншотах, где видны конечные значения по ключевым метрикам.

Критерии оценивания финальной работы

- **Полнота выполнения обработки данных на Spark** — Оценивается корректность загрузки, очистки и подготовки данных с использованием PySpark. Код должен эффективно работать с большим объёмом данных, поддерживать масштабируемость и соответствовать требованиям производительности.

- **Корректность оркестрации в AirFlow** — Проверяется структура и оптимизация DAG. Оценивается правильное разбиение на задачи, последовательность и корректность взаимодействия с Yandex Data Proc и другими внешними сервисами. Скриншоты выполнения DAG и задач должны подтверждать успешное завершение процессов.
- **Эффективность SQL-запросов** — Оценивается правильность и оптимизация SQL-запросов для обработки данных, которые выгружаются в ClickHouse. Запросы должны быть оптимизированы для быстрой обработки данных без потери корректности.
- **Аналитическая точность** — Оценивается полнота и точность ответов на аналитические вопросы (например, зоны с высоким спросом, предпочтительные способы оплаты и т.д.). Анализ должен быть обоснованным и соответствовать ключевым бизнес-выводам проекта.
- **Качество и точность бизнес-предложений** — предложения должны опираться на данные и показывать стратегические выводы, такие как динамическое ценообразование, групповое путешествие или изменения в оплате.
- **Оценка экономического эффекта предложений** — Оценивается обоснование экономической выгоды, предложенной в проекте. Модель оценки должна учитывать конкретные показатели и потенциальное влияние на прибыль компании.