

Проблемы качества данных в архитектуре больших данных: методы оценки и улучшения

Саяпин Никита Анатольевич

Приамурский государственный университет им.Шолом-Алейхема

Студент

Аннотация

В статье рассматривается проблема качества данных в контексте архитектур больших данных (Big Data). Описываются основные проблемы качества, такие как пропуски, дубликаты, выбросы и несогласованные типы, их влияние на анализ и принятие решений. В практической части представлен эксперимент на Python с использованием библиотеки Pandas, демонстрирующий методы оценки и улучшения качества на примере симулированного датасета. Результаты подчеркивают важность регулярной проверки и очистки данных для повышения надежности систем Big Data.

Ключевые слова: большие данные, качество данных, оценка данных, улучшение данных, Python, Pandas, очистка данных, Big Data архитектура.

Problems of data quality in big data architecture: methods of assessment and improvement

Sayapin Nikita Anatolyevich

Sholom-Aleichem Priamursky State University

Student

Abstract

This article discusses the problem of data quality in the context of big data architectures. It describes the main quality issues, such as missing values, duplicates, outliers, and inconsistent types, and their impact on analysis and decision-making. In the practical part, an experiment using Python and the Pandas library is presented, demonstrating methods for assessing and improving quality on a simulated dataset example. The results highlight the importance of regular data validation and cleaning to enhance the reliability of Big Data systems.

Keywords: big data, data quality, data assessment, data improvement, Python, Pandas, data cleaning, Big Data architecture.

1. Введение

1.1 Актуальность

В эпоху больших данных (Big Data) объемы информации растут экспоненциально, но качество этих данных часто остается низким, что приводит к ошибкам в аналитике, машинном обучении и бизнес-решениях. Проблемы качества, такие как пропущенные значения, дубликаты, выбросы и

несогласованные форматы, усугубляются в распределенных архитектурах (например, Hadoop или Spark), где данные поступают из множественных источников. Методы оценки и улучшения качества, включая автоматизированные скрипты на Python, позволяют минимизировать риски. В данной статье исследуется, как эти проблемы влияют на устойчивость систем Big Data, и проводится эксперимент для оценки времени и эффективности очистки данных.

1.2 Обзор исследований

В статье Д.Э. Тыщенко и Г.О. Хорожева рассказывается о применении статистических методов при анализе больших данных. Рассматриваются различные этапы обработки и анализа больших объемов информации, начиная от сбора данных до построения прогнозных моделей. [1]. В работе Е.А.Поповой рассматривается определение понятия "большие данные" (Big Data) и обосновывается важность обработки, хранения и анализа таких объемов информации. Рассматриваются примеры использования больших данных в различных областях, таких как анализ социальных медиа данных, прогнозирование спроса, диагностика и генетический анализ. Проводится сравнительный анализ языков программирования для обработки больших данных: Python, R и Java. [2]. Е.Х. Абу говорит о возможностях и проблемах, с которыми сталкивается бухгалтерия в связи с появлением больших данных. [3]. В работе О.Л. Солдаткиной рассматриваются проблемы сбора больших данных с позиции законодательства о персональных данных [4].

1.3 Цель исследования

Целью исследования является разработка и демонстрация простого инструмента для анализа проблем качества данных в архитектуре Big Data с использованием Python и библиотеки Pandas. Работа направлена на изучение уязвимостей низкокачественных данных, предоставление практического примера для начинающих специалистов по data engineering и иллюстрацию важности методов оценки и улучшения для надежности аналитики в больших объемах данных.

2 Материалы и методы

В данном исследовании используется интерпретируемая среда Python 3.x для разработки скрипта анализа качества данных. Применяются библиотеки Python: pandas для обработки и анализа данных, numpy для числовых операций и matplotlib для визуализации.

2.1 Подготовка

Необходимо установить Python версии 3.x на систему. Библиотеки pandas, numpy и matplotlib устанавливаются с помощью pip: `pip install pandas numpy matplotlib`. Следует создать новый файл с именем `data_quality_analyzer.py` в текстовом редакторе или среде разработки, например, PyCharm или Visual Studio Code (Рис.1).

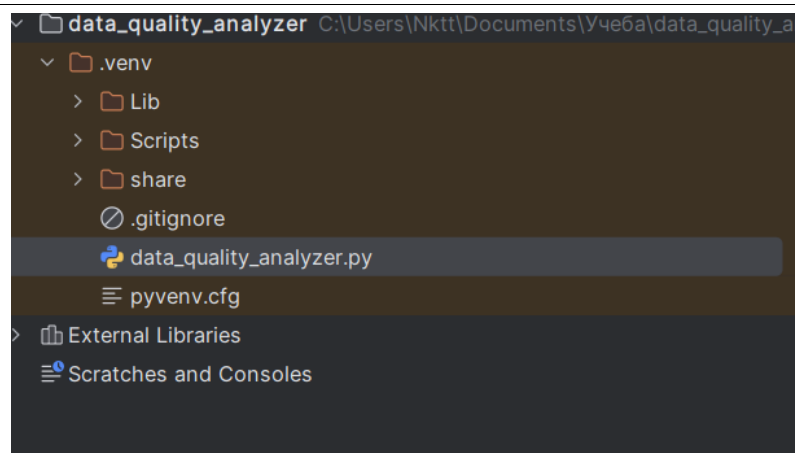


Рисунок 1 – Создание файла data_quality_analyzer.py в редакторе кода

3. Результаты

Для демонстрации проблем качества данных в условиях анализа больших массивов информации был сгенерирован искусственный набор данных, включающий типовые ошибки, встречающиеся в реальных системах. В исходный датасет намеренно внесены следующие искажения: дублирующиеся записи, пропуски в полях, некорректные значения в атрибуте количества (строковое представление числа, отрицательные значения и значительные выбросы). Сформированный датасет сохраняется в файл sales_data.csv и используется в дальнейшем для проведения эксперимента (Рис.2)

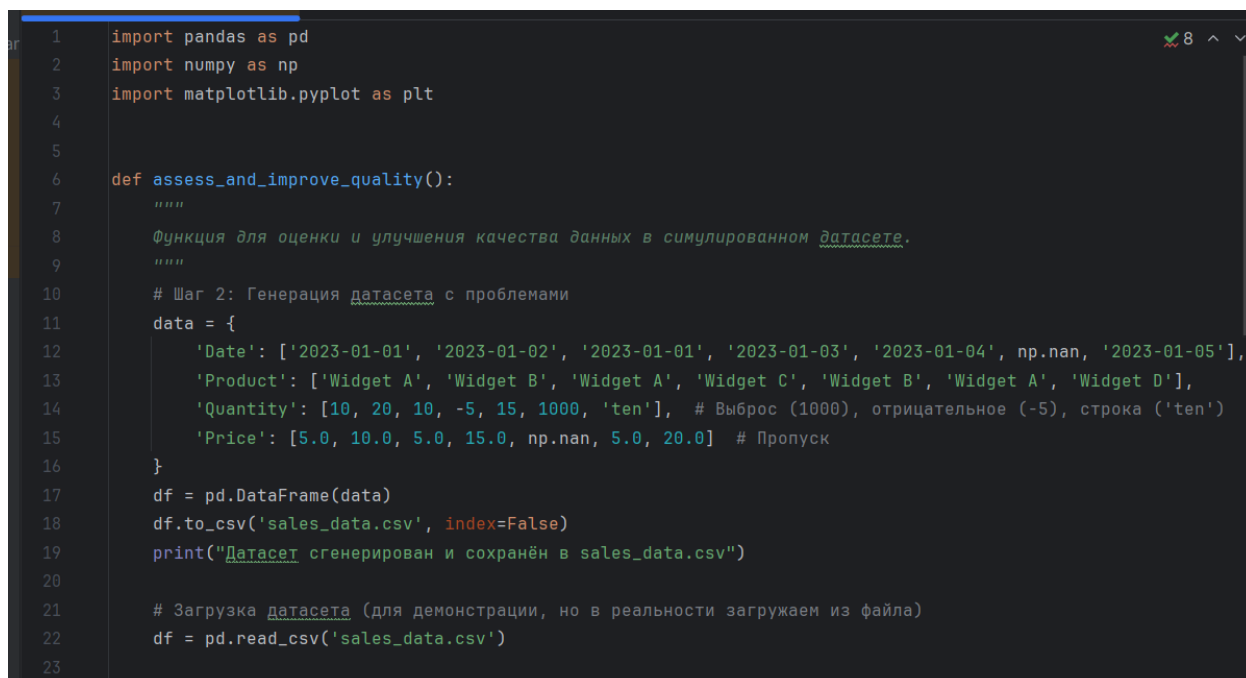


Рисунок 2 – Исходный набор данных

На первом этапе осуществляется загрузка подготовленного датасета и его первичная оценка. Для этого выводится общая информация о структуре и типах данных, производится подсчёт пропущенных значений, определяется

количество дублирующихся строк и рассчитываются описательные статистики по количественным признакам. Подобный подход позволяет выявить основные проблемы качества данных и зафиксировать наличие аномалий (Рис. 3).

```
21      # Загрузка датасета (для демонстрации, но в реальности загружаем из файла)
22      df = pd.read_csv('sales_data.csv')
23
24      # Шаг 3: Оценка качества данных
25      print("\n=== Оценка качества данных ===")
26
27      # 1. Общая информация
28      print("Общая информация о датасете:")
29      print(df.info()) # Покажет типы и пропуски
30
```

Рисунок 3 – Вывод общей информации о датасете

Для повышения наглядности дополнительно строится диаграмма размаха (boxplot) по атрибуту Quantity, где визуально отображаются выбросы и некорректные значения (Рис.4).

```
41      # Конвертируем Quantity в числовой тип, ошибки в NaN
42      df['Quantity'] = pd.to_numeric(df['Quantity'], errors='coerce')
43      print("\nСтатистика по Quantity:")
44      print(df['Quantity'].describe()) # Мин, макс, среднее – выявит выбросы
45
46      # Визуализация выбросов "до"
47      plt.figure(figsize=(8, 6))
48      plt.boxplot(df['Quantity'].dropna())
49      plt.title('Boxplot Quantity (до очистки)')
50      plt.ylabel('Quantity')
51      plt.show()
52
53      # Шаг 4: Улучшение качества данных
54      print("\n=== Улучшение качества данных ===")
55
56      # 1. Удаление дубликатов
57      df = df.drop_duplicates()
58      print(f"После удаления дубликатов: {df.shape[0]} строк")
59
60      # 2. Обработка пропусков
61      df['Date'].fillna('Unknown', inplace=True) # Для дат – 'Unknown'
62      df['Price'].fillna(df['Price'].mean(), inplace=True) # Средним для цены
63      df['Quantity'].fillna(df['Quantity'].median(), inplace=True) # Медианой для количества
64
65      # 3. Обработка выбросов в Quantity
66      # Заменяем отрицательные на 0, а >500 на среднее
67      mean_quantity = df['Quantity'].mean()
68      df['Quantity'] = df['Quantity'].apply(lambda x: 0 if x < 0 else (mean_quantity if x > 500 else x))
69
```

Рисунок 4 – Диаграмма размаха до очистки

Следующий этап посвящён очистке данных. В первую очередь удаляются дубликаты строк, что позволяет сократить объём избыточной информации. Далее выполняется обработка пропусков: в поле Date

отсутствующие значения заменяются текстовым индикатором Unknown, в поле Price они корректируются средним значением, а в поле Quantity — медианой, что обеспечивает минимизацию искажений распределения. Особое внимание уделяется выбросам: отрицательные значения количества заменяются нулём, а аномально большие значения, превышающие порог в 500 единиц, корректируются средним арифметическим значением по данному признаку. После проведения описанных процедур формируется очищенный набор данных, сохраняемый в файл `cleaned_sales_data.csv` (Рис.5)

```
70      # 4. Сохранение очищенного датасета
71      df.to_csv('cleaned_sales_data.csv', index=False)
72      print("\nОчищенный датасет сохранён в cleaned_sales_data.csv")
73
74      # Визуализация после
75      plt.figure(figsize=(8, 6))
76      plt.boxplot(df['Quantity'])
77      plt.title('Boxplot Quantity (после очистки)')
78      plt.ylabel('Quantity')
79      plt.show()
80
```

Рисунок 5 – Результат после обработки данных

Завершающий этап включает повторную оценку качества данных (Рис.6). Проверяется отсутствие дубликатов и пропусков, а также производится повторный расчёт статистических показателей по количественным данным. Полученные результаты свидетельствуют о том, что проведённая очистка позволила устранить аномалии и сделать структуру датасета более однородной. Для подтверждения корректности изменений строится обновлённая диаграмма размаха, на которой видно отсутствие экстремальных выбросов (Рис.7).

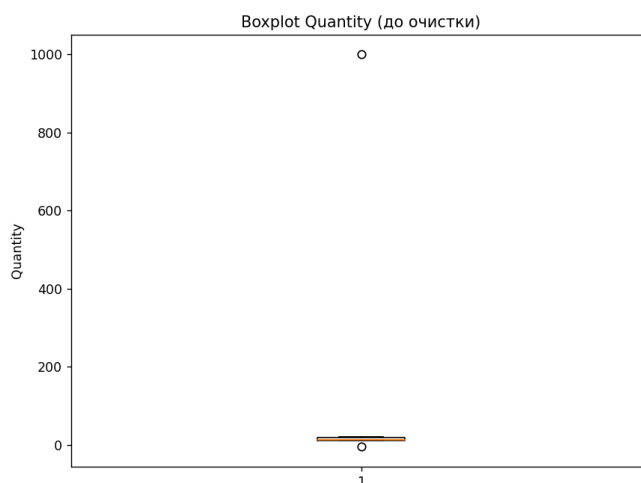


Рисунок 6 – График данных до очистки

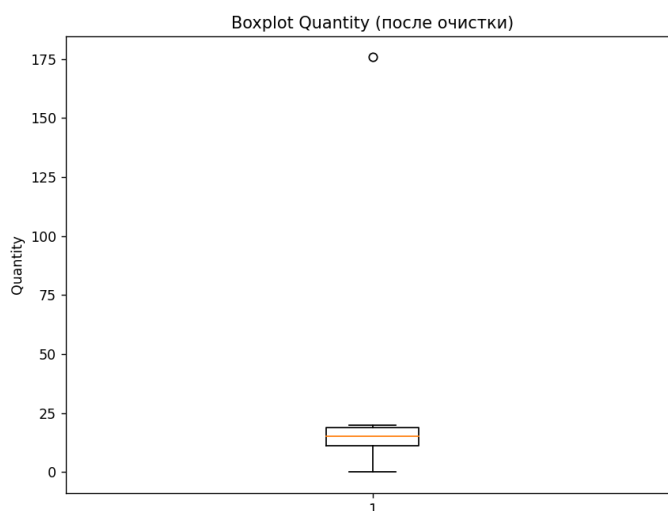


Рисунок 7 – График данных после очистки

4 Выводы

Проведённая работа наглядно демонстрирует необходимость системной оценки и обработки данных перед их дальнейшим использованием в аналитических задачах. В ходе эксперимента были выявлены и устранены типичные проблемы, характерные для реальных массивов данных: дубликаты, пропуски, некорректные и выбросные значения. Итоговый набор данных приобрёл целостный и корректный вид, что делает его пригодным для дальнейшего анализа и построения моделей.

Библиографический список

1. Тыщенко Д. Э. Анализ больших данных: статистические подходы к обработке и анализу больших данных // Аллея науки. 2024. Т. 1. № 11(98). С. 3-14
2. Полуэктов А. В., Попова И. В., Журавлева А.В. Большие данные (big data): обработка, хранение и анализ больших объемов данных // Перспективные аспекты моделирования систем и процессов: Материалы Всероссийской научно-практической конференции. Воронеж: Воронежский государственный лесотехнический университет им. Г.Ф. Морозова, 2023. С. 368-374.
3. Абу Е. Х. Большие данные и мир бухгалтерского учета: как повлияет появление больших данных на профессию бухгалтера? // Учетно-контрольные и аналитические процессы в условиях цифровизации экономики: Межвузовский сборник научных трудов и результатов совместных научно-исследовательских проектов. Том Часть 1. М.: Компания КиноРус, 2019. С. 19-22.
4. Иванов Д.С. Методы анализа безопасности веб-приложений // Журнал информационной безопасности. 2023. № 4. С. 25–31.