Модели построения информационно-аналитических средств киберфизических систем предприятий ТЭК

Д. П. Плахотников[1], Е. Е. Котова[2]
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]dimapl21@yandex.ru, [2]apu_kotova@mail.ru

***Аннотация*. Киберфизические системы представляют связанные между собой физические и информационные объекты. Для того, чтобы обрабатывать и отображать создаваемую данными системами информацию, необходимо построить информационно-аналитические средства. Могут быть использованы как простые модели, состоящие из одного инструмента, так и модели, состоящие из нескольких инструментов, взаимодействующих между собой с помощью различных методов.**

Ключевые слова: киберфизические системы; анализ данных; информационно-аналитические средства; большие данные; бизнес-аналитика

I. Введение

Киберфизические системы представляют собой объединение физических сущностей и информационных систем в единое целое. К киберфизическим системам можно отнести автоматизированные системы управления, сети электроснабжения, предприятия топливно-энергетического комплекса и многое другое. Главным фактором построения таких систем является глубокая интеграция информационных систем в физический мир с помощью различного рода сенсоров, датчиков, контроллеров.

Информационно-аналитические средства позволяют увеличить эффективность и оптимизировать работу киберфизических систем предприятий ТЭК [1].

II. Этапы построения информационно-аналитических средств

Для построения информационно-аналитических средств необходимо выполнить 4 основных этапа:

1) Определение целевых систем и получение информации с них;

2) Очистка и обработка полученных данных;

3) Визуализация и отображение данных;

4) Аналитика данных – поиск закономерностей, построение прогноза и т. д.

Для выполнения всех этапов построения предлагается несколько вариантов моделей информационно-аналитических средств:

1) Использование коммерческих систем бизнес-аналитики;

2) Использование свободного программного обеспечения;

3) Разработка собственного инструмента (с помощью языка программирования Python).

III. Построение информационно-аналитических средств с помощью систем бизнес-аналитики

Системы бизнес-аналитики включают набор методик, процессов, архитектур и технологий для переработки исходных данных в полезную информацию, используемую для эффективного принятия обоснованных решений на стратегическом, тактическом и операционном уровнях [2].

Бизнес-аналитика в широком смысле слова определяет:

- процесс превращения данных в информацию для поддержки принятия решений;
- информационные технологии, методы и средства сбора данных, консолидацию информации визуализации;
- углубленный анализа данных с помощью встроенных инструментов.

В основе технологии бизнес-аналитики лежит организация доступа конечных пользователей и анализ структурированных количественных по своей природе данных. Основные возможности систем бизнес-аналитики развиваются по четырем основным направлениям: хранение данных, интеграция данных, анализ данных и представление данных. Информация в хранилище данных, включая исторические данные, собирается из различных транзакционных систем и структурируется специальным образом для более эффективного анализа и обработки запросов (в отличие от обычных баз данных, где информация организована таким образом, чтобы оптимизировать время обработки текущих транзакций). Для решения более узких, конкретных задач из общего хранилища могут выделяться подмножества данных называемые витринами данных [3].

Среди преимуществ систем бизнес-аналитики:

- использование нескольких аналитических решений для различных направлений анализа;
- извлечение, анализ и консолидация данных из большого количества источников;

- обеспечение масштабируемости, эффективности и производительности;
- выстраивание и поддерживание в масштабах всего предприятия сквозных процедур и процессов обработки, единые централизованные аналитические модели и проекты;
- обеспечение доступа к данным и аналитическим инструментам большего числа пользователей.

Основной недостаток данного типа систем, что они являются коммерческими, с высокой стоимостью.

В основе систем бизнес-аналитики лежит ETL (Extract, Transform, Load) модуль, предназначенный для:

- загрузки данных из различных источников – файлов, реляционных и нереляционных баз данных, службы каталогов, веб-сайтов и других источников;
- очистки данных от ошибок с помощью специальных фильтров и условий;
- определения соответствия данных между справочниками и целевой системой;
- агрегации данных – объединение различных строк из таблиц с помощью специальных функций.

Схема ETL-процесса изображена на рис. 1.

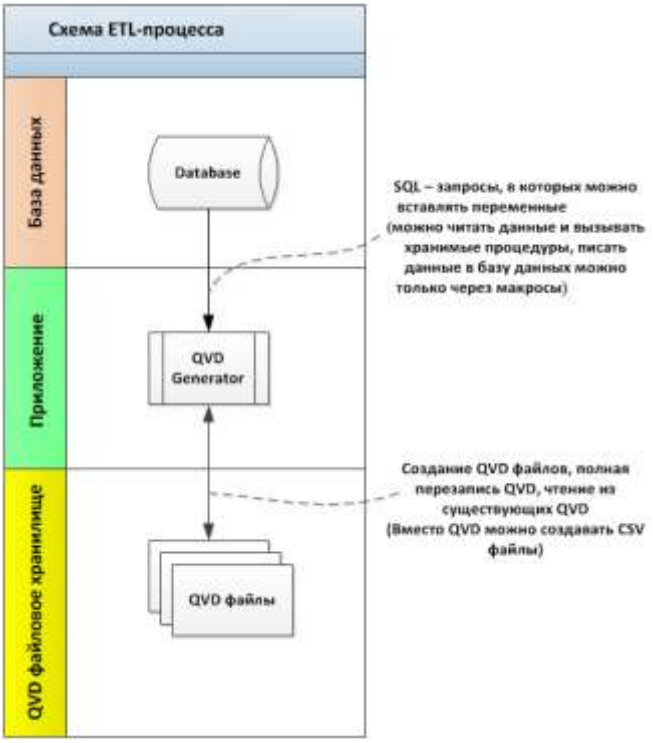

Рис. 1. Схема ETL процесса систем бизнес-аналитики

Так же платформы бизнес-аналитики поддерживают модель построения данных DAR: структура аналитических приложений строится таким образом, что пользователь переходит от общих ключевых показателей (Dashboard) к детализированным отчетно-аналитическим (Analysis) и отчетам (Reporting). Данная структура представлена на рис. 2.

Ключевые показатели отображают только самую необходимую информацию для понимания полной картины в целом и являются наименее интерактивной частью аналитического приложения. Страницы, посвященные анализу, являются более интерактивными и предназначены для полноценного исследование данных. В отчетах представлена наиболее детальная информация в виде таблиц.

В целом, трехступенчатая модель приложения DAR позволяет эффективно работать с информацией, упростить разработку и выполнения первичного анализа информации.

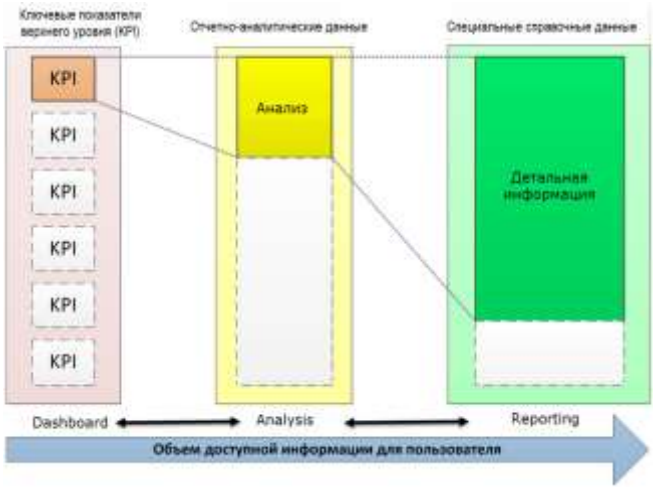

Рис. 2. Модель построения данных DAR

Системы бизнес-аналитики имеют встроенные и внешние инструменты глубокого анализа –машинное обучение, сегментация, прогнозирование, анализ сезонности, корреляция [4].

Подводя итог можно сказать, что модель построения информационно-аналитических средств с помощью систем бизнес-аналитики заключается в разработке ETL процесса для получения данных, их очистки и обработки, далее в использовании модели разработки визуализаций DAR и выполнение глубокого анализа с помощью встроенных или подключенных инструментов.

Инструментов бизнес-анализа достаточно для обработки больших данных создаваемых киберфизическими системами предприятиями топливно-энергетического комплекса, но для них нужны увеличенные вычислительные ресурсы из-за наличия более универсального инструментария используемого для широкого спектра задач.

IV. Построение информационно-аналитических средств с помощью свободного программного обеспечения

A. *Получение информации с баз данных и текстовых логов и обработки данных*

Существует разнообразное количество инструментов для получения информации с баз данных. Примером такого инструмента является ViXtract – сборка на основе популярных открытых инструментов обработки данных, которая помогает самостоятельно выгружать, очищать и преобразовывать данные.

В основе инструмента лежат три ключевых компонента: Jupyter – интерактивная среда для работы, PETL – библиотека преобразования данных, и Cronicle – планировщик с графическим интерфейсом. Источниками данных могут быть любые файловые источники и СУБД, а также API [5].

Особенность данного инструмента заключается в том, что он скриптовый. То есть требуется владение дополнительными навыками для написания процедур (по сравнению с системами бизнес-аналитики).

Альтернативным вариантом является использование инструмента с графическим интерфейсом получения и обработки данных – Apache NiFi. Данный инструмент содержит высокоуровневые функции для преобразования данных. Он поддерживает системную логику и масштабируемые графы маршрутизации данных [6].

Данный инструмент поддерживает настройку пропускного потока, устойчивости к потерям, динамическую расстановку приоритетов и безопасность передачи информации.

B. Визуализация полученных данных и их анализ

Для визуального отображения и анализа данных была использована платформа Grafana. Платформа с открытым исходным кодом для визуализации, мониторинга и анализа данных. Этот инструмент, в сочетании с Graylog, – часть двухсторонней системы мониторинга поведения пользователей и производительности системы. Grafana позволяет пользователям создавать дашборды с панелями, каждая из которых отображает определенные показатели в течение установленного периода времени.

С помощью специальных плагинов, таких как Sankey Map, Hierarchical View, FlowCharting и с помощью специальных формул формата *count_values_over_time(station_m [1m], 5)* есть возможность провести поверхностный анализ полученной информации.

К сожалению, среди открытых информационно-аналитических средств мало инструментов расширенной аналитики полученных данных. В основном используются внешние модули, разработанные на языке программирования Python.

В следующей главе описаны данные модули и их возможности для проведения анализа.

V. Построение информационно-аналитических средств с помощью разработки программного обеспечения

В качестве языка программирования разработки собственных информационно-аналитических средств использован Python, поскольку он имеет большое количество свободно распространяемых модулей для получения и аналитики больших данных.

A. Получение информации с баз данных и текстовых отчетов оборудования ТЭК

Для получения информации с баз данных необходимо использовать специальных модуль взаимодействия в зависимости от используемого типа базы данных. Список модулей представлен в табл. 1.

ТАБЛИЦА I. Перечень модулей python для работы с БД

База данных	**Модуль Python**
ODBC	pyodbc
PostgreSQL	psycopg2
SQLite	sqlite3
MySQL	mysql.connector

Все представленные библиотеки являются свободными к использованию

Первый тип баз данных представленный в таблице (ODBC) является универсальным. С помощью него можно подключаться к различным типам баз данных с помощью специального интерфейса.

Для запроса в базу киберфизической системы предприятия ТЭК необходимо использовать команду *cnc = psycopg2.connect(host=имя_хоста, user=имя_польз., password=пароль, dbname=наименования_базы)*

Представленная команда создает соединение с базой данных. Следующим этапом является создание курсора – специального объекта для выполнения запросов и получения их результатов: *cur = cnc.cursor()*

Для выполнения запроса необходимо выполнить команду формата *cur.execute("SELECT Station, Time, Value, FROM TranactionTable ORDER BY Time DESC LIMIT 1000")*

Результат выполненного запроса получается с помощью команды *data = cur.fetchall()*

Другим вариантом получения результата является специальный объект для анализа данных (dataframe): *df = DataFrame(cur.fetchall())*

Этот объект является одним из элементов библиотеки Pandas. Она позволяет обрабатывать, фильтровать и агрегировать и сохранять данные.

Библиотека Pandas позволяет обрабатывать текстовые отчеты оборудования. Пример команды чтения таких файлов: *df = pd.read_csv('sensor_data.txt', sep=" ", header=None)*

В следующем подразделы описаны примеры расширенной аналитики с помощью полученных данных.

B. Выполнение расширенного анализа данных

В качестве примера расширенного анализа данных были решены задачи прогнозирования данных и поиск выбросов.

Для прогнозирования данных киберфизических систем предприятий ТЭК был выбран метод сезонного авторегрессионного интегрированного скользящего среднего (SARIMA). Он прогнозирует следующий шаг в последовательности как линейную функцию разностных наблюдений, ошибок, разностных сезонных наблюдений и сезонных ошибок на предыдущих временных шагах [7].

Для загрузки библиотеки используется строка *from statsmodels.tsa.statespace.sarimax import SARIMAX*. Далее идет обучение и оценка модели: *model = SARIMAX(df,*

order=(1, 1, 1), seasonal_order=(1, 1, 1, 1)) model_fit = model.fit(disp=False)

И после построение прогноза: *yhat = model_fit.predict(len(data), len(data))*

Результат прогноза в виде графика представлен на рис. 3.

Рис. 3. Прогнозирование количество транзакций методом SARIMA

Следующий пример расширенной аналитики с помощью Python демонстрирует поиск выбросов. Обнаружение выбросов относится к определению редких элементов, которые отклоняются от общего распределения данных. Существующие подходы требуют высокой вычислительной сложности, либо имеют низкую предсказательную способность и ограниченную интерпретируемость.

В качестве решения проблемы ресурсов в сентябре 2020 года был представлен новый алгоритм обнаружения выбросов под названием COPOD, который вдохновлен копулами (copulas) для моделирования многомерного распределения данных. Алгоритм является вероятностным (probabilistic). COPOD сначала строит эмпирический пакет (empirical copula), а затем использует его для прогнозирования вероятностей хвоста каждой заданной точки данных, чтобы определить её уровень «экстремальности» [7].

Для загрузки библиотеки необходимо загрузить ее: *from pyod.models.copod import COPOD*

Следующим этапом является выделение необходимого поля для поиска выбросов: *raw=df['field'].values.reshape(-1,1).astype('float64').*

После этого выполняется поиск выбросов с помощью соответствующей функции *clf=COPOD()* и с помощью функции *clf.fit(raw)* выполняется построение графика представлено на рис. 4.

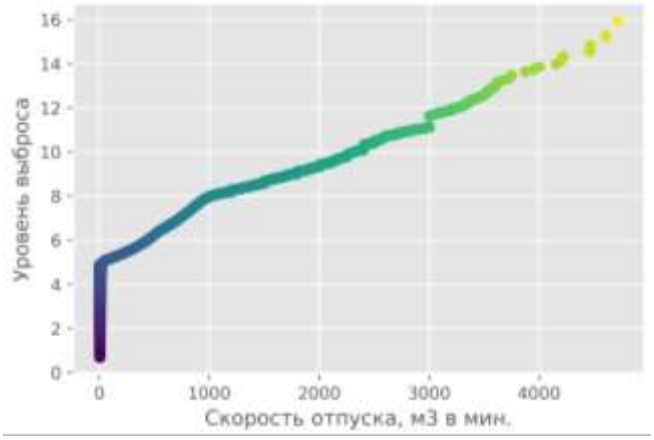

Рис. 4. Поиск аномалий методом Copula-Based Outlier Detection (COPOD)

VI. Заключение

Для построения информационно-аналитических средств киберфизических систем предприятий ТЭК существует различные модели различающиеся по сложности реализации, эффективности, возможности анализа и другим параметрам. В зависимости от поставленных задач и ресурсов может быть использована та или иная модель.

Список литературы

[1] Плахотников Д.П. «Методика повышения эффективности функционирования киберфизических систем предприятий ТЭК» // «Оригинальные исследования (ОРИС)». Том 13, № 2, 2023 г. с. 93-96.

[2] Business Intelligence [Электронный ресурс]. – Интернет-сайт. – URL: https://www.forrester.com/report/Topic-Overview-Business-Intelligence/RES39218 (дата обращения: 10.04.2023).

[3] Ralph Kimball, Margy Ross. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling - 3rd edition // Wiley. 2013. 601 p.

[4] Plakhotnikov D.P. "Ways of Forecasting Cyber-Physical Systems Characteristics," // 2021 IV International Conference on Control in Technical Systems (CTS), 2021, pp. 238-241, doi: 10.1109/CTS53513.2021.9562908.

[5] Vixtract [Электронный ресурс]. – Интернет-сайт. – URL: https://ru.visiology.su/ecosystem/vixtract (дата обращения: 10.04.2023).

[6] Apache NiFi [Электронный ресурс]. – Интернет-сайт. – URL: https://nifi.apache.org/ (дата обращения: 10.04.2023).

[7] A Gentle Introduction to SARIMA for Time Series Forecasting in Python [Электронный ресурс]. – Интернет-сайт. – URL: https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/ (дата обращения: 10.04.2023).

[8] Li, Zheng & Zhao, Yue & Botta, Nicola & Ionescu, Cezar & Hu, Xiyang. (2020). COPOD: Copula-Based Outlier Detection. 10.1109/ICDM50108.2020.00135.