Электронный научный журнал "Математическое моделирование, компьютерный и натурный эксперимент в естественных науках" http://mathmod.esrae.ru/
URL статьи: **mathmod.esrae.ru/41-163**
Ссылка для цитирования этой статьи:
Иванов Д.А. Разработка схемы использования программных средств для обработки и хранения данных в защищенном исполнении // Математическое моделирование, компьютерный и натурный эксперимент в естественных науках. 2023. №1

УДК 004.414 DOI: 10.24412/2541-9269-2023-1-28-33

РАЗРАБОТКА СХЕМЫ ИСПОЛЬЗОВАНИЯ ПРОГРАММНЫХ СРЕДСТВ ДЛЯ ОБРАБОТКИ И ХРАНЕНИЯ ДАННЫХ В ЗАЩИЩЕННОМ ИСПОЛНЕНИИ

Иванов Д.А.
Саратовский государственный технический университет имени Гагарина Ю.А.,
Россия, Саратов, d.ivanov.sstu@yandex.ru

DEVELOPMENT OF THE SCHEME OF USING SOFTWARE TOOLS FOR PROCESSING AND STORING DATA IN A PROTECTED VERSION

Ivanov D.A.
Yuri Gagarin State Technical University of Saratov, Russia,
Saratov, d.ivanov.sstu@yandex.ru

Аннотация. Для современных систем обработки и хранения данных основными условиями являются скорость и безопасность обработки информации. В связи с чем основной задачей при разработке таких систем является построение схемы обработки данных, компонентами которой должны являться программные средства способные обеспечить выполнение основных условий. Было рассмотрено построение типовой схемы обработки данных с учётом необходимого ПО для торговой сети, выбранной в качестве модели.

Ключевые слова: информационная система; фреймворк; хранилище данных; обработка данных; торговая сеть.

Abstract. For modern data processing and storage systems, the main conditions are the speed and security of information processing. In this connection, the main task in the development of such systems is the construction of a data processing scheme, the components of which should be software tools capable of ensuring the fulfillment of the basic conditions. The construction of a typical data processing scheme was considered, taking into account the necessary software for the trading network chosen as a model.

Keywords: information system; big data; framework; data store; data processing; commercial network.

В современных информационных системах со временем неизбежно растет количество обрабатываемых и хранимых данных. В то же время из-за

возрастающего количества хакерских атак и перехватов данных растет и количество угроз безопасности. Зачастую таким системам становится всё сложнее обрабатывать большие объемы поступающей информации, а также обеспечивать её защиту и сохранность. В связи с чем для информационных систем данного типа существуют проблемы своевременной обработки данных, результаты которой позволяют отразить реальное положение дел в организации, а также обеспечение защищенности этих данных при их обработке и хранении.

Для того, чтобы решить данные проблемы и минимизировать возникающие риски необходимо разработать систему, которая будет устойчива к растущей нагрузке и реализациям актуальных угроз безопасности.

В данном случае главным вопросом будет выбор средств разработки, которые смогут обеспечить корректное функционирование информационной системы, а также построение схемы обработки данных с учётом выбранных средств.

В качестве модели будем рассматривать организацию, которая представляет собой торговую сеть, состоящую из пресс-службы, службы персонала, отдела маркетинга, отдела программы лояльности, отдела развития, IT-отдела, отдела бухгалтерии, отдела логистики и отдела аренды. Организация обладает проблемами учёта продукции по причине возрастающего товарооборота, прогнозирования спроса на товары, а также обеспечения сохранности обрабатываемых и хранящихся данных.

Для достижения целей создаваемой системы необходимо решить следующие задачи:

- Обеспечить своевременную обработку, как поступающего, так и проданного товара;
- Обеспечить защищенность и сохранность персональных и иных данных об организации при их обработке и хранении.
- Еженедельно составлять прогноз о спросе на определенные товары, основываясь на котором будет возможность составлять список товаров для следующих поставок наиболее оптимально;
- Предлагать постоянным клиентам торговой сети персонализированные скидки и предложения, а также оповещать о предстоящих акциях;

Так как разрабатываемая информационная система должна быть в защищенном исполнении, то необходимо предварительно составить модель угроз, с помощью которой можно определить потенциальных нарушителей и угрозы, которые нарушители могут воспроизвести для доступа к системе и её данным. Составление модели угроз будет проводиться в соответствии с методическим документом «Методика оценки угроз безопасности», утвержденным ФСТЭК России 5 февраля 2021 года [2].

При разработке схемы обработки и хранения данных были рассмотрены основные составляющие, которыми являются обработка потоковых и

исторических данных, хранение и визуализация информации.

При потоковой обработке данных система получает информацию из какого-либо источника и записывает их в брокер сообщений. Сообщения в брокере должны хранится в отсортированном порядке, что позволяет гарантировать очередность. Далее необходима обработка данных из брокера в реальном времени. При этом обработка данных происходит распределено, что позволяет обрабатывать большие объемы данных с высокой эффективностью. После обработки все данные должны записываться в распределенное хранилище.

Обработка исторических данных отличается от потоковой тем, что данные уже хранятся и нет необходимости в дополнительном упорядочивании. Необходимо прочитать определенные данные из файлового хранилища и также как и в случае с потоковой обработкой составить скрипт с необходимой структурой обработки. Обработка при этом также происходит распределенно. После данного этапа данные заливаются в распределенное хранилище.

Все данные в информационной системе должны хранится в распределенном файловом хранилище, которое может гарантировать отказоустойчивость.

Также в информационной системе должна применятся реляционная база данных, которая необходима для дополнительной обработки небольшого количества данных и создания витрины, с которой в последствие очень удобно работать при отображении различных графиков и диаграмм.

На финальном этапе необходимо обеспечить визуализацию обработанных данных для отображения зависимостей.

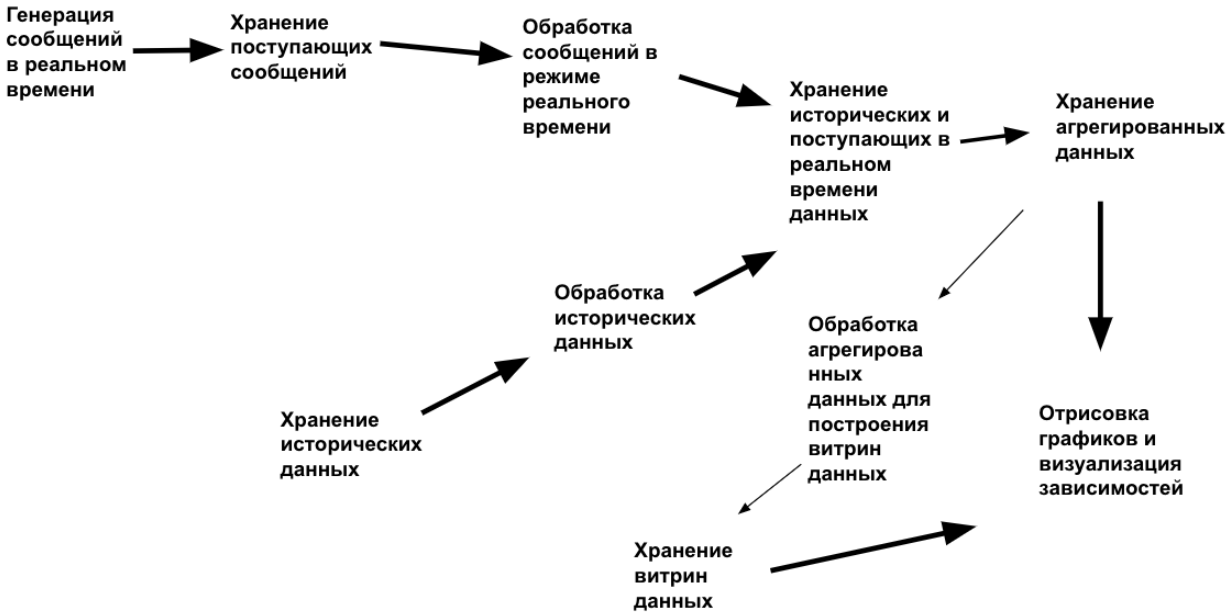

Рис. 1 - Схема обработки и хранения данных

Исходя из представленных выше требований общая схема обработки и хранения данных в информационной системе будет выглядеть следующим образом:

Необходимо рассмотреть каждую составляющую более подробно и выбрать конкретные средства разработки.

Система получает потоковые данные из генератора, который написан на языке Python и записывает их в брокер сообщений Kafka. Python позволяет написать наиболее лаконичный скрипт, который позволяет неоднократно использовать его в последующих итерациях, а брокер сообщений Kafka позволяет получать и сохранять большое количество событий, а также поддерживает протоколы шифрования, что очень необходимо в данной системе [7]. Затем фреймворк Spark с помощью Streaming библиотеки получает данные из брокера и обрабатывает их по заданному заранее образу в режиме реального времени. Выбор данного фреймворка обусловлен тем, что существует способность обработки больших объемов данных в реальном времени на распределенных машинах, а также благодаря библиотекам для использования криптографических методов можно обеспечить защищенность обработки [1]. После обработки все данные записываются в распределенное хранилище HDFS в виде файлов с заданной структурой.

Для обработки исторических данных также используется фреймворк Spark, поскольку имеются те же преимущества, что и при пакетной обработке (распределенность, хранение промежуточных данных в памяти).

Все данные в автоматизированной системе хранятся в файловом хранилище HDFS. Данное файловое хранилище является отказоустойчивым. Отказоустойчивость обеспечивается за счёт реплицирования данных на несколько серверов (по умолчанию 3), что позволяет в случае отказа одной машины восстановить данные с другой. Безопасности хранения данных можно достигнуть, используя шлюз Apache Knox Gateway, набор сервисов Apache Atlas и инфраструктуру для обеспечения, мониторинга и управления комплексной безопасностью Apache Ranger. Все данные хранятся в виде файлов с древовидной заданной структурой [3].

Также в автоматизированной системе применяется реляционная база данных PostgreSQL, так как она позволяет настраивать коннект с фреймворком Spark и читать данные из HDFS. Данные из файлового хранилища путем обработки фреймворком Spark попадают в базу данных после первой обработки [5-6].

Для визуализации данных применяется программная система Grafana. Данное программное обеспечение позволяет путём создания запросов на UI странице и выбора необходимых таблиц, представлений и файлов в файловом хранилище, которые хранятся в виде витрин реляционной базы данных создавать различные диаграммы (например, зависимость трат в месяц от времени года и т.д.).

Таким образом, итоговая схема использования технологий обработки

информации выглядит следующим образом:

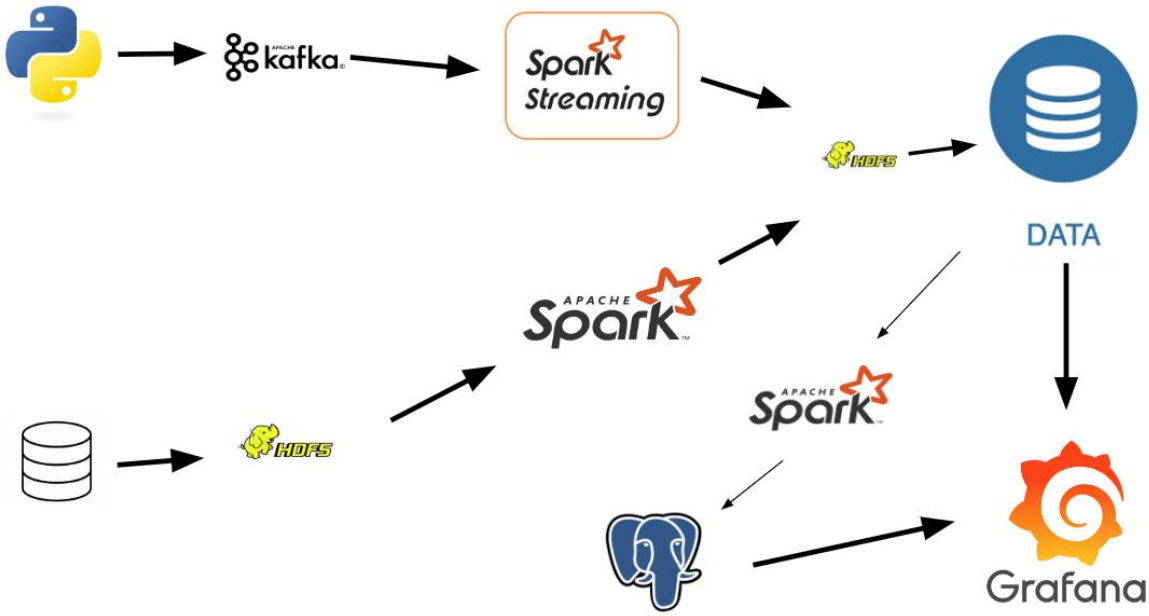

Рис. 2 - Итоговая схема использования средств обработки и хранения информации

В результате был произведен выбор средств разработки, которые отвечают поставленным задачам и позволяют предотвратить реализацию актуальных угроз, определенных в модели угроз, а также составлена схема обработки и хранения данных с учётом выбранных средств.

Данная схема обработки данных может применяться для типового построения информационных систем подобного типа, которые могут применяться в торговле, банковской отрасли, сфере оказания услуг связи и других сферах.

Литература

1. Изучаем Spark: молниеносный анализ данных / Х. Карау, Э. Конвински, П. Венделл, М. Захария. — Москва : ДМК Пресс, 2015. — 304 с. — ISBN 978-5-97060-323-9. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/90118 (дата обращения: 16.04.2023). — Режим доступа: для авториз. пользователей.
2. Методический документ ФСТЭК. «Методика оценки угроз безопасности информации».
3. Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 2-е изд., стер. — Санкт-Петербург : Лань, 2022. — 188 с. — ISBN 978-5-8114-9690-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/198599 (дата обращения: 16.04.2023). — Режим доступа: для авториз. пользователей.

4. Методы и модели исследования сложных систем и обработки больших данных: монография / И. Ю. Парамонов, В. А. Смагин, Н. Е. Косых, А. Д. Хомоненко; под редакцией В. А. Смагинаи А. Д. Хомоненко. — Санкт-Петербург: Лань, 2020. — 236 с. — ISBN 978-5-8114-4006-1. — Текст: электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/126938 (дата обращения: 24.04.2023). — Режим доступа: для авториз. пользователей.
5. Полякова, Л. Н. Основы SQL : учебное пособие / Л. Н. Полякова. — 2-е изд. — Москва : ИНТУИТ, 2016. — 273 с. — ISBN 978-5-94774-649-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/100348 (дата обращения: 24.12.2022). — Режим доступа: для авториз. пользователей.
6. Шёниг, Г. -. PostgreSQL 11. Мастерство разработки / Г. -. Шёниг ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2020. — 352 с. — ISBN 978-5-97060-671-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131714 (дата обращения: 24.04.2023). — Режим доступа: для авториз. пользователей.
7. Скотт, Д. Kafka в действии : руководство / Д. Скотт, В. Гамов, Д. Клейн ; перевод с английского А. Н. Киселева. — Москва : ДМК Пресс, 2022. — 310 с. — ISBN 978-5-93700-118-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/314888 (дата обращения: 24.04.2023). — Режим доступа: для авториз. пользователей.