УДК [004.4+658.5.011]:658.512

УПРАВЛЕНИЕ ЖИЗНЕННЫМ ЦИКЛОМ БОЛЬШИХ ДАННЫХ

В.В. Малиновская
Магистрант БГУИР,
проектный менеджер в игровой мобильной индустрии

В.Ф. Алексеев
Доцент кафедры проектирования информационных компьютерных систем, кандидат технических наук, доцент

Кафедра проектирования информационно-компьютерных систем факультета компьютерного проектирования Белорусского государственного университета информатики и радиоэлектроники, Республика Беларусь
Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
ООО «СейГеймс», Республика Беларусь
E-mail: viktoria.malinovskaya7@gmail.com

В.В. Малиновская
Окончила Белорусский государственный университет информатики и радиоэлектроники. Магистрант кафедры проектирования информационных компьютерных систем БГУИР. Работает в SayGames в должности проектного менеджера. Проводит исследования организации процессов управления на предприятии.

В.Ф. Алексеев
Окончил Минский радиотехнический институт. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем, исследованием проблем тепловой нестационарности полупроводниковых структур, изучением проблем обеспечения электромагнитной совместимости радиоэлектронных средств, организацией учебного и научно-исследовательского процессов в техническом университете.

Аннотация. В статье представлен архитектурный обзор управления жизненным циклом больших данных. Понимание этого процесса имеет важное значение для проектирования информационных систем. Правильное проектирование решения обеспечивает более быстрое его создание и позволяет решить многие проблемы эксплуатации на начальном этапе. Описанные в статье этапы позволяют наиболее удачное спроектироватьBig Data решение.

Ключевые слова: Аналитика больших данных, менеджмент, технологии, Data Science.

Введение.

В последнее время становится очевидным необходимость обработки больших объемов данных. По это причине всё более популярным становится термин Big Data. Очевидно, что появление этого понятия так или иначе связано с резким ростом количества доступной для анализа информации. Действительно, в большинстве статей на тему Big Data рассказ о новой технологии начинается с обсуждения проблемы экспоненциального роста объема данных. Объем данных не может являться точным критерием того, являются ли они большими. В информационной индустрии и научных кругах существует достаточно много различных определений больших данных. Авторами предлагается использовать наиболее краткое, но всеобъемлющее определение: «Большие данные – это большие объемы, высокоскоростные и разнообразные информационные активы, которые требуют экономичных, инновационных форм обработки информации для лучшего понимания и принятия решений».

Единственным отсутствующим ключевым словом в этом определении является «правдивость».

С точки зрения архитектуры и проектирования решений типичное решение для больших

данных, аналогичное традиционному жизненному циклу данных, может включать целый ряд различных этапов в общий процесс решения для жизненного цикла данных [1–8].

Анализ рынка Big Data, проведенный IDC, показывает, что возможно множество комбинаций программного и аппаратного обеспечения, сервисов, посредством которых реализуются успешные решения по анализу «больших данных».

Аналитические задачи (например, интеллектуальный анализ данных, многомерный анализ, визуализация данных) – наиболее частый пример использования Big Data, однако далеко не единственный. Технологии Big Data могут применяться также для поддержки социальных медиа- и игровых приложений, рассчитанных на огромное число пользователей [7].

Критерий последней категории представлен на рисунке 1. Здесь «Значение для бизнеса» формулируется как постоянное извлечение ценной информации для бизнеса [7].

Рисунок 1. Методика отнесения ИТ-проектов к Big Data (источник: IDC Russia)

Архитекторы решений Big Data участвуют во всех фазах жизненного цикла, обеспечивая различные входные данные и производя различные выходные данные для каждой фазы [1–6].

Эти фазы могут быть реализованы под различными именами в различных группах решения данных. При этом необходимо учитывать эффективность использования Big Data в жизненном цикле управления деятельностью компании.

В информационной индустрии отсутствует строгий универсальный системный подход к жизненному циклу больших данных, поскольку эта область все еще развивается [8].

Общий подход заключается в том, что опыт традиционного управления данными переносится и расширяется для конкретных сценариев использования решения.

При рассмотрении вопросов управления жизненным циклом больших данных можно предложить следующие основные этапы.

Этап 1: Основные моменты.

В процессе управления данными базовый этап включает различные аспекты, такие как понимание и проверка требований к данным, объем решения, роли и обязанности, подготовка инфраструктуры данных, технические и нетехнические соображения и понимание правил передачи данных в организации.

На этом этапе требуется детальный план, который в идеале был бы разработан менеджером проекта решения для обработки данных с существенным вкладом архитектора решения Big Data и некоторых специалистов по доменам данных.

Проект решения Big Data включает такие детали, как планы, финансирование, рекламные ролики, ресурсы, риски, допущения, проблемы и зависимости в отчет по определению проекта (PDR). Руководители проектов составляют и составляют PDR. Однако обзор решения в этом критическом артефакте предоставляется архитектором больших данных.

Фаза 2: Сбор данных.

Наборы данных могут быть получены из различных источников. Эти источники могут быть как внутренними, так и внешними для бизнес-организаций.

Данные могут быть предоставлены в структурированной форме, таких как переносимые из хранилища данных, различных операционных систем или полуструктурированных источников, таких как логи веб-приложений, системные журналы, или неструктурированные источники, такие как исходящие из медиафайлов, состоят из видео, аудио и картин.

Несмотря на то, что сбор данных осуществляется различными специалистами по данным и администраторами баз данных, архитектор больших данных играет существенную роль в ускорении и упрощении этого этапа.

Ведущий архитектор решения Big Data во взаимодействии с различными корпоративными и бизнес-архитекторами ведет и документирует стратегию сбора данных, требования пользователей, архитектурные решения, сценарии использования и технические спецификации на этом этапе.

В сложных решениях крупных бизнес-организаций ведущий архитектор Big Data может делегировать некоторые из этих видов деятельности различным архитекторам и специалистам по данным.

Этап 3: Подготовка данных.

На этапе подготовки данных собранные данные в необработанном формате находятся в состоянии “is cleaned” или “cleansed” – эти два термина взаимозаменяемы в различных методах обработки данных различных бизнес-организаций.

На данном этапе данные тщательно проверяются на наличие несоответствий, ошибок и дубликатов. Удаляются избыточные, дублированные, неполные и неверные данные. Цель этапа состоит в том, чтобы иметь чистые и пригодные для использования наборы данных.

Архитектура решения Big Data облегчает этот этап. Однако большинство задач по уборке данных из-за детализации операций могут выполняться специалистами по обработке данных, которые обучены методам подготовки и очистки данных.

Этап 4: Ввод данных и доступ к ним.

Ввод данных означает отправку данных в запланированные целевые репозитории данных, системы или приложения.

Например, мы можем отправить чистые данные к различным сервисам, такие как приложение CRM (управление взаимоотношениями с клиентами) или хранилище данных для использования определенными отделами. На этой фазе специалисты по данным преобразуют необработанные данные в пригодный для использования формат.

Доступ к данным можно получить с помощью различных методов. Эти методы могут включать использование реляционных баз данных, плоских файлов или NoSQL. Наиболее актуальным является NoSQL, который широко используется для решений Big Data в различных бизнес-организациях.

Несмотря на то, что этот этап возглавляет архитектор Big Data, они обычно делегируют подробные операции специалистам по данным и администраторам баз данных, которые могут выполнять требования к вводу и доступу на этой фазе.

Этап 5: Обработка данных.

Фаза обработки данных начинается с обработки неструктурированных данных. Затем преобразуем данные в читаемый формат, давая им форму и контекст. После завершения этой операции можно интерпретировать данные с помощью выбранных инструментов аналитики данных для нашей бизнес-организации.

Популярными инструментами обработки данных в режиме реального времени в большинстве решений были HBase, а инструментом обработки данных в режиме реального времени – Spark Streaming. На рынке существует множество инструментов с открытым исходным кодом и запатентованных инструментов.

Обработка данных также включает такие операции, как аннотирование данных, интеграция данных, агрегирование данных и представление данных.

Интеграция данных направлена на объединение данных, существующих в различных источниках, и обеспечивает унифицированное представление данных потребителям данных.

Представление данных относится к способу обработки, передачи и сохранения данных. Эти три важные функции отображают представление данных в жизненном цикле.

Агрегирование данных направлено на компиляцию данных из баз данных в объединенные наборы данных, которые будут использоваться для обработки данных.

На этапе обработки данных данные могут изменяться в соответствии с требованиями потребителей. Обработанные данные могут использоваться в различных выходах данных в озерах данных, в корпоративных сетях и подключенных устройствах.

Обработка больших данных требует участия различных членов команды с различными наборами навыков.

В то время как ведущий архитектор решения *Big Data* возглавляет этап обработки, большинство задач выполняются специалистами по обработке данных, управляющими данными, инженерами по обработке данных и учеными по обработке данных.

Архитектура решения *Big Data* упрощает сквозной процесс для этой фазы.

Этап 6: Вывод и интерпретация данных.

На данной фазе данные находятся в формате, готовом для использования бизнес-пользователями. Можно преобразовать данные в пригодные для использования форматы, такие как обычный текст, графики, обработанные изображения или видеофайлы.

На фазе вывода данные объявляются готовыми к использованию и передаются на следующую фазу для хранения. Эта фаза в некоторых методах обработки данных и бизнес-организации также называется приемом данных. Например, процесс приема данных предназначен для импорта данных для немедленного или будущего использования или сохранения их в формате базы данных.

Процесс приема данных может выполняться в режиме реального времени или в пакетном формате. Некоторыми стандартными инструментами приема больших данных, которые обычно использовались в решениях, были потоковая передача *Sqoop, Flume* и *Spark*. Это популярные инструменты с открытым исходным кодом.

Одной из операций является интерпретация проглоченных данных. Для выполнения этой операции необходимо проанализировать полученные данные и извлечь из них информацию или смысл, чтобы ответить на вопросы, связанные с бизнес-решениями *Big Data*.

Этап 7: Место хранения данных.

После завершения фазы вывода данных они сохраняются в спроектированных и назначенных единицах места хранения. Эти блоки являются частью проектирования платформы данных и инфраструктуры с учетом всех нефункциональных архитектурных аспектов, таких как емкость, масштабируемость, безопасность, соответствие нормативам, производительность и доступность.

Инфраструктура может состоять из сетей места хранения (*SAN*), сетевых мест хранения данных (*NAS*) или форматов места хранения данных прямого доступа (*DAS*). Администраторы данных и баз данных могут управлять сохраненными данными и предоставлять доступ к определенным группам пользователей [9].

Место хранения больших данных может включать в себя базовые технологии, такие как кластеры баз данных, реляционное место хранения данных или расширенное место хранения данных, например *HDFS* и *HBASE*, которые являются системами с открытым исходным кодом.

Кроме того, форматы файлов, такие как текстовый, двоичный или другие специализированные форматы, такие как *Sequence, Avro* и *Parquet*, должны учитываться на этапе проектирования места хранения данных.

Этап 8: Интеграция данных.

В традиционных моделях после сохранения данных процесс управления данными завершается. Однако в случае больших данных может возникнуть необходимость интеграции хранимых данных в различные системы для различных целей.

Интеграция данных является сложной и важной архитектурной задачей в процессе решения больших данных. Архитекторы больших данных участвуют в проектировании и проектировании различных соединителей данных для интеграции решений больших данных.

Могут существовать примеры использования и требования для многих соединителей, таких как *ODBC, JDBC, Kafka, DB2, Amazon S3, Netezza, Teradata, Oracle* и многих других, основанных на источниках данных, используемых в решении.

Некоторые модели данных могут потребовать интеграции озер данных с хранилищем данных или витринами данных. Также могут существовать требования к интеграции приложений для решений *Big Data*.

Например, некоторые операции интеграции могут включать интеграцию больших данных с панелями мониторинга, таблицей, веб-сайтами или различными приложениями визуализации данных. Эта операция может перекрываться со следующей фазой – аналитикой данных.

Этап 9: Анализ и визуализация данных.

Интегрированные данные могут быть полезными и продуктивными для анализа и визуализации данных.

Аналитика данных является важным компонентом процесса управления большими данными. Этот этап имеет решающее значение, поскольку именно здесь решения для больших данных приносят пользу бизнесу.

Визуализация данных является одной из ключевых функций этой фазы.

Можно использовать множество инструментов для анализа и визуализации, основанных на требованиях решения.

Архитекторы решений для больших данных играют ограниченную роль на этом этапе, однако они тесно сотрудничают с учеными по данным, чтобы обеспечить соответствие практики и платформ аналитики бизнес-целям.

Архитекторы *Big Data* должны обеспечить выполнение этапов жизненного цикла с архитектурной жесткостью.

Фаза 10: Потребление данных.

После выполнения анализа, данные преобразуются в информацию, готовую для использования внутренними или внешними пользователями, включая клиентов бизнес-организации.

Потребление данных требует архитектурного входа для политики, правил, инструкций, принципов и рекомендаций. Например, потребление данных может быть основано на процессе предоставления услуг. Органы управления данными разрабатывают правила предоставления услуг.

Ведущий архитектор решений для больших данных возглавляет и облегчает создание этих политик, правил, принципов и руководств с использованием архитектурной структуры, выбранной в бизнес-организациях.

Этап 11: хранение, резервное копирование и архивирование.

Для обеспечения защиты и соответствия отраслевым требованиям необходимо резервное копирование критически важных данных.

Необходимо использовать установленные стратегии резервной копии данных, методы и инструменты. Архитектор решения *Big Data* должен определять, документировать и получать одобрение решений по хранению, резервному копированию и архивированию.

Архитектор решения *Big Data* может делегировать детальное проектирование этого этапа архитектору инфраструктуры, которому помогают несколько специалистов по доменам данных, баз данных, мест хранения и восстановления [10].

Некоторые данные по нормативным или иным бизнес-причинам должны храниться в течение определенного периода времени. Стратегия хранения данных должна быть задокументирована и утверждена руководящим органом, особенно корпоративными архитекторами, а также внедрена архитекторами инфраструктуры и специалистами по месту хранения.

Фаза 12: Уничтожение данных.

Могут существовать нормативные требования к уничтожению определенного типа данных после определенного количества раз.

Требования к уничтожению могут меняться в зависимости от отраслей.

Необходимо подтвердить требования к уничтожению с помощью команды управления данными в бизнес-организациях.

Заключение.

Существует хронологический порядок управления жизненным циклом, для создания решений для больших данных. Некоторые фазы могут слегка перекрываться и могут выполняться параллельно.

Жизненный цикл, предложенный в этой статье, является лишь ориентиром для осознания всего процесса. Процесс можно настроить на основе структуры группы решений для обработки данных, уникальных платформ организационных данных, требований к решениям для обработки данных, сценариев использования и динамики организации-владельца, ее отделов или общей экосистемы предприятия.

Список литературы

[1] Фрэнкс Билл. Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики / Билл Фрэнкс. – М.: Альпина Паблишер, 2016. – 430 с.

[2] Вайгенд Андреас. BIG DATA. Вся технология в одной книге / Андреас Вайгенд. – Москва: Эксмо, 2018. – 384 с.

[3] Алексеев, В.Ф. Информационная поддержка управления инновационной деятельностью предприятия / В.Ф. Алексеев, Д.В. Лихачевский, В.В. Хорошко // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня: сб. материалов VI Междунар. науч.-практ. конф., Минск, 20-21 мая 2020 года: в 3 ч. Ч. 3 / редкол.: В.А. Богуш [и др.]. – Минск: Бестпринт, 2020. – С. 412-417.

[4] Алексеев, В.Ф. Разработка онлайн платформы оценки и финансирования инновационных проектов / В.Ф. Алексеев, Д.В. Лихачевский, Г. А. Пискун // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018 / editorial board: M. Batura [etc.]. – Minsk, BSUIR, 2018. – P. 398 – 404.

[5] Алексеев, В.Ф. Взаимосвязь операционного и логистического менеджмента в структуре маркетинговой и производственной стратегии фирмы / В.Ф. Алексеев // Экономическое развитие общества: инновации, информатизация, системный подход. Международная научно-практическая конференция. Тезисы докладов. − Минск: Изд-во «ПАРАДОКС», 2008. – С. 310-312.

[6] Алексеев, В.Ф. Анализ системы маркетинга на предприятии и её совершенствование с использованием Internet-технологий / В.Ф. Алексеев [и др.] // Современные информационные компьютерные технологии: Сб. науч. ст. в 2ч. Ч.1 − Гродно: ГрГУ, 2008. – С. 118=122.

[7] Большие данные: насколько они большие? [Электронный ресурс]. – Режим доступа: https://compress.ru/article.aspx?id=23469

[8] Радченко, И.А. Технологии и инфраструктура Big Data / И.А. Радченко, И.Н. Николаев. – СПб: Университет ИТМО, 2018. – 52 с.

[9] CoderLesson.com [Электронный ресурс]. – Режим доступа: https://coderlessons.com

[10] BuranGroup [Электронный ресурс]. – Режим доступа: http://buran.group/solutions/big_data

[11] ITEnterprise [Электронный ресурс]. – Режим доступа: https://www.it.ua

BIG DATA LIFECYCLE MANAGEMENT

V.V. MALINOVSKAYA
BSUIR Master,
Project Manager
SayGames

V.F. ALEKSEEV
Associate Professor, Department of Information Computer Systems Design, Candidate of Technical sciences, Associate Professor

Department of Information and Computer Systems Design
Faculty of Computer Engineering
Belarusian State University of computer science and Radio Electronics, Republic of Belarus
EPAM Systems, Republic of Belarus
E-mail: viktoria.malinovskaya7@gmail.com

Abstract. This article provides an architectural overview of big data lifecycle management. Understanding this process is important for designing big data solutions. The correct design of the solution enables faster creation and solves many operational problems in the initial phase. The steps described in the article allow the most successful design of the Big Data solution.

Keywords: Management, Technology, Data Science, Big Data Analytics.