УДК 004.056
МРНТИ 20.53.17

РАЗРАБОТКА СИСТЕМЫ ОЦЕНКИ ИНФОРМАЦИОННО-ТЕХНОЛОГИЧЕСКОГО ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ БОЛЬШИХ ДАННЫХ

И. УВАЛИЕВА[1], Г. СОЛТАН[2]

[1]Восточно-Казахстанский государственный технический университет им. Д. Серикбаева
[2]Казахский агротехнический университет им. Сакена Сейфуллина

Аннотация: *В данной работе описана система безопасности больших данных, основанная на алгоритме оценки информационно-технологического обеспечения безопасности информации. Алгоритм основан на вычислении комплексного показателя информационной безопасности, позволяющий оценивать состояние выполнения организационных и технических требований по обеспечению защиты информации. Результаты оценки представляются достаточно наглядно и обеспечивают возможность руководству организаций принимать обоснованные решения по дальнейшему совершенствованию системы защиты информации.*

Ключевые слова: *информационная безопасность, защита информации, система безопасности, обеспечение безопасности, большие данные*

DEVELOPMENT OF THE SYSTEM FOR ASSESSMENT OF INFORMATION AND TECHNOLOGICAL SUPPORT OF BIG DATA SECURITY

Abstract: *This paper describes a big data security system based on an algorithm for evaluating information technology security of information security. The algorithm is based on the calculation of a comprehensive information security indicator, which allows assessing the state of compliance with organizational and technical requirements for ensuring information security. The results of the assessment are presented quite clearly and provide an opportunity for the management of organizations to make informed decisions to further improve the information security system.*

Keywords: *information security, information protection, safety system, security, big data*

ҮЛКЕН ДЕРЕКТЕР ҚАУІПСІЗДІГІНІҢ АҚПАРАТТЫҚ-ТЕХНОЛОГИЯЛЫҚ ҚОЛДАУЫН БАҒАЛАУДЫҢ ЖҮЙЕСІН ҚҰРУ

Аңдатпа: *Бұл мақалада ақпараттық қауіпсіздікті бағалаудың алгоритміне негізделген үлкен деректерді қорғау жүйесі сипатталған. Алгоритм ақпараттық қауіпсіздікті қамтамасыз етудің ұйымдастырушылық және техникалық талаптарына сәйкестігін бағалауға мүмкіндік туғызатын толық ақпараттық қауіпсіздіктің кешенді көрсеткішін есептеуге негізделген. Бағалаудың нәтижелері басқарушыға айқын түрде ұсынылып, ақпараттық қауіпсіздікті одан әрі жетілдіру бойынша шешімдер қабылдайды.*
Ақуыздың үшінші реттік құрылымдық деректер жинағының физикалық және химиялық қасиеттеріне байланысты көрсеткіштердің интеллектуалдық өңделуін іске асыру айқындалады. Деректерді өңдеуді жүргізуде тазалау, спектралдық анализ және кластерлеу әдістері қолданылды. Кластерлеудің интеллектуалдық әдісін жүзеге асыру нәтижесінде ақуыздың үшінші реттік құрылым көрсеткіштерінің үш үлкен және негізгі кластерлері алынды.

Түйінді сөздер: *ақпараттық қауіпсіздік, ақпаратты қорғау, қауіпсіздік жүйесі, қауіпсіздікті қамтамасыз ету, үлкен деректер*

Введение

Современные технологии, которые позволяют накапливать огромное количество информации, создают новые формы бизнеса. Так, по данным исследований International Data Corporation, ежегодно объемы хранимой информации вырастают на 40%. В этих условиях информация перестает быть самоценностью, на передний план выходят способы ее обработки и использования [1]. В результате использования технологий Big Data компании имеют возможность получать важную информацию за несколько секунд, что позволяет повышать эффективность экономических решений, быстрее реагировать на изменения в поведении клиентов, в режиме реального времени выявить рыночные тренды на самых ранних этапах [2].

В большинстве случаев работа с большими данными подразумевает стандартный рабочий процесс: от сбора необработанных данных и до получения пригодной для использования информации. Основная цель работы с большими данными – это получение на их основании ценных аналитических выводов для практического применения. В идеале большие данные должны становиться доступными для всех заинтересованных сторон, чтобы они получали возможность легко и быстро изучать пакеты данных с помощью инструментов бизнес-аналитики и настраиваемой визуализации, рассчитанных на самостоятельное использование. В зависимости от типа аналитики конечным пользователям могут предоставляться готовые результаты в форме данных статических «прогнозов» (в случае прогнозирующей аналитики) или рекомендованных действий (в случае предписывающей аналитики) [3].

1 Особенности систем с большими данными

Большие данные получили широкое распространение во многих отраслях бизнеса. Их используют в здравоохранении, телекоммуникации, торговле, логистике, в финансовых компаниях, а также в государственном управлении. Технологии больших данных применяются в основном для анализа клиентской среды. Так, HSBC использует технологии Больших данных для противодействия мошеннических операций с пластиковыми картами. С помощью Big Data компания увеличила эффективность службы безопасности в 3 раза, распознавание мошеннических инцидентов – в 10 раз. Экономический эффект от внедрения данных технологий превысил 10 млн долл. США [4].

Большие данные, помимо уже известных и распространенных задач, можно использовать в том числе для борьбы с заболеваниями и отслеживания роста эпидемии, считают эксперты. В Германии, например, уже сегодня благодаря технологиям Больших данных онкологические заболевания либо предрасположенность к ним, выявляются по анализу крови пациентов и доноров. Существует ряд аппаратно-программных комплексов, предоставляющих предконфигурированные решения для обработки больших данных: Aster MapReduce appliance (корпорации Teradata), Oracle Big Data appliance, Greenplum appliance. Эти комплексы поставляются как готовые к установке в центры обработки данных телекоммуникационные шкафы, содержащие кластер серверов и управляющее программное обеспечение для массово-параллельной обработки [5].

Аппаратные решения для резидентных вычислений, прежде всего, для баз данных в оперативной памяти и аналитики в оперативной памяти, в частности, предлагаемой аппаратно-программными комплексами Hana и Exalytics, также иногда относят к решениям из области больших данных, несмотря на то, что такая обработка изначально не является массово-параллельной, а объёмы оперативной памяти одного узла ограничиваются несколькими терабайтами.

Кроме того, иногда к решениям для больших данных относят и аппаратно-программные комплексы на основе традиционных реляционных систем управления базами данных – Netezza, Teradata, Exadata, как способные эффективно обрабатывать терабай-

ты и эксабайты структурированной информации, решая задачи быстрой поисковой и аналитической обработки огромных объёмов структурированных данных. Отмечается, что первыми массово-параллельными аппаратно-программными решениями для обработки сверхбольших объёмов данных были машины компаний Britton Lee, впервые выпущенные в 1983 году, и Teradata [6].

Система хранения данных (СХД) представляет собой конгломерат программного обеспечения и специализированного оборудования, предназначенный для хранения и передачи информации больших объемов. Особенностью СХД является оптимальное распределение ресурсов при хранении информации на дисковых площадках. Надежное хранение данных и быстродействие доступа к ним требуют организации средств хранения, как отдельной подсистемы вычислительных комплексов. Аппаратные решения DAS – систем хранения данных, напрямую присоединённых к узлам, в условиях независимости узлов обработки в SN-архитектуре также иногда относят к технологиям больших данных. Именно с появлением концепции больших данных связывают всплеск интереса к DAS-решениям в начале 2010-х годов, после вытеснения их в 2000-е годы сетевыми решениями классов NAS и SAN [7-9].

Наглядное представление результатов анализа больших данных имеет принципиальное значение для их интерпретации. Не секрет, что восприятие человека ограничено, и ученые продолжают вести исследования в области совершенствования современных методов представления данных в виде изображений, диаграмм или анимаций [10].

На рисунке представлены популярные технологии при использовании больших данных [3].

На сегодняшний день некоторые из перечисленных в предыдущем подразделе подходов или определенную их совокупность позволяют реализовать на практике аналитические движки для работы с большими данными. Из свободных или относительно недорогих открытых систем анализа Big Data можно порекомендовать: 1010data; Apache Chukwa; Apache Hadoop; Apache Hive; Apache Pig!; Jaspersoft; LexisNexis Risk Solutions HPCC Systems; MapReduce; Revolution Analytics (на базе языка R для мат.статистики).

Как мы видим на рисунке 30% выбрали решение на базе In-memory платформы (SAP, HANA, Oracle Exadata и др.). 18% – платформы NoSQL. 15% – Log-file аналитические программы (Splunk, InTrust Dell и др.). 12% – Columnar платформы (1010data, Calpont, HP Vertica и др.). 11% – Hadoop/MapReduce. 14% – другие [7].

MapReduce – это модель распределенной обработки данных, предложенная компанией Google для обработки больших объёмов данных на компьютерных кластерах [8]. Прин-

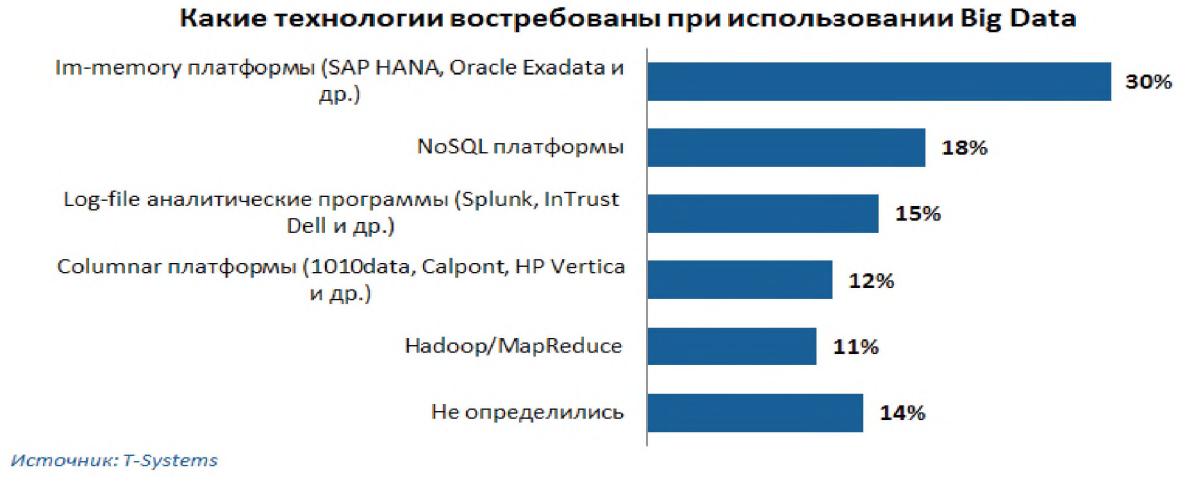

Рис. 1 – Технологии использования big data

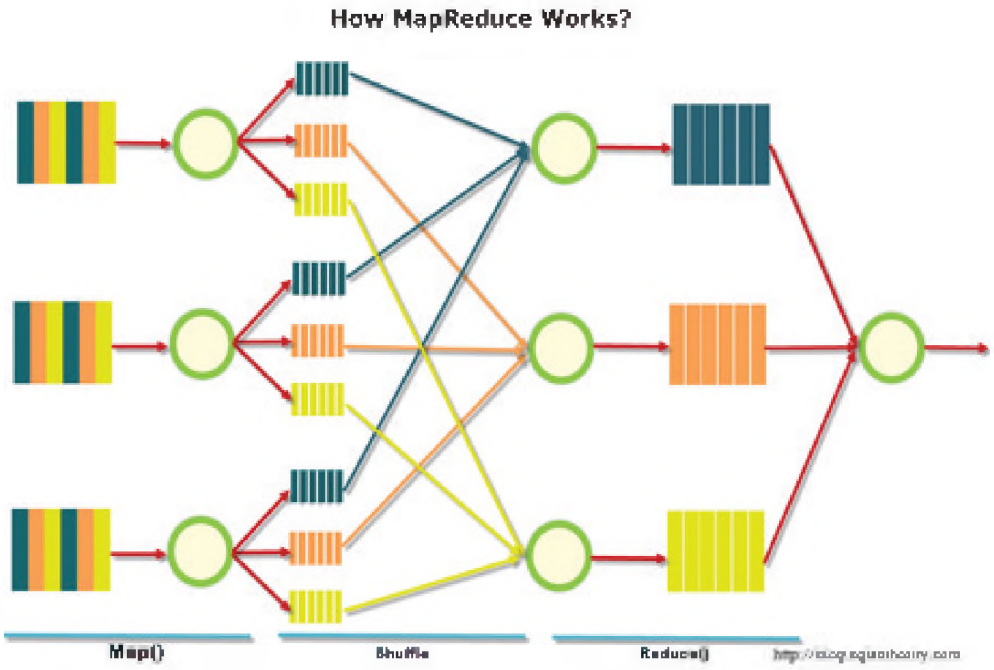

Рис. 2 – Схема функционирования MapReduce

цип работы технологии MapReduce представлен на рисунке 2.

MapReduce предполагает, что данные организованы в виде некоторых записей. Обработка данных происходит в 3 стадии. На стадии Map данные предобрабатываются при помощи функции map(), которую определяет пользователь. Работа этой стадии заключается в предобработке и фильтрации данных. Работа очень похожа на операцию map в функциональных языках программирования – пользовательская функция применяется к каждой входной записи [7].

Функция map(), примененная к одной входной записи, и выдаёт множество пар ключ-значение. Множество – т.е. может выдать только одну запись, может не выдать ничего, а может выдать несколько пар ключ-значение. Что будет находиться в ключе и в значении – решать пользователю, но ключ – очень важная вещь, так как данные с одним ключом в будущем попадут в один экземпляр функции reduce [8].

Стадия Shuffle проходит незаметно для пользователя. В этой стадии вывод функции map «разбирается по корзинам» – каждая корзина соответствует одному ключу вывода стадии map. В дальнейшем эти корзины послужат входом для reduce.

2 Алгоритм оценки информационно-технологического обеспечения безопасности информации

Для реализации алгоритма оценки информационно-технологического обеспечения безопасности информации [11] вначале необходимо ввести исходные данные. Далее m присваивается первоначальное значение 1, на 4-м шаге переменной $S_{\text{ф}}^{\text{рс}}$ первоначальное значение $S_{\text{ф}}^{\text{рс}} = 0$. Далее проверяется условие $m \leq M$ и вычисляется значение $S_{\text{ф}}^{\text{рс}} = S_{\text{ф}}^{\text{рс}} + p_m^c$. Значение m увеличивается на 1.

При невыполнении условия $m \leq M$ выполняется вычисление значения $W^c = S_{\text{ф}}^{\text{рс}} / S_{\text{т}}^{\text{рс}}$ и присваиваются первоначальные значения переменным q = 1, $S_{\text{ф}}^{\text{рк}} = 0$.

На 12 шаге проверяется условие $q \leq Q$.

При выполнении условия, $S_{\text{ф}}^{\text{рк}} = S_{\text{ф}}^{\text{рк}} + p_q^k$ (13 шаг), q=q+1 (14 шаг), и возврат на 12 шаг. При невыполнении условия $S_{\text{ф}}^{\text{рк}} = Q$.

Далее вычисляются $W^{\text{с}}$ и $W^{\text{о}}$ по формуле $W^{\text{к}} = S_{\text{ф}}^{\text{рк}} / S_{\text{т}}^{\text{рк}}$ и $W^{\text{о}} = (W^{\text{с}} + W^{\text{к}})/2$, после чего производится проверка условия s≤ S. При выполнении условия на шаге 20, $S_{\text{ф}}^{\text{нп}} = S_{\text{ф}}^{\text{нп}} + p_s^{\text{нп}}$ (21 шаг), s=s+1 (22 шаг) с последующим возвратом на 20 шаг. При невыполнении условия на шаге 20 значению присваивается $S_{\text{ф}}^{\text{нп}} = S$ (23 шаг).

На 24 шаге значение $W^{\text{с}}$ рассчитывается по формуле $W^{\text{нп}} = S_{\text{ф}}^{\text{нп}} / S_{\text{т}}^{\text{нп}}$. Присваиваем изначальное значение для n = 1 (25 шаг) и $S_{\text{ф}}^{\text{нд}} = 0$ (26 шаг).

Далее начинается проверка выполнения условия n≤ N (27 шаг). При выполнении условия, значение $S_{\text{ф}}^{\text{нп}}$ рассчитывается по формуле $S_{\text{ф}}^{\text{нп}} = S_{\text{ф}}^{\text{нп}} + p_s^{\text{нп}}$ (28 шаг), значение n увеличивается на единицу n=n+1 (29 шаг), после чего идет возврат на 27 шаг.

При выполнении условия на 27 шаге, значение $S_{\text{ф}}^{\text{нд}}$ рассчитывается по формуле $S_{\text{ф}}^{\text{нд}} = S_{\text{ф}}^{\text{нд}} + p_n^{\text{нп}}$ (28), а n увеличивается на единицу n=n+1 (29), после чего идет возврат на 27 шаг. При невыполнении условия на 27 шаге, $S_{\text{т}}^{\text{нд}} = N$, значение $W^{\text{нд}}$ рассчитывается по формуле $W^{\text{нд}} = S_{\text{ф}}^{\text{нд}} / S_{\text{т}}^{\text{нд}}$ (31 шаг), значение $W^{\text{п}}$ по формуле $W^{\text{п}} = (W^{\text{нд}} + W^{\text{нп}})/2$ (32 шаг). Изначальные значения h_{ri}=1 (33 шаг), $r_i = 1$ (34 шаг), $i = 1$ (35 шаг), $S_{irh}^{\text{фп}} = 0$ (36 шаг).

На следующем шаге производится проверка выполнения условия $h_{ri} \le H_{ri}$ (37 шаг). При выполнении условия, $S_{irh}^{\text{фп}} = S_{irh}^{\text{фп}} + p_{irh}^{\text{фп}}$, (38 шаг), h_{ri}=h_{ri} +1 (39 шаг), после чего идет возврат на 37 шаг. При невыполнении условия на 37 шаге $S_{irh}^{\text{тп}} = H$ri (40).

Далее рассчитывается значение $W^{\text{нд}}$ по формуле $W_{ir}^{\text{п}} = \frac{S_{irh}^{\text{фп}}}{p_{irh}^{\text{тп}}}$ (41 шаг). Значению M1 присваивается значение $W_{ir}^{\text{п}}$, (M1=[$W_{ir}^{\text{п}}$]), $r_i = r_i + 1$ (43 шаг). После чего начинается проверка выполнения условия $r_i \le R_i$(44 шаг), при выполнении условия происходит возврат на (37 шаг). В противоположном исходе, $W^{\text{исi}}$ рассчитывается по формуле $W^{\text{исi}} = \frac{\sum_r \sum_h p_{rh}^{\text{фпi}}}{\sum_r \sum_h p_{rh}^{\text{тпi}}}$ (45 шаг). Значение M2=[$W_i^{\text{ис}}$] (46 шаг). Значение i увеличивается на единицу (i=i+1).

В конце алгоритма производится проверка выполнения условия $i \le N$ (48 шаг). Значение $W^{\text{ис}}$ исчисляется по формуле $W^{\text{ис}} = \sum_i a_i * W_i^{\text{ис}}$ (49 шаг). Итоговое значение $W^{\text{СИБ}} = (b_1 * W^{\text{о}} + b_2 * W^{\text{п}} + b_3 * W^{\text{ис}})$ (50 шаг). На 51 шаге идет вывод результатов. 52 шаг – конец алгоритма.

Графическое отображение алгоритма представлено на рисунке 3.

Алгоритм состоит из 52 этапов, по выполнению которых мы получаем итоговое значение комплексного показателя информационной безопасности системы. Таким образом, алгоритм легко применим и для систем, где необходимо обрабатывать большие данные.

3 Архитектура системы оценки информационно-технологического обеспечения безопасности больших данных

Экспертная система реализуется в виде программного продукта. Последовательность проведения работ с использованием экспертной системы представлена на рисунке 4.

На первом этапе выполняется первичное заполнение базы данных, после чего идет определение и ввод коэффициентов весомости и характеристик информационной системы.

На втором этапе формируется и проводится контроль реализации требований, по-

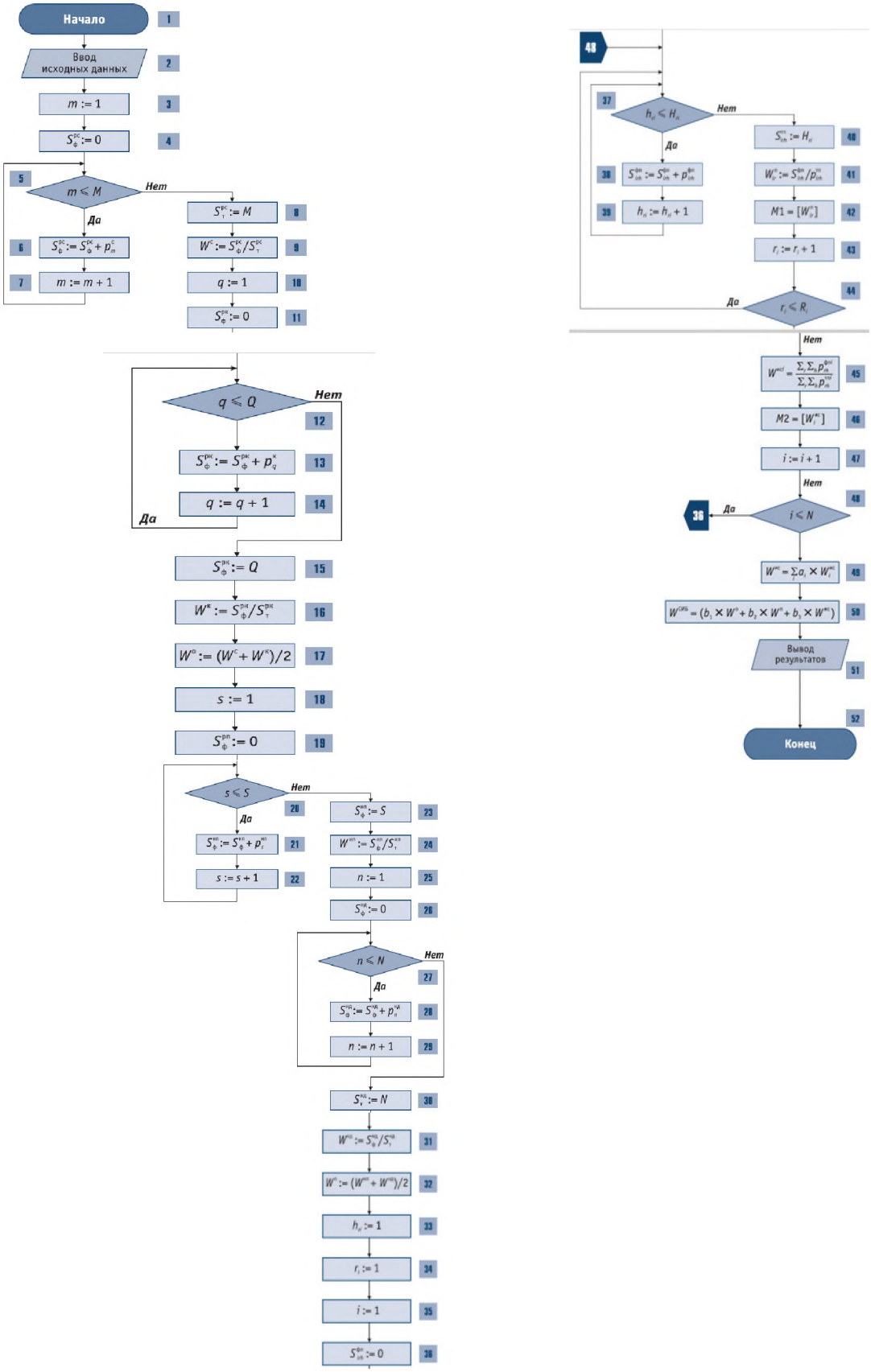

Рис. 3 – Алгоритм рассчета показателей эффективности

сле чего следует ввод результатов контроля в базу данных.

Третий этап – это расчет комплексных показателей оценки степени выполнения требований для каждой функциональной подсистемы.

Четвертый этап – оценка и анализ состояния системы обеспечения информационной безопасности с последующим завершением работы.

Экспериментальные исследования программы оценки состояния системы обеспечения информационной безопасности выполнялись с использованием принципов модульного тестирования. В ходе проведения экспериментальных исследований было определено, что алгоритм оценки степени выполнения требований обладает такими существенными свойствами как дискретность, детерминированность, конечность, результативность.

Расчет показателей безопасности системы будет осуществляться на основе формул из алгоритма расчета информационной безопасности, формируемый экспертной системой. Схема работы системы по расчету информационной безопаности системы с Big Data.

Экспериментальное исследование по проведению оценки эффективности защиты информации осуществляется путем выполнения следующих процедур:

этап 1 – подготовка исходных данных;

этап 2 – проведение контроля реализации требований;

этап 3 – расчет комплексных показателей оценки состояния системы защиты информации.

На этапе 1 осуществляется подготовка следующих исходных данных:

- перечень видов защищаемой информации;

- перечни требований к составу организационно-распорядительных документов (ОРД);

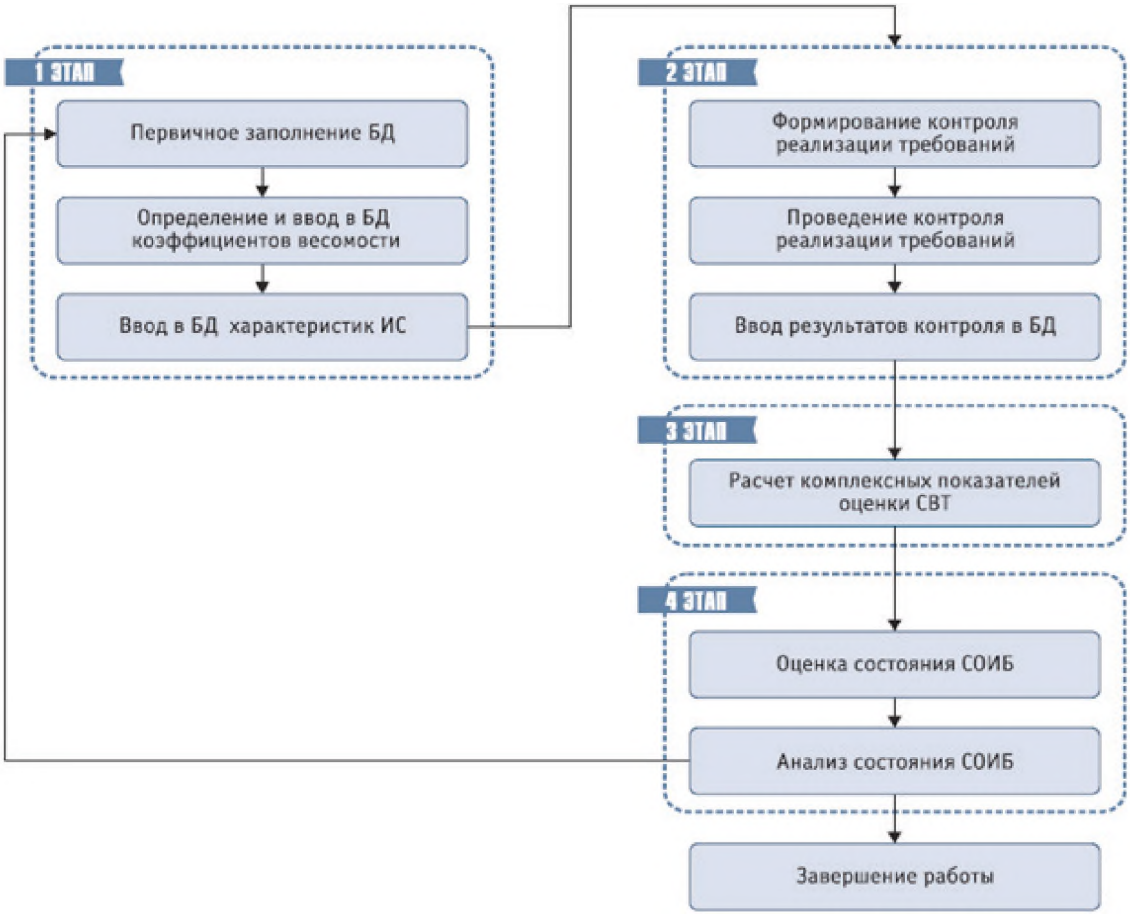

Рис. 4 – Последовательность выполнения работ с использованием экспертной системы

- перечни требований к подразделениям обеспечения безопасности информации и квалификации сотрудников данных подразделений;

- перечень типов информационных систем;

- перечень требований, предъявляемых к системе защиты информации информационных систем и СЗИ информационных систем персональных данных;

- классы защищенности информационных систем;

- типы информационных систем, функционирующих в организации;

- количество информационных систем каждого типа.

На этапе 2 осуществляются следующие мероприятия:

Шаг 1. Формирование анкет для проведения контроля реализации требований.

Шаг 2. Непосредственное проведение контроля реализации требований, осуществляемое методом анкетирования.

Данные требования определяются с учетом класса защищенности информационной системы и базового набора требований к системе защиты информации информационной системы. Количество анкет с требованиями к системам защиты информации информационных систем соответствует количеству информационных систем, функционирующих в организации.

Шаг 3. Контроль реализации требований выполняется членами комиссии, отвечающими за обеспечение информационной безопасности в организации. В ходе контроля проверяется выполнение требований по каждой анкете, и в графе «Единичные показатели, фактическое выполнение» проставляется «1» в случае выполнения требования и «0» в противном случае. На этапе 3 осуществляется расчет показателей эффективности защиты информации.

Заключение

Разработка алгоритма существенно осложняется, если разработчик не придерживается с самого начала некоторой дисциплины, позволяющей на каждом этапе разработки четко выделить необходимые подцели и прослеживать взаимосвязь между ними. Такой дисциплиной, получившей в последние годы широкое распространение, является метод пошаговой разработки. Предложенный алгоритм разрабатывается «по шагам», начиная с его спецификации, полученной в результате анализа задачи. На каждом этапе принимается небольшое число решений, приводящих к постепенной детализации управляющей и информационной структуры алгоритма. Таким образом, получается последовательность все более детальных спецификаций алгоритма, приближающихся к окончательной версии программы.

ЛИТЕРАТУРА

1. Большие данные. Революция, которая изменит то, как мы живем работаем и мыслим. Виктор Майер-Шенбергер и Кеннет Кукьер, 2010 г.
2. Корытникова Н.В. Online Big Data как источник аналитической информации в online-исследованиях // Социс. – 2015. – № 8. – С. 14-24.
3. Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики. Билл Фрэнкс, 2016 г.
4. Виктор Майер-Шенбергер и Кеннет Кукьер. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. 2013 г.
5. Отчет AAPOR о «Больших данных» / Л. Джапек, Ф. Крейтер, М.Берг [и др.] / Американская ассоциация исследователей общественного мнения, 2015 г.
6. Черняк Л. «Большие данные» – новая теория и практика // Открытые системы. – 2011. – № 10.

7. BigData проблемы, технологии, рынок: http://compress.ru/article.aspx?id=22725/. Дата обращения: 27.12.2018 года.
8. Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce»: https://habrahabr.ru/company/dca/blog/267361/. Дата обращения: 26.10.2017 г.
9. Big data от А до Я. Часть 3: Приемы и стратегии разработки MapReduce-приложений: https://habrahabr.ru/company/dca/blog/270453/. Дата обращения: 28.10.2017 года.
10. Толстова Ю.Н. Социология и компьютерные технологии // Социс. – 2015. – № 8. – С. 3-13.
11. Люльченко А. Н. Экспертная система оценки эффективности защиты информации // Защита информации. INSIDE. – 2016. – №4. – С. 20-24.