Применение кластерных вычислений для решения задачи многомерной рандомизации данных в риск-анализе

Шишкин В.М.[1], Савков С.В.[2]

[1] *Санкт-Петербургский институт информатики и автоматизации РАН, 14 линия ВО,39, Санкт-Петербург, Россия*
[2] *Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Кронверкский пр., 49, Санкт-Петербург, Россия*

vms@iias.spb.su, sergsavkov@gmail.com

Abstract. *The subjects of the research are the models and tools of risk analysis under incomplete and heterogeneous information. On substantive reasons, a priori, there is a requirement to obtain resulting indices in a stochastic form. In such conditions it is impossible to use the classic methods for operating with uncertain information which is supposed to be homogenous. The situation is complicated by the fact that the analyzed system is a complex causal structure which itself can be under uncertainty. The use of analytical methods for calculations becomes impossible, and the only method that provides the solution of the problem is a resource-intensive statistical modeling. The article describes different types of initial data setting, methods of homogenization and unification for further calculation and randomizing. A necessary technological part of the developed process is an HPC system which hasn't been used before for this purpose. There is also web-resource for solving risk analysis tasks in remote access mode.*

Ключевые слова

Оценка рисков, гетерогенная информация, многомерная рандомизация, ресурсоемкость, кластерные вычисления.

1 Введение

Современные информационные системы зачастую представляют собой сложные комплексы взаимосвязанных компонентов. Задача анализа безопасности, оценки рисков в таких системах усложняется тем, что эксперту неизвестны точные значения характеристик анализируемой системы. Большинство существующих методик предполагают задание приближенных точечных оценок, что снижает достоверность получаемых также точечных результирующих показателей.. Естественное желание повышения доверия к оценкам предполагает какую-либо вариативность результатов. Методически это непростая задача, требующая значительных вычислительных ресурсов.

Говоря об экспертном оценивании, мы сталкиваемся с различными видами неопределенности, и основной задачей является ее совместное моделирование. Аналогичная проблема возникает и при выборе наиболее эффективного комплекса средств противодействия угрозам информационной безопасности. Для ее решения используется методика рандомизации оценок факторов с последующим отбором результатов, удовлетворяющих исходным данным.

Наряду с неопределенностью оценок возможна и структурная неопределенность, то есть неполнота знаний о наличии или отсутствии отношений между факторами. С учетом структурной неопределенности, а также согласования мнения нескольких экспертов, задача моделирования факторов риска усложняется, а время рандомизации значительно возрастает. Приемлемое время расчета стохастического профиля рисков обеспечивается использованием высокопроизводительного кластерного вычислительного ресурса, позволяющего распараллелить алгоритм рандомизации исходных данных.

2 Обзор аналогов

В мире разработано и широко применяется не много алгоритмов и систем полного анализа информационных рисков [1]. Наиболее известные из них, например, британский CRAMM и американский RiskWatch, модификации которых распространены и в отечественной практике; в России — это ГРИФ и АванГард.

Метод CRAMM (CCTA Risk Analysis and Managment Method) был разработан Агентством по компьютерам и телекоммуникациям Великобритании (Central Computer and Telecommunications Agency) и взят на вооружение в качестве государственного стандарта. Преимуществом данной системы является то обстоятельство, что она изначально ориентировалась на поддержание стандартов, которые стали основой стандартов ISO, принятых позднее почти без изменений уже в качестве Российских стандартов.

Программное обеспечение RiskWatch является довольно мощным средством анализа и управления рисками. В семейство RiskWatch входят программные продукты для проведения различных видов аудита безопасности. В отличие от CRAMM, программа RiskWatch более ориентирована на количественную оценку соотношения потерь от угроз безопасности и затрат на создание системы защиты. Надо также отметить, что в этом продукте риски в сфере информационной и физической безопасности компьютерной сети предприятия рассматриваются совместно.

Комплексная экспертная система (КЭС) «АванГард позволяет построить структурную модель ИС, модель угроз и модель событий рисков, связанных с отдельными составляющими ИС и, таким образом, выявить те сегменты и объекты, риск нарушения безопасности которых является неприемлемым, то есть критическим. Помимо этого, он позволяет построить модель защиты — систему мер и требований, которые должны выполняться, чтобы обеспечить безопасность ИС, а также выработать комплекс мероприятий по защите. Комплекс «АванГард-Контроль» позволяет проводить мониторинг-контроль выполнения требований по защите критических сегментов ИС и определять «узкие» места в защите и обеспечении безопасности ИС.

Перечисленные системы и комплексы обладают рядом существенных недостатков: не учитывается неопределенность задания исходных данных, а также результаты расчета представляются в точечном виде, что снижает доверие к результатам анализа. Преодоление указанных недостатков являлось мотивом для выполнения представляемой работы.

3 Алгоритм оценки факторов риска

В структурной метамодели, положенной в основу разрабатываемой программной системы [2], определяется три категории риск-факторов: субъекты, объекты и воздействия первых на вторые. Соответственно категориям выделяются три непересекающихся непустых подмножества множества $M_0 = M_s \cup M_e \cup M_c$ элементов модели:

независимые активные субъекты, «источники угроз» - множество M_s;

проводники воздействий, события, порождаемые источниками угроз, «угрозы» нарушения безопасности - множество M_e, в котором выделяется подмножество так называемых «событий риска» - угроз, наносящих непосредственно ущерб объекту;

«компоненты» объекта - множество M_c.

На множестве M_0 определяется хотя бы один тип отношений: бинарное отношение причинности R со свойством транзитивности, к которому можно свести многие связи, имеющие импликативный характер. R упорядочивает M_0 и задает на нем структуру, фиксирующую каналы распространения потоков угроз от источников до объекта, и порождает квадратную матрицу отношений $\mathbf{W}_0$.

Цель реализует «система защиты информации» - СЗИ, которая представима в виде множества элементов S, каждый из которых осуществляет воздействие на элементы из M_0. Между элементами множеств S и M_0 устанавливается отношение, формально сводимое к R, порождающее прямоугольную матрицу отношений $\mathbf{R}_0$.

Простейшая количественная интерпретация метамодели отображает $\mathbf{W}_0$ в арифметическую матрицу $\mathbf{W} = (w_{ij})$, элементы которой можно рассматривать как весовые коэффициенты, имеющие смысл меры влияния i-го элемента на j-ый. Она содержит все исходные данные для расчетов на модели, полученные тем или иным способом.

Далее рассчитываются показатели v_{ij}, аналогичные по смыслу w_{ij}, но уже с учетом транзитивности отношений. В результате определяется матрица $\mathbf{V}$, структурно эквивалентная $\mathbf{W}$. При отсутствии рефлексии элементов, если $\mathbf{W}$ считать взвешенной матрицей смежности некоторого графа, они легко рассчитываются на графе в соответствующих терминах, как суммы по всем путям из i-ой в j-ую вершину произведений оценок дуг каждого пути, что равносильно матричному преобразованию $\mathbf{V} = (\mathbf{I} - \mathbf{W})^{-1} - \mathbf{I}$.

Последний, всегда ненулевой, z-ый столбец $\mathbf{v}_z$ матрицы $\mathbf{V}$ содержит искомые показатели $\{v_{iz}\}$ влияния любого i-го фактора риска на объект, представляя профиль риска. Эти показатели должны целенаправленно ориентировать создание системы защиты информации на противодействие наиболее значимым факторам риска. Процесс преобразования данных при построении профиля риска можно схематически показать на рисунке 1. Рассмотрим каждый этап процесса более подробно.

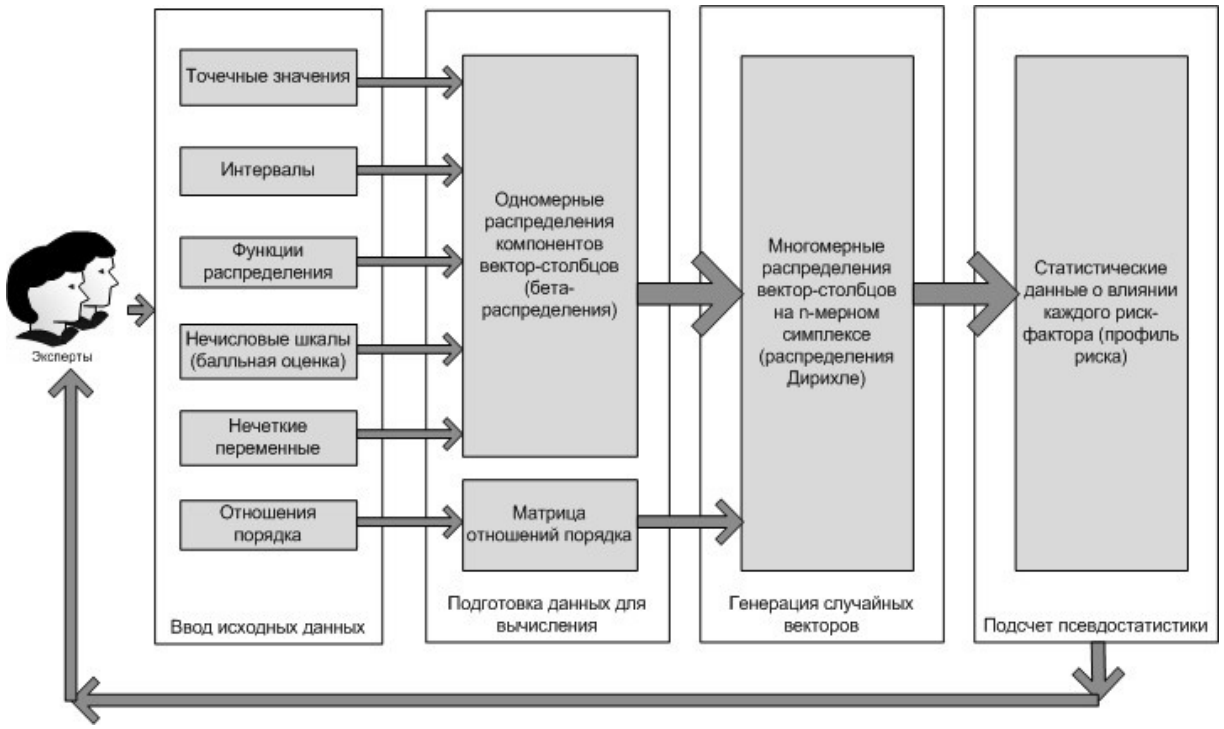

Рис. 1. Преобразование данных при построении профиля риска

4 Способы задания исходной информации

Информация о каждой оценке w_{ij} для арифметизации матрицы $\mathbf{W}_0$ может быть представлена отдельным значением (точка), диапазоном значений (интервал), некоторым распределением вероятностей значений, нечеткой величиной, иной величиной в нечисловых шкалах.

В общем случае, при интервальном оценивании уместно использовать бета-распределение, частным случаем которого является равномерное [3]. Бета-распределение представляет собой двухпараметрическое семейство непрерывных распределений, плотность вероятности которого имеет вид:

$$f(w) = \frac{1}{B(\alpha, \beta)} w^{\alpha-1}(1-w)^{\beta-1}$$

где $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$ - бета-функция.

При $\alpha = \beta = 1$, $f_{ij}(w)$ обращается в стандартное равномерное распределение, которое используется, если изначально известны только границы интервала, на котором распределена случайная величина (таблица 1).

В случае, когда в качестве распределения с максимальной энтропией для какого-либо фактора выбирается нормальное с параметрами *(m, s2)*, его можно также аппроксимировать бета-распределением, рассчитав соответствующие коэффициенты (α, β). Для аппроксимации приравняем моменты соответствующих распределений:

$$\frac{\alpha}{\alpha+\beta} = m\,,\ \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = s^2\,,$$

откуда получаем выражения для коэффициентов:

$$\alpha = \left(\frac{m(1-m)}{s^2} - 1\right)m$$

$$\beta = \left(\frac{m(1-m)}{s^2} - 1\right)(1-m)$$

Графически пример аппроксимации показан на рисунке 2.

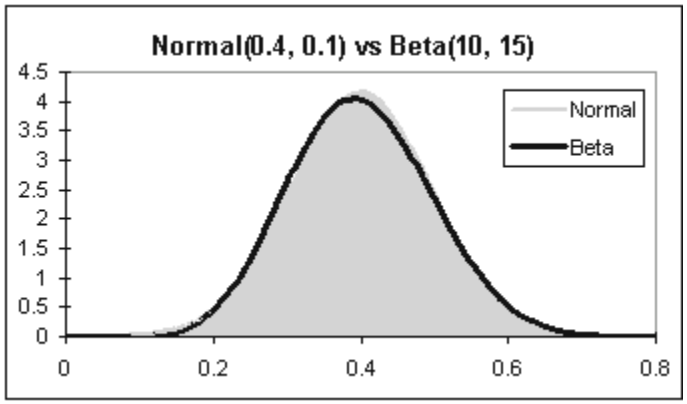

Рис. 2. Аппроксимация нормального распределения бета-распределением

Зачастую информация о факторе может быть представлена в терминах нечетких переменных. В таком случае вместо функции плотности распределения можно рассматривать функцию принадлежности, которая принимает значения $0 \le \mu(x) \le 1$ и характеризует степень принадлежности каждого члена пространства рассуждения данному нечеткому множеству. В качестве функции принадлежности могут быть выбраны абсолютно любые функции, однако существуют стандартные классы функций принадлежности, используемые для решения большинства задач описания нечетких переменных. К таким классам относят:

- кусочно-линейные функции принадлежности;
- S и Z-бразные функции принадлежности;
- П-образные функции принадлежности.

Как показано в [4], нечеткие множества правомерно рассматривать как проекции случайных множеств, распределение случайного множества с независимыми элементами полностью определяется его проекцией, и там же дан метод сведения нечетких множеств к случайным. Рассмотрение отношений, связывающих два и более коэффициента w_{ij} для любого индекса j на множестве всех индексов i, ограничим отношением порядка (ординальным отношением). Его для нашего случая можно определить следующим образом. Обозначим $w_1 = w_{i_1 j}$ и $w_2 = w_{i_2 j}$ - пару весовых коэффициентов *j*-го столбца, для которых задано бинарное отношение порядка. Тогда это отношение будет определять угловая координата $\alpha = arctg(w_2 / w_1)$.

При задании нечеткого отношения определяется функция принадлежности для параметра α. Тогда при последующей генерации векторов в процессе рандомизации плотность их распределения окажется неравномерной, и будет зависеть от заданной функции распределения (рисунок 3).

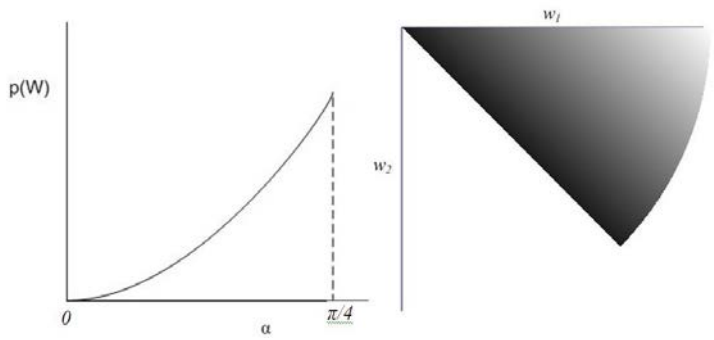

Рис. 3. Графическая иллюстрация нечеткого отношения порядка

Все отмеченное разнообразие представления исходной информации сводимо к распределениям вероятностей значений w_{ij}, и информация может быть в этом смысле гомогенизирована. В результате происходит идентификация w_{ij} уже как случайных величин $\tilde{w}_{ij}$, связанных в допустимых реализациях условием $\sum_i w_{ij}^* = 1$. Первоначально [3] неявная гомогенизации и последующая рандомизация векторов $\mathbf{W}_j$ выполнялась путем независимой генерации случайных значений $\tilde{w}_{ij}$ по каждой из координат с минимальным преобразованием исходных данных и без учета их зависимости, то есть на единичном гиперкубе.

Из полученных на гиперкубе композиций $< w_{i_1 j}, ..., w_{i_k j} >$ затем происходил отбор тех из них, что принадлежат симплексу Sj, и дополнительно, уже на симплексе, окончательных реализаций случайных векторов $\tilde{\mathbf{w}}_j$, допустимых, согласно ординальным и корректирующим ограничениям. В результате этих действий для всех j получалась реализация случайной матрицы $\tilde{\mathbf{W}}$, на которой выполнялись операции как на обычной числовой матрице $\mathbf{W}$ согласно базовому алгоритму преобразования $\mathbf{W} \to \mathbf{V}$. Заданное количество реализаций $\tilde{\mathbf{W}}$ обеспечивало рандомизацию $\mathbf{V}$ с получением в итоге на множестве реализаций $\tilde{\mathbf{V}}$ стохастических оценок показателей для принятия решений.

Ясно, что этот процесс, хотя в нем наиболее полно и точно отражается вся заданная экспертом информация, крайне ресурсоемкий. Поэтому в настоящее время, с учетом всех обстоятельств выбран иной путь, и сочтено целесообразным гомогенизировать информацию явным образом на основе бета-распределения, унифицируя стандартным его видом. Тогда генерацию векторов $\tilde{\mathbf{w}}_j$ становится возможным выполнять через распределение Дирихле непосредственно на симплексе, что резко снижает ресурсоемкость без существенной потери качества.

Во всех описанных случаях мы считали, что структура взаимосвязей между факторами этой системы задана определенным фиксированным образом, а неопределенность свойственна только информации об оценках. Но при решении практических задач часто невозможно с уверенностью утверждать о наличии той или иной связи между факторами. Здесь мы сталкиваемся с проблемой структурной неопределенности, когда в распоряжении аналитика имеется только перечень факторов, но недостаточно информация об их взаимосвязях. Эта задача пока не реализована. Описанная проблема может быть решена разными способами, в частности рандомизацией матриц $\mathbf{V}_0$, с помощью аппарата нечетких графов.

5 Алгоритм рандомизации исходных данных

Алгоритм получения результирующего множества допустимых векторов состоит из нескольких последовательных этапов:

1. Генерация случайного вектора на симплексе с заданным распределением;
2. Отбор векторов, удовлетворяющих исходной информации о каждом факторе;
3. Отбор векторов, удовлетворяющих исходной информации об отношениях.

Изначально планировалось использование равномерного генератора на гиперкубе с последующим отбором векторов, принадлежащих окрестностям симплекса, но с ростом числа факторов n время генерации возрастало

в факториальной зависимости. Поэтому, алгоритм был модифицирован таким образом, что генерация производилась непосредственно на симплексе.

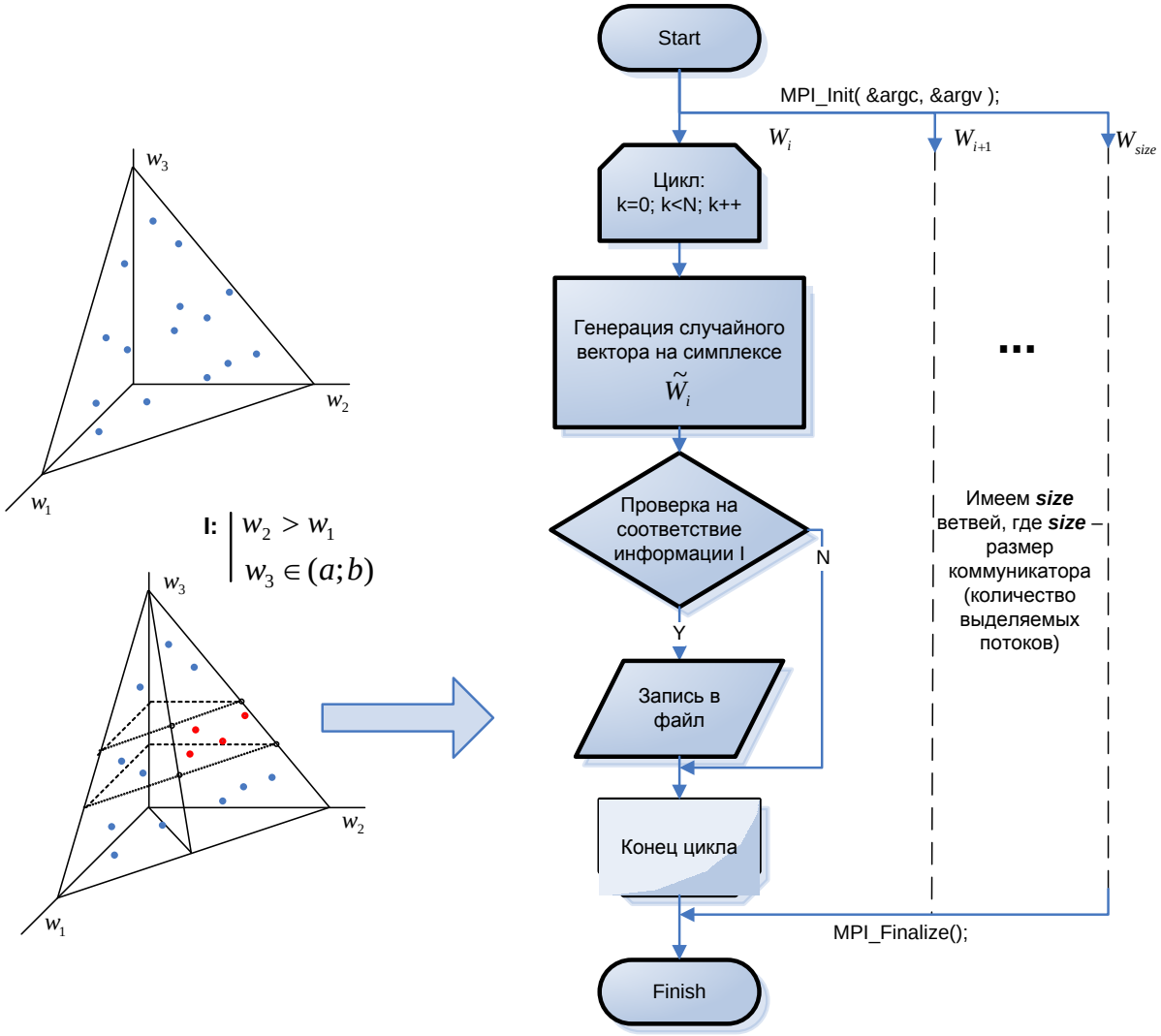

Рис. 4. Функциональная схема алгоритма рандомизации

На рисунке 4 *i*-я ветвь алгоритма, порожденная функцией MPI_Init() генерирует множество допустимых векторов, для столбца $\mathbf{W}_i$ матрицы $\mathbf{W}$. Каждая ветвь содержит цикл из k итераций, где k определяет количество генерируемых векторов. С увеличением k возрастает точность получаемого результата. На каждой итерации генерируется случайный вектор $\tilde{\mathbf{W}}_i$, имеющий равномерное распределение на симплексе S в m-мерном пространстве. Далее, вектор $\tilde{\mathbf{W}}_i$ проверяется на соответствие исходной информации I [5].

При таком подходе время вычисления будет зависеть от того, какая часть сгенерированного множества векторов будет отсеяна в результате отбора. Алгоритм завершает работу, как только будет получен нужный объем выборки. Следовательно, в случаях, когда значения отдельных факторов будут ограничены узкими интервалами, большая часть генераций будет отброшена. Для уменьшения времени работы алгоритма был принят другой способ, использующий распределение, уже учитывающее исходную информацию об отдельных факторах, а именно, распределение Дирихле:

$$f(x_1,...,x_{n-1};\alpha_1,...,\alpha_n) = \frac{1}{B(\alpha)}\prod_{k=1}^{n} x_k^{\alpha_k-1}$$, где x_k представляют собой в нашем случае w_{ij} для всех i.

Выбор распределения Дирихле обуславливается тем, что оно является обобщением бета-распределения на многомерный случай и позволяет проводить генерацию непосредственно на симплексе. Коэффициенты распределения Дирихле выбираются на том основании, что его частными одномерными распределениями являются бета-распределения, а, следовательно, например величина x_1 будет распределена по бета-закону с

коэффициентами $(\alpha_1, \sum_{k=2}^{N} \alpha_k)$. Это справедливо для любого x_k, что позволяет найти необходимые коэффициенты непосредственно из параметров исходных бета-распределений.

На последнем этапе отбираются вектора согласно заданным четким и нечетким отношениям порядка. Каждый отобранный вектор записывается в файл результатов.

В результате серии генераций получаем множество случайных реализаций арифметизированной матрицы факторов риска W с нормированными по столбцам весовыми коэффициентами, на каждой из которых по определенному алгоритму с учетом транзитивности отношения выполняется расчет реализации детерминированного профиля риска. В совокупности множество этих реализаций предоставляет псевдостатистику для идентификации параметров стохастического профиля риска анализируемого объекта.

6 Технология и архитектура

Программная система, схематически изображенная на рисунке 5 находится в стадии разработки и базируется на аппаратной платформе СПИИРАН. Она включает в себя клиентскую часть, представленную веб-интерфейсом, сервер приложений, сервер баз данных и высокопроиводительный вычислитель, который позволяет вести расчеты в параллельном режиме, что снимает ограничение на размерность моделей, а также значительно ускоряет получение результатов вычисления [6].

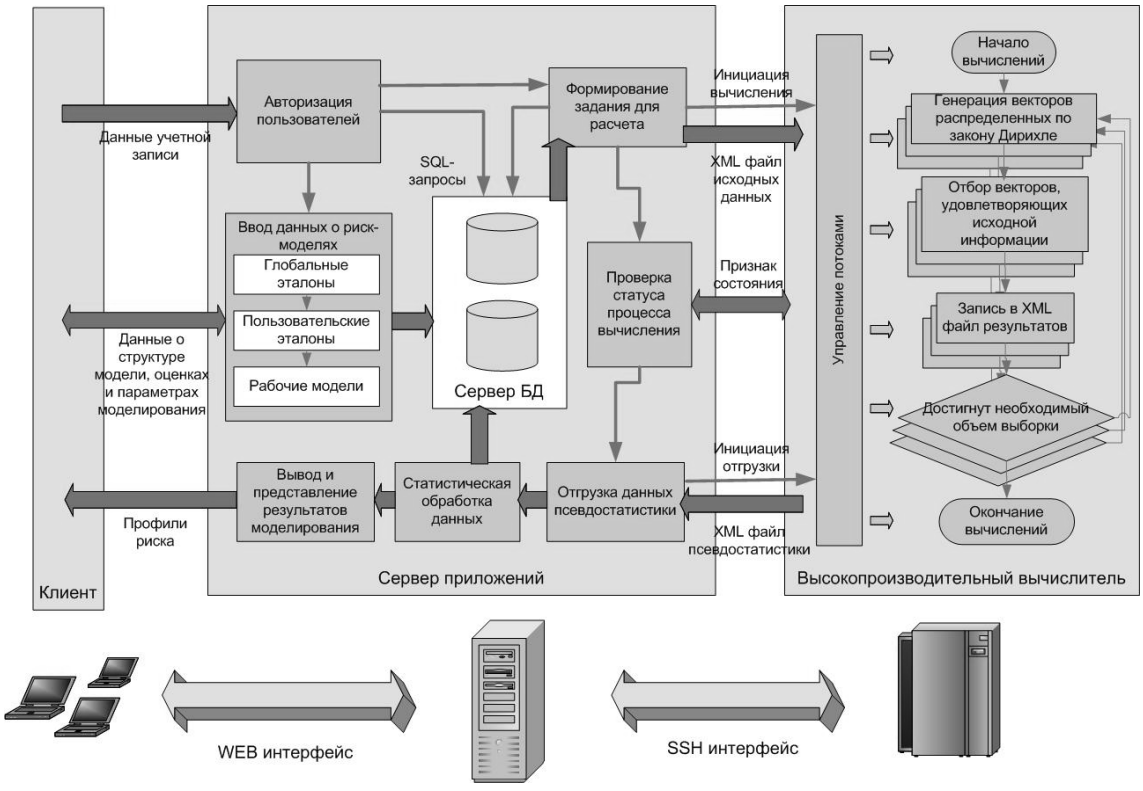

Рис. 5. Архитектура комплекса оценки рисков

Клиентская часть программы реализована посредством веб-интерфейса, она обеспечивает пользователю возможность ввода, редактирования и отображения данных в различных режимах и с разными правами доступа к моделям. В соответствии с иерархическим принципом выделяются эталонные и пользовательские риск-модели. Эталонные модели доступны для чтения любому пользователю. Редактирование эталонных моделей не допускается. Эталонные модели могут использоваться в качестве шаблонов при подготовке пользовательских риск-моделей.

Каждая модель включает набор факторов, отношения между ними, а также различную информацию об этих отношениях. Пользователь на этапе ввода модели имеет право выбрать наиболее удобный для него способ задания информации, при этом осуществляется контроль ее непротиворечивости. Информация о модели

заносится в БД, так что пользователь (группа) в любой момент может возобновить редактирование своей риск-модели.

Сервер приложений исполняет функции посредника между клиентом, БД и вычислителем. На него возложены задачи по гомогенизации данных и подготовке к их отправке на вычислительный ресурс, инициация вычисления и контроль его состояния, загрузка результатов моделирования и их статистическая обработка, а также представление в удобном для пользователя виде. Обмен данных между кластером и сервером приложений осуществляется через текстовые файлы XML формата. При этом на кластер передается информация о параметрах бета-распределения для каждого из заданных отношений, а с кластера загружается файл, содержащий выборку сгенерированных векторов, удовлетворяющих этой информации. Контроль состояния вычисления производится путем периодического опроса кластера не предмет завершения расчетов. Таким образом, во взаимодействии между сервером приложений и кластером все запросы инициируются только сервером приложений и выполняются кластером в режиме пакетной обработки данных и осуществляются по протоколам SSH/SCP.

7 Заключение

Использование кластерных вычислений для целей, подобных рассмотренной нами, может показаться неожиданным. Действительно, существующие системы анализа рисков, даже наиболее громоздкие, CRAMM, например, не требуют значительных вычислительных мощностей. Более привычно и уже традиционно их применение в тех предметных областях, где возникают задачи с высокой вычислительной сложностью, многомерные и ресурсоёмкие. Такие задачи в области гидрометеорологии, океанологии, теории чисел [7] также решаются и на кластере СПИИРАН, который служит технологическим ресурсом для создаваемой системы.

Необходимость параллельных вычислений в нашей задаче первоначально возникла, когда в результате экспериментов на макете системы риск-анализа и оценочных расчётов ресурсоёмкости выяснилось, что применявшаяся нами тогда методика рандомизации, хотя и наиболее точно позволяла отобразить исходную гетерогенную информацию, в реализации приводила к факториальному росту времени вычислений и даже на структурных моделях средней размерности становилась практически неприемлемой.

Последующее развитие методики позволило радикально сократить объём «холостых» вычислений, но одновременно сняло ограничения на размерность и открыло качественно новые возможности для усложнения самой модели риск-анализа, допустив, в частности, не только оценочную, но и структурную неопределённость, что резко увеличит объем вычислений. Поэтому потребность в использовании высокопроизводительного вычислителя не исчезнет, и уже очевидно, хотя оценка ресурсоёмкости пока не проводилась, что придётся решать задачу оптимизации загрузки процессоров

Литература

[1] Савков С.В., Шишкин В.М. Сравнительный анализ методик комплексного оценивания рисков в информационных системах *Региональная информатика (РИ-2010) / XII Санкт–Петербургская Международная конференция. Санкт-Петербург, 20–22 октября 2010 г.: Труды конференции* / СПОИСУ. — СПб. 2010. — С. 139.

[2] Шишкин В.М. Метамодель анализа, оценки и управления безопасностью информационных систем // *Проблемы управления информационной безопасностью: Сборник трудов ИСА РАН / Под ред. Д.С.Черешкина*. — М.: Едиториал УРСС, 2002. – С. 92-105.

[3] Шишкин В.М., Савков С.В. Методика арифметизации неполной гетерогенной исходной информации для идентификации профиля рисков // *Моделирование и анализ безопасности и риска в сложных системах: Труды Междунар. научн. школы МА БР–2010* / — СПб.: ГУАП, 2010. — С. 295-300.

[4] Орлов А.И. Нечисловая статистика *М.: МЗ-Пресс*, 2004. – 513 с

[5] Савков С.В., Шишкин В.М. Разработка системы интервального оценивания информационных рисков *Научно-технический журнал "Приборостроение", №9*, 2011 г., С. 38-44

[6] Шишкин В.М., Савков С.В. Веб-ориентированный комплекс оценки рисков в сложных системах // *Труды конгресса по интеллектуальным системам и информационным технологиям «IS&IT'12». Научное издание в 4-х томах. – М.: Физматлит, 2012*. - Т. 2. – С. 39-43.

[7] Yuri Matiyasevich. New Conjectures about Zeroes of Riemann's Zeta Function ./ *University of Leicester Department of Mathematics. Research report MA12-03*. [Электронный ресурс] - http://www.newton.ac.uk/preprints/NI12088.pdf