МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

СОГЛАСОВАНО
Генеральный директор
ЗАО «АйТи»

_____ Бакиев О.Р.
"29" декабря 2011 г.

УТВЕРЖДАЮ
Ректор НИУ ИТМО

_____ Васильев В.Н.
"29" декабря 2011 г.

МНОГОПРОФИЛЬНАЯ ИНСТРУМЕНТАЛЬНО-ТЕХНОЛОГИЧЕСКАЯ ПЛАТФОРМА СОЗДАНИЯ И УПРАВЛЕНИЯ РАСПРЕДЕЛЕННОЙ СРЕДОЙ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ CLAVIRE

ТЕХНОЛОГИЧЕСКАЯ ПЛАТФОРМА ПОТОКОВОЙ ОБРАБОТКИ СВЕРХБОЛЬШИХ ОБЪЕМОВ ДАННЫХ И ИЗВЛЕЧЕНИЯ ИЗ НИХ ЗНАНИЙ НА ОСНОВЕ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ (МИТП-Д)

Описание применения

ЛИСТ УТВЕРЖДЕНИЯ

RU.СНАБ.80066-06 31 06-ЛУ

Представители
Организации-разработчика

Руководитель разработки,
профессор НИУ ИТМО

_____ Бухановский А.В.
"29" декабря 2011 г.

Ответственный исполнитель,
с.н.с. НИУ ИТМО

_____ Луценко А.Е.
"29" декабря 2011 г.

Нормоконтролер
ведущий инженер НИУ ИТМО

_____ Позднякова Л.Г.
"29" декабря 2011 г.

2011

Инв.№ подл.	Подп. и дата	Взам.инв.№	Инв.№ дубл.	Подп. и дата

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

УТВЕРЖДЕН
RU.СНАБ.80066-06 31 06-ЛУ

МНОГОПРОФИЛЬНАЯ ИНСТРУМЕНТАЛЬНО-ТЕХНОЛОГИЧЕСКАЯ ПЛАТФОРМА СОЗДАНИЯ И УПРАВЛЕНИЯ РАСПРЕДЕЛЕННОЙ СРЕДОЙ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ CLAVIRE

ТЕХНОЛОГИЧЕСКАЯ ПЛАТФОРМА ПОТОКОВОЙ ОБРАБОТКИ СВЕРХБОЛЬШИХ ОБЪЕМОВ ДАННЫХ И ИЗВЛЕЧЕНИЯ ИЗ НИХ ЗНАНИЙ НА ОСНОВЕ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ (МИТП-Д)

Описание применения

RU.СНАБ.80066-06 31 06

ЛИСТОВ 33

2011

Инв.№ подл.	Подп. и дата	Взам.инв.№	Инв.№ дубл.	Подп. и дата

АННОТАЦИЯ

Документ содержит описание применения технологической платформы потоковой обработки сверхбольших объемов данных и извлечения из них знаний на основе облачных вычислений (МИТП-Д) RU.СНАБ.80066-06 01 45. Технологическая платформа МИТП-Д входит в состав многопрофильной инструментально-технологической среды CLAVIRE (Cloud Applications Virtual Environment) RU.СНАБ.80066-06. Она предназначена для поддержки разработки и исполнения композитных приложений сбора и обработки данных из распределенных источников, объединенных сетями общего назначения (Интернет) на основе единых стандартов взаимодействия, в целях решения специфических задач (например, анализа медиаконтента в социальных сетях). МИТП-Д разработана в ходе выполнения проекта «Создание распределенной вычислительной среды на базе облачной архитектуры для построения и эксплуатации высокопроизводительных композитных приложений» (Договор № 21057 от 15 июля 2010 г., шифр 2010-218-01-209) в рамках реализации постановления Правительства РФ № 218 «О мерах государственной поддержки развития кооперации российских высших учебных заведений и организаций, реализующих комплексные проекты по созданию высокотехнологичного производства».

СОДЕРЖАНИЕ

1. НАЗНАЧЕНИЕ ПРОГРАММЫ 4
1.1. Функциональное назначение....4
1.2. Область применения 4
1.3. Основные характеристики....5
1.4. Ограничения, накладываемые на область применения 6
2. УСЛОВИЯ ПРИМЕНЕНИЯ....7
2.1. Условия развертывания программы 7
2.2 Необходимые технические средства управляющей подсистемы МИТП-Д 9
2.3. Необходимые технические средства подсистемы вычислительной инфраструктуры МИТП-Д 10
3. ОПИСАНИЕ ЗАДАЧИ....12
3.1 Определение задачи 12
3.2. Методы решения задачи 14
3.2.1. Унифицированное описание прикладных пакетов в МИТП-Д....15
3.2.2. Унифицированное описание композитных приложений в МИТП-Д....18
3.2.3. Организация процесса исполнения композитного приложения в МИТП-Д 20
3.2.4. Организация сбора и анализа данных в социальных сетях в Интернете 23
3.2.5. Особенности создания и управления потоковой обработки сверхбольших объемов данных и извлечения из них знаний на основе облачных вычислений 25
3.2.6. Решение типовой задачи сбора и обработки данных с использованием МИТП-Д 27
4. ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ 30
ПЕРЕЧЕНЬ СОКРАЩЕНИЙ 32
ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ....33

1. НАЗНАЧЕНИЕ ПРОГРАММЫ

Технологическая платформа потоковой обработки сверхбольших объемов данных и извлечения из них знаний на основе облачных вычислений (МИТП-Д) RU.СНАБ.80066-06 01 45 входит в состав многопрофильной инструментально-технологической среды CLAVIRE (Cloud Applications Virtual Environment) RU.СНАБ.80066-06. Она предназначена для поддержки разработки и исполнения композитных приложений сбора и обработки данных из распределенных источников, объединенных сетями общего назначения (Интернет) на основе единых стандартов взаимодействия, в целях решения специфических задач (например, анализа медиаконтента в социальных сетях).

1.1. Функциональное назначение

МИТП-Д представляет собой комплекс программного обеспечения для разработки, настройки и эксплуатации сред распределенных вычислений и обработки больших объемов данных, предназначенный для:

1) эффективного управления вычислительными, информационными и программными ресурсами среды распределенного сбора и обработки данных, включая собственные (выделенные) вычислительные ресурсы центров и ресурсы внешних провайдеров (в том числе глобальных коллаборативных сред);
2) создания, исполнения, управления и предоставления сервисов доступа к предметно-ориентированным высокопроизводительным композитным приложениям для сбора и обработки данных в распределенных источниках;
3) обеспечения функционирования программно-аппаратного комплекса поддержки инфраструктуры облачных вычислений для сбора и обработки больших объемов данных.

1.2. Область применения

Технологическая платформа МИТП-Д предназначена для создания программной инфраструктуры центров компетенции, ориентированных на сбор и обработку большого объема данных. Она выступает как надстройка над низкоуровневыми распределенными средами обработки данных и позволяет обеспечить унифицированный интерфейс взаимодействия с пользователем при организации сбора и использования (в расчетах, для инициализации моделей, для визуализации) данных из разнородных источников. МИТП-Д применима для решения следующих классов задач.

1. Сбор и потоковая обработка данных в регулярно обновляемых распределенных хранилищах с целью извлечения из них знаний.
2. Сбор данных и мониторинг состояния объектов в социальных сетях в Интернете (включая анализ текстовых и мультимедийных данных).
3. Сбор, анализ и усвоение (assimilation) данных в вычислительных моделях, функционирующих в оперативном режиме (системы слежения или прогнозирования).

1.3. Основные характеристики

Технологическая платформа МИТП-Д реализована на основе МИТП CLAVIRE в рамках концепции iPSE (Intelligent Problem Solving Environment). Она ориентирована на развитие интеллектуальных технологий поддержки жизненного цикла проблемно-ориентированных сред распределенных вычислений, обеспечивающих сбор, обработку и анализ больших объемов данных в распределенных источниках.

МИТП-Д обеспечивает совместное решение двух задач – унификации средств доступа к различным источникам данных (как на формальном, так и на содержательном уровне) и обеспечения их «бесшовного» использования в качестве параметров и входных данных для вычислительных моделей, реализуемых прикладными сервисами МИТП-Д. В целом МИТП-Д поддерживает две категории прикладных сервисов – сбора и обработки данных, а также моделирования. Сервисы сбора и обработки функционируют на собственных ресурсах МИТП-Д; они используют для своей работы низкоуровневую среду распределенной обработки данных Hadoop. Сервисы моделирования в МИТП-Д могут строиться на основе различных пакетов, в зависимости от решаемых задач. Для их функционирования используются как собственные (корпоративные) ресурсы центров компетенции в области обработки данных, так и привлекаемые ресурсы внешних провайдеров (выделенные суперкомпьютеры, Грид-среды, среды облачных вычислений первого поколения).

МИТП-Д обеспечивает выполнение следующих функций:

1) Поддержка разработки и исполнения композитных приложений сбора и обработки данных из распределенных источников, объединенных сетями общего назначения (Интернет) на основе единых стандартов взаимодействия, в целях решения специфических задач.
2) Возможность развертывания в существующих коллаборативных распределенных средах, включая существующие инфраструктуры Грид I поколения.

3) Динамическое управление (мониторинг состояния, запуск приложений, передача данных, распределение нагрузки, миграция данных и задач) в автоматическом режиме набором приложений для сбора и обработки данных из распределенных источников.
4) Автоматическая оптимизация по времени процесса использования доступных вычислительных ресурсов и прикладных сервисов для сбора и обработки данных из распределенных источников.
5) Представление описания композитных приложений для сбора и обработки данных из распределенных источников на основе цепочек заданий (workflow), обеспечивающих запуск, выполнение, остановку и возобновление работы цепочки заданий в ручном и автоматическом режимах.
6) Поддержка процесса установки и первоначальной конфигурации технологической платформы для сбора и обработки данных из распределенных источников, и ее составных частей на ресурсах коллаборативной распределенной среды.
7) Поддержку многопользовательского режима при решении задач сбора и обработки данных из распределенных источников.
8) Квотирование, биллинг и тарификация использования данных из распределенных источников и вычислительных сервисов их обработки.
9) Каталогизация входных данных пользователей на основе метаданных.
10) Администрирование и контроль работы с дифференцированными правами администраторов в рамках многоуровневой политики доступа к ресурсам распределенного хранения данных.
11) Модификация знаний, используемых системой, как в ручном, так и в автоматическом режимах.
12) Функционирование сервисов резервирования и отката исправлений для результатов работы вычислительных сервисов обработки данных в удаленном хранилище в составе распределенной среды хранения данных.
13) Функционирование механизмов конвертирования данных между различными прикладными сервисами обработки данных, по заданию пользователя.

1.4. Ограничения, накладываемые на область применения

Специфика решения задачи сбора и обработки данных накладывает на применение МИТП-Д следующие ограничения:

1) в состав МИТП-Д не входят прикладные пакеты для моделирования и обработки данных; они регистрируются в МИТП-Д исходя из специфики деятельности конкретного центра обработки данных;
2) прикладные сервисы для сбора и первичной обработки данных связаны с используемыми источниками данных; МИТП-Д не предусматривает поддержку каких-либо видов источников данных на системном уровне (для любого источника данных необходима разработка приложения доступа к нему);
3) прикладные сервисы сбора и первичной обработки данных в социальных сетях в Интернете, входящие в состав самой платформы МИТП-Д, функционируют на основе платформы Hadoop, развернутой на ресурсах ПАК. Допускается использование других технологий работы с данными в рамках развития и внедрения в МИТП-Д новых прикладных сервисов.

2. УСЛОВИЯ ПРИМЕНЕНИЯ

2.1. Условия развертывания программы

Установка и настройка системных компонентов МИТП-Д производится посредством компонента развертывания и конфигурирования RU.СНАБ.80066-06 01 36. Данный компонент предоставляет графический интерфейс для решения следующих задач:

1) полуавтоматическое развертывание компонентов МИТП-Д;
2) конфигурирование компонентов МИТП RU.СНАБ.80066-06 01 01 при подготовке и настройке технологической платформы МИТП-Д, а также во время эксплуатации;
3) автоматизированная проверка корректности развертывания компонентов за счет выполнения тестов для развернутых системных сервисов МИТП-Д.

Развертывание МИТП-Д производится из установочного пакета МИТП RU.СНАБ.80066-06 01, который содержит данный компонент и готовые к установке пакеты системных компонентов, включая:

- компонент хранения знаний RU.СНАБ.80066-06 01 17;
- компонент диалога поддержки принятия решений RU.СНАБ.80066-06 01 18;
- компонент разбора скрипта EasyFlow RU.СНАБ.80066-06 01 19;
- компонент интерпретации WF RU.СНАБ.80066-06 01 20;
- компонент взаимодействия с пользователем RU.СНАБ.80066-06 01 21;
- компонент серверной визуализации RU.СНАБ.80066-06 01 22;
- компонент событийного взаимодействия RU.СНАБ.80066-06 01 23;

- компонент мониторинга RU.СНАБ.80066-06 01 24;
- компонент контроля доступа RU.СНАБ.80066-06 01 26;
- компонент обеспечения доступа к инфраструктуре RU.СНАБ.80066-06 01 27;
- компонент планирования исполнения WF RU.СНАБ.80066-06 01 28;
- компонент исполнения WF RU.СНАБ.80066-06 01 29;
- компонент информационного портала RU.СНАБ.80066-06 01 31;
- компонент хранения профилей исполнения WF RU.СНАБ.80066-06 01 32;
- компонент-база ресурсов RU.СНАБ.80066-06 01 33;
- компонент учета использования ресурсов RU.СНАБ.80066-06 01 34;
- компонент-база пакетов RU.СНАБ.80066-06 01 35;
- компонент хранения данных RU.СНАБ.80066-06 01 37;
- компонент доступа к вычислительным ресурсам RU.СНАБ.80066-06 01 38;
- компонент сбора данных в социальных сетях в Интернете RU.СНАБ.80066-06 01 39.

Для развертывания компонентов МИТП-Д необходима вычислительная система под управлением ОС Windows (XP и выше), с установленной средой Silverlight 4.0, или Linux (с ядром 2.6.22 и выше), с установленной средой Mono Framework с поддержкой библиотек .NET 2.0 и выше (рекомендуется версия Mono Framework 2.6 или выше). Для корректного функционирования необходимо наличие установленного web-сервера с поддержкой технологии ASP .NET WebServices, WCF, Silverlight и удаленного развертывания сервисов (с использованием технологии WebDeploy). Примером web-сервера, соответствующего предъявленным требованиям может служить Microsoft IIS версии 7.0 или выше.

Дополнительно для функционирования МИТП-Д должен быть установлен сервер баз данных: MongoDB версии 1.6.5. В ходе установки и настройки используются стандартные конфигурации указанных программных средств, не требующие отдельной настройки. После установки необходимо осуществить запуск сервера баз данных для локального использования (localhost). СУБД MongoDB используется компонентами CLAVIRE/Ginger RU.СНАБ.80066-06 01 21 – для хранения данных о пользовательских проектах; CLAVIRE/Eventing RU.СНАБ.80066-06 01 23 – для журналирования произошедших в системе событий; CLAVIRE/Monitoring RU.СНАБ.80066-06 01 24 – в качестве хранилища актуальных данных о платформе; CLAVIRE/GateKeeper RU.СНАБ.80066-06 01 26 – для хранения учетных данных пользователей; CLAVIRE/InfraAccess RU.СНАБ.80066-06 01 27 – для хранения данных о

зарегистрированных компонентах; CLAVIRE/Provenance RU.СНАБ.80066-06 01 32 – для хранения профилей исполнения композитных приложений; CLAVIRE/Billing RU.СНАБ.80066-06 01 34 – для хранения пользовательских счетов, тарифов и истории операций; CLAVIRE/Storage RU.СНАБ.80066-06 01 37 – для хранения сервисной информации, используемой центральным модулем хранения данных, а также для хранения метаинформации, соответствующей объектам хранения.

Для работы компонента информационного портала RU.СНАБ.80066-06 01 31 требуется установка СУБД MySQL (версии 5.0 или выше) и поддержка web-сервером интерпретатора языка PHP (версии 5.2 или выше). Для работы компонента хранения знаний RU.СНАБ.80066-06 01 17 требуется установка СУБД Microsoft SQLServer Compact Edition (версии 3.5 или выше). Также должен быть установлен web-сервер Glassfish версии 3.0.1, обеспечивающий поддержку технологии WebServices, необходимой для функционирования варианта реализации хранилища онтологической структуры RunLib. На ту же вычислительную систему должен быть установлен интерпретатор онтологической структуры Pellet версии 2.2.2, необходимый для функционирования хранилища знаний.

Для работы компонента сбора данных в социальных сетях в Интернете RU.СНАБ.80066-06 01 39 необходимы развернутые библиотеки с открытым исходных кодом Apache Hadoop (версии 0.20.2 и выше).

2.2 Необходимые технические средства управляющей подсистемы МИТП-Д

Компоненты МИТП-Д функционируют на вычислительной системе – серверной ЭВМ со следующими минимальными характеристиками:

- тип процессоров: Intel-совместимый;
- количество ядер – не менее 4;
- количество процессоров – не менее 2;
- тактовая частота каждого процессора – не менее 2.0 ГГц;
- оперативная память (на ядро) – не менее 2.0 ГБ;
- дисковая подсистема – не менее 5×250 ГБ RAID5;
- пропускная способность сетевых интерфейсов – не менее 1 Гбит/с.

Для взаимодействия с другими модулями системы требуется наличие выхода в Интернет или локальную сеть (если web-сервисы других подсистем доступны из локальной сети) с соответствующей поддержкой со стороны оборудования.

Для функционирования компонента развертывания и конфигурирования необходима рабочая станция с видеоадаптером и дисплеем, способным отображать WPF-приложение с размером окна 800×600 пикселов, со следующими минимальными характеристиками:

- архитектура процессора – x86, x86_64, IA64;
- объем оперативной памяти – 1 ГБ;
- объем свободного пространства на жестком диске – 1 ГБ;
- тактовая частота процессора – 1 ГГц.

В целях увеличения производительности и реактивности МИТП-Д отдельные компоненты могут функционировать на разных вычислительных системах в рамках общей локальной сети центра обработки данных.

2.3. Необходимые технические средства подсистемы вычислительной инфраструктуры МИТП-Д

В состав комплекса технических средств подсистемы вычислительной инфраструктуры МИТП-Д входят как собственные ресурсы центра компетенции в области обработки данных, так и ресурсы внешних провайдеров, например, в составе глобальных сред распределенных вычислений:

1. *Вычислительные кластеры.* Доступные высокопроизводительные ресурсы центра компетенции и внешних провайдеров; предназначены для установки (в т.ч. автоматической) и последующего использования прикладных сервисов МИТП-Д. Ввиду того что МИТП-Д предоставляет возможность унифицированной работы с ресурсами, обладающими различными классами архитектур, ключевыми требованиями к таким ресурсам являются возможность доступа по стандартным сетевым протоколам в рамках корпоративной сети, программная совместимость с прикладными пакетами, а также возможность использования стандартных средств управления ресурсами.
2. *Корпоративная или глобальная облачная инфраструктура.* Виртуальная вычислительная инфраструктура, конфигурируемая по запросу со стороны МИТП-Д или пользователя системы (при организации доступа к уже сконфигурированным виртуальным ресурсам). Со стороны МИТП-Д работа с облачной инфраструктурой происходит не только на уровне абстрактных прикладных сервисов (реализуются подбор и конфигурация существующих статических ресурсов), но и на уровне

абстрактных вычислительных ресурсов (осуществляется динамическая конфигурация ресурса в соответствии с предъявляемыми требованиями).

3. *Грид-инфраструктура первого поколения.* Данный класс ресурсов реализует концепцию виртуальных организаций. При этом специфика доступа к таким ресурсам определяется технологическими особенностями сервисной среды Грид, а также высокой изменчивостью структуры и характеристик этих ресурсов. Тем не менее распределенный характер и высокая суммарная производительность позволяют эффективно задействовать данный класс ресурсов при решении ряда вычислительных задач.
4. *Прочие виды корпоративных ресурсов* (рабочие станции, серверные ЭВМ и пр.). Интеграция широкого спектра разнородных ресурсов позволяет сформировать инфраструктуру, обеспечивающую исполнение заданий, оптимизированных для различных архитектур. Минимальные требования к корпоративным ресурсам для использования в МИТП-Д:
    - архитектура: SMP, MPP, GPGPU, CBEA;
    - тип процессоров: Intel-совместимый;
    - количество ядер – не менее 4;
    - количество процессоров – не менее 1;
    - количество вычислительных узлов – не менее 1;
    - тактовая частота каждого процессора – не менее 2.0 ГГц;
    - оперативная память (на ядро) – не менее 1.0 ГБ;
    - дисковая подсистема – не менее 250 ГБ на узел;
    - системы управления Torque, Ganglia
    - операционные системы: Windows, Linux.
5. *Специализированные хранилища данных.* МИТП-Д обеспечивает унифицированный доступ как к локальным, так и к распределенным хранилищам и источникам данных при условии их нахождения в локальной сети центра компетенции совместно с управляющей подсистемой МИТП-Д. Использование инфраструктуры внешних провайдеров для хранения данных МИТП-Д нецелесообразно по соображениям: (а) безопасности и (б) надежности.
6. *Вычислительная инфраструктура низкоуровневой среды обработки данных* необходима для развертывания системы Apache Hadoop и настройки системного сервиса сбора данных в социальных сетях в Интернете. В состав инфраструктуры включаются минимум 4 ЭВМ со следующими характеристиками:

- тип процессоров: Intel-совместимый;
- количество ядер – не менее 4;
- количество процессоров – не менее 1;
- тактовая частота каждого процессора – не менее 2.0 ГГц;
- оперативная память (на ядро) – не менее 2.0 ГБ;
- дисковая подсистема – не менее 0,5 ТБ;
- пропускная способность сетевых интерфейсов – не менее 1 Гбит/с.

3. ОПИСАНИЕ ЗАДАЧИ

3.1 Определение задачи

Основной задачей МИТП-Д является обеспечение работы пользователя с облачными сервисами и композитными приложениями для сбора и обработки больших объемов данных в распределенных источниках. В целом организация процесса создания и исполнения композитного приложения под управлением МИТП-Д в рамках концепции iPSE сводится к последовательной формализации наборов описаний в терминах потоков заданий (workflow, WF). На первом этапе процесса проектирования композитного приложения создается мета-WF (MWF). Пользователь может осуществлять выбор классов сервисов, которые доступны в облачной среде, и уточнять их по мере ввода дополнительной информации. Указанные пользователем классы сервисов будут использоваться на следующем этапе для подбора конкретных сервисов. На основе MWF создается поток заданий, в котором уже зафиксированы конкретные реализации вычислительных сервисов, однако еще ничего не известно об условиях их выполнения (AWF). Следующим этапом процесса проектирования является построение расписания и создание сценария выполнения в терминах конкретного WF (CWF), который представляет собой поток заданий с полностью определенными блоками. Для блоков действий указаны сервисы и узлы для исполнения, а для блоков данных – конкретное местоположение необходимых данных. Более полное описание процесса создания и исполнения композитного приложения, общего для всех технологических платформ в составе МИТП CLAVIRE, приведено в документе RU.СНАБ.80066-06 31 01.

Применительно к МИТП-Д процесс создания и использования приложений сбора и обработки данных определяется следующим алгоритмом (рис. 3.1).

1. Пользователь авторизуется (используя сертификат) в МИТП-Д через портал провайдера услуг МИТП, что дает ему возможность доступа к соответствующим ресурсам, сервисам и источникам данных.
2. Пользователь МИТП-Д через соответствующий web-интерфейс может выбрать конкретные сервисы или шаблоны композитных приложений в форме WF, а также получить (при необходимости) доступ к технической и эксплуатационной документации. За подбор, каталогизацию и аннотирование сервисов отвечают специалисты провайдера МИТП-Д.
3. Выбрав необходимые сервисы, пользователь средствами МИТП-Д конструирует соответствующее композитное приложение в форме описания WF, которое определяет правила сбора, обработки и анализа данных. При этом может использоваться как графическая, так и текстовая форма представления композитных приложений.
4. Для подготовленного описания композитного приложения пользователь конфигурирует условия вычислений: определяет требуемые параметры WF, редактирует (при необходимости) его описание, готовит и загружает в хранилище МИТП-Д входные данные для расчетов. В ряде случаев такие данные могут предоставляться провайдером МИТП-Д (через соответствующий раздел хранилища).
5. Пользователь определяет режим исполнения задачи в МИТП-Д (утверждает предлагаемые ему варианты) в соответствии с требованиями к временным характеристикам расчета и правилами доступа к различным источникам данных. При этом пользователю предлагаются различные тарифные варианты, определяемые использованием разных социальных сетей и привлечением ресурсов сред распределенных вычислений.
6. Пользователь запускает задачу на исполнение в среде МИТП-Д. Использование вычислительных ресурсов и сервисов производится с учетом единого сертификата МИТП-Д, с которым в средах внешних провайдеров распределенных вычислений ассоциированы сертификаты «суперпользователей» МИТП-Д.
7. В процессе сбора данных из хранилищ социальных сетей в локальные хранилища данных, связанные с элементами WF, доступ к данным обеспечивается путем авторизации через соответствующий сервис оператора социальной сети. Авторизация проводится по сертификату

«суперпользователя» МИТП-Д; персональные данные реального пользователя (при необходимости) передаются оператору социальной сети постфактум.

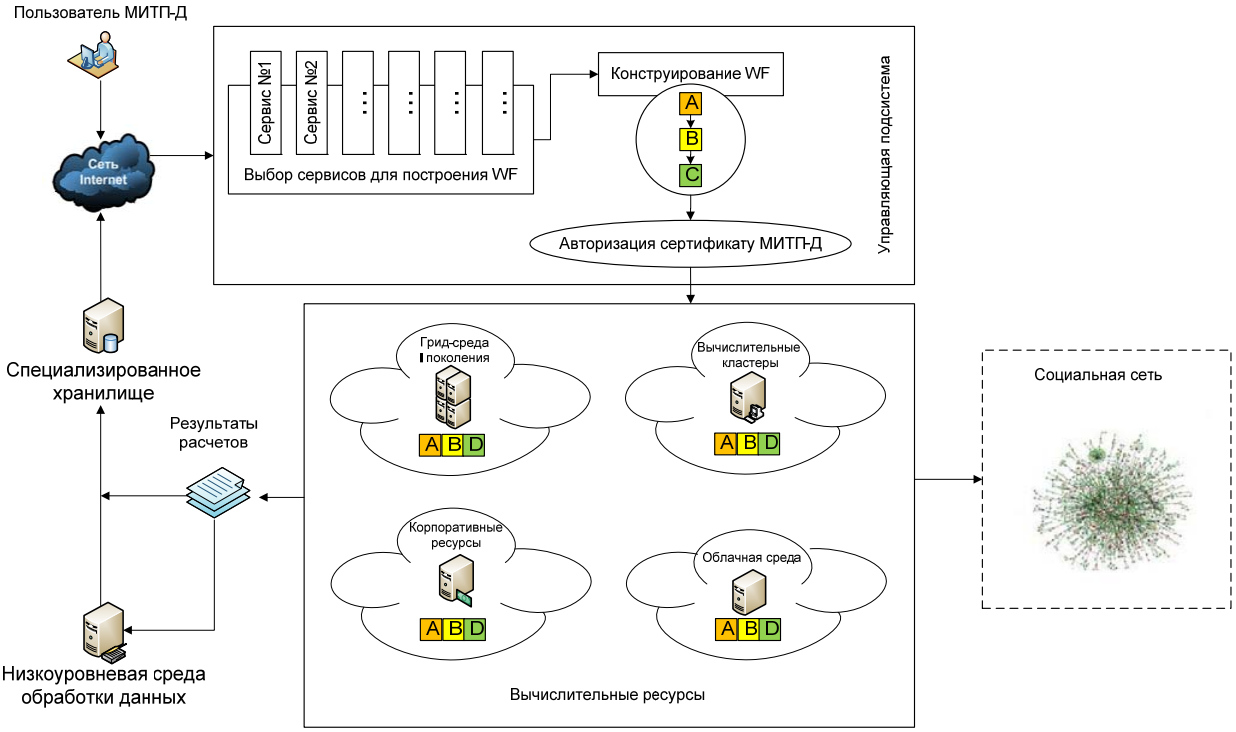


Рисунок 3.1 Принцип функционирования МИТП-Д
(на примере мониторинга социальных сетей в Интернете)

8. Пользователь (при необходимости) осуществляет мониторинг процесса исполнения (в форме динамического отображения WF); при этом прогнозируется время завершения вычислений. Пользователь может на время расчетов завершить рабочую сессию с МИТП-Д и начать ее только при необходимости использования результатов.
9. Когда все данные собраны и расчет задачи завершен, результаты помещаются в хранилище данных МИТП-Д; пользователю отправляется соответствующее уведомление. Пользователь может получить доступ к результатам расчетов через интерфейс МИТП-Д.

Таким образом, основная задача МИТП-Д сводится к технологическому обеспечению операций алгоритма.

3.2. Методы решения задачи

В данном разделе рассматриваются основные методы решения задачи раздела 3.1, характерные для МИТП-Д, они предназначены для:

1) унифицированного описания прикладных пакетов с целью организации единообразного доступа к ним в форме облачных сервисов;
2) описания структуры композитных приложений, состоящих из нескольких взаимодействующих сервисов в распределенной среде;
3) организации всего процесса исполнения композитного приложения под управлением МИТП-Д.
4) выполнения процедуры сбора данных в социальных сетях в Интернете.

3.2.1. Унифицированное описание прикладных пакетов в МИТП-Д

В МИТП-Д простейшая форма WF представляет собой описание исполнения одного вычислительного пакета на ресурсе корпоративной вычислительной среды, с загрузкой входных данных и получением выходных. Однако унифицированное описание этого действия осложнено тем, что разные вычислительные пакеты используют свою стратегию работы с данными (использование конфигурационного файла, командной строки аргументов, переменных окружения, проектов, хранящихся в структуре директорий и файлов). Ситуация осложняется требованием единообразных принципов работы с одним и тем же пакетом, установленным на ресурсах с различными операционными системами, средами управления и исполнения и пр.

В МИТП-Д для решения задачи унификации описаний пакетов использован предметно-ориентированный язык (Domain Specific Language, DSL) EasyPackage, позволяющий описывать пакеты в наглядной форме, понятной специалистам-предметникам, и поддающийся программной обработке. EasyPackage разработан на основе реализации языка Ruby (IronRuby), он является интерпретируемым со строгой динамической типизацией и явным приведением типов. Его базовые элементы идентичны элементам Ruby.

Описание пакета представляет собой один или несколько текстовых файлов. Оно использует следующие понятия: пакет, входной/выходной параметр, входной/выходной файл, режим запуска. *Пакет* – это исполняемое приложение, запускаемое в пакетном режиме (модель IPO – Input–Process–Output), которое принимает на входе определенный набор файлов, параметров командной строки, переменных окружения и других источников данных, а на выходе генерирует набор выходных файлов. *Параметр пакета* – это элемент данных, имеющий имя, тип и значение. Параметр может быть входным или выходным, а также может быть параметром исполнения. Тип параметра может быть одним из базовых: строка, логический тип, число с плавающей точкой, перечислимый тип,

целое число, список. Режим запуска характеризуется набором используемых в нем параметров.

Структура описания пакета состоит из раздела объявления расширений, общего описания пакета, секционного описания входных и выходных данных пакета (секции *inputs* и *outputs*), описания параметров исполнения (рис. 3.2). Раздел объявления расширений предназначен для определения процедур, позволяющих расширить функциональные возможности базовой библиотеки языка. Общее описание пакета включает в себя набор полей, несущих общую информацию о пакете: имя, версия, лицензия, поставщик и т.д. (строки 1–6). Раздел секционного описания содержит определение входных и выходных параметров и файлов. Параметры характеризуются следующим набором полей: тип, значение по умолчанию, процедура проверки значения параметра на корректность (например, параметр в строках 15–21). Параметры могут быть вычислимыми (строки 22–27), тогда для них указывается процедура вычисления из рабочего контекста – *evaluator* (строка 26).

```
1  name "TESTP"
2  display_as "Testp"
3  vendor "SPbSU ITMO"
4  url "http://escience.ifmo.ru"
5  license "GPLv3"
6  description "Simple package example"
7  inputs {
8      raw file {
9          name "inf"
10         filename "arg.txt"
11         place "/"
12         extractor IntegerFileExtractor.new("in")
13         assembler ObjectToSAssembler.new("in")
14     }
15     meta param {
16         name "in"
17         required
18         type int
19         validator lambda { |val, ctx| val > 0 and val < 10000 }
20         validation_error_msg "num have to be in [0; 10000]"
21     }
22     meta param {
23         name "abs_plus_3"
24         required
25         type int
26         evaluator { |ctx| ctx.in.abs + 3 }
27     }
28     cmdline lambda { |ctx| "{0} arg.txt out.txt" }
29 }
30 outputs {
31     auto file {
32         name "output_num"
33         required
34         filename "out.txt"
35         place "/"
36         extractor IntegerFileExtractor.new("out")
37     }
38     auto param {
39         name "out"
40         required
41         display_as "Output number"
42         type int
43     }
44 }
45 prepare_package
```

Рисунок 3.2 – Фрагмент описания прикладного пакета на языке EasyPackage

Контекст работы представляет собой набор уже вычисленных значений параметров (*ctx*). Файловые параметры дополнительно имеют следующий набор полей (строки 8–14): имя файла, путь до файла, процедура извлечения данных из файла (*extractor*), процедура сборки файла (*assembler*). Процедура сборки файла позволяет создавать входной файл, основываясь на значениях входных параметров. На практике используются стандартные процедуры, например, сборка файла по шаблону (библиотека ERB). Процедура извлечения данных из файла, как правило, применяется для выходных файлов пакета с целью определения значений выходных параметров (строка 12). Базовый набор процедур извлечения значений из файлов и их сборки из параметров можно дополнять за счет

написания своих процедур в секции расширений. Последним в файле описания является раздел параметров исполнения, который позволяет при работе с пакетом не учитывать неоднородность ресурсов (различных ОС, архитектур). К параметрам данного раздела относятся: скрипт запуска пакета (точнее, процедура его сборки), командная строка, переменные окружения.

Таким образом, описание на языке EasyPackage позволяет не только задать правила обращения к конкретному пакету в распределенной вычислительной среде, но и корректно интерпретировать его входные и выходные данные (посредством процедур *extractor* и *assembler*). Это обеспечивает совместимость (по данным) пакетов различных разработчиков в составе WF.

Подробно особенности языка EasyPackage изложены в документе RU.СНАБ.80066-06 33 01.

3.2.2. Унифицированное описание композитных приложений в МИТП-Д

Описание композитных приложений, состоящих из нескольких прикладных пакетов, требует определения не только правил работы с пакетами (см. раздел 3.2.1), но и структуры взаимодействия между ними. Специализированный язык EasyFlow, поддерживаемый МИТП-Д, позволяет упростить процедуру задания композитных приложений. Он предоставляет конечному пользователю гибкие возможности по заданию различных форм WF, в рамках которых выполняются различные прикладные пакеты, происходит генерация выходных данных, их получение, конвертация и обработка.

Характерной чертой языка является полное абстрагирование от особенностей распределенной вычислительной среды, в которой работает пользователь. Фактически EasyFlow – это высокоуровневый язык описания AWF. Такой подход позволяет описывать саму решаемую задачу, а не способ ее исполнения на конкретной вычислительной архитектуре. На рис. 3.3 приведен пример описания простого AWF, представляющего собой скрипт. Тело скрипта состоит из описания вызовов прикладных пакетов – *шагов*, которые задаются с помощью директивы *step* и представляют собой узлы графа WF. Для описания каждого шага необходимо задать его имя (в примере это *Step1*, *Step2*, *Step3*), название запускаемого пакета (*EmptyPackage*, *Package1* и *Package2*) и перечень предметных параметров этого пакета (см. раздел 3.2.1).

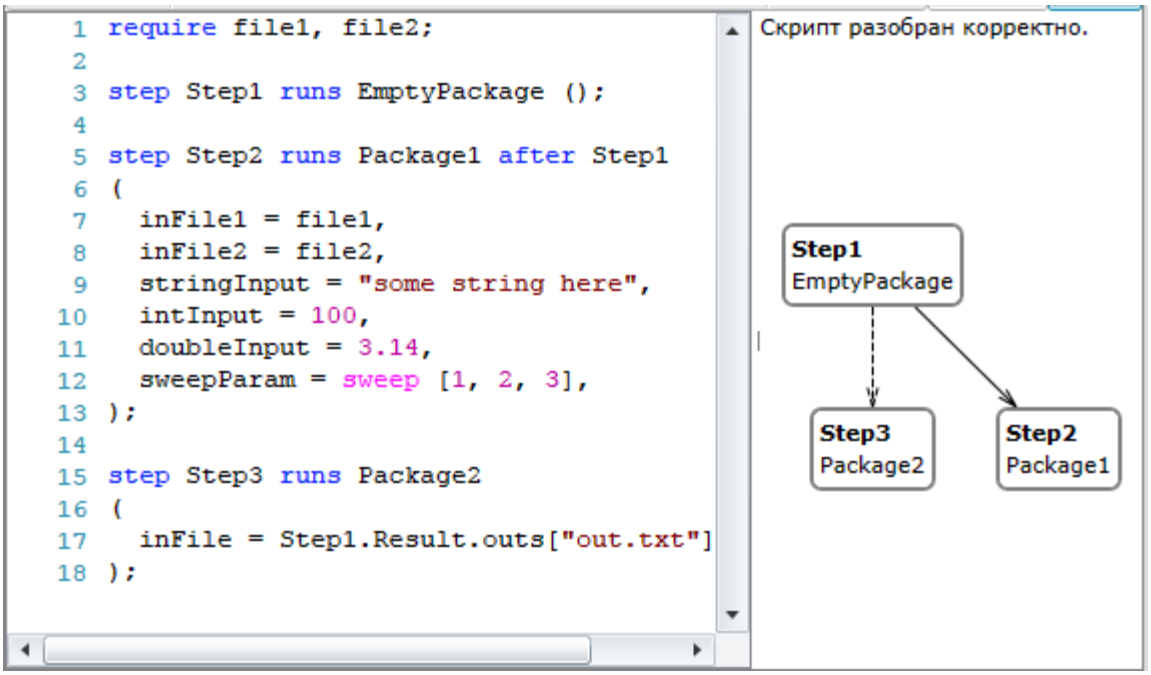


Рисунок 3.3 – Пример описания композитного приложения на языке EasyFlow и его графическое представление в МИТП-Д

Язык EasyFlow позволяет задавать параметры для следующих типов данных: целое число, строка, число с плавающей точкой, список, структура, указание на использование файла (см. описание шага *Step3*).

Большинство прикладных пакетов помимо параметров принимает и генерирует входные и выходные файлы, поэтому в EasyFlow предусмотрена поддержка работы с файлами. Их задание в скрипте представляет собой лишь абстрактное указание с помощью директивы *require*, что освобождает пользователя от необходимости указания абсолютных путей к файлам. В этой директиве через запятую перечислены файловые переменные, которые могут быть указаны в качестве значений параметров при описании шага (см. параметры *inFile1* и *inFile2* в описании шага *Step2*). В рамках одного скрипта директива требования файлов может появляться неограниченное число раз.

Так как WF представляет собой ориентированный граф, в EasyFlow введены механизмы определения порядка выполнения шагов, позволяющие организовать его структуру: зависимости по управлению и зависимости по данным.

Зависимости по управлению представляют собой явные указания на то, что один шаг должен начать свое исполнение после завершения другого. Это делается с помощью директивы *after* (см. рис. 3.3, шаг *Step2*).

Зависимости по данным представляют собой неявные указания на зависимости между шагами, которые анализируются при интерпретации скрипта EasyFlow. Они

выражаются в том, что некоторые шаги могут использовать данные других шагов, что неявно влияет на последовательность их запуска. Такие зависимости могут присутствовать в описываемом WF одновременно с зависимостями по управлению, что позволяет очень гибко настраивать порядок выполнения шагов. Пример зависимостей по данным содержится в описании шага *Step3* (строка 17), где указано, что в качестве входного файла используется файл out.txt, полученный в результате выполнения шага *Step1*.

Еще одной полезной возможностью EasyFlow является автоматическое варьирование параметров (*parameter sweep*). Такая задача часто возникает, когда необходимо запустить один и тот же вычислительный пакет, закрепив одни и варьируя другие параметры. Для этого в язык введена директива *sweep*, которая принимает список параметров для варьирования и из одного шага создает *N* шагов, где *N* соответствует числу элементов в декартовом произведении списков варьирования для различных параметров.

Пример варьирования параметра приведен на рис. 3.3 (строка 12). В этом примере будут запущены три шага *Step2* с параметром *sweepParam*, равным соответственно единице, двум и трем при прочих зафиксированных параметрах.

Таким образом, WF, описанные на языке EasyFlow, полностью независимы от конкретной архитектуры вычислений и хранения данных, что позволяет пользователям распределенной среды беспрепятственно обмениваться ими и запускать их на различных вычислительных ресурсах.

Подробно особенности языка EasyFlow изложены в документе RU.СНАБ.80066-06 33 02.

3.2.3. Организация процесса исполнения композитного приложения в МИТП-Д

Рассмотренные выше подходы (см. разделы 3.2.1–3.2.2) к описанию пакетов, композитных приложений и интерфейсов работы с ними позволяют организовать процесс исполнения приложения в рамках МИТП-Д. Композитное приложение представляется в виде скрипта описания WF на языке EasyFlow, который может быть параметризован набором входных параметров и файлов, а также параметров исполнения, т.е. один и тот же WF может быть исполнен для разного набора входных данных, а также в различных условиях исполнения. За разбор скрипта WF и исполнение WF в целом отвечает компонент интерпретации WF CLAVIRE/FlowSystem RU.СНАБ.80066-06 01 20. После получения скрипта он «разбирает» его и преобразует

во внутреннее представление (предварительно проверив его корректность). Обработка представления WF производится непрерывно, согласно событийной модели функционирования, т.е. интерпретация WF происходит в рамках цикла обработки поступающих событий. При запуске отдельной задачи происходит интерпретация параметров узла WF и формируется описание запуска задачи, после чего сформированное описание передается в очередь компоненту исполнения WF CLAVIRE/Executor RU.СНАБ.80066-06 01 29. Далее компонент CLAVIRE/Executor подготавливает данные для пакета и производит запуск пакета в среде Грид, после чего обрабатывает выходные данные с помощью компонента хранения данных CLAVIRE/Storage RU.СНАБ.80066-06 01 25.

Для подготовки пакета к запуску и обработки его результатов используется компонент CLAVIRE/PackageBase RU.СНАБ.80066-06 01 35, который позволяет сопоставить абстрактные и фактические правила работы с каждым пакетом в составе WF. На рис. 3.4(а) приведена структура CLAVIRE/PackageBase (см. подробнее документ RU.СНАБ.80066-06 13 35). Компонент состоит из интерфейсной библиотеки и репозитория пакетов. Описание пакета в такой схеме хранится в виде файла со скриптом EasyPackage в репозитории. При этом системные модули для работы с данным описанием получают скрипты и сопутствующие файлы, и интерпретируют их на своей стороне, используя лишь необходимую им информацию. Данный подход выгодно отличается от применения централизованного хранилища информации о пакетах, построенного на сервисно-ориентированной модели, так как позволяет легко масштабировать систему на большее количество пользователей за счет перераспределения нагрузки.

На рис. 3.4(б) представлен сценарий исполнения композитного приложения под управлением МИТП-Д с использованием подходов, описанных в разделах 3.2.1–3.2.2. Характеристиками на данной схеме считаются вычислимые параметры, необходимые только для планирования запуска в распределенной среде. После получения и обработки результатов данные о завершении работы WF передаются обратно в компонент CLAVIRE/FlowSystem RU.СНАБ.80066-06 01 20, там они становятся доступными пользователю (через соответствующий интерфейс человеко-компьютерного взаимодействия CLAVIRE/Ginger RU.СНАБ.80066-06 01 21).

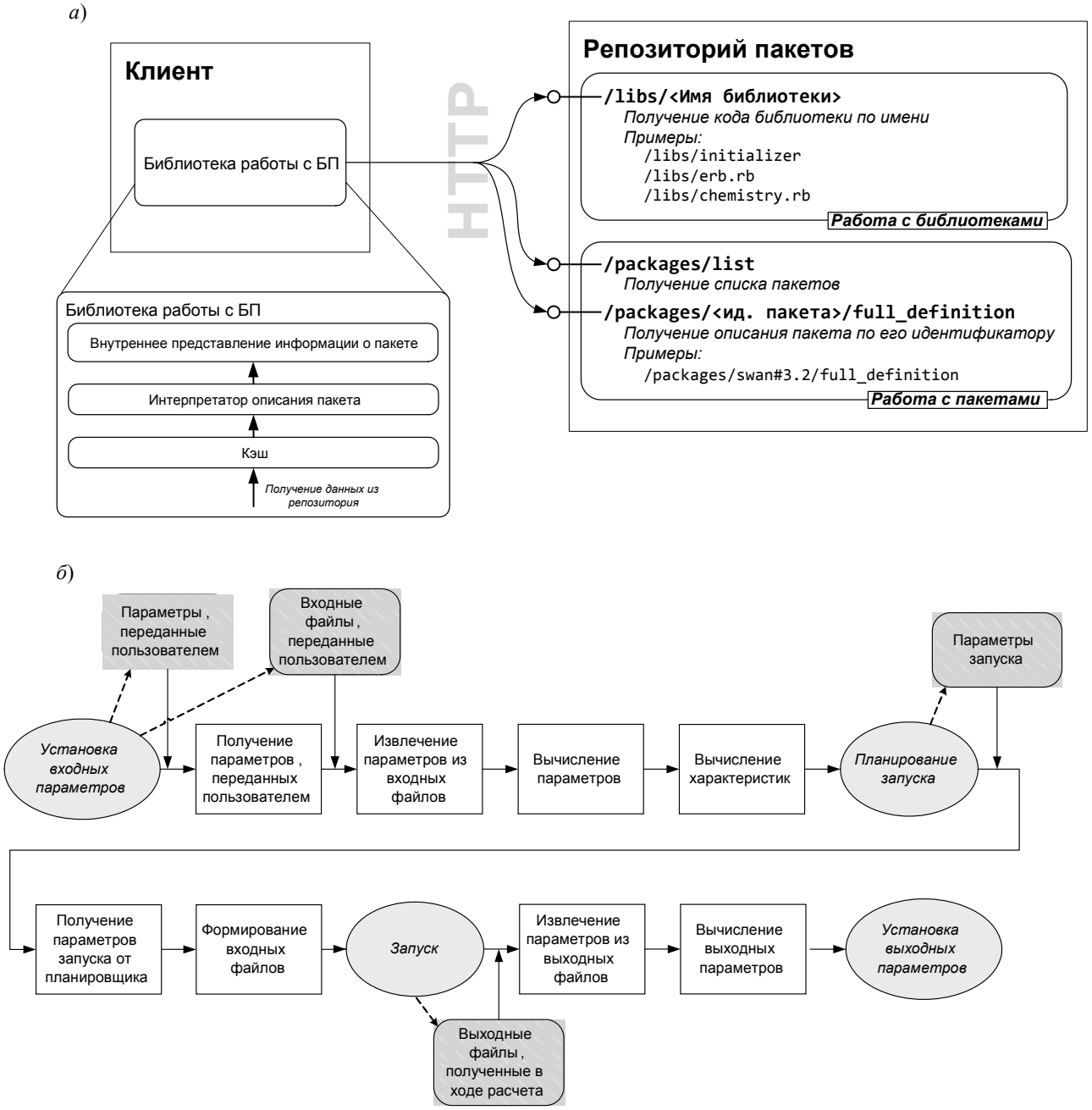


Рисунок 3.4 – Интерпретация и исполнение WF: (а) структура CLAVIRE/PackageBase ; (б) процесс исполнения композитного приложения под управлением МИТП-Д

Таким образом, рассмотренные в разделах 3.2.1–3.2.3 методы и технологии обеспечивают поддержку разработки и исполнения прикладных сервисов и композитных приложений на их основе для сбора и обработки данных. При этом сервисы сбора данных исполняются по той же схеме (рис. 3.4), что и сервисы обработки и моделирования.

3.2.4. Организация сбора и анализа данных в социальных сетях в Интернете

На рис. 3.5 приведена общая схема процедуры краулинга – сбора данных в социальных сетях в Интернете, поддерживаемая МИТП-Д.

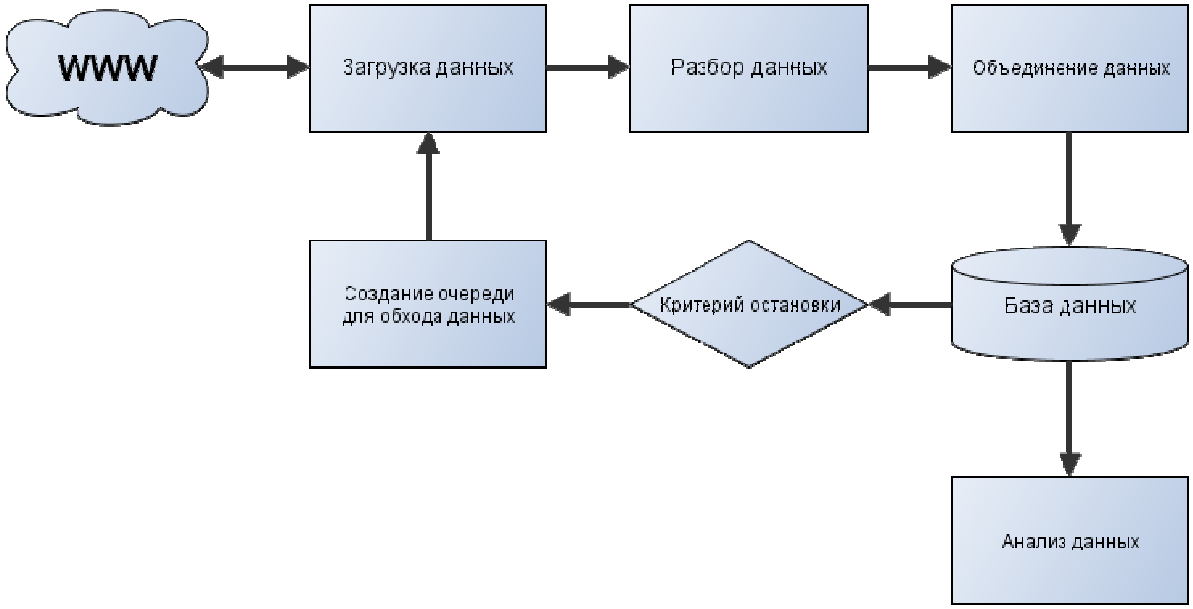


Рисунок 3.5 – Общая схема работы краулера, реализованная в МИТП-Д

Краулер МИТП-Д работает итерационно: на каждой итерации он посещает фиксированное число пользователей, после чего происходит анализ собранной информации. На рис. 3.6 приведена схема одной итерации краулера, эллипсами на ней изображены процессы, проводимые над данными, а прямоугольниками – данные. Процессы выполняются последовательно друг за другом, и их взаимное расположение по горизонтали указывает на порядок выполнения.

На первом этапе работы краулера создается список новых узлов, найденных на каждой итерации. Для этого анализируются все ребра, найденные на текущей итерации, и проверяется, что они ведут в еще не посещенные узлы, для чего в память загружается список уже посещенных узлов. На втором этапе берется список узлов для обхода, построенный на предыдущей итерации, и от него отфильтровываются узлы, не посещенные на текущей итерации. Если обнаружилось, что какой-либо узел не посещен вследствие ошибки, счетчик числа попыток посещения увеличивается. Затем оба списка узлов объединяются, при этом происходит обновление атрибутов узла: числа попыток, затраченных на его посещение, и его приоритета, также происходит удаление из очереди узлов, для которых счетчик числа попыток посещения превысил заданное значение. После

этого полученный список узлов ранжируется согласно значению приоритета, и он становится списком узлов для обхода на новой итерации.

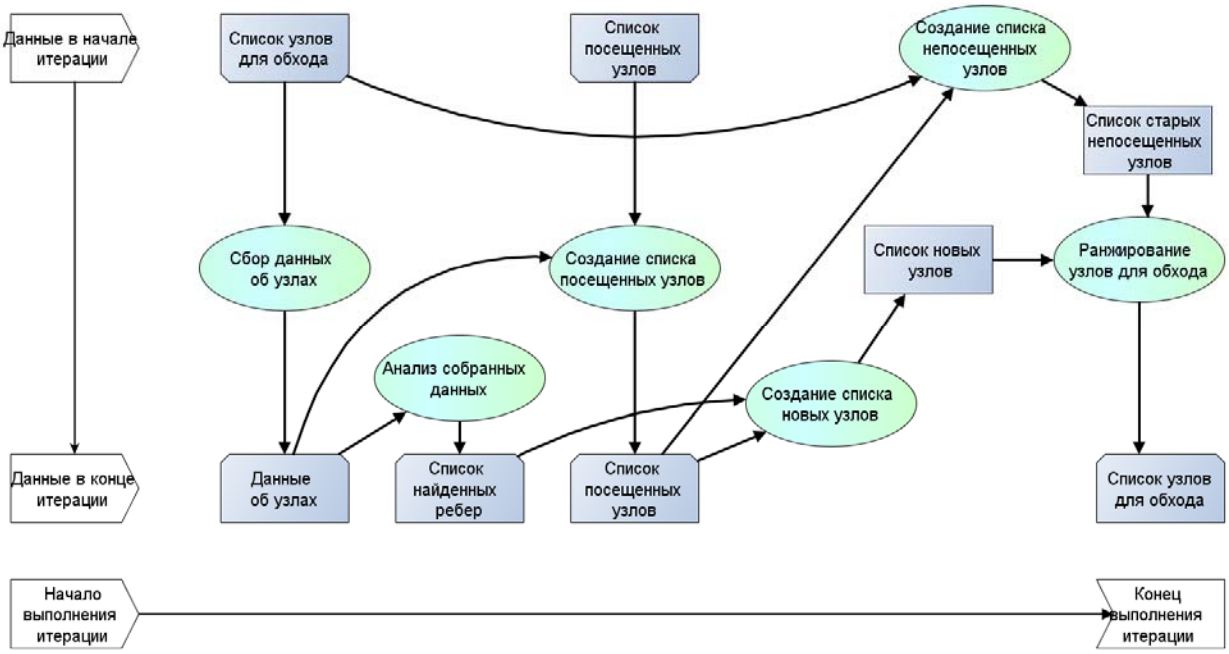


Рисунок 3.6 – Схема одной итерации краулера

Приоритет узла считается равным числу ссылок, указывающих на него. Чем больше ссылок найдено, тем выше приоритет узла и тем вероятнее он содержит информацию о заданной тематике. Число ссылок легко поддерживать в узле от итерации к итерации, оно равно сумме числа ссылок, найденных на предыдущих итерациях, и числа ребер, найденных на текущей итерации, указывающих на это ребро.

Социальные сети устанавливают различные правила, регулирующие автоматический сбор данных. В МИТП-Д поддерживаются два способа сбора данных: посредством специального API и на основе непосредственного анализа HTML-кода, выдаваемого сервисом для обычных пользователей социальных сетей.

В краулере, поддерживаемом МИТП-Д, используется подход, основанный на анализе контекста пользователя социальной сети (его интересов и его документов), в предположении, что люди, интересующиеся схожими темами, образуют неявные сообщества в социальной сети, и между ними существуют ссылки. Таким образом, политика обхода сети выглядит следующим образом.

1. Анализ контекста пользователя (его интересов или документов) и проверка их соответствия заданным темам.
2. Если контекст пользователя отнесен к определенной теме, то его друзья добавляются в очередь узлов для обхода.

3. Если контекст пользователя не отнесен к определенной теме, то продолжается поиск новых узлов.

Поэтому можно говорить, что политика обхода фактически определяется алгоритмами анализа контекстов пользователя и их фильтрации.

Для содержательного анализа текстов и интересов пользователя в краулере МИТП-Д используются три подхода:

1) на основе списка ключевых слов (задаются экспертно). При наличии *хотя бы одного ключевого слова* считается, что документ посвящен заданной теме, это инициирует обход «друзей» текущего пользователя;
2) на основе списка ключевых фраз – неупорядоченных наборов ключевых слов, которые свидетельствуют о тематической принадлежности текста. При наличии *хотя бы одной фразы* считается, что документ посвящен заданной теме, это инициирует обход «друзей» текущего пользователя;
3) на основе взвешенной оценки текста по всем ключевым словам и фразам (по достижении весом определенного порога инициируется обход «друзей» текущего пользователя).

Таким образом, рассмотренная процедура позволяет собрать и поместить в локальное хранилище краулера (на основе Apache Hadoop) все содержимое web-страниц пользователей социальных сетей, соответствующих заданной теме (включая мультимедийный контент). Непосредственное перемещение этих данных в хранилище МИТП-Д не всегда целесообразно в силу их неоднородности и большого объема. Потому первичная обработка выполняется непосредственно в процессе краулинга; данные, необходимые для детального анализа или усвоения в моделях, передаются в МИТП-Д через стандартный механизм WF, описанный в разделах 3.2.1–3.2.2.

3.2.5. Особенности создания и управления потоковой обработки сверхбольших объемов данных и извлечения из них знаний на основе облачных вычислений

Использование МИТП-Д связана с такими задачами, как сбор и обработка данных из социальных сетей. Специфика выполнения количественных исследований на основе социальных сетей связана с рядом особенностей, ограничивающих доступность таких данных для широкого круга исследователей; отметим некоторые из них.

1. Доступ к данным глобальных социальных сетей регламентируется политикой оператора сети и соответствующим законодательством в области персональных

данных. Для масштабного сбора и анализа соответствующих данных необходимо наличие предварительных соглашений с оператором. Данные соглашения являются прерогативой оператора МИТП-Д, который предоставляет права их использования потребителям (клиентам МИТП).

2. Социальные сети имеют технологически различные интерфейсы доступа к данным, существуют разные принципы обхода сетей. Как следствие, проведение измерений характеристик различных сетей требует разработки специализированных средств сбора данных, для отдельных исследователей это весьма трудоемкий процесс. Потому МИТП-Д обеспечивает унифицированный интерфейс для доступа к различным социальным сетям.
3. Сбор данных в социальных сетях является достаточно ресурсоемкой операцией и требует соответствующих выделенных вычислительных ресурсов. Регулярное выполнение таких операций различными пользователями увеличивает нагрузку на инфраструктуру оператора сети, что нежелательно. МИТП-Д обеспечивает единую инфраструктуру сбора и хранения таких данных на основе облачной среды; при этом отдельным пользователям может предоставляться доступ как к сервисам сбора, так и сервисам доступа к данным, уже сохраненным в хранилище МИТП.
4. Алгоритмы обработки и анализа таких социальных сетей во многих случаях имеют нелинейную сложность, поскольку описывают взаимоотношения „каждого с каждым“. Поэтому исследование сетей достаточно большого объема требует применения соответствующих вычислительных ресурсов и программного обеспечения, допускающего эффективное распараллеливание. Т.е. МИТП-Д обеспечивает доступ к соответствующим сервисам, функционирующим на высокопроизводительных вычислительных системах.

С точки зрения пользователя, к особенностям создания и управления потоковой обработки сверхбольших объемов данных и извлечения из них знаний относят следующие пункты.

1) Создание типовых коротких WF направленных на массовый расчет сверхбольших объемов данных.
2) Разработка WF с использованием специализированных сервисов доступа к внешним системам обработки данных при реализации типовых WF .
3) Использование возможности интегрированной системы для распределенных вычислений Apache Hadoop, с целью анализа сверхбольших объемов данных

полученных путем краулинга содержимого web-страниц социальных сетей Интернета.

4) Использование данных полученных из реляционных СУБД через специализированные сервисы, реализованные в виде прикладных пакетов.
5) Использование принципов long running WF для длительных расчетов с возможностью корректировки параметров и этапов выполнения, не прерывая при этом самого процесса исполнения композитного приложения.
6) Оптимизация работы распределенного хранилища данных (репликации, оптимизация трафика инфраструктуры и т.д.).

3.2.6. Решение типовой задачи сбора и обработки данных с использованием МИТП-Д

В качестве характерного композитного приложения для МИТП-Д рассмотрим WF поиска латентного сообщества, связанного с потреблением наркотиков, в социальной сети LiveJournal с целью его анализа на предмет динамики и эффективности распространения информации (слухов). Результатом моделирования процесса распространения слухов на выявленной сети являются интервальные характеристики (минимальные и максимальные порядковые статистики, медиана, размах выборки) покрытия сети при заданной структуре. На основании этих характеристик возможно исследовать скорость распространения информации в определенной среде в целях обеспечения поддержки принятия решений при прогнозировании результатов утечки информации и волнений в данной маргинальной группе.

Для поиска лиц, связанных с потреблением наркотиков, в социальной сети и для получения топологической структуры сообщества используется программный компонент Hadrawler (в составе МИТП-Д). Для его запуска пользователю необходимо составить список ключевых слов и фраз, которые позволяют определять семантику текстов, и на основе этого выяснять, принадлежит ли узел социальной сети к группе людей, связанных с потреблением наркотиков. Экспертный список ключевых слов о наркотиках создается на основе статистических данных и представляет собой полный словарь различных терминов о наркотиках, способах их употребления и их эффектах. Наличие такого словаря позволяет сформировать как список начальных узлов, с которых краулер начнет обходить сеть, так и набор правил для фильтрации пользователей во время краулинга. Начальный список узлов формируется при помощи внешних поисковых систем.

Для моделирования процесса распространения слухов используется модель Далея-Кендалла, реализованная в программном компоненте ISM. Все множество вершин условно разделяется на три группы: неинформированные, информированные-распространители и невосприимчивые. На каждом шаге алгоритма моделирования производится расчет числа новообразованных распространяющих и невосприимчивых узлов. Если распространяющий узел контактирует с неинформированным, то с вероятностью λ второй узел также становится распространителем. Если распространяющий узел контактирует с распространяющим или невосприимчивым узлом, то он с вероятностью α становится невосприимчивым.

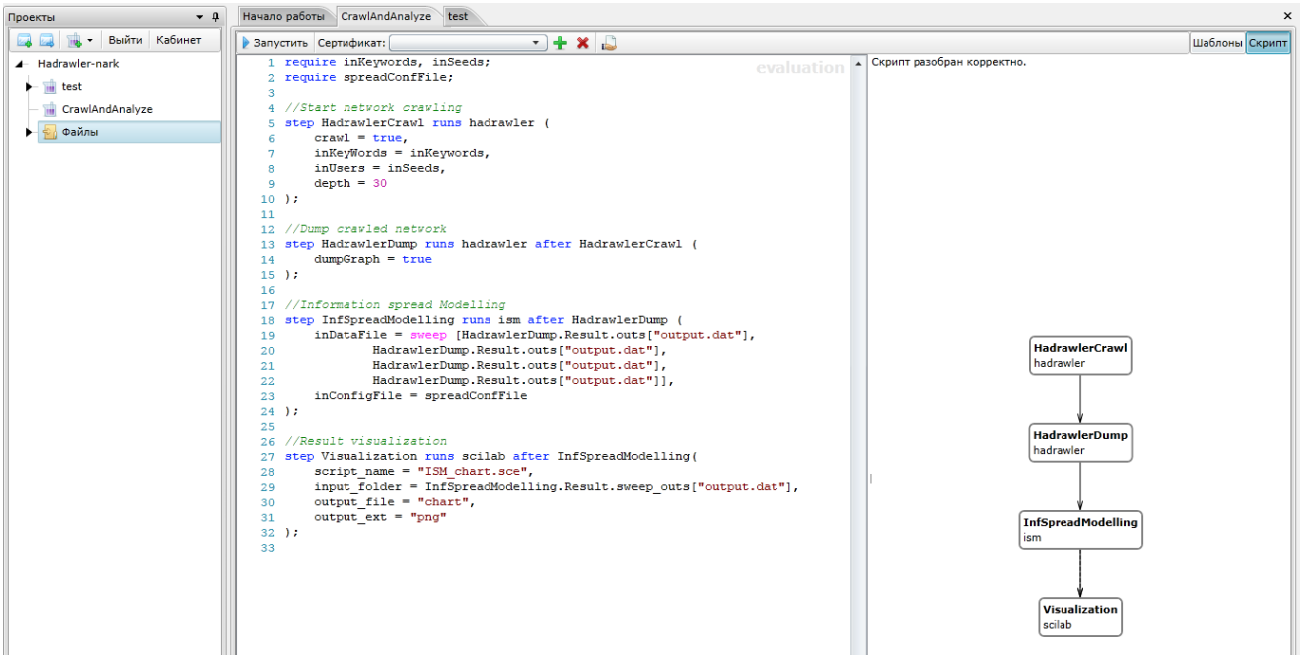
В листинге 3.1 приведен код соответствующего композитного приложения.

Листинг 3.1 – Код композитного приложения моделирования процессов распространения информации в виртуальных сообществах в Интернете, связанных с потреблением наркотиков

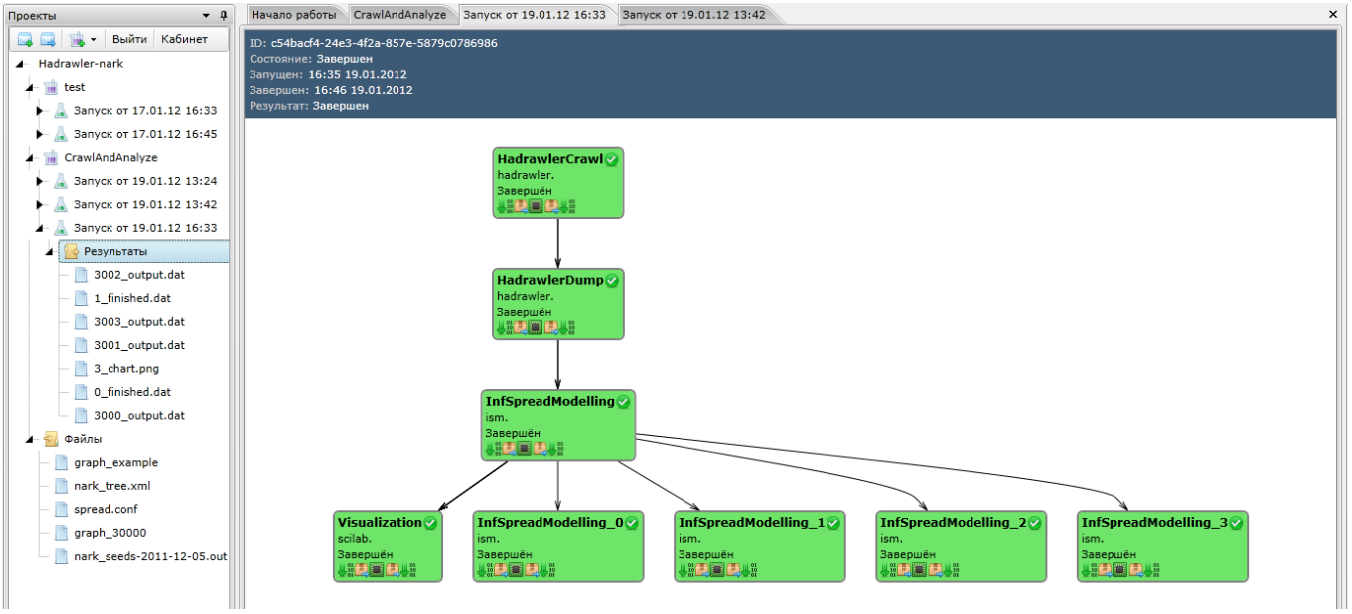
```
require inKeywords, inSeeds;
require spreadConfFile;
//Start network crawling
step HadrawlerCrawl runs hadrawler (
 crawl = true,
 inKeyWords = inKeywords,
 inUsers = inSeeds,
 depth = 30
);

//Dump crawled network
step HadrawlerDump runs hadrawler after HadrawlerCrawl (
 dumpGraph = true
);

//Information spread Modelling
step InfSpreadModelling runs ism after HadrawlerDump (
 inDataFile = sweep [HadrawlerDump.Result.outs["output.dat"],
   HadrawlerDump.Result.outs["output.dat"],
   HadrawlerDump.Result.outs["output.dat"],
   HadrawlerDump.Result.outs["output.dat"]],
 inConfigFile = spreadConfFile
);

//Result visualization
step Visualization runs scilab after InfSpreadModelling(
 script_name = "ISM_chart.sce",
 input_folder = InfSpreadModelling.Result.sweep_outs["output.dat"],
 output_file = "chart",
 output_ext = "png"
);
```

На рис. 3.7(а) приведено окно интерпретации композитного приложения в МИТП-Д с указанием всех шагов выполнения моделирования, а на рис. 3.7(б) – результат исполнения WF. Видно, что для снижения времени работы МИТП-Д проводит распараллеливание вычислений на этапе моделирования. Распараллеливание операции

краулинга выполняется внутренними средствами МИТП-Д и не отображается на WF. Для наглядности на рис. 3.7(б) приведены только 4 параллельных запуска блока моделирования.



(а)



(б)

Рисунок 3.7 – Разработка и исполнение композитного приложения моделирования процессов распространения информации в виртуальных сообществах в Интернете,

связанных с потреблением наркотиков, с использованием МИТП-Д: (а) окно интерпретации WF, (б) результат исполнения WF

На рис. 3.8 показаны результаты моделирования процесса распространения слухов на полученной в процессе краулинга сети из 30 000 вершин (ограничение, установленное пользователем при применении МИТП-Д). На графике отображаются минимальная, максимальная и усредненная величины покрытия сети информацией на определенный час. Видно, что наивысшая скорость распространения достигается для интервала 5–10 часов, а максимальное покрытие составляет порядка 22 000 членов сети (около 75 %) при 35 часах протекания процесса распространения.

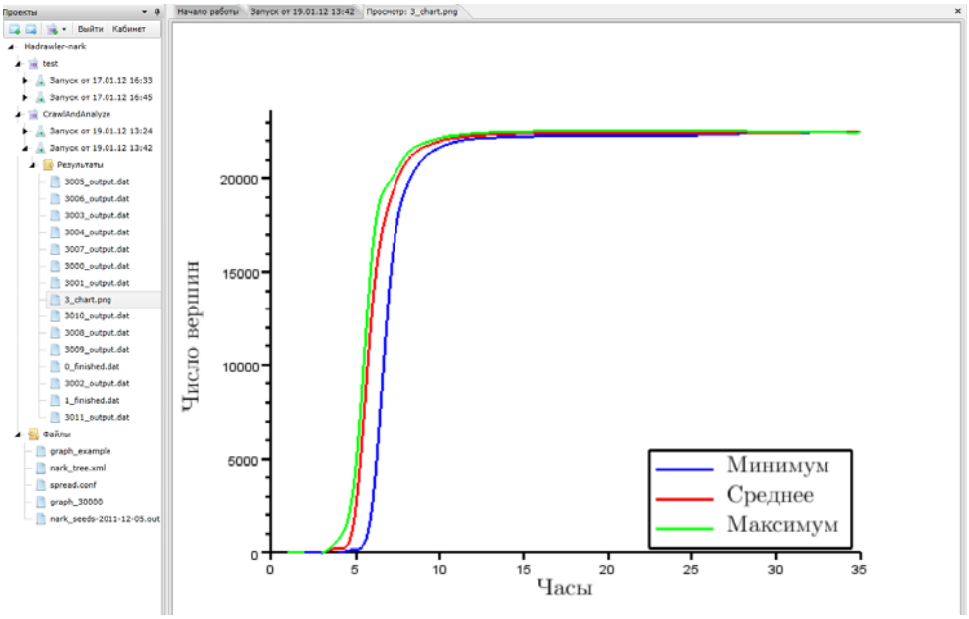


Рисунок 3.8 – Графики распространения информации по контактной сети сообщества LiveJournal, связанного с потреблением наркотиков

Подробное описание рассмотренной задачи и композитного приложения приведено в документе RU.СНАБ.80066-06 13 56.

4. ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ

Как видно из примера в разделе 3.2.5, для работы с МИТП-Д не требуются специальные виды входных данных. В ходе работы входными данными могут являться любые входные файлы, соответствующие по формату запускаемым прикладным сервисам (в текстовом, цифровом, графическом виде), а также данные, вводимые пользователем с клавиатуры по запросу сервиса, или в скрипте композитного приложения. В случае

несоответствия данных условиям их использования будет выдано соответствующее системное сообщение. Для приведения к общему формату используется технология описания на основе языка EasyPackage (раздел 3.2.1).

В качестве выходных данных МИТП-Д предоставляет результаты расчетов, загруженные с удаленного хранилища CLAVIRE/Storage RU.СНАБ.80066-06 01 25 (в форме текстового, графического или цифрового файла, размещаемого в директории, указываемой пользователем через соответствующее диалоговое окно). Формат файла соответствует тому сервису, посредством которого был произведен расчет. Для обеспечения единого формата в целях унификации процесса передачи данных между сервисами используется технология описания на основе языка EasyPackage (раздел 3.2.1).

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ

МИТП	Многопрофильная инструментально-технологическая платформа
МИТП-Д	Технологическая платформа потоковой обработки сверхбольших объемов данных и извлечения из них знаний на основе облачных вычислений (на основе CLAVIRE)
ОС	Операционная система
ПАК	Программно-аппаратный комплекс
СУБД	Система управления баз данных
ЭВМ	Электронная вычислительная машина
AaaS	Application as a Service, модель облачных вычислений
AWF	Абстрактный WF
CLAVIRE	Cloud Applications Virtual Environment, наименование МИТП
CWF	Конкретный WF
DSL	Domain Specific Language, предметно-ориентированный язык
iPSE	Intelligent Problem Solving Environment, концепция
MWF	Мета-WF
SaaS	Software as a Service, модель облачных вычислений
WF	Поток заданий, workflow

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

Лист регистрации изменений									
Изм.	Номера листов (страниц)				Всего листов (страниц) в докум.	№ докумен та	Входящий № сопроводи-тельного докум. и дата	Подп.	Дата
	изме-ненных	заме-ненных	но-вых	аннули-рован-ных					