



Inpolus ETL

**Инструменты интеграции и управления
потоками данных**

127287, Москва, 2-я Хуторская 38А, стр. 9
+7 (495) 274-01-91
info@inpolus.ru
www.inpolus.ru

Задача: управление разнородными потоками данных в крупных компаниях

- Много разнородных источников/получателей данных с разными структурами
- Разнотипные процессы обработки данных (потокковая и пакетная обработка)
- Использование разнородных инструментов обработки данных: ETL/ELT-инструменты, пользовательские приложения, SQL-скрипты с прямым доступом к БД
- Сложность мониторинга результатов и этапов обработки/анализа данных
- Периодически потери или дублирование данных в процессе
- Периодические проблемы с производительностью и масштабируемостью решений
- Отсутствие единого контроля доступа ко всем сервисам
- Необходимость использовать большой набор разнородных инструментов, в т.ч. Legacy, что повышает требования к специалистам и увеличивает длительность “периода вхождения” для новичков



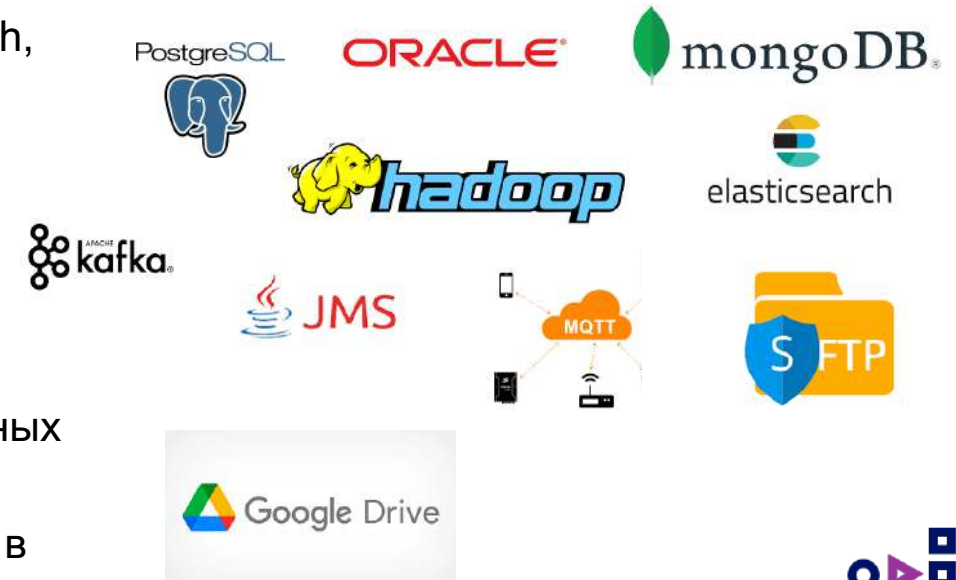
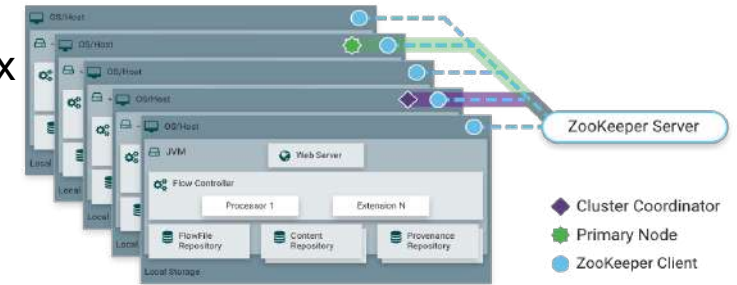
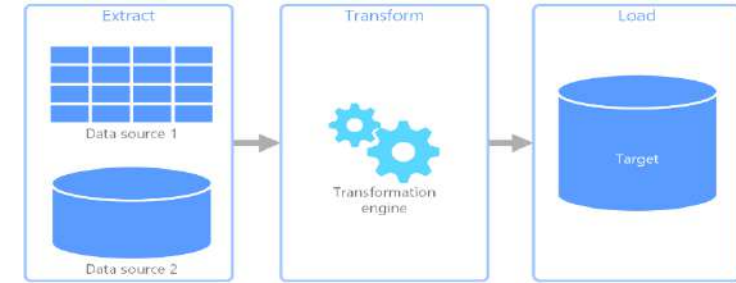
Решение: Inpolus ETL

- Надежный ETL/ELT инструмент на базе широко используемых и отказоустойчивых Open Source решений
- Предоставляет возможность обработки произвольных потоков данных
- Обеспечивает мониторинг выполнения процессов и информирует о сбоях в системе
- Повышает прозрачность работы за счет сбора и анализ метрик и логов
- Предоставляет гибкую модель разграничения прав доступа на основе ролевой модели и интеграции с корпоративными Identity Providers
- Low Code – решение, позволяющее управлять и создавать процессы обработки данных через удобный визуальный интерфейс
- Позволяет разрабатывать расширения функциональности, подключаемые по стандартному интерфейсу



Inpolus ETL - назначение инструмента

- ✓ Платформа для визуального проектирования и исполнения ETL/ELT-процессов – извлечения, трансформации, обогащения и загрузки данных, позволяющая интегрировать произвольные системы хранения и обработки данных
- ✓ Поддерживает кластерную конфигурацию и параллельную обработку данных
- ✓ Оптимизирована под быструю потоковую обработку данных
- ✓ Предоставляет встроенную поддержку разнородных источников и потребителей данных:
 - Реляционные БД
 - Распространенные СУБД NoSQL: MongoDB, Cassandra, ElasticSearch, DynamoDB
 - Продукты экосистемы Hadoop: HBase, HDFS
 - Apache Kafka, JMS, MQTT
 - FTP/SFTP
 - Внешние сервисы хранения – Google Drive, DropBox
 - Возможность подключить нестандартный источник/потребитель данных
- ✓ Позволяет расширять стандартные механизмы обработки данных при помощи подключения пользовательских/сторонних обработчиков данных в проектируемые ETL/ELT-процессы



Inpolus ETL – обзор принципов работы

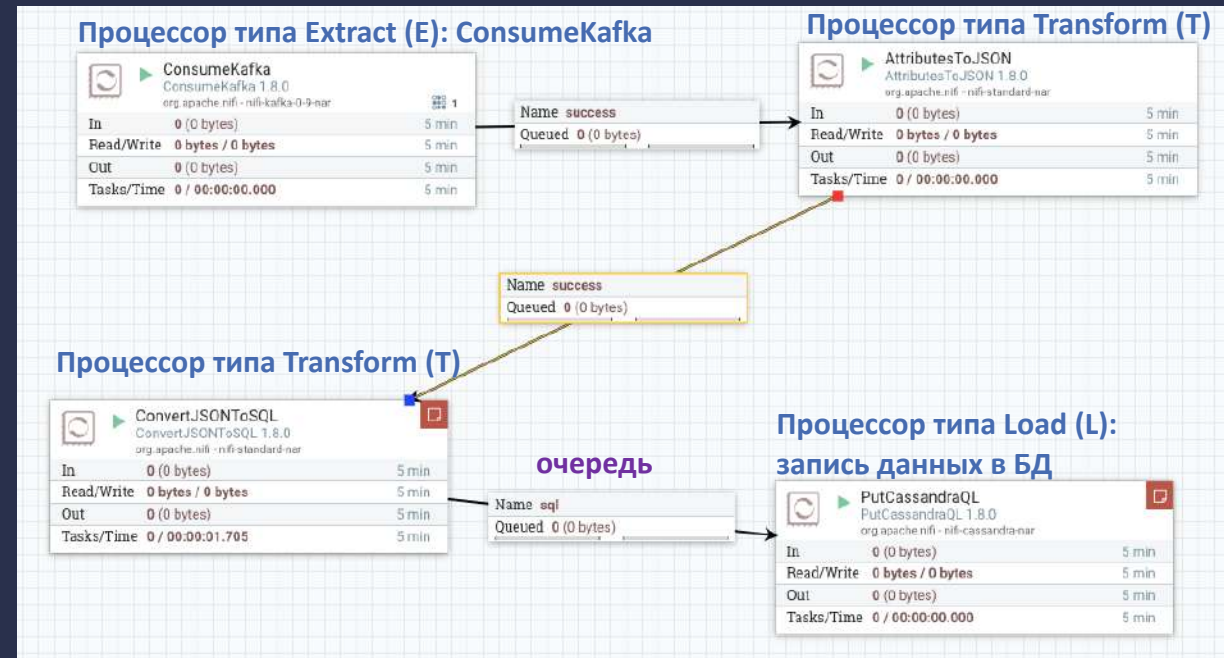
Каждый проектируемый ETL/ELT процесс состоит из следующих компонентов:

- Процессоры (обработчики): выполняют один шаг процесса – операцию по загрузке, преобразованию/обогащению данных или отправку данных потребителю
- Очереди: связывают два процессора (шага обработки), используются для организации передачи данных между шагами
- FlowFile: объект потока данных – помещается в очередь и передается между процессорами, содержит атрибуты и ссылку на сами обрабатываемые данные (контент)

Платформа предоставляет порядка 200 встроенных процессоров следующих типов:

- Загрузка данных из источников;
- Отправка данных потребителям (запись)
- Работа с БД: чтение и запись (выделены в отдельный тип)
- Трансформация данных
- Объединение/разделение потоков данных

Пример процесса: чтение данных из Kafka topic, их преобразование и запись в БД

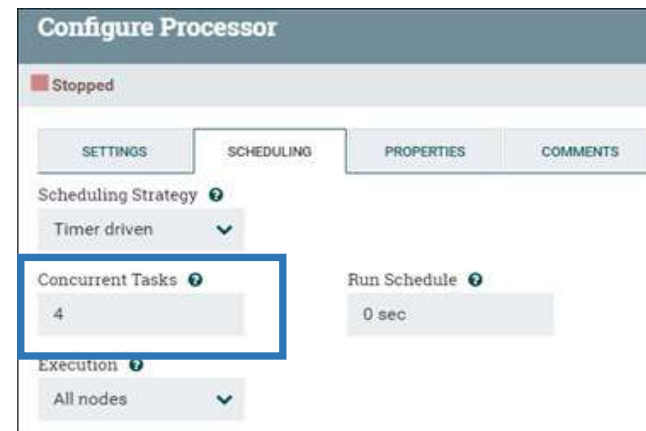
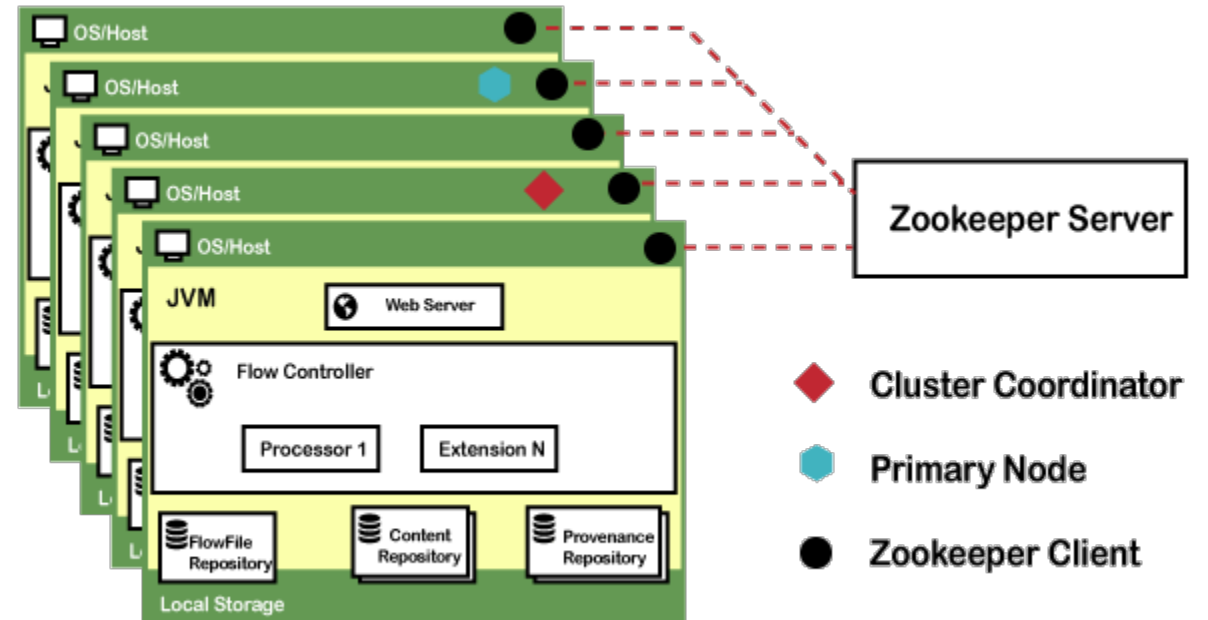


- Процессы маршрутизации – выбор маршрута обработки в зависимости от условий (значений атрибутов операции и самих данных)
- Процессоры взаимодействия с системой (вызов скриптов, других процессов, команд ОС)



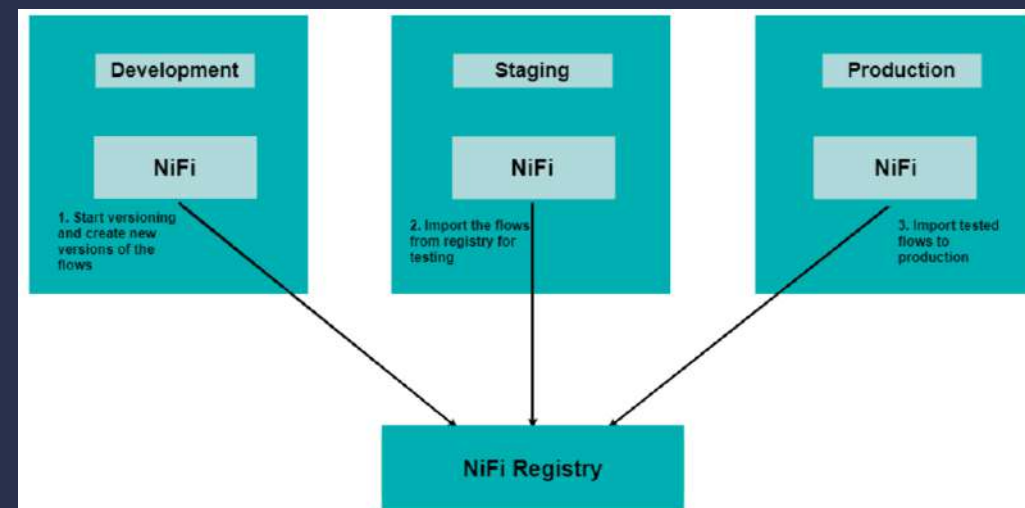
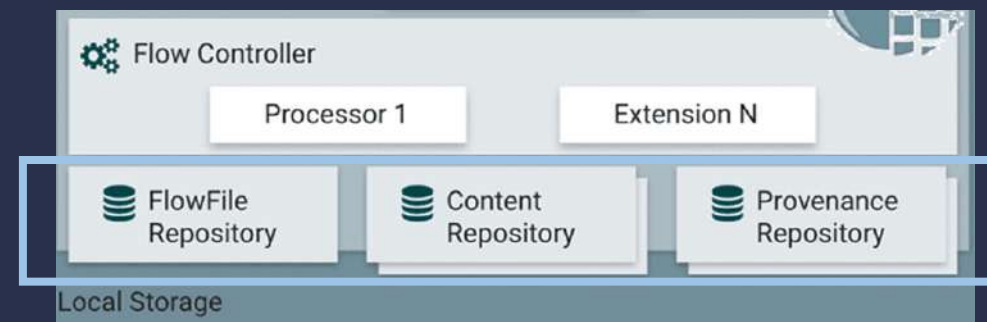
Возможность распределенной обработки данных

- ✓ Inpolus ETL предоставляет возможность использовать **кластерную конфигурацию** с балансировкой нагрузки между узлами кластера, что позволяет распараллеливать обработку данных (как пример – вычитывание выделенным узлом большого объема из источника, разбиение данных на несколько потоков и их дальнейшая параллельная обработка на разных узлах)
- Поддерживаемые **стратегии балансировки нагрузки**:
 - Round-Robin – равномерное распределение обрабатываемых данных по узлам
 - Partition by attribute – распределение по узлам в зависимости от значения атрибута в объекте данных (FlowFile)
- ✓ Помимо распределения нагрузки по разным узлам кластера есть возможность распараллелить обработку в рамках одного процессора, сконфигурировав обработку данных в несколько потоков (Concurrent tasks).



Отказоустойчивость и надежность

- ✓ Inpolus ETL хранит информацию о состоянии процесса обработки данных (включая историю изменений состояния процесса) в трех репозиториях. В случае аварийного завершения работы и последующего перезапуска процессы рестартуют с того шагов обработки, на которых произошел сбой
 - FlowFile Repository - репозиторий файлов потоков данных, хранит информацию о последнем консистентном состоянии потока
 - Content Repository – репозиторий непосредственно самих обрабатываемых данных
 - Provenance Repository – “репозиторий происхождения”, хранит историю изменений файла потока (FlowFile)
- ✓ Инструмент предоставляет встроенные процессоры **поиска и устранения дубликатов** в данных (DetectDuplicate, DeduplicateRecord)
- ✓ Модуль Registry обеспечивает версионирование метаданных процессов и возможность “откатиться” к одной из предыдущих версий, а также перенос конфигураций процессов между средами



Логирование и мониторинг

✓ Благодаря использованию репозитория, в частности – Provenance Repository, есть возможность получить **состояние потоков данных** практически на любой момент времени

✓ Поддерживаются **генерация метрик** в процессе выполнения и возможность экспорта отчета по метрикам

✓ Inpolus ETL обеспечивает **мониторинг выполнения процессов** как и из собственного интерфейса (UI), так и позволяет выгружать **отчеты по мониторингу** и интегрироваться с такими распространенными инструментами, как Prometheus, DataDog и т.п.

✓ **Статистика** собирается и отображается в UI на **разных уровнях**:

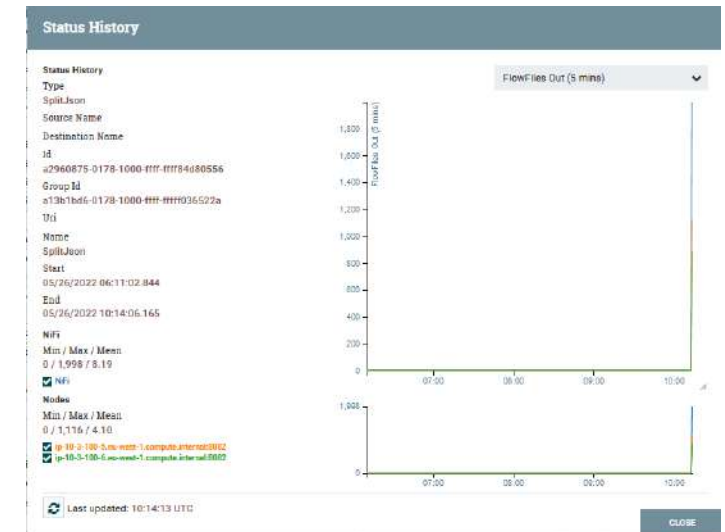
- Кластера в целом
- Конкретного процессора
- Конкретного потока данных (FlowFile)

✓ Также доступен общий **журнал системных событий** (так называемая Bulletin Board – “доска объявлений”), содержащий информацию обо всех ошибках и предупреждениях, сгенерированных процессорами Inpolus ETL

Аннотации к скриншоту:

- Индикатор состояния процессора (красный квадратик в углу)
- Имя процессора (PutSQL.SaveLogToDB)
- Индикатор ошибки (Bulletin Board) (красный квадратик в углу)
- Тип процессора (иконка процессора)
- Статистика (1 раз в 5 минут) (таблица с метриками)

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.271	5 min



Общая статистика по кластеру:

- 2 / 2: Количество узлов кластера подключенных/ всего
- 0: Количество активных потоков
- 56,795 (14.62 GB): количество и объем потоковых файлов в очередях
- 0: Удаленная группа процессов (Передача отключена)
- 0: Удаленная группа процессов (Передача включена)
- 63: Количество запущенных компонентов
- 81: Количество остановленных компонентов
- 4: Количество некорректных компонентов
- 102: Количество отключенных компонентов
- 3: Версионированные группы процессов локально измененные
- 1: Версионированные группы процессов в актуальном
- 0: Доступна более новая версия
- 0: локальные изменения и доступна новая версия
- 0: Ошибка синхронизации
- 08:27:15 UTC: Время последнего обновления

Общая статистика по кластеру



Безопасность

- ✓ Для аутентификации и разграничения доступа Inpolus ETL предоставляет ролевую модель и поддерживает интеграцию с LDAP и Kerberos
- ✓ Шифрование данных обеспечивается при помощи криптографических процессоров, реализующих стандарт OpenPGP

Policy	Action
Component policy for process group NiFi Flow	read
Component policy for process group NiFi Flow	write
Component policy for processor GetFile	read
Component policy for processor GetFile	write
Global policy to access all policies	write
Global policy to access all policies	read
Global policy to access restricted components	write
Global policy to access the controller	write
Global policy to access the controller	read
Global policy to access users/user groups	read
Global policy to access users/user groups	write
Global policy to view the user interface	read

Some policies may be inherited by descendant components unless explicitly overridden.

CLOSE

EncryptContent
org.apache.nifi - nifi-standard-nar

In	1 (2.31 KB)	5 min
Read/Write	2.31 KB / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	1 / 00:00:00.030	5 min

Configure Processor | EncryptContent 1160

Invalid

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

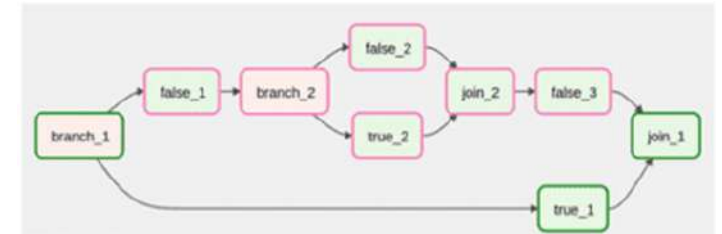
Property	Value
Mode	Encrypt
Key Derivation Function	NIFI Legacy KDF
Encryption Algorithm	MDS_128AES
Allow insecure cryptographic modes	Not Allowed
Password	No value set
Raw Key (hexadecimal)	No value set
Public Keyring File	No value set
Public Key User Id	No value set
Private Keyring File	No value set
Private Keyring Passphrase	No value set
PGP Symmetric Cipher	AES_128

CANCEL APPLY

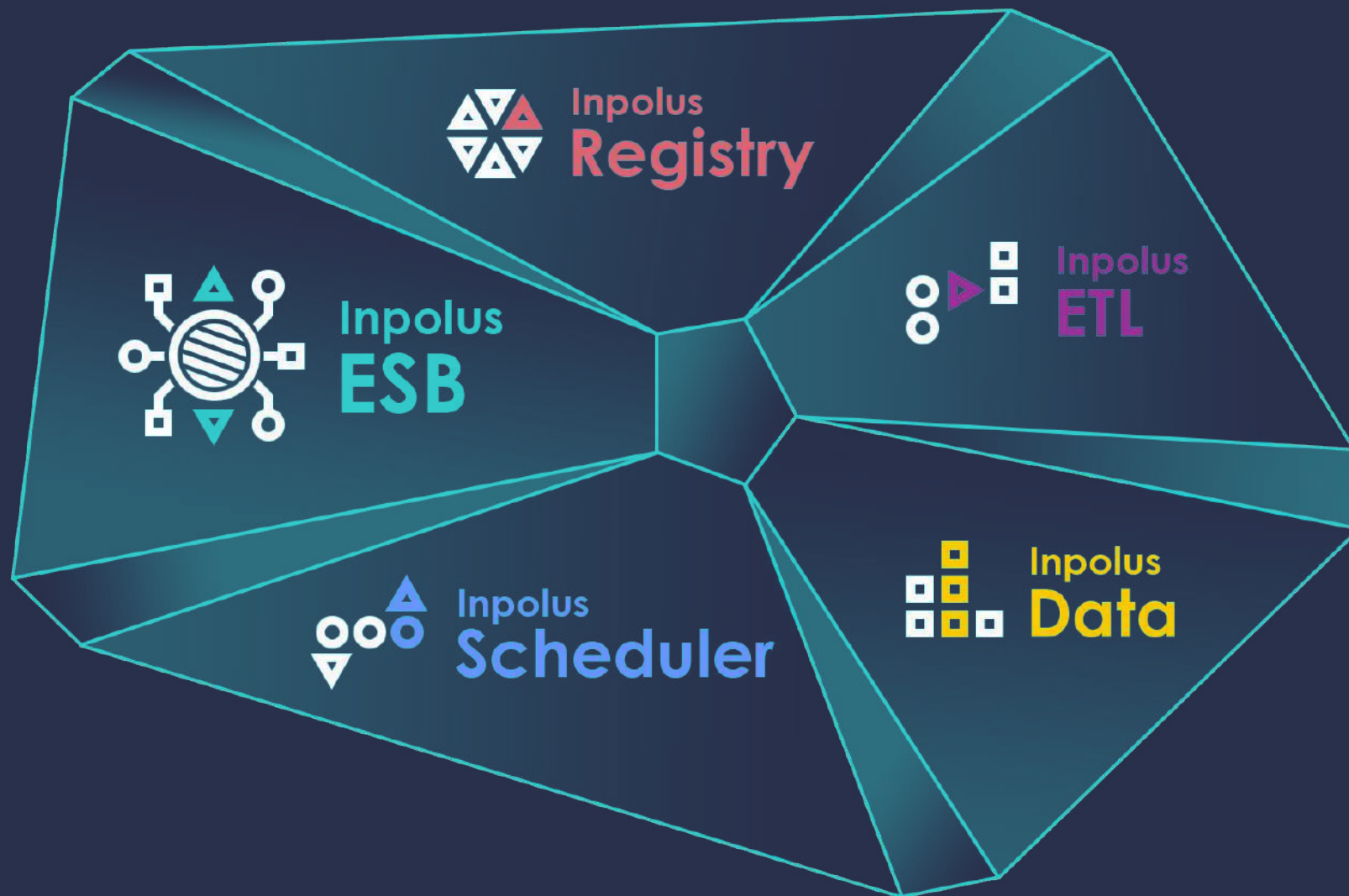


Дополнительный модуль Inpolus ETL - оркестратор

- ✓ Inpolus ETL - это расширяемая платформа, предоставляющая понятные средства проектирования ETL/ELT-процессов и включающая средства мониторинга выполнения процессов; часто используется в задачах обработки больших данных;
- ✓ Основная задача ETL/ELT-платформы – трансформация данных и передачи их между информационными системами, для большинства ETL/ELT-процессов достаточно одного базового инструмента;
- ✓ В некоторых случаях может потребоваться спроектировать более сложный процесс, состоящий из подпроцессов, выполняемых разными средствами, для этого можно использовать дополнительный модуль - **оркестратор**;
- ✓ Для оркестровки потоков операций используется представление в виде направленного ациклического графа (DAG); собранная в граф группа операций может запускаться событию или по расписанию;
- ✓ С учетом дополнительного модуля платформа обеспечивает полноценные возможности оркестрации, проектирования, запуска и мониторинга выполнения практически произвольных ETL/ELT-процессов и может быть использована в задачах обработки больших данных.

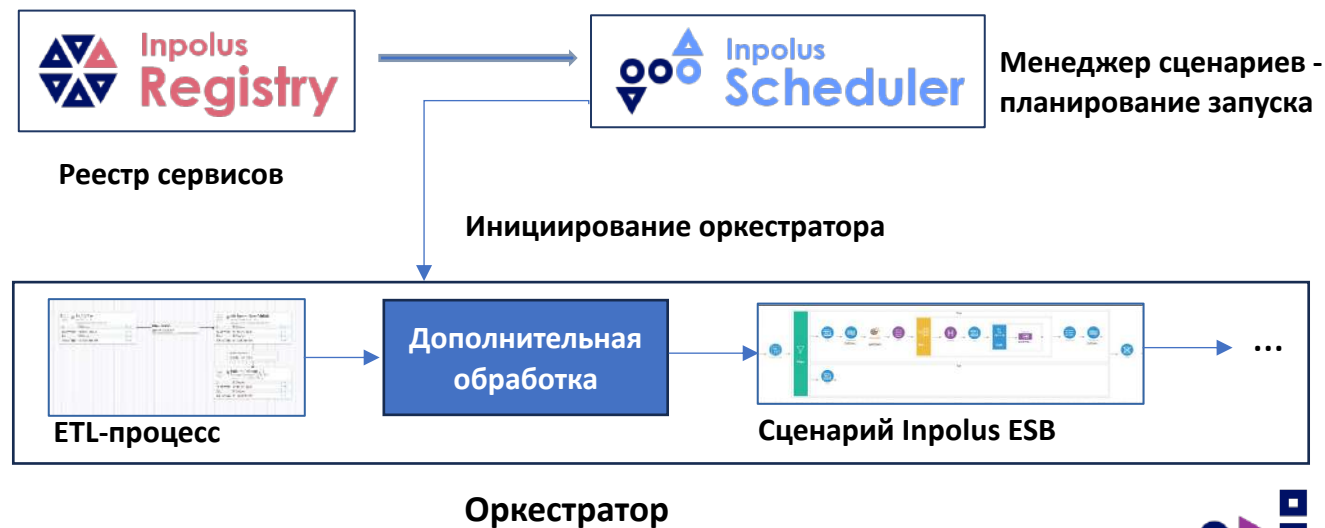
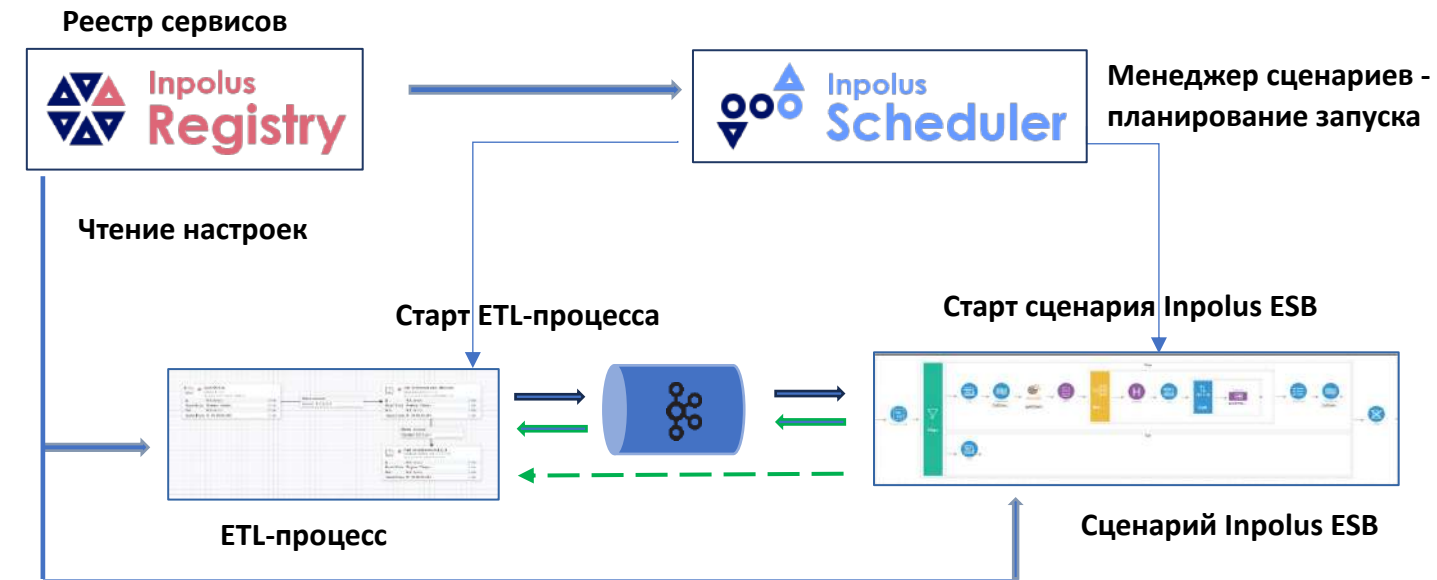


Взаимодействие ETL и других продуктов INPOLUS Integration Platform



Взаимодействие ETL и других продуктов INPOLUS Integration Platform

- Вся конфигурационная информация как для ETL-процессов, так и сценариев Inpolus ESB хранится в **Реестре сервисов**
- Запуск как ETL-процессов, так и сценариев Inpolus ESB может быть запланирован в **Менеджере сценариев**
- При необходимости завершение ETL-процесса Inpolus ETL может инициировать вызов сценария шины Inpolus ESB (например, Inpolus ETL помещает полученные и преобразованные данные в Kafka-топик, который “слушает” процесс Inpolus ESB)
- Сценарий Inpolus ESB также может инициировать вызов ETL-процесса по событию (через HTTP/S listener, публикацию в Kafka/RabbitMQ)
- В сложных случаях взаимодействия возможно использование дополнительного модуля - **оркестратора**

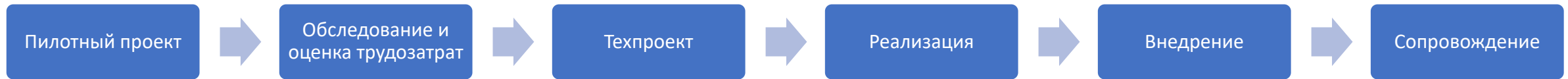


Inpolus ETL - выводы

- ✓ Inpolus ETL – полнофункциональный инструмент для реализации разнотипных ETL/ELT-процессов:
 - Поточковой передачи данных между системами
 - Агрегирования данных из разных источников для дальнейшей трансформации и записи в хранилище
 - Взаимодействия с BigData – инструментами, например – экосистемой Hadoop
- ✓ Часть **общей интеграционной платформы INPOLUS**
- ✓ Ориентирован, в первую очередь, на потоковую, а не на пакетную передачу данных, тем не менее обработка большого пакета данных также возможна при корректном распараллеливании обработки
- ✓ Возможность **параллельной обработки** данных как на уровне кластера в целом, так и при помощи параллельных потоков, выполняющих обработку в рамках конкретного процессора
- ✓ Поддержка порядка **200 встроенных процессоров** загрузки, трансформации/обогащения и записи данных
- ✓ Возможность **визуального проектирования** интеграционных процессов
- ✓ Возможность реализации **собственных процессоров**
- ✓ Возможность подключать **внешний код** (Скрипты, Внешние процессы, Команды ОС)
- ✓ Встроенные **средства мониторинга**
- ✓ Встроенная поддержка **шифрования и внешних Identity Managers** на базе LDAP или Kerberos



Подход компании Инполюс к выполнению проектов



- ✓ На этапе **пилотного проекта** специалисты компании реализуют один из реальных ETL/ELT-процессов, используемых у заказчика, при помощи инструмента Inpolus ETL;
- ✓ На этапе **обследования** производится анализ существующих ETL/ELT-процессов Заказчика, инфраструктуры и дополнительных требований по расширению существующих механизмов обработки данных, а также проводится оценка трудозатрат; на данном этапе необходимо интенсивное вовлечение представителей Заказчика;
- ✓ В случае заключения договора на этапе **Техпроекта** формируются следующие документы:
 - ✓ Описание архитектуры решения;
 - ✓ Уточненное описание функциональных и нефункциональных требований;
 - ✓ Детальный план реализации и внедрения;
 - ✓ Требования к инфраструктуре;
 - ✓ Программа и методика испытаний (ПМИ);
 - ✓ Описание процесса управления изменениями;
- ✓ После согласования документов Техпроекта осуществляется **реализация**:
 - ✓ развертывание инструментов на предоставленных средах заказчика (как правило - DEV, TEST, PROD, но возможно уточнение на этапе техпроекта);
 - ✓ доработка платформы (при необходимости);
 - ✓ реализация процессов Заказчика;
- ✓ На этапе **внедрения** осуществляются:
 - ✓ Опытная эксплуатация решения;
 - ✓ Миграция существующих процессов на новую платформу (при необходимости);
 - ✓ Проведение обучения ключевых специалистов заказчика;
 - ✓ Приемка решения на основании ПМИ, утвержденной Заказчиком;
- ✓ В рамках отдельного договора на сопровождение команда специалистов Inpolus выполняет работы по поддержке внедренного решения.



Почему Inpolus ETL?

- Инструмент создан на основе одних из наиболее часто используемых Open Source продуктов для реализации ETL/ELT-процессов
- Является частью общей интеграционной платформы INPOLUS
- Интегрируется с разнотипными и часто используемыми источниками данных (СУБД – реляционные и NoSQL, Kafka, RabbitMQ, JMS, экосистема Hadoop, FTP/SFTP, MQTT и др.)
- Может быть использован как составляющая BigData-решений
- Кластерная конфигурация позволяет масштабировать решение
- Все исходные коды и компоненты, использованные в нашем решении, размещаются на собственных ресурсах на территории России
- Мы в компании Инполус проделали большую работу по анализу и изучению продуктов и готовы предоставить нашу версию популярного решения



О компании

ООО «Инполюс» - российская ИТ компания, с 2009 года поставляет решения, услуги и программное обеспечение в области консалтинга, ИТ сервисов и интеграционных технологий со специализацией по следующим направлениям:

Внедрение и сопровождение собственной корпоративной интеграционной (сервисной) платформы Inpolus Integration Platform (Enterprise Service Bus, ESB) , которая позволяет централизованно управлять жизненным циклом сервисов предприятия и обеспечивает гарантированное взаимодействие всех компонентов информационных систем организации.

Разработка специализированных прикладных сервисов и систем, базирующиеся на принципах сервис-ориентированной архитектуры (Service Oriented Architecture, SOA), в том числе систем автоматизации бизнес-процессов (Business Process Management System, BPMS).

Решения в области Искусственного Интеллекта.

ИТ консалтинг и разработка приложений на заказ (JAVA, JavaScript, .NET, PHP).

Разработка, внедрение и поддержка веб-проектов (HTML, CSS, JavaScript, Java, PHP, Symfony, Twig, Doctrine, произвольные SQL-БД).

Оптимизация ИТ инфраструктуры. Разработка и внедрение технологических стандартов.

Решения в области импортозамещения.

Outstaffing, Outsourcing.

Поставки лицензионного программного обеспечения.

Оказание услуг технической поддержки.

15 лет успеха!



Больше информации:



<https://www.inpolus.ru/solutions>

