ETL Polus

# Инструменты интеграции и управления потоками данных

ООО «Инполюс»
127287, Москва, 2-я Хуторская 38А, стр. 9
Тел./Факс: +7 495 274 01 91
E-Mail: info@inpolus.ru
www.inpolus.ru

# **Задача:** управление разнородными потоками данных в крупных компаниях

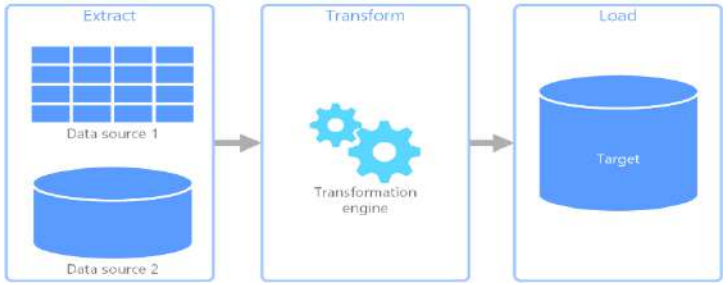
- Много разнородных источников/получателей данных с разными структурами
- Разнотипные процессы обработки данных (потоковая и пакетная обработка)
- Использование разнородных инструментов обработки данных: ETL/ELT-инструменты, пользовательские приложения, SQL-скрипты с прямым доступом к БД
- Сложность 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: Polus ETL

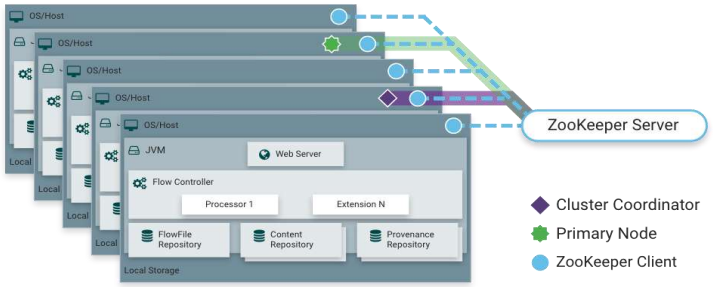
- Надежный ETL/ELT инструмент на базе широко используемых и  отказоустойчивых Open Source решений
- Предоставляет возможность обработки  произвольных потоков данных
- Обеспечивает мониторинг выполнения процессов и информирует о сбоях в системе
- Повышает прозрачность работы за счет сбора и анализ метрик  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

ETL Polus

# Polus 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




PostgreSQL ORACLE mongoDB

hadoop elasticsearch

kafka JMS MQTT SFTP

Google Drive

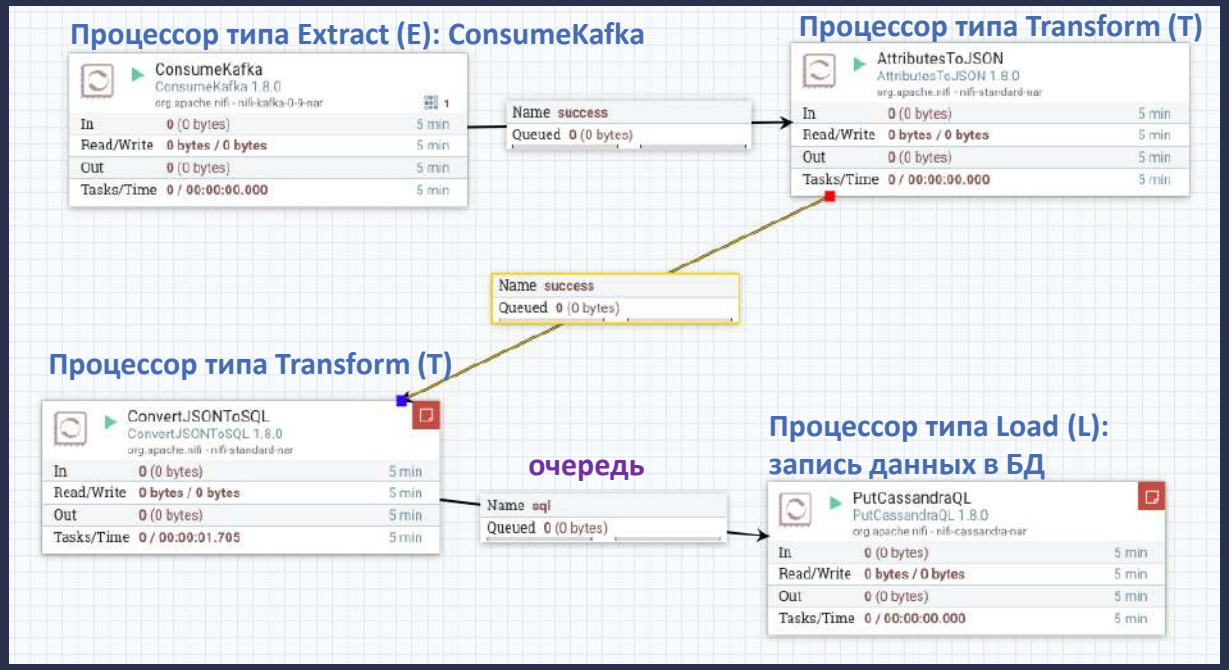
ETL Polus

# Polus 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)

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




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
- 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)

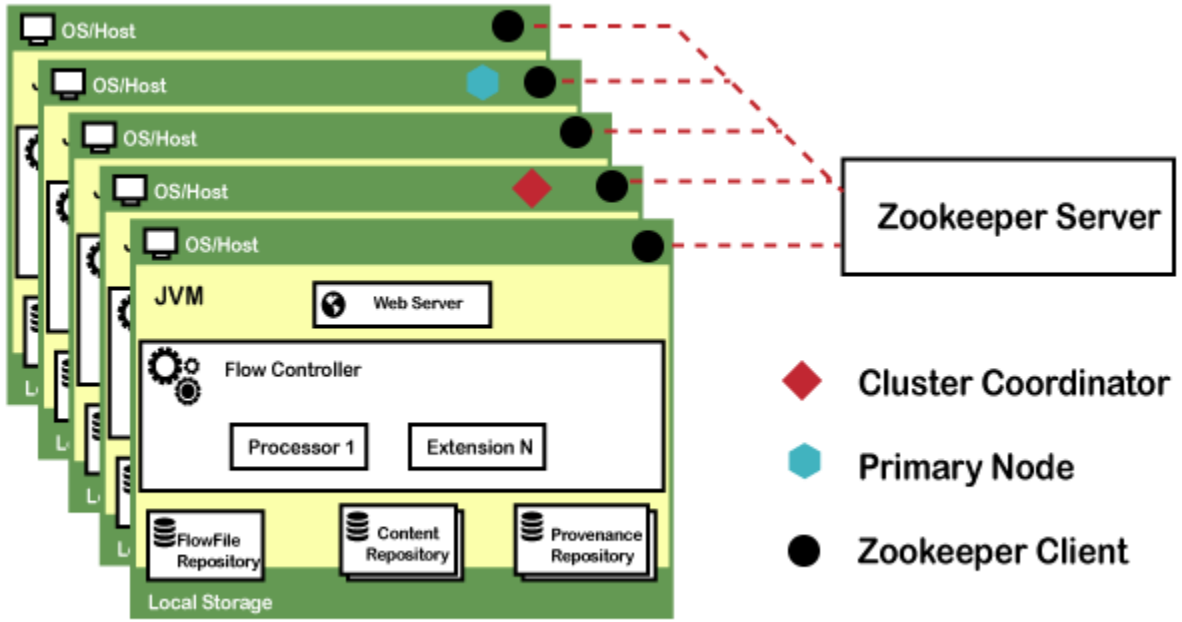
ETL Polus

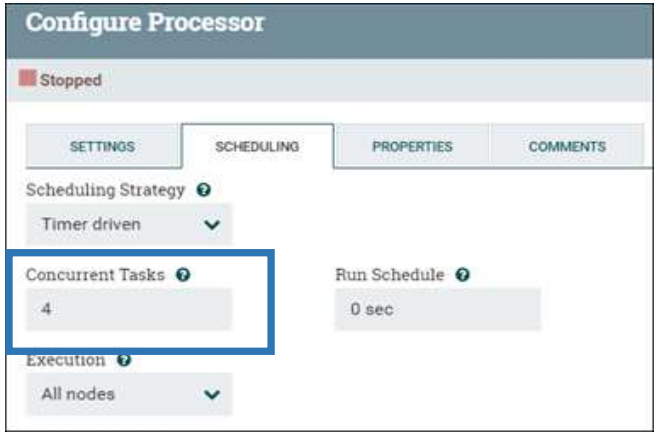
# Возможность распределенной обработки данных

- ✓ Polus 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).

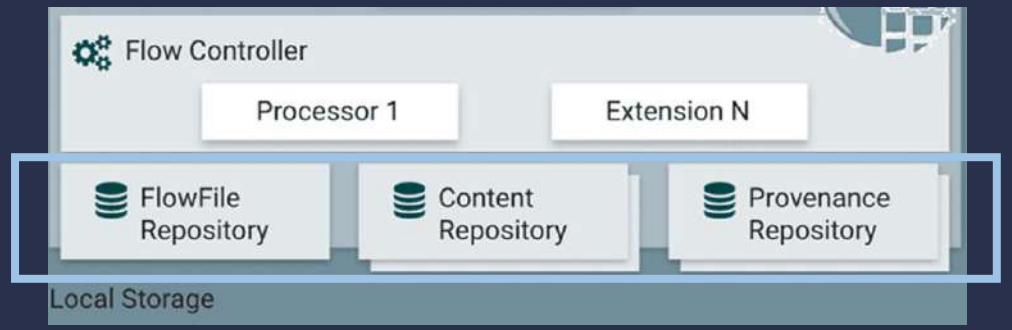


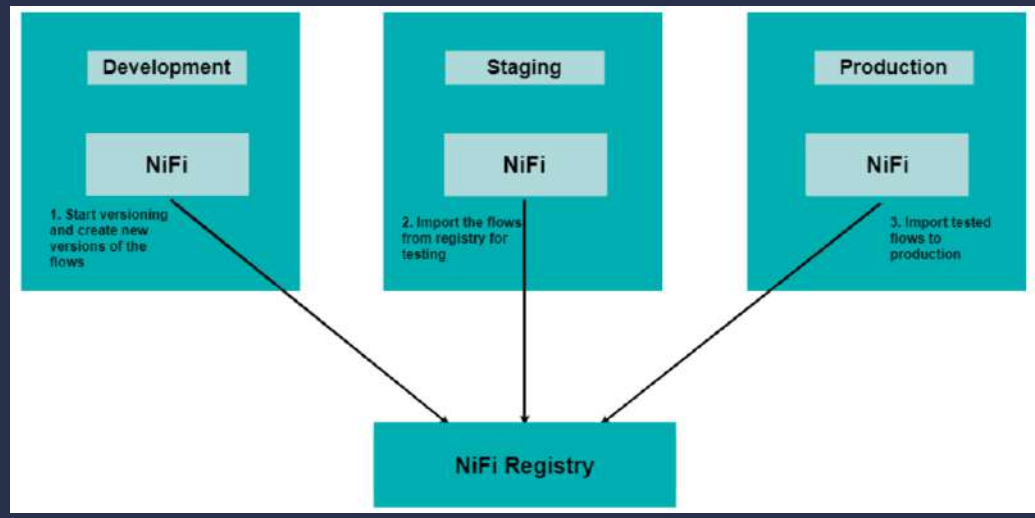




ETL Polus

# Отказоустойчивость и надежность

- ✓ Polus 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

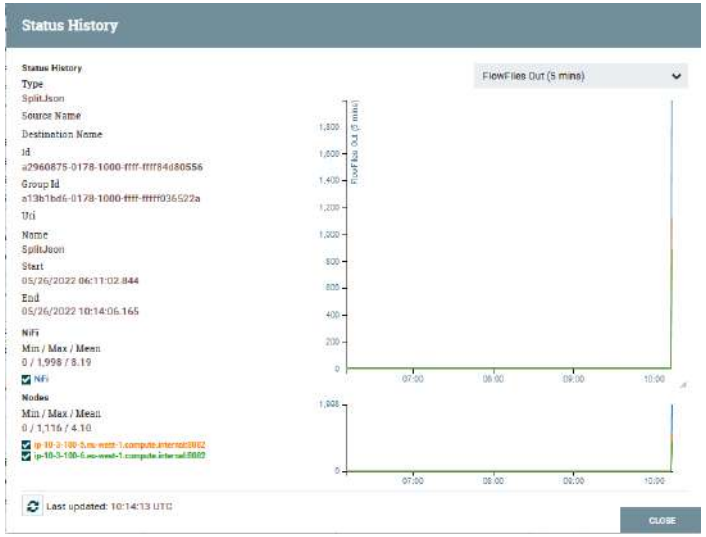


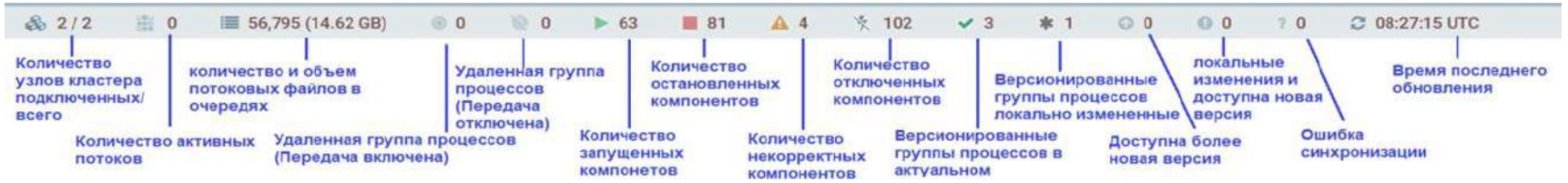




ETL Polus

# Логирование и мониторинг

- ✓ Благодаря использованию репозиториев, в частности – Provenance Repository, есть возможность получить **состояние потоков данных** практически на любой момент времени
- ✓ Polus ETL поддерживает **генерацию метрик** в процессе выполнения и возможность экспорт отчета по метрикам
- ✓ Polus ETLобеспечивает **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 Polus ETL

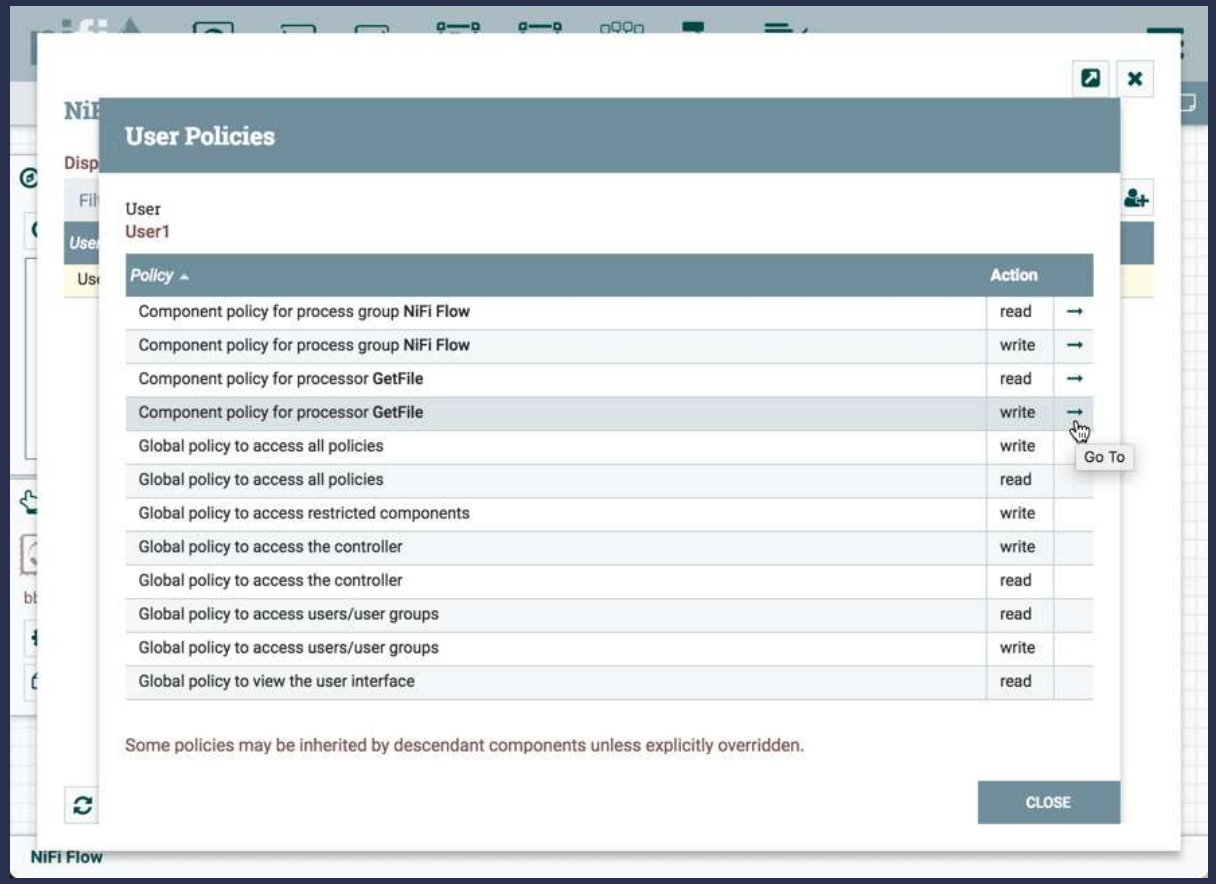



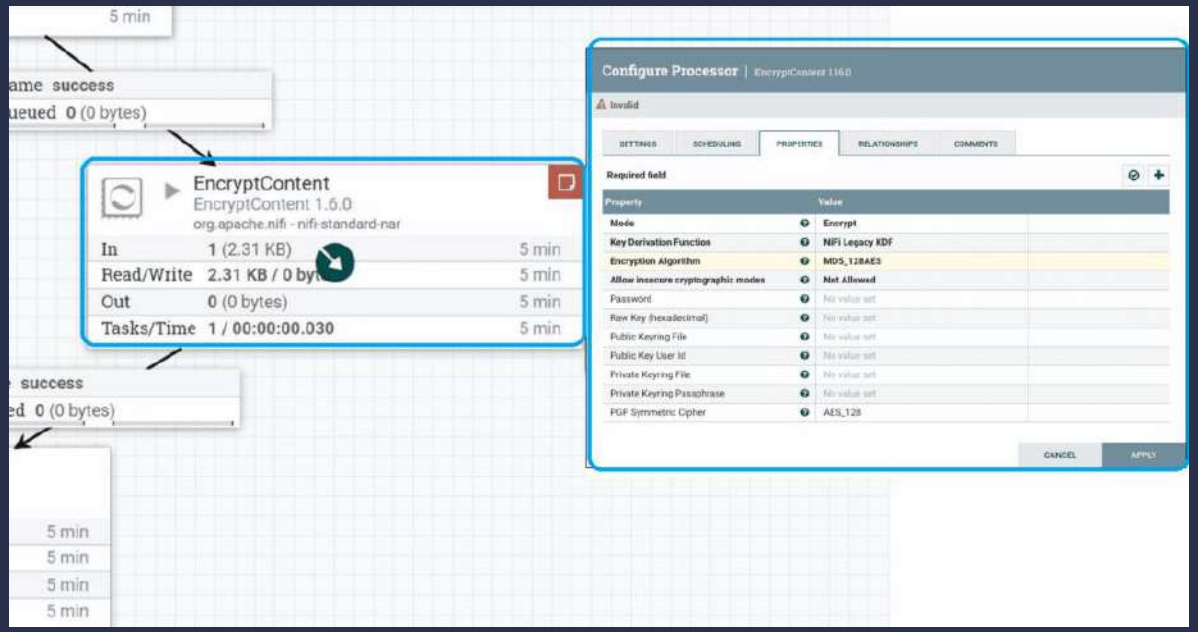

**Общая статистика по кластеру**

ETL Polus

# Безопасность

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




ETL Polus

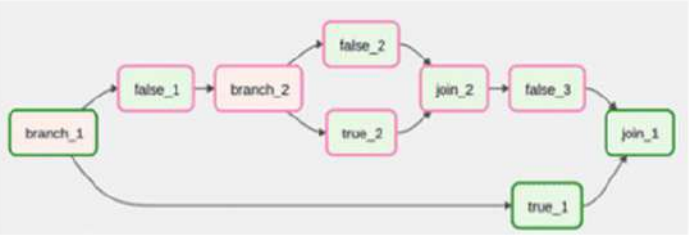
# Дополнительный модуль Polus ETL - оркестратор

- ✓ Polus 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**;

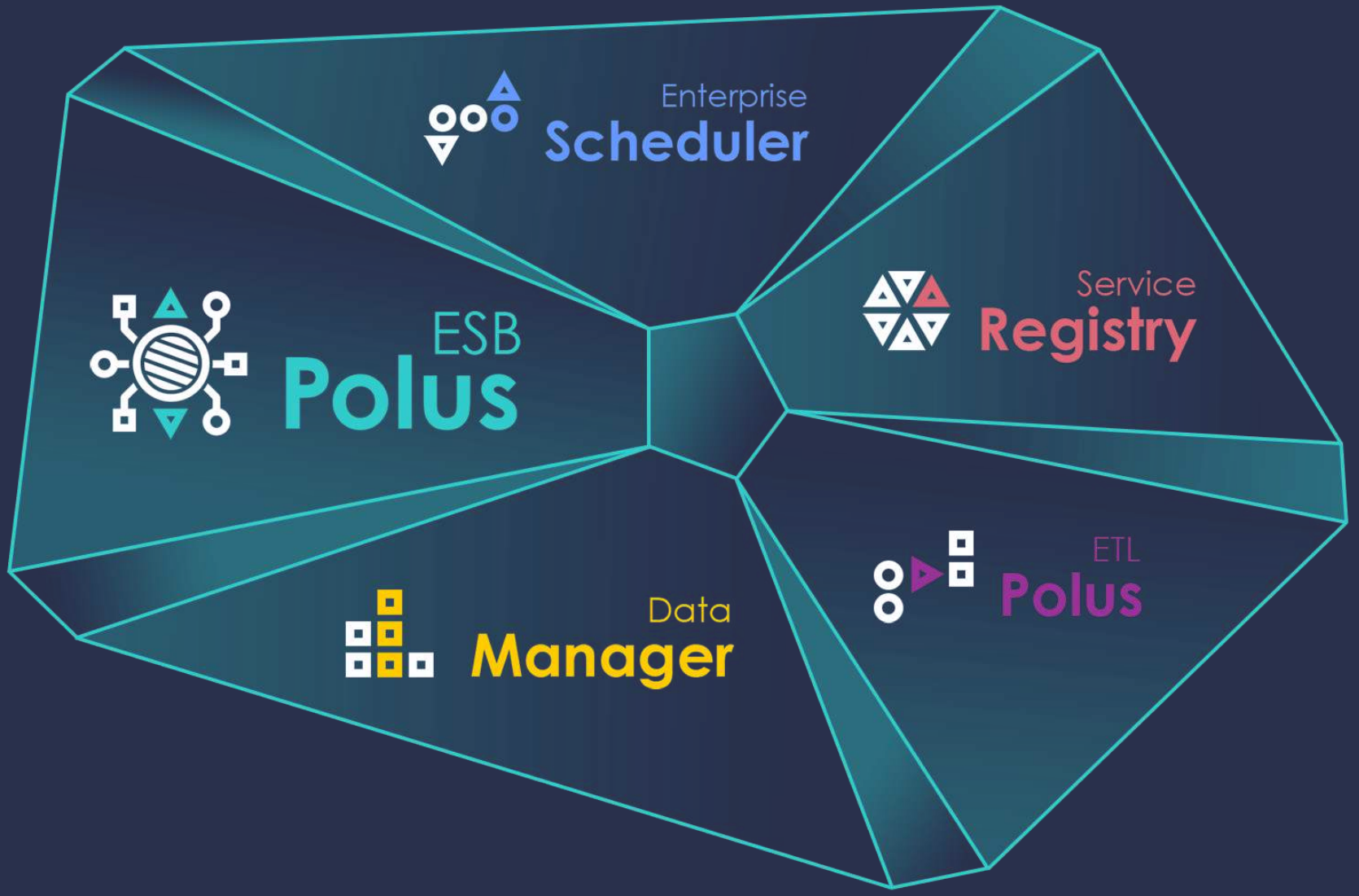
- ✓ Для оркестровки потоков операций используется представление в виде направленного ациклического графа (DAG); собранная в граф группа операций может запускаться по событию или по расписанию;




- ✓ 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.

ETL Polus

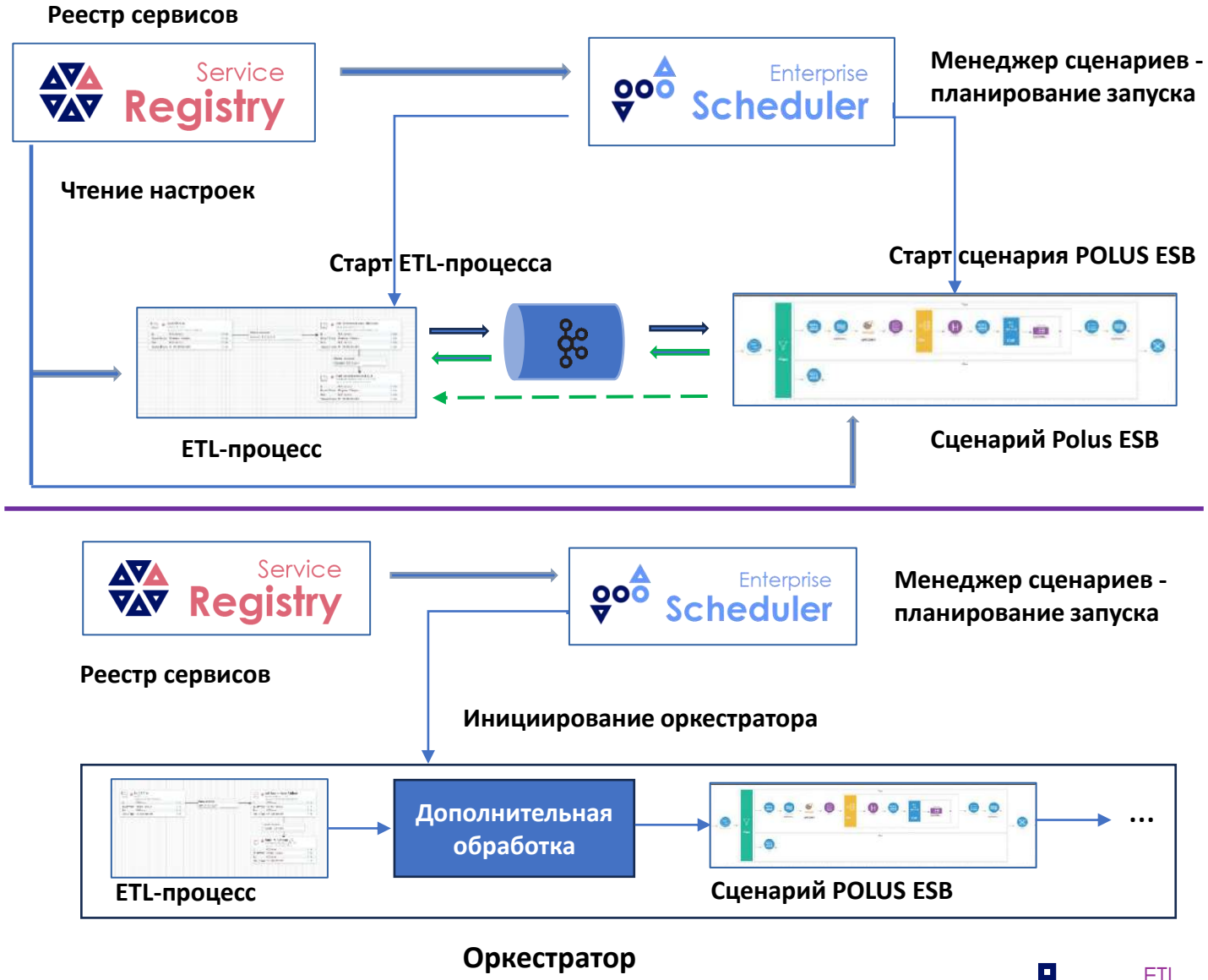
# Взаимодействие ETL и других продуктов интеграционной платформы POLUS




ETL
Polus

# Взаимодействие ETL и других продуктов интеграционной платформы POLUS

- Вся конфигурационная информация как для ETL-процессов, так и сценариев Polus ESB хранится в **Реестре сервисов**
- Запуск как ETL-процессов, так и сценариев Polus ESB может быть запланирован в **Менеджере сценариев**
- При необходимости завершение ETL-процесса Polus ETL может инициировать вызов сценария шины Polus ESB (например, Polus ETL помещает полученные и преобразованные данные в Kafka-топик, который “слушает” процесс Polus ESB)
- Сценарий Polus 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**



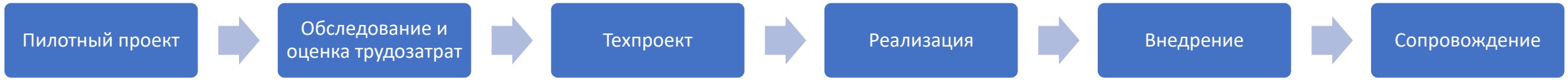

ETL Polus

# Polus ETL - выводы

- ✓ Polus ETL – полнофункциональный инструмент для реализации разнотипных ETL/ELT-процессов:
  - Потоковой 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 POLUS**
- ✓ 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

ETL Polus

# Подход компании Inpolus к выполнению проектов




- ✓ На этапе **пилотного проекта** специалисты компании реализуют один из реальных ETL/ELT-процессов, используемых у заказчика, при помощи инструмента Polus 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.

ETL Polus

# Почему Polus ETL?

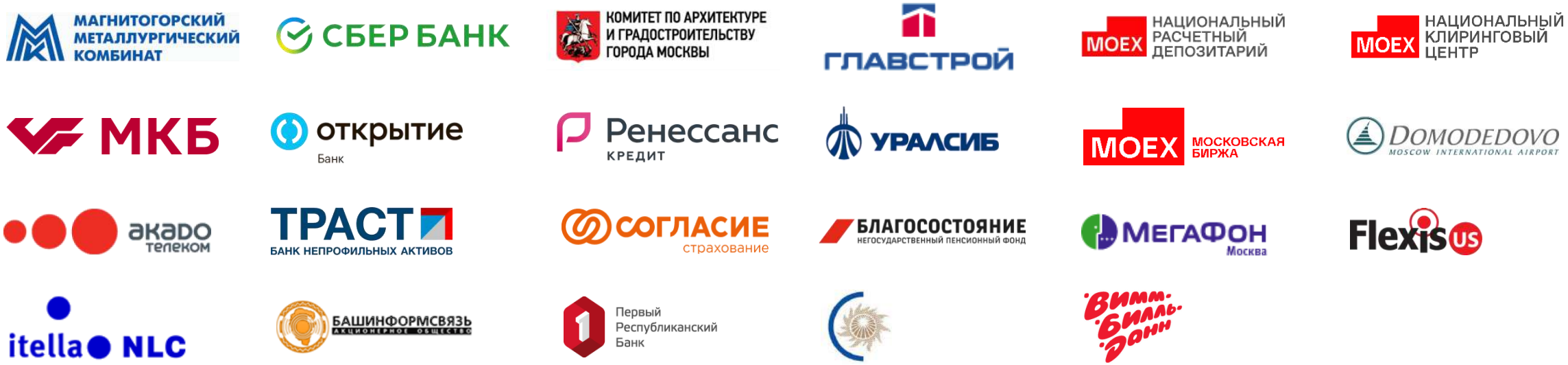
- Инструмент создан на основе одних из наиболее часто используемых Open Source продуктов для реализации ETL/ELT-процессов
- Является частью общей интеграционной платформы POLUS
- Интегрируется с разнотипными и часто используемыми источниками данных (СУБД – реляционные и NoSQL, Kafka, RabbitMQ, JMS, экосистема 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 Inpolus проделали большую работу по анализу и изучению продуктов и готовы предоставить нашу версию популярного решения

ETL Polus

# О компании «Инполюс»

«Инполюс» - российская ИТ компания, с 2009 года поставляет решения, услуги и программное обеспечение в области консалтинга, ИТ сервисов, безопасности и интеграционных технологий

## Нам доверяют



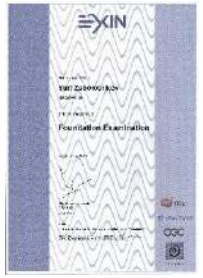
# Компетенции:

- Разработка специализированных прикладных сервисов и систем, базирующиеся на принципах сервис-ориентированной архитектуры (Service Oriented Architecture, SOA), в том числе корпоративных информационных шин (Enterprise Service Bus, ESB), систем автоматизации бизнес-процессов (Business Process Management System, BPMS)
- ИТ консалтинг и разработка приложений на заказ (JAVA, JavaScript, J2EE, PHP)
- Разработка, внедрение и поддержка веб-проектов (HTML, CSS, JavaScript, Java, PHP, Symfony, Twig, Doctrine, произвольные SQL-БД)
- Проектирование и построение сервисной модели функционирования ИТ подразделения компании на основе методик ITIL/ITSM
- Поставки лицензионного программного обеспечения
- Оказание услуг технической поддержки

# Экспертиза:

- Специалисты с опытом  работы в ИТ индустрии от 15 лет
- **Техническая сертификация:**
  - Java, .NET , Tibco, IBM, Oracle
- **Проектная сертификация:**
  - Project Management Professional (PMI PMP)
  - ITIL Foundation
  - IBM Project Management







## Больше информации:

https://www.inpolus.ru/solutions