МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Инструменты интеграции данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Лекция 1.
Введение в интеграцию данных

Важность интеграции данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Среднестатистическая корпоративная вычислительная среда состоит из сотен или даже тысяч разрозненных и изменяющихся компьютерных систем, которые были построены, приобретены или только приобретаются.

Данные из этих различных систем необходимо интегрировать для:

- **составления отчетов и анализа,**
- **совместно использовать для обработки бизнес-транзакций**
- **преобразовывать из одного системного формата в другой**

при замене старых систем и приобретении новых систем.

Важность интеграции данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Основное внимание в управлении данными **уделяется** данным, хранящимся в таких структурах, как **базы данных** и **файлы**, и гораздо **меньшее** внимание **уделяется** данным, **перемещающимся между структурами данных**.

Управление интерфейсами передачи данных в организациях быстро становится главной **проблемой руководства** бизнесом и информационными технологиями

Важность интеграции данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Традиционная разработка интерфейса быстро **приводит** к **неуправляемому уровню сложности**. Количество интерфейсов между приложениями и системами может стать экспоненциальным фактором количества систем.

На практике не каждая система должна взаимодействовать друг с другом, но между системами может существовать множество интерфейсов для разных типов данных или потребностей.

Для организации с 100 приложениями может быть около **5000 интерфейсов**.

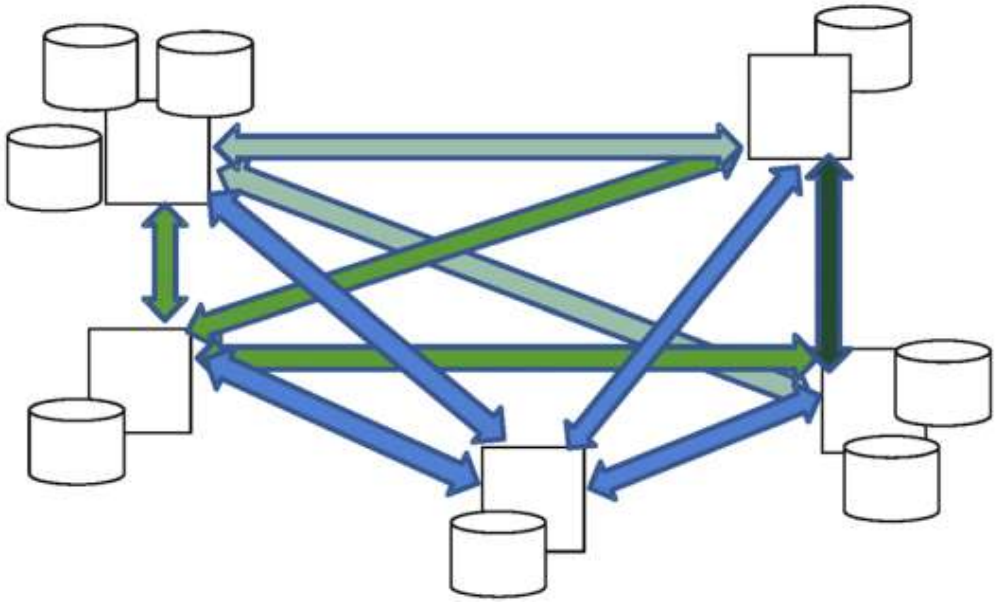
Портфель из 1000 приложений может предоставлять **полмиллиона интерфейсов** для управления.

Важность интеграции данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Внедрение передовых методов интеграции данных может сделать **управление интерфейсами организации намного более разумным**, чем традиционные решения для интеграции данных “**point to point**”, которые порождают управленческие проблемы.

Организация, разрабатывающая интерфейсы без стратегии интеграции корпоративных данных, может быстро обнаружить, что **управление** огромным количеством интерфейсов **невозможно**.



Увеличение количества приобретенных ПП от поставщиков

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

COTS (коммерческое готовое программное обеспечение). Большинство программных приложений, внедряемых в организациях в настоящее время, представляют собой приобретенные пакеты поставщиков, работа и процесс интеграции конкретного портфолио программного обеспечения, используемого в конкретной организации, является одним из немногих оставшихся видов деятельности по индивидуальной разработке.

Поставщики программного обеспечения могут разрабатывать системы таким образом, чтобы поддерживать интеграцию и взаимодействие с другими системами в портфолио, но конкретный портфель систем, требующих интеграции в организации, и, следовательно, решение для интеграции данных уникальны для каждой организации.

Увеличение количества приобретенных ПП от поставщиков

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Хотя практика приобретения прикладных решений вместо создания на заказ несколько **упрощает управление портфелем приложений** и его постоянную поддержку, она также **усложняет требуемую интеграцию данных** во всем портфеле приложений, чем если бы все приложения разрабатывались на заказ и использовали общие структуры данных.

Ключевое преимущество больших данных и виртуализации

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

В **облачной обработке** и **виртуализации данных** важнейшими компонентами внедрения этих технологий и решений являются **методы интеграции данных**.

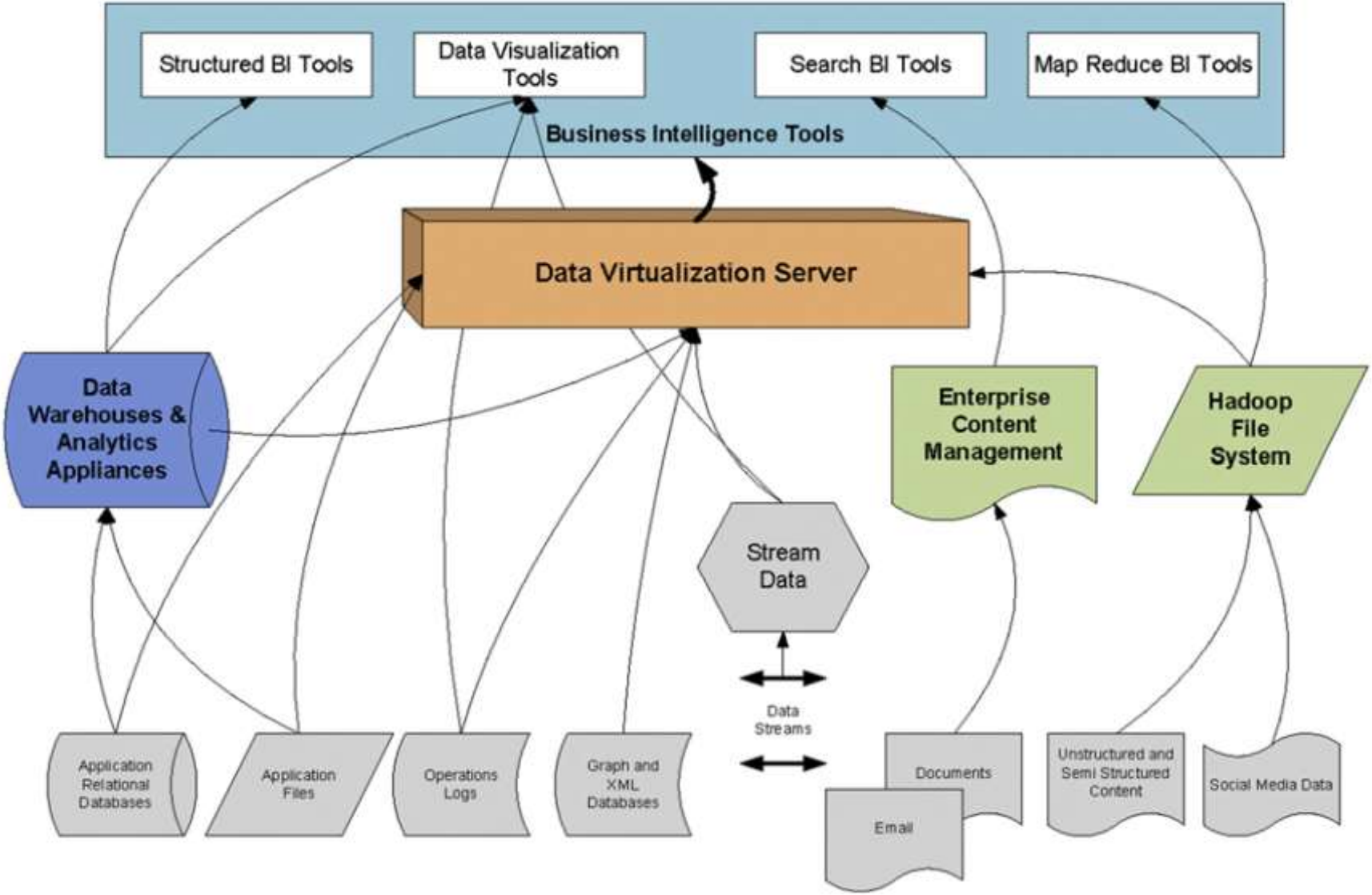
При работе с большими данными часто лучшим решением, вместо **консолидации данных** перед анализом, является **оставить большие объемы** и **различные типы данных** там, где они есть, и **распределить обработку между данными**, то есть использовать решение для параллельной обработки.

Ключевое преимущество больших данных и виртуализации

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Интеграция данных имеет важное значение для решений в области больших данных, но эти решения могут существенно отличаться от традиционной интеграции данных.

Big Data Architecture



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Ключевое преимущество больших данных и виртуализации

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ

Облачные архитектуры с решениями для внешних и виртуальных серверов, репликацией данных и потребностью в отказоустойчивых решениях **полагаются на решения интеграции данных**.

Однако, опять же, внедрение решений по интеграции данных в облачной среде **сильно отличается от решений в традиционных центрах обработки данных**, но основывается на базовых концепциях, разработанных за последние два десятилетия в области интеграции данных.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Что такое Интеграция данных?

Что такое Интеграция данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

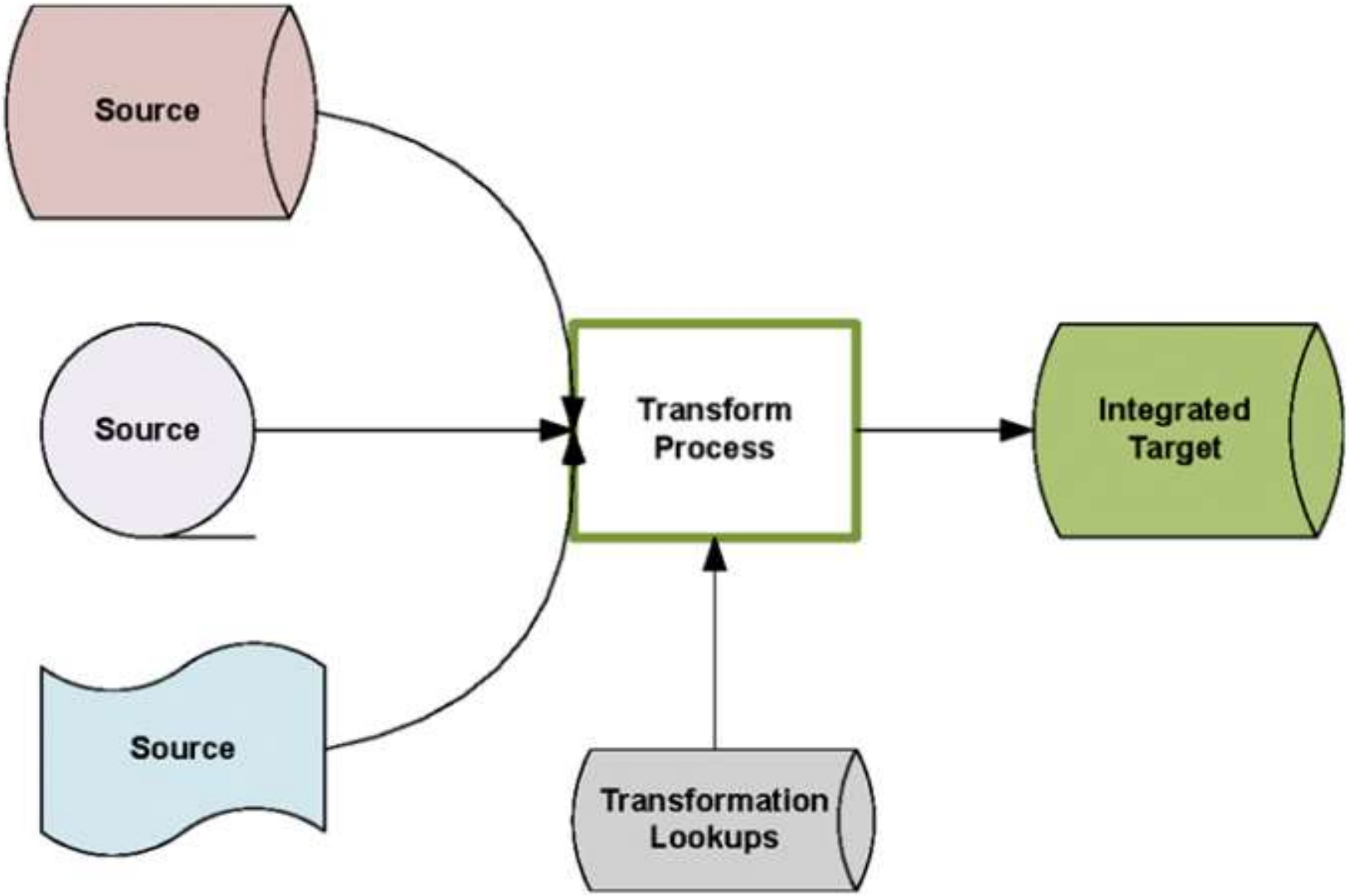
- **Планирование управлением данными** в хранилищах данных связано с “постоянными” данными, которые остаются неизменными.
- **Управление данными**, которые мигрируют между системами, приложениями, хранилищами данных и организациями — “**данные в движении**” — занимает **центральное место в эффективности** любой организации.

Что такое Интеграция данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

- Преобразование данных в универсальный формат **ПОНИМАНИЯ** структуры объединенных данных требует как технического, так и бизнес-понимания данных и структур данных, чтобы определить, как данные должны быть преобразованы.

Трансформация данных в универсальный формат



Что такое Интеграция данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Когда приложение в организации **заменяется новым пользовательским приложением** или приобретенным ППП, данные из старой системы необходимо перенести в новое приложение.

Новое приложение может уже использоваться в производстве и добавляются дополнительные данные, или приложение может еще не использоваться, а добавляемые данные заполняют пустые структуры данных.

Что такое Интеграция данных?

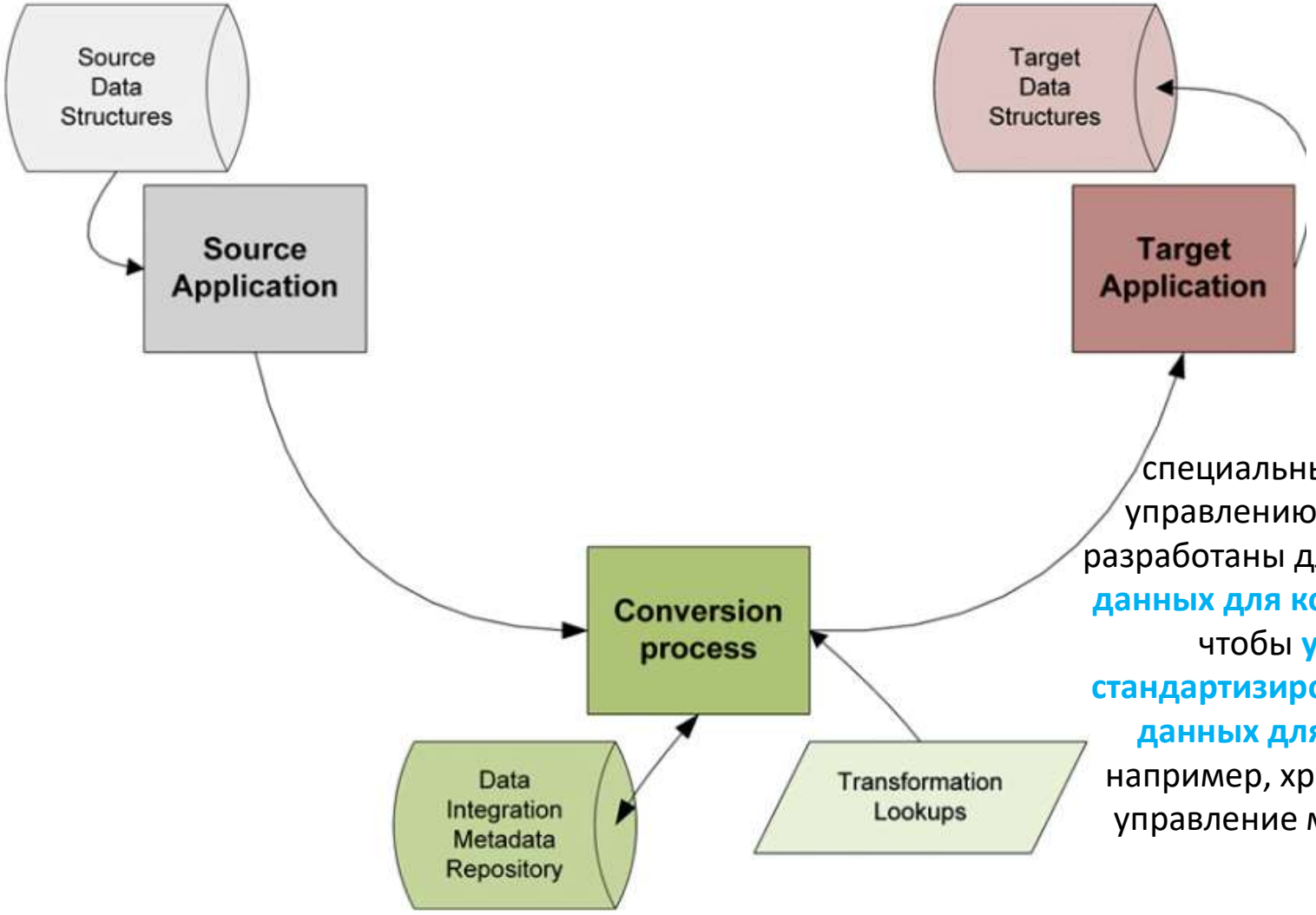
МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Процесс преобразования данных взаимодействует с исходной и целевой прикладными системами для перемещения и преобразования из технического формата, необходимого исходной системе, в формат и структуру, необходимые целевой системе.

Это наилучшая практика, особенно для того, чтобы происходило обновление данных

Миграция данных из одного приложения в другое

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ



специальные решения по управлению данными были разработаны для **централизации данных для конкретных целей**, чтобы **упростить и стандартизировать интеграцию данных для организации**, например, хранение данных и управление мастер-данными

Что такое Интеграция данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Большинство организаций имеют сотни или тысячи приложений, каждое со своими собственными базами данных.

Независимо от того, являются ли хранилища данных традиционными технологиями и системами управления базами данных или другими типами структур, такими как документы, сообщения или аудиофайлы, для организации

очень важно, чтобы эти приложения могли обмениваться информацией между собой.

Что такое Интеграция данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Решения для **интеграции данных**, как правило, внедряются в качестве сопутствующих решений для хранения данных, таких как:

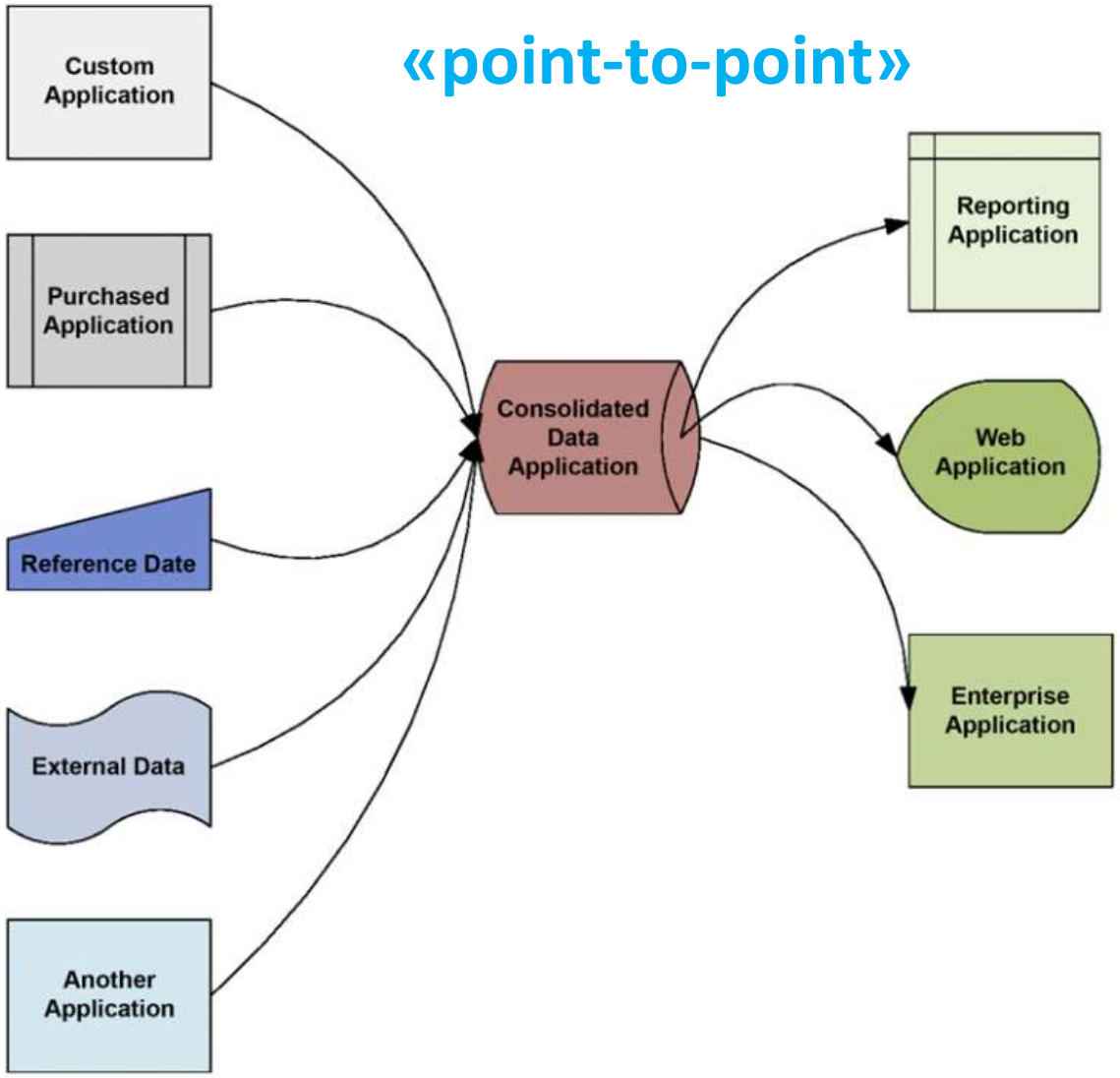
- **хранилища данных,**
- **управление основными данными,**
- **бизнес-аналитика,**
- **репозитории метаданных.**

Что такое Интеграция данных?

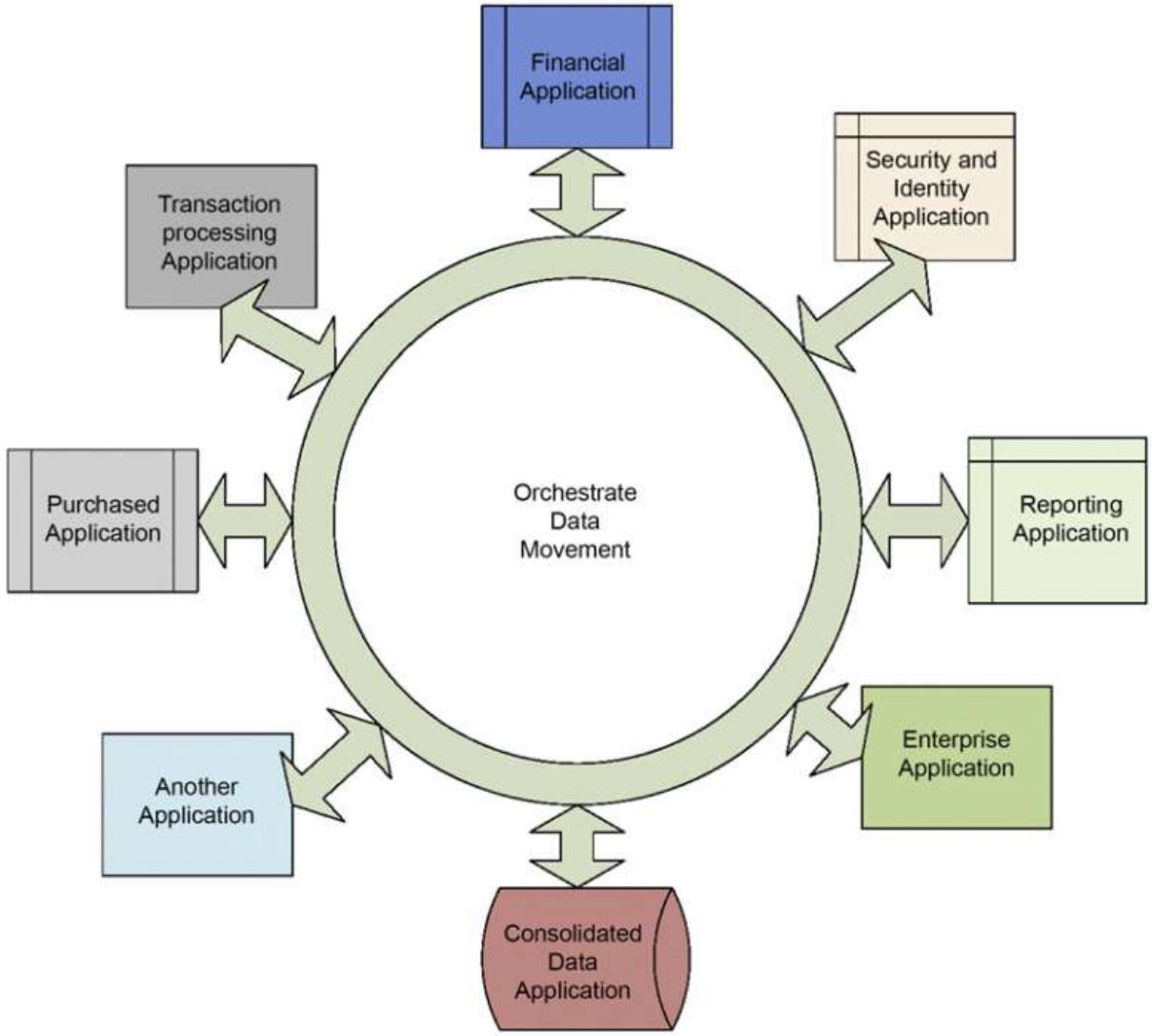
МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Стратегии и решения для **интеграции данных** в режиме **реального времени** теперь включают схемы миграции данных, которые значительно более эффективны, чем **«point-to-point».**

Перемещение данных в центральные точки консолидации



Перемещение данных по Организации



Что такое Интеграция данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

В прошлом большинство проектов по **интеграции данных** касались почти исключительно данных, хранящихся в базах данных.

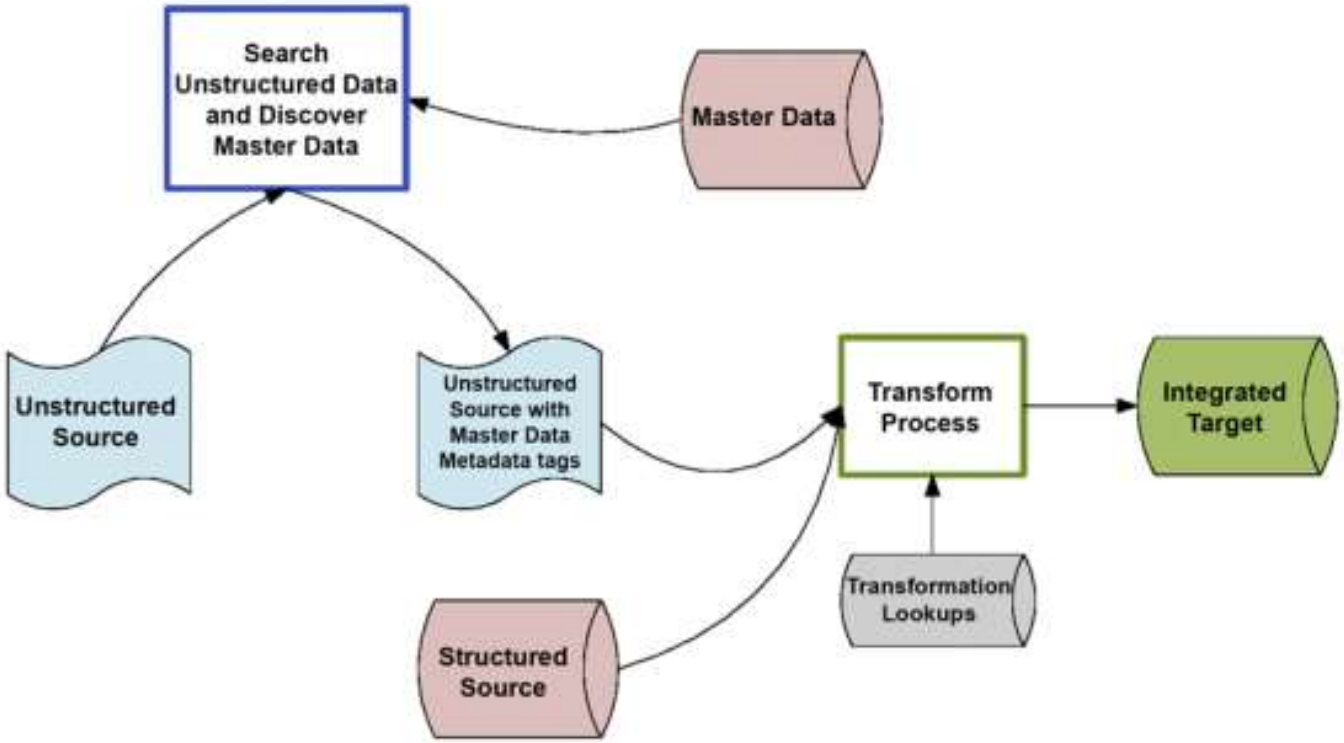
Теперь организациям крайне важно **интегрировать свои базы данных** (или структурированные) данные **с данными в документах, электронной почте, веб-сайтах, социальных сетях, аудио- и видеофайлах**.

Перемещение данных по Организации

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ

Данные, поиск которых происходит вне баз данных, таких как документы, электронная почта, аудио- и видеофайлы, могут использоваться для поиска клиентов, продуктов, сотрудников или другие важные ссылки.

Ссылки на мастер-данные прикрепляются к неструктурированным данным в виде тегов метаданных, которые затем могут быть использованы для **интеграции данных** с другими источниками и типами.

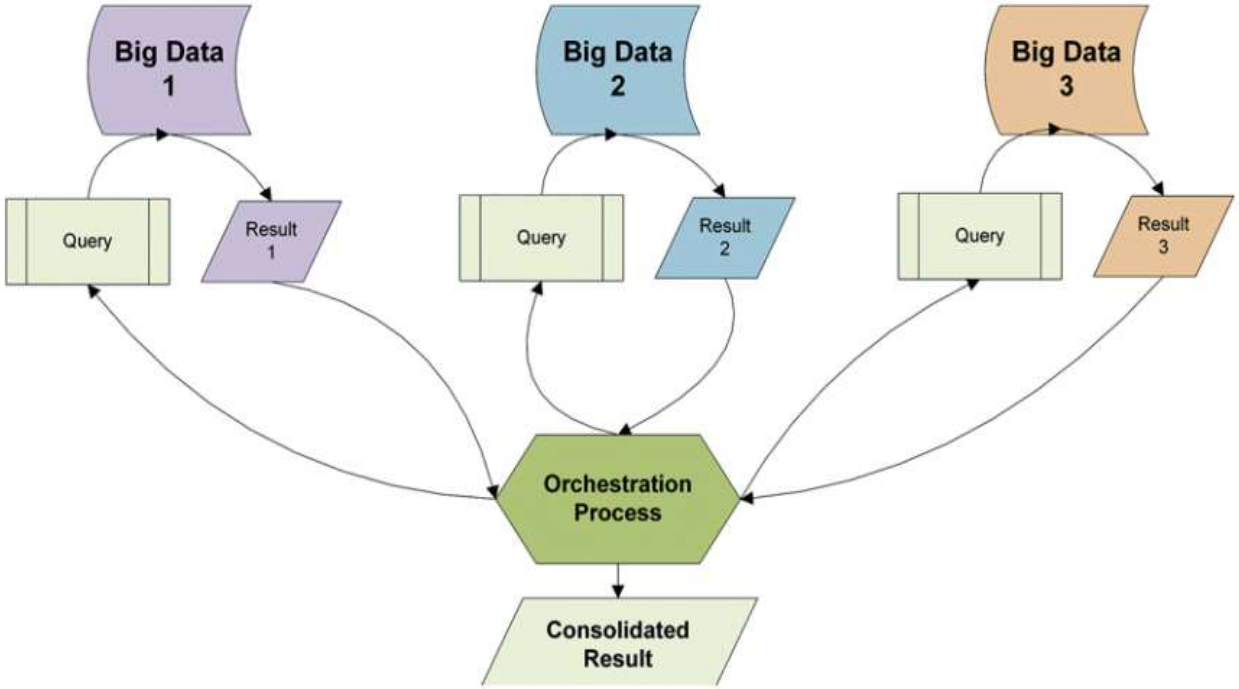


Перемещение процесса в данные

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

В некоторых случаях работы с очень большими объемами данных **более эффективно перенести процесс на данны** а **затем консолидировать гораздо меньшие результать**

Полученные решения для больших данных в основном используются программистам или специалистами по данны



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Типы и сложность интеграции данных

Типы и сложность интеграции данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Доступ к данным и управление безопасностью являются основными задачами как для **постоянных**, так и для **перемещаемых** данных.

Типы и сложность интеграции данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

- В случае с **постоянными данными** проблемы связана с **моделью** или **структурой хранимых данных**.
- При управлении **перемещаемыми данными** - проблема заключается в том, **как связывать, отображать и преобразовывать данные** между различными системами.

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Существует реализация решения по **интеграции данных**, она включает:

- **моделирование данных в пути,**
- **использование центральной модели данных.**

.

Интеграция пакетных данных происходит, когда данные, которые должны быть переданы из источника к получателю, группируются вместе и отправляются периодически, например, ежедневно, еженедельно или ежемесячно.

Большинство интерфейсов между системами в прошлом использовались в форме передачи большого файла данных из одной системы в другую на периодической основе.

Каноническое моделирование

Передача данных между двумя системами, при которой система-отправитель передает данные целевой системе-получателю, называется **«point-to-point».**

Файл данных будет обрабатываться принимающей системой в какой-то момент времени, не обязательно мгновенно; таким образом, интерфейс будет «асинхронным», поскольку отправляющая система не будет ждать немедленного подтверждения, прежде чем транзакция будет считаться завершенной.

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

«Пакетный» подход к интеграции данных по-прежнему подходит и эффективен для очень больших взаимодействий с данными, таких как преобразование данных и загрузка моментальных снимков данных в хранилища данных.

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

При управлении большими портфелями прикладных систем **предпочтительнее иметь более слабую связь системных интерфейсов**, чтобы разрешить внесение изменений в приложения, которые не нарушают работу других систем и не требуют такой тщательной координации одновременных изменений. Поэтому обычно предпочтительно, чтобы решения для интеграции данных были «**слабо связанными**».

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Интерфейсы, которые обмениваются моментально данными для выполнения одной бизнес-транзакции, называются интерфейсами «**реального времени**» (**потоковая передача данных**). Обычно они включают гораздо меньший объем данных, передаваемых в форме «**сообщения**». Большинство интерфейсов реального времени по-прежнему являются двухточечными между отправляющей и принимающей системами и тесно связаны между собой, потому что отправляющая и принимающая системы по-прежнему имеют конкретное соглашение относительно формата, так что любые изменения должны вноситься в две системы одновременно. интерфейсы обычно называют синхронными, потому что транзакция будет ожидать завершения обработки интерфейса данных как в отправляющей, так и в принимающей системах.

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Термин «большие данные» указывает на наличие больших объемов данных, а также данных различных типов.

Принимая во внимание дополнительные объемы и различные форматы, **интеграция данных «больших данных»** может включать в себя параллельное распределение обработки данных, которое должно выполняться по исходным данным и только интеграцию результатов, поскольку предварительная консолидация данных может занять слишком много времени, дополнительного места для хранения.

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Интеграция структурированных и неструктурированных данных включает связывание между ними общей информации, которая, вероятно, представлена в виде мастер-данных или ключей структурированных данных в базах данных и в виде тегов метаданных или встроенного содержимого в неструктурированных данных.

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Виртуализация данных включает в себя использование различных методов интеграции данных для консолидации данных в режиме реального времени из различных источников, а не только структурированных данных.

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

«Хранилище данных» — это практика управления данными, при которой данные копируются из различных операционных систем в постоянное хранилище данных в согласованном формате для использования в целях анализа и составления отчетов.

Каноническое моделирование

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Технологии виртуализации данных делают возможной интеграцию данных в режиме реального времени для анализа, особенно при использовании в сочетании с хранилищем данных.

Новые технологии, использующие **хранилища данных в оперативной памяти** и другие подходы к виртуализации, **позволяют создавать решения для очень быстрой интеграции данных**, которые не должны полагаться на промежуточные постоянные хранилища данных, такие как хранилища данных и витрины данных.

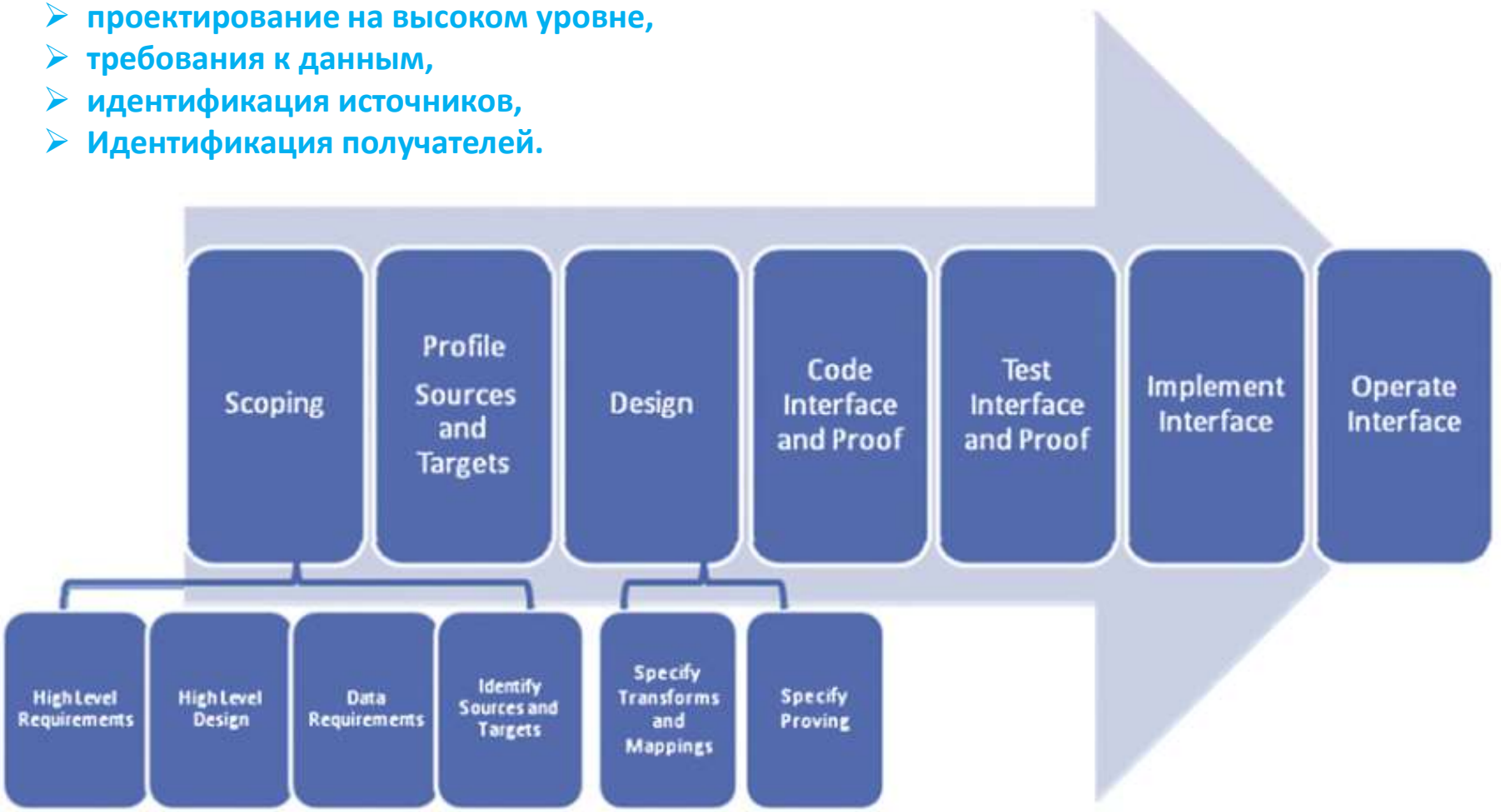
МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Процесс разработки интеграции данных

Data Integration Life Cycle

Главная часть **жизненного цикла** — определение структуры проекта:

- **проектирование на высоком уровне,**
- **требования к данным,**
- **идентификация источников,**
- **Идентификация получателей.**



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

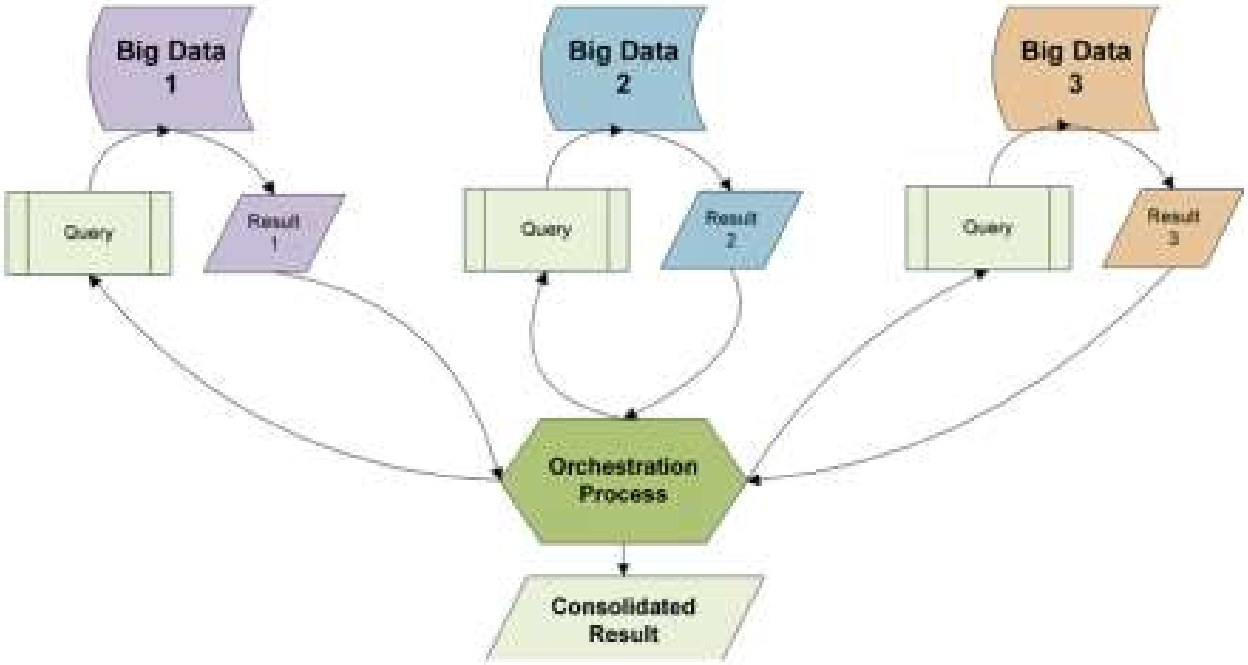
Переход от процесса к данным

Moving process to data

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

В эпоху огромного расширения объема данных, доступных для организации (большие данные), иногда более эффективно распределять обработку по нескольким местоположениям данных, а не собирать данные вместе (и, следовательно, дублировать) для их обработки.

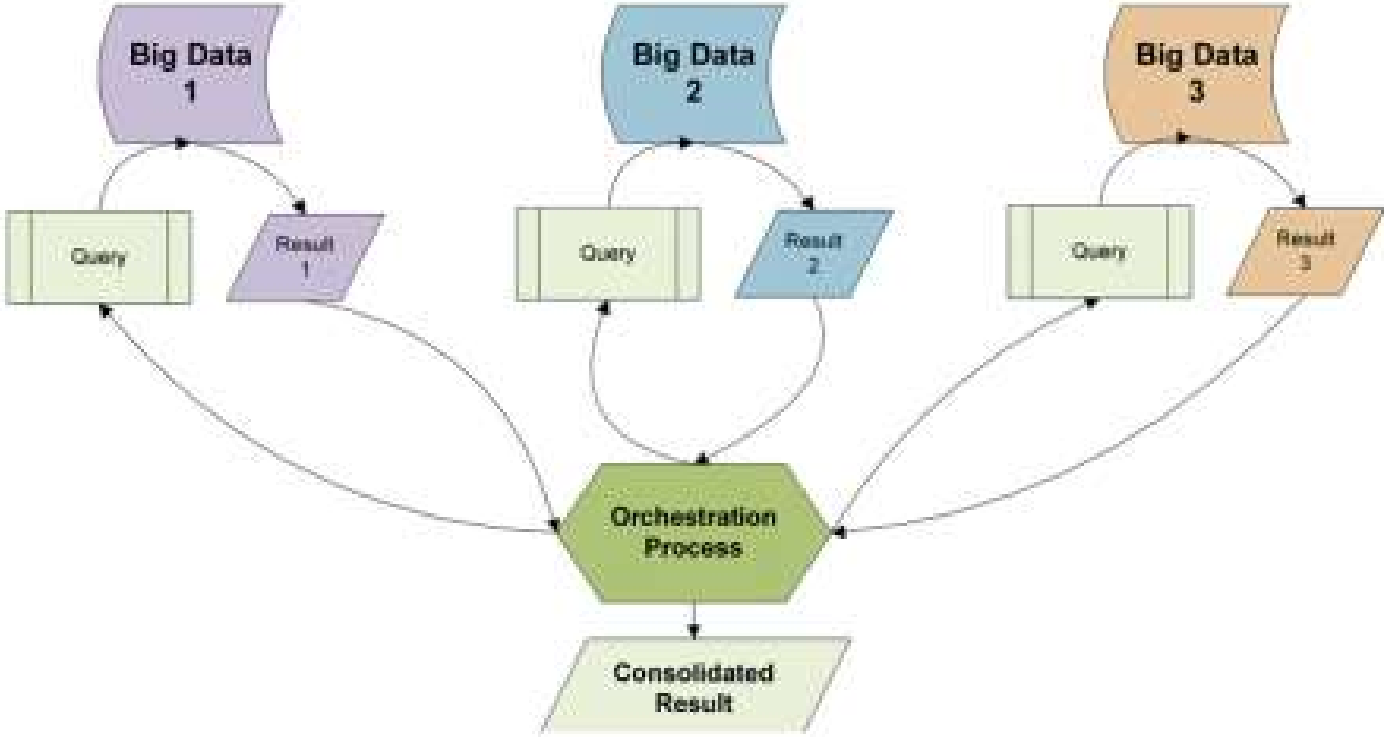
Решения для больших данных часто подходят к интеграции данных с совершенно иной точки зрения, чем традиционные решения для интеграции данных.



Moving process to data

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Как показано на рис., в некоторых случаях работы с очень большими объемами более эффективно перенести процесс на данные, а затем консолидировать гораздо меньшие результаты. Появляющиеся решения для больших данных в основном используются программистами и технологами или высококвалифицированными специалистами, такими как специалисты по данным.



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Типы и сложность интеграции данных

Moving process to data

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Доступ к данным и управление безопасностью являются основными задачами как для постоянных, так и для перемещаемых данных.

Постоянная безопасность данных обычно управляется на уровнях:

- **физическом,**
- **сетевом,**
- **серверном,**
- **прикладном,**
- **в хранилище данных.**

Пакетная интеграция данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Интеграция пакетных данных происходит, когда данные, которые должны быть переданы от источника к цели, группируются вместе и отправляются периодически, например, ежедневно, еженедельно или ежемесячно.

Файл данных из одной системы в другую на периодической основе. Содержимое файла будет представлять собой записи согласованного макета, а отправляющая и принимающая прикладные системы будут соглашаться с этим форматом и понимать его.

Пакетная интеграция данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Передача данных между двумя системами, при которой система-отправитель передает данные целевой системе-получателю, называется «**точка-точка**».

Файл данных будет обрабатываться принимающей системой в какой-то момент времени, не обязательно мгновенно; таким образом, интерфейс будет «асинхронным», поскольку отправляющая система не будет ждать немедленного подтверждения, прежде чем транзакция будет считаться завершенной.

Пакетная интеграция данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

«**Пакетный**» подход к интеграции данных по-прежнему подходит и эффективен для очень больших взаимодействий с данными, таких как преобразование данных и загрузка моментальных снимков данных в хранилища данных.

Пакетная интеграция данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

При управлении большими портфелями прикладных систем предпочтительнее иметь более слабую связь системных интерфейсов, чтобы разрешить внесение изменений в приложения, которые не нарушают работу других систем и не требуют такой тщательной координации одновременных изменений. Поэтому обычно предпочтительно, чтобы решения для интеграции данных были «**слабо связанными**».

Интеграция данных в режиме реального времени

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Интерфейсы, которые необходимы между системами немедленно для выполнения одной бизнес-транзакции, называются интерфейсами «реального времени». Обычно они включают гораздо меньший объем данных, передаваемых в форме «сообщения».

Большинство интерфейсов реального времени по-прежнему являются двухточечными между отправляющей и принимающей системами и тесно связаны между собой, потому что отправляющая и принимающая системы по-прежнему имеют конкретное соглашение относительно формата, так что любые изменения должны вноситься в две системы одновременно, интерфейсы обычно называют **синхронными**, потому что транзакция будет ожидать завершения обработки интерфейса данных как в отправляющей, так и в принимающей системах.

Интеграция больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Термин «большие данные» указывает на наличие больших объемов данных, а также данных различных технологий и типов. Принимая во внимание дополнительные объемы и различные форматы, интеграция данных больших данных может включать в себя параллельное распределение обработки данных, которые должны выполняться, по исходным данным и только интеграцию результатов, поскольку предварительная консолидация данных может

занять слишком много времени и средств.

слишком много дополнительного места для хранения.

Интеграция больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Термин «большие данные» указывает на наличие больших объемов данных, а также данных различных технологий и типов. Принимая во внимание дополнительные объемы и различные форматы, интеграция данных больших данных может включать в себя параллельное распределение обработки данных, которые должны выполняться, по исходным данным и только интеграцию результатов, поскольку предварительная консолидация данных может

занять слишком много времени и средств.

слишком много дополнительного места для хранения.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Пакетная интеграция данных

Что такое пакетная интеграция данных?

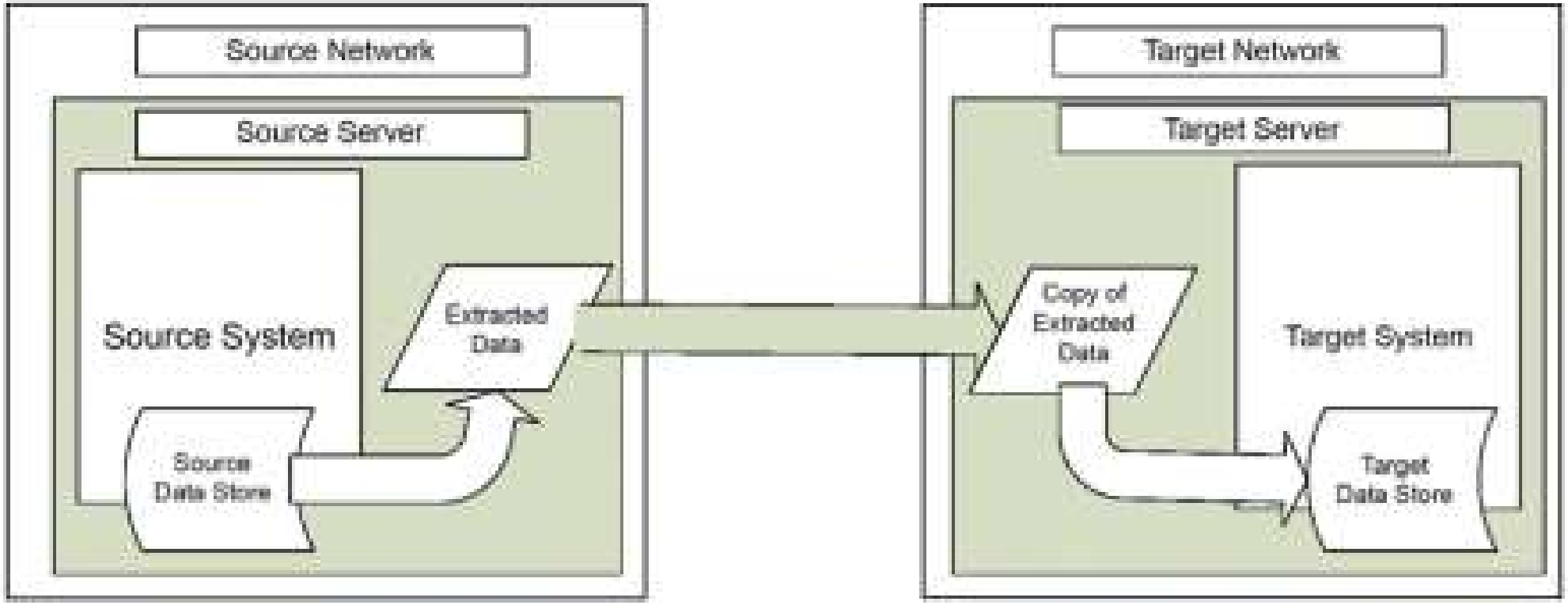
МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Большинство интерфейсов между системами традиционно представляли собой передачу большого файла данных из одной системы в другую на периодической основе, например, ежедневно, еженедельно или ежемесячно. Содержимое файла будет представлять собой записи согласованного макета, а формат будет согласован и понятен между отправляющей и принимающей прикладными системами.

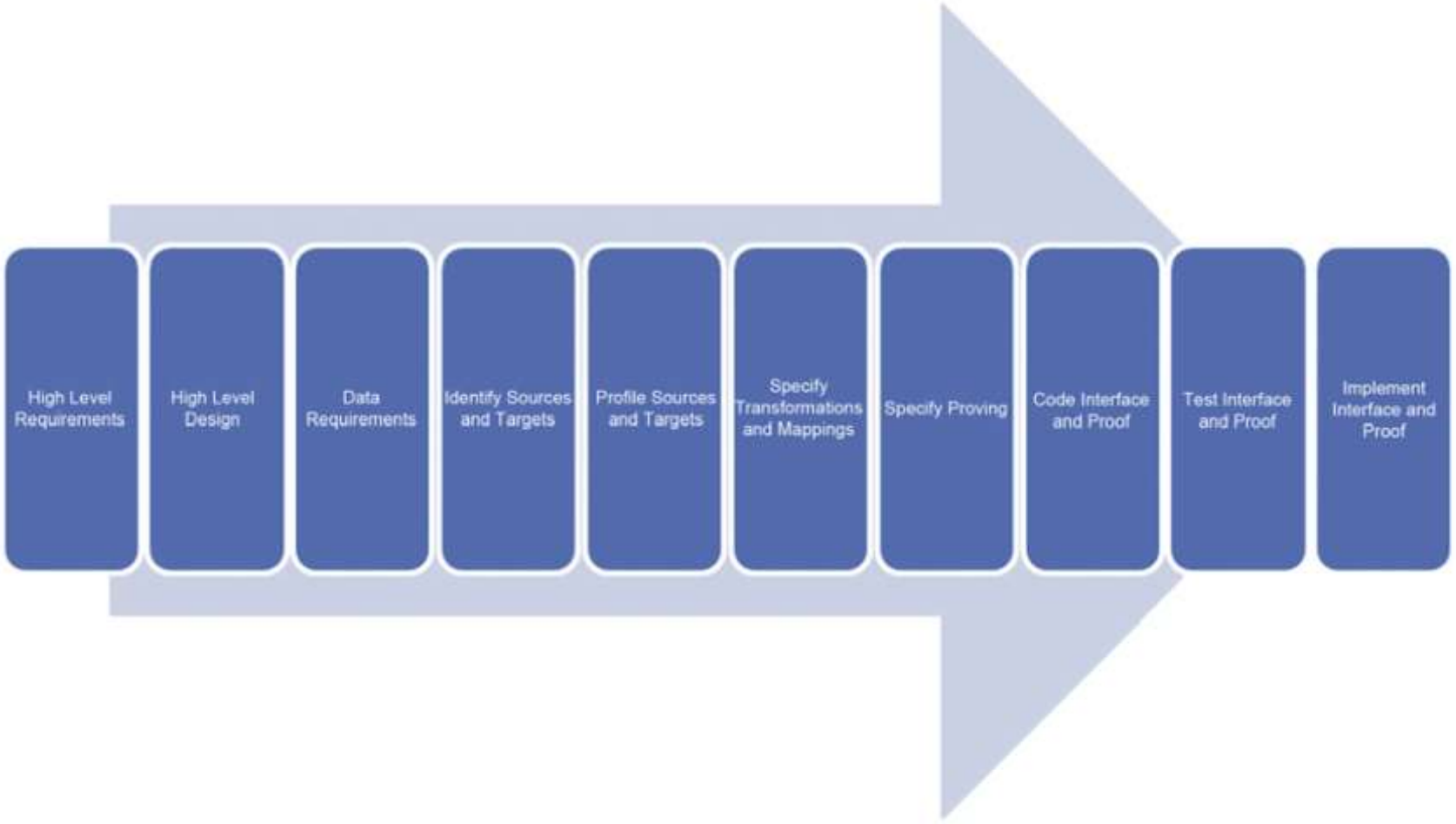
Этот процесс называется пакетным режимом, потому что данные «пакетируются» в группы и отправляются периодически, а не по отдельности в режиме реального времени.

Что такое пакетная интеграция данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ



Жизненный цикл пакетной интеграции данных



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Конверсия данных

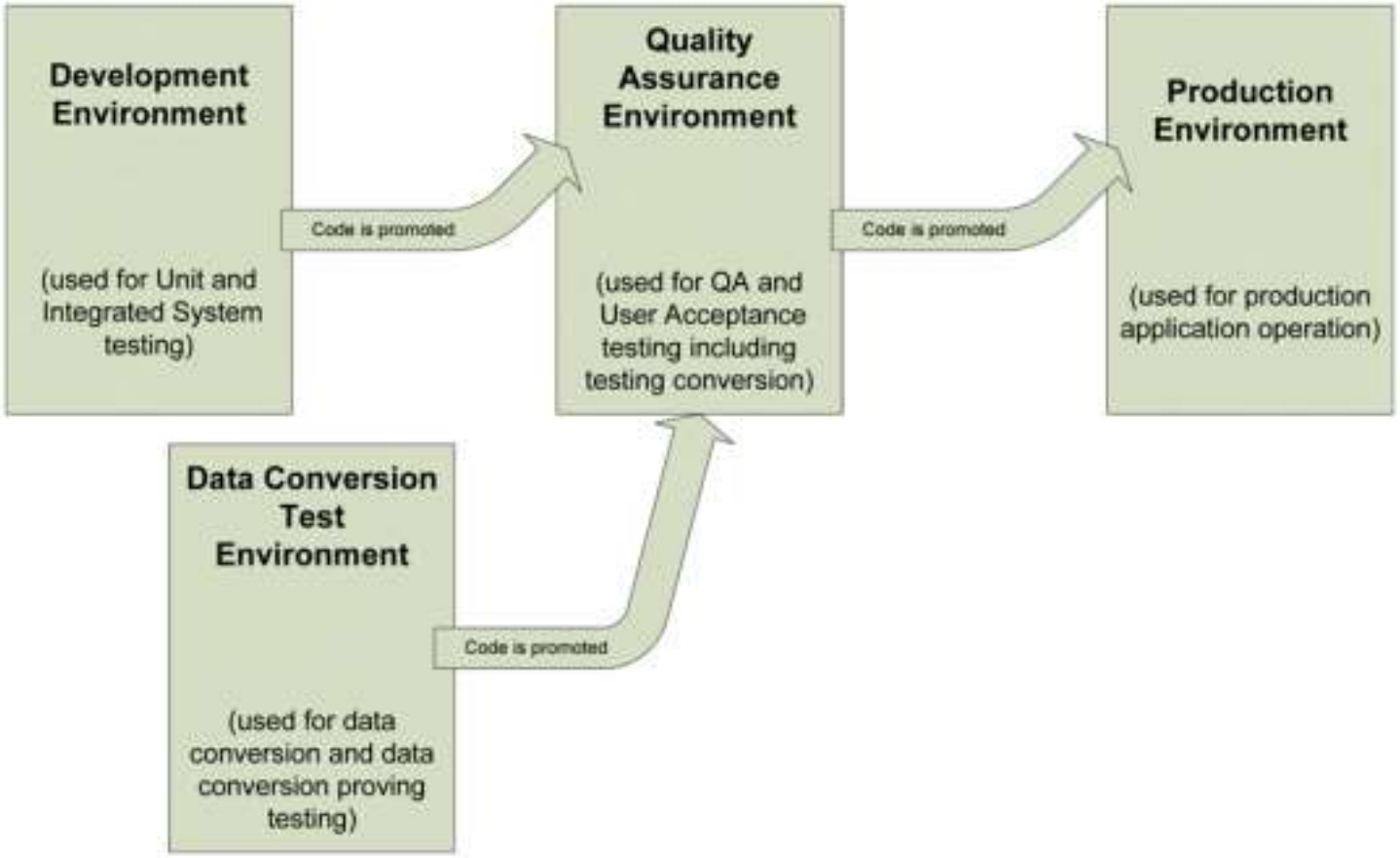
Что такое преобразование данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

При внедрении новой прикладной системы или изменении операций с одной прикладной системы на другую необходимо **заполнить структуры данных новой прикладной системы**.

Иногда структуры данных новой системы приложений пусты, а иногда при консолидации приложений в новой структуре данных уже есть данные и их необходимо добавить.

Поток данных среды



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Архитектура интеграции пакетных данных

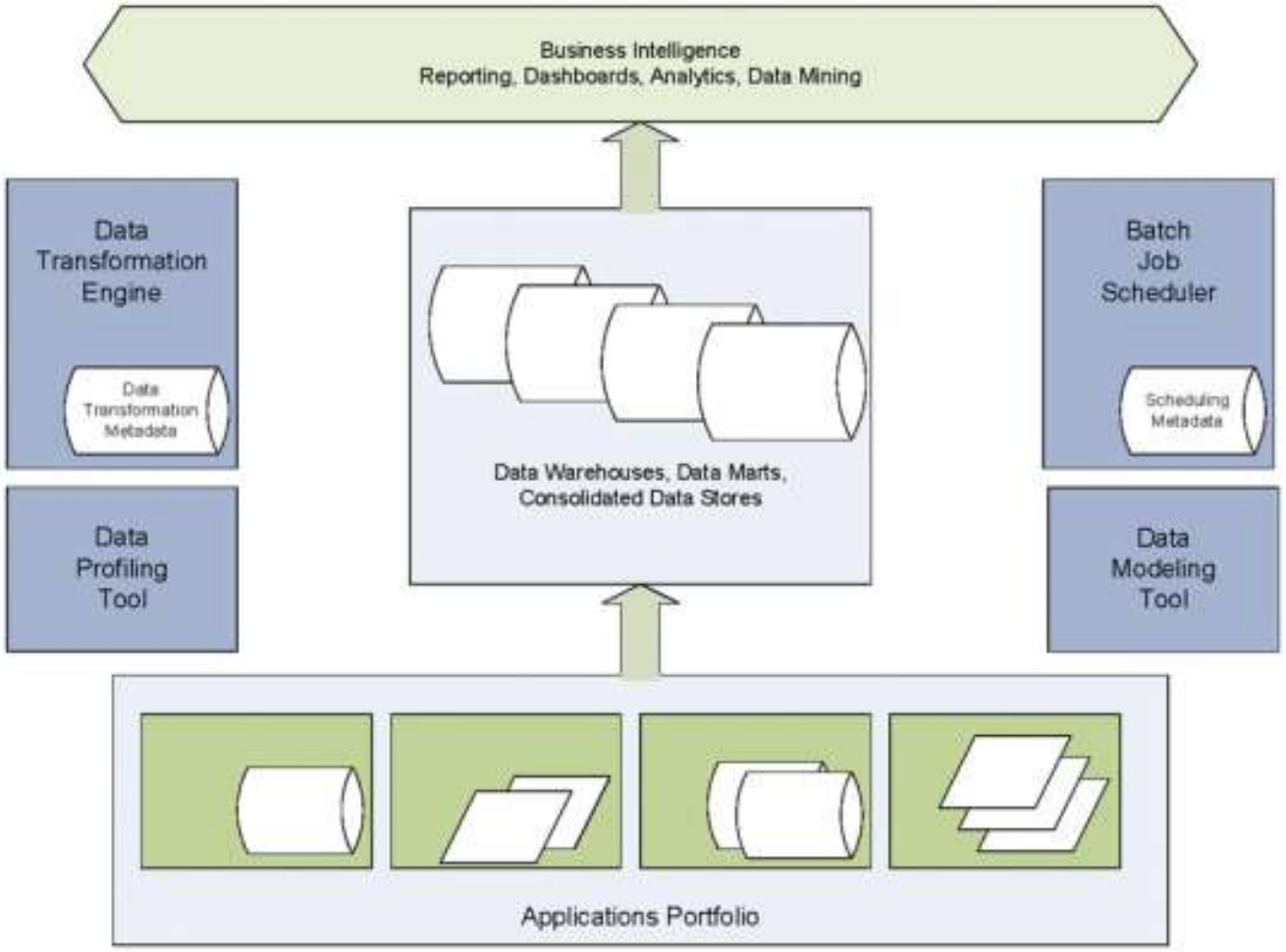
Что такое архитектура интеграции пакетных данных?

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Чтобы включить пакетную интеграцию данных, необходимо иметь инструменты для поддержки анализа и разработки кода интеграции данных, а также инструменты для поддержки операций по интеграции данных.

Очень сложные возможности имеют относительно высокую цену и требуют, чтобы ими управляли специалисты, хорошо разбирающиеся в использовании инструментов. Большинство средних и крупных организаций вкладывают средства в сложную возможность интеграции пакетных данных по крайней мере для одной области потребностей, например для загрузки хранилищ данных и киосков данных.

Batch Data Integration Architecture



инструменты и системы, необходимые для реализации возможности интеграции пакетных данных.

МОСКОВСКИЙ
ГОРОДСКОЙ
УНИВЕРСИТЕТ
МГПУ

СПАСИБО ЗА ВНИМАНИЕ