МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Инструменты аналитика данных для решения прикладных задач
лекция 2

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Инструменты интеграции и удаление файлов (ETL)

Критерии идеального ETL (Extract-Transform-Load) инструмента

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

1. **ETL-инструмент должен быть простым в освоении.**

Специалист не должен тратить полжизни на изучение нового ПО, а просто взять и практически сразу начать работать с ним.

Критерии идеального ETL (Extract-Transform-Load) инструмента

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

2. В нём должно быть предусмотрено максимальное количество готовых коннекторов. Мы все пользуемся плюс-минус одними и теми же системами: от **1С** до **SAP**, **Oracle**, **AmoCRM**, **Google Analytics**. И никто не хочет программировать коннекторы к ним с нуля.

Критерии идеального ETL (Extract-Transform-Load) инструмента

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

3. Инструмент должен быть универсальным и работать с разными BI системами.

Это облегчает переход аналитиков и разработчиков из одной компании в другую — если на прошлом месте работы, например, использовали **QlikView**, а на новом — **Visiology,** желательно сохранить возможность пользоваться тем же **ETL-**инструментом.

Критерии идеального ETL (Extract-Transform-Load) инструмента

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

4. ETL не должен ограничивать развитие аналитики.
У многих ETL-инструментов есть критическая проблема — в них несложно реализовать простенькие вещи, но для более сложных задач приходится искать новый инструмент, который сможет расти вместе с тобой.

Критерии идеального ETL (Extract-Transform-Load) инструмента

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

5. Получить недорогой (а лучше — полностью бесплатный) инструмент, причем не только на время “пробного периода”, а насовсем, чтобы пользоваться им без ограничений.

ETL: Extract, Transform, Load

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

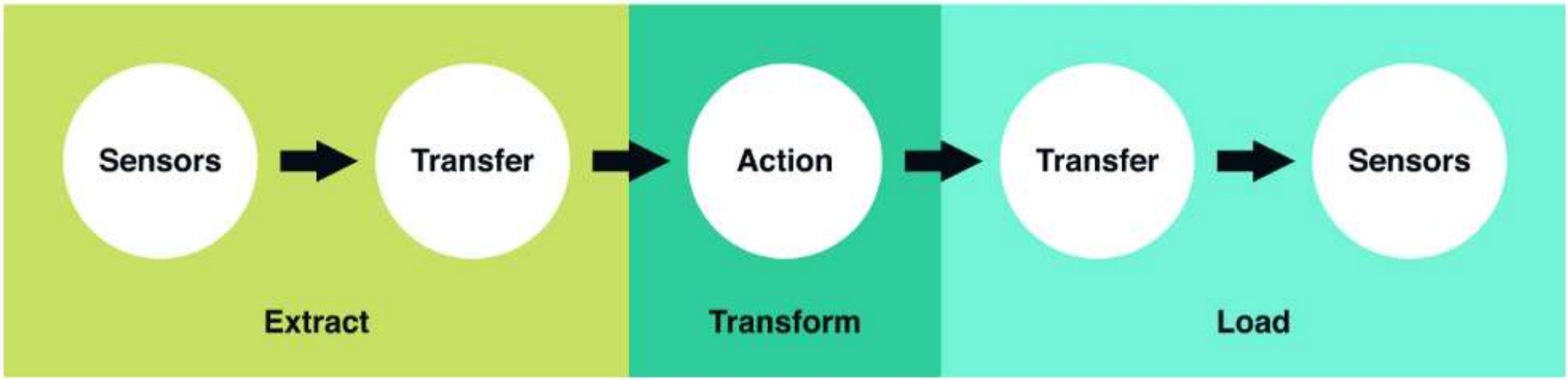
Extract, Transform и Load — это 3 концептуально важных шага, определяющих, каким образом устроены большинство современных пайплайнов данных.

На сегодняшний день это базовая модель того, как сырые данные сделать готовыми для анализа.

ETL: Extract, Transform, Load

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

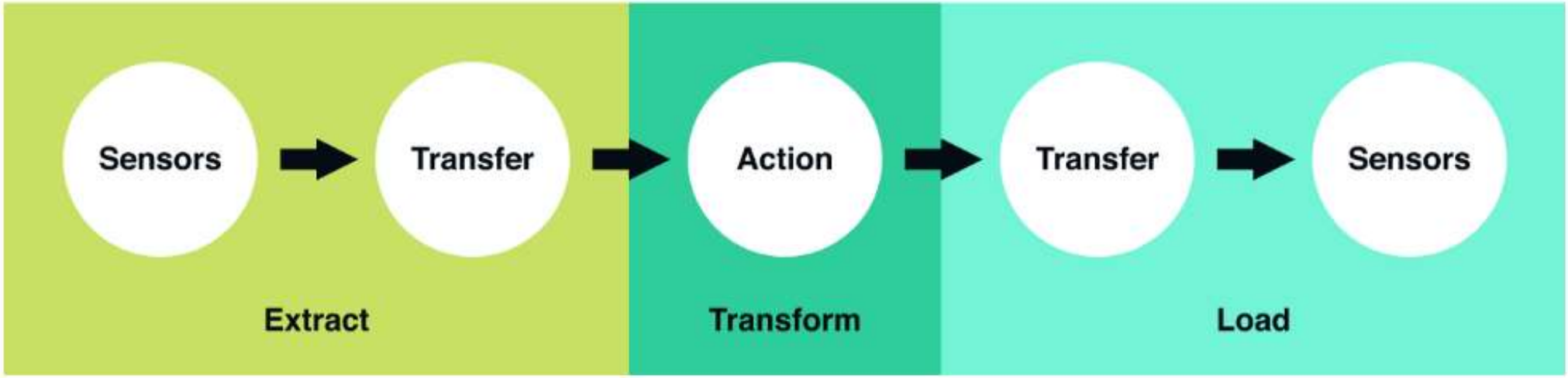
Extract. Это шаг, на котором датчики принимают на вход данные из различных источников (**логов пользователей, копии реляционной БД, внешнего набора данных** и т.д.), а затем передают их дальше для последующих преобразований.



ETL: Extract, Transform, Load

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

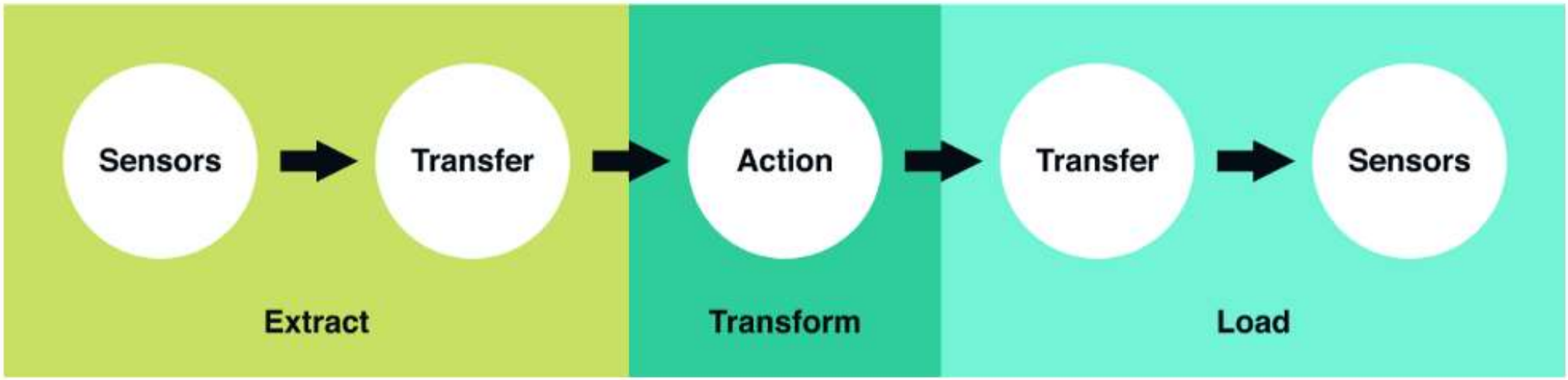
Transform. Это «сердце» любого ETL, этап, когда мы применяем бизнес-логику и делаем фильтрацию, группировку и агрегирование, чтобы преобразовать сырые данные в готовый к анализу датасет. Эта процедура требует понимания бизнес задач и наличия базовых знаний в области.



ETL: Extract, Transform, Load

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Load. Загрузка обработанных данных.
Полученный набор данных может быть использован конечными пользователями, а может являться входным потоком к еще одному ETL.



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Работа ETL системы

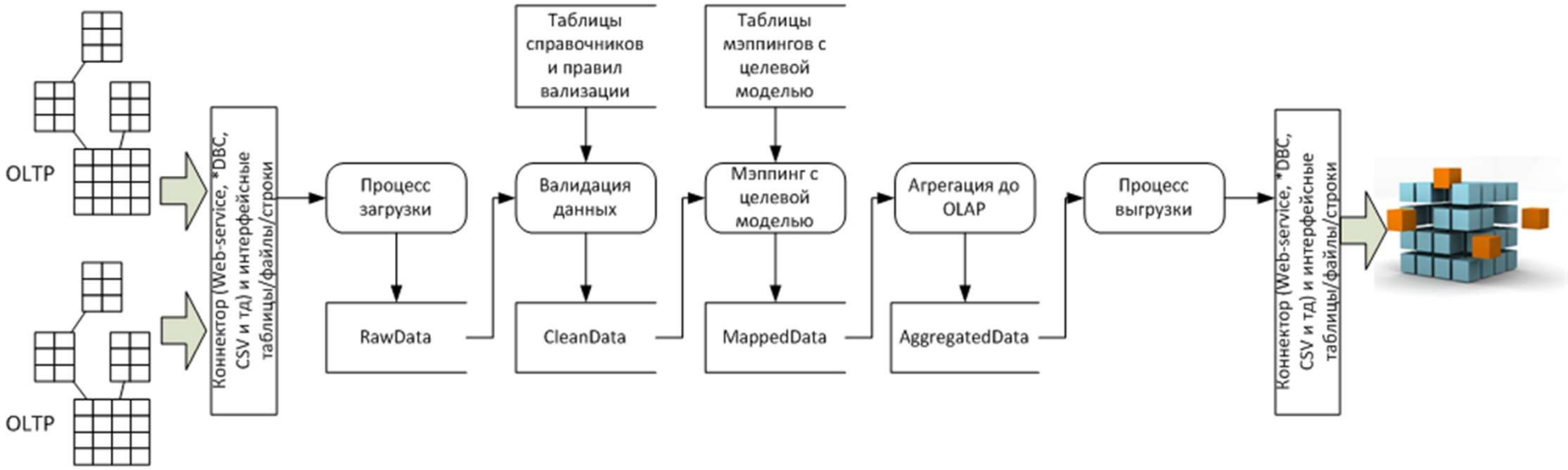
Функции ETL системы

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Стадии преобразования в ETL

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

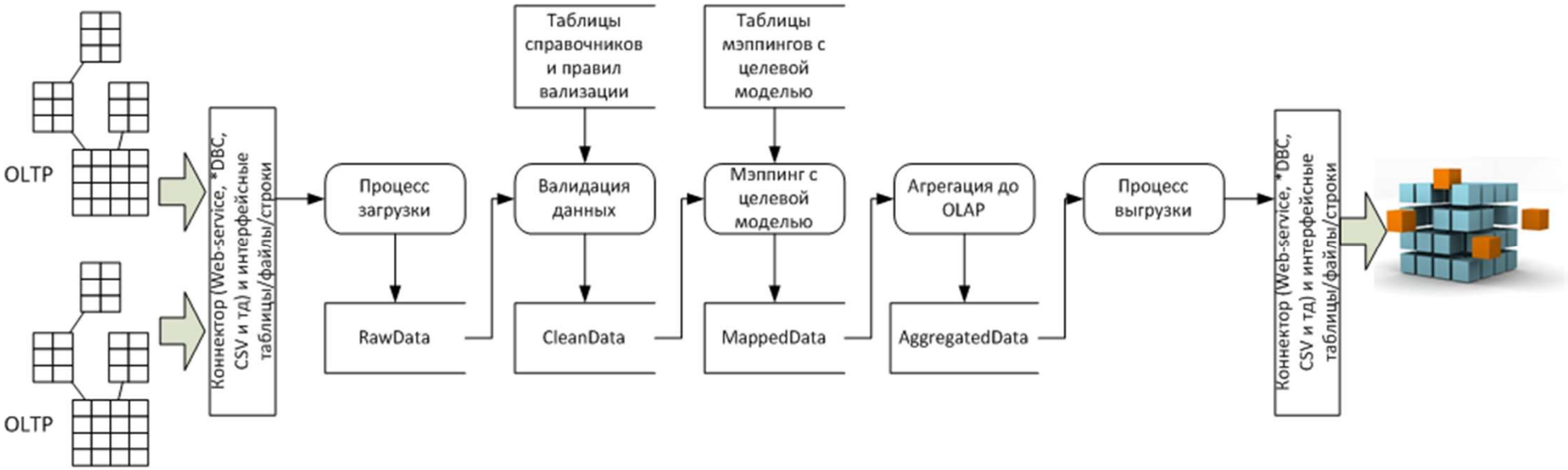
Поток данных из нескольких систем-источников (обычно **OLTP**) и система приемник (обычно **OLAP**)



Стадии преобразования в ETL

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

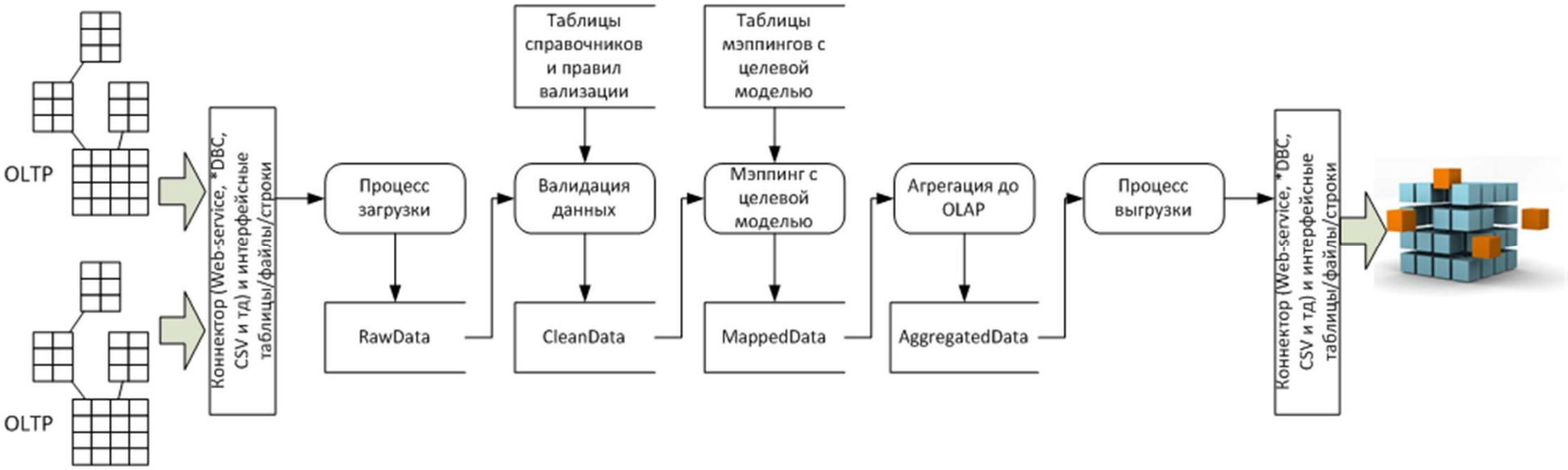
1.Процесс загрузки – Его задача затянуть в **ETL** данные произвольного качества для дальнейшей обработки, на этом этапе важно сверить суммы пришедших строк, если в исходной системе больше строк, чем в **RawData** то значит — загрузка прошла с ошибкой;



Стадии преобразования в ETL

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

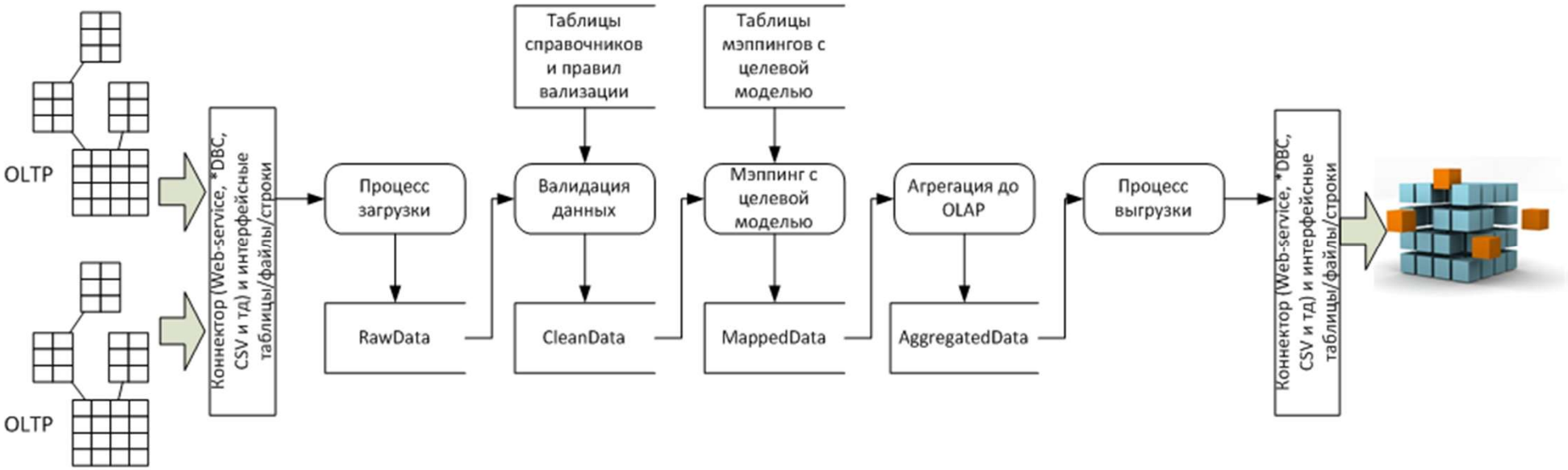
2. Процесс валидации данных – на этом этапе данные последовательно проверяются на корректность и полноту, составляется отчет об ошибках для исправления;



Стадии преобразования в ETL

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

3. Процесс мэппинга данных с целевой моделью – на этом этапе к валидированной таблице пристраивается еще n-столбцов по количеству справочников целевой модели данных, а потом по таблицам мэппингов в каждой пристроенной ячейке, в каждой строке проставляются значения целевых справочников. Значения могут проставляться как 1:1, так и *:1, так и 1:* и *:*, для настройки последних двух вариантов используют формулы и скрипты мэппинга, реализованные в **ETL**-инструменте;



Стадии преобразования в ETL

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

4. Процесс агрегации данных – этот процесс нужен из-за разности детализации данных в **OLTP** и **OLAP** системах.

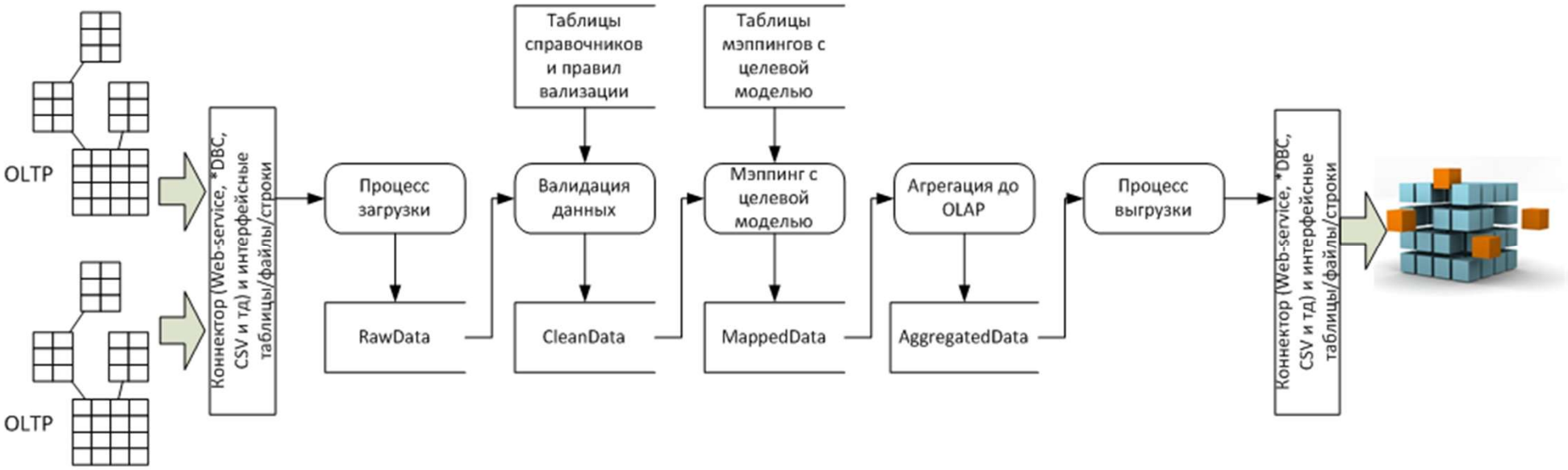
OLAP-система — это полностью денормализованная таблица фактов и окружающие ее таблицы справочников, максимальная детализация сумм OLAP – это количество перестановок всех элементов всех справочников.

OLTP-система может содержать несколько сумм для одного и того же набора элементов справочников.
Можно было-бы убивать OLTP-детализацию еще на входе в ETL, но тогда мы потеряли бы «аудиторский след». Этот след нужен для построения Drill-down отчета, который показывает — из каких строк OLTP, сформировалась сумма в ячейке OLAP-системы. Поэтому сначала делается мэппинг на детализации OLTP, а потом в отдельной таблице данные «схлопывают» для загрузки в OLAP;

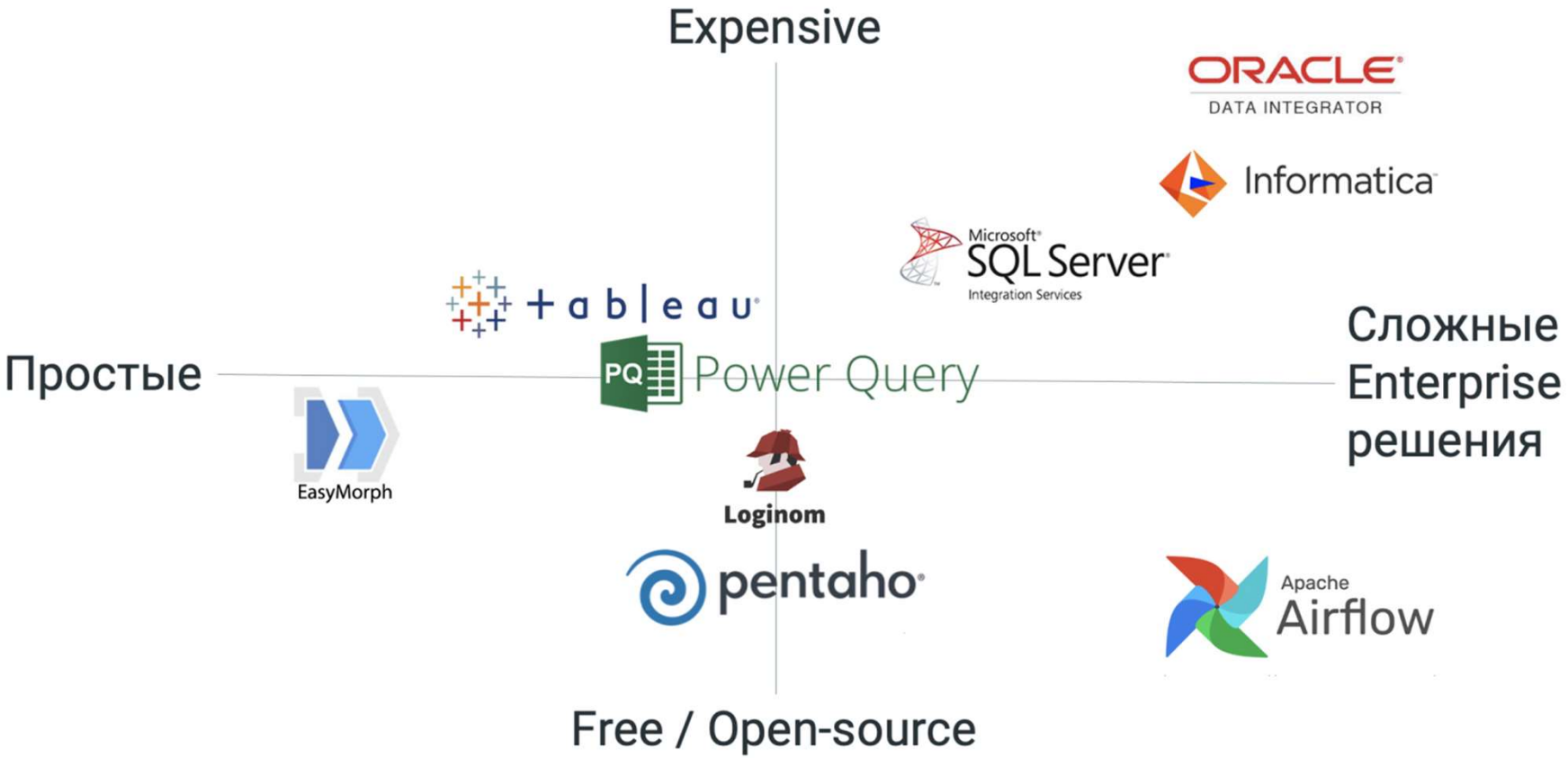
Стадии преобразования в ETL

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

5. Выгрузка в целевую систему — это технический процесс использования коннектора и передачи данных в целевую систему.



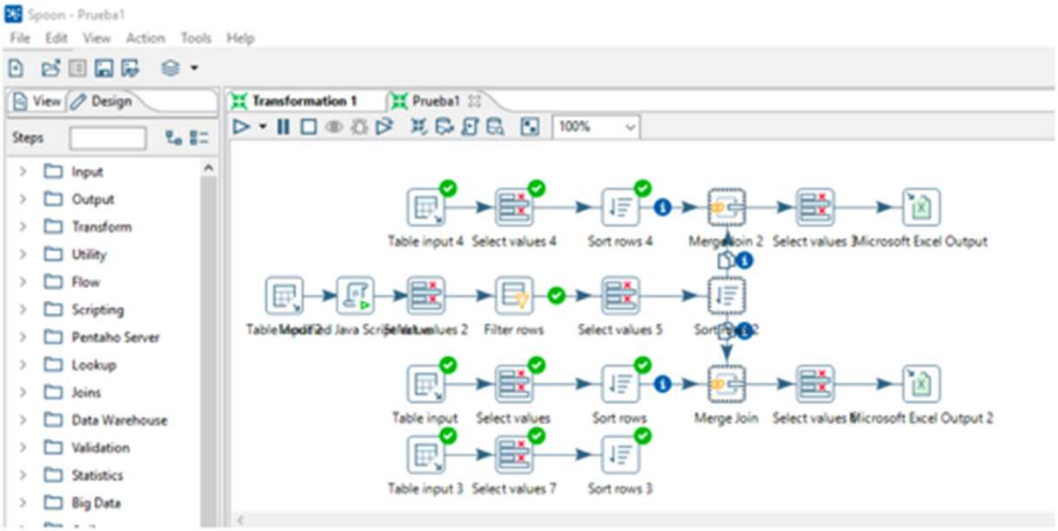
Рынок ETL


Expensive
ORACLE
DATA INTEGRATOR
Informatica
Microsoft
SQL Server
Integration Services
+ableau
Простые
PQ
Power Query
Сложные
Enterprise
решения
EasyMorph
Loginom
pentaho
Apache
Airflow
Free / Open-source

ETL-инструменты

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

VISUAL ETL vs SCRIPT ETL



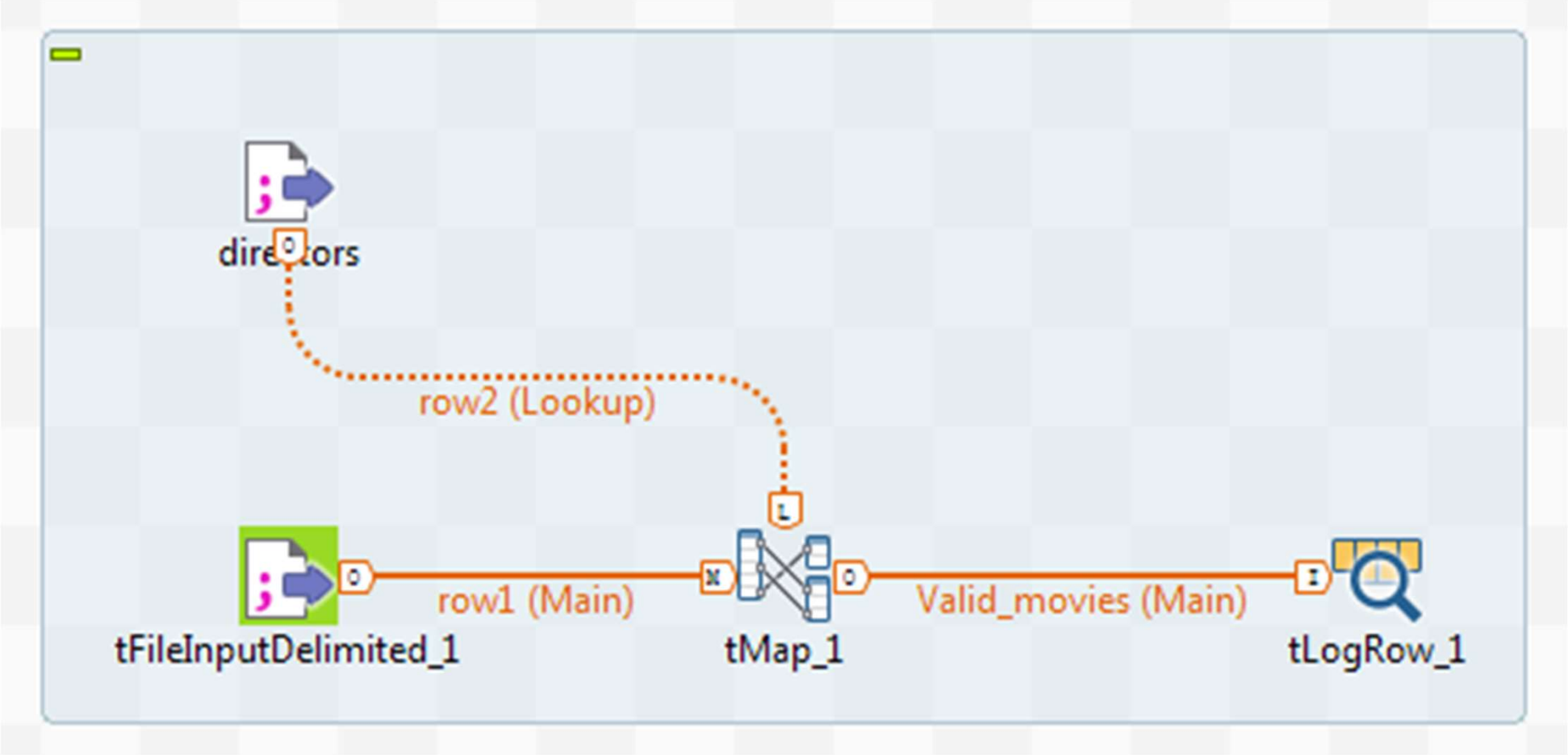

```
[0COSTCENTER_TEXT]:
LOAD
 [DATEFROM] as [DATEFROM.Valid-From Date],
 [DATETO] as [DATETO.Valid-to date],
 [KOKRS] as [KOKRS.Controlling Area],
 [KOSTL] as [KOSTL.Cost Center],
 [LANGU] as [LANGU.Language Key],
 [TXTMD] as [TXTMD.Medium description],
 [TXTSH] as [TXTSH.Short description];
SQL EXTRACTOR 0COSTCENTER_TEXT
TFRMETHOD T   // tRFC transfer method
//UPDMODE F   // full extractor
UPDMODE C   // initial extraction, to be followed by delta extractions
//UPDMODE D   // delta extraction
//INITRNR <NR>  // Resend extraction
//IDOC <NR>  // Resend single IDoc
EXTRLANGUAGE E
LOGSYS QTQVCEXTR2
WHERE
KOKRS I EQ 1000
;
//STORE * FROM [0COSTCENTER_TEXT] INTO FULL_0COSTCENTER_TEXT.QVD;
STORE * FROM [0COSTCENTER_TEXT] INTO INIT_0COSTCENTER_TEXT.QVD;
//LET vDate=Replace(now(),':','');
//STORE * FROM [0COSTCENTER_TEXT] INTO DELTA_0COSTCENTER_TEXT$(vDate).QVD;
DROP TABLE [0COSTCENTER_TEXT];
```

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Инструменты аналитики
Работа в ETL-системе Talend

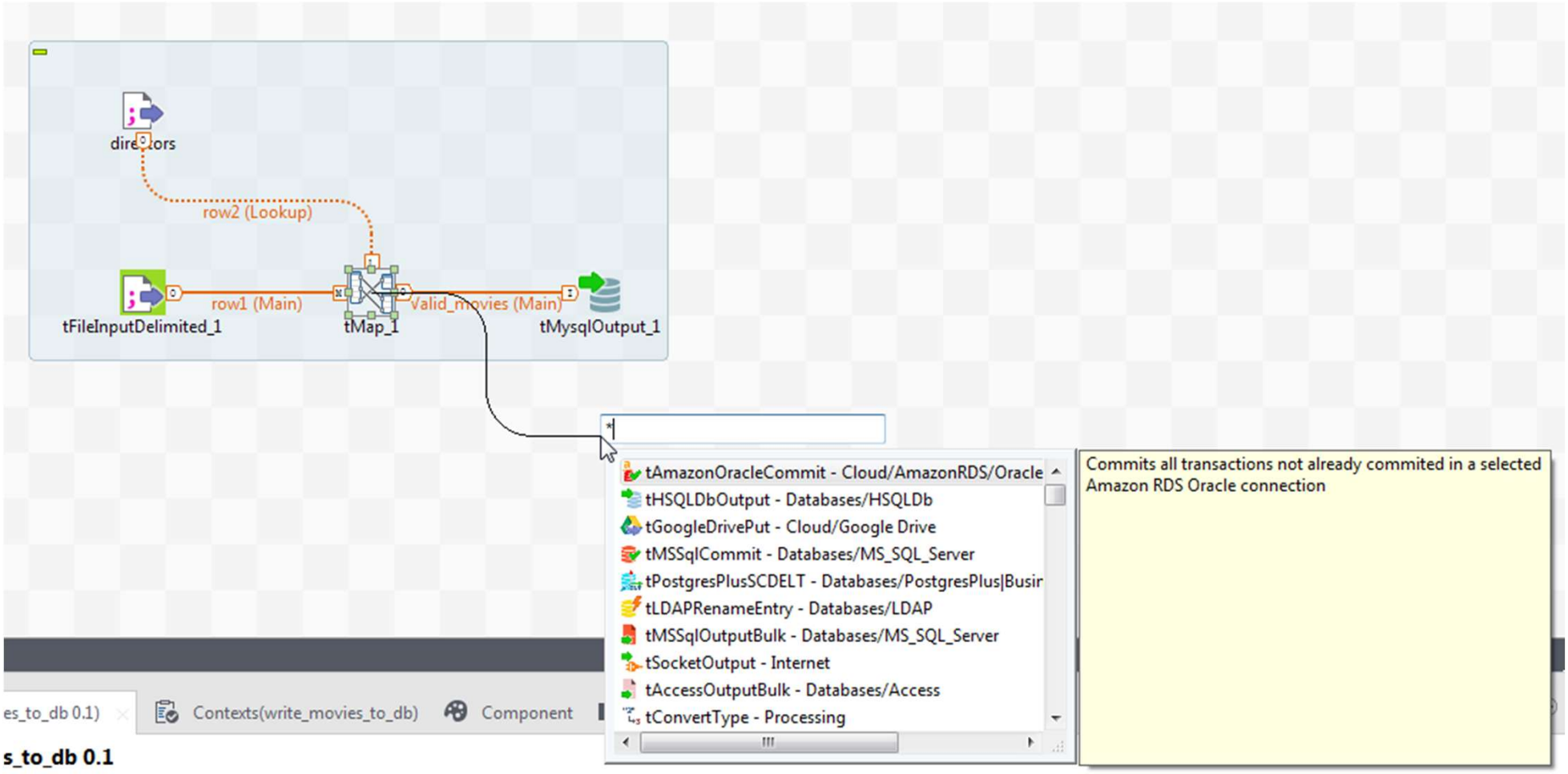
ETL Job1

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ



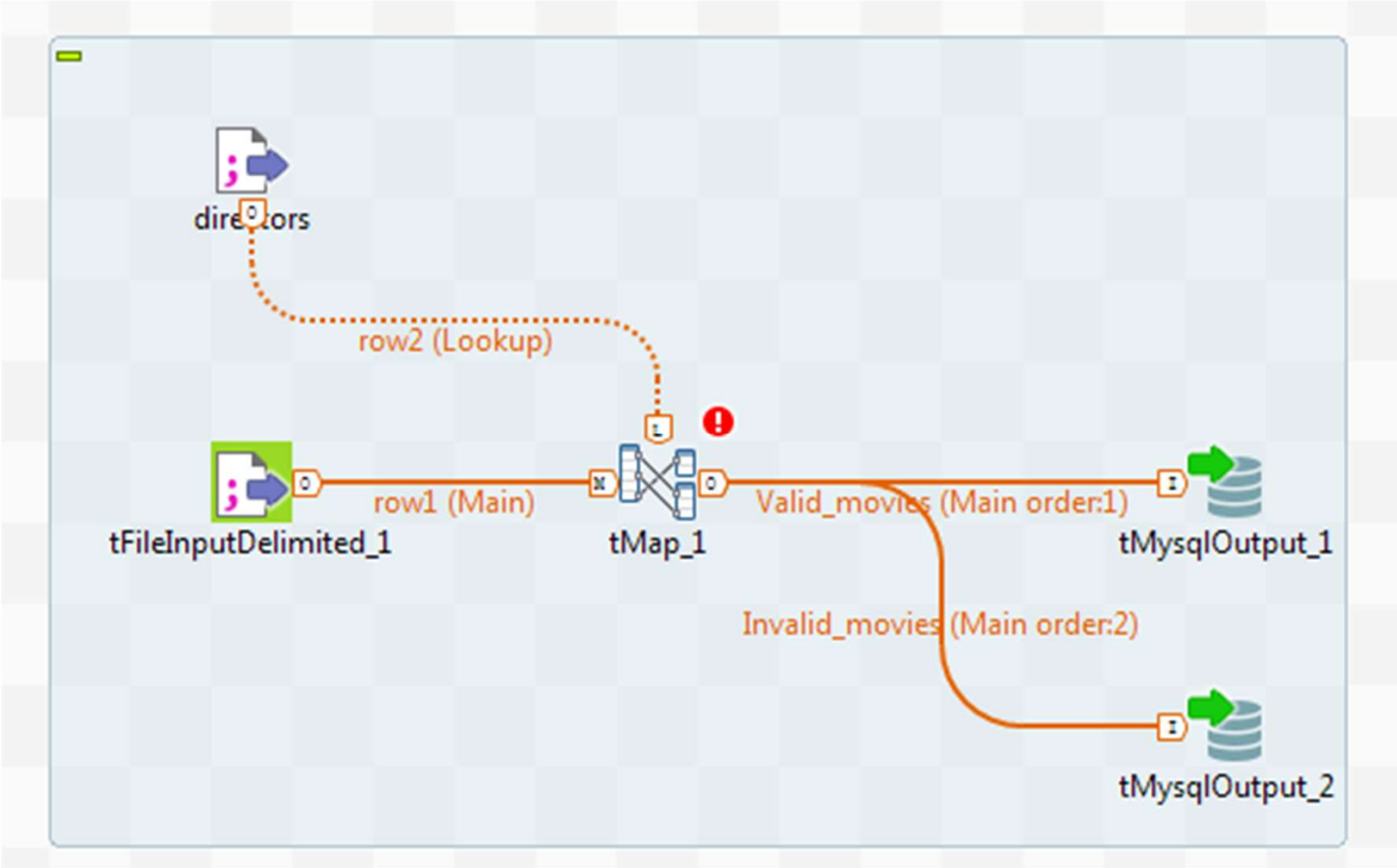
ETL Job2

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ



ETL Job3

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ



МОСКОВСКИЙ
ГОРОДСКОЙ
УНИВЕРСИТЕТ
МГПУ

СПАСИБО ЗА ВНИМАНИЕ