

|  |
| --- |
| Data Warehouse  Издательский дом «МИФ» |
|  |

| Related Artifacts | |
| --- | --- |
| Ref. | Name |
|  |  |
|  |  |

|  |  |
| --- | --- |
| Abbreviations and Acronyms | |
|  |  |
|  |  |

Contents

[1 Example of a 1st lvl heading 4](#_Toc499225911)

[1. Business Description 6](#_Toc499225912)

[1.1. Business background 6](#_Toc499225913)

[1.2. Problems because of poor data management 6](#_Toc499225914)

[1.3. Benefits from implementing a Data Warehouse 6](#_Toc499225915)

[2. Dimensions of a Business 7](#_Toc499225916)

[2.1. Business process 7](#_Toc499225917)

[2.2. The grain 7](#_Toc499225918)

[2.3. Dimensions 7](#_Toc499225919)

[2.3.1. Dim\_Time\_Day 7](#_Toc499225920)

[2.3.2. Dim\_Customers 8](#_Toc499225921)

[2.3.3. Dim\_Employees\_SCD 9](#_Toc499225922)

[2.3.4. Dim\_Stores 9](#_Toc499225923)

[2.3.5. Dim\_Products 10](#_Toc499225924)

[2.3.6. Dim\_Discounts\_SCD 10](#_Toc499225925)

[2.3.7. Dim\_Payments 11](#_Toc499225926)

[2.4. Фактовая таблица 12](#_Toc499225927)

[3. Logical Scheme 13](#_Toc499225928)

[4. Data Flow 14](#_Toc499225929)

[5. Fact Table Partitioning Strategy 15](#_Toc499225930)

[6. Strategy of Parallel Load 16](#_Toc499225931)

# Business Description

## Business background

«Манн, Иванов и Фербер» — московское издательство деловой литературы, начавшее свою деятельность в [2005 году](https://ru.wikipedia.org/wiki/2005_%D0%B3%D0%BE%D0%B4). Основатели: [Игорь Манн](https://ru.wikipedia.org/w/index.php?title=%D0%9C%D0%B0%D0%BD%D0%BD,_%D0%98%D0%B3%D0%BE%D1%80%D1%8C_%D0%91%D0%BE%D1%80%D0%B8%D1%81%D0%BE%D0%B2%D0%B8%D1%87&action=edit&redlink=1" \o "Манн, Игорь Борисович (страница отсутствует)), Михаил Иванов и Михаил Фербер (по первым буквам фамилий основателей и произошло название). Издательство специализируется на издании книг по [маркетингу](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%80%D0%BA%D0%B5%D1%82%D0%B8%D0%BD%D0%B3) и [менеджменту](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%BD%D0%B5%D0%B4%D0%B6%D0%BC%D0%B5%D0%BD%D1%82), [управлению временем](https://ru.wikipedia.org/wiki/%D0%A3%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%B5%D0%BC), адресованных собственникам бизнеса, менеджерам, консультантам, студентам и слушателям программы [MBA](https://ru.wikipedia.org/wiki/MBA). Кроме того, развиваются направления книг о спорте, психологии и кулинарии.

На сегодняшний день издательство «МИФ» является одним из самых прогрессивных на рынке СНГ: ведет активную маркетинговую политику, внедряет передовые технологии и использует нестандартный подход в решении бизнес-задач.

## Problems because of poor data management

Неправильное хранение и управление данными может привести к большим потерям ресурсов компании: времени, клиентов, партнеров – все это ведет к значительным денежным убыткам.

Список проблем, которые возникли у заказчика:

* Потеря клиентов, в связи с устаревшей или некорректной информацией.
* Провальные рекламные кампании.
* Убытки, вызванные решениям, принятыми на основе неверных данных.
* Упущенная прибыль: неосуществленные возможности получения прибыли в связи с неудачным выбором способа действий.

## Benefits from implementing a Data Warehouse

* **Историчность**: позволяет хранить данные за все время операционной деятельности компании, что позволяет анализировать данные за различные промежутки времени, отслеживать тренды, а также делать прогнозы о будущем развитии, нуждах и возможных проблемах.
* **Консистенность данных**: все данные хранятся в едином формате, что позволяет сделать анализ максимально точным.
* **Принятие решений**: позволяет максимально быстро получить доступ к необходимым данным и на их основе принять взвешенное решение.
* **Эффективность работы:** увеличивает эффективность бизнес-процессов организации.

# Dimensions of a Business

## Business process

Бизнес-процессом для издательского дома «МИФ» является учет продаж изданных книг.

## The grain

Одна строка из таблицы фактов представляет собой одну транзакцию в чеке: определенную книгу, которую работник продал потребителю в конкретном магазине в какой-либо день со скидкой (или без), а также каким способом осуществлялась оплата.

## Dimensions

Хранилище данных будет включать следующие измерения:

* Даты (Dim\_\_Time\_Day)
* Клиенты (Dim\_Customers)
* Работники (Dim\_Employees\_SCD)
* Магазины (Dim\_Stores)
* Продукты (Dim\_Products)
* Скидки (Dim\_Discounts\_SCD)
* Оплата (Dim\_Payments)

Перед описанием каждого измерения будут указаны бизнес-правила, которые были применены (если таковые имеются).

### Dim\_Time\_Day

* В данном измерении были использованы три индикатора (флага):
  + Weekend Indicator
  + Last Day Of Week Indicator
  + Last Day Of Month Indicator

Они могут принимать только два значения: true/false. Однако, были выбраны «говорящие» значения (напр. Weekend/weekday, а не Y/N), так как в дальнейшем такие значения будут удобнее для пользователей в процессе формирования отчетов, применения фильтров итд.

| Dim\_Time\_Day | | | |
| --- | --- | --- | --- |
| # | Name | Type | Description |
| 1 | Full Date DT | DATE | Primary Key: 12-NOV-2017 |
| 2 | Full Date Description | VARCHAR2(20) | 12 November 2017 |
| 3 | Day Of Week | NUMBER(1) | День недели (1-7) |
| 4 | Day Of Month | NUMBER(2) | День месяца (1-31) |
| 5 | Day Of Year | NUMBER(3) | День в году (1-366) |
| 6 | Day Name | VARCHAR2(10) | Monday-Sunday |
| 7 | Weekend Indicator | VARCHAR2(40) | Weekend/Weekday |
| 8 | Last Day Of Week Indicator | VARCHAR2(40) | Last day of the week/Not last day of the week |
| 9 | Last Day Of Month Indicator | VARCHAR2(40) | Last day of the month/Not last day of the month |
| 10 | Week Of Month | NUMBER(1) | Номер недели в месяце (1-5) |
| 11 | Week Of Year | NUMBER(2) | Номер недели в году (1-52) |
| 12 | Week Ending DT | DATE | Дата конца недели (воскресенье) |
| 13 | Month Number | NUMBER(2) | Номер месяца в году (1-12) |
| 14 | Month Name | VARCHAR2(30) | Название месяца |
| 15 | Month Year | VARCHAR2(30) | Формат: November-2017 |
| 16 | Quarter Number | NUMBER(3) | Номер квартала (1-4) |
| 17 | Quarter Year | VARCHAR2(30) | Формат: 4-2017 |
| 18 | Year Number | NUMBER(4) | Год |

### Dim\_Customers

* Так как анализ клиента в геогрфическом разрезе очень важен для заказчика, все суррогатные ключи на географические данные были сохранены.

|  |  |  |  |
| --- | --- | --- | --- |
| Dim\_Customers | | | |
| # | Name | Type | Description |
| 1 | Customer ID | NUMBER(8) | Primary Key (уникален для каждой записи) |
| 2 | Customer code | VARCHAR2(15) | Natural Key (может повторяться) |
| 3 | Customer Name | VARCHAR2(35) | Имя клиента |
| 4 | Customer Surname | VARCHAR2(35) | Фамилия клиента |
| 5 | Customer Email | VARCHAR2(65) | Емейл клиента |
| 6 | Customer Phone | VARCHAR2(30) | Номер мобильного телефона |
| 7 | Customer Age | NUMBER(3) | Возраст |
| 8 | Customer Card | VARCHAR2(20) | Номер карты: 4929 7798 1456 9983 |
| 9 | Customer Region ID | NUMBER(8) | PK региона |
| 10 | Customer Region | VARCHAR2(50) | Название региона |
| 11 | Customer Country ID | NUMBER(8) | PK страны |
| 12 | Customer Country | VARCHAR2(50) | Название страны |
| 13 | Customer City ID | NUMBER(8) | PK города |
| 14 | Customer City | VARCHAR2(50) | Название города |
| 15 | Customer Address ID | NUMBER(8) | PK адреса |
| 16 | Customer Address | VARCHAR2(50) | Адрес |
| 17 | Customer Postal Code | NUMBER(8) | Индекс |
| 18 | Insert\_DT | DATE | Дата добавления |
| 19 | Update\_DT | DATE | Дата измений аттрибута(ов) |

### Dim\_Employees\_SCD

* Это второй тип SCD, те здесь добавлена возможность точного отслеживания всех измененийю
* Если IsActive=Y, то end\_dt будет 31-DEC-9999’
* При заполнении данных будут добавлены значения по умолчанию для N/A (not applicable)и N/D (not defined).

|  |  |  |  |
| --- | --- | --- | --- |
| Dim\_Employees\_SCD | | | |
| # | Name | Type | Decription |
| 1 | Employee ID | NUMBER(8) | Primary Key (уникален для каждой записи) |
| 2 | Employee code | VARCHAR2(15) | Natural Key (может повторяться) |
| 3 | Employee Name | VARCHAR2(35) | Имя работника |
| 4 | Employee Surname | VARCHAR2(35) | Фамилия работника |
| 5 | Employee Email | VARCHAR2(65) | Емейл работника |
| 6 | Employee Phone | VARCHAR2(30) | Номер мобильного телефона |
| 7 | Department Name | VARCHAR2(30) | Название отдела |
| 8 | Manager ID | NUMBER(8) | РК менеджера (Employee Key) |
| 9 | Manager Name | VARCHAR2(70) | Имя и фамилия менеджера |
| 10 | Store ID | NUMBER(8) | РК магазина |
| 11 | Store Name | VARCHAR2(35) | Название магазина |
| 12 | Store City | VARCHAR2(50) | Город, где находится магазин |
| 13 | Store Address | VARCHAR2(50) | Адрес, где находится магазин |
| 14 | Start DT | DATE | Дата, когда состояние вступило в силу |
| 15 | End DT | DATE | Дата, когда состояние изменилось |
| 16 | IsActive | VARCHAR2(10) | Y (действителен)/N (не действителен) |
| 17 | Insert\_DT | DATE | Дата добавления |

### Dim\_Stores\_SCD

* Это второй тип SCD, те здесь добавлена возможность точного отслеживания всех изменений.
* Если IsActive=Y, то end\_dt будет 31-DEC-9999’
* При заполнении данных будут добавлены значения по умолчанию для N/A (not applicable)и N/D (not defined).

|  |  |  |  |
| --- | --- | --- | --- |
| Dim\_Stores | | | |
| # | Name | Type | Decription |
| 1 | Store Key | NUMBER(8) | Primary Key (уникален для каждой записи) |
| 2 | Store ID | VARCHAR2(15) | Natural Key (может повторяться) |
| 3 | Store Name | VARCHAR2(35) | Название магазина |
| 4 | Store Email | VARCHAR2(65) | Емейл магазина |
| 5 | Store Phone | VARCHAR2(30) | Номер магазина |
| 8 | Manager Key | NUMBER(8) | РК менеджера (Employee Key) |
| 6 | Manager Name | VARCHAR2(70) | Имя и фамилия менеджера |
| 7 | Store Region ID | NUMBER(8) | PK региона |
| 8 | Store Region | VARCHAR2(50) | Название региона |
| 9 | Store Country ID | NUMBER(8) | PK страны |
| 10 | Store Country | VARCHAR2(50) | Название страны |
| 11 | Store City ID | NUMBER(8) | PK города |
| 12 | Store City | VARCHAR2(50) | Название города |
| 13 | Store Address ID | NUMBER(8) | PK адреса |
| 14 | Store Address | VARCHAR2(50) | Адрес |
| 15 | Insert\_DT | DATE | Дата добавления |
| 16 | Update\_DT | DATE | Дата измений аттрибута(ов) |

### Dim\_Products

* При заполнении данных будут добавлены значения по умолчанию для N/A (not applicable)и N/D (not defined): аттрибуты, которые не предназначены для какой-либо категории будут заполнятся как N/A.

|  |  |  |  |
| --- | --- | --- | --- |
| Dim\_Products | | | |
| # | Name | Type | Decription |
| 1 | Product ID | NUMBER(8) | Primary Key (уникален для каждой записи) |
| 2 | Product code | VARCHAR2(15) | Natural Key (может повторяться) |
| 4 | Category Name | VARCHAR2(25) | Printed book, audiobook, e-book. |
| 5 | Category Description | VARCHAR2(150) | Описание категории |
| 6 | Product Name | VARCHAR2(50) | Название книги |
| 8 | Product Description | VARCHAR2(150) | Описание книги |
| 9 | Product Genre | VARCHAR2(50) | Жанр |
| 10 | Product Author | VARCHAR2(50) | Автор |
| 11 | Product Weight (kg) | NUMBER(3,3) | Вес книги: только для печатных книг |
| 12 | Product Weight (MB) | NUMBER(8,3) | Вес книги: для электронных и аудиокниг |
| 13 | Product Number of symbols | NUMBER(8) | Кол-во символов: лоя электронных книг |
| 14 | Product Durability (min) | NUMBER(4) | Длительность книги: для аудиокниг |
| 15 | Product Reader | VARCHAR2(70) | Кто читает книгу: для аудиокниг |
| 16 | Insert\_DT | DATE | Дата добавления |
| 17 | Update\_DT | DATE | Дата измений аттрибута(ов) |

### Dim\_Payments

* Это будет “Junk Dimension” – измерение, куда включены несколько логически несвязанных между собой значений. Это уместно, так как все эти значения имеют маленькую мощность (т.е. малое количество уникальных значений), а потому нет смысла хранить их в различных измерениях. В таком случае уменьшается размер фактовой таблицы, а также упрощается работа с моделью.

|  |  |  |  |
| --- | --- | --- | --- |
| # | Name | Type | Decription |
| 1 | Payment ID | NUMBER(8) | Primary Key (уникален для каждой записи) |
| 2 | Payment Type Group | VARCHAR2(25) | * Наличные * Кредитная карта, * Электронная валюта |
| 3 | Payment Type Description | VARCHAR2(25) | Наличные:   * BYN * USD   Кредитные карты:   * Visa, * MasterCard   Электронная валюта:   * Bitcoin * Etherum |
| 4 | Prepayment | VARCHAR2(25) | Предоплата:   * Prepayment * non-prepayment |
| 5 | Insert\_DT | DATE | Дата добавления |
| 6 | Update\_DT | DATE | Дата измений аттрибута(ов) |

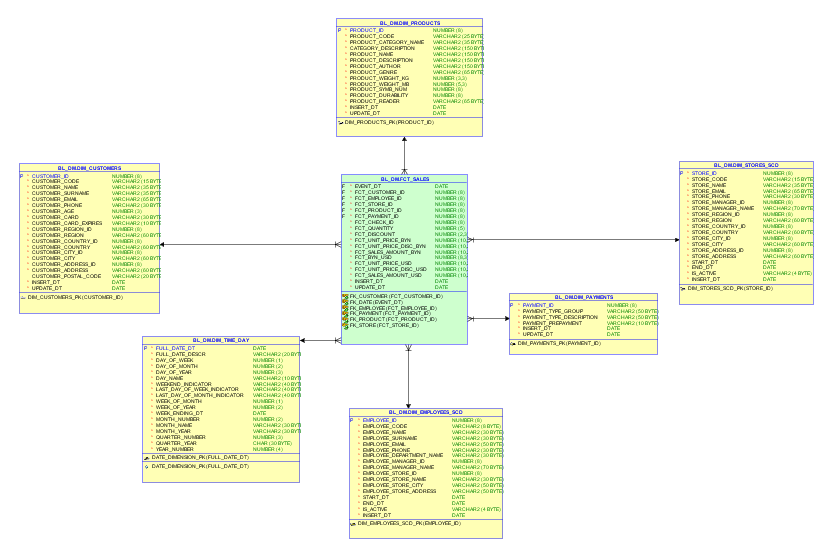
## Фактовая таблица

|  |  |  |  |
| --- | --- | --- | --- |
| # | Name | Type | Decription |
| 1 | Date ID | NUMBER(8) | FK на измерение |
| 2 | Customer ID | NUMBER(8) | FK на измерение |
| 3 | Employee ID | NUMBER(8) | FK на измерение |
| 4 | Store ID | NUMBER(8) | FK на измерение |
| 5 | Product ID | NUMBER(8) | FK на измерение |
| 6 | Discount ID | NUMBER(8) | FK на измерение |
| 7 | Currency ID | NUMBER(8) | FK на измерение |
| 8 | Payment ID | NUMBER(8) | FK на измерение |
| 9 | Check ID | NUMBER(8) | Дегенеративное измерение |
| 10 | Quantity | NUMBER(5) | Кол-во |
| 11 | Unit Price BYN | NUMBER(10,2) | Цена за штуку |
| 12 | Unit Price With Discount BYN | NUMBER(10,2) | Цена за штуку со скидкой |
| 13 | Sales Amount BYN | NUMBER(10,2) | Кол-во\*Цену за штуку |
| 14 | BYN\_USD | NUMBER(8,3) | Курс доллара к белорусскому рублю |
| 15 | Unit Price USD | NUMBER(10,2) | Цена за штуку в долларах |
| 16 | Unit Price With Discount USD | NUMBER(10,2) | Цена за штуку со скидкой в долларах |
| 17 | Sales Amount USD | NUMBER(10,2 ) | Кол-во\*Цену за штуку в долларах |
| 18 | Insert\_DT | DATE | Дата добавления |
| 19 | Update\_DT | DATE | Дата измений аттрибута(ов) |

# Logical Scheme

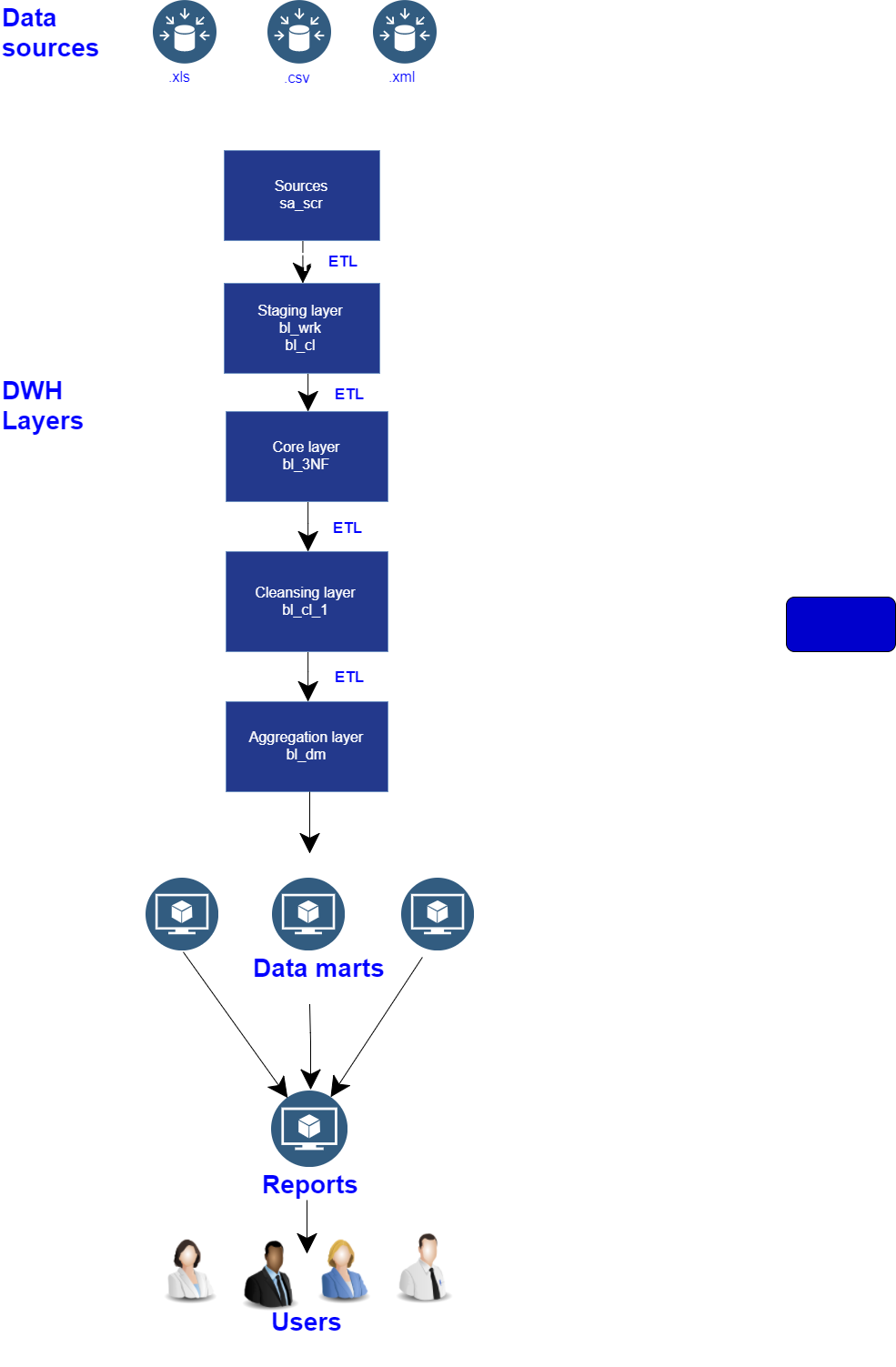
## Star scheme

Была выбрана схема Звезда, так как:

* Увеличивается производительность: так как все данные денормализованы, то уменьшается количество операций объединения таблиц (join) в запросах.
* Относительно небольшое количество данных позволяет хранить все в денормализованном виде.
* Быстрее осуществляется загрузка данных в таблицы: так как измерения и фактовые таблицы разделены, можно часто добавлять данные в фактовую таблицу, при этом записи в измерениях добавлять/изменять лишь при необходимости.  
    
  

## 3NF scheme

# Data Flow



## SA\_SRC

На этом этапе все источники в виде внешних таблиц (таблицы, доступныу только для чтения, метаданные которых хранятся в базе данных, а данные – вне базы данных) добавляются в базу данных.

## STAGING LAYER

### BL\_WRK

На этом этапе все данные из внешних таблиц загружаются в таблицы бд: это делается для увеличения производительности, так как считывание из бд быстрее, чем считывание с диска.

Таким образом, структуры таблиц sa\_src и bl\_wrk совпадают.

## BL\_CL

На этом этапе все данные очищаются,фильтруются, все null-значения заменяются на предопределенные значения (N/A, N/D). Именно на этом этапе осуществляется маппинг данных из разных таблиц.

Структура bl\_cl таблиц такая же как у bl\_3NF.

## BL\_3NF

На этом этапе уже полностью очищенные и подготовленные данные расидываются по таблицам в третьей нормальной форме для их дальнейшей загрузки в DWH. Здесь уже появляются все необходимые ограничение (PK, FK) для поддержания ссылочной целостности.

## BL\_CL\_DM

## BL\_DM

# Fact Table Partitioning Strategy

Партицирование будет осуществляться по полю Full\_Date\_DT, так как клиента в основном будут интересовать заказы, соврешенные за последние 5 лет.

Будет использоваться Range partition.

Внутри также будет партицирование (composite partitioning) по Sales\_amount, так как заказчику часто необходимо разграничивать крупных, средних и мелких покупателей.

# Strategy of Parallel Load

| REVISION HISTORY | | | | | |
| --- | --- | --- | --- | --- | --- |
| Ver. | Description of Change | Author | Date | Approved | |
| Name | Effective Date |
| n.n |  |  | dd-Mmm-yyyy |  | dd-Mmm-yyyy |
|  |  |  |  |  |  |
|  |  |  |  |  |  |