

Образовательная автономная некоммерческая организация высшего
образования «Московский технологический институт»

Практическая работа на тему:
«Корпоративные информационные системы»

Подготовил студент группы:

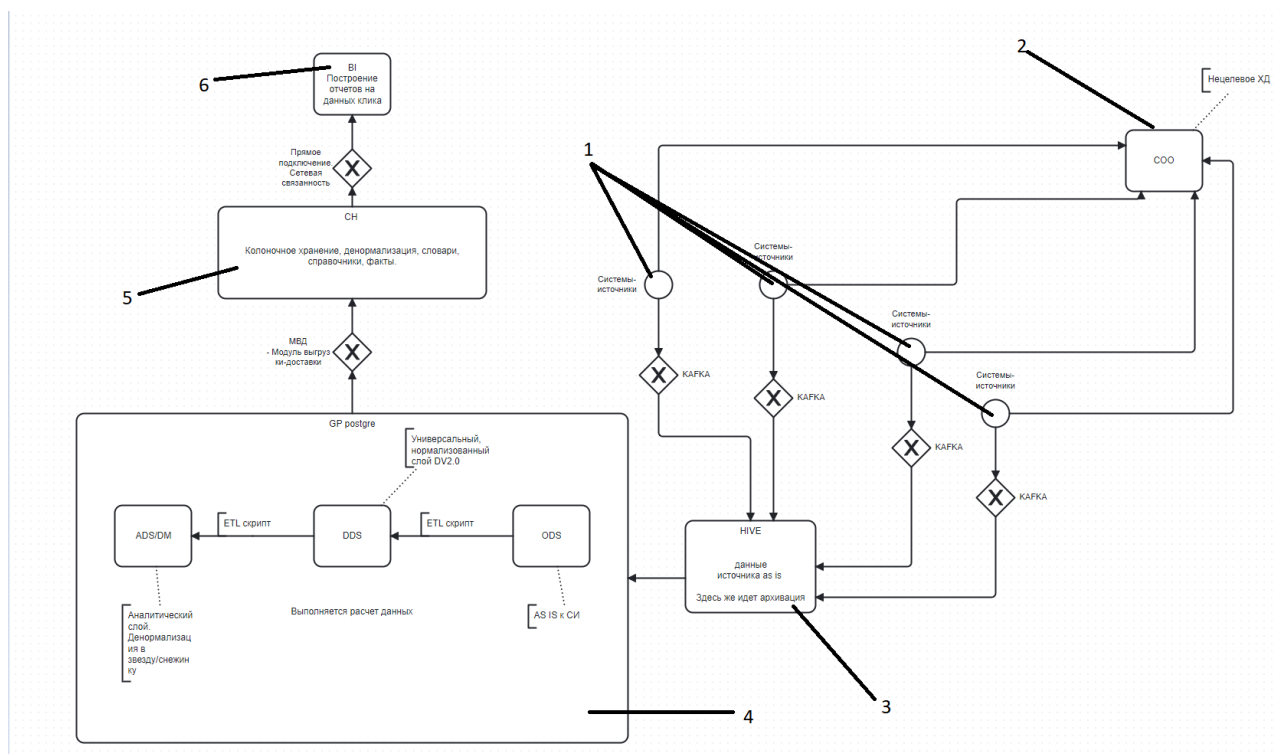
ОДБП-402э

А.Ю. Руденко

Образовательная автономная некоммерческая организация высшего образования «Московский технологический институт»

Корпоративные информационные системы в данном документе будут разобраны на примере реальных задач в рамках миграции нецелевых баз данных в централизованное хранилище данных (далее – ЦХД)

На рисунке ниже представлена схема жизненного цикла данных в ЦХД.



Краткое описание движения данных в рамках миграции

Данные с систем-источников – СИ (1 на рисунке) поступают напрямую в нецелевое хранилище Систему Объективной Отчетности – COO (2 на рисунке)

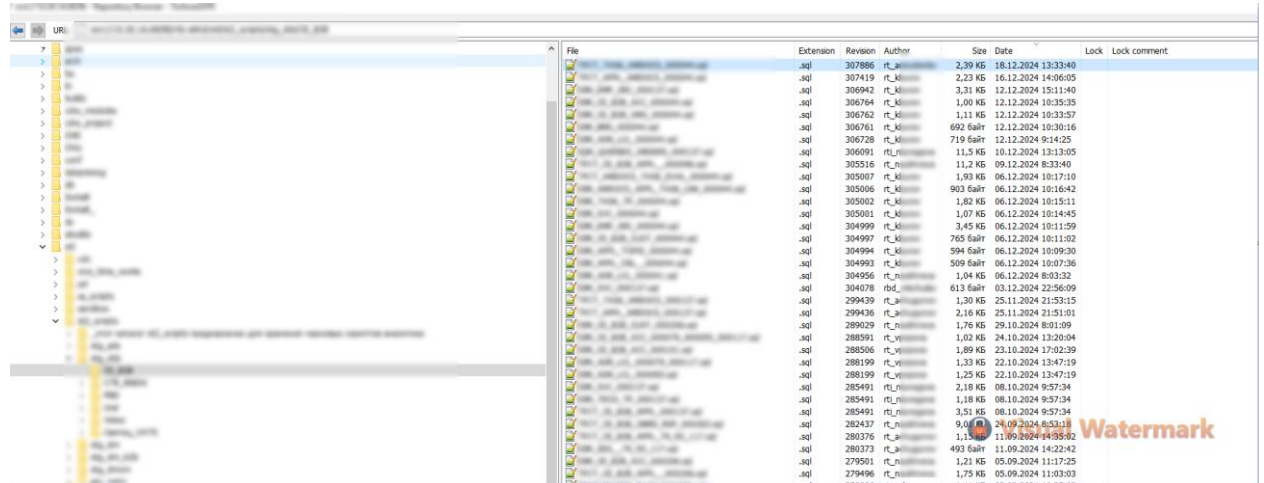
В рамках реверс-инженеринга данные с СИ через KAFKA (собственной сборки) поступают на контур HIVE (3 на рисунке, часть ЦХД, используется для архивации), затем попадает в БД Postgre под СУБД GreenPlum (4 на фото) где происходят расчеты данных на различных слоях (об этом будет написано дальше) и через Модуль выгрузки-доставки (далее МВД) попадают в БД ClickHouse (5 на рисунке, поколоночное хранение) для того чтобы в дальнейшем отдавать данные пользователям BI, который представлен superset от apache собственной сборки (6 на рисунке)

Основным этапом миграции является сбор, нормализация и подготовка данных в GP для дальнейшего экспорта в CH.

В рамках рабочих задач аналитики и разработчики пользуются внутренним файловым обменником TortoiseSVN – самый обыкновенный файловый репозиторий с отслеживанием версий. В нем хранятся все артефакты необходимые для создания, описания и эксплуатации

Образовательная автономная некоммерческая организация высшего образования «Московский технологический институт»

объектов (БФТ, са-скрипты, патчи БД, ETL)



File	Extension	Revision	Author	Size	Date	Lock	Lock comment
...	.sql	307886	rt.kl	2,39 KB	18.12.2024 13:33:40		
...	.sql	307419	rt.kl	2,23 KB	16.12.2024 14:06:05		
...	.sql	306942	rt.kl	3,31 KB	12.12.2024 15:11:40		
...	.sql	306764	rt.kl	1,00 KB	12.12.2024 10:35:25		
...	.sql	306762	rt.kl	1,11 KB	12.12.2024 10:33:57		
...	.sql	306761	rt.kl	692 байт	12.12.2024 10:30:16		
...	.sql	306728	rt.kl	719 байт	12.12.2024 9:14:25		
...	.sql	306091	rt.kl	11,5 KB	10.12.2024 13:13:05		
...	.sql	305516	rt.kl	11,2 KB	09.12.2024 8:33:40		
...	.sql	305067	rt.kl	1,93 KB	06.12.2024 10:17:10		
...	.sql	305006	rt.kl	903 байт	06.12.2024 10:16:42		
...	.sql	305002	rt.kl	1,82 KB	06.12.2024 10:15:11		
...	.sql	305001	rt.kl	1,07 KB	06.12.2024 10:14:45		
...	.sql	304999	rt.kl	3,45 KB	06.12.2024 10:11:59		
...	.sql	304997	rt.kl	765 байт	06.12.2024 10:11:02		
...	.sql	304994	rt.kl	594 байт	06.12.2024 10:09:30		
...	.sql	304993	rt.kl	509 байт	06.12.2024 10:07:36		
...	.sql	304956	rt.kl	1,04 KB	06.12.2024 8:03:32		
...	.sql	304078	rt.kl	613 байт	03.12.2024 22:56:09		
...	.sql	299439	rt.kl	1,30 KB	25.11.2024 21:53:15		
...	.sql	299438	rt.kl	2,16 KB	25.11.2024 21:51:01		
...	.sql	289029	rt.kl	1,76 KB	29.10.2024 8:01:09		
...	.sql	288591	rt.kl	1,02 KB	24.10.2024 13:20:04		
...	.sql	288506	rt.kl	1,89 KB	23.10.2024 17:02:39		
...	.sql	288199	rt.kl	1,33 KB	22.10.2024 13:47:19		
...	.sql	288199	rt.kl	1,25 KB	22.10.2024 13:47:19		
...	.sql	285491	rt.kl	2,18 KB	08.10.2024 9:57:34		
...	.sql	285491	rt.kl	1,18 KB	08.10.2024 9:57:34		
...	.sql	285491	rt.kl	3,51 KB	08.10.2024 9:57:34		
...	.sql	282437	rt.kl	9,06 KB	24.09.2024 8:53:18		
...	.sql	280376	rt.kl	1,13 KB	11.09.2024 14:58:02		
...	.sql	280373	rt.kl	493 байт	11.09.2024 14:22:42		
...	.sql	279501	rt.kl	1,21 KB	05.09.2024 11:17:25		
...	.sql	279496	rt.kl	1,75 KB	05.09.2024 11:03:03		

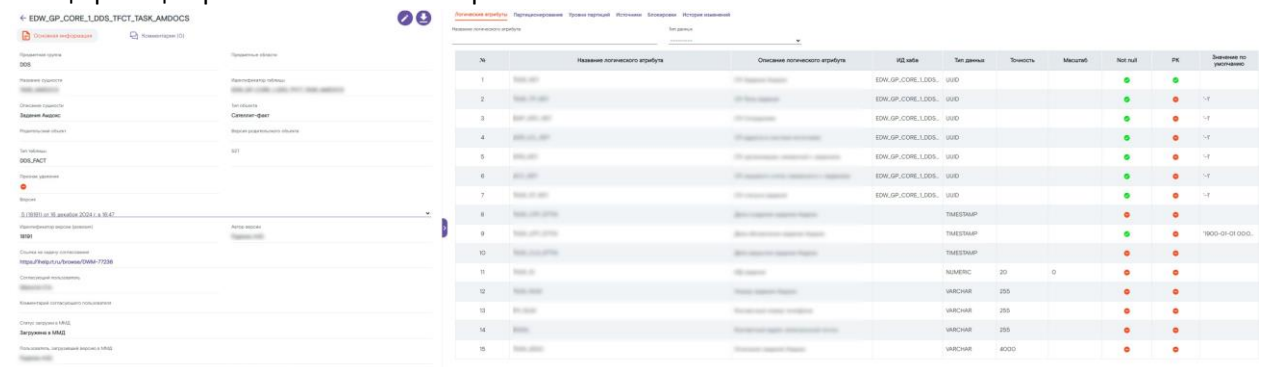
Просматривая схему (рис1) видно, что в целом процесс является ELT – Extract Load Transform.

Но внутри GP из-за особенностей хранения данных послойно – Оперативный (ODS), Детальный (DDS), Аналитический (ADS/DM) движение данных реализуется посредством ETL – Extract, transform Load через промежуточные объекты - STG-слой (чуть более подробно будет описано дальше).

Итак, имея Бизнес-функциональное требование (БФТ) прежде всего поговорим о создании объектов (таблиц) в нашей базе данных.

Любой объект (таблица/витрина/представление) формируется строго на его метаданных - т.е. сначала необходимо “скормить” его спецификацию в Управляющий механизм (УМ) БД.

Спецификация реализована в веб-сервисном ПО DataGovernance



№	Название логического атрибута	Описание логического атрибута	ID атрибута	Тип данных	Длина	Масштаб	Not null	PK	Значения по умолчанию
1	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	UUID					
2	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	UUID					-1
3	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	UUID					-1
4	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	UUID					-1
5	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	UUID					-1
6	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	UUID					-1
7	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	UUID					-1
8	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	TIMESTAMP					
9	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	TIMESTAMP					9999-01-01 00:00:00
10	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	TIMESTAMP					
11	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	NUMERIC	20	0			
12	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	VARCHAR	255				
13	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	VARCHAR	255				
14	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	VARCHAR	255				
15	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	...	EDW_GP_CORE_1.DDS_TPCT_TASK_ANDCOS	VARCHAR	4000				

После этого создается патч базы данных с процедурой postgresSQL которая физически создаст таблицу на узлах кластера.

Образовательная автономная некоммерческая организация высшего образования «Московский технологический институт»

```

-- EDW_GP_CORE_1_DDS
-- <BLOCK ENTITY="edw_core_1">
-- "TYPE="setup" KIND="DDL" COMPONENT="CDWH_GP" LAYER="dds" SOURCE="000001",
-- "TYPE="table">
do $$
declare
    r integer;
begin
    edw_stg_ddl = '
CREATE TABLE edw_stg_ddl (
    id VARCHAR(512) NOT NULL,
    client_type VARCHAR(512) NOT NULL default ''-1'',
    client_type_link VARCHAR(512) NOT NULL default ''-1'',
    client_type_link_id VARCHAR(100) NOT NULL default ''-1'',
    client_type_link_id_id VARCHAR(512) NOT NULL default now(),
    client_type_link_id_id_id VARCHAR(512) NOT NULL default -1,
    client_type_link_id_id_id_id VARCHAR(512) NOT NULL default -1,
    client_type_link_id_id_id_id_id VARCHAR(512) NOT NULL default ''00000000000000000000000000000000'',
    client_type_link_id_id_id_id_id_id VARCHAR(512) NOT NULL default ''1900-01-01 00:00:00::timestamp without time zone'',
    client_type_link_id_id_id_id_id_id_id VARCHAR(512) NOT NULL default ''2999-12-31 00:00:00::timestamp without time zone''
);
';
    with (appendonly=true, compressstype=zstd, compresslevel=3)
distributed randomly
;
comment on table edw_stg_ddl is 'Таблица для хранения данных о клиентах (version 12.1)';
comment on column edw_stg_ddl.id is 'Ид клиента';
comment on column edw_stg_ddl.client_type is 'Тип клиента';
comment on column edw_stg_ddl.client_type_link is 'Связь клиента с типом клиента (version 12.1)';
comment on column edw_stg_ddl.client_type_link_id is 'Ид связи клиента и типа клиента';
comment on column edw_stg_ddl.client_type_link_id_id is 'Системная дата и время создания записи';
comment on column edw_stg_ddl.client_type_link_id_id_id is 'Идентификатор системы источника';
comment on column edw_stg_ddl.client_type_link_id_id_id_id is 'Идентификатор запуска потока загрузки';
comment on column edw_stg_ddl.client_type_link_id_id_id_id_id is 'MD5 значений записи';
comment on column edw_stg_ddl.client_type_link_id_id_id_id_id_id is 'Идентификатор запуска потока загрузки, обновившего запись';
comment on column edw_stg_ddl.client_type_link_id_id_id_id_id_id_id is 'Дата и время начала действия записи';
comment on column edw_stg_ddl.client_type_link_id_id_id_id_id_id_id_id is 'Дата и время окончания действия записи';
alter table edw_stg_ddl.t_000
do $$
declare
    r integer;

```

После того, как структуры подготовлены – необходимо создать второй артефакт объекта – s2t – описание трансформации витрины. Простыми словами – как и какие данные должны в нее попадать. А еще проще – написать селект к таблицам на которых строится объект.

Данный документ так же формируется в DataGovernance

EDW-STG-MDM-PUT-DWG-DM-DM-WF-T4		
Добавить +		
Основная информация История изменений Блокировки Источники Структура dim_... PUT_...		
<div> <div>✎ ⚙ 📄</div> <div>ЦХД</div> <div>EDW_STG_MDM . PUT_...</div> </div>		
Описание		Версия
Данные по фискалам СИ для маппинга на эталон		От 21.06.2023 00:37 (32 версия)
Отчеты из реестра		Спецификация
{272_Test}		Слой
Параметры партнеризации		EDW_STG_MDM
Параметры распределения		Ссылка на S-BT
Бизнес-сегмент		
Все сегменты		
Согласующая роль	Статус согласования	Комментарий
Ведущий функциональный разработчик	Согласовано	

Главным здесь является написание аналитиком sql-скрипта, который далее будет обёрнут в патч ETL и передан на PROD в команду сопровождения.

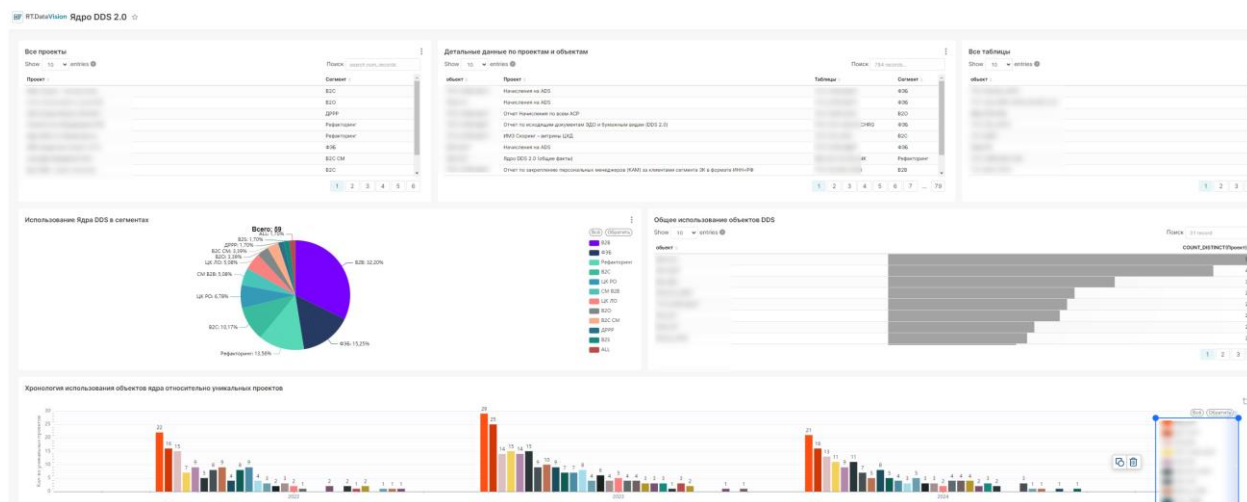
Патч ETL еще называют ETL-скриптом – именно за счет вшитых в скрипт правил (сценариев и подписок) - таблица будет автоматически загружаться по заданным правилам.

г. Москва, 2024 г.

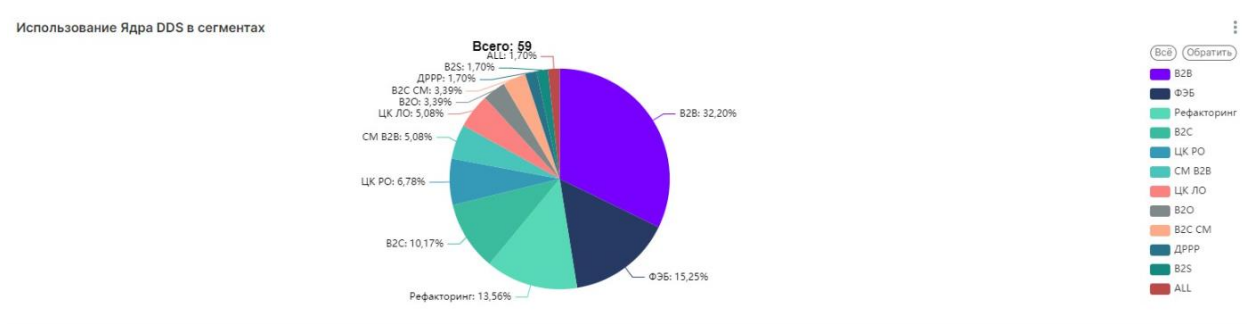
Образовательная автономная некоммерческая организация высшего образования «Московский технологический институт»

Сам BI реализован на корпоративной ИС (RTDV) – это сборка суперсет от Апачи. Пока что находится в доработках, но уже используется юзерами

Как выглядит стандартный BI

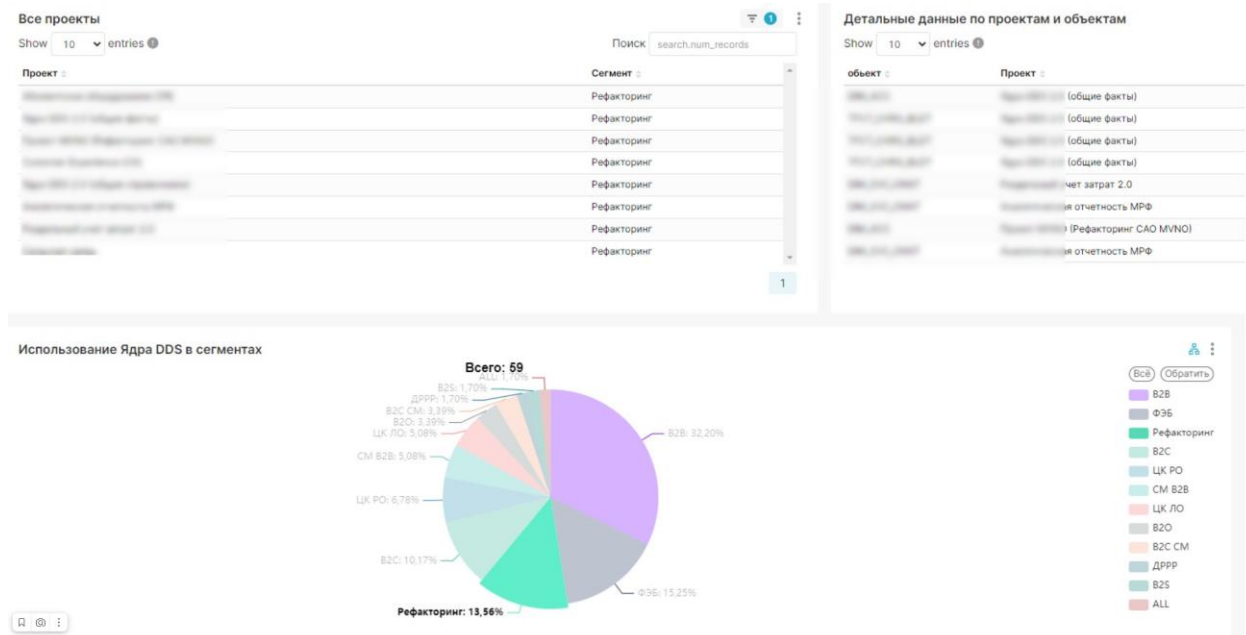


В данном случае – представлена информация об использовании объектов(таблиц) детального слоя DDS2.0 в различных проектах



Весь дашборд интерактивный – т.е. например при выборе сегмента команды – получаем фильтрацию на остальных графиках и сводках.

Образовательная автономная некоммерческая организация высшего образования «Московский технологический институт»



Краткий итог проведенной работы – жизненный цикл данных в хранилище включает в себя – генерацию данных на источнике, загрузку, обработку, нормализацию, представление и архивацию.

За счет Корпоративных ИС, таких как SVN, DataGovernance, RTDV, Kafka – специалистам удастся слаженно и своевременно выполнять свою работу, проекты и в конечном счете передавать готовые для анализа данные в бизнес, чтобы он продолжал развиваться, принимать решения, богатеть чтобы повышать зарплаты тем, кто обеспечивает технические и технологические процессы.