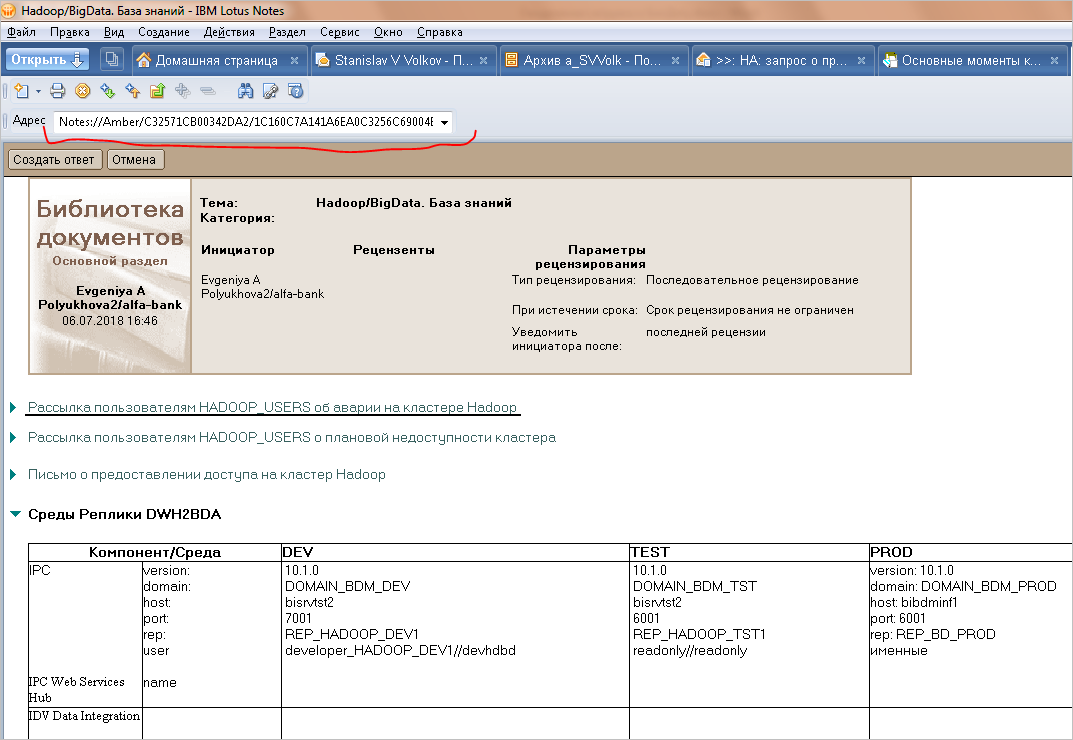
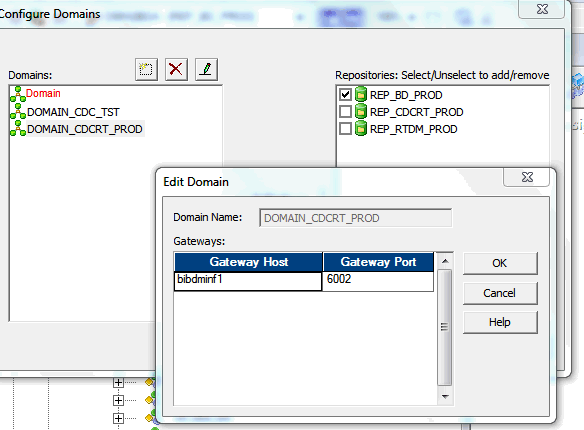
Ежедневная загрузка в БигДату

**База знаний Hadoop:** <Notes://Amber/C32571CB00342DA2/1C160C7A141A6EA0C3256C69004BAF87/4EFD5454F568955A432582C2004BABB8>



Репозиторий REP\_BD\_PROD (9.6.1) **support / support**

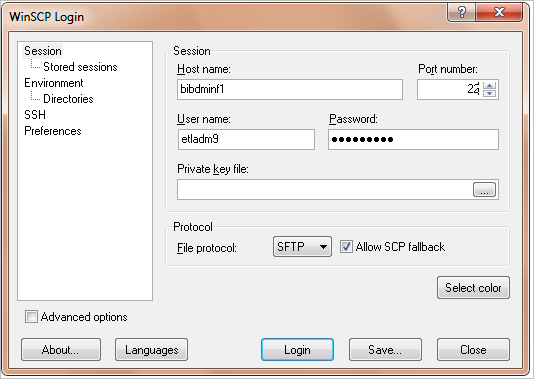


Сервер: **bibdminf1 ( etladm9** / **vpd92smed** учетка с большими привилегиями, которую дал Макс – осторожно!**)**

После перехода на **10**-ую информатику **etladmin // aci32stn**

Папка с файлами /app/informatica/pc**10**/server/infa\_shared/BD\_PROD/TgtFiles

Подключаться через WinSCP:



**hue консоль** (он же доступ в hadoop):

логин/пароль: **u\_m0g4q** / **u\_m0g4q** (маленькими буквами)

[http://bda**1**1node04:8888/metastore/tables/s\_dwh](http://bda11node04:8888/metastore/tables/s_dwh) - **tech\_oozie // tech\_oozie1**

**второй кластер:**

[http://bda**2**1node04:8888/metastore/tables/s\_dwh](http://bda21node04:8888/metastore/tables/s_dwh)

**третий кластер:**

[http://bda**3**1node04:8888/metastore/tables/s\_dwh](http://bda31node04:8888/metastore/tables/s_dwh)

**Поток переехал на Информатику 10.1.0:**

(как к ней подключиться описано в D:\MDM\Подключение к Sandbox ( REP\_SANDBOX\_DEV1) Информатика 10.1.1.docx)

Письма Макса:

Коллеги, создал репозиторий REP\_BD\_PLT:

version: 10.1.0

domain: DOMAIN\_BDM\_PROD

host: bibdminf1

port: 6005

пользователи именные, Стас, тебя создал **SVVolkov // SVVolkov** пароль поменяешь, Жене дал полномочия на этот репозиторий.

Что требуется сделать - взять пару поточков и прогнать на этом репозитории, проверить что всё хорошо. Скрипты менять не надо, прям как есть.

Заодно кое что попробуем. Как будете стартовать - напишите мне.

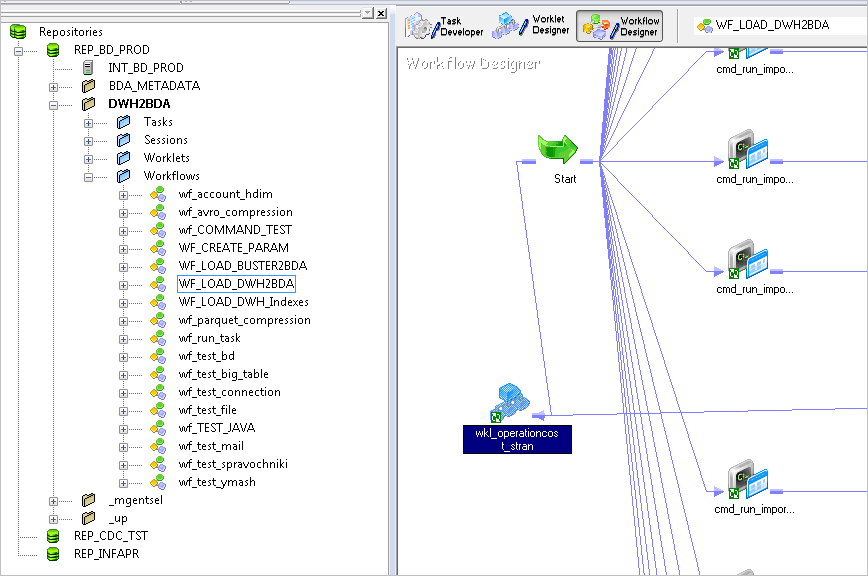
---  
С уважением,  
Максим Генцель  
6268.

Ещё момент, не PLT, а **PROD**, говорю же, как и было.

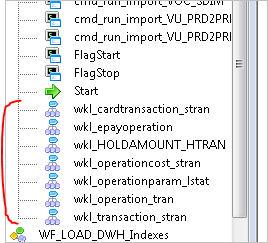
---  
С уважением,  
Максим Генцель  
6268.

etladmin // aci32stn

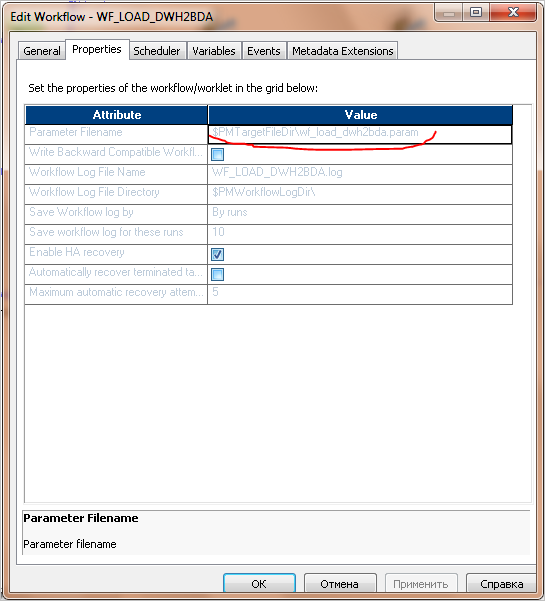
Основной поток **WF\_LOAD\_DWH2BDA** отвечающий за загрузку таблиц:

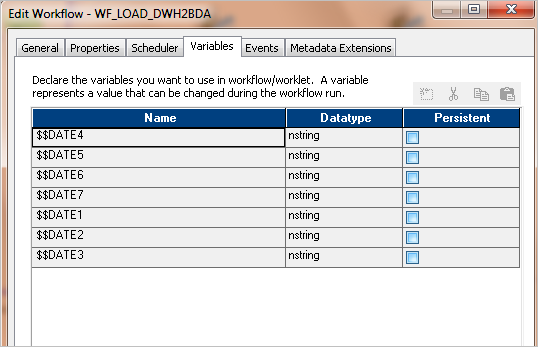


Помимо cmd-шников в нем запускаются ворклеты:

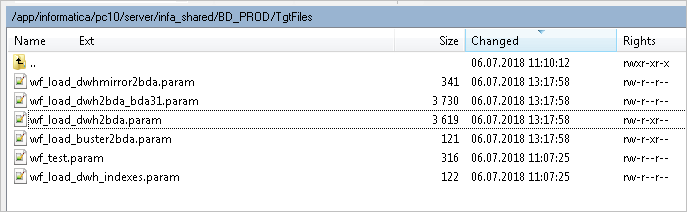


Поток берет параметры (даты) из файла:

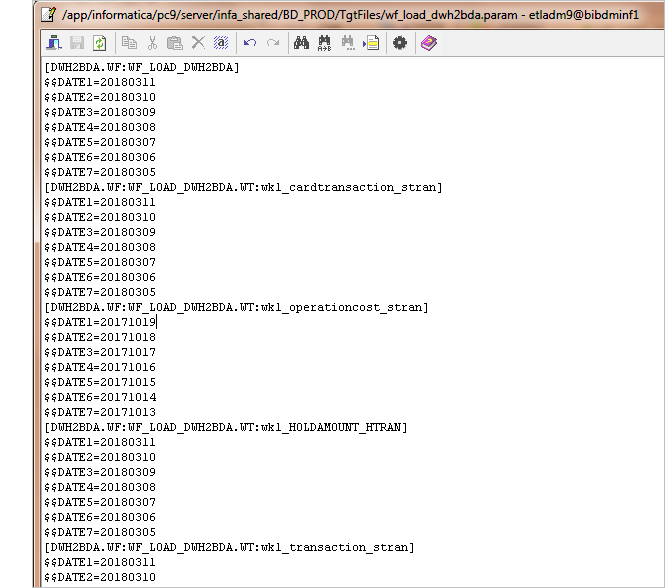




Сам файл параметров находится здесь:

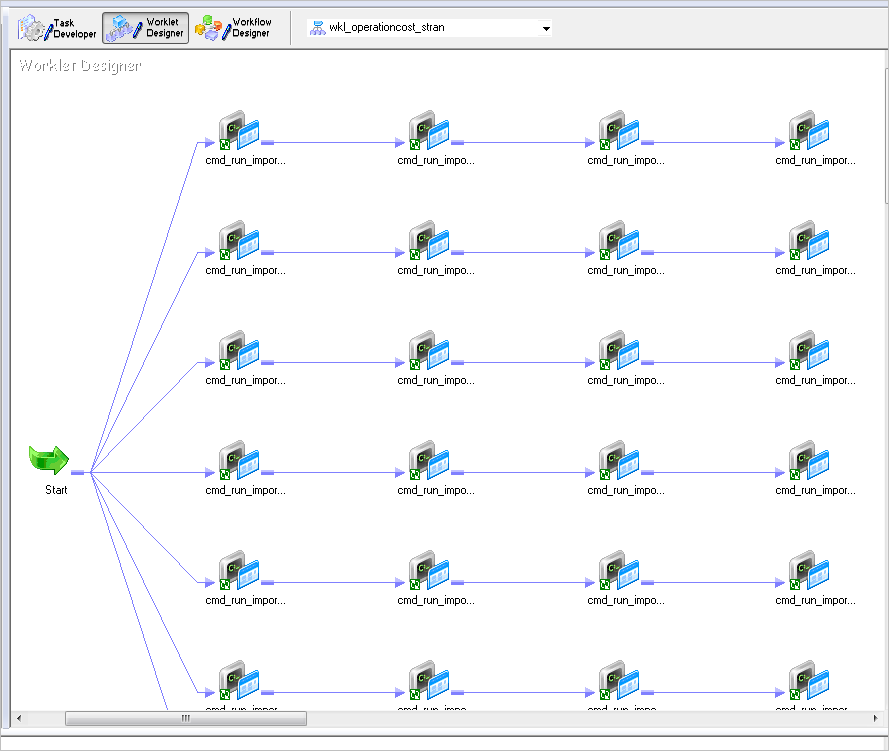


и выглядит так:

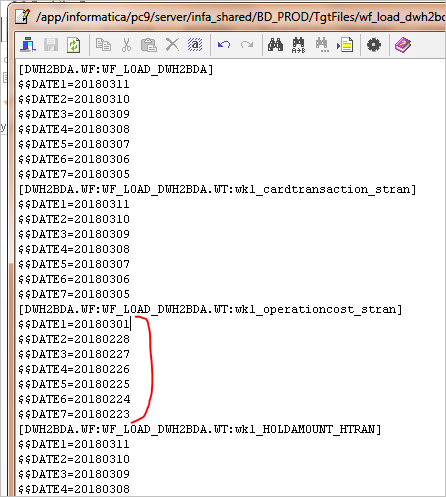


Как добавить загрузку таблицы.

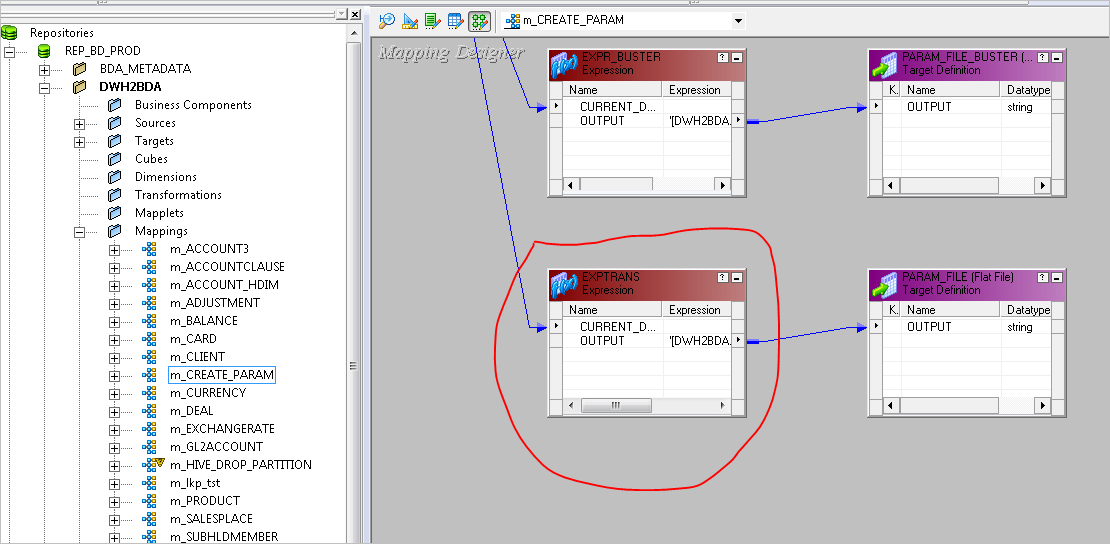
На примере загрузки таблицы OPERATIONCOST\_STRAN, надо создать worklet **wkl\_operationcost\_stran** в котором семь веток. Одна ветка один день.



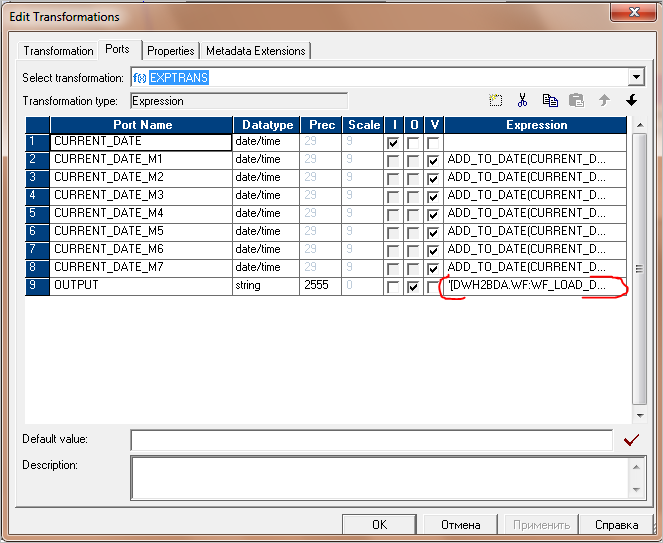
Для того чтобы загрузить неделю, выставляем даты в файле параметров(если в ручную):



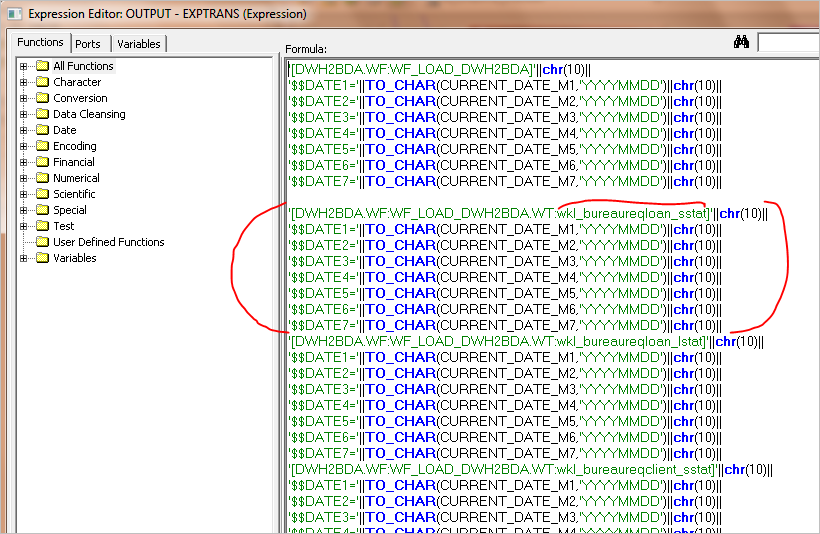
Для того чтобы параметры формировались потоком, в мапинг m\_CREATE\_PARAM добавляем в экспрешн той ветки мапинга которая отвечает за поток загрузки:



В значении PORT – OUTPUT:



Строки для нового ворклета:

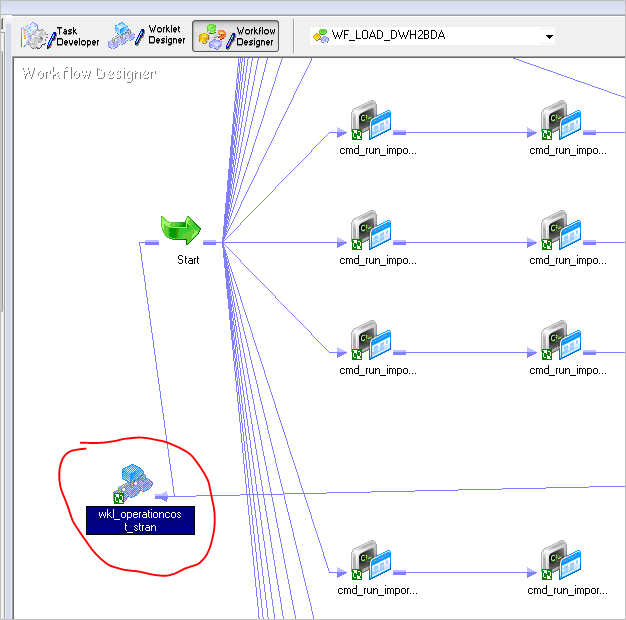


И в потоке **WF\_LOAD\_DWH2BDA** запускаем отдельно ворклет **wkl\_operationcost\_stran** (правая кнопка - StartTask)

Если запустить весь поток **WF\_LOAD\_DWH2BDA** то начнеться загрузка всех таблиц (работает около 5 часов)

Посмотреть загруженные партиции по дням можно в **hue консоли: (**u\_m0g4q/u\_m0g4q**)**

<http://bda11node04:8888/filebrowser/view=/storage/s_dwh/db/OPERATIONCOST_STRAN#/storage/s_dwh/db/OPERATIONCOST_STRAN>



Для замены формата загрузки данных на parquet

надо в таске, например **cmd\_run\_import\_MODULE\_LDIM** поменять скрипт

с такого:

$PMRootDir/Scripts/quotes\_insert.pl 'ssh etladmin@bda11node06 ^sqoop import --connect \^jdbc:oracle:thin:@(DESCRIPTION= (ADDRESS\_LIST=(ADDRESS=(PROTOCOL=TCP)(HOST=exa1-scan)(PORT=1521))(ADDRESS=(PROTOCOL=TCP)(HOST=exa2-scan)(PORT=1521))) (CONNECT\_DATA=(SERVICE\_NAME=DWSTPROD\_TAF)))\^ --username $DWH\_USER --password $DWH\_PASS --query \^select \* from MODULE\_LDIM where deleted\_flag=~N~ and \\\$CONDITIONS\^ -m 1 --as-avrodatafile --target-dir /storage/s\_dwh/db/MODULE\_LDIM/ --append^'

на такой:

$PMRootDir/Scripts/quotes\_insert.pl 'ssh etladmin@bda11node06 ^sqoop import --connect \^jdbc:oracle:thin:@(DESCRIPTION=(ADDRESS\_LIST=(ADDRESS=(PROTOCOL=TCP)(HOST=exa1-scan)(PORT=1521))(ADDRESS=(PROTOCOL=TCP)(HOST=exa2-scan)(PORT=1521)))(CONNECT\_DATA=(SERVICE\_NAME=DWSTPROD\_TAF)))\^ --username $DWH\_USER --password $DWH\_PASS --query \^select \* from MODULE\_LDIM where deleted\_flag=~N~ and \\\$CONDITIONS\^ -m 1 --compression-codec GZIP --as-parquetfile --hive-import --hive-database s\_dwh --hive-table MODULE\_LDIM --target-dir /storage/s\_dwh/db/MODULE\_LDIM/ --hive-overwrite^'

За формат данных отвечает параметр **--as-parquetfile**

Параметр **--hive-overwrite** указывает на то, что таблица перезаписывается, т.е. удаляется и сразу создается.

**Запрос скрипта создания таблицы:**

show create table s\_dwh.cardtransaction\_htran

**Создание ПАРТИЦИОНИРОВАННОЙ таблицы:**

CREATE TABLE `s\_dwh.NEW\_DEALSTATE\_SSTAT`(

`REGTYPE\_CCODE` string,

`CHANNEL\_SALE\_UK` string,

`CHANNEL\_MAIN\_ATTRACT\_UK` string,

`CHANNEL\_PRIOR\_ATTRACT\_UK` string,

`CLIENT\_PIN` string,

`CLIENT\_UK` string,

`ACCOUNT\_UK` string,

`DEAL\_ACTIVE\_FLAG` string,

`DEAL\_CNT` string,

`DEAL\_END\_DATE` string,

`DEAL\_PARENT\_REF` string,

`DEAL\_PARENT\_UK` string,

`DEAL\_REF` string,

`DEAL\_START\_DATE` string,

`DEAL\_UK` string,

`EQUITY\_RUR\_AMT` string,

`EQUITY\_RUR\_AMT\_30\_AVG` string,

`EQUITY\_RUR\_AMT\_AVG` string,

`EQUITY\_USD\_AMT` string,

`EQUITY\_USD\_AMT\_AVG` string,

`GENERAL\_DEBT\_RUR\_AMT` string,

`GENERAL\_DEBT\_RUR\_AMT\_30\_AVG` string,

`GENERAL\_DEBT\_RUR\_AMT\_AVG` string,

`GENERAL\_DEBT\_USD\_AMT` string,

`GENERAL\_DEBT\_USD\_AMT\_AVG` string,

`PRODUCT\_ACTIVE\_FLAG` string,

`PRODUCT\_CNT` string,

`PRODUCT\_PARENT\_UK` string,

`PRODUCT\_UK` string,

`PRODUCT\_UPSELL\_FLAG` string,

`PROFITCENTER\_UK` string,

`SALARY\_FLAG` string,

`TK` string,

`TURN\_30\_CNT` string,

`TURN\_CM\_CNT` string,

`VALUE\_DAY` string,

`WRITE\_OFF\_DEBT\_RUR\_AMT` string,

`WRITE\_OFF\_DEBT\_RUR\_AMT\_AVG` string,

`WRITE\_OFF\_DEBT\_USD\_AMT` string,

`WRITE\_OFF\_DEBT\_USD\_AMT\_AVG` string,

`DATAFLOW\_CCODE` string,

`DELETED\_FLAG` string,

`EMIX` string,

`JOB\_INSERT` string,

`JOB\_UPDATE` string,

`AS\_OF\_DAY` string)

PARTITIONED BY (

`date\_part` int)

ROW FORMAT SERDE

'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'

STORED AS INPUTFORMAT

'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'

LOCATION

'hdfs://bda11/storage/s\_dwh/db/NEW\_DEALSTATE\_SSTAT'

TBLPROPERTIES (

'avro.schema.url'='hdfs://bda11/storage/s\_dwh/db/NEW\_DEALSTATE\_SSTAT/date\_part=20180502/.metadata/schemas/1.avsc',

'kite.compression.type'='snappy',

'transient\_lastDdlTime'='1523957504')

**Добавление партиции:**

alter table BUREAUREQCLIENT\_LSTAT add partition (date\_part=20180502)

**Удаление партиции:**

alter table cardtransaction\_stran drop partition (date\_part='20160101')

**Удаление таблицы:**

DROP TABLE `s\_dwh.BUREAUREQCLIENT\_LSTAT`

**Запрос с партицией:**

select date\_part from HOLDAMOUNT\_HTRAN where (date\_part=20180623)

select distinct(date\_part) from l\_deriveddata.card\_transactions where (date\_part >=20181001 and date\_part <=20181031);

select distinct(value\_day) from s\_dwh.CARDTRANSACTION\_STRAN where (date\_part =20170307) ;

**Вспомогательные запросы для создания репликации:**

select \* from

DMPP.NEW\_DEALSTATE\_SSTAT

where value\_day=to\_date('20180502','yyyymmdd') and deleted\_flag='N';

select min(VALUE\_DAY) from *-- 06.12.2017*

DMPP.NEW\_DEALSTATE\_SSTAT

where deleted\_flag='N';

select max(VALUE\_DAY) from *-- 02.05.2018*

DMPP.NEW\_DEALSTATE\_SSTAT

where deleted\_flag='N';

select distinct(VALUE\_DAY) from

DMPP.NEW\_DEALSTATE\_SSTAT

where deleted\_flag='N'

order by VALUE\_DAY desc;

select REGTYPE\_CCODE, CHANNEL\_SALE\_UK, CHANNEL\_MAIN\_ATTRACT\_UK, CHANNEL\_PRIOR\_ATTRACT\_UK, CLIENT\_PIN, CLIENT\_UK, ACCOUNT\_UK, DEAL\_ACTIVE\_FLAG, DEAL\_CNT, to\_char(DEAL\_END\_DATE, ~dd.mm.yyyy~) as DEAL\_END\_DATE, DEAL\_PARENT\_REF, DEAL\_PARENT\_UK, DEAL\_REF, to\_char(DEAL\_START\_DATE, ~dd.mm.yyyy~) as DEAL\_START\_DATE, DEAL\_UK, EQUITY\_RUR\_AMT, EQUITY\_RUR\_AMT\_30\_AVG, EQUITY\_RUR\_AMT\_AVG, EQUITY\_USD\_AMT, EQUITY\_USD\_AMT\_AVG, GENERAL\_DEBT\_RUR\_AMT, GENERAL\_DEBT\_RUR\_AMT\_30\_AVG, GENERAL\_DEBT\_RUR\_AMT\_AVG, GENERAL\_DEBT\_USD\_AMT, GENERAL\_DEBT\_USD\_AMT\_AVG, PRODUCT\_ACTIVE\_FLAG, PRODUCT\_CNT, PRODUCT\_PARENT\_UK, PRODUCT\_UK, PRODUCT\_UPSELL\_FLAG, PROFITCENTER\_UK, SALARY\_FLAG, TK, TURN\_30\_CNT, TURN\_CM\_CNT, to\_char(VALUE\_DAY, ~dd.mm.yyyy~) as VALUE\_DAY, WRITE\_OFF\_DEBT\_RUR\_AMT, WRITE\_OFF\_DEBT\_RUR\_AMT\_AVG, WRITE\_OFF\_DEBT\_USD\_AMT, WRITE\_OFF\_DEBT\_USD\_AMT\_AVG, DATAFLOW\_CCODE, DELETED\_FLAG, EMIX, JOB\_INSERT, JOB\_UPDATE, to\_char(AS\_OF\_DAY, ~dd.mm.yyyy~) as AS\_OF\_DAY from DMPP.NEW\_DEALSTATE\_SSTAT where value\_day=to\_date(~$$DATE1~,~yyyymmdd~) and deleted\_flag=~N~

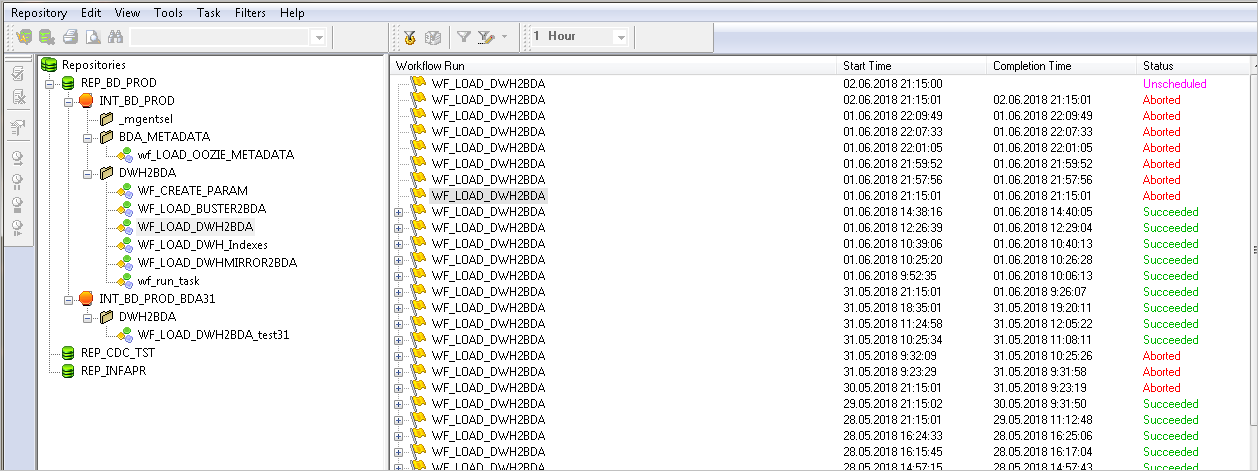
Александр Маисеев

Оксана Ханкина

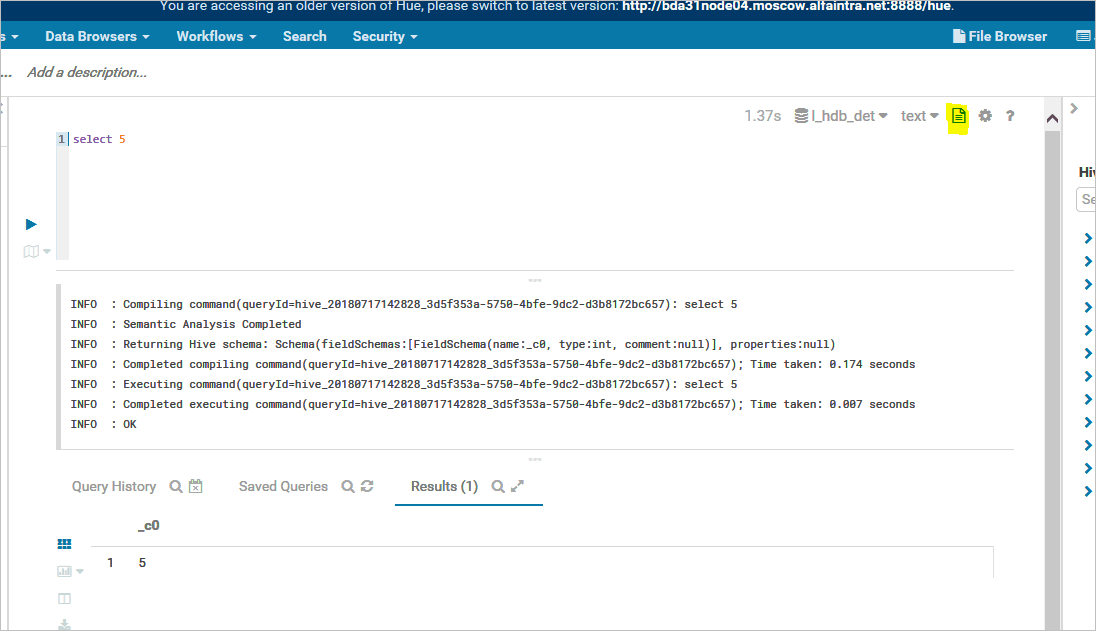
Игорь Радивоз

Сергей Богута

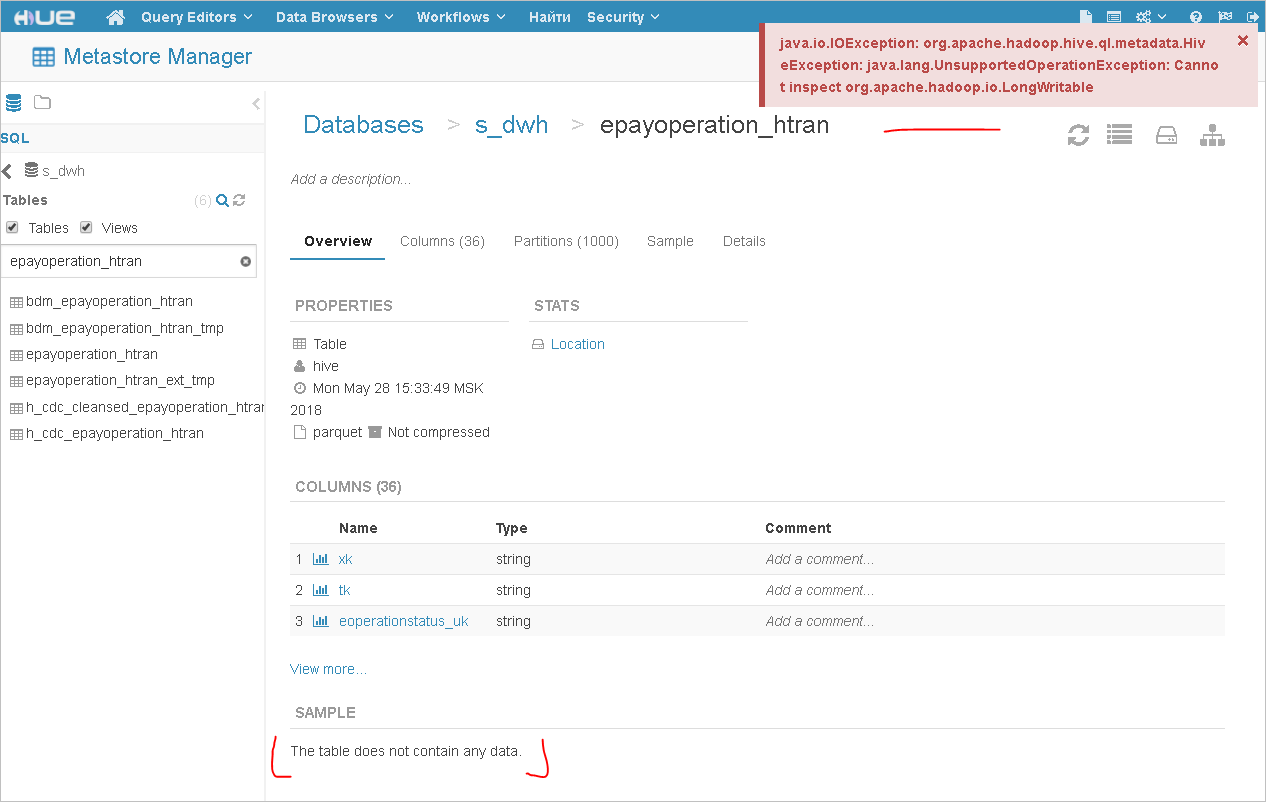
Если на сервере заканчивается место, то поток начинает падать вот так:



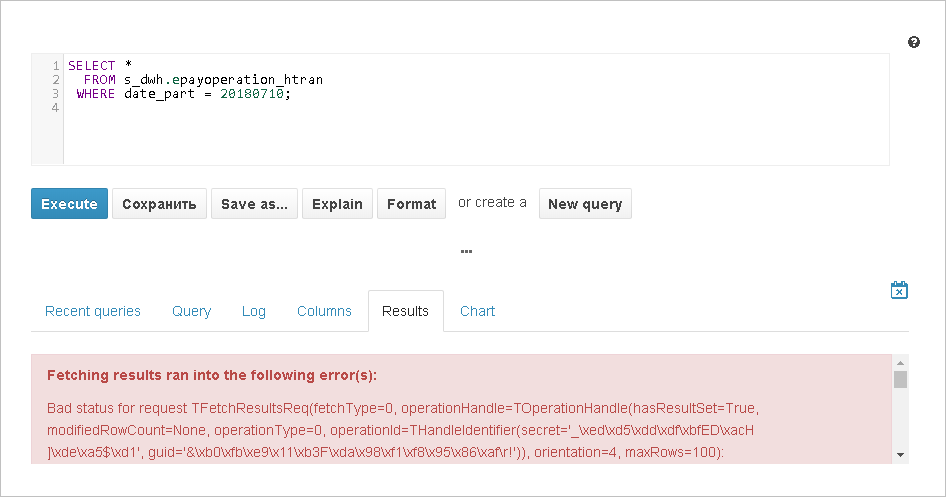
Кнопка "Show logs" над текст-боксом выводит или скрывает лог с экрана:



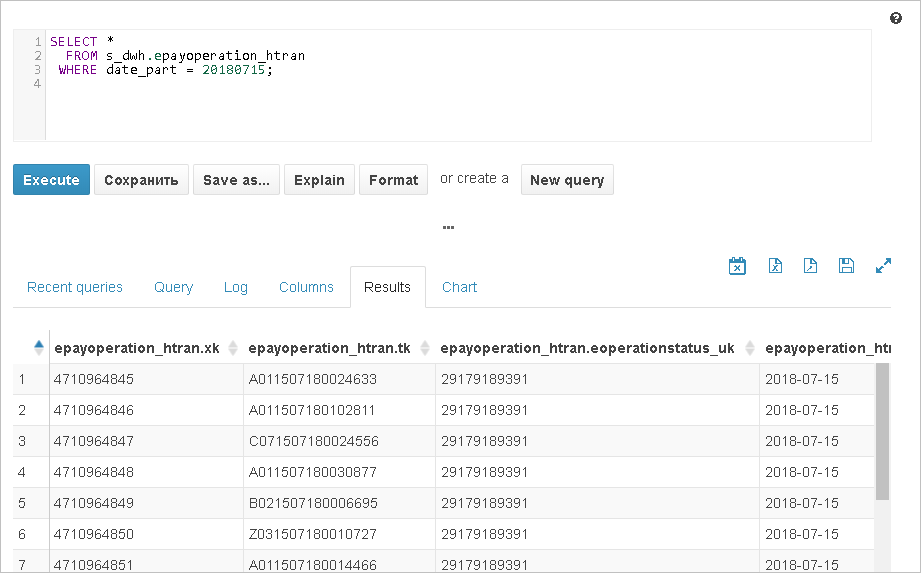
Если все поля таблицы типа стринг а в селекте загрузки стоит \* то даты заливаются в типе bigint и при заходе в таблицу возникает ошибка:



При выполнении запроса к партиции где есть данные типа bigint ошибка такая:

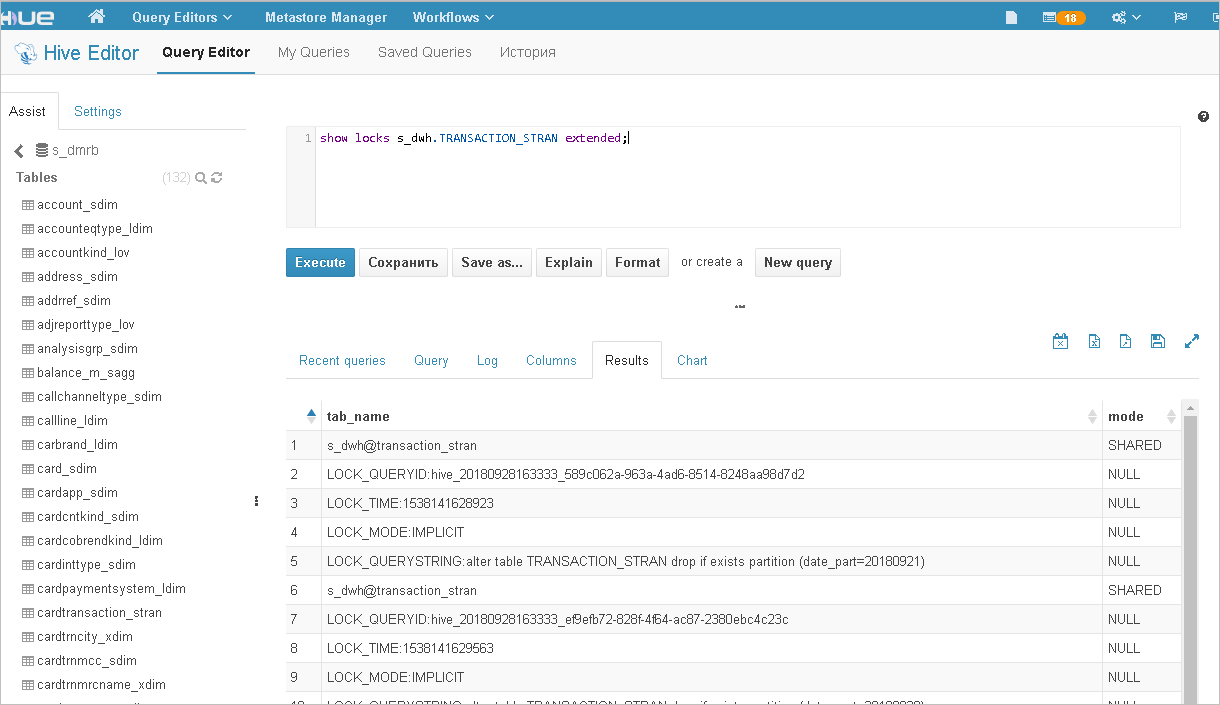


Если переделать \* в список полей, с преобразованием даты в стринг и выполнить загрузку, то при обращении к партиции запрос выполняется без ошибок:



Посмотреть блокировки таблицы можно так:

show locks s\_dwh.TRANSACTION\_STRAN extended;



Убить все блокировки можно так:

unlock table s\_dwh.TRANSACTION\_STRAN;

**Причиной зависания загрузок в информатике и oozie может быть запрос:**

select \* from l\_hdb\_det.wsrm\_log

