

Joseph Norell Guttman

Platform by Per Scholas

Data Engineer Cohort 02: CASE STUDY

Instructor: Muhammad Haseeb



2.2 Functional Requirements – ETL of Data

2.2.1 Data Extraction and Transportation Module

2.2.1 Data Extraction and Transportation with Sqoop

Credit Card System Req-2.2.1	Data Extraction and Transportation with Sqoop
Functional Requirements	<p>Utilize Sqoop to extract the following data according to the specifications found in the mapping document:</p> <ol style="list-style-type: none">1. Branch data into CDW_SAPP_BRANCH2. Credit Card Data into CDW_SAPP_CREDITCARD3. Time data into CDW_SAPP_TIME4. Customer Data into CDW_SAPP_CUSTOMER <p>Notes:</p> <ul style="list-style-type: none">• Data Engineers will be required to transform the data based on requirements found in the Mapping Document prior to loading the data into Hadoop.• TIMEID is a field that the Data Engineers should create based on the DAY, MONTH, and TIME fields located in the CUSTOMER table. Format should be YYYYMMDD. For instance, January 4th, 2017 would become 20170104• Data Engineers should extract the above data to the /Credit_Card_System/ folder in the Hadoop Filesystem

Target Table/File name	Target Field names	Target Data Type	Description	Mapping Logic
CDW_SAPP_D_TIME	TIMED	VARCHAR2(6)	The unique key defines a day	Convert DAY, MONTH
CDW_SAPP_D_TIME	DAY	NUMBER(2)	Day of a month	Move from DAY column

1. Branch data into CDW_SAPP_BRANCH

#	BRANCH_CODE	BRANCH_NAME	BRANCH_STREET	BRANCH_CITY	BRANCH_STATE	BRANCH_ZIP	BRANCH_PHONE
1	1	Example Bank	Bridle Court	Lakeville	MN	55044	1234565276

```
[maria_dev@sandbox ~]$ sqoop import --connect jdbc:mysql://localhost/CDW_SAPP --
driver com.mysql.jdbc.Driver --username root --password password --query 'SELECT
branch_code, branch_name, branch_street, branch_city, branch_state, branch_zip,
branch_phone FROM CDW_SAPP_BRANCH WHERE $CONDITIONS' --split-by BRANCH_CODE --t
arget-dir /user/maria_dev/CDW_SAPP_BRANCH | tee Sqoop.Output.CDW_SAPP_Branch.txt
```

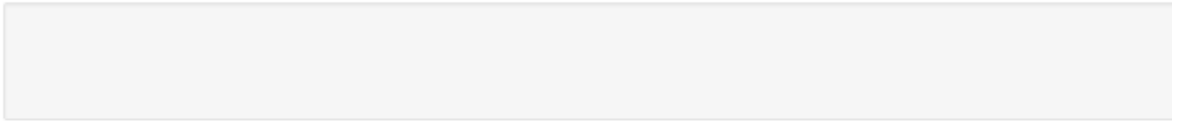


/ >

user >

maria_dev >

CDW_SAPP_BRANCH



Name >

Size >



_SUCCESS

0.1 kB

part-m-00000

2.6 kB

part-m-00001

1.9 kB

part-m-00002

1.4 kB

part-m-00003

0.9 kB

File Preview

/user/maria_dev/CDW_SAPP_BRANCH/part-m-00000

```
1,Example Bank,Bridle Court,Lakeville,MN,55044,1234565276
2,Example Bank,Washington Street,Huntley,IL,60142,1234618993
3,Example Bank,Warren Street,SouthRichmondHill,NY,11419,1234985926
4,Example Bank,Cleveland Street,Middleburg,FL,32068,1234663064
5,Example Bank,14th Street,KingOfPrussia,PA,19406,1234849701
7,Example Bank,Jefferson Street,Paterson,NJ,7501,1234144890
8,Example Bank,B Street,Pittsford,NY,14534,1234678272
9,Example Bank,Jefferson Court,Wethersfield,CT,6100,1234675210
```

2. Credit Card Data into CDW_SAPP_CREDITCARD

#	TRANSACTION_ID	DAY	MONTH	YEAR	CREDIT_CARD_NO	CUST_SSN	BRANCH_CODE	TRANSACTION_TYPE	TRANSACTION_VALUE
1	1	14	2	2018	4210653349028689	123459988	114	Education	78.900

All the transactions appear to be in the same year of 2018 so I split it by month:

```
sqoop import --connect jdbc:mysql://localhost/CDW_SAPP --driver  
com.mysql.jdbc.Driver --username root --password password --query 'SELECT  
transaction_id, day, month, year, credit_card_no, cust_ssn, branch_code,  
transaction_type, transaction_value FROM CDW_SAPP_CREDITCARD WHERE  
$CONDITIONS' --split-by month --target-dir  
/user/maria_dev/CDW_SAPP_CREDITCARD |tee  
Sqoop.Output.CDW_SAPP_CREDITCARD.txt
```




/ > user > maria_dev > CDW_SAPP_CREDITCARD			Total: 5 files or folders
Name >	Size >	Last Modified >	
_SUCCESS	0.1 kB	2018-06-21 16:32	
part-m-00000	691.1 kB	2018-06-21 16:32	
part-m-00001	690.4 kB	2018-06-21 16:32	
part-m-00002	694.6 kB	2018-06-21 16:32	
part-m-00003	712.0 kB	2018-06-21 16:32	

File Preview	
/user/maria_dev/CDW_SAPP_CREDITCARD/part-m-00000	
0	1,14,2,2018,4210653349028689,123459988,114,Education,78.900
1	2,20,3,2018,4210653349028689,123459988,35,Entertainment,14.240
2	9,18,3,2018,4210653349028689,123459988,166,Entertainment,93.260
	23,17,3,2018,4210653349028689,123459988,44,Bills,7.970
	25,3,3,2018,4210653349028689,123459988,119,Entertainment,35.020
	40,16,1,2018,4210653349028689,123459988,69,Entertainment,96.290
	41,7,3,2018,4210653349028689,123459988,26,Education,25.010
	46,21,1,2018,4210653349028689,123459988,104,Bills,4.950
	50,2,1,2018,4210653349028689,123459988,76,Entertainment,53.490
	51,4,3,2018,4210653349028689,123459988,93,Gas,73.450
	54,3,2,2018,4210653349028689,123459988,59,Healthcare,31.870

2. Credit card data into CDW_SAPP_CREDITCARD
3. Time data into CDW_SAPP_TIME

```
sqoop import --connect jdbc:mysql://localhost/CDW_SAPP --driver
com.mysql.jdbc.Driver --username root --password password --query "SELECT
DATE_FORMAT(STR_TO_DATE(CONCAT(YEAR, LPAD(MONTH , 2, '0' ) , LPAD(DAY
```

```
, 2, '0' ) ), '%Y%m%d' ) , '%Y%m%d') AS TIMEID, DAY, MONTH,  
QUARTER(STR_TO_DATE(CONCAT(YEAR, LPAD(MONTH , 2, '0' ) , LPAD(DAY , 2,  
'0' ) ), '%Y%m%d' )) AS QUARTR, YEAR FROM CDW_SAPP_CREDITCARD WHERE \  
$CONDITIONS" -m 1 --target-dir /user/maria_dev/CDW_SAPP_TIME
```

Name >	Size >	Last Modified >
		
 _SUCCESS	0.1 kB	2018-06-21 16:55
 part-m-00000	954.5 kB	2018-06-21 16:55

user

maria_dev

CDW_SAPP_TIME

File Preview

/user/maria_dev/CDW_SAPP_TIME/part-m-00000

20180214,14,2,1,2018

20180320,20,3,1,2018

20180708,8,7,3,2018

20180419,19,4,2,2018

20181010,10,10,4,2018

20180528,28,5,2,2018

20180519,19,5,2,2018

20180808,8,8,3,2018

20180318,18,3,1,2018

20180903,3,9,3,2018

20180821,21,8,3,2018

20181224,24,12,4,2018

20180403,3,4,2,2018

20180415,15,4,2,2018

20180517,17,5,2,2018

20180706,6,7,3,2018

20180928,28,9,3,2018

20180704,4,7,3,2018

20180424,24,4,2,2018

4. Customer Data into CDW_SAPP_CUSTOMER

#	FIRST_NAME	MIDDLE_NAME	LAST_NAME	SSN	CREDIT_CARD_NO	APT_NO	STREET_NAME	CUST_CITY	CUST_STATE	CUST_COUNTRY	CUST_ZIP	CUST_PHONE	CUST_EMAIL
1	Alec	Wm	Hooper	123456100	4210653310061055	656	Main Street North	Natchez	MS	United States	39120	1237818	AHooper@example.com
2	Etta	Brendan	Holman	123453023	4210653310102868	829	Redwood Drive	Wethersfield	CT	United States	06109	1238933	EHolman@example.com
3	Wilber	Ezequiel	Dunham	123454487	4210653310116272	683	12th Street East	Huntley	IL	United States	60142	1243018	WDunham@example.com
4	Eugenio	Trina	Hardy	123459758	4210653310195948	253	Country Club Road	NewBerlin	WI	United States	53151	1243215	EHardy@example.com
5	Wilfred	May	Ayers	123454431	4210653310356919	301	Madison Street	ElPaso	TX	United States	79930	1242074	WAyers@example.com
6	Beau	Ambrose	Woodard	123454202	4210653310395982	3	Colonial Drive	NorthOlmsted	OH	United States	44070	1242570	BWoodard@example.com

```
2 SELECT distinct CUST_STATE FROM CDW_SAPP_CUSTOMER LIMIT 50;
```







#	CUST_STATE
9	MI
10	SC
11	FL
12	PA
13	MD
14	NJ
15	IA
16	GA
17	MT
18	MN
19	NY
20	NC
21	MA
22	KY
23	WA
24	IN
25	AR
26	AL

```
sqoop import -Dorg.apache.sqoop.splitter.allow_text_splitter=true --connect
jdbc:mysql://localhost/CDW_SAPP --driver com.mysql.jdbc.Driver --username
root --password password --query 'SELECT first_name, middle_name,
last_name, ssn, credit_card_no, apt_no, street_name, cust_city, cust_state,
cust_country, cust_zip, cust_phone, cust_email FROM CDW_SAPP_CUSTOMER
WHERE $CONDITIONS' --split-by CUST_STATE --target-dir
/user/maria_dev/CDW_SAPP_CUST
```

code for allowing splitter by CUST_STATE :

<http://discuss.itversity.com/t/running-sqoop-query-with-text-field-filter/6249/3>

   / > user > maria_dev > CDW_SAPP_CUST

Name >	Size >
	
 _SUCCESS	0.1 kB
 part-m-00000	8.5 kB
 part-m-00001	30.3 kB
 part-m-00002	60.7 kB
 part-m-00003	18.1 kB

2.2.2 Data Loading with Hive

Credit Card System Req-2.2.2	Data Loading with Hive
Functional Requirements	Utilize Hive to create tables in the Hadoop Filesystem and then load the data extracted via Sqoop into those tables. Data Engineers will be map to transform the data based on requirements found in the Mapping Document.

CREATE EXTERNAL TABLE CDW_SAPP_BRANCH

(

BRANCH_CODE int,

BRANCH_NAME varchar(50),

BRANCH_STREET varchar(150),

BRANCH_CITY varchar(50),

BRANCH_STATE varchar(50),

BRANCH_ZIP int,

BRANCH_PHONE varchar(20)

)

**ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location
'/user/maria_dev/CDW_SAPP_BRANCH/';**

Query Editor

Worksheet *

```
1 CREATE EXTERNAL TABLE CDW_SAPP_BRANCH
2 (
3   BRANCH_CODE int,
4   BRANCH_NAME varchar(50),
5   BRANCH_STREET varchar(150),
6   BRANCH_CITY varchar(50),
7   BRANCH_STATE varchar(50),
8   BRANCH_ZIP int,
9   BRANCH_PHONE varchar(20)
10 )
11 ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/aria_dev/CDW_SAPP_BRANCH/';
12
13
```

Execute

Explain

Save as...

New Worksheet

Query Process Results (Status: SUCCEEDED)

Save results... ▾

Query Editor

Worksheet * x

cdw_sapp_branch sample x

1

SELECT * FROM cdw_sapp_branch LIMIT 100;

SQL

TEZ

Execute

Explain

Save as...

New Worksheet

Query Process Results (Status: SUCCEEDED)Save results... ▾

LogsResults

Filter columns...previousnext

cdw_sapp_branch.branch_code	cdw_sapp_branch.branch_name	cdw_sapp_branch.branch_street	cdw_sapp_br
1	Example Bank	Bridle Court	Lakeville
2	Example Bank	Washington Street	Huntley
3	Example Bank	Warren Street	SouthRichmor
4	Example Bank	Cleveland Street	Middleburg
5	Example Bank	14th Street	KingOfPrussia
7	Example Bank	Jefferson Street	Paterson
8	Example Bank	B Street	Rittsford
9	Example Bank	Jefferson Court	Wethersfield

```
CREATE EXTERNAL TABLE CDW_SAPP_CREDITCARD
(
  TRANSACTION_ID int,
  DAY int,
  MONTH int ,
  YEAR int ,
  CREDIT_CARD_NO varchar(16) ,
  CUST_SSN int ,
  BRANCH_CODE int ,
  TRANSACTION_TYPE varchar(30) ,
  TRANSACTION_VALUE decimal(20,7)
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location
'/user/maria_dev/CDW_SAPP_CREDITCARD/';
```

Query Editor

CDW_SAPP_BRANCH-hive1 xcdw_sapp_branch sample xCDW_SAPP_BRANCH-hive1 x

cdw_sapp_branch sample xWorksheet x

```
1 CREATE EXTERNAL TABLE CDW_SAPP_CREDITCARD
2 (
3   TRANSACTION_ID int,
4   DAY int,
5   MONTH int,
6   YEAR int,
7   CREDIT_CARD_NO varchar(16),
8   CUST_SSN int,
9   BRANCH_CODE int,
10  TRANSACTION_TYPE varchar(30),
11  TRANSACTION_VALUE decimal(20,7)
12 )
13 ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/maria_dev/CDW_SAPP_CREDITCARD/';
```

Execute

Explain

Save as...

New Worksheet

SQL

TEZ

Query Process Results (Status: SUCCEEDED)

Save results... ▾

Logs

Results

Query Process Results (Status: SUCCEEDED)

Save results... ▾

Logs

Results

Filter columns...

previous

next

cdw_sapp_creditcard.transaction_id	cdw_sapp_creditcard.day	cdw_sapp_creditcard.month	cdw_sapp_creditcar
1	14	2	2018
2	20	3	2018


```
CREATE EXTERNAL TABLE CDW_SAPP_CUSTOMER
(
  FIRST_NAME varchar(70),
  MIDDLE_NAME varchar(70),
  LAST_NAME varchar(70),
  SSN int,
  CREDIT_CARD_NO varchar(16),
  APT_NO varchar(20),
  STREET_NAME varchar(70),
  CUST_CITY varchar(70),
  CUST_STATE varchar(70),
  CUST_COUNTRY varchar(70),
  CUST_ZIP varchar(10),
  CUST_PHONE int,
  CUST_EMAIL varchar(150)
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/maria_dev/CDW_SAPP_CUST/';
```

Query Editor

Worksheet x

cdw_sapp_customer sample x

cdw_sapp_customer sample x

```
1 CREATE EXTERNAL TABLE CDW_SAPP_CUSTOMER
2 (
3   FIRST_NAME varchar(70),
4   MIDDLE_NAME varchar(70),
5   LAST_NAME varchar(70),
6   SSN int,
7   CREDIT_CARD_NO varchar(16),
8   APT_NO varchar(20),
9   STREET_NAME varchar(70),
10  CUST_CITY varchar(70),
11  CUST_STATE varchar(70),
12  CUST_COUNTRY varchar(70),
13  CUST_ZIP varchar(10),
14  CUST_PHONE int,
15  CUST_EMAIL varchar(150)
16 )
17 ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/maria_dev/CDW_SAPP_CUSTOMER/';
```

Execute

Explain

Save as...

New Worksheet

Query Process Results (Status: SUCCEEDED)

Save results... ▾

Logs

Results

Query Process Results (Status: SUCCEEDED)

Save results... ▾

Logs

Results

Filter columns...

previous

next

cdw_sapp_customer.first_name	cdw_sapp_customer.middle_name	cdw_sapp_customer.last_name	cdw_sapp_c
Etta	Brendan	Holman	123453023
Wendy	Ora	Hurley	123453875
Patty	Angelita	Thomas	123455343
Elden	Carolina	Murphy	123453988

```

CREATE TABLE CDW_SAPP_TIME (
    TIMEID VARCHAR(40),
    DAY int,
    MONTH int ,
    QUARTR int,
    YEAR int
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/maria_dev/CDW_SAPP_TIME/';

```

Query Process Results (Status: SUCCEEDED)					Save results... ▾
<div> <div>Logs</div> <div>Results</div> </div>					
<div>Filter columns...</div>					<div>previous</div> <div>next</div>
cdw_sapp_time.timeid	cdw_sapp_time.day	cdw_sapp_time.month	cdw_sapp_time.quartr	cdw_sapp_time.year	
20180214	14	2	1	2018	
20180320	20	3	1	2018	
20180708	8	7	3	2018	

Hive

Query

Saved Queries

History

Database Explorer



default



Search tables...

Databases

default

cdw_sapp_branch
cdw_sapp_creditcard
cdw_sapp_customer
cdw_sapp_time
cdw_sapp_time



2.2.3 Process Automation Module

2.2.3 Automating the Process with Oozie

Credit Card System Req-2.2.3	Automating the Process with Oozie
Functional Requirements	<ol style="list-style-type: none">1) Create an Oozie Workflow that will automate the processes of steps 2.2.1 and 2.2.2.<ul style="list-style-type: none">• Each of the files in step 2.2.1 should be deleted before the workflow is executed in order to prevent storage of redundant data• The tables created in step 2.2.2 should be dropped before executing the hive workflow in order to prevent redundancy.

```
drop table cdw_sapp_branch;  
drop table cdw_sapp_branch;  
drop table cdw_sapp_creditcard;  
drop table cdw_sapp_customer;  
drop table cdw_sapp_time;
```

```
sqoop import --connect jdbc:mysql://localhost/CDW_SAPP --driver  
com.mysql.jdbc.Driver --username root --password password --query 'SELECT  
branch_code, branch_name, branch_street, branch_city, branch_state, branch_zip,  
branch_phone FROM CDW_SAPP_BRANCH WHERE $CONDITIONS' --split-by  
BRANCH_CODE -target-dir /user/maria_dev/CDW_SAPP_BRANCH |tee  
Sqoop.Output.CDW_SAPP_Branch.txt
```

```
sqoop import --connect jdbc:mysql://localhost/CDW_SAPP --driver  
com.mysql.jdbc.Driver --username root --password password --query 'SELECT  
transaction_id, day, month, year, credit_card_no, cust_ssn, branch_code,  
transaction_type, transaction_value FROM CDW_SAPP_CREDITCARD WHERE  
$CONDITIONS' --split-by month --target-dir  
/user/maria_dev/CDW_SAPP_CREDITCARD |tee  
Sqoop.Output.CDW_SAPP_CREDITCARD.txt
```

```
sqoop import --connect jdbc:mysql://localhost/CDW_SAPP --driver
com.mysql.jdbc.Driver --username root --password password --query "SELECT
DATE_FORMAT(STR_TO_DATE(CONCAT(YEAR, LPAD(MONTH , 2, '0' ) ,
LPAD(DAY , 2, '0' ) ), '%Y%m%d' ) , '%Y%m%d') AS TIMEID, DAY, MONTH,
QUARTER(STR_TO_DATE(CONCAT(YEAR, LPAD(MONTH , 2, '0' ) , LPAD(DAY , 2,
'0' ) ), '%Y%m%d' )) AS QUARTR, YEAR FROM CDW_SAPP_CREDITCARD WHERE \
$CONDITIONS" -m 1 --target-dir /user/maria_dev/CDW_SAPP_TIME
```

```
sqoop import -Dorg.apache.sqoop.splitter.allow_text_splitter=true --connect
jdbc:mysql://localhost/CDW_SAPP --driver com.mysql.jdbc.Driver --username root
--password password --query 'SELECT first_name, middle_name, last_name, ssn,
credit_card_no, apt_no, street_name, cust_city, cust_state, cust_country, cust_zip,
cust_phone, cust_email FROM CDW_SAPP_CUSTOMER WHERE $CONDITIONS'
--split-by CUST_STATE --target-dir /user/maria_dev/CDW_SAPP_CUST
```

```
CREATE EXTERNAL TABLE CDW_SAPP_BRANCH
(
  BRANCH_CODE int,
  BRANCH_NAME varchar(50),
  BRANCH_STREET varchar(150),
  BRANCH_CITY varchar(50),
  BRANCH_STATE varchar(50),
  BRANCH_ZIP int,
  BRANCH_PHONE varchar(20)
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location
'/user/maria_dev/CDW_SAPP_BRANCH/';
```

```
CREATE EXTERNAL TABLE CDW_SAPP_CUSTOMER
(
  FIRST_NAME varchar(70),
  MIDDLE_NAME varchar(70),
  LAST_NAME varchar(70),
  SSN int,
  CREDIT_CARD_NO varchar(16),
  APT_NO varchar(20),
  STREET_NAME varchar(70),
  CUST_CITY varchar(70),
  CUST_STATE varchar(70),
  CUST_COUNTRY varchar(70),
  CUST_ZIP varchar(10),
```

```
CUST_PHONE int,  
CUST_EMAIL varchar(150)  
)  
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location  
'/user/maria_dev/CDW_SAPP_CUST/';
```

```
CREATE TABLE CDW_SAPP_TIME (  
  TIMEID VARCHAR(40),  
  DAY int,  
  MONTH int ,  
  QUARTR int,  
  YEAR int  
)  
  ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location  
'/user/maria_dev/CDW_SAPP_TIME/';
```

```
CREATE EXTERNAL TABLE CDW_SAPP_CREDITCARD  
(  
  TRANSACTION_ID int,  
  DAY int,  
  MONTH int ,  
  YEAR int ,  
  CREDIT_CARD_NO varchar(16) ,  
  CUST_SSN int ,  
  BRANCH_CODE int ,  
  TRANSACTION_TYPE varchar(30) ,  
  TRANSACTION_VALUE decimal(20,7)  
)  
  ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location  
'/user/maria_dev/CDW_SAPP_CREDITCARD/';
```

Still trying to work on the XML workflow....

Query Editor

Worksheet ✕ Worksheet (1) ✕ cdw_sapp_creditcard sample ✕

```
1 SELECT cdw_sapp_branch.branch_zip,
2        sum(cdw_sapp_creditcard.transaction_value)
3 FROM cdw_sapp_branch JOIN cdw_sapp_creditcard
4 ON (cdw_sapp_branch.branch_code=cdw_sapp_creditcard.branch_code)
5 GROUP BY cdw_sapp_branch.branch_zip
6 ORDER BY sum(cdw_sapp_creditcard.transaction_value)
7 limit 20;
```

Execute Explain Save as...

New Worksheet

Query Process Results (Status: RUNNING)

Could not get this to run so could not do the visualization....