

Dataprocessing & ETL Process - Tools – Tips

Uwe Geercken

uwe.geercken@swissport.com



Swissport Services

- Aircraft maintenance
- Aircraft servicing and cleaning
- Airport aviation security
- Cargo and mail handling (on-/off-airport)
- Catering services
- e-Services
- Executive aviation handling
- Flight operations and crew administration
- Fueling
- Load control
- Operation of airport lounges
- Outsourcing and global packages
- Passenger and baggage handling
- Ramp services
- Station supervision and administration
- Surface transport of passengers and crews
- Unit Load Device control and management

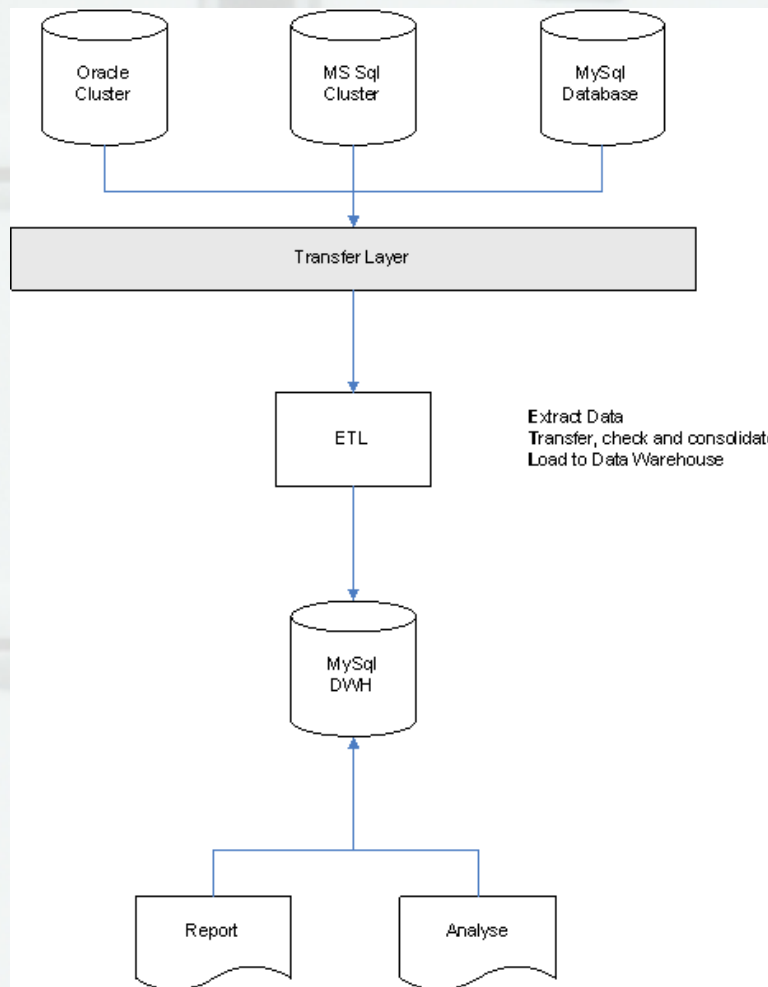
Swissport Facts

- Countries: 41
- Employees: over 33 000
- Flights handled: over 2.5 million
- Cargo handled (tonnes): over 3.5 million
- Airports served: 179
- Customer airlines: over 650
- Passengers handled: over 70 million

3 Questions

- Who knows Linux well?
- Who would say that she/he knows what ETL and Datawarehousing means
- Who has already worked with MySQL and Infobright?

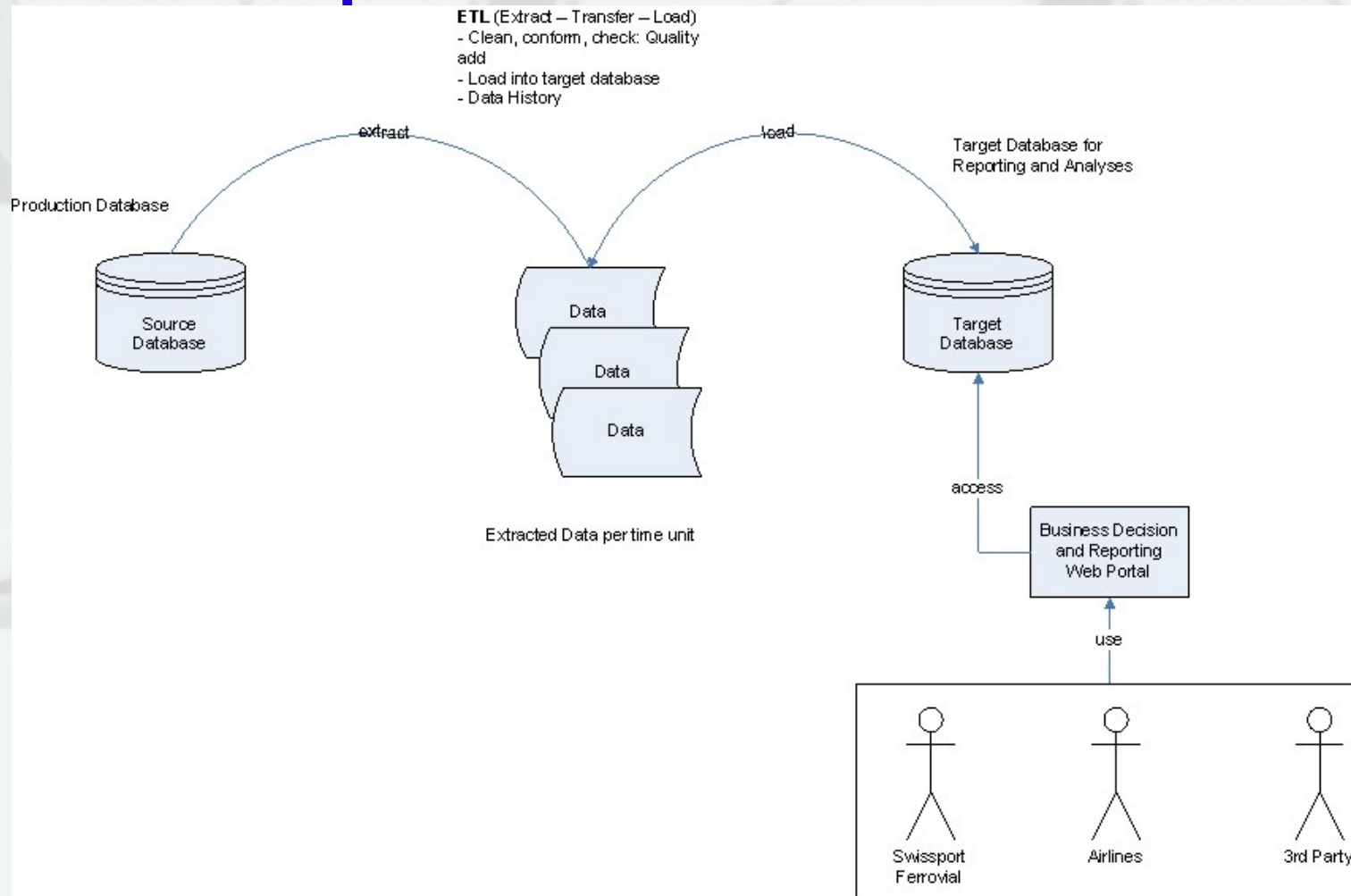
Data Warehouse Overview



DWH System+Components

- RedHat Enterprise Linux 64bit (Virtual, VmWare), Hosted by Provider
- MySQL 5.1 (3x)
- Infobright Enterprise (2x)
- Pentaho BI Suite

ETL process Overview



ETL process pre-requisites

- Modular, re-usable
- Flexible
- Integrate many systems
- Easy
- Fast
- Linux !

ETL Process

- 6 step approach:

- 1) Copy data
- 2) Pre-process
- 3) Transform
- 4) Post-process
- 5) Load into DWH
- 6) cleanup

General Tools

- Awk, Sed, Grep etc :
- Pentaho PDI
- Java
- Groovy, Beanshell, Shell scripts
- Apache Velocity
- MySQL, Infobright

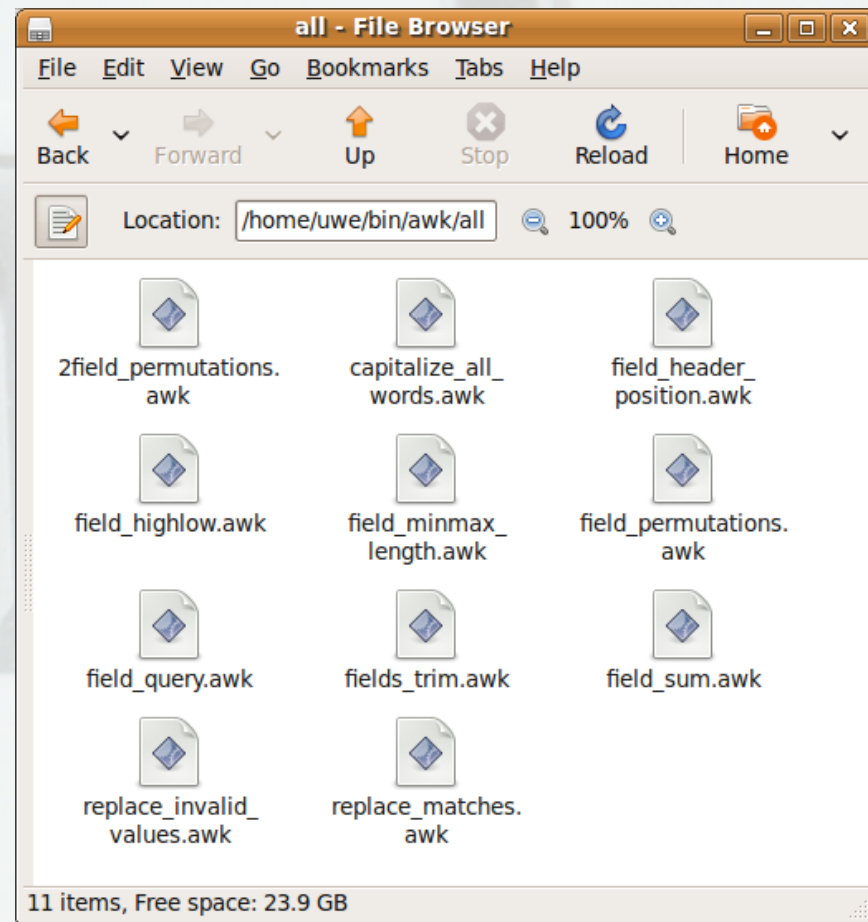
Additional Tools

- Data Generator (Java) *
- JaRE – Java Rule Engine *
- Datawriter (Java)
- Awk Scripts *

* download:
<http://www.datamelt.com>

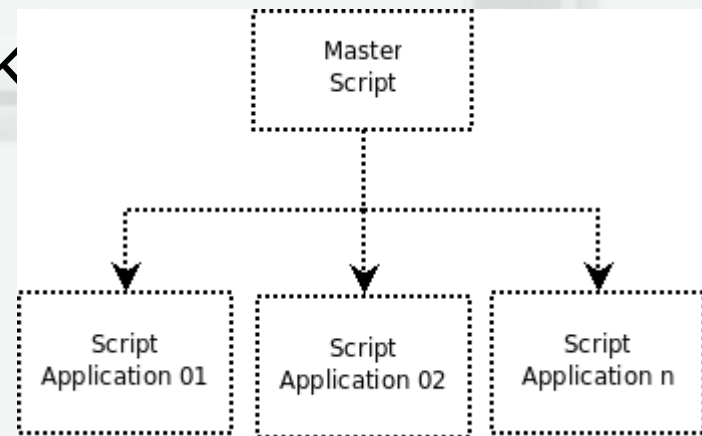
Awk Scripts

- Field permutations
- Field high/low
- Field length
- Field query
- Replace values
- Replace + match



Practical Part

- 6 step approach based on example data
 - 1) Copy Source file (CSV)
 - 2) Pre-Process with JaRE and Awk
 - 3) Transform with Pentaho PDI (Spoon)
 - 4) Post-Process with Awk
 - 5) Load into Infobright
 - 6) Clean up



Practical Part – Step 1

- Step 1: Copy Source file

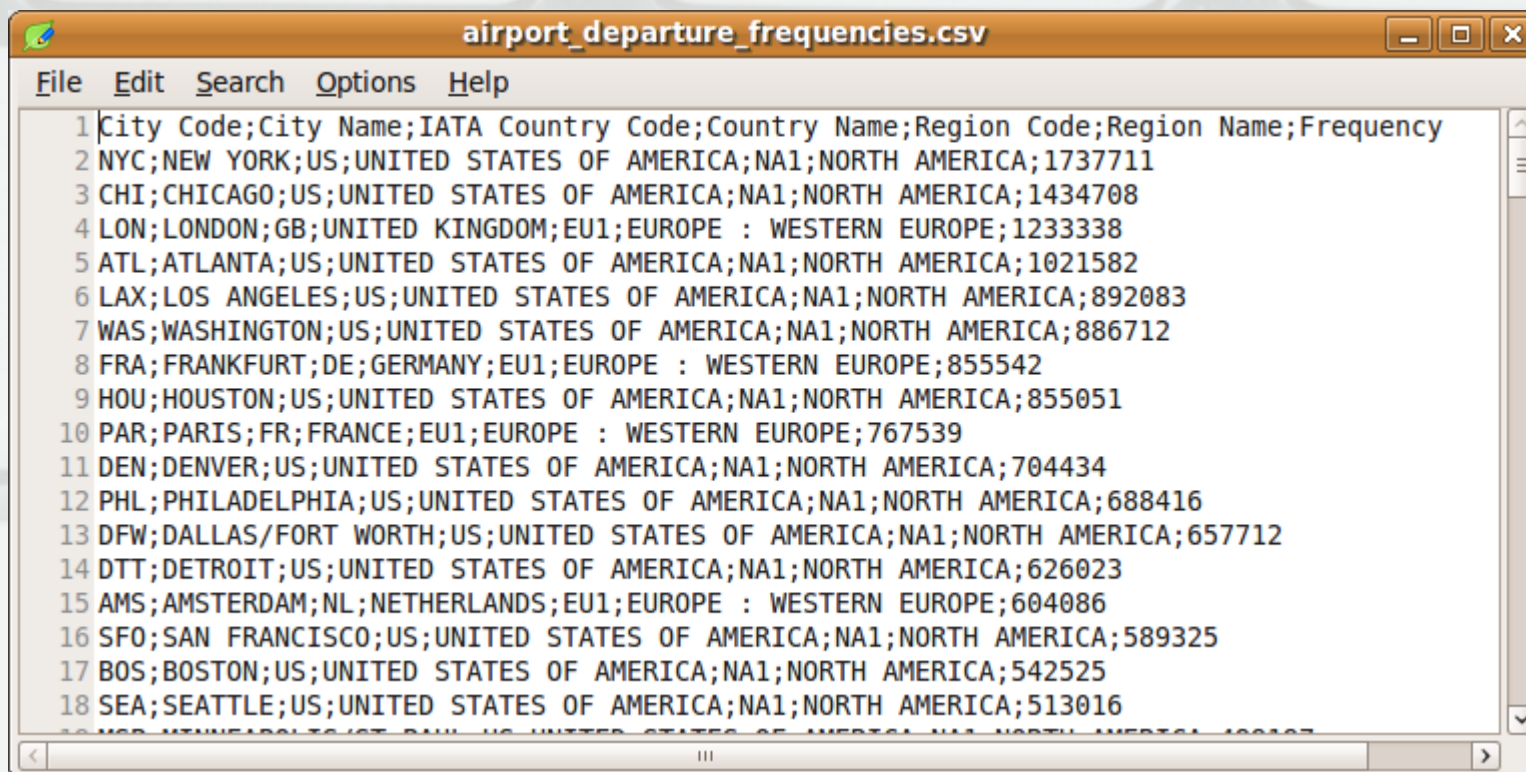
```
##### step 1
# copy files to a given folder
# step_name=step_01

.....

# action is to copy files from a certain location to the
# target folder for further processing
cp ${folder_copy_from}/${filename_01} ${folder_copy_to}
```

Practical Part – Step 1.1

Original source file



```
airport_departure_frequencies.csv
File Edit Search Options Help
1 City Code;City Name;IATA Country Code;Country Name;Region Code;Region Name;Frequency
2 NYC;NEW YORK;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;1737711
3 CHI;CHICAGO;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;1434708
4 LON;LONDON;GB;UNITED KINGDOM;EU1;EUROPE : WESTERN EUROPE;1233338
5 ATL;ATLANTA;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;1021582
6 LAX;LOS ANGELES;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;892083
7 WAS;WASHINGTON;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;886712
8 FRA;FRANKFURT;DE;GERMANY;EU1;EUROPE : WESTERN EUROPE;855542
9 HOU;HOUSTON;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;855051
10 PAR;PARIS;FR;FRANCE;EU1;EUROPE : WESTERN EUROPE;767539
11 DEN;DENVER;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;704434
12 PHL;PHILADELPHIA;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;688416
13 DFW;DALLAS/FORT WORTH;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;657712
14 DTT;DETROIT;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;626023
15 AMS;AMSTERDAM;NL;NETHERLANDS;EU1;EUROPE : WESTERN EUROPE;604086
16 SFO;SAN FRANCISCO;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;589325
17 BOS;BOSTON;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;542525
18 SEA;SEATTLE;US;UNITED STATES OF AMERICA;NA1;NORTH AMERICA;513016
```

Practical Part – Step 2

- Step 2: Pre-Process with JaRE and Awk

```
##### step 2
# pre-process files
...
# we are pre-processing files: replacing invalid values, checking data,
  matching data
${application_process_folder}/${step_name}/preprocess_01.awk $
  {application_output_folder}/${previous_step_name}/${filename_01} > $
  {application_output_folder}/${step_name}/${filename_01}
.....

# at this point we also run the business rule engine JaRE
# it will check the data of the file
${application_process_folder}/${
  {step_name}/jare/run_airport_departures_check.sh
```

Practical Part – Step 2.1

Field 6 and 7 manipulated with Awk

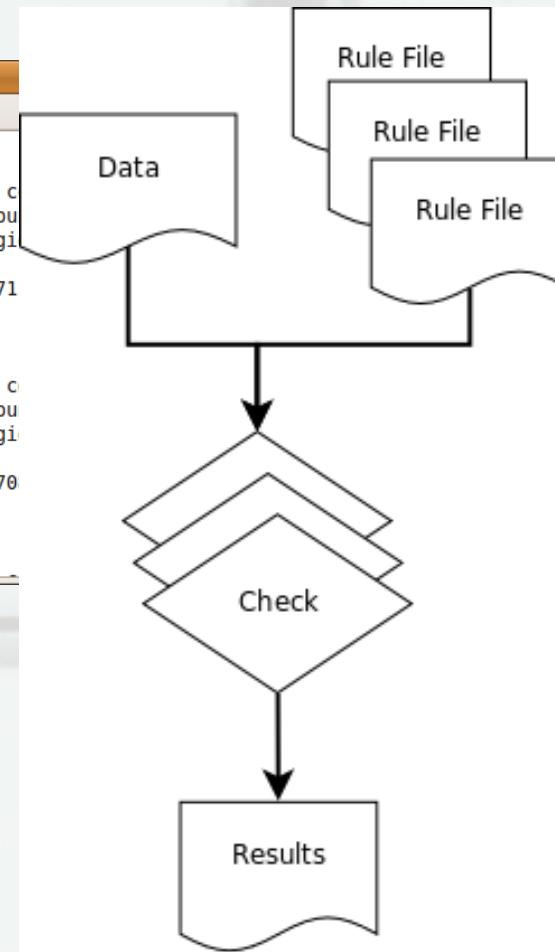


The screenshot shows a terminal window titled 'airport_departure_frequencies.csv'. The window contains a list of 18 lines of data, each representing an airport. The data is formatted as a CSV-like string with fields separated by semicolons. The fields include airport code, city, country, continent, region, and frequency. The data is as follows:

```
1 NYC;NEW YORK;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;1737711
2 CHI;CHICAGO;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;1434708
3 LON;LONDON;GB;UNITED KINGDOM;EU1;EUROPE : WESTERN EUROPE;1233338
4 ATL;ATLANTA;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;1021582
5 LAX;LOS ANGELES;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;892083
6 WAS;WASHINGTON;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;886712
7 FRA;FRANKFURT;DE;GERMANY;EU1;EUROPE : WESTERN EUROPE;855542
8 HOU;HOUSTON;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;855051
9 PAR;PARIS;FR;FRANCE;EU1;EUROPE : WESTERN EUROPE;767539
10 DEN;DENVER;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;704434
11 PHL;PHILADELPHIA;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;688416
12 DFW;DALLAS/FORT WORTH;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;657712
13 DTT;DETROIT;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;626023
14 AMS;AMSTERDAM;NL;NETHERLANDS;EU1;EUROPE : WESTERN EUROPE;604086
15 SFO;SAN FRANCISCO;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;589325
16 BOS;BOSTON;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;542525
17 SEA;SEATTLE;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;513016
18 MSP;MINNEAPOLIS/ST PAUL;US;UNITED STATES OF AMERICA;NA1;AMERICA : NORTH AMERICA;499197
```

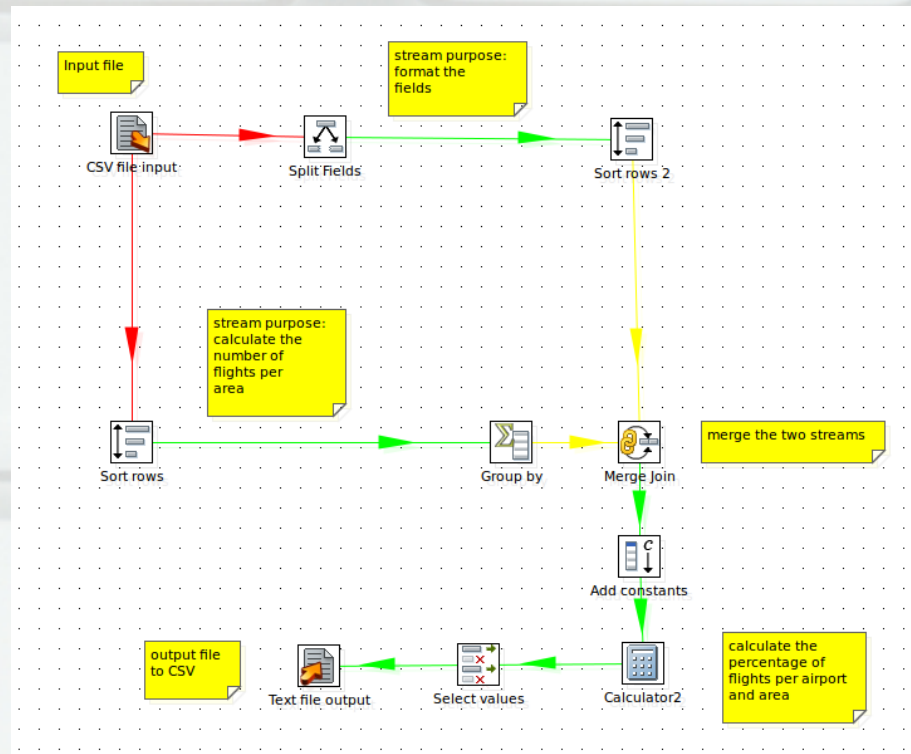
Practical Part – Step 2.2

```
airport_departure_frequencies_checks_20090528132842.log
File Edit Search Options Help
1|row: 0 group_1: failed=[false]
2|   subgroup subgroup_1 failed: [false]
3|     rule rule_city_code_0 failed: [false] - correct length: [3] of city c
4|     rule rule_country_code_0 failed: [false] - correct length: [2] of cou
5|     rule rule_region_code_0 failed: [false] - correct length: [3] of regi
6|   subgroup subgroup_2 failed: [false]
7|     rule rule_frequency_code_0 failed: [false] - correct numeric: [173771
8|
9|row: 1 group_1: failed=[false]
10|  subgroup subgroup_1 failed: [false]
11|    rule rule_city_code_0 failed: [false] - correct length: [3] of city c
12|    rule rule_country_code_0 failed: [false] - correct length: [2] of cou
13|    rule rule_region_code_0 failed: [false] - correct length: [3] of regi
14|  subgroup subgroup_2 failed: [false]
15|    rule rule_frequency_code_0 failed: [false] - correct numeric: [143470
16|
17|row: 2 group_1: failed=[false]
18|  subgroup subgroup_1 failed: [false]
19|    rule rule_city_code_0 failed: [false] - correct length: [3] of city c
```



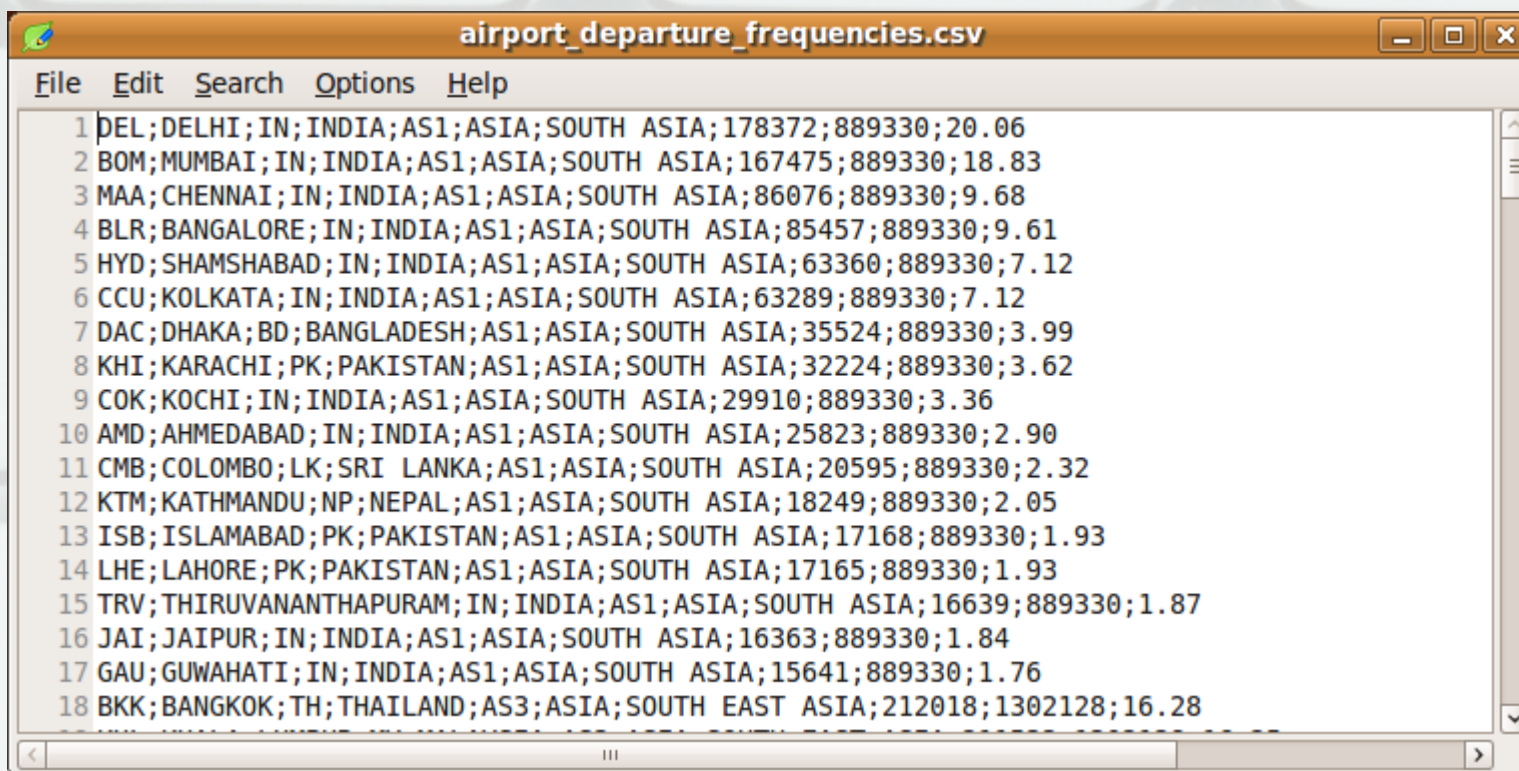
Practical Part – Step 3

- Step 3: Transform with Pentaho PDI (Spoon)



Practical Part – Step 3.1

Transformed with Pentaho PDI (Spoon)



```
airport_departure_frequencies.csv
File Edit Search Options Help
1 DEL;DELHI;IN;INDIA;AS1;ASIA;SOUTH ASIA;178372;889330;20.06
2 BOM;MUMBAI;IN;INDIA;AS1;ASIA;SOUTH ASIA;167475;889330;18.83
3 MAA;CHENNAI;IN;INDIA;AS1;ASIA;SOUTH ASIA;86076;889330;9.68
4 BLR;BANGALORE;IN;INDIA;AS1;ASIA;SOUTH ASIA;85457;889330;9.61
5 HYD;SHAMSHABAD;IN;INDIA;AS1;ASIA;SOUTH ASIA;63360;889330;7.12
6 CCU;KOLKATA;IN;INDIA;AS1;ASIA;SOUTH ASIA;63289;889330;7.12
7 DAC;DHAKA;BD;BANGLADESH;AS1;ASIA;SOUTH ASIA;35524;889330;3.99
8 KHI;KARACHI;PK;PAKISTAN;AS1;ASIA;SOUTH ASIA;32224;889330;3.62
9 COK;KOCHI;IN;INDIA;AS1;ASIA;SOUTH ASIA;29910;889330;3.36
10 AMD;AHMEDABAD;IN;INDIA;AS1;ASIA;SOUTH ASIA;25823;889330;2.90
11 CMB;COLOMBO;LK;SRI LANKA;AS1;ASIA;SOUTH ASIA;20595;889330;2.32
12 KTM;KATHMANDU;NP;NEPAL;AS1;ASIA;SOUTH ASIA;18249;889330;2.05
13 ISB;ISLAMABAD;PK;PAKISTAN;AS1;ASIA;SOUTH ASIA;17168;889330;1.93
14 LHE;LAHORE;PK;PAKISTAN;AS1;ASIA;SOUTH ASIA;17165;889330;1.93
15 TRV;THIRUVANANTHAPURAM;IN;INDIA;AS1;ASIA;SOUTH ASIA;16639;889330;1.87
16 JAI;JAIPUR;IN;INDIA;AS1;ASIA;SOUTH ASIA;16363;889330;1.84
17 GAU;GUWAHATI;IN;INDIA;AS1;ASIA;SOUTH ASIA;15641;889330;1.76
18 BKK;BANGKOK;TH;THAILAND;AS3;ASIA;SOUTH EAST ASIA;212018;1302128;16.28
```

Practical Part – Step 4

- Step 4: Post-Process with Awk

```
##### step 4
# post-process files
.....

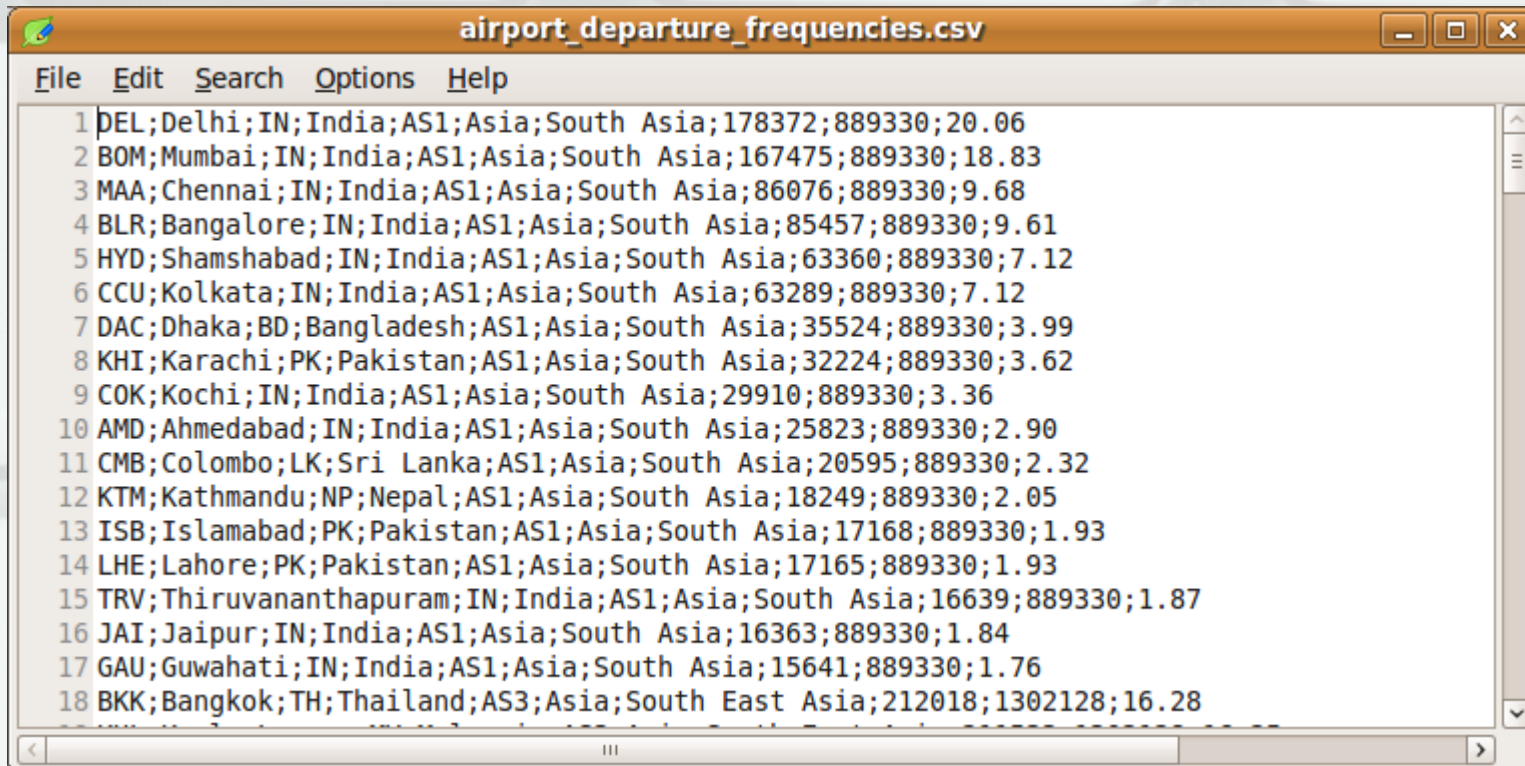
# we are pre-processing files: replacing invalid values, checking data,
  matching data
${application_process_folder}/${step_name}/capitalize_all_words.awk $
  ${application_output_folder}/${previous_step_name}/${filename_01} > $
  ${application_output_folder}/${step_name}/${filename_01}

# copy the final data file to a final location, such as the transfer drive
folder_copy_from=${application_output_folder}/${step_name}
folder_copy_to=${data_target_folder}

cp ${folder_copy_from}/${filename_01} ${folder_copy_to}
```

Practical Part – Step 4.1

Words capitalized with Awk



```
airport_departure_frequencies.csv
File Edit Search Options Help
1 DEL;Delhi;IN;India;AS1;Asia;South Asia;178372;889330;20.06
2 BOM;Mumbai;IN;India;AS1;Asia;South Asia;167475;889330;18.83
3 MAA;Chennai;IN;India;AS1;Asia;South Asia;86076;889330;9.68
4 BLR;Bangalore;IN;India;AS1;Asia;South Asia;85457;889330;9.61
5 HYD;Shamshabad;IN;India;AS1;Asia;South Asia;63360;889330;7.12
6 CCU;Kolkata;IN;India;AS1;Asia;South Asia;63289;889330;7.12
7 DAC;Dhaka;BD;Bangladesh;AS1;Asia;South Asia;35524;889330;3.99
8 KHI;Karachi;PK;Pakistan;AS1;Asia;South Asia;32224;889330;3.62
9 COK;Kochi;IN;India;AS1;Asia;South Asia;29910;889330;3.36
10 AMD;Ahmedabad;IN;India;AS1;Asia;South Asia;25823;889330;2.90
11 CMB;Colombo;LK;Sri Lanka;AS1;Asia;South Asia;20595;889330;2.32
12 KTM;Kathmandu;NP;Nepal;AS1;Asia;South Asia;18249;889330;2.05
13 ISB;Islamabad;PK;Pakistan;AS1;Asia;South Asia;17168;889330;1.93
14 LHE;Lahore;PK;Pakistan;AS1;Asia;South Asia;17165;889330;1.93
15 TRV;Thiruvananthapuram;IN;India;AS1;Asia;South Asia;16639;889330;1.87
16 JAI;Jaipur;IN;India;AS1;Asia;South Asia;16363;889330;1.84
17 GAU;Guwahati;IN;India;AS1;Asia;South Asia;15641;889330;1.76
18 BKK;Bangkok;TH;Thailand;AS3;Asia;South East Asia;212018;1302128;16.28
```

Practical Part – Step 5

- Step 5: Load into Infobright

```
##### step 5
# load files into the data warehouse

- .....

inputfile=${application_output_folder}/${step_name}/${filename_01}
export inputfile

# call script to load data
. ${application_process_folder}/${step_name}/load_data.sh
```

Practical Part – Step 5.1

Load into infobright

```
uwe@CHZLM0SGNB3017: ~  
File Edit View Terminal Help  
mysql> describe departure_frequencies;  
+-----+-----+-----+-----+-----+-----+  
| Field | Type | Null | Key | Default | Extra |  
+-----+-----+-----+-----+-----+-----+  
| Airport_Code | char(3) | YES | | NULL | |  
| Airport_Name | varchar(40) | YES | | NULL | |  
| Country_Code | char(2) | YES | | NULL | |  
| Country_Name | varchar(40) | YES | | NULL | |  
| Region_Code | char(3) | YES | | NULL | |  
| Region_Name | varchar(40) | YES | | NULL | |  
| Subregion_Name | varchar(40) | YES | | NULL | |  
| Airport_Departure_Frequency | int(11) | YES | | NULL | |  
| Region_Departure_Frequency | int(11) | YES | | NULL | |  
| Airport_Frequency_Percentage_of_Region | decimal(5,2) | YES | | NULL | |  
+-----+-----+-----+-----+-----+-----+  
10 rows in set (0.00 sec)  
  
mysql> select * from departure_frequencies limit 1\G  
***** 1. row *****  
Airport_Code: DEL  
Airport_Name: Delhi  
Country_Code: IN  
Country_Name: India  
Region_Code: AS1  
Region_Name: Asia  
Subregion_Name: South Asia  
Airport_Departure_Frequency: 178372  
Region_Departure_Frequency: 889330  
Airport_Frequency_Percentage_of_Region: 20.06  
1 row in set (0.00 sec)  
  
mysql>
```

Practical Part – Step 6

- Step 6: Clean up

```
##### step 6
# clean up step: remove all obsolete files
.....

# list of files that will be removed
remove_file_step_01=${application_output_folder}/step_01/${filename_01}
remove_file_step_02=${application_output_folder}/step_02/${filename_01}
remove_file_step_03=${application_output_folder}/step_03/${filename_01}
remove_file_step_04=${application_output_folder}/step_04/${filename_01}

# remove the files that we don't need anymore
#rm -f ${remove_file_step_01} ${remove_file_step_02} $
    {remove_file_step_03} ${remove_file_step_04}
```

Tips

- ETL
 - here we have to get things right!
 - Keep it flexible
 - Batch processing/scheduling
- DWH
 - for the Business not for IT
 - A lot of details (Granularity)
 - use Infobright (limitations)

Thanks!

Presentation and Screencast available

Contact me:

Uwe Geercken

uwe.geercken@swissport.com