Data Integration Design Report

Group F
Mariana Narvaez Cubillos
Nisrine Ferahi
Jonas Hellevang
Daniel Bilitewski
Rabiga Shangereyeva
Guillermo Chacon Clericus

Master in Business Analytics and Big Data October 2019 – Section 1
Business Intelligence & Data Warehousing – Professor Josep Curto

A – Instructions

1. Download and unzip the folder *data_integration_groupF.zip* uploaded on campus online along with these instructions.

2. Download the data from https://public.enigma.com/datasets/federal-communications-commission-licenses/9ee2d718-cb8f-4e35-b751-540ca590f2a0?filter=%2B%5B%5D%E2%86%93callsign.

3. Rename the csv file to *fcc_licenses.csv* and place it in the *Input* folder which is in the *data_integration_groupF* folder.

4. Execute the jobs in the following order:

   - Job_FCC_Hubs_Satellites.kjb
   - Job_Geolocation_Hub_Satellite.kjb
   - Job_FCC_Links.kjb
   - Job_Link_Antenna_Building_Geolocation.kjb
   - Job_Link_Antenna_OperatingEntity.kjb
   - Job_Link_License_LicenseHolder_OperatingEntity_Service.kjb

5. All the jobs take about three hours to run. Please refer to Appendix A for screenshots of the metrics of every job.

6. Due to the size of the dataset, we expect PDI to have at least 10GB of available RAM assigned to it. Should there be issues with the size of the dataset, download a subset of the data from dropbox.com and go back to step 3.
   https://www.dropbox.com/s/31x0aaavb7be0rn/fcc_licenses.csv?dl=0

B – Changes to the Data Warehouse Schema

Three major changes were made to the data warehouse schema.

Tables. The tables *hub_antenna_model* and *satellite_antenna_model* was removed, and the columns added to *satellite_antenna*. The reason behind this is that there was not enough justification to have the columns *antenna_make, antenna_model* and *antenna_type* in a separate table, while columns such as *azimuth* or *structure_height* remain in the *satellite_antenna* table. This way, all columns describing the antenna can be found in a single table.

Primary Keys. Since the *antenna_id*, *frequency_id, emission_id,* etc. are already unique identifiers to their respective tables, they are now also used as primary keys. Since the original dataset has a size of almost 19 million rows and nearly 100 columns, optimizing the computational efficiency of the data integration process is essential. Using these columns as primary keys has the advantage that a link can be loaded without the *Combination Lookup/Update* block. This block is very slow, because it needs to look up information in at least two satellites, to find out which satellite primary keys to match as foreign keys in the link. Having columns such as *antenna_id*, which already exists in the original dataset, as primary keys means that the *Combination Lookup/Update* block can be avoided, which significantly speeds up the transformations creating the links.

Data Types. Various data types were changed if they were too small to load the data. The most typical examples are changing INT to BIGINT and increasing the amount of characters in a VARCHAR.

C – Data Integration Approach

Default Extraction. The data is loaded from the csv file using the *CSV file input* block and then only the required columns are kept using the *Select values* block. Since there is only one data source, the aforementioned csv file, this step is simple and consistent across all transformations loading the hubs and satellites.

Data Mapping. After extraction, the data is sorted, and only unique rows are kept. A number of things are considered here:

- Sort size: The sort size is always set to 20 million. If the *Sort rows* block doesn't take all rows into memory at the same time, it won't properly sort because it can't consider all rows. Not sorting properly would mean that there could be duplicate rows in the table.

- Sorting speed: Sorting algorithms typically don't scale linearly with the number of values to sorts, which means that reducing the number of values to sort significantly speeds up this step. Therefore, the number of copies of this step is set to four, so that each has to sort no more than approximately 4.5 million rows. To properly sort the output, the *Sorted merge* block is used.

- Columns to sort: Since columns such as *antenna_id*, etc. are enough to uniquely identify a row, the sorting and finding unique values will only be done on such columns, instead of all columns of the table, to speed up this process. This is also done because the *Sort rows* block can't properly deal with NULLs, which means that if it were used on all columns, these columns would need to first be filled with a default value. Doing this would significantly slow down the transformation and significantly increase the RAM required to run it, meaning sorting only on one column is a required strategy to avoid this.

- Unique rows: The *Unique rows* block is one of the few blocks were the number of copies is not set to four. This is because if there were four copies, each would only look at a quarter of the rows, which means the loaded database could end up having duplicate rows in the tables.

Metadata. The metadata is created using a *Get system info* block which takes the datestamp as it is in the computer loading the data. It also takes the hostname of the computer loading the data for the *source* column in each table. This way, when loading rows or tables at different times from different systems, this information can be retained. In this case, the computer loading the data is saved, since the csv file containing the original data resides on said computer. However, in a productive system, this information will be the source system from which the data is written to the data warehouse.

Loading Data. The hubs and satellites are loaded in the same transformation using the *Load table* block instead of the *Insert / update* block because the *Load table* block is approximately four times faster than the *Insert / update* block. This is even after adding statements to the database connection to force it to load data in bulk[1], which already gave a significant speed increase compared to the default method used by PDI, which loads each row individually to the database.

Data Quality. All transformations loading the hubs and satellites except *TR_Geolocation* follow the structure described above and don't do any additional cleaning. This is for two reasons:

[1] useServerPrepStmts=false&rewriteBatchedStatements=true

first, since this data is used for regulatory purpose, there is a high incentive to keep data in its original form to make sure that no information gets lost; and second, some of the data is unclean and the added value from cleaning it is so low that it's better to keep it in the original form. An example of the second reason is the contact information in the *satellite_operating_entity* table.

However, a significant effort is made to clean the geolocation data, because it would otherwise be impossible to create dashboards using maps. In the original data, the geolocation is divided into the four columns degrees, minutes, seconds and directions of longitude and latitude respectively. Some of the columns for minutes and seconds have NULLs, these will be replaced with zero, assuming that NULL means that only the degrees column is enough to describe the location. Next, there are a few rows where the degrees are above 180°, the minutes above 60´ and the seconds also above 60´´. In these cases, the value will be set to the respective maximum.

Tableau requires geolocation data in degrees as a decimal, with negative numbers for south and west. In the first step, the data is converted into degrees using the following formula

$$degrees_{decimal} = degrees + \frac{minutes}{60} + \frac{seconds}{3600}$$

Next, the data for the direction has to be cleaned. The most dominant categories are north and west, because this is where the US territory is. There are a lot of missing values and a few values not corresponding to 'N' or 'S' for latitude and 'W' or 'E' for longitude. Therefore, all values for latitude not corresponding to 'S' will be replace by 'N', since that is the most likely category. Likewise, for longitude, all values not corresponding to 'W' will be replaced by 'E'. This way, for all data where the real value can't be certain, it will be assumed it is north and east. After this is cleaned, the new degrees column will be multiplied by -1 if the latitude is 'S' and longitude is 'E' for the respective directions.

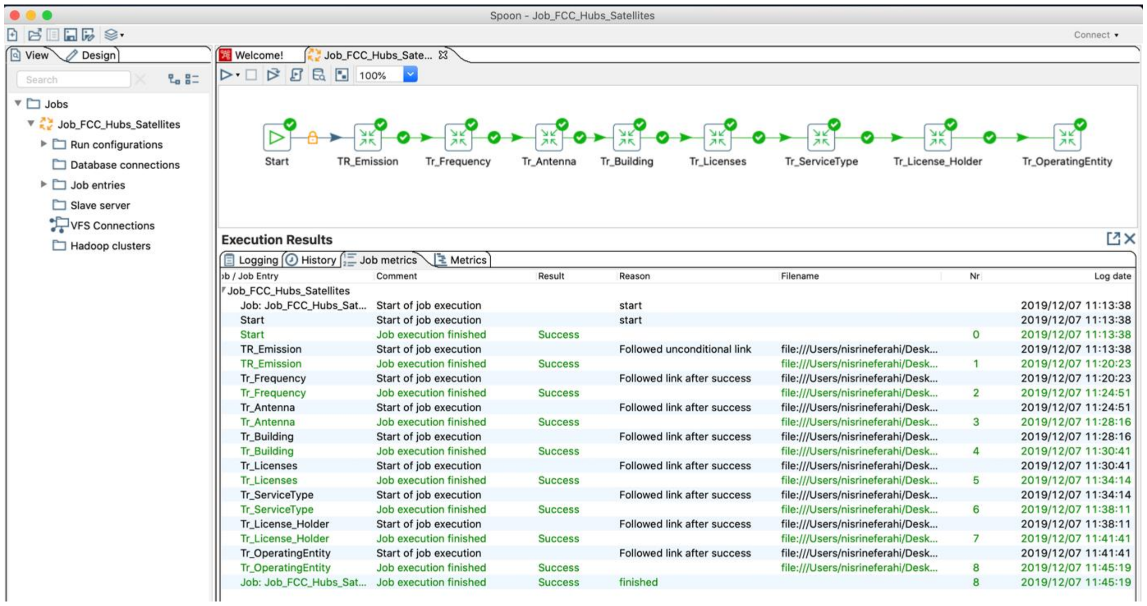
To avoid using a *Combination Lookup/Update* block, a primary key will be created by concatenating the degrees, minutes, seconds and direction columns for both latitude and longitude. To be able to keep it as a primary key of type BIGINT, the direction was previously encoded as a numeric Boolean variable (0 and 1). The advantage of this is that in the link, the same columns in the original dataset have to just be concatenated again to get the primary key, instead of doing the very computationally expensive lookup. The remaining steps in the transformation are the same as for all others.

<u>Links.</u> The transformations for the links have a similar structure to the transformations for the hubs and satellites. In case where the primary keys it takes from the satellites are columns that also exist in the original datasets, such as *antenna_id, frequency_id, emission_id,* etc., then they can just be selected from the csv file. If the primary key is a created auto-incremental value, then the *Combination Lookup/Update* block is used.

<u>Other Considerations.</u> Since the transformation loading *hub_geolocation* and *satellite_geolocation* requires more than 10 GB of RAM, which is difficult to assign to PDI even on the computers with 16 GB of RAM, this transformation has been split into four parts. To do this, the data is split on the degrees latitude column to ensure no duplicates in the final table. For example, the first transformation is for all rows where degrees latitude is smaller than 30. For loading the links, the transformation for loading *Link_Antenna_OperatingEntity* has been split into four parts as well using the antenna id column to split the data. Also, the transformation for loading *Link_Antenna_Building_Geolocation* has been split into four part using the degrees latitude column. And the transformation *Link_License_LicenseHolder_OperatingEntity_Service* has been split into four parts as well using license id column.

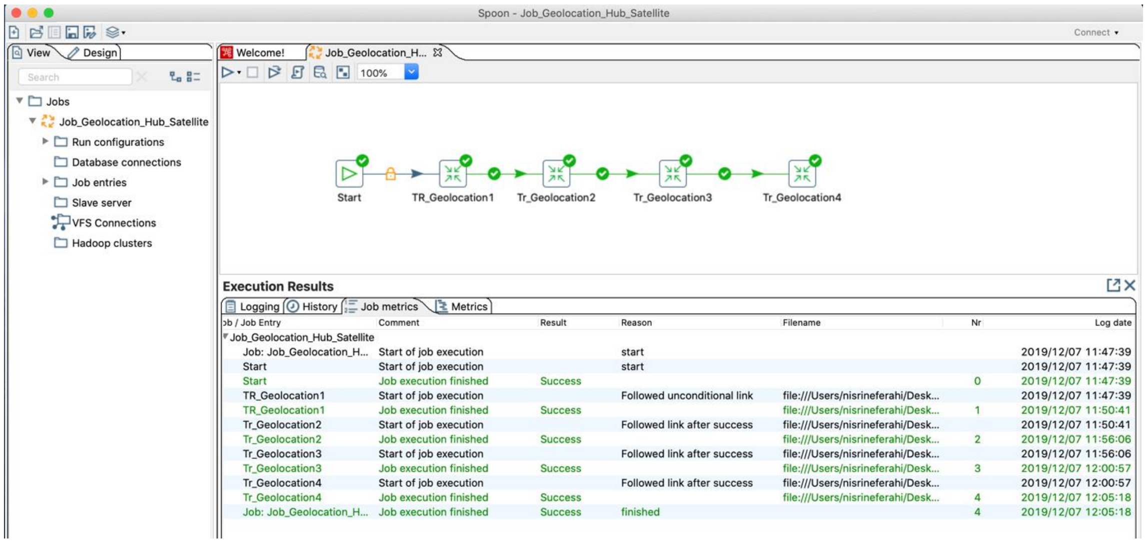
Appendix A – Metrics for Running the Jobs in Pentaho.

- Job_FCC_Hubs_Satellites.kjb



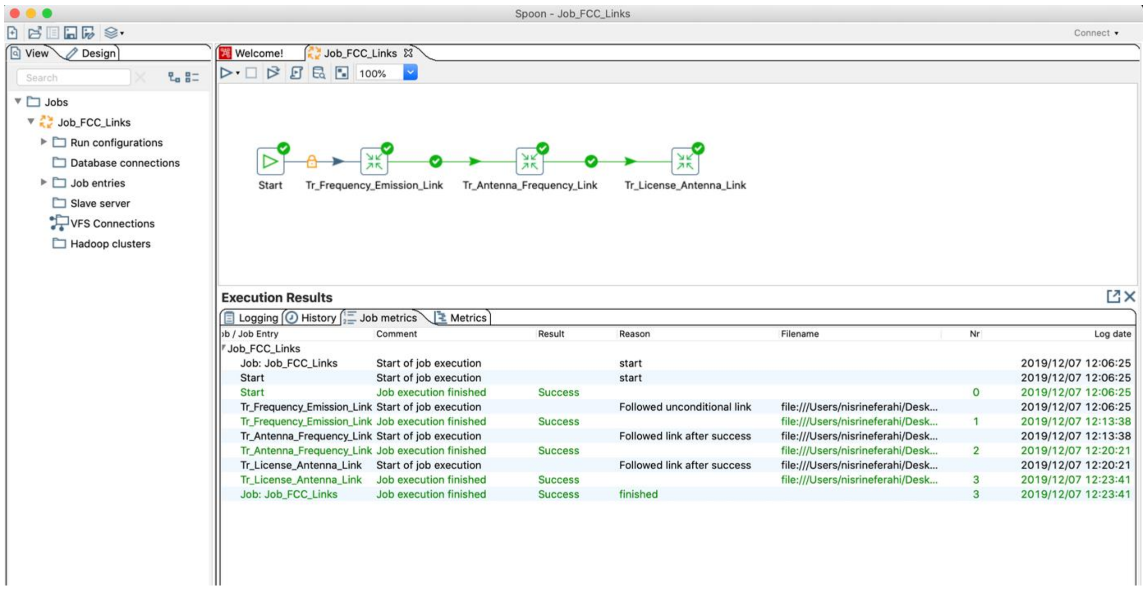
›b / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job_FCC_Hubs_Satellites						
Job: Job_FCC_Hubs_Sat...	Start of job execution		start			2019/12/07 11:13:38
Start	Start of job execution		start			2019/12/07 11:13:38
Start	Job execution finished	Success			0	2019/12/07 11:13:38
TR_Emission	Start of job execution		Followed unconditional link	file:///Users/nisrineferahi/Desk...		2019/12/07 11:13:38
TR_Emission	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	1	2019/12/07 11:20:23
Tr_Frequency	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:20:23
Tr_Frequency	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	2	2019/12/07 11:24:51
Tr_Antenna	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:24:51
Tr_Antenna	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	3	2019/12/07 11:28:16
Tr_Building	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:28:16
Tr_Building	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	4	2019/12/07 11:30:41
Tr_Licenses	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:30:41
Tr_Licenses	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	5	2019/12/07 11:34:14
Tr_ServiceType	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:34:14
Tr_ServiceType	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	6	2019/12/07 11:38:11
Tr_License_Holder	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:38:11
Tr_License_Holder	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	7	2019/12/07 11:41:41
Tr_OperatingEntity	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:41:41
Tr_OperatingEntity	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	8	2019/12/07 11:45:19
Job: Job_FCC_Hubs_Sat...	Job execution finished	Success	finished		8	2019/12/07 11:45:19

- Job_Geolocation_Hub_Satellite.kjb



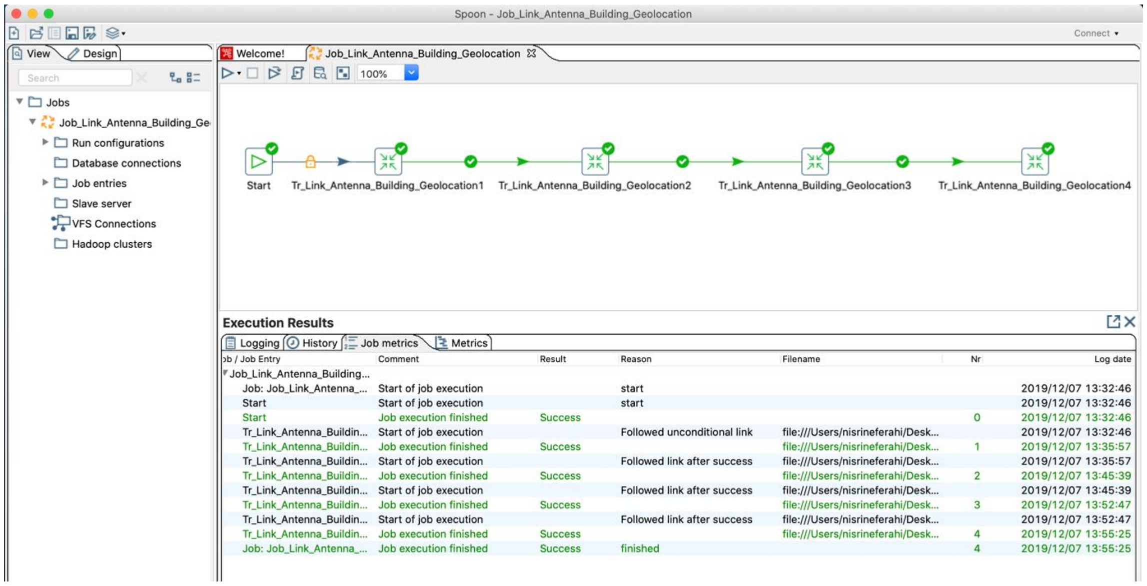
›b / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job_Geolocation_Hub_Satellite						
Job: Job_Geolocation_H...	Start of job execution		start			2019/12/07 11:47:39
Start	Start of job execution		start			2019/12/07 11:47:39
Start	Job execution finished	Success			0	2019/12/07 11:47:39
TR_Geolocation1	Start of job execution		Followed unconditional link	file:///Users/nisrineferahi/Desk...		2019/12/07 11:47:39
TR_Geolocation1	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	1	2019/12/07 11:50:41
Tr_Geolocation2	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:50:41
Tr_Geolocation2	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	2	2019/12/07 11:56:06
Tr_Geolocation3	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 11:56:06
Tr_Geolocation3	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	3	2019/12/07 12:00:57
Tr_Geolocation4	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 12:00:57
Tr_Geolocation4	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	4	2019/12/07 12:05:18
Job: Job_Geolocation_H...	Job execution finished	Success	finished		4	2019/12/07 12:05:18

- Job_FCC_Links.kjb



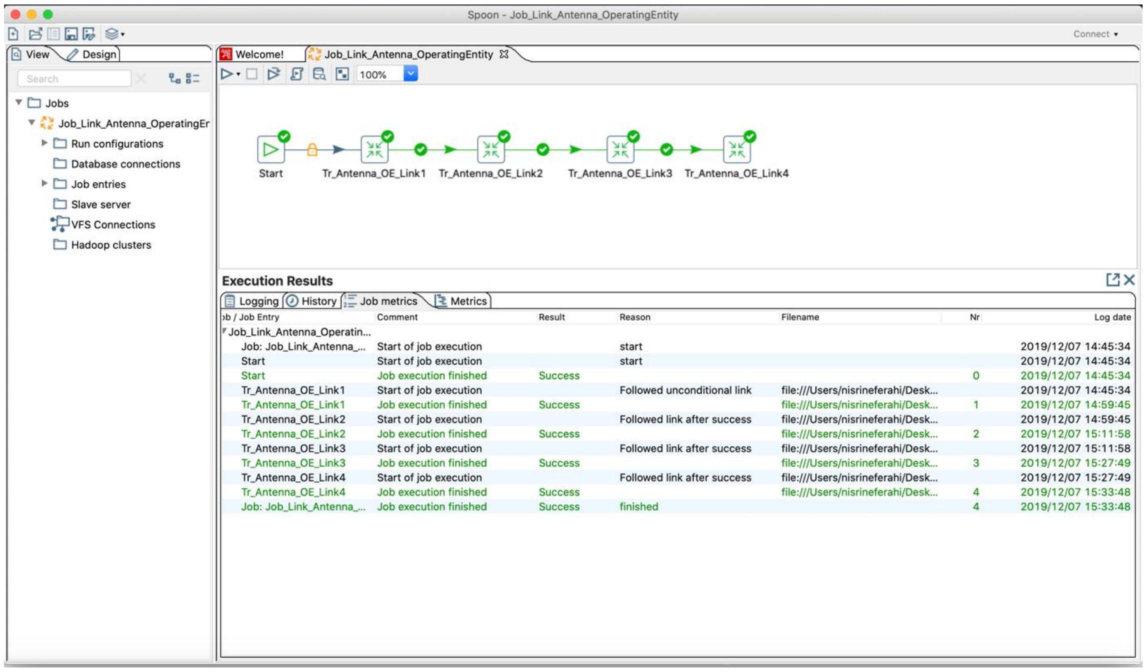
ob / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job_FCC_Links						
Job: Job_FCC_Links	Start of job execution		start			2019/12/07 12:06:25
Start	Start of job execution		start			2019/12/07 12:06:25
Start	Job execution finished	Success			0	2019/12/07 12:06:25
Tr_Frequency_Emission_Link	Start of job execution		Followed unconditional link	file:///Users/nisrineferahi/Desk...		2019/12/07 12:06:25
Tr_Frequency_Emission_Link	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	1	2019/12/07 12:13:38
Tr_Antenna_Frequency_Link	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 12:13:38
Tr_Antenna_Frequency_Link	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	2	2019/12/07 12:20:21
Tr_License_Antenna_Link	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 12:20:21
Tr_License_Antenna_Link	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	3	2019/12/07 12:23:41
Job: Job_FCC_Links	Job execution finished	Success	finished		3	2019/12/07 12:23:41

- Job_Link_Antenna_Building_Geolocation.kjb


Spoon - Job_Link_Antenna_Building_Geolocation
Connect
View
Design
Welcome!
Job_Link_Antenna_Building_Geolocation
100%
Jobs
Job_Link_Antenna_Building_Ge
Run configurations
Database connections
Job entries
Slave server
VFS Connections
Hadoop clusters
Start
Tr_Link_Antenna_Building_Geolocation1
Tr_Link_Antenna_Building_Geolocation2
Tr_Link_Antenna_Building_Geolocation3
Tr_Link_Antenna_Building_Geolocation4
Execution Results
Logging
History
Job metrics
Metrics

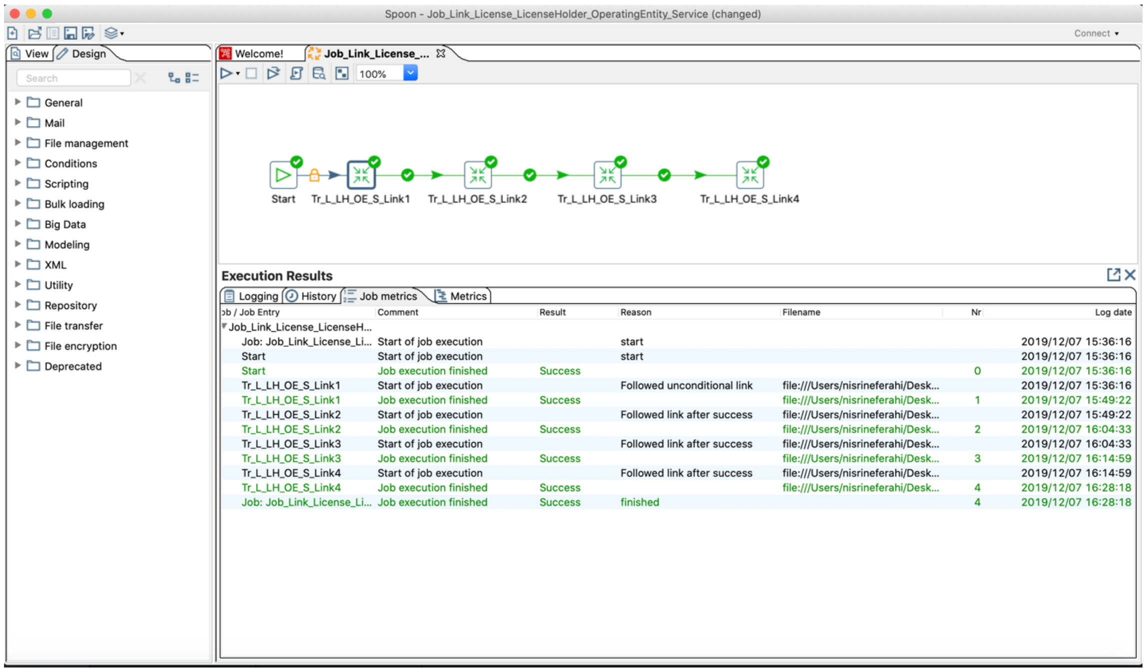
ob / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job_Link_Antenna_Building...						
Job: Job_Link_Antenna_...	Start of job execution		start			2019/12/07 13:32:46
Start	Start of job execution		start			2019/12/07 13:32:46
Start	Job execution finished	Success			0	2019/12/07 13:32:46
Tr_Link_Antenna_Buildin...	Start of job execution		Followed unconditional link	file:///Users/nisrineferahi/Desk...		2019/12/07 13:32:46
Tr_Link_Antenna_Buildin...	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	1	2019/12/07 13:35:57
Tr_Link_Antenna_Buildin...	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 13:35:57
Tr_Link_Antenna_Buildin...	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	2	2019/12/07 13:45:39
Tr_Link_Antenna_Buildin...	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 13:45:39
Tr_Link_Antenna_Buildin...	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	3	2019/12/07 13:52:47
Tr_Link_Antenna_Buildin...	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 13:52:47
Tr_Link_Antenna_Buildin...	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	4	2019/12/07 13:55:25
Job: Job_Link_Antenna_...	Job execution finished	Success	finished		4	2019/12/07 13:55:25

- Job_Link_Antenna_OperatingEntity.kjb



ob / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job_Link_Antenna_Operatin...						
Job: Job_Link_Antenna_...	Start of job execution		start			2019/12/07 14:45:34
Start	Start of job execution		start			2019/12/07 14:45:34
Start	Job execution finished	Success			0	2019/12/07 14:45:34
Tr_Antenna_OE_Link1	Start of job execution		Followed unconditional link	file:///Users/nisrineferahi/Desk...		2019/12/07 14:45:34
Tr_Antenna_OE_Link1	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	1	2019/12/07 14:59:45
Tr_Antenna_OE_Link2	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 14:59:45
Tr_Antenna_OE_Link2	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	2	2019/12/07 15:11:58
Tr_Antenna_OE_Link3	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 15:11:58
Tr_Antenna_OE_Link3	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	3	2019/12/07 15:27:49
Tr_Antenna_OE_Link4	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 15:27:49
Tr_Antenna_OE_Link4	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	4	2019/12/07 15:33:48
Job: Job_Link_Antenna_...	Job execution finished	Success	finished		4	2019/12/07 15:33:48

- Job_Link_License_LicenseHolder_OperatingEntity_Service.kjb



ob / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job_Link_License_LicenseH...						
Job: Job_Link_License_Li...	Start of job execution		start			2019/12/07 15:36:16
Start	Start of job execution		start			2019/12/07 15:36:16
Start	Job execution finished	Success			0	2019/12/07 15:36:16
Tr_L_LH_OE_S_Link1	Start of job execution		Followed unconditional link	file:///Users/nisrineferahi/Desk...		2019/12/07 15:36:16
Tr_L_LH_OE_S_Link1	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	1	2019/12/07 15:49:22
Tr_L_LH_OE_S_Link2	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 15:49:22
Tr_L_LH_OE_S_Link2	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	2	2019/12/07 16:04:33
Tr_L_LH_OE_S_Link3	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 16:04:33
Tr_L_LH_OE_S_Link3	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	3	2019/12/07 16:14:59
Tr_L_LH_OE_S_Link4	Start of job execution		Followed link after success	file:///Users/nisrineferahi/Desk...		2019/12/07 16:14:59
Tr_L_LH_OE_S_Link4	Job execution finished	Success		file:///Users/nisrineferahi/Desk...	4	2019/12/07 16:28:18
Job: Job_Link_License_Li...	Job execution finished	Success	finished		4	2019/12/07 16:28:18