Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 33 (2016) 626 – 634

Procedia Environmental Sciences

The 2nd International Symposium on LAPAN-IPB Satellite for Food Security and Environmental Monitoring 2015, LISAT-FSEM 2015

Extraction, Transformation, and Loading (ETL) module for hotspot spatial data warehouse using geokettle

Winda Astriani, Rina Trisminingsih*

Department of Computer Science, Bogor Agricultural University, Kampus IPB Darmaga, Bogor 16680, Indonesia

Abstract

Spatial data warehouse technology is one solution to the problem of big spatial data. Accumulation In the process of making spatial data warehouse, extraction, transformation, and loading (ETL) process has an important role to determine the quality of data. Manual ETL process requires a long time and makes a lot of queries. Therefore, this research uses Geokettle as a spatial ETL tool to integrate spatial data. This research used hotspot dataset of Indonesia from 2006 to 2014 and administrative districts data in Indonesia. This research performed ETL modeling with the simplification, adjustment, and design of ETL scenarios. The result of this research is ETL modeling implemented using Geokettle. SpagoBI Studio was used to create multidimensional data cubes. Moreover ETL testing was conducted using Geokettle, and spatial data warehouse testing was done by comparing the total number of hotspots between SQL query result and spatial analysis hotspot result on Quantum GIS.

Peer-review under responsibility of the organizing committee of LISAT-FSEM2015

Keywords: ETL; hotspot; spatial data warehouse.

1. Introduction

Forest is a natural resource potential and has an important role on earth. Serious problems experienced by Indonesia in forest management is forest fires. One of indicators used as the possibility of forest fires is hotspots. There are several satellites and remote sensing systems that can be used to monitor hotspots from the sky. Sensors that are used to monitor hotspots occurence, especially in Indonesia are NOAA-AVHRR sensor and Terra-MODIS

* Corresponding author. Tel.: +6281-2986-0241.
E-mail address: rina.ilkomipb@gmail.com.

Peer-review under responsibility of the organizing committee of LISAT-FSEM2015
doi:10.1016/j.proenv.2016.03.117

sensor [1]. Both sensors are manufactured by the US space agency (NASA), which provides data in realtime hotspot.

Hotspot data contains a dimension of time and location, consequently data collected would be big datasets, Spatial data warehouse technology is one solution to the problem accumulation of big spatial data. In the development of spatial data warehouse, the process of extraction, transformation, loading (ETL) has an important role. ETL process is a cornerstone of a data warehouse. An ETL design well will extract the data from the source systems, maintain data quality, applying standard rules, and presenting data in a variety forms that can be used in the decision-making process [2]. In this research, the ETL process automatically using spatial ETL tool that supports vector geometry data types and provides a consistent data integration that is Geokettle. In addition to the preprocessing of data, this study also makes a model that aims to design ETL scenarios, customize and simplify the mapping between the attributes in the data source with the attributes of the data warehouse tables. ETL application of using Geokettle expected to facilitate data warehouse developers in performing preprocessing data automatically that allows regulate the insertion of new data and update data without generating a lot of queries.

2. Data and research steps

2.1 Data

The data used in this research is hotspot datasets from 2006 to 2014, as the geometric vector data in with shapefile format (.shp). This hotspot data can be obtained at http://firms.modaps.eosdis.nasa.gov. Meanwhile the administrative district data provided by Geospatial Information Agency (BIG).

2.2 Researh Steps

The research step can be seen in Fig. 1.

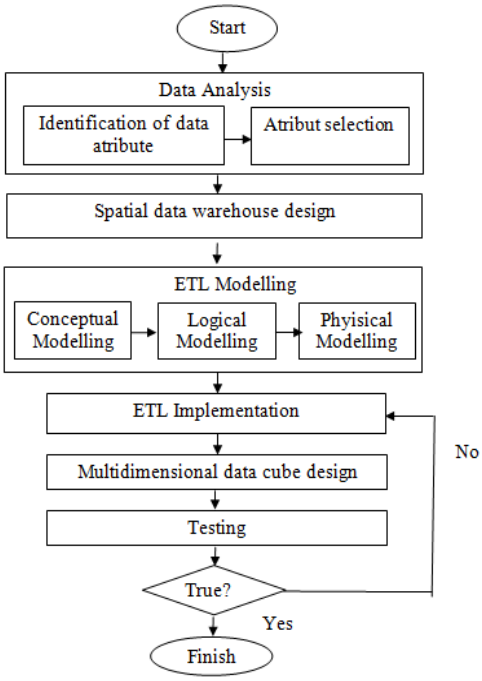


Fig. 1 Research steps

3. Results and discussions

3.1. Data analysis

- Identification of data attribute

Hotspot data is point geometry data type. Hotspot data has thirteen attribute, such as the_geom, latitude, longitude, brightness, scan, track, acq_date, acq_time, satellite, confidence, version, bright_t31 and FRP. Hotspaot data analysis was conducted based on the year, quartiles and months.

Administrative data is polygon geometry data type. Administrative data has eleven attribute, such as the_geom, object id_1, objectid, sub district, district, shape_leng, shape_le_1, shape_area, area, and province. Administrative data has not district code and island attribute. Information about island and district code obtained from the Central Statistics Agency (BPS), which is then extracted into the form of an excel (.xls). After getting district code and the island attribute, then the file is merged with the administrative map use the Quantum GIS (QGIS). On the data found a null value as much as 542 rows and 37 columns district lines on province column of 10 484 lines. The null data can be filled using the Quantum GIS by selecting the rows in the column are null districts and provinces magnification then performed on the area map, after it completes the data with the district and nearby provinces that exist in the map area.

- Attribute selection

The selected attributes data from hotspot data are the_geom hotspot, acq_date, and satellite. Attributes are selected on administrative data are the_geom, district code, island, province and district.

3.2. Spatial data warehouse design

At this step, a multidimensional scheme is made through a star schema that consist of a fact table (fact_forestfire), the time dimension (tb_waktu), satellite dimension (tb_satelit), and the location dimension (tb_lokasi). In the fact_forestfire table there is one measure. The measure is jumlah_hotspot as a numerical measure. The star schema result shows in Fig.2.

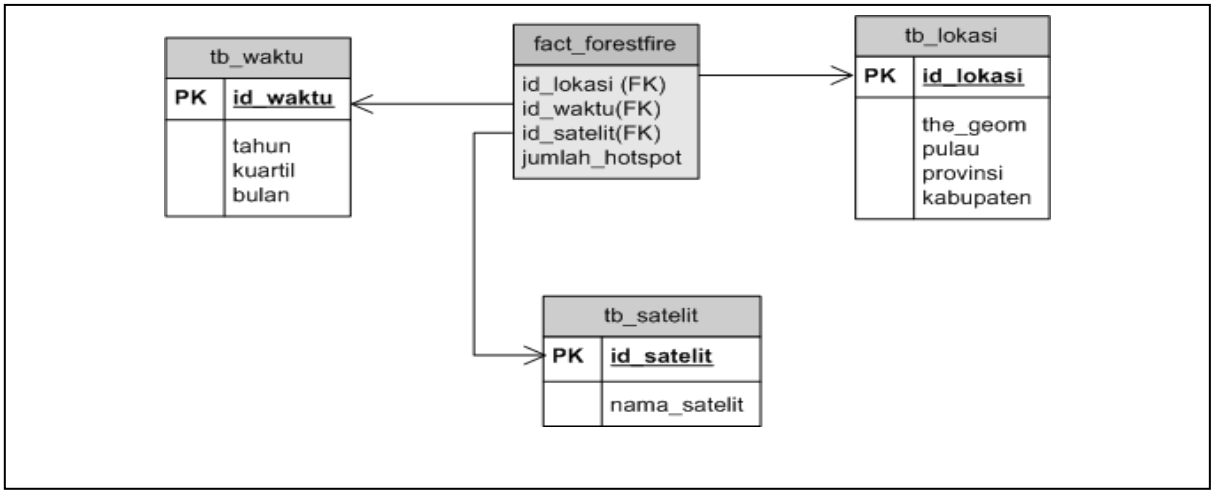


Fig. 2. Star schema result

3.3. ETL modelling

- Conceptual modelling

This step perform conceptual modelling for data warehouse table. Conceptual modeling aims to create a conceptual model for ETL process that describes the mapping of attributes from the data source to the attributes of the data warehouse tables 3. Conceptual modelling models is illustrated with the notation and templates. The result of conceptual modelling for fact_forestfire table shows in Fig. 3.

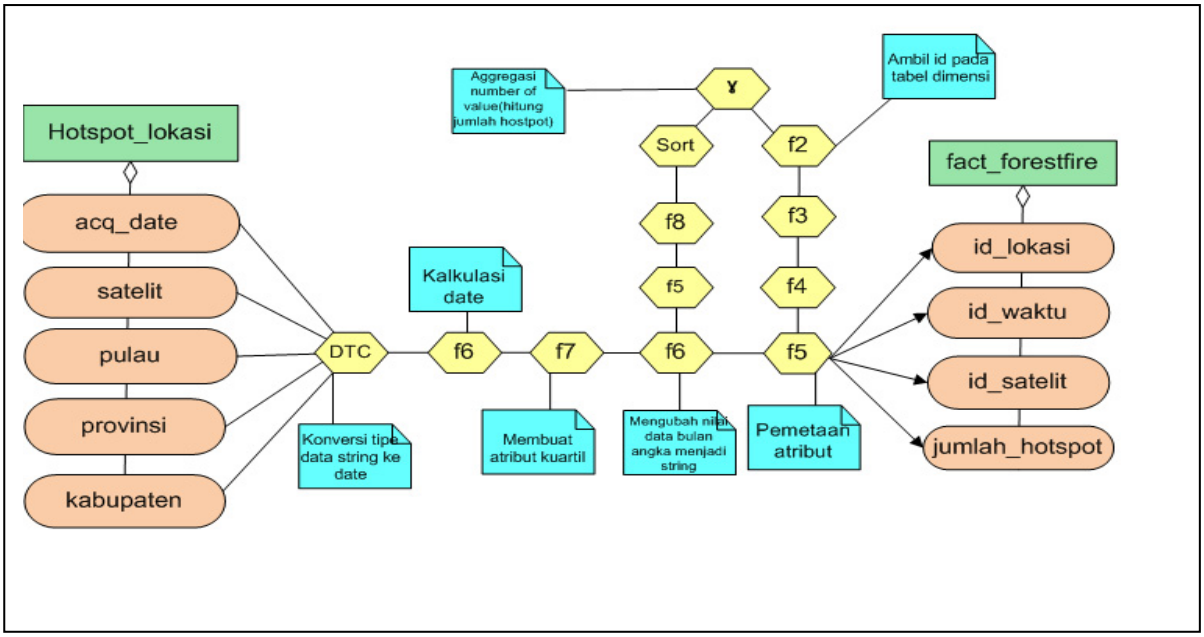


Fig. 3. Conceptual modelling for fact_forestfire table

Fig. 3. Shows mapping the attributes of the data sources from database to data warehouse table. The data source is hotspot_lokasi and data warehouse table is fact_forestfire. Fact_forestfire is fact table from data warehouse. The mapping through several transformation step, including the selection attributes, data type conversion, data consistency, data sorting, aggregation, retrieval id in each dimension that aims to get the foreign key of the dimension tables, and mapping attributes to match source the attributes of the fact_forestfire table.

- Logical modelling

This step perform logical modelling. Logical modeling to concentrate on the flow of data from the source towards the data warehouse through a process that ended in data storage [3]. Logical modelling is the development of the conceptual model, by changing the existing notation on conceptual modeling into logical modelling notations. Logical modelling for fact_forestfire tabel shown in Fig. 4.

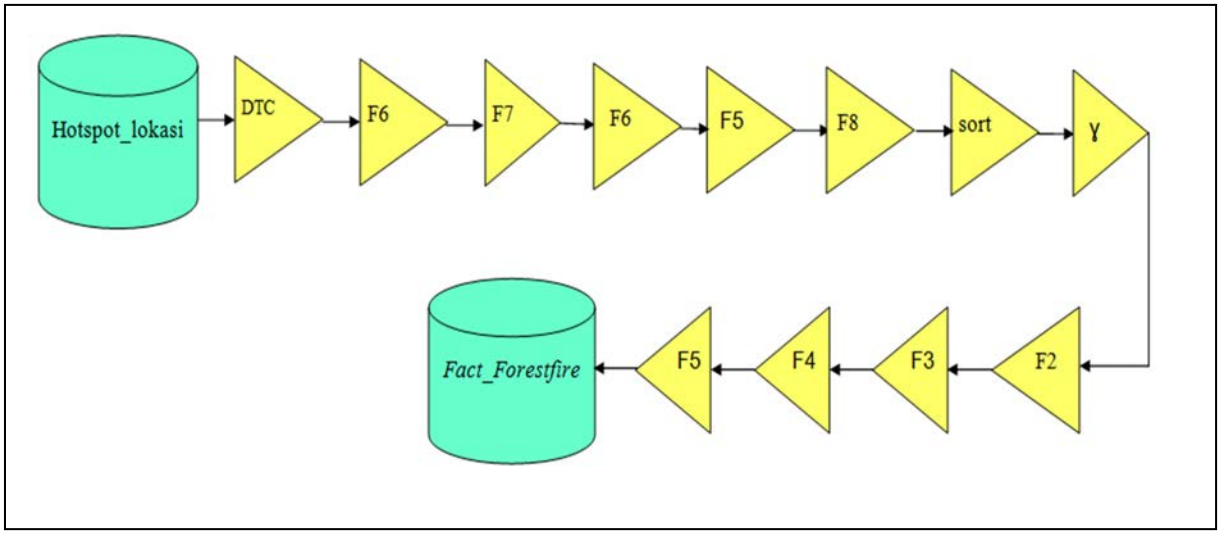


Fig. 4. Logical Modelling for fact_forestfire table

- Physical modellling

The physical modeling is the development of the conceptual model, by mapping the results of the transformation into an existing table in the data warehouse. At the specified physical modeling data types for each of the attributes of the data warehouse. Fisical modelling for fact_forestfire table shows in Fig. 5.

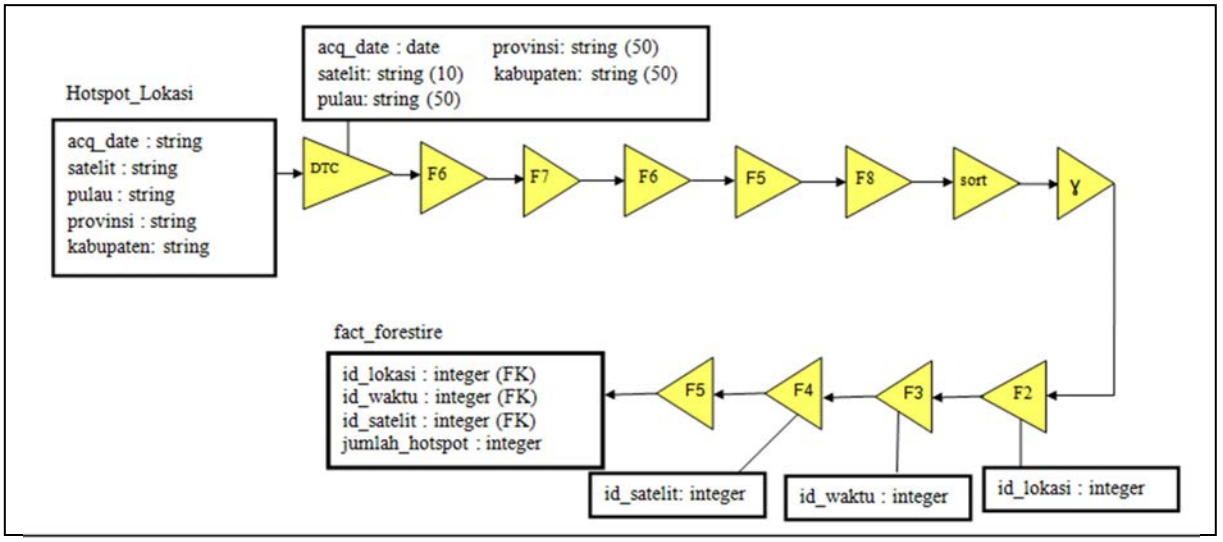


Fig. 5. Physical modelling for fact_forestfire table

Fig. 5 shows the physical modeling for fact_forestfire table. Fact_forestfire tabel contains of id_location, id_time and id_satelit attribute as a foreign key. The foreign key has integer data type. Fact forestfire also contains measure. The measure is total number of hotspot (jumlah_hotspot) as numerical measure. The measure has integer data type. The transformation process in physical modeling same as transformation process in the conceptual and logical modeling.

- ETL implementation

ETL process is implemented using Geokettle. Geokettle is spatial ETL tool and support the geometry vektor data, for example line, polygon and point. At this step, perform transformation modul for dimension tables and fact tables and a job module. The ETL implementation of fact_forestfire table shows in Fig. 6.

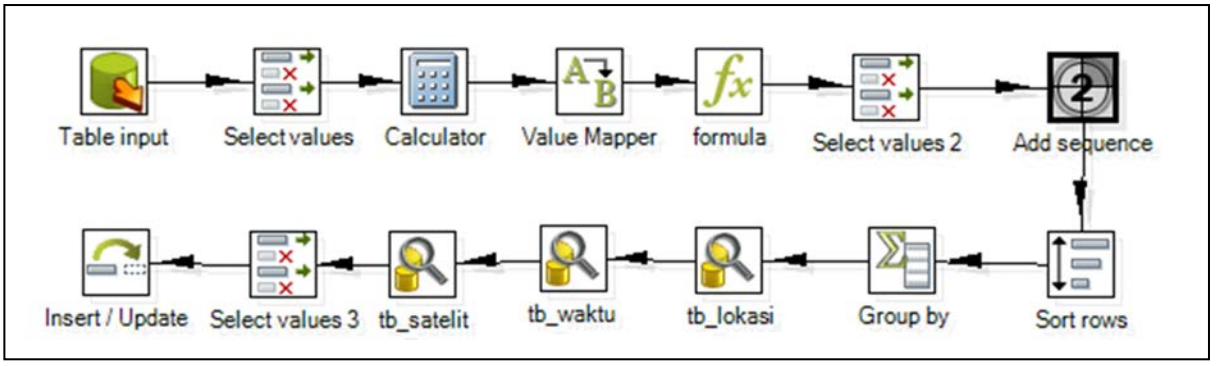


Fig. 6. ETL implementation for fact_forestfire table

The first step of Fig. 6. Is the data extraction from database source using table input step. The data source is hotspot_lokasi table. Hotspot_lokasi table is overlay result between administrative district data and hotspot data. Overlay is a merger of two or more maps and generating new information. Overlay process is performed by using spatial query. Query is used for spatial operations is as follows:

```
CREATE TABLE hotspot_lokasi
as select h.acq_date,h.satellite,l.pulau, l.provinsi, l.kabupaten
FROM forestfire as h, administratif as l
WHERE ST_WITHIN (h.hotspot_geom, l.the_geom);
```

Further transformation to make adjustments based on star schema attribute names that have been created using the select values step. Data type on attribute acq_date is string. Tab metadata on the select values step can change data type on acq_date attribute into date data type. In the next step carried solving attribute acq_date into year and month using the calculator step. Quartile attribute created through value mapper step. Value data of the month attribute still be an integer, for example 1 was changed to January using the formula step. Select values step is used to select the attributes that correspond to the data warehouse. The next step is to make an add sequence attribute which has a unique value and is used for group functions by or aggregation. This research use nonspatial aggregation. Before performing aggregation, the data must be sorted in advance to generate valid data, using sort rows step. Aggregation is performed to calculate the amount of certain hot spots using group by step based on kode_hotspot column. The fact table will show the id of each dimension and measurenya. Therefore, use a database lookup step to take the id of each dimension table. Table dimension consists of tb_lokasi, tb_waktu and tb_satelit. After getting the id of each dimension, attribute mapping from data source to the attributes of the fact_forestfire table using the select values step and the last stage is load the transformation result to data warehouse using the insert / update step. Insert/update step used to facilitate the insertion of new data or update data.

The whole transformation modules that have been made, then run the job module. Job module serves to regulate the order of transformation, scheduling transformation, and send notifications via email. The module begins with a job at a start step as the initialization to start the job. In executing the transformation, is used to run the transformation step transformation module location dimension, the dimension of time, and the dimensions of the satellite and fact_forestfire table. Mail step used to send notifications via email and generate a log of the ETL process. Success step is used to indicate the process works. If the transformation successfully updated, reported success through succes step, if not successful then the transformation that failed will be reported through the mail failure. Job module shows in Fig. 7.

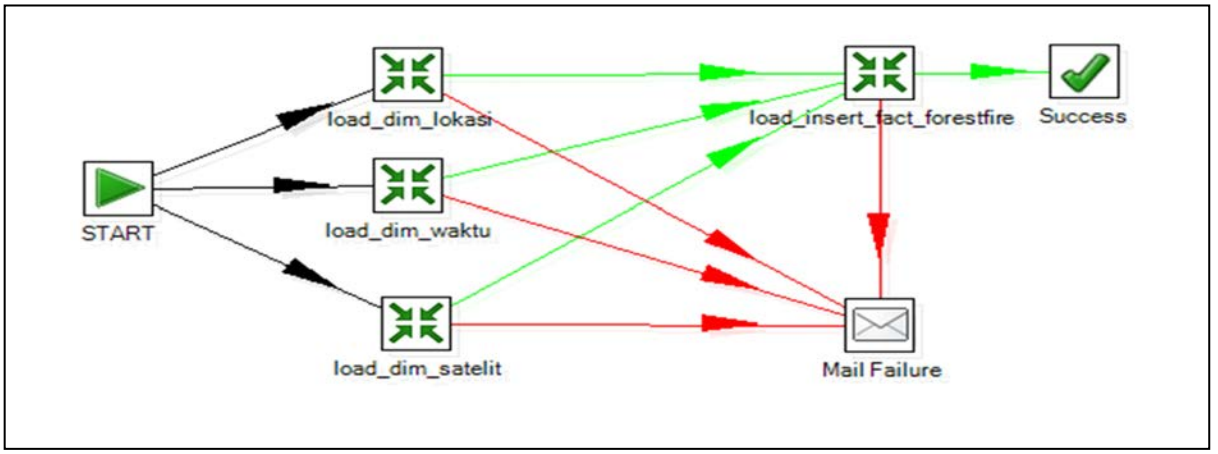


Fig. 7. Job module

- Multidimensional data cube design

At this step, the multidimensional data cube design using Spago BI Studio 5.1.0. The result is one data cube that is fact_forestfire. In the fact table is selected attributes to be a function measure, whereas the dimension tables arranged in order of type dimension and created a hierarchy of each dimension.. The result of data cube on SpagoBI shows in Fig. 8.

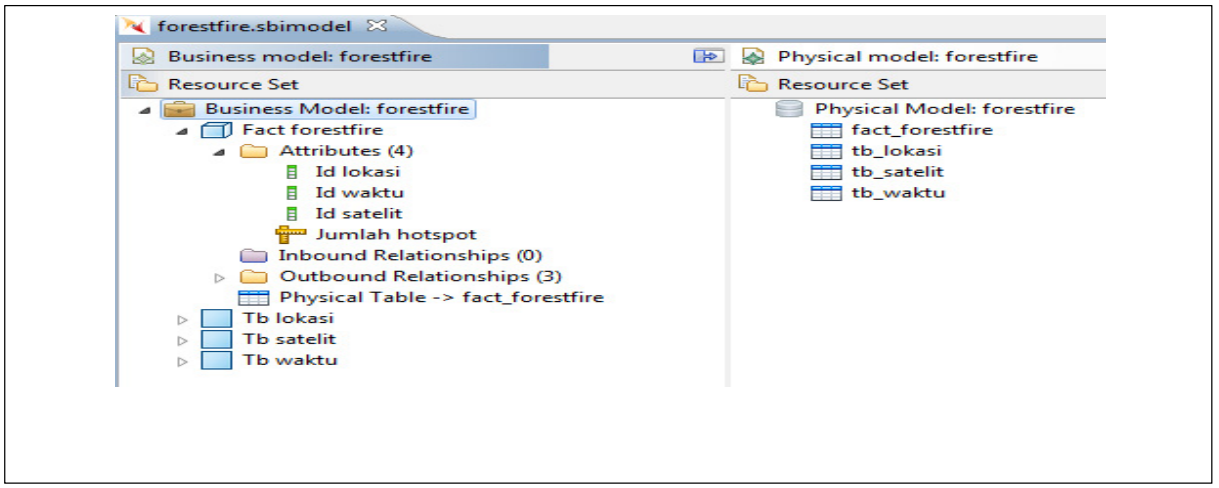


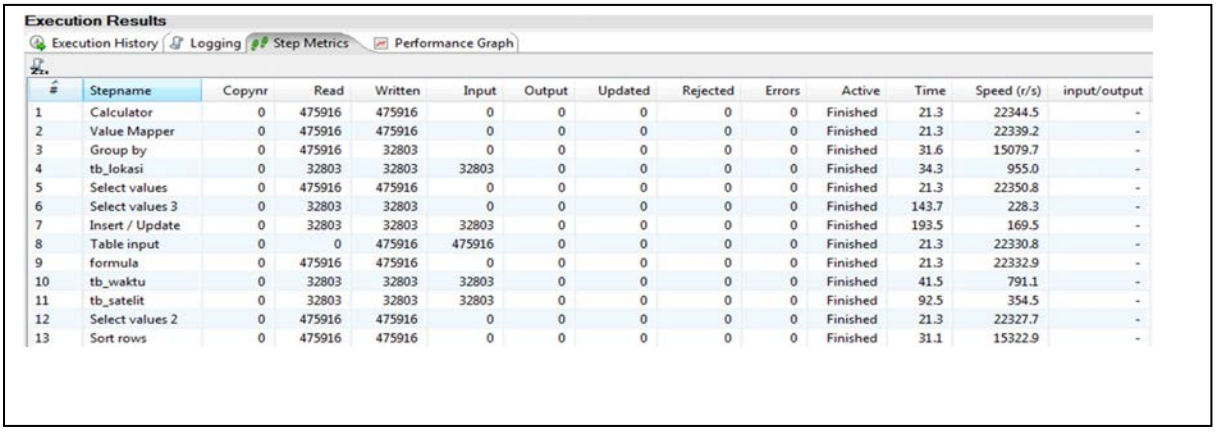
Fig. 8 . Multidimensional data cube on SpagoBI

- Testing

Testing is done through two phases are ETL testing and spatial ETL data warehouse testing.

- ETL Testing

ETL testing can be seen in the step metrics tab. Fact_forestfire table produce 32 803 rows with the status finished, which means the entire transformation successfully executed without error. The result of ETL testing shows in Fig. 9.



Execution Results

Execution History | Logging | Step Metrics | Performance Graph

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Calculator	0	475916	475916	0	0	0	0	0	Finished	21.3	22344.5	-
2	Value Mapper	0	475916	475916	0	0	0	0	0	Finished	21.3	22339.2	-
3	Group by	0	475916	32803	0	0	0	0	0	Finished	31.6	15079.7	-
4	tb_lokasi	0	32803	32803	32803	0	0	0	0	Finished	34.3	955.0	-
5	Select values	0	475916	475916	0	0	0	0	0	Finished	21.3	22350.8	-
6	Select values 3	0	32803	32803	0	0	0	0	0	Finished	143.7	228.3	-
7	Insert / Update	0	32803	32803	32803	0	0	0	0	Finished	193.5	169.5	-
8	Table input	0	0	475916	475916	0	0	0	0	Finished	21.3	22330.8	-
9	formula	0	475916	475916	0	0	0	0	0	Finished	21.3	22332.9	-
10	tb_waktu	0	32803	32803	32803	0	0	0	0	Finished	41.5	791.1	-
11	tb_satelit	0	32803	32803	32803	0	0	0	0	Finished	92.5	354.5	-
12	Select values 2	0	475916	475916	0	0	0	0	0	Finished	21.3	22327.7	-
13	Sort rows	0	475916	475916	0	0	0	0	0	Finished	31.1	15322.9	-

Fig. 9. ETL testing result

- Spatial data warehouse testing

Spatial data warehouse testing is done by comparing the number of hotspots between SQL query result and analysis hotspot result in QuantumGIS. At QuantumGIS added layer of data vectors for hotspots (points) and administrative data (polygon). Once all the data is added, then select the tab and selecting equipment vector analysis. In the analysis equipment, chosen point within the polygon to count the number of hotspots that exist in the polygon. The SQL uery result and the analysis hotspot result in QuantumGIS shows in Fig. 10.

Fig. 10 (a) SQL query result; (b) Analysis hotspot result in QuantumGIS.

The total number of hotspot for Indragiri Hulu district in JmlHotspot attribute on the Fig. 11 is 5764. The result same with the number of hotspots in the sql query.

4. Conclusions

This work has successfully created ETL modeling which implemented using Geokettle. ETL testing conducted use Geokettle and spatial data warehouse testing is done by comparing the total number of hotspots between sql query result and spatial analysis hotspot result on QuantumGIS.

Acknowledgements

The authors would like to thank all the research participants for their time, effort, and contribution to the research.

References

1. [IRI] International Research Institute. 2009. Sistem Peringatan Dini untuk Manajemen Kebakaran di Kalimantan Tengah [Online]. Available: http://crk.iri.columbia.edu/fire/exercises/fire_BA HASA.pdf.
2. Kimbal R, Caserta J. 2004. The Data Warehouse ETL Toolkit. USA: Willey Publishing, Inc.
3. Simitsis A. 2003. Modelling and Managing ETL Process. Di dalam: Scholl MH, Grust T, editor. Proceedings of the {VLDB} 2003 PhD Workshop. Co-located with the 29th International Conference on Very Large Data Bases (VLDB); 2003 September 12-13; Berlin, Germany. Berlin (DE) : CEUR-WS.org.