MODUL *EXTRACT, TRANSFORM, LOAD* UNTUK *DATA WAREHOUSE* KOMODITAS PERTANIAN INDONESIA MENGGUNAKAN TALEND

Rina Trisminingsih[1)], Intan Yuli Kiswari[2)]

[1,2]Departemen Ilmu Komputer, Institut Pertanian Bogor
e-mail: rina.ilkomipb@gmail.com[1], ykintan@gmail.com[2]

Abstrak

Kementerian Pertanian Indonesia menghimpun data hasil komoditas pertanian Indonesia dalam fail yang tidak saling terintegrasi. Solusi terbaik untuk melakukan integrasi data tersebut adalah data warehouse dengan proses *extract, transform, load* (ETL). Penelitian ini membangun ETL *data warehouse* hasil komoditas pertanian Indonesia yang memungkinkan praproses dan pembersihan data lebih cepat sebelum dimuat ke *data warehouse*. Pembangunan ETL diawali dengan merancang *data warehouse* menggunakan skema bintang dan merancang pemodelan ETL untuk melakukan transformasi. Transformasi dilakukan dengan membagi fail menjadi *header* dan *body file*. Implementasi model transformasi dilakukan menggunakan *tool* Talend. Hasil pengujian transformasi menunjukkan bahwa proses ETL berjalan dengan baik. Pengujian nilai menunjukkan bahwa nilai keluaran pada DBMS dan operasi OLAP menghasilkan nilai yang sama dengan nilai masukan yang berasal dari fail masukan.

Kata Kunci : *data warehouse*, ETL, komoditas pertanian, Talend.

Abstract

The Indonesian Ministry of Agriculture stores data of Indonesian agricultural commodities in files that are not integrated with each other. The best solution for the integration of these data is to create a data warehouse with extract, transform, load (ETL) processes. This study aimed to create an ETL data warehouse of Indonesian agricultural commodities that provides the possibility of faster pre-processing and data cleaning before being stored in the data warehouse. ETL development begins with designing a data warehouse using a star schema and designing the ETL model to perform the transformation. The transformation is done by dividing the file into a header and a body. The Implementation of the transformation model was done using Talend. The transformation test results showed that the ETL process runs well. The data value test showed that the resulting value of the DBMS and OLAP operations are the same values inputted from the inserted file.

Keywords*: agricultural commodities, data warehouse, ETL, Talend.*

I. PENDAHULUAN

Kementerian Pertanian (Kementan) Republik Indonesia menghimpun data komoditas hasil dari subsektor tanaman pangan, hortikultura, perkebunan dan peternakan di seluruh Indonesia. Data tersaji dalam situs Kementan yang dapat diakses berdasarkan nilai ukuran terpilih dengan identitas data, satuan dan periode tahun tertentu. Pengguna tidak dapat menampilkan informasi yang lebih besar atau lebih kecil karena data setiap indikator tersebut tidak saling terintegrasi. Pengguna dapat menerapkan praproses data manual untuk integrasi, namun akan membutuhkan waktu yang sangat lama. Integrasi data hasil komoditas pertanian dapat dilakukan dengan menggabungkannya ke dalam *data warehouse*. *Data warehouse* menawarkan arsitektur dan *tools* bagi para eksekutif bisnis untuk mengorganisir secara sistematik, memahami dan menggunakan data dalam pengambilan keputusan [1].

Pembangunan *data warehouse* hasil komoditas pertanian untuk subsektor tanaman hortikultura telah dilakukan pada penelitian Online Analytical Processing (OLAP) berbasis *web* pada tanaman hortikultura menggunakan Palo yang dilakukan [2] dan menggunakan SpagoBI oleh [3]. Namun permasalahan yang dihadapi adalah data yang diperoleh dari situs Kementan merupakan data yang masih memerlukan proses transformasi dari data mentah sehingga menjadi data yang sesuai dengan format *data warehouse*. Hal ini karena data yang dapat dimuat ke dalam *data warehouse* hanya data terstruktur dan sesuai dengan format. Sehingga diperlukan proses transformasi data untuk merapikannya ke dalam format yang sesuai dengan format data dalam *data warehouse* yang dituju.

Data warehouse umumnya dicirikan dengan adanya proses *Extract, Transform, Load* (ETL) yang

memungkinkan penggabungan data dari berbagai sumber, penyesuaian format dan pembuatan *datamart* untuk berbagai kebutuhan. Kesuksesan dalam pembangunan data warehouse bergantung pada kesuksesan proses *Extract, Transform, Load* (ETL) dari basis data Online Transactional Processing (OLTP) ke dalam *data warehouse* [4]. Meskipun proses ETL dalam *data warehouse* sangat penting, penelitian terkait bidang ini masih terbilang sedikit dilakukan. Hal ini karena sulit dan kurangnya model formal untuk mewakili aktivitas ETL yang memetakan data mentah dari sumber data yang berbeda ke dalam format yang sama untuk dipetakan ke dalam *data warehouse* [5]. Tugas utama ETL adalah melakukan ekstraksi, transformasi dan integrasi seluruh data yang kemudian dibersihkan sebelum dipetakan ke dalam *data warehouse* [6].

Pembangunan *data warehouse* hasil komoditas pertanian pada subsektor tanaman hortikultura oleh [2] dan [3] tersebut kemudian dilanjutkan oleh [7] dengan menambahkan modul ETL menggunakan Kettle. Penelitian tersebut masih melakukan transformasi data secara manual untuk menghasilkan data dengan format yang sesuai dengan format data warehouse, sehingga memerlukan waktu praproses data yang cukup lama.

Penelitian ini mengembangkan modul ETL untuk *data warehouse* komoditas pertanian Indonesia yang meliputi subsektor tanaman pangan, tanaman hortikultura, perkebunan dan peternakan. Data mentah yang digunakan dalam penelitian diperoleh dari situs resmi Kementrian Pertanian Indonesia. Peneliti memilih *tool* Talend Open Studio untuk mengimplementasikan pemodelan dalam melakukan transformasi data mentah hasil komoditas pertanian Indonesia.

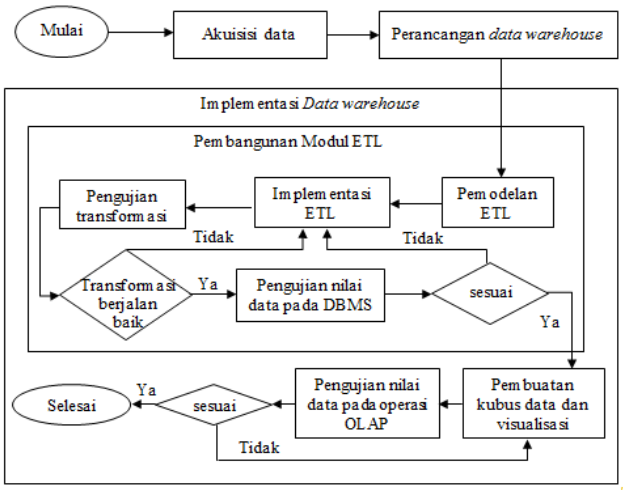
II. BAHAN DAN METODE

Data

Data yang digunakan dalam penelitian adalah data hasil komoditas pertanian yang diperoleh dari situs Kementrian Pertanian Republik Indonesia pada alamat http://aplikasi.pertanian.go.id/bdsp/newlok.asp. Penelitian ini memproses data dari subsektor tanaman pangan, tanaman hortikultura, peternakan dan perkebunan dari 30 komoditas dari 505 kabupaten.

Tahapan Penelitian

Arsitektur *data warehouse* terdapat empat **LAPISAN** yaitu sumber data, proses ETL, penyimpanan *data warehouse*, dan *end user*. Berkaitan dengan proses ETL *data warehouse*, *data warehouse* sebagai portal data yang canggih harus memiliki alat integrasi yang mampu mengakses data baik data terstruktur (*database*) maupun data tak terstruktur (dokumen). Proses ETL *data warehouse* terdiri dari ekstraksi data dari sumber data, transformasi dan pembersihan data dalam *Data Staging Area* (DSA), serta pemuatan data ke dalam *data warehouse* [6]. Tahapan penelitian yang dilakukan ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

1. Akuisisi Data

Seluruh data diperoleh dari situs Kementrian Pertanian Republik Indonesia dengan alamat http://aplikasi.pertanian.go.id/bdsp/index.asp. Data yang ditampilkan adalah nilai indikator dari setiap komoditas dengan identitas data yang meliputi keterangan subsektor, komoditas, indikator, satuan, level, status angka dan pada periode tahun tertentu. Indikator yang dimaksud adalah luas panen, produksi dan produktivitas, populasi dan pemotongan ternak.

2. Perancangan *Data warehouse*

Perancangan *data warehouse* dimulai dari perancangan skema model *data warehouse*, yaitu model data yang memiliki banyak dimensi (*multidimensional*). Setiap dimensi memiliki tabel asosiasi yang disebut tabel dimensi. Skema yang digunakan pada penelitian ini adalah skema bintang hasil pengembangan dari skema bintang *data warehouse* tanaman hortikultura oleh penelitian [7]. Skema bintang [7] dikembangkan dengan menambahkan *measure.*

3. Implementasi *Data warehouse*

Tahap implementasi *data warehouse* dilakukan dengan membangun modul ETL dan dan membangun kubus data untuk visualisasi OLAP pada *data warehouse*. Tahap pembangunan modul ETL meliputi pemodelan ETL nya dan

pengimplementasian pemodelan tersebut pada *tool* yang digunakan untuk membangun modul ETL.

a. Pemodelan ETL
Pemodelan ETL *data warehouse* meliputi pemodelan konseptual, pemodelan logika dan pemodelan fisik. Pemodelan konseptual menggabarkan konsep transformasi yang akan dilakukan. Pemodelan logika ETL menggambarkan alur kerja ETL yang fokus pada proses aliran data dari sumber data hingga menuju *data warehouse*. Gambaran alur ETL, aktivitas yang terlibat, kumpulan data, dan fungsi digambarkan menggunakan notasi *architecture graph* [8]. Pemodelan fisik dalam proses ETL fokus pada analisi data dan pemodelan setiap entitas basis data dalam DBMS.

b. Implementasi ETL
Implementasi dari pemodelan ETL dilakukan menggunakan *tools* Talend Open Studio. Tahap ini menghasilkan basis data dalam DBMS melalui proses ETL. *Extract*, *transform*, *load* merupakan kombinasi 3 fungsi yang secara otomatis mengambil data dari suatu sumber data dan menempatkannya ke basis data lain yang lebih besar [9]. Pada tahap implementasi ETL, penelitian ini memfokuskan pada proses *transform* untuk melakukan transformasi data mentah hasil komoditas pertanian Indonesia.

c. Pembuatan kubus data
Pembuatan kubus data dilakukan dalam SpagoBI Studio. Kubus data menampilkan data dalam banyak dimensi. Kubus data kemudian divisualisasikan dalam OLAP menggunakan SpagoBI *Server*.

4. Pengujian

Tahap pengujian melakukan tiga tahap pengujian yaitu pengujian transformasi pada *job* ETL, pengujian nilai pada DBMS dan pada operasi OLAP. Pengujian transformasi menunjukan apakah seluruh transformasi berjalan dengan baik atau tidak. Pengujian fungsi transformasi berhasil jika selama transformasi tidak terjadi *error* sistem. Pengujian nilai data pada DBMS menunjukkan apakah nilai data yang terdapat pada fail masukan sama dengan nilai keluaran pada DBMS setelah dilakukan transformasi. Pengujian nilai data pada operasi OLAP menunjukkan apakah nilai yang ditampilkan pada operasi OLAP *data warehouse* sama dengan nilai data pada fail masukan, maka pengujian telah berhasil dilakukan.

III. HASIL DAN PEMBAHASAN

Akuisisi Data

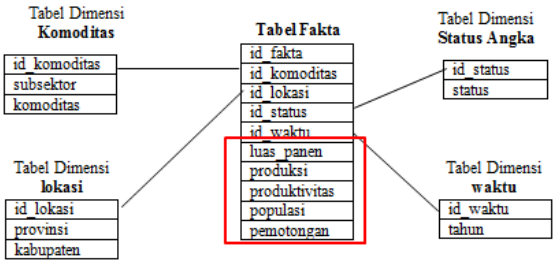
Data hasil komoditas pertanian Indonesia diperoleh dari situs Kementan dengan alamat http://aplikasi.pertanian.go.id/bdsp/newlok.asp. Fail tersebut terdiri dari enam fail dari subsektor perkebunan, delapan fail dari subsektor dari subsektor peternakan, tiga fail dari subsektor tanaman pangan dan 15 fail yang diambil dari subsektor tanaman hortikultura. Seluruh fail yang diunduh kemudian dilakukan perubahan format fail dari format .asp.xls menjadi format fail .xlsx. Perubahan format dilakukan agar fail dapat diproses dengan mudah ditransformasi untuk memperoleh data yang diperlukan dalam pembangunan *data warehouse*. Contoh tampilan fail data hasil komoditas pertanian Indonesia dapat dilihat pada Gambar 2.

	A	B	C	D	E	F	G	H	I	J	K
1	Hasil Pencarian Lokasi										
2	**Sub Sektor**	:	Perkebunan								
3	**Komoditi**	:	Jambu Mete								
4	**Indikator**	:	Produksi								
5	**Satuan**	:	Ton								
6	**Level**	:	Kabupaten/Kota								
7	**Provinsi**	:	Daerah Istimewa Yogyakarta								
8	**Status Angka**	:	Angka tetap								
9	**Tahun**	:	2000-2009								
10											
11											
12	**Lokasi**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
13	Kab. Kulon Progo	0	6	7	0	0	0	10	14	16	14
14	Kab. Bantul	0	3	33	0	0	0	123	125	124	124
15	Kab. Gunung Kidul	0	218	279	0	0	0	529	479	561	541
16	Kab. Sleman	0	12	13	0	0	0	12	12	6	24
17	Kota Yogyakarta	0	0	0	0	0	0	1	1	1	1
18											
19	**Sumber Data**	:	Kementerian Pertanian								

Gambar 2 Contoh tampilan fail data hasil komoditas pertanian Indonesia

Perancangan *Data warehouse*

Data warehouse dirancang menggunakan skema bintang dengan satu tabel fakta dan empat tabel dimensi yaitu dimensi komoditas, dimensi status angka, dimensi lokasi dan dimensi waktu. Perancangan *data warehouse* hasil komoditas pertanian Indonesia mengadopsi penelitian yang dilakukan oleh [7] dengan penambahan *measure*. *Measure* yang ditambahkan adalah populasi dan pemotongan. Skema bintang perancangan *data warehouse* setelah dilakukan penambahan *measure* dapat dilihat pada Gambar 3.



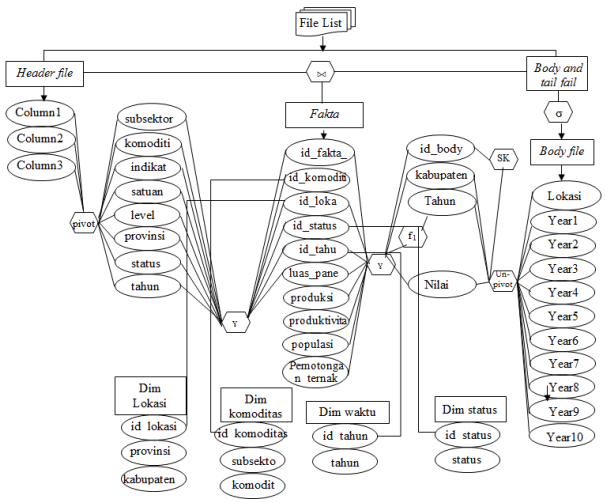
Gambar 3 Skema bintang perancangan data warehouse hasil komoditas pertanian

Pembangunan Modul ETL

1. Pemodelan ETL

Pemodelan konseptual pada tahap ini adalah membuat desain ETL untuk mengambil informasi dari setiap fail yang dieksekusi. Pemodelan konseptual ETL *data warehouse* hasil komoditas pertanian dapat dilihat pada Gambar 4. Fail masukan yang ditempatkan pada satu folder (*file list*) dieksekusi melalui dua *job flow. Job flow* pertama mengambil bagian *header* dan *jobflow* kedua mengambil bagian *body* dan *tail.*

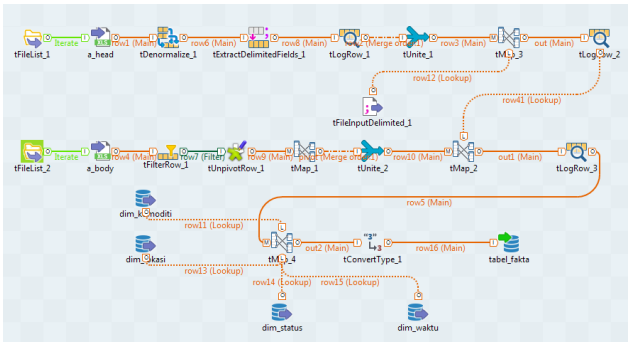
Header file diberikan perlakuan *pivoting* untuk mengubah struktur tabel pada *header file. Pivoting* mengubah tabel data 8 baris 3 kolom menjadi 3 baris 8 kolom yang kemudian dari ketiga baris tersebut dipilih baris header, baris data dan baris tidak terpakai yang dibuang. Informasi yang diperoleh dari *header file* digunakan sebagai *lookup* tabel untuk melakukan *join* dengan nilai keluaran dari *body file.*



Gambar 4 Pemodelan konseptual ETL *data warehouse*

2. Implementasi ETL

Pemodelan ETL yang telah dibuat diimplementasikan menggunakan *tool* Talend. Implementasi pemodelan konseptual pada Gambar 4 untuk modul ETL *data warehouse* hasil komoditas pertanian ditunjukan pada Gambar 5.



Gambar 5 Implementasi pemodelan ETL dalam Talend

Header file berisi informasi subsektor komoditas, indikator, satuan, level, provinsi, status angka dan tahun. Informasi tersebut disajikan di setiap baris, sehingga harus dilakukan *pivoting* untuk memindahkan posisi baris ke kolom dan *header file* yang telah dibersihkan dari kolom-kolom tidak terpakai terdiri dari dua kolom saja, yaitu kolom A dan kolom C. Kolom A dipilih sebagai *input column* yang diproyeksikan menjadi *header table*, sedangkan kolom C menjadi nilai yang tersimpan pada tabel. Hasil denormalisasi menggunakan tDenormalize menyimpan kedelapan baris yang diputar *(transpose)* masih berada dalam satu kolom yang dipisahkan oleh operator “;”. Kemudian komponen tExtractDelimitedFields akan melakukan *splitting* berdasarkan operator “;” sehingga menghasilkan delapan kolom hasil *splitting*. Nilai yang dihasilkan digabungkan oleh komponen tUnite dengan menunggu iterasi dilakukan terhadap seluruh fail masukan. Nilai tersebut kemudian digabungkan dengan dimensi lokasi yang memiliki atribut id_lokasi, Kabupaten dan Provinsi untuk mendapatkan nama kabupaten di setiap provinsi sehingga menghasilkan tabel identitas *job flow* kedua memproses *body* dan *tail file*. *Body file* merupakan tabel komoditas pada lokasi dan tahun tertentu. Informasi yang dikandung sudah terangkum dalam 11 kolom, yaitu kolom lokasi yang berisi nama kabupaten di Indonesia dan 10 kolom identitas tahun pada periode tertentu yang berbeda setiap failnya. *Tail file* berisi informasi sumber data. Nama kabupaten dan kota pada kolom tersebut selalu diawali dengan awalan “Kab.” atau “Kota”. *Filtering* untuk mendapatkan *body file* dilakukan dengan memilih *rows* yang diawali dengan tulisan ”Kab.” Atau “Kota”. Tabel ini dinormalisasi menggunakan komponen *tUnpivotRow* Talend. Hasil *unpivoting* di disimpan sementara oleh tUnite kemudian melakukan *join* dengan *lookup* tabel dari *job flow* pemrosesan *header file*.

Dari keluaran kedua *job flow* di atas komponen utama tabel fakta dibentuk dan disimpan oleh DBMS. Kolom identitas hasil pemrosesan *header file* selanjutnya digabungkan dengan tabel hasil pemrosesan *body file* sehingga terbentuk tabel fakta yang ditunjukkan pada Gambar 6.

id [PK] i	Sub_sektor character varyi	Komoditi characte	Indikator character var	Provinsi character varying	Kabupaten character vary	Tahu chara	Luas_ charac	Pro cha	Produk charac	Popu chara	Pemotongan_ character var
2098	Tanaman Panga	Jagung	Luas Panen	Sumatera Barat	Kab. Pasaman	2007	26707	0	0	0	0
2099	Tanaman Panga	Jagung	Luas Panen	Sumatera Barat	Kab. Pasaman	2008	41874	0	0	0	0
2100	Tanaman Panga	Jagung	Luas Panen	Sumatera Barat	Kab. Pasaman	2009	43493	0	0	0	0
2101	Tanaman Panga	Jagung	Luas Panen	Sumatera Barat	Kota Padang	2000	10	0	0	0	0
2102	Tanaman Panga	Jagung	Luas Panen	Sumatera Barat	Kota Padang	2001	0	0	0	0	0

Gambar 6 Tabel fakta hasil pemrosesan fail masukan

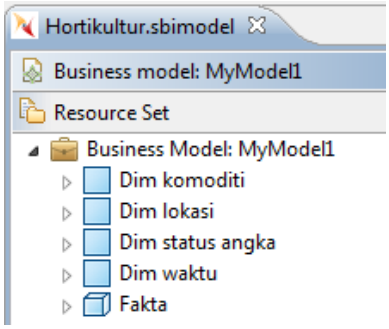
Tabel fakta hasil pemrosesan selanjutnya dilakukan pengkodean terhadap tabel fakta tersebut dengan menggunakan tabel dimensi sebagai *lookup kode* berdasarkan *id number*. Nilai keluaran tabel fakta dan keempat tabel dimensi disimpan sementara pada DBMS postgresql. Hasil keluaran dari tabel fakta yang telah dikodekan dengan tabel dimensi menghasilkan tabel fakta dengan pengkodeannya. Tabel fakta hasil pemetaan akhir yang dihasilkan pada DBMS ditunjukkan pada Gambar 7.

id_fakta [PK] integer	id_komoditi character var	id_lokasi character var	id_status character var	id_tahun character var	luas_panen integer	produksi integer	produktivitas integer	populasi integer	pemotongan_ integer
8	1	53	1	48	0	14	0	0	0
9	1	53	1	49	0	16	0	0	0
10	1	53	1	50	0	14	0	0	0
11	1	51	1	41	0	0	0	0	0

Gambar 7 Tabel fakta *data warehouse* hasil komoditas pertanian

Pembangunan Kubus Data

Kubus data untuk *data warehouse* hasil komoditas pertanian Indonesia dibuat menggunakan BI *platform* SpagoBI. Kubus data terdiri atas satu tabel fakta dan empat tabel dimensi. Pembuatan kubus data dimulai dengan menentukan dimensi dan kubus serta memilih atribut – atribut pada tabel kubus yang menjadi *measure*. Kemudian membuat hierarki dari masing-masing dimensi dan menentukan relasi antara tabel dimensi dan kubus. Kubus data yang dihasilkan dari SpagoBI Studio ditunjukkan pada Gambar 8.

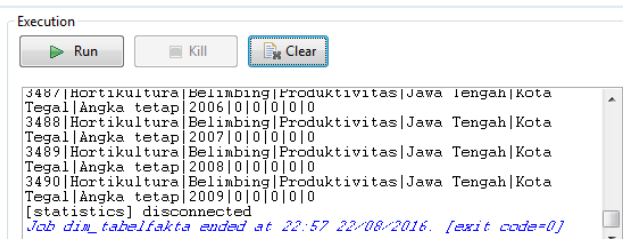


Gambar 8 Kubus *data warehouse* hasil komoditas pertanian

Pengujian

1. Pengujian Transformasi

Pengujian fungsi transformasi dilakukan dengan mengekseskusi semua *job* ETL pada Talend Open Studio. Pengujian transformasi berhasil dilakukan tanpa terjadi *error* pada jendela eksekusi yang ditunjukkan oleh Gambar 9.



Gambar 9 Hasil eksekusi *job* ETL untuk tabel fakta

2. Pengujian nilai data pada DBMS

Pengujian nilai data menunjukan apakah nilai yang terdapat pada DBMS dan nilai data yang terdapat pada tampilan hasil operasi OLAP telah sesuai dengan nilai pada data masukan sebelum melalui proses transformasi. Pengujian nilai data dimulai dengan membandingkan nilai data masukan dengan nilai data yang dihasilkan pada DBMS.

Berdasarkan data masukan produksi pepaya di Kab. Buton pada tahun 2005 adalah 511 Ton. Setelah dilakukan pengecekan bahwa komoditas pepaya memiliki kode 14 pada dimensi komoditas, kabupaten Buton memiliki kode 14 pada dimensi lokasi, status angka tetap memiliki kode 1 dan tahun 2005 memiliki kode 46. Tabel fakta dengan kode komoditas 14, kode lokasi 408, kode status 1 dan kode tahun 46 memiliki nilai 511, itu artinya nilai yang tersimpan pada tabel fakta sama dengan nilai yang terdapat pada data masukan.

3. Pengujian nilai data pada OLAP

Hasil Pengujian nilai data terhadap nilai data pada tampilan hasil operasi OLAP memiliki nilai yang sama dengan data masukan. Nilai data yang terdapat pada fail masukan memiliki nilai yang sama dengan nilai yang ditampilkan oleh operasi OLAP.

IV. KESIMPULAN DAN SARAN

Penelitian ini berhasil membangun modul ETL *data warehouse* untuk mentransformasikan data hasil komoditas pertanian Indonesia sehingga dapat diintegrasikan ke dalam *data warehouse*. Penelitian ini menghasilkan lima *job flow* transformasi untuk membangun empat tabel dimensi dan satu tabel fakta

yang digunakan untuk kebutuhan dalam membangun *data warehouse*. Transformasi telah berhasil di lakukan untuk lima *job* yang dihasilkan dan nilai yang dihasilkan oleh transformasi telah sesuai dengan nilai awal pada fail unduhan hasil komoditas pertanian Indonesia.

DAFTAR PUSTAKA

[1] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Ed ke-3. Massachusetts (US): Morgan Kaufman, 2011.

[2] F. Dwiprianti, "*Online analytical processing* (OLAP) berbasis *web* untuk tanaman hortikultura menggunakan Palo", skripsi, Departemen Ilmu Komputer, Institut Pertanian Bogor, Bogor (ID), 2015.

[3] E.R. Permana ER. "Aplikasi *online analytical processing* (OLAP) berbasis *web* dari data tanaman hortikultura menggunakan SpagoBI", skripsi, Departemen Ilmu Komputer, Institut Pertanian Bogor, Bogor (ID), 2015.

[4] A. Amborowati, "Analisis faktor – faktor yang mempengaruhi proses ETL pada data warehouse" dipresentasikan pada *Seminar Nasional Teknik Informatika*, UPN Veteran Yogyakarta, Yogyakarta, Indonesia, 2010 Mei 22

[5] S.H.A. El-Sappagh, A.M. Hendawi, dan El-A.H. Bastawissy AH, "A proposed model for data warehouse ETL processes". *Journal of King Saud – Computer and Information Sciences*. 23(2011): 91–104. doi:10.1016/j.jksuci.2011.05.005.

[6] Vassiliadis P, Simitsis A, Skiadopoulos S. 2002. Conceptual modeling for ETL processes. *5th ACM International Workshop on Data Warehousing and OLAP* (*DOLAP 2002*)*;* 2002 Nov 8; McLean, Virginia, USA. McLean (US): ACM. hlm 14-21.

[7] R. Hartomo, "Modul *extract, transform,* dan *load* untuk *data warehouse* tanaman hortikultura menggunakan Kettle", skripsi, Departemen Ilmu Komputer, Institut Pertanian Bogor, Bogor (ID), 2015.

[8] Simitsis A. 2005. Mapping conceptual to logical model for ETL processes. *8th ACM International Workshop on Data Warehousing and OLAP* (*DOLAP 2005*)*;* 2005 Nov 4-5; Bremen, Jerman. Bremen (DE): ACM. hlm 67-76.

[9] V. Gour, S.S. Sarangdevot, G.S. Tanwar, A. Sharma. "Improve performance of extract, transform and load (ETL) in data warehouse". *International Journal on Computer Science and Engineering (IJCSE)* 02(03):786-789, 2010