Perancangan Data Warehouse untuk Data Transaksi Penjualan Menggunakan *Schema Snowflake Studi Kasus : Online Market Dataset*

Ivan Rivaldo Marbun
Program Teknik Informatika
Universitas Kristen Satya Wacana
Salatiga, Indonesia
ivanr.marbun@gmail.com

Ramos Somya
Program Teknik Informatika
Universitas Kristen Satya Wacana
Salatiga, Indonesia
ramos.somya@uksw.edu

***Abstrak* – Database sebagai penyimpanan data banyak di implementasi pada berbagai bidang saat ini, seperti penggunaan database untuk menyimpan data transaksi penjualan. Banyaknya data yang disimpan pada sebuah database akan mempengaruhi kinerja dari database, seperti saat melakukan pencarian dan load data. Perancangan desain database yang tidak cermat dapat menyebabkan hilangnya data, data yang tidak konsisten dan redudansi data.Berdasarkan permasalah diatas diperlukan sebuah perancangan data warehouse yang terdiri dari beberapa dimensi tabel yang terintegrasi dan mempunyai sisi rentang waktu. Perancangan data warehouse ini menggunakan metode *nine-steps* Kimball dan skema *snowflake* sehingga dapat memodelkan tabel menjadi dimensi – dimensi yang terintegrasi. Untuk melakukan proses pengolahan data (ETL) menggunakan *Talend Open Studio* dan *Cloudera* sebagai platform penyimpanan data. Pada perancangan data warehouse ini juga menggunakan *Tableau* untuk menampilkan data yang sudah diproses dalam bentuk dashboard dan report. Perancangan ini akan menghasilkan data warehouse yang memuat data ke dalam dimensi – dimensi tabel sehingga data menjadi rapi dan terpusat serta data yang ditampilkan dalam bentuk report dan dashboard, akan mempermudah melakukan pencarian dan analisis data.**

Kata kunci— Data warehouse, skema snowflake, nine-steps Kimball, ETL (Extract, Transform dan Loading).

I. PENDAHULUAN

Tindakan memanfaatkan wadah seperti internet untuk menemukan dan menganalisa masalah menjadi suatu trend dimasa kini. Pemanfaatan ini tidak menutup kemungkinan pengembangan – pengembangan baru yang dapat dilakukan oleh seseorang. Melakukan tindakan analisa terhadap masalah yang sebelumnya telah diidentifikasi merupakan tindakan yang penting dalam merumuskan keputusan dari masalah tersebut. Analisa merupakan kegiatan berpikir dalam menguraikan keseluruhan data menjadi komponen yang setiap komponennya ketika digabungkan memiliki hubungan dan fungsi yang sepadan dan terpadu [1]. Identifikasi dan analisa masalah ketika dikonversi ke dalam sistem basis data, merupakan proses yang dapat memberikan hasil yang beragam. Data yang tersedia melalui internet dijadikan sebagai pilar untuk mengembangkan suatu masalah.

Desain sebuah database sangatlah berpengaruh besar, termasuk desain data, tipe data dan relasi antar tabel. Perancangan model desain database yang tidak cermat akan mempengaruhi data yang disimpan, seperti data yang tidak konsisten, redudansi data, proses update yang lambat dan banyak hal lain. Database merupakan jantung dari sistem informasi, oleh sebab itu dibutuhkan perancangan model desain database yang memenuhi efesiensi penyimpanan data dan efisiensi pembacaan maupun update data. Data warehouse yang terdiri dari beberapa tabel dimensi dan fakta yang terintegrasi, mampu membentuk data menjadi sebuah informasi trend atau analisis bisnis yang lebih baik. Pemanfaatan dataset online untuk dilakukan perancangan data warehouse, sehingga menjadi informasi penjualan produk – produk yang menjadi trend atau negara yang paling banyak melakukan pembelian dapat berguna untuk kemajuan bisnis.

Berdasarkan latar belakang yang telah dijelaskan, maka akan dilakukan penelitian yang bertujuan untuk melakukan perancangan data warehouse untuk data transaksi sebuah online market. Pada perancangan data warehouse ini akan menggunakan *schema snowflake* untuk *data modeling*. Penggunaan *schema snowflake* pada perancangan ini bertujuan untuk meminimalkan adanya data yang berlebihan (*data redundancy*) [2]. Perancangan data warehouse ini juga memanfaatkan teknologi Cloudera untuk HDFS (*Hadoop Distributed File System*) atau Sistem File untuk Distribusi Hadoop, dan memanfaatkan *impala* dan *hive* untuk melakukan query pada HUE salah satu *cluster Cloudera Manager* [3]. Pada penelitian ini menggunakan *Talend Data Platform* untuk melakukan *data processing* (ETL) [4] dan *Tableau* untuk menampilkan data yang informatif dari data warehouse yang telah dirancang [5].

II. TINJAUAN PUSTAKA

Pengembangan sistem analisis data menggunakan aplikasi OLAP Data Warehouse, membentuk data warehouse dengan menggunakan model dimensional yang memberikan kemudahan dan fleksibilitas untuk melakukan analisis dari berbagai sudut pandang bisnis [6]. Perancangan *Data Warehouse* yang di implementasikan untuk mengetahui *trend* produk, laporan analisa dalam bentuk tabel dan pivot grafik didapat dari hasil rancangan *data warehouse* menggunakan skema *snowflake*, yang menormalisasikan tabel – tabel dimensi dalam bentuk hirarki sebuah *cube* [7].

Pembentukan Data Warehouse menggunakan skema *snowflake* untuk perancangan *data warehouse* perpustakaan di perguruan tinggi, dalam tahapan perancangan skema memilih skema *snowflake* karena dalam *snowflake* schema, setiap tabel dimensi dapat memiliki sub-tabel lagi, yang bertujuan untuk meminimalkan adanya data yang berlebihan

(*redudancy* data). Dimensi data tersebut akan menjadi subjek informasi dalam pengambilan keputusan, karena pada setiap tabel dimensi data memungkinkan pemecahan yang lebih mendetail lagi sehingga dapat membuat informasi bisa menjadi lebih banyak dan detail [2].

Pembentukan desain *data warehouse* untuk pengukuran kinerja kereta api dalam mendistribusikan jumlah penumpang kereta api, yang berfokus pada pembagian tabel - table fakta dan dimensi ke dalam skema untuk mendapatkan desain organisasi *data warehouse* yang baik dan membuat proses integrasi data source ke *data warehouse* menjadi lebih efisien. Untuk mendapatkan output berupa report yang mampu menganalisa pembagian kinerja kereta api secara optimal untuk kebutuhan jumlah penumpang [1].

III. METODE PENELITIAN

A. Analisis

Terdapat 2 analisis yang dilakukan dalam tahap ini, yaitu analisis masalah dan analisis kebutuhan data. Analisis masalah dari penelitian ini adalah perancangan data warehouse untuk mengetahui *trend* dengan memanfaatkan teknologi untuk melakukan *data processing* (ETL), penyimpanan data dan pembentukan report. Analisis kebutuhan data dari penelitian ini adalah, dibutuhkan data mentah dengan format .csv, .xlx atau dataset lainnya.

B. Pengumpulan Data

Metode yang digunakan adalah studi *literature* untuk menemukan data yang akurat dan memenuhi krtiteria analisis kebutuhan data. Data yang digunakan merupakan data transaksi sebuah online market sebanyak 159.042 data dari sumber data set url :
(https://archive.ics.uci.edu/ml/datasets/online+retail)

C. Pemodelan Data

Pemodelan data menggunakan metode nine-steps Kimball, dengan tahap – tahap sebagai berikut :

- Menentukan proses bisnis (*Choosing the process*), sesuai dengan hasil analisis berdasarkan data yang digunakan, maka proses bisnis yang ditetapkan terkait penelitian ini adalah proses pengolahan data transaksi online market, dapat dilihat di Tabel I.

Tabel I. *CHOOSING THE PROCESS*

Proses Bisnis	Deskripsi	Fungsi yang terlibat
Pengolahan Data Transaksi Online Market	Mengumpulkan data transaksi dari sebuah online market, baik transaksi yang sukses maupun yang tidak.	Data transaksi, data pembeli, data produk,

- Menentukan *granularity* (*Choosing the grain*), pada bagian ini memutuskan secara pasti apa yang akan dipresentasikan oleh sebuah tabel fakta. Pada tahap ini ditentukan tingkat detail data yang bisa didapatkan dari model dimensional. Sehingga *granularity* yang dipilih adalah informasi transaksi penjualan, dapat dilihat di Tabel II.

Tabel II. *CHOOSING THE GRAIN*

Grain	Deskripsi	Proses bisnis yang terlibat
Informasi transaksi penjualan	Informasi transaksi penjualan online market terdiri dari tanggal transaksi, pembeli yang melakukan transaksi dan banyak barang yang dibeli baik transaksi yang sukses maupun yang tidak.	Data transaksi online market

- Identifikasi dan menyesuaikan dimensi (*Identifying and conforming the dimensions*), langkah ketiga dalam perancangan data warehouse yaitu mengidentifikasikan dimensi yang berhubungan dengan tabel fakta, dapat dilihat di Tabel III.

Tabel III. *IDENTIFYING AND CONFORMING THE DIMENSIONS*

Tabel Dimensi	Deskripsi	Grain
dim_calendar	Mempunyai atribut date_transaction, day, month dan year	Informasi transaksi penjualan
dim_product	Mempunyai atribut product_id, product_name dan id_category	Informasi transaksi penjualan
dim_category	Mempunyai atribut id_category dan category	Informasi transaksi penjualan
dim_customer	Mempunyai atribut customer_id dan country_id	Informasi transaksi penjualan
dim_country	Mempunyai id_country dan country	Informasi transaksi penjualan
dim_transaction	Mempunyai atribut transaction_id, payment_method dan transaction_status	Informasi transaksi penjualan

- Menentukan fakta (*Choosing the fact*), pada tahap ini pemilihan tabel fakta merupakan tabel yang dapat mengimplementasikan semua *grain* yang digunakan. Dalam hal ini fakta yang ditentukan adalah : *fact_sale,* dapat dilihat di Tabel IV.

Tabel IV. *CHOOSING THE FACT*

Fakta	Deskripsi	Dimensi
fact_sale	fact_sale berisikan informasi transaksi penjualan baik transaksi yang sukses maupun yang tidak.	dim_calendar, dim_product, dim_category, dim_customer, dim_country, dim_transaction

- Menyimpan hasil perhitungan sementara pada tabel fakta (*Storing pre-calculations in the fact table*), pada tahap ini dilakukan proses kalkulasi terhadap tabel fakta, dan menyimpan hasil pre-kalkulasi tersebut, sebagai berikut : *fact_sale* , tanggal transaksi, jumlah barang, harga dan diskon perlu dihitung dan disimpan sementara.
- Melengkapi tabel – tabel dimensi (*Rounding-out the* dimension *tables*), pada tahap ini dilakukan untuk melengkapi tabel dimensi dengan atribut dan keterangan, dapat dilihat di Tabel V.

TABEL V. ROUNDING-OUT THE DIMENSION TABLES

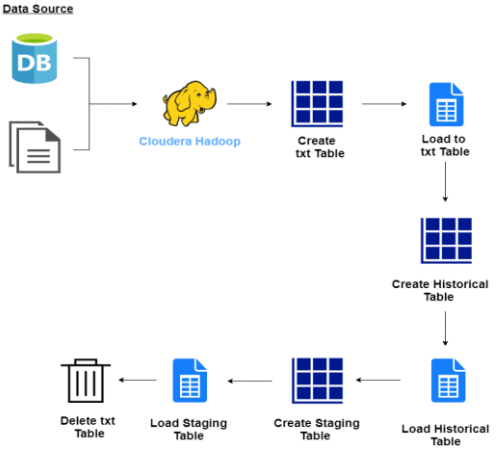
Dimensi	Atribut	Type (length)	Keterangan
dim_calendar	date_transaction	date	Tanggal transaksi
	day	int (10)	Hari transaksi
	month	int (10)	Bulan transaksi
	year	int (10)	Tahun transaksi
dim_customer	customer_id	varchar (10)	ID pelanggan
	country_id	varchar (10)	ID negara

dim_country	id_country	varchar (10)	ID negara
	country	varchar (30)	Nama negara
dim_transaction	transaction_id	varchar (10)	ID transaksi
	payment_method	varchar (10)	Metode pembayaran
	transaction_status	varchar (20)	Status transaksi
dim_product	product_id	int (10)	ID produk
	product_name	varchar (40)	Nama produk
	category_id	varchar (10)	ID kategori
dim_category	id_category	varchar (10)	ID kategori
	category	varchar (30)	Nama kategori

- Menentukan durasi dimensi (*Choosing the duration of the dimension*), untuk perancangan data warehouse ini ditetapkan data yang digunakan adalah data transaksi dari tanggal 01 – 01 – 2011 sampai 31 – 05 – 2011, dengan jumlah data sebanyak 159.042. Penentuan durasi dapat dilakukan dengan penambahan partisi pada data mart sebelum ditarik untuk pembentukan data warehouse.
- Menelusuri dimensi yang termasuk *slowly changing dimension* (*Tracking slowly changing dimension*), pada tahap ini memperhitungkan dimensi yang perlahan dapat telusuri perubahannya. Pada perancangan data warehouse ini, data harga dan diskon selalu dapat berubah dengan perlahan mengikuti standar harga pasaran. Selain itu data transaksi dan data pelanggan selalu berubah secara dinamis.
- Memutuskan prioritas *query* dan bentuknya (*Deciding the query priorities and the query modes*), tahap ini menggunakan perancangan fisik untuk menghasilkan data warehouse yang siap diimplementasikan. Pada tahap ini juga dibuat ketetapan query – query atau laporan (*reporting*) untuk dapat menampilkan data yang diinginkan [8].

D. Proses ETL (Extract, Transform dan Loading)

Proses ETL menggunakan Talend Open Studio sebagi data platform. Proses Extract merupakan tahap penarikan data dan cleaning data dari source data (.txt, .xls atau database lain). Proses Transform dilakukan untuk mengubah format tabel sesuai dengan model yang telah dirancang mengikuti metode nine-steps Kimball. Proses Loading dilakukan setelah proses Extract dan Transform selesai data akan tujukan ke dalam database yang baru atau sistem yang baru. Proses ETL pada Talend dapat dilihat pada Gambar 1.

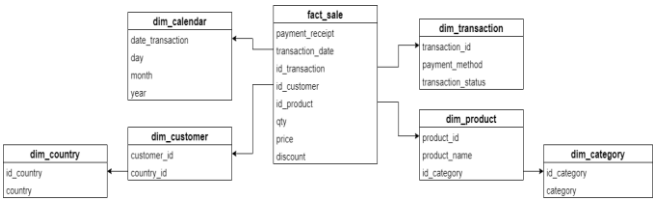


Gambar 1. Proses ETL Talend

Gambar 1 merupakan proses penarikan data dari *data source* ke dalam *database staging* dan *historical* dengan menambahkan partisi *date* dalam proses *load*-nya. *Database historical* dan *staging* disini merupakan *data lake* yang akan digunakan untuk pembentukan data warehouse. *Database historical* merupakan penampungan seluruh data yang ditarik dari *data source*, sedangkan *database staging* hanya menampung data sesuai dengan partisi *date* yang ingin dimuat. Setelah melalui proses ETL untuk pembentukan *data lake* dan *data warehouse*, selanjutnya data akan ditampilkan dalam bentuk *report* dan *dashboard* menggunakan Tableau. Untuk menguji pembentukan tabel, penarikan data dan kesamaan data akan dilakukan pengujian menggunakan metode *black box*.

IV. PEMBAHASAN DAN HASIL

Berikut merupakan hasil dari perancangan model skema data warehouse berdasarkan metode *nine-steps Kimball*. Skema yang dipakai dalam perancangan data warehouse ini adalah skema *snowflake*, penggunaan skema *snowflake* dikarenakan skema ini terorganisir sehingga membuat data tidak akan mengalami *redundant* dan dalam segi kompleksitas *query* skema *snowflake* lebih unggul dari skema lain seperti skema *star*.



Gambar 2. Skema Snowflake

Pada Gambar 2 dapat dilihat merupakan skema data warehouse yang terdiri dari 6 tabel dimensi dan 1 tabel fakta, dimana setiap tabel memiliki *primary key* (PK) untuk mengindentifikasi setiap baris data dengan menggunakan suatu data yang unik dan beberapa tabel memiliki *foreign key* (FK) sebagai *relation* untuk menunjukkan ke induk tabel biasanya di gunakan pada saat melakukan *query join* serta ada beberapa atribut sebagai *mandatory* yang dimana nilai data pada atribut tersebut tidak boleh kosong atau *null* seperti yang ditunjukkan pada Tabel VI.

Tabel VI. DETAIL TABEL

No	Nama Tabel	Deskripsi	Atribut
1	fact_sale	Berisi informasi lengkap seluruh transaksi, seperti tanggal, customer, produk dan jumlah produk.	a. payment_receipt (PK) b. transaction_date (FK) c. id_transaction (FK) d. id_customer (FK) e. id_product (FK) f. qty g. price h. discount
2	dim_transaction	Berisi informasi ID transaksi, metode pembayaran dan status transaksi.	a. transaction_id (PK) b. payment_method c. transaction_status
3	dim_calendar	Berisi seluruh tanggal transaksi.	a. date_transaction (PK) b. day c. month d. year
4	dim_customer	Berisi informasi seluruh pelanggan.	a. customer_id (PK) b. country_id (FK)
5	dim_country	Berisi Negara	a. id_country (PK)

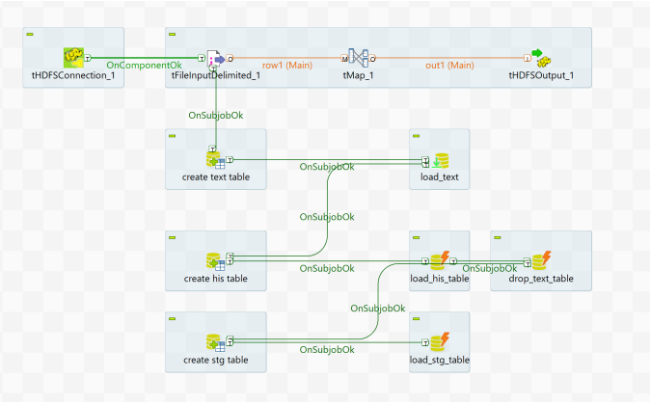
		dari pelanggan.	b. country
6	dim_produt	Berisi seluruh produk yang dijual.	a. product_id (PK) b. product_name c. id_category (FK)
7	dim_category	Berisi seluruh kategori yang dijual.	a. id_category (PK) b. category

A. Proses ETL (Extract, Transform dan Loading)

Pada prose ETL menggunakan Talend sebagai data platform, cloudera sebagai server penyimpanan data dan tableau untuk menampilkan data dalam bentuk report.

- Extract

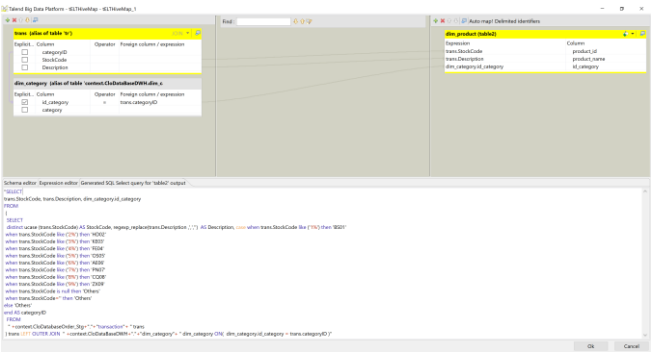
  Proses penarikan data atau extraction data source dari file .csv untuk dibentuk tabel pada database historical dan staging.



Gambar 2. Job create table untuk data historical dan staging

- Trasnform

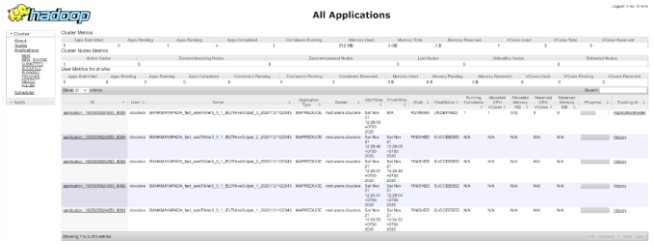
  Proses transformation tabel berdasarkan model perancangan data warehouse, dengan melakukan join antar tabel menggunakan komponen Talend tELTHiveMap



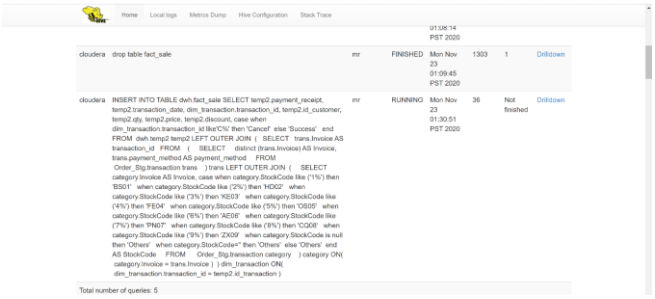
Gambar 3. Proses Transformation tabel

- Loading

  Tabel yang telah dibentuk akan disimpan ke dalam cloudera server, cloudera memiliki beberapa cluster yang berguna untuk developer dalam pembentukan table. Salah satunya adalah cluster untuk melakukan monitoring pembentukan tabel dan query pada server cloudera.



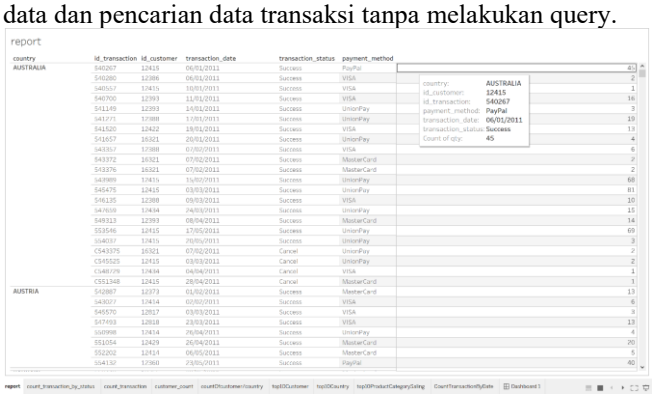
Gambar 4. Cloudera's cluster for monitoring job development



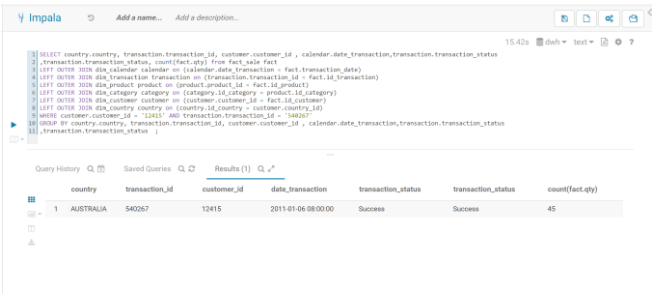
Gambar 5. Cloudera's cluster for query monitoring.

B. Report

Data yang telah dibentuk ke dalam tabel – tabel dimensi dan fakta akan ditampilkan dalam bentuk report menggunakan Tableau, penggunaan Tableau untuk pembentukan report mempermudah untuk menampilkan data dan pencarian data transaksi tanpa melakukan query.



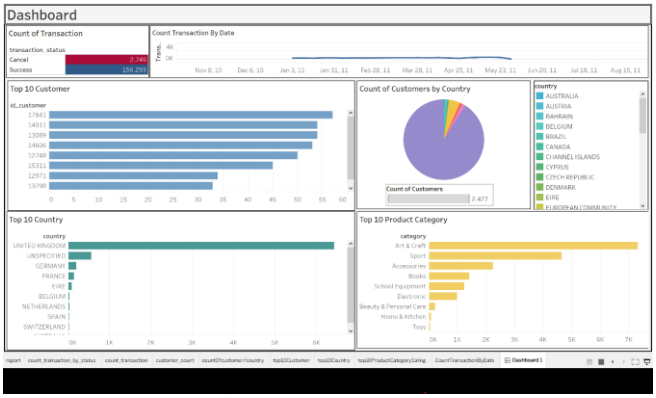
Gambar 6. *Tableau's Report.*



Gambar 7. *Query compare data Tableau report*

C. Dashboard

Pembentukan Dashboard menggunakan Tableau untuk menampilkan berbagai informasi seperti kategori produk yang paling banyak dibeli, negara paling banyak melakukan transaksi dan jumlah transaksi setiap bulannya.



Gambar 8. *Tableau's Dashboard.*

D. Pengujian Sistem

Pengujian sistem dilakukan untuk memastikan sistem berjalan dengan baik atau belum. Pengujian dilakukan pada laptop Razer Stealth Blade 2016 dengan spesifikasi sebagai berikut : (1) Sistem Operasi Windows 10 Pro 64bit; (2) Processor Intel(R) Core(TM) i7-7500U CPU @ 2.70Ghz - 2.90GHz; (3) RAM DDR3 16GB. Pengujian dilakukan dengan cara alpha testing dengan menggunakan metode *black box,* poin – poin pengujian dapat dilihat pada Tabel VII.

TABEL VII. PENGUJIAN MENGGUNAKAN METODE *BLACK BOX.*

No.	Daftar Proses Pengujian	Hasil yang diharapkan	Hasil Uji	Status Pengujian
1.	Test Koneksi pada server cloudera	a. Jika berhasil maka Talend dapat melakukan data processing di server cloudera. b. Jika gagal maka akan muncul pada Talend dan perlu pemeriksaan pada *set-up* koneksi server.	Koneksi Talend dengan server cloudera berhasil.	Berhasil
2.	Job Penarikan *data source* dan pembentukan *data lake*	a. Jika berhasil maka data dibentuk pada database Hive. b. Jika gagal maka akan muncul error pada job talend.	Tabel berhasil dibentuk dan berisi data.	Berhasil
3.	Job pembentukan tabel dimensi dan fakta	a. Jika berhasil maka tabel akan dibentuk dan berisi data. b. Jika gagal maka akan muncul error pada job talend atau data kosong.	Tabel berhasil dibentuk dan berisi data.	Berhasil
4.	Query count row pada data lake dan data warehouse.	a. Jika berhasil maka jumlah data dari data warehouse dan data lake sama. b. Jika gagal maka perlu melakukan pemeriksaan pada job ETL.	Data dari database staging dan database datawarehouse sama.	Berhasil
5.	Data ditampilkan dalam bentuk report dan dashboard pada Tableau.	a. Jika berhasil maka data akan muncul dan dapat dibentuk menjadi *report* dan *dashboard*. b. Jika gagal maka perlu melakukan pemeriksaan koneksi pada *cloudera server*.	Server terkoneksi dan dapat membentuk report dari database data warehouse.	Berhasil
6.	Validasi data pada Tableau dan database data warehouse	a. Jika berhasil maka data yang ditampilkan Tableau dan query database akan sama. b. Jika gagal maka perlu dilakuka pemeriksaan pada koneksi Tableau atau query.	Data dari Tableau dan database datawarehouse sama.	Berhasil

V. PENUTUP

Penelitian ini dilakukan untuk merancang sebuah Data warehouse data transaksi online market dengan menggunakan Cloudera sebagai server untuk penyimpanan data, Talend sebagai tool data platform untuk proses ETL, Tableau sebagai tool untuk menampilkan data dan melakukan analisis dan skema snowflake sebagai model data sangat membantu dalam proses perancangan data warehouse pada penelitian dan hasil dari perancangan data warehouse dapat diaplikasikan pada Tableau yang membuat data – data yang diolah menjadi informasi yang berguna. Adapun saran yang diperlukan adalah dibutuhkannya perangkat dengan spesifikasi lebih untuk mempermudah beban pekerjaan saat proses ETL dan pembentukan tabel.

DAFTAR PUSTAKA

[1] Musadek, Ahmad & Tjahyanto, A. (2008). Desain Data *Warehouse* Pengukur Kinerja Setiap KA Penumpang dengan Distibusi Jumlah Penumpang – Studi Kasus Daop I-IX.

[2] Dahlan, A. (2013). Perancangan Data *Warehouse* Perpustakaan Perguruan Tinggi XYZ Menggunakan Metode *Snowflake Schema*.

[3] O'Driscoll, A., Daugelaite, J., & D.Sleator, R. (2013). *Journal of Biomedical Informatics*.*'Big Data', Hadoop and Cloud Computing in Genomics*, *46*(5), Hal. 774–781.

[4] *Kimball, R., & Margy, R. (2010). The Kimball Group Reader: Relentlessly Practical Tools For Data Warehousing and Business Intelligence. First Edition. Indianapolis: John Wiley & Sons.*

[5] Sellis, Timos, & Miller, Renne J. (2011) ' SIGMOD 2011 : *Proceeding of the 2011 ACM SIGMOD International Conference on Management of data', An analytic data engine for visualization in tableau, Hal.1185 – 1194.*

[6] Supriyatna, A. (2016). Sistem Analis Data Mahasiswa Menggunakan Aplikasi *Online Analytical Processing* (Olap) Data *Warehouse*.

[7] Busiarli, Novia & Hayati, M. (2017). Perancangan dan Implementasi Data *Warehouse* Menggunakan *Snowflake Schema*. Untuk Mengetahuin *Trend* Produksi dan Pemasaran Produk.

[8] Suni, E. K., & Ridwan, W. (2018). 'Jurnal Teknik Informatika', Analisis dan Perancangan Data *Warehouse* untuk Mendukung Keputusan Redaksi Televisi Menggunakan Metode *Nine-Step* Kimball (Studi Kasus pada Redaksi Kompas TV Jakarta), 11(2), Hal. 197.