Simposium Nasional Ilmiah dengan tema: (Peningkatan Kualitas Publikasi Ilmiah melalui Hasil Riset dan Pengabdian kepada Masyarakat), 7 November 2019, hal: 526-532 ISBN: 978-623-90151-7-6 DOI: 10.30998/simponi.v0i0.545

# PENERAPAN DATA LINEAGE UNTUK MEMINIMALISIR KESALAHAN PROSES ETL DARI BASIS-DATA SUMBER KE BASIS-DATA TARGET

Muhammad Firdaus
Universitas Indraprasta PGRI
dasurichi@gmail.com

## ABSTRAK

Di era penggunaan teknologi informasi yang semakin meluas, kebutuhan data menjadi hal yang sangat mendasar dan penting untuk dijaga keabsahannya. Apalagi banyaknya tuntutan dari perusahaan agar data yang di kumpulkan dan diterima dari basis data sumber harus terjaga keasliannya. Penerapan *data lineage* memungkinkan terhindarnya dari kesalahan proses ETL (Extract-Transform-Load) dari Basis-Data sumber ke Basis-Data target ataupun bisa dimanfaatkan sebagai *audit trail*. Pada penelitian ini mencoba mengkaji pemanfaatannya dengan menggunakan metode eksperimen dan kuantitatif, dimana data-data akan dikumpulkan dan di simulasikan menggunakan perangkat lunak khusus, serta observasi lebih dalam hasil pengujiannya, dan mencari relevansinya dengan studi pustaka terkait. Dengan adanya penelitian ini, diharapkan akan dapat memberikan solusi terbaik dan dapat mencari akar permasalahan yang kerap kali terjadi pada data sumber akibat tidak konsistennya data yang di dapat, sehingga dikemudian hari bisa dijadikan pembelajaran dari hasil analisa tersebut.
**Kata kunci:** *data lineage*, basis-data, sumber, target.

## *ABSTRACT*

*In an era of expanding information technology, data needs become very basic and important to keep in legitimacy. Moreover, many demands from the company so that the data collected and received from database source must be kept in authenticity. Data lineage implementation allows to prevent from inevitable error of ETL (Extract-Transform-Load) process from database source into target, or it can be utilized as an audit trail. In this study tried to assess its utilization using experimental and quantitative methods, where data will be collected and simulate using special software, as well as further observation of its testing results, and seeking its relevance from references. With this research, hopefully will be able to provide the best solution and can find the root cause that often occur in the source due to inconsistent data that can be, so that one day it can be used as a lesson learned from result of analysis.*
***Keyword*****:** *data lineage, database, source, target*.

## PENDAHULUAN

Dijaman era revolusi industri ke-4 saat ini, kebutuhan terhadap data semakin meningkat, apalagi dengan hadirnya solusi teknologi baru yang dapat digunakan dengan mudah dan aman, tanpa harus khawatir apakah data yang disimpan dan di olah nantinya aman terhadap pihak-pihak yang tidak berwenang atau mungkin virus baru yang bisa berubah bentuk dan dapat menyebar melalui jaringan. Sejarah tercatat bahwa revolusi industry pertama kali muncul pada tahun 1800 di Inggris, dimana pada waktu itu perekonomian di negara tersebut yang semula agraris berubah menjadi industri. Sektor industri mengalami perkembangan yang luar biasa di era tersebut, dengan adanya penemuan mesin bertenaga uap, peran manusia mulai tergantikan, dan proses produksi meningkat, serta lebih efisien dan efektif (B. Prasetyo & Trisyanti, 2018).

Di era revolusi industri 4.0 inilah penggunaan IoT (Internet of Things) dan IoS (Internet of Service) dapat dimaksimalkan dan dimanfaatkan dengan baik oleh setiap pemangku

kepentingan, baik secara internal maupun antar organisasi, sehingga tercapainya kreasi nilai baru ataupun optimasi nilai yang sudah ada hadir di setiap proses di semua lini industri (H. Prasetyo & Sutopo, 2018). Hal ini juga berdampak pada kebutuhan akan data, dimana saat ini konsep integrasi data dan terpusat menjadi perhatian penting dan perlu ditangani dengan baik. Penggunaan *data warehouse* dan *data mart* memungkinkan integrasi berbagai macam jenis data dari berbagai macam aplikasi atau sistem (Supriyatna, 2016). Disamping itu juga, kebutuhan terhadap *Data Lineage* menjadi perhatian utama kebutuhan organisasi dalam mengungkapkan permasalahan pada data berdasarkan silsilah garis keturunan pada data.

Data Lineage sendiri adalah suatu teknik yang dapat membantu untuk mendokumentasikan perbedaan proses, peraturan bisnis, ketergantungan terhadap atribut lainnya yang dapat menjelaskan dari mana data tersebut berasal, dan bagaimana data tersebut dapat digunakan untuk mengkalkulasikan suatu hasil / informasi yang dapat dimanfaatkan oleh pengguna (Sweet, 2016). Dalam menelusuri suatu data berdasarkan silsilah juga dibutuhkan fungsi metadata yang berperan dalam menjelaskan informasi yang menggambarkan karakteristik data terutama isi, kualitas, kondisi, dan cara perolehannya (Subli, Sugiantoro, & Prayudi, 2017), sehingga hal tersebut menjadi perhatian penting dan dasar pemikiran utama bagi organisasi, maupun badan usaha dalam menjaga keaslian data. Permasalahan yang biasa terjadi dikarenakan perpindahan data dari sumber ke target yang dituju seringkali menimbulkan konflik, akibat dari kurangnya informasi yang terdapat di metadata sumber maupun di target. Ketika membangun koneksi perpindahan data tersebut menggunakan perangkat lunak integrasi data seperti SSIS (SQL Server Integration Services), IBM Infosphere DataStage, Talend Studio, dan lain-lainnya kerap kali kesulitan dalam menelusuri petunjuk dari mana data tersebut berasal dan bagaimana caranya untuk mengetahui adanya permasalahan di setiap data maupun tipe data yang digunakan, walaupun sudah di verifikasi melalui jalur ETL (Extract – Transform – Load) yang terhubung.

Proses ETL yang melibatkan lebih dari satu data sumber, pada umumnya memiliki beberapa permasalahan, diantaranya: perlunya waktu eksekusi yang cukup lama, penggunaan memori yang besar, serta seringnya terjadi perubahan struktur data, sehingga hal ini berdampak pada eksekusi *package* atau *job* berulang-ulang. *Package* ETL itu sendiri berisi hasil *query* pemetaan terhadap beberapa *table* atau basis data ataupun diambil dari hasil *import* dari file external lainnya, seperti flat file atau pun file excel, yang nanti nya akan diubah menjadi bentuk *table* dan di petakan atau di pindahkan ke target tujuan sesuai dengan kebutuhan pengguna.

## METODE

Metode penelitian yang digunakan dalam mengkaji fungsi dan manfaat pada data lineage terhadap proses pengembangan ETL ini adalah eksperimen dan kuantitatif, dimana data-data akan dikumpulkan dan di simulasikan menggunakan perangkat lunak SSIS versi 2013 dan data lineage tools “MANTA” (perangkat lunak ini bisa digunakan tanpa perlu di

*install* di *desktop* PC atau Laptop, serta dapat di akses secara langsung ke alamat url: https://getmanta.com/live-manta/), serta di observasi lebih dalam hasil pengujiannya, dan mencari relevansinya dengan studi pustaka terkait. Pada penelitian ini ada beberapa tahap yang akan dilakukan, diantaranya sebagai berikut ini : 1. Membangun *query* pembanding yang diperlukan untuk menganalisa setiap perubahan yang terjadi pada pemetaan data (data mapping), sehingga akan mempermudah melihat hasil simulasi sebelum diterapkan ke dalam pengembangan ETL, 2. Analisa hasil pemetaan yang telah dikeluarkan hasilnya oleh perangkat lunak "MANTA", 3. Kaji permasalahan dari hasil Analisa yang telah dilakukan sebelumnya untuk mendapatkan akar permasalahan (root cause) yang nantinya akan berguna dalam melihat dampaknya terhadap bisnis proses perusahaan atau organisasi.

## HASIL

Pada kegiatan penelitian ini dilakukan simulasi penyelesaian masalah dengan menggunakan perangkat lunak data lineage "MANTA" dengan rincian detail tahapan pelaksanaan sebagai berikut :

### 1. Membangun *Query* Pembanding

Pada tahapan ini dibuat *query* pembanding dengan memetakan kolom-kolom pada table yang tersedia di data sumber (dalam hal ini di *Data Warehouse*) dengan kolom-kolom yang ada pada data target (Datamart), sebelum nantinya dibangun package ETL berdasarkan hasil pemetaan sebelumnya, seperti pada gambar dibawah ini.

```
select
    Current_timestamp as Tgl_data,
    MP.NIK AS NIK,
    MP.Nama AS Nama_Karyawan,
    DP.Jabatan AS Jabatan,
    DP.Department AS Department,
    DP.Atasan_Langsung Melapor_ke,
    NP.Performance_Appraisal AS Nilai_PA,
    DP.Gaji AS Gaji,
    (case when DP.Gaji < 5000000 then (DP.Gaji*NP.Performance_Appraisal)
          when DP.Gaji > 5000000 then (DP.Gaji*NP.Performance_Appraisal)
    else NULL
    end) AS Bonus,
    ((NP.Performance_Appraisal * DP.Gaji)/DP.Gaji) AS Persen_Bonus
from dbo.Master_Pegawai MP
LEFT JOIN Detail_Pegawai DP on MP.NIK = DP.NIK
LEFT JOIN Nilai_Pegawai NP on MP.NIK = NP.NIK
```

**Sumber**: pribadi

Gambar 1. *Query* Pembanding

*Query* diatas menghasilkan tabel pembanding yang sama dengan tabel laporan yang terdapat di datamart dengan bentuk sebagai berikut :

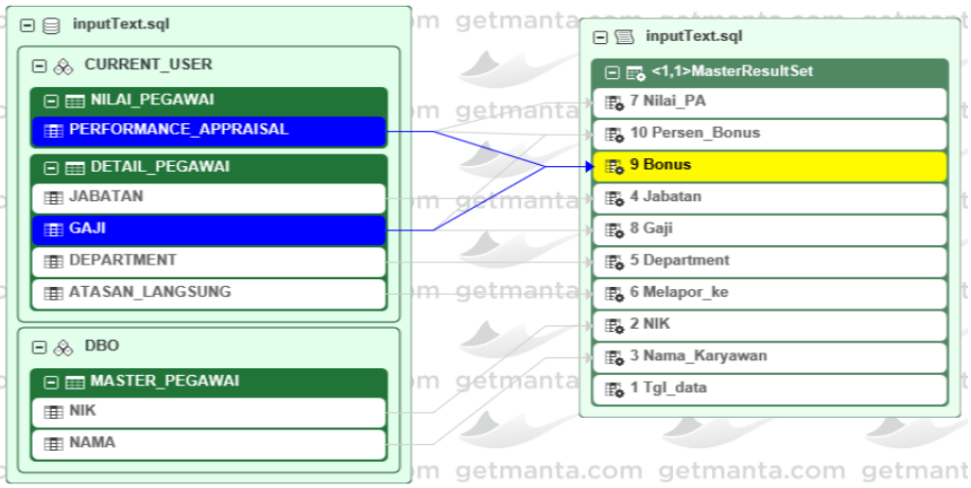
Tabel 1. Hasil *Query* Pembanding

| | Tgl_data | NIK | Nama_Karyawan | Jabatan | Department | Melapor_ke | Nilai_... | Gaji | Bonus | Persen_Bonus |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2019-11-07 02:59:01.863 | 11900... | Geri | Senior Consult... | PSAK | Mario | 5 | 50000... | NULL | 5.000000 |
| 2 | 2019-11-07 02:59:01.863 | 11900... | Rudi | Consultant | RR | Waluyo | 4 | 70000... | 280000... | 4.000000 |
| 3 | 2019-11-07 02:59:01.863 | 11900... | Budi | IT Support | Data Manageme... | Rahmat | 3 | 90000... | 270000... | 3.000000 |

**Sumber data**: pribadi

## 2. Analisa Hasil Pemetaan

Apabila hasil *query* diatas menunjukkan hasil error, secara otomatis memang perangkat lunak SQL Server dapat memberikan informasi dimana letak error tersebut. Namun misalkan kita ingin mengetahui kenapa pada *field* (kolom) "Bonus", kenapa karyawan atas nama Geri bernilai "NULL", kadang kala untuk query yg rumit terlalu susah untuk di telusuri sumber datanya, sedangkan apabila kita petakan ke dalam perangkat lunak MANTA (seperti pada gambar dibawah ini), kita bisa lebih mudah mengetahui dari table dan kolom mana saja perhitungan bonus karyawan bisa di dapat, tanpa perlu mengetahui isi pada table tersebut.

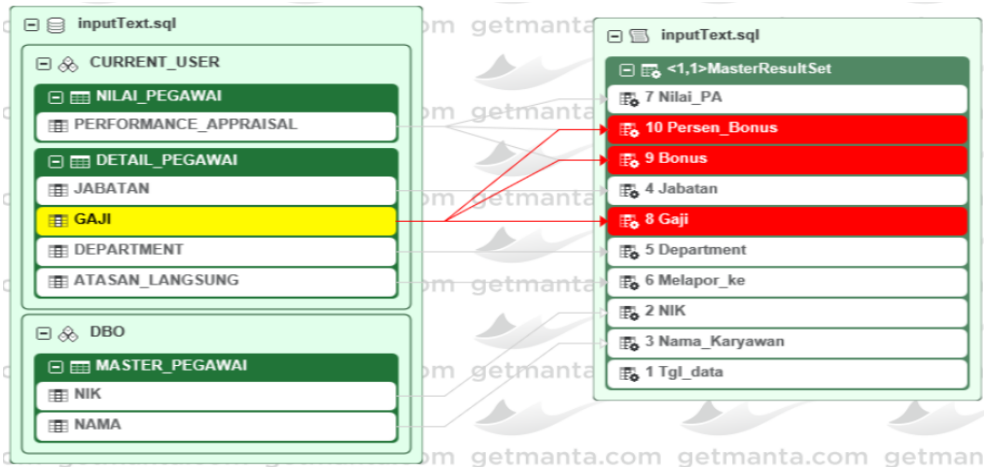



**Sumber data**: https://getmanta.com/live-manta/#demo

Gambar 2. Pemetaan hasil query dengan MANTA

Sebelumnya, untuk memproses pemetaan hasil *query* pada Gambar 1 menjadi table pemetaan pada Gambar 2 juga tidak memerlukan detail isi pada masing-masing table dan *field* tersebut, tetapi hanya memerlukan *query* pembanding yang sudah lengkap dengan penggabungan beberapa table dan *field* tersebut.

Dari perangkat lunak MANTA juga bisa kita dapatkan penelusuran jejak dari komponen *field* "GAJI" ini dipecah dan digunakan untuk beberapa *field* apa sajakah di table target di datamart. Seperti pada Gambar 3 dibawah ini.

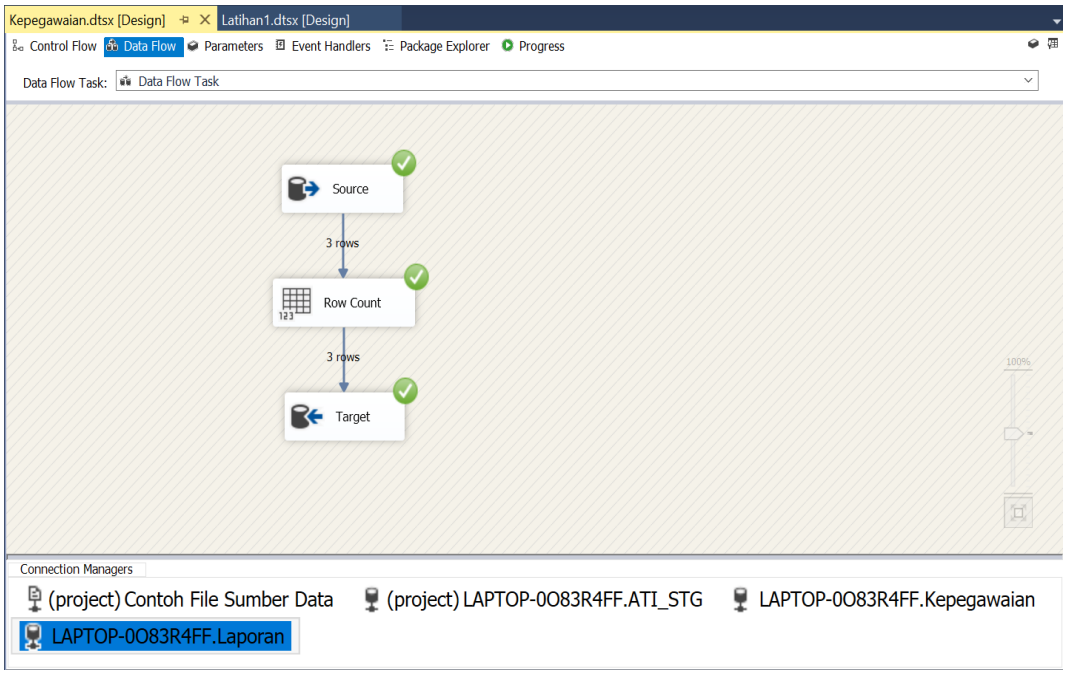



**Sumber data**: https://getmanta.com/live-manta/#demo

Gambar 3. Pemetaan Komponen *Field* "Gaji"

## 3. Kaji Permasalahan Hasil Analisa

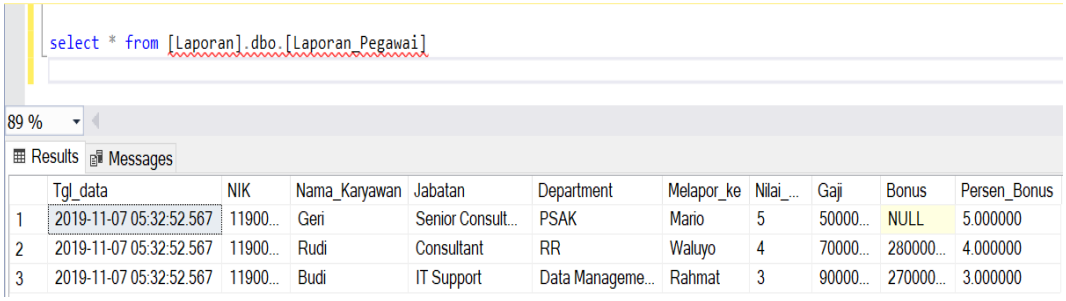
Dari hasil Analisa diatas kita bisa telusuri sumbernya, sehingga bisa dijadikan *audit trail* untuk mengetahui akar permasalahannya, sehingga akan lebih mudah untuk dilakukan *troubleshooting* maupun hanya sekedar untuk menunjukkan kepada pihak pengguna bisnis mengenai sumber data yang menjadi topik pembicaraan pada saat membahas rekam jejak data. Ketika *query* pembanding tersebut diterapkan di perangkat lunak SSIS, seperti pada gambar dibawah ini.




**Sumber data**: pribadi

Gambar 4. Implementasi SSIS

Sehingga apabila di cek dan verifikasi dari data target di basis data Laporan, untuk table Laporan_Pegawai, maka hasilnya akan sama dengan yang terdapat pada sumber data yang terdapat di basis data Kepegawaian, seperti pada gambar dibawah ini.



```
select * from [Laporan].dbo.[Laporan_Pegawai]
```

89 %

Results | Messages

| | Tgl_data | NIK | Nama_Karyawan | Jabatan | Department | Melapor_ke | Nilai_... | Gaji | Bonus | Persen_Bonus |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2019-11-07 05:32:52.567 | 11900... | Geri | Senior Consult... | PSAK | Mario | 5 | 50000... | NULL | 5.000000 |
| 2 | 2019-11-07 05:32:52.567 | 11900... | Rudi | Consultant | RR | Waluyo | 4 | 70000... | 280000... | 4.000000 |
| 3 | 2019-11-07 05:32:52.567 | 11900... | Budi | IT Support | Data Manageme... | Rahmat | 3 | 90000... | 270000... | 3.000000 |

**Sumber data**: pribadi

Gambar 5. Cek Hasil Query table Laporan Pegawai

## SIMPULAN

Dari hasil penelitian tersebut memberikan gambaran bahwa perangkat lunak MANTA ini sangatlah membantu pengembang dalam memberikan jawaban atas permasalahan yang sering dihadapi pada saat menelusuri rekam jejak data dari sumber data ke target data, sehingga bukti rekam jejak tersebut bisa dijadikan dasar investigasi lebih lanjut pada proses audit berikutnya. Kesalahan *input* data ataupun kesalahan yang kerap kali terjadi pada saat menggabungkan 2 atau lebih table data dapat di hindarkan, dan kejadian tersebut tidak akan terulang kembali di pengembangan ETL berikutnya.

Diharapkan nantinya penelitian-penelitian mengenai data *lineage* akan terus berlanjut dengan studi kasus yang berbeda dan lebih menantang, dan diharapkan bisa bermanfaat bagi perusahaan, organisasi, maupun masyarakat luas.

## UCAPAN TERIMAKASIH

Dalam kesempatan ini saya ucapkan terima kasih yang sedalam-dalamnya kepada Direktorat Riset dan Pengabdian Kepada Masyarakat, Istri dan anakku, serta ayah dan ibuku yang telah mendukung selama proses penyusunan jurnal saya sehingga dapat diselesaikan dengan baik, walaupun saya sadar bahwa kesalahan datang dari saya dan semua kesempurnaan ini hanya milik Allah semata.

## DAFTAR RUJUKAN

Prasetyo, B., & Trisyanti, U. (2018). Revolusi Industri 4.0 Dan Tantangan Perubahan Sosial. *IPTEK Journal of Proceedings Series*, *0*(5), 22–27. https://doi.org/10.12962/j23546026.y2018i5.4417

Prasetyo, H., & Sutopo, W. (2018). INDUSTRI 4.0: TELAAH KLASIFIKASI ASPEK DAN ARAH PEREMBANGAN RISET. *Jurnal Teknik Industri*, *13*(4), 372. https://doi.org/10.2307/1782970

Subli, M., Sugiantoro, B., & Prayudi, Y. (2017). METADATA FORENSIK UNTUK MENDUKUNG PROSES INVESTIGASI DIGITAL. *Data Manajemen Dan Teknologi Informasi (DASI)*, *18*(1), 44–50. Retrieved from http://www.ghbook.ir/index.php?name=فرهنگ و زبانه های

فوق&option=com_dbook&task=readonline&book_id=13650&page=73&chkhashk=ED9C9491B4&Itemid=218&lang=fa&tmpl=component

Supriyatna, A. (2016). Sistem Analisis Data Mahasiswa Menggunakan Aplikasi Online Analytical Processing (Olap) Data Warehouse. *None*, *12*(1), 62–71.

Sweet, E. (2016). Data Lineage and Compliance. *ISACA*, *5*, 1–4.