Laporan Temuan Permasalahan Data Lake House pada Perusahaan E-Commerce XYZ (Nama Disamarkan)

Harvian Dito Syahputra, Samuel Yuma Krismata, Hafiz Akmaldi Santosa, Nur Azka Rahardiansyah, Naufan Zaki Luqmanulhakim

1. Pendahuluan

Data Lakehouse telah menjadi pendekatan inovatif dalam pengelolaan data modern, menggabungkan fleksibilitas penyimpanan data tak terstruktur dari data lake dengan kemampuan manajemen terstruktur dari data warehouse. Pendekatan ini dirancang untuk menangani berbagai jenis data baik yang terstruktur, semi-terstruktur, maupun tidak terstruktur dan mendukung analisis yang lebih efisien serta pengambilan keputusan berbasis data. Namun, seperti sistem manajemen data lainnya, implementasi Data Lakehouse di organisasi seringkali menghadapi berbagai tantangan yang dapat mempengaruhi efektivitas dan efisiensinya.

Salah satu permasalahan utama yang sering dihadapi adalah kinerja arsitektur dan proses Extract, Transform, Load (ETL) yang lambat. Proses ETL yang tidak optimal dapat memperlambat integrasi dan pemrosesan data, menghambat penyajian informasi, dan mempengaruhi pengalaman pengguna secara keseluruhan. Selain itu, kualitas data yang buruk—seperti duplikasi data dan pembersihan yang tidak standar dapat mengurangi validitas serta keandalan analisis dan model machine learning yang bergantung pada data tersebut. Monitoring dan tuning performa aktivitas business intelligence (BI) juga menjadi tantangan, terutama pada saat beban kerja tinggi (peak time), di mana query yang tidak efisien dapat membebani memori sistem dan memperpanjang waktu respons.

Dalam konteks implementasi Data Lakehouse di perusahaan, evaluasi maturity level menggunakan Capability Maturity Model Integration (CMMI) menunjukkan adanya variasi tingkat kematangan manajemen data. Beberapa aktivitas penting, seperti *Define and Maintain the DW-BI Architecture*, *Process Data for Business Intelligence*, serta *Monitor and Tune BI Activity and Performance*, masih berada pada Level 2 (Managed). Sementara itu, aktivitas lainnya, seperti *Implement BI Tools and User Interfaces* dan *Monitor and Tune Data Warehousing Processes*, telah mencapai Level 3 (Defined). Adanya *gap* ini menunjukkan perlunya peningkatan manajemen data secara menyeluruh untuk mencapai target Level 3 pada semua aktivitas, sesuai dengan ekspektasi perusahaan.

Laporan ini bertujuan untuk mengidentifikasi permasalahan utama yang dihadapi oleh sistem Data Lakehouse, menyusun kerangka solusi yang komprehensif, dan memberikan rekomendasi strategis untuk mengoptimalkan manajemen data. Analisis dilakukan berdasarkan data operasional perusahaan, wawancara dengan para pemangku kepentingan, serta evaluasi penggunaan alat dan teknologi yang mendukung Data Lakehouse. Hasilnya diharapkan dapat membantu perusahaan mencapai sistem pengelolaan data yang lebih efisien, andal, dan mendukung pengambilan keputusan berbasis data secara real-time.

2. Identifikasi Masalah

2.1. Arsitektur dan Proses ETL (Extract, Transform, Load) yang Lambat
Arsitektur dan proses ETL (Extract, Transform, Load) yang lambat dapat mempengaruhi sistem secara keseluruhan. Akibatnya, integrasi dan pemrosesan data menjadi terhambat. Hal ini tentunya mempengaruhi pengalaman pengguna.

2.2. Monitoring dan Tuning Performa Business Intelligence Activity
Query yang dilakukan dapat menjadi buruk dalam waktu-waktu tertentu terutama pada *peak time*. Perlu dilakukan efisiensi dalam hal query sehingga tidak memakan beban memory yang terlalu besar pada saat jam-jam tertentu terutama pada *peak time*.

2.3. Kualitas Data yang Buruk
Pembersihan dan transformasi data yang buruk dapat mempengaruhi validitas dan reliabilitas data di dalam Data Lakehouse. Proses Cleansing yang buruk dan tidak terstandar dapat menyebabkan data keluaran menjadi tidak baik yang dapat mempengaruhi performa serta keakuratan machine learning yang akan digunakan selanjutnya serta mengurangi manfaat dari segi bisnis dari Data Lakehouse tersebut.

3. Data Pendukung

1. Data Operasional Perusahaan:
   a. Jenis Data:
      Data pelanggan, transaksi, produk, dan aktivitas bisnis lainnya.
   b. Volume Data:
      Perusahaan menghadapi pertumbuhan data yang signifikan akibat transaksi harian. Data ini berasal dari beberapa sumber internal dan eksternal.
   c. Masalah:
      Duplikasi data dan kualitas data rendah menjadi kendala besar.
      Proses penyajian laporan membutuhkan waktu lama (hingga 1 bulan).

2. Tingkat Kematangan Sistem DW-BI:
Evaluasi menggunakan Capability Maturity Model Integration (CMMI) menunjukkan bahwa manajemen data di perusahaan berada di tingkat kematangan yang bervariasi:
   a. Tiga aktivitas berada pada Level 2 (Managed):
      - Define and Maintain the DW-BI Architecture

- Process Data for Business Intelligence
- Monitor and Tune BI Activity and Performance

b. Empat aktivitas berada pada Level 3 (Defined):
- Understand Business Intelligence Information Needs
- Implement Data Warehouses and Data Marts
- Implement BI Tools and User Interfaces
- Monitor and Tune Data Warehousing Processes

c. Gap Analisis:
Target perusahaan adalah mencapai Level 3 untuk semua aktivitas, sehingga masih ada gap pada tiga aktivitas Level 2.

3. Masalah Spesifik Berdasarkan Wawancara:
   a. Define and Maintain the DW-BI Architecture:
   Metadata tidak tersedia secara lengkap, menyebabkan kurangnya transparansi dan kesulitan dalam pelacakan lineage data.
   b. Process Data for BI:
   Kegiatan cleansing data kurang optimal karena kurangnya koordinasi antara data engineer (eksekusi teknis) dan data analyst (pemahaman bisnis).
   c. Monitor and Tune BI Activity and Performance:
   Waktu query yang tidak konsisten, terutama saat workload tinggi, berdampak pada pengalaman pengguna BI tools.

4. Alat dan Teknologi yang Digunakan:
   a. Business Intelligence Tools:
   Tableau, Google BigQuery, dan Google Data Studio digunakan untuk analisis dan visualisasi data.
   b. Sistem ETL Tradisional:
   Mengelola integrasi data dari berbagai sumber ke dalam data warehouse.
   c. Kendala Teknologi:
   Performa tools seperti BigQuery cenderung menurun saat beban tinggi.
   Tidak ada tools khusus untuk mendukung metadata dan data lineage.

4. Kerangka Solusi

4.1. Pengembangan Arsitektur *Data Lakehouse*

Mengintegrasikan fitur dari *data warehouse* dan *data lake* untuk membuat sistem manajemen data yang lebih efisien. Arsitektur ini harus mendukung *storage* dan pemrosesan data dengan data yang *structured, semi-structured,* dan *unstructured* secara bersamaan.

4.2. Peningkatan *Metadata Extraction*

Menerapkan mekanisme pengelolaan metadata yang mendalam dengan *semantic annotation, data indexing,* dan *data lineage*. Hal ini diperlukan untuk meningkatkan kemampuan analisis serta pengelolaan data secara menyeluruh

4.3. Pengurangan *Data Swamp*

Menggunakan metode transformasi data yang tepat untuk melakukan *data cleaning*, meringkas, serta mempersiapkan data sebelum digunakan. Ini juga termasuk mekanisme deteksi penyimpangan (*deviation detection)* untuk data baru.

4.4. Ingest dan Transformasi Data

Membuat pipeline ingestion yang terintegrasi dengan menggunakan tools seperti Apache Kafka. Pipeline ini memungkinkan pengelolaan data secara menyeluruh - baik data *batch*, *near real-time* maupun *streaming* - dari berbagai sumber data secara efisien.

4.5. Penerapan *Real-Time Analytics* dan *Machine Learning*

Menggunakan algoritma machine learning untuk menganalisis pola dalam data yang besar dan beragam, meningkatkan pengambilan keputusan berbasis data melalui analisis prediktif, serta melakukan *real-time analytics* dengan bantuan *tools* seperti Apache Kafka untuk melakukan *query real-time*.