**Data Engineer Final Project**

## Project Background

Salah satu hal terpenting dalam peminjaman bank adalah memastikan debitur mampu membayar kembali pinjaman tersebut. Kebutuhan untuk menentukan kemungkinan pembayaran menggunakan perhitungan kompleks akan kemudian disingkatkan dalam satu variabel yang disebut sebagai Nilai Kredit atau “Credit Score”. Nilai Kredit digunakan agar memudahkan penentuan kemungkinan pembayaran dimana Nilai Kredit yang lebih tinggi menunjukkan kemungkinan pembayaran yang lebih tinggi. Kuantifikasi kemungkinan pembayaran juga dapat digunakan untuk menentukan titik pembatasan yang dapat diterima, maupun menggunakan variabel tersebut untuk perhitungan langkah-langkah berikutnya seperti jika Nilai Kredit yang baik akan dapat ditawarkan lebih banyak pinjaman meskipun dengan bunga yang lebih kompetitif.

## Problem Statement

Kecepatan adalah salah satu faktor yang semakin diprioritaskan pelanggan jaman ini. Respon yang cepat dapat meningkatkan kesenangan pelanggan sementara respon yang pelan akan dapat menyebabkan pelanggan untuk tidak melanjutkan permintaan mereka. Perhitungan yang dibutuhkan untuk nilai kredit tidak dapat dilakukan secara cepat menggunakan kalkulator. Oleh sebab itu, dibutuhkan suatu sistem baru yang dapat melakukan perhitungan nilai kredit secara cepat, efektif, dan konsisten.

## Data Platform Design

Prioritas pertama dalam data platform design yang sesuai dengan Problem Statement adalah menciptakan model dengan cepat. Oleh sebab itu, arsitektur yang dipilih berupa Bottom-Up Architecture (Kimball) karena hal ini dapat dilaksanakan dengan cepat dan fleksibel. Data akan diterima melalui OLTP dan perhitungan nilai kredit dilakukan melalui OLAP yang kemudian akan dikirim menggunakan Streaming untuk mendapatkan hasil secara cepat.

Medallion Architecture akan digunakan untuk memastikan data diolah bertahap dan dikelola secara terstruktur dan efisien. Data akan pertama diterima dalam bentuk Excel sesuai dengan format yang sebelumnya ditentukan agar dapat diproses lebih cepat kedalam data mart yang tersedia sebagai lapisan Bronze. Data tersebut kemudian akan diproses oleh Airflow melalui task guna menghasilkan nilai kredit yang akan dipakai kedalam lapisan Silver**.(Tolong konfirmasi apa masih diproses oleh Airflow untuk scheduling task/job atau pakai Kafka saja)** Data tersebut kemudian akan digabungkan dengan data-data lain untuk menghasilkan data mart Gold yang dapat digunakan oleh RM secara mudah dan cepat.

## Data Understanding

Sumber data dari Excel berdasarkan pengisian baik dari RM atau pelanggan sendiri. Data ini berstruktur Excel dengan header sebagai baris pertama agar memudahkan bagi RM. Kualitas data berdasarkan pemasukkan secara manual sehingga ada kemungkinan untuk human error.

## Transformation

Data akan diproses menggunakan ETL/ELT sebelum di batching/streaming. Spark jobs akan dilakukan untuk memindahkan data ke table baru serta melakukan Analisa Nilai Kredit.

## Data Modeling

Menggunakan Medallion Architecture, ada 3 lapis data yang digunakan:

**Bronze** - Data “mentah” yang didapatkan dari penerimaan Excel berisi parameter yang ditentukan untuk perhitungan serta ID untuk identifikasi. Data bersifat Normalized.

Create external table **Mart\_Loan\_Offer** (

Customer\_ID bigint,

Loan\_ID bigint,

##(Param musti ada Kuantitatif & Kualitatif)

Param\_1 varchar(255)

Param\_2 int

Param\_3

Param\_4

) Partitioned\_by (Request\_Date date)

**Silver** – Data yang sudah diproses sehingga dapat digunakan untuk melakukan Analisa menggunakan table-table yang tersedia. Data mulai bersifat Denormalized karena akan mulai digunakan untuk analisa.

Create external table **Mart\_Credit\_Score** (

Customer\_ID bigint,

Loan\_ID bigint,

Credit\_Score decimal(6,3),

Request\_Date date

) Partitioned\_by (Date\_Process date)

Create external table **Mart\_Customer** (

Customer\_ID bigint,

First\_Name varchar(50),

Last\_Name varchar(50),

Pendapatan\_Per\_Bulan decimal(28,3)

) Partitioned\_by (date\_process date)

Insert into Mart\_Credit\_Score

Select

L.Customer\_ID,

L.Loan\_ID,

C.Full\_name,

Credit\_Score,

Pendapatan\_Per\_bulan,

Request\_Date,

C.Date\_Process

FROM

(

Select

Customer\_ID,

Loan\_ID,

Credit\_Score,

Request\_Date,

Date\_Process

FROM MART\_CREDIT\_SCORE

Where date\_process = current\_date()

)A

LEFT JOIN

(

Select

Customer\_ID,

Concat\_ws(first\_name, last\_name) as full\_name,

Pendapatan\_per\_bulan

FROM MART\_CUSTOMER

WHERE date\_process = current\_date()

)B

ON A.Customer\_ID = B.Customer\_ID

AND A.date\_process = B.date\_process

**Gold** – Datamart yang sudah dirapihkan sehingga dapat digunakan secara langsung oleh user. Data bersifat denormalized agar memudahkan penggunaan oleh user.

Create external table **Customer\_Credit\_Score\_Result** (

Customer\_ID bigint,

Loan\_ID bigint,

Full\_Name varchar(255),

Amount\_Permintaan decimal(28,3),

Pendapatan\_Per\_Bulan decimal(28,3),

Credit\_Score decimal(6,3),

Request\_Date date

) Partitioned\_by (Date\_Process date)

Data yang digunakan memakai Snowflake Schema untuk Bronze, dan Star Schema mulai dari Silver sampai Gold.

UAT dengan user (RM) dapat dilakukan untuk mendapatkan feedback bila ada field yang ingin ditambahkan seperti Usia, dll.

Machine Learning – Analisa Nilai Kredit menggunakan (Regression/Classification)

## Kesimpulan

Nilai kredit dapat menentukan variabel yang dapat digunakan untuk mempermudah persetujuan pinjaman, akan tetapi kebaikan variabel tersebut tergantung pada parameter yang dimasukkan. Jika ada parameter yang dimasukkan secara tidak benar, mungkin karena ketidakpengertian parameter yang dimaksud, nilai kredit mungkin tidak seakurat yang diinginkan. Selain itu, nilai kredit tidak memfaktorkan hal-hal seperti kondisi makro sektor bisnis baik secara ekonomis maupun politis.

Hal-hal ini diluar cakupan nilai kredit tetapi berdampak besar kepada kemungkinan pembayaran sehingga RM tidak boleh hanya menggunakan nilai kredit sebagai satu-satunya pacuan untuk menyetujui pinjaman.