Data and Business Understanding

Canrakerta & Sigit Hariyanto

PJJ DA Akt.IV - September 2022

KEMENTERIAN KEUANGAN
REPUBLIK INDONESIA

KEMENTERIAN KEUANGAN
REPUBLIK INDONESIA

Kompetensi Dasar Materi Data and Business Understanding – PJJ DA 2022

01.

Menguraikan hubungan tujuan organisasi dan proses bisnis dengan identifikasi masalah

02.

Menguraikan konsep ETL (Extract-Transform-Load)

03.

Merinci sumber-sumber data yang dapat digunakan

04.

Melakukan ekstrak data dari sumbernya menggunakan Jupyter Notebook

05.

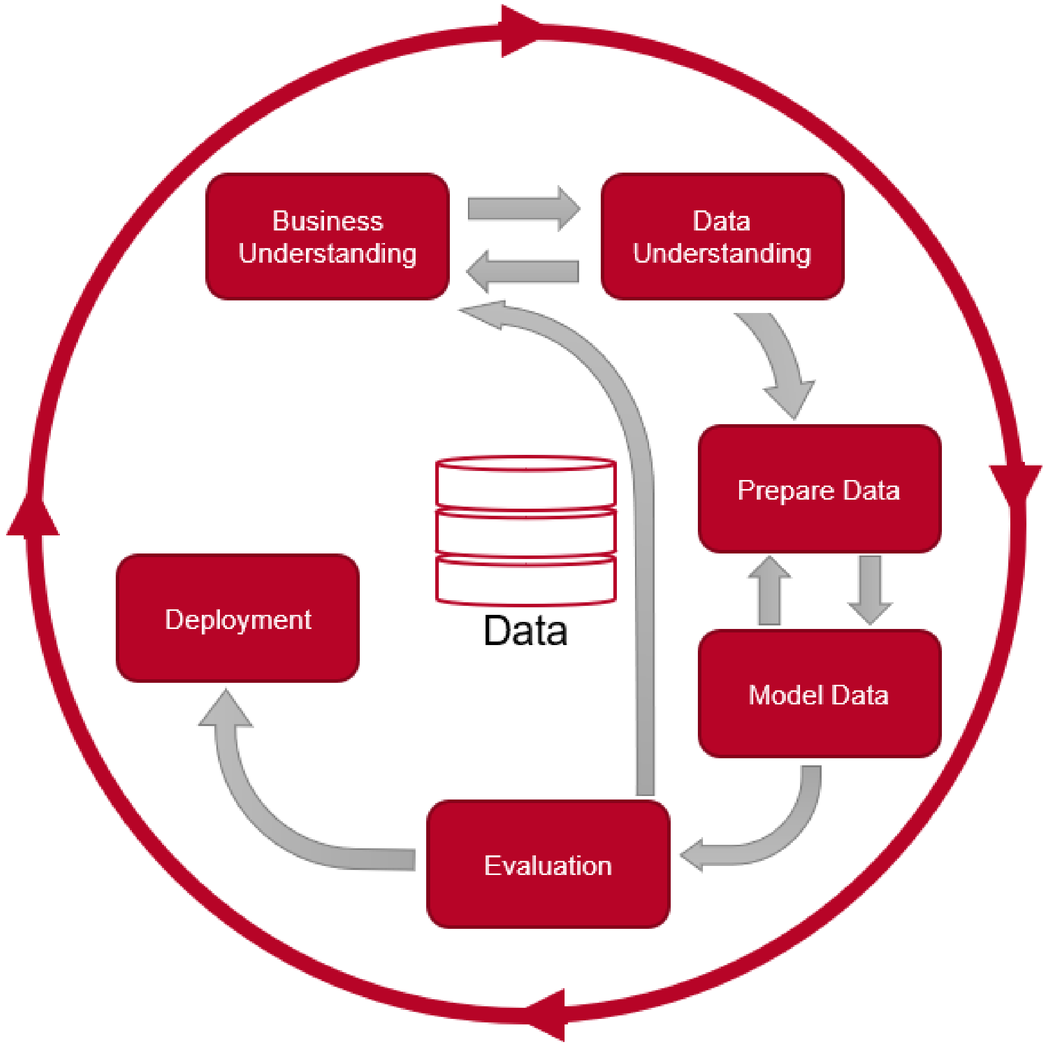
Melakukan transform data sesuai format dengan Jupyter Notebook

06.

Melakukan load kepada target data dalam Jupyter Notebook

KEMENTERIAN KEUANGAN
REPUBLIK INDONESIA

Proses Pengembangan Data Analitik Berdasarkan Standar Proses CRISP-DM

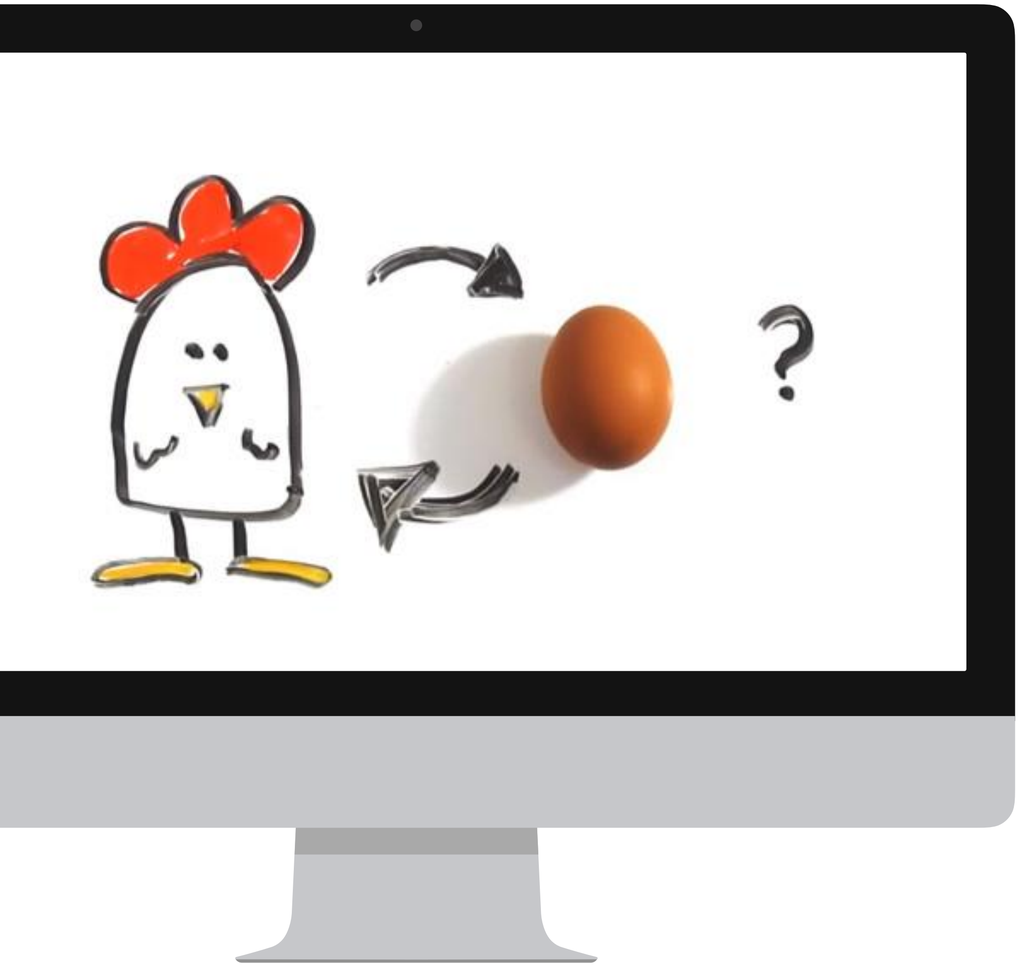


THE FAMOUS CRISP-DM

- **Business Understanding**
- **Data Understanding**
- Data Preparation
- Modeling
- Evaluation
- Deployment

Perhatikan Iterasinya ☺

KEMENTERIAN KEUANGAN
REPUBLIK INDONESIA



Big Question

Mana yang lebih dahulu, memahami problem bisnis atau pemenuhan atas kebutuhan data dahulu yang harus dipenuhi?

KEMENTERIAN KEUANGAN
REPUBLIK INDONESIA

Mendefinisikan Sebuah Masalah pada Proses Bisnis adalah Masalah Tersendiri?

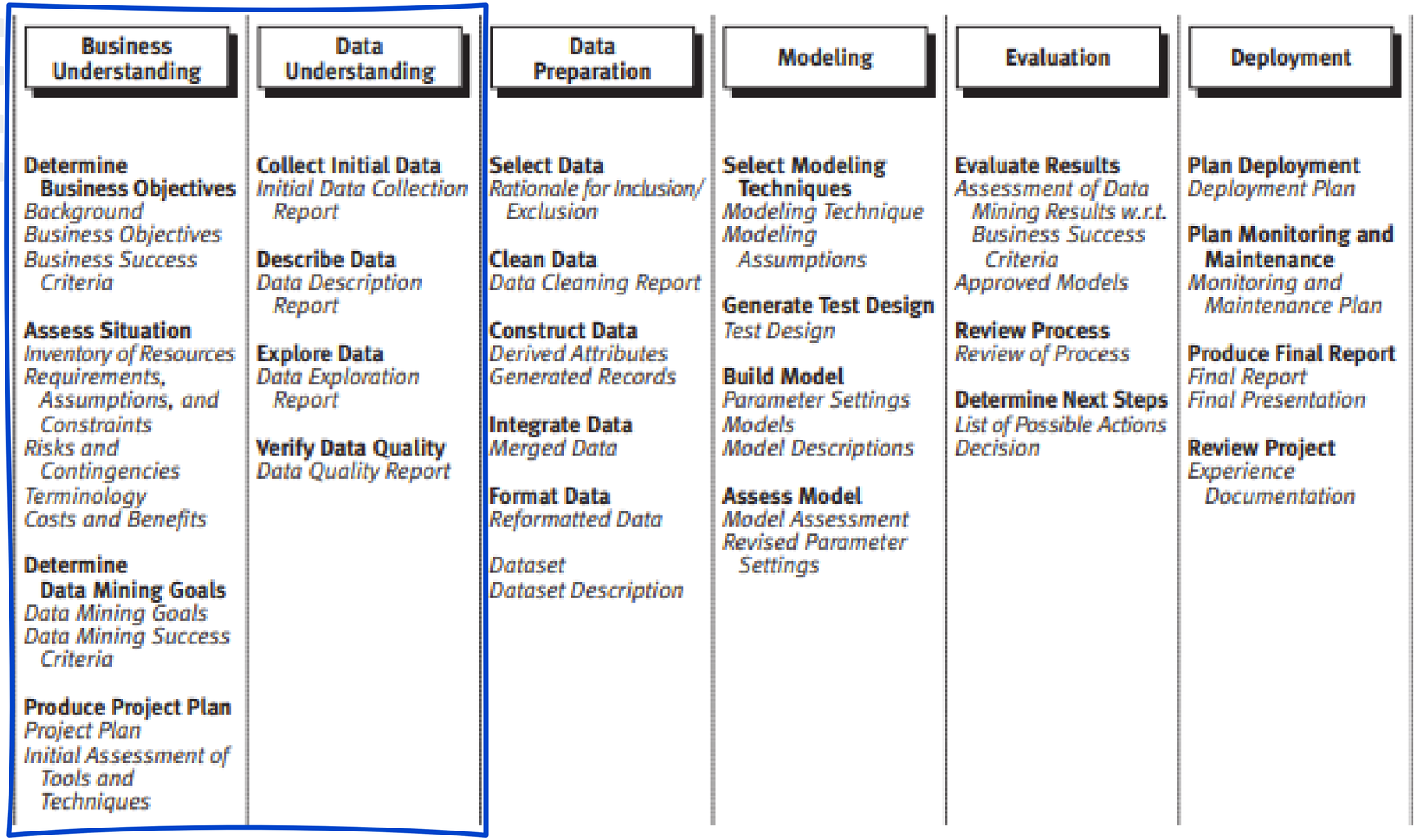
Quotation

"Identifikasi masalah yang tidak baik merupakan penyebab kegagalan proyek data analitik yang paling banyak terjadi"

- Big Data/Analytics Project Failure: A Literature Review (Reggio & Astesiano, 2020)

- Berikan **perhatian yang cukup** untuk **mendefinisikan tujuan, objektif, dan *problem statement*** yang dihadapi dalam memberikan layanan
- Lakukan **pendalaman terhadap problem** yang dihadapi untuk **menentukan akar masalah**
- Lakukan **studi literatur** terhadap *use cases* yang sama, baik melalui **proyek yang pernah ada** sebelumnya maupun melalui ***paper***

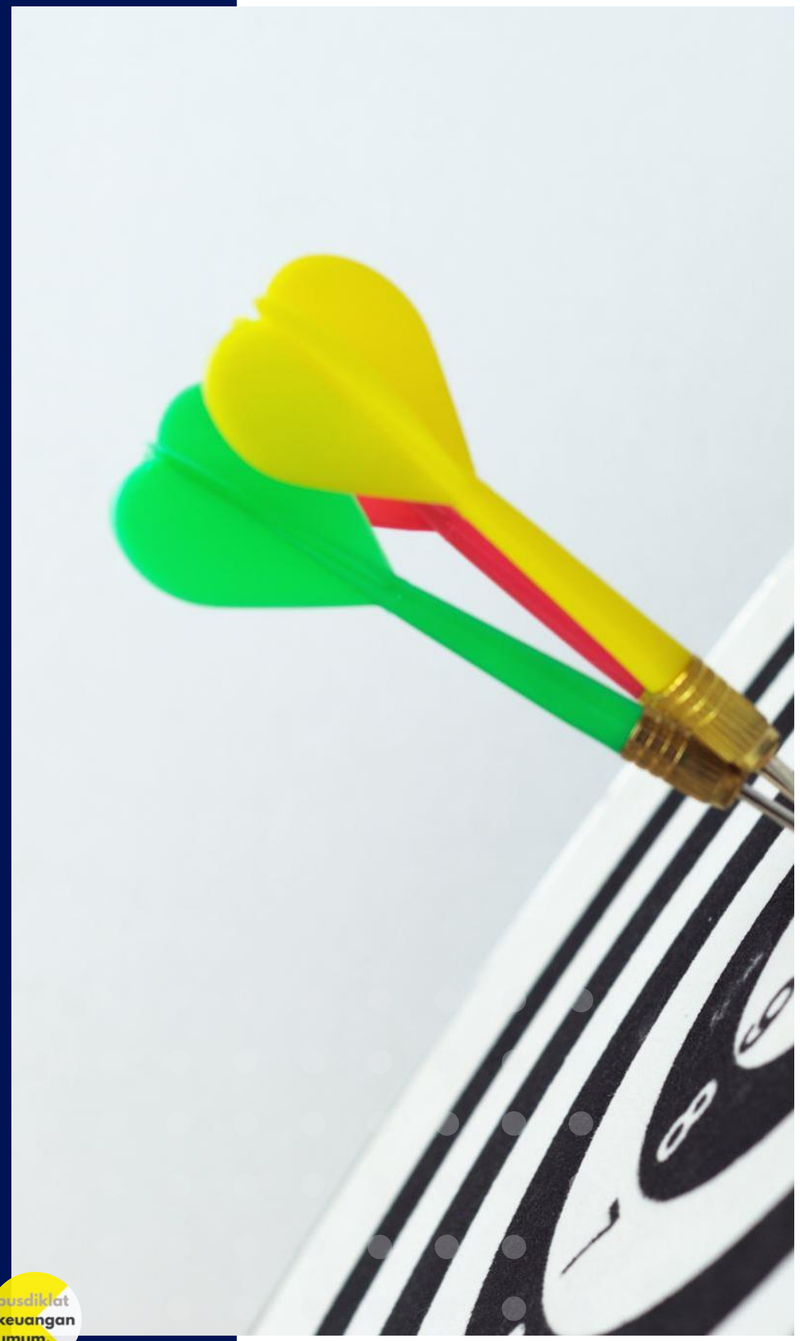
Petunjuk Proses Data and Business Understanding pada CRISP-DM



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

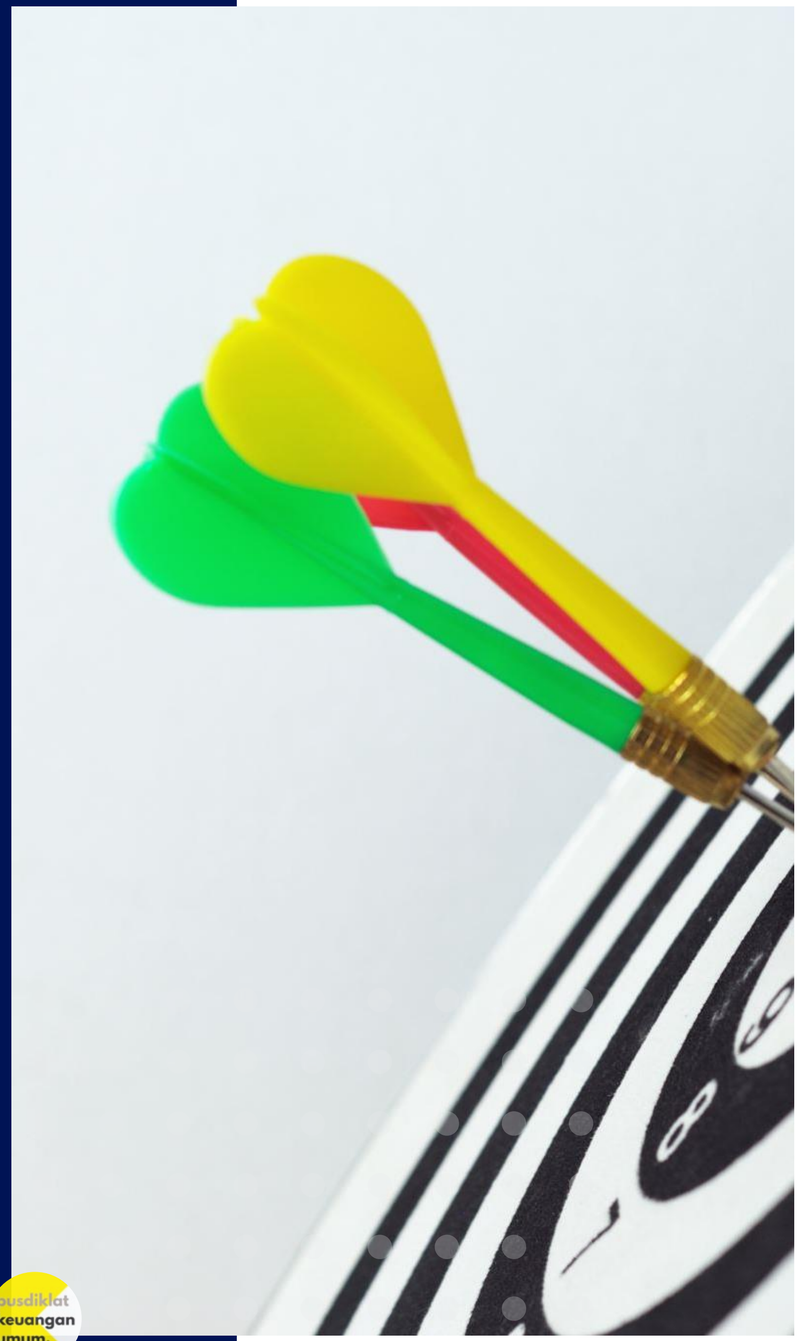
CRISP-DM 1.0 – Chapman et al. (2000)



Determine Business Objective

Business Objective

- Menjelaskan **sasaran utama** berdasarkan perspektif bisnis.
- Biasanya diturunkan dari **visi**, **misi**, **tujuan organisasi**, **sasaran strategis** dan turunannya
- Contoh:
  - Meningkatkan *tax ratio*;
  - Meningkatkan penyelesaikan piutang negara;
  - Menyediakan informasi kas yang valid;
  - Efektivitas penyaluran bantuan sosial;
  - Mengembangkan pengetahuan pegawai;
  - Meningkatkan peringkat indeks EoDB.



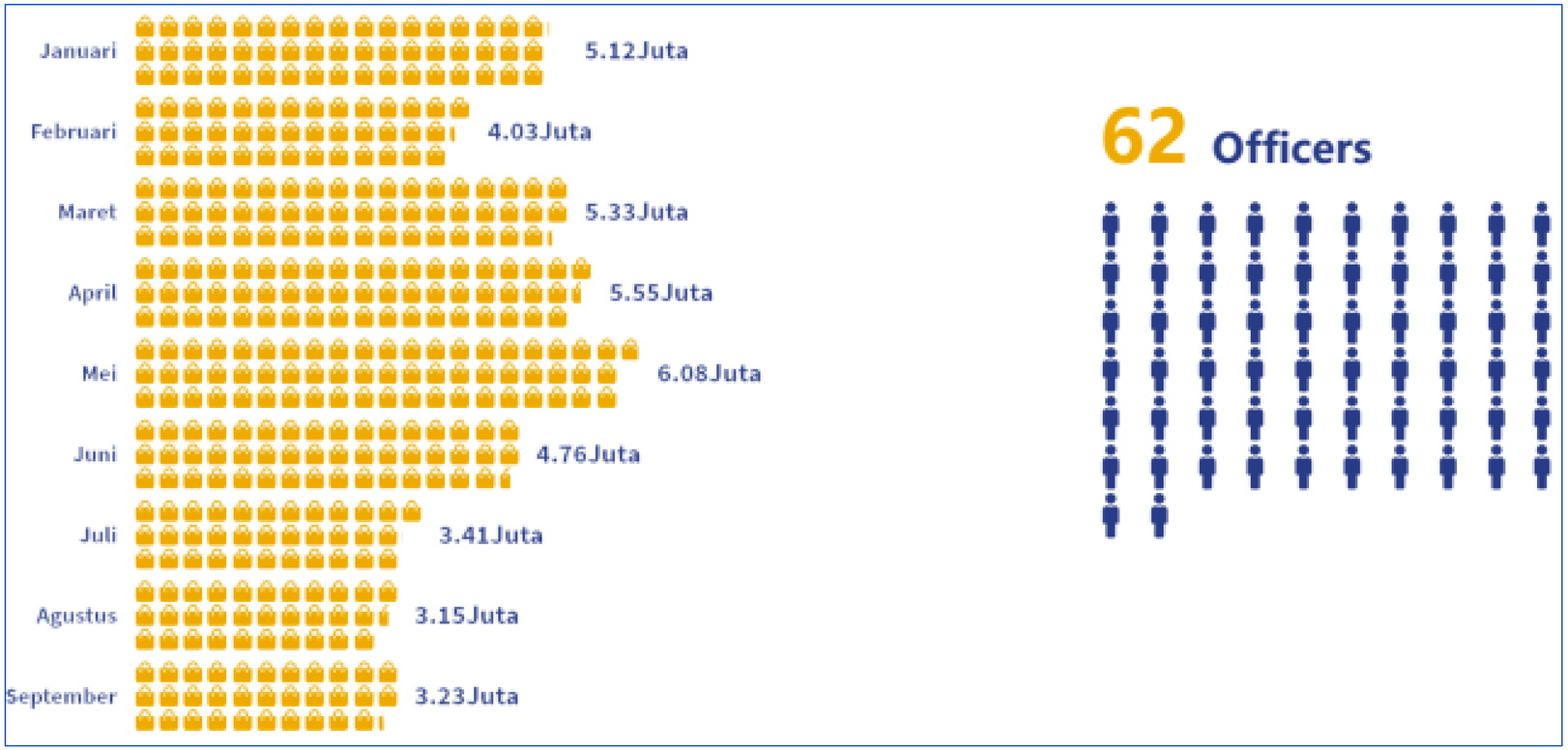
Determine Business Objective (2)

Business Success Criteria

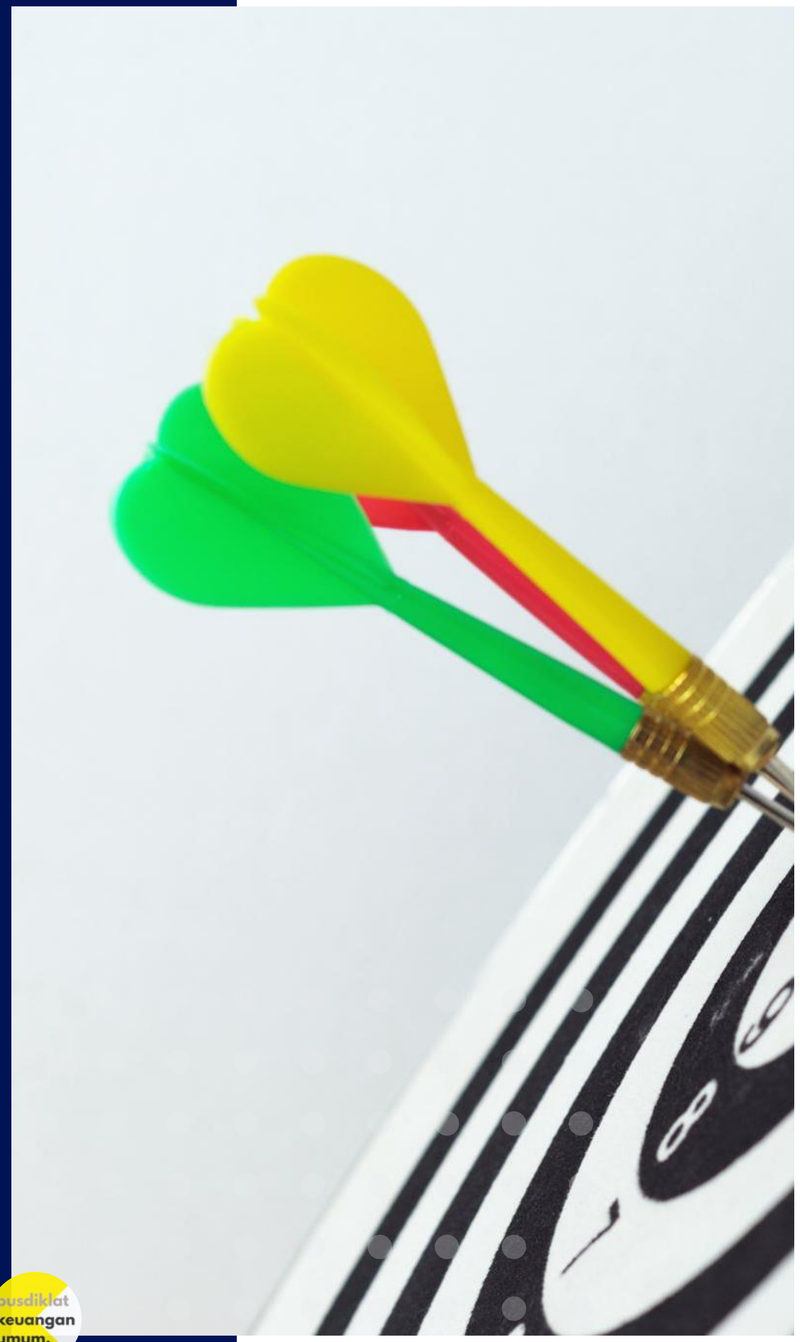
- Menjelaskan **kriteria** sukses atau tidaknya proyek berdasarkan perspektif bisnis.
- Kriteria harus memenuhi pendekatan ***Specific***, ***Measurable***, ***Achievable***, ***Relevant***, **dan** ***Time-bound*** **(SMART)**
- Contoh:
  - Menemukan 3 faktor utama penyebab penyaluran bansos tidak efektif;
  - Meningkatkan hit rate pemeriksaan barang hingga 20%
  - Menurunkan waktu proses *clearance* di pelabuhan menjadi 1 hari
  - Meningkatkan persentase penyelesaian piutang negara hingga 70% di tahun 2023

Use Case: Pelayanan Barang Kiriman di Soekarno-Hatta

Bagaimana memberikan pelayanan yang efektif dan efisien?



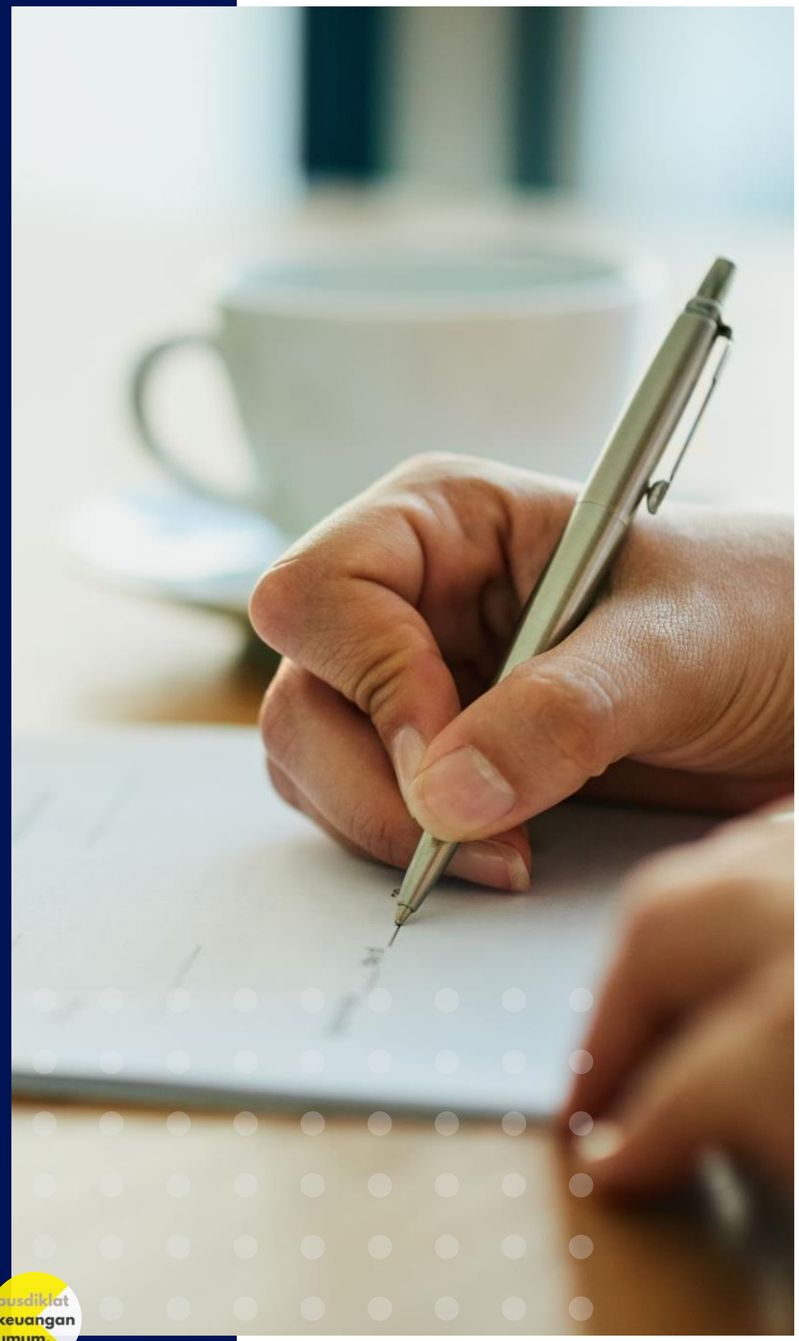
Berdasarkan **data jumlah paket barang kiriman per bulan** dan **pegawai yang tersedia**, bagaimana caranya kantor Soekarno-Hatta dapat **menentukan paket** yang harus **diperiksa secara fisik** atau **tidak diperiksa**? Dengan harapan, pelayanan yang efektif dan efisien dapat **meningkatkan kepatuhan perusahan** pengiriman paket dan **penambahan penerimaan** bagi negara.



Determine Business Objective (3)

Template dan Contoh output

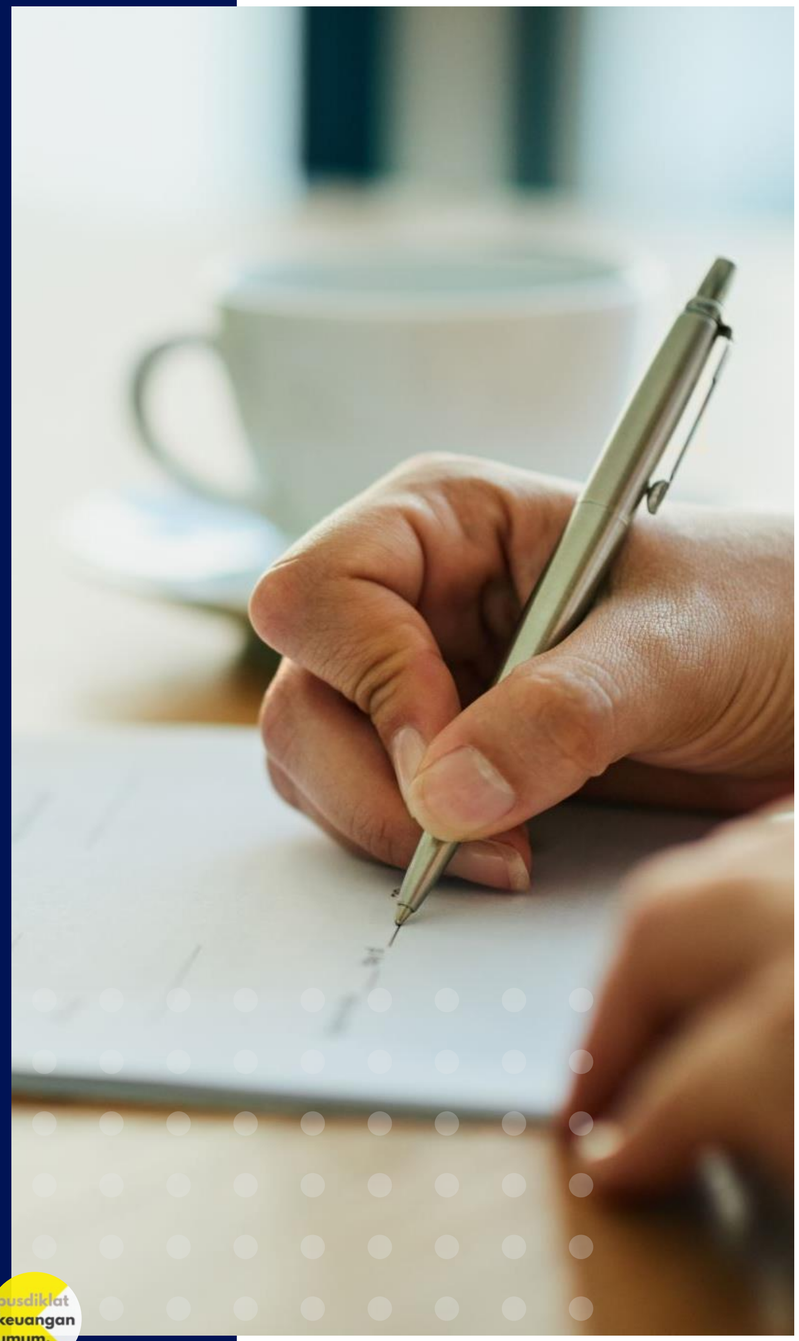
BU.1. Determine Business Objectives	
This task depicts what the customer really wants to accomplish from a business view.	
Outputs	
Background	Record the known information about the business situation.
	• Jumlah petugas tidak sebanding dengan jumlah Dokumen • Masih terdapat dokumen yang dilakukan pemeriksaan, namun tidak terdapat temuan apapun • Masih ada ketidakpatuhan perusahaan dalam menyampaikan dokumen
Business Objectives	Describe the customer's primary business objectives.
	• Kepatuhan perusahaan meningkat • Penambahan Penerimaan
Business Success Criteria	From a business point of view, describe the criteria for a successful or useful outcome to the project. This should be specific enough to be measured objectively.
	• Persentase temuan sebanyak 20% dari total dokumen yang harus diperiksa • Meningkatkan penerimaan bea masuk sebesar 10%



Assess Situation

Inventory of Resources

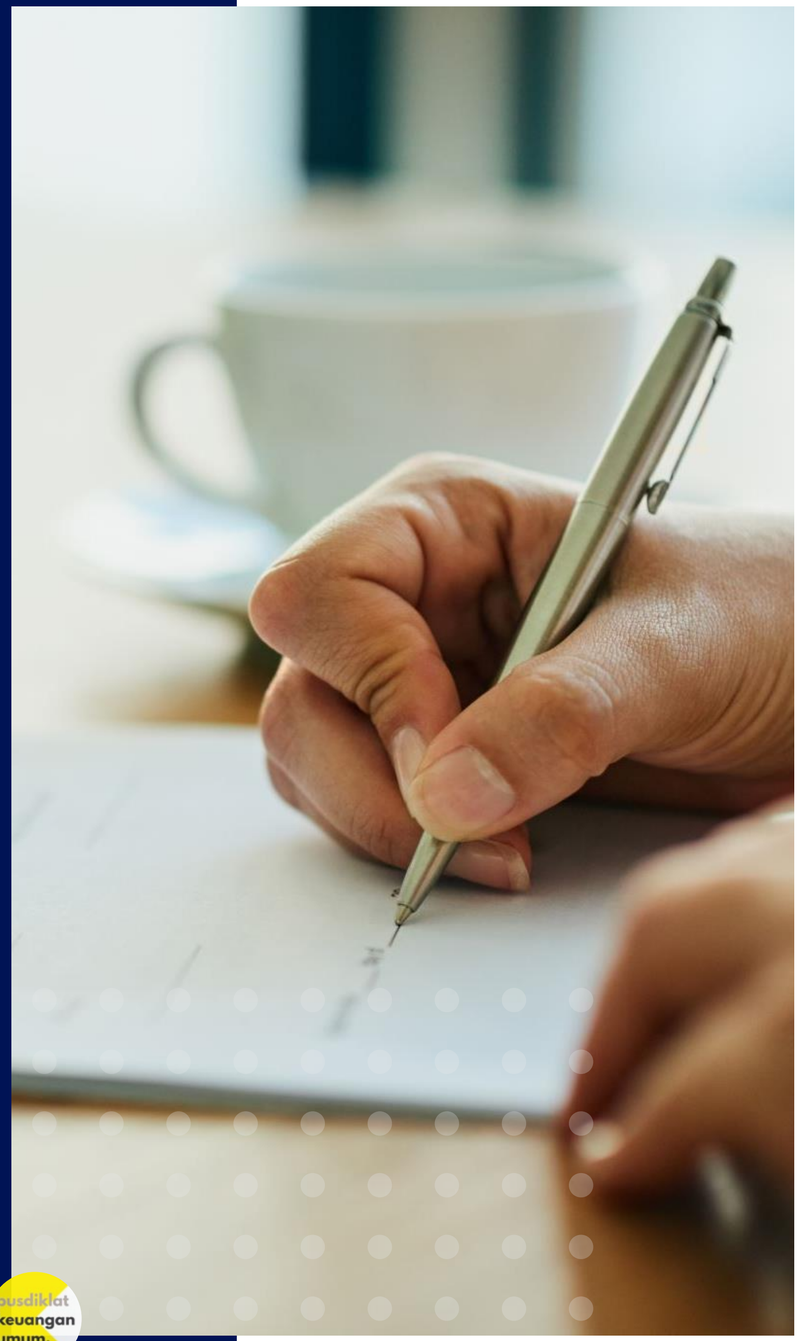
- Melakukan **inventarisir terhadap sumber daya** yang dibutuhkan:
  - **SDM**: Mendata ketersediaan dan kemampuan Project Manager, Data Engineer, Data Scientist, Data Analyst, dan Domain Expert
  - **Data**: Apakah data dapat diakses untuk digunakan? Apakah data sudah tersedia di data warehouse atau masih transaksional? Berapa volume data yang dibutuhkan?
  - ***Computing resources***: Apakah tersedia perangkat keras untuk mengolah data?
  - ***Software***: Tools apa saja yang diperlukan?



Assess Situation (2)

Requirements, Assumptions, and Constraints

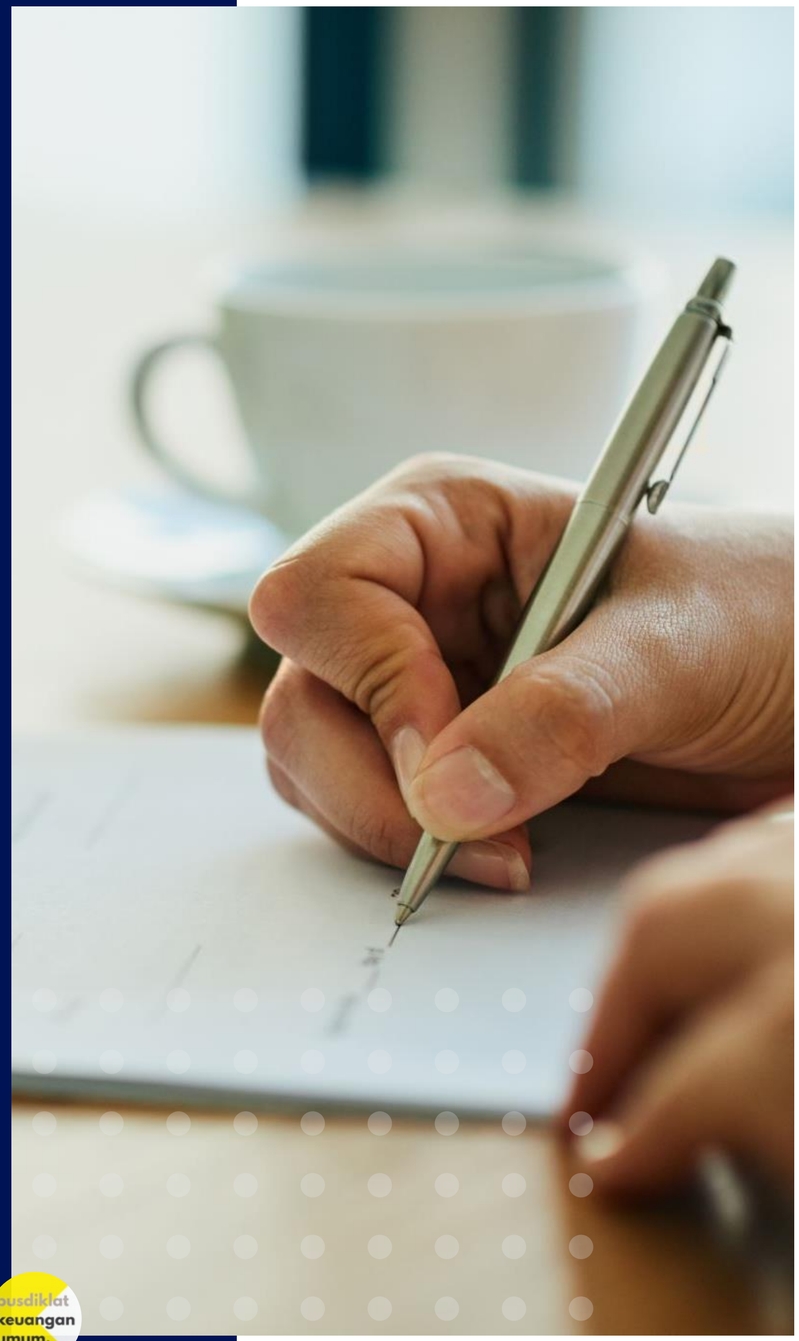
- Melakukan **inventarisir terhadap kebutuhan proyek**:
  - Kebutuhan jadwal penyelesaian proyek
  - Kebutuhan akan kualitas hasil yang diharapkan
  - Kebutuhan akan keamanan data
  - Kebutuhan legalitas penggunaan data
- Melakukan **inventarisir terhadap asumsi dan batasan proyek**:
  - Batasan ketersediaan sumber daya termasuk penggunaan teknologi, kapasitas media penyimpanan, kemampuan server, dan kecepatan *processor*
  - Batasan waktu
  - Batasan data yang digunakan



Assess Situation (3)

Cost and Benefit

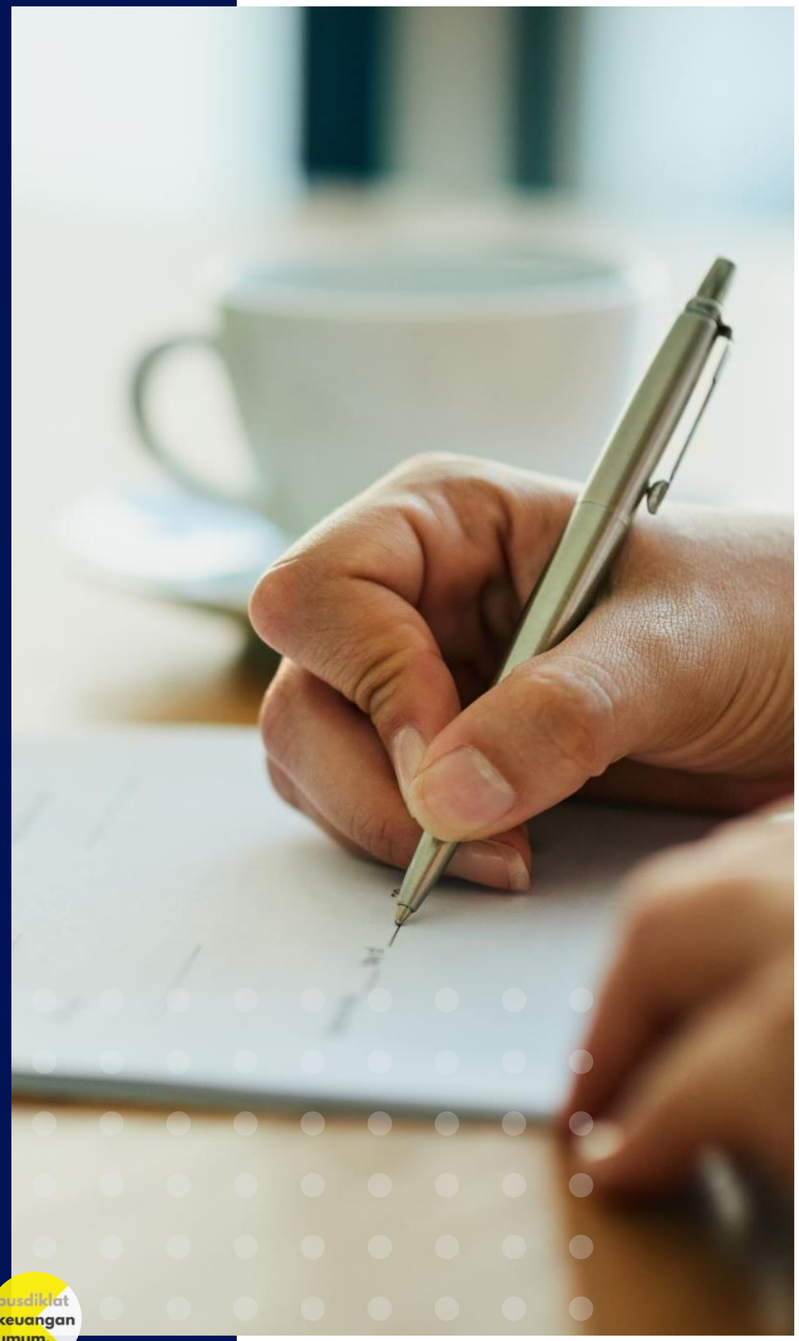
- Melakukan **penilaian terhadap *cost and benefit* proyek**:
  - Membuat *cost benefit analysis* sebelum memulai proyek
  - Pastikan perhitungan dilakukan secara spesifik
  - Bandingkan *cost* yang dibutuhkan dengan *benefit* yang akan didapatkan



Assess Situation (4)

Template dan contoh output

BU.2. Assess Situation	
This task involves more detailed investigation of the resources, constraints, assumptions, and other factors that affect data analysis goal and project plan.	
Outputs	
Inventory of resources	List of available resources such as personnel, data, computing resources and software.
Team	Project Manager: Arik Sutiawan (PT) Data Scientist: Fajar Hidayat dan M Abdul Basit (FT) Data Engineer: Toto Andriyanto (PT)
Data	Data Dokumen Barang Kiriman periode 2021 Data Hasil Pemeriksaan Dokumen Barang Kiriman periode 2021 Data Barang pada Marketplace Hijau Data Barang pada Marketplace Orange
Hardware	2 Unit PC i7 RAM 16Gb
Software	Python versi 3.8 Jupiter Notebook



Assess Situation (4)

Template dan contoh output (cont.)

Requirements, assumptions and constraints	List of project requirements, such as completion schedule, quality of results, security and legal issues. Make sure that you can use the data.
Requirements	Deadline Project Desember 2022 Solusi yang dihasilkan tidak mengganggu layanan yang sudah ada Solusi yang dihasilkan tiidak menambah dwelling time
Assumptions	Tidak ada perubahan peraturan selama proyek
Constraints	Data yang digunakan hanya periode 2021 Hanya 2 perusahaan yang akan dijadikan piloting
Risks and contingencies	List the risks or events that might delay the project or cause it to fail. Plans and actions will be taken if these risks take place.
Risks	• Bertambahnya pengajuan keberatan atas hasil pemeriksaan oleh perusahaan • Pemblokiran akses marketplace terhadap *crawling* data yang dilakukan
Contigencies	-
Cost and Benefit	Cost: 20 Cloud Server x USD 60 x Kurs 15.000 = 18 Juta/bulan Benefit: Menambah penerimaan Bea Masuk sekitar 2 Miliar/bulan

Determining Data Mining Goals

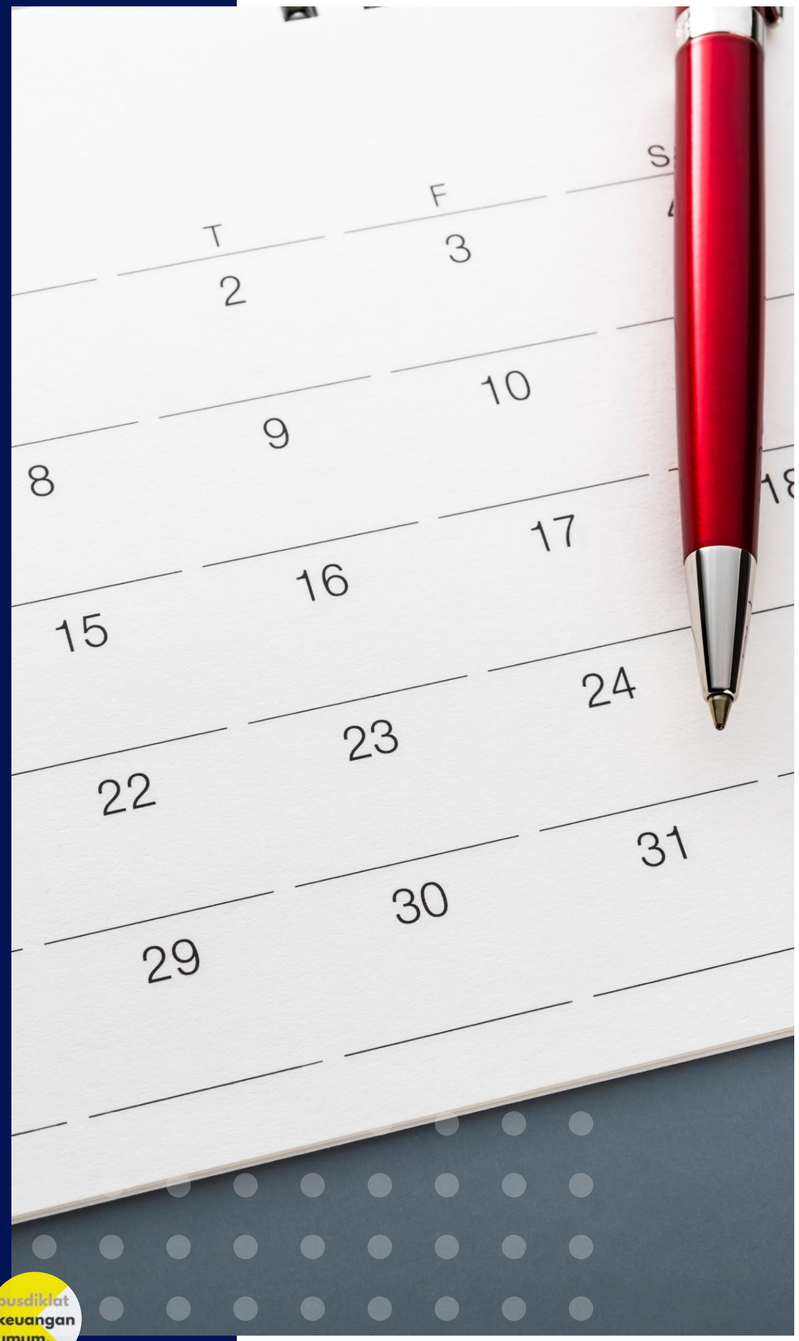
Data Mining Goals and Data Mining Success Criteria

- **Menentukan pendekatan data analitik** yang dibutuhkan untuk menyelesaikan masalah dalam perspektif bisnis:
  - Menerjemahkan *business objective* ke dalam pendekatan data analitik yang dibutuhkan
  - Menjelaskan proses dan output dari data analitik yang dikembangkan
  - Contoh:
    - Pendekatan klasifikasi digunakan untuk memprediksi perusahan yang akan melakukan fraud
    - Pendekatan regresi digunakan untuk menghitung perkiraan harga rumah
    - Menyusun kebutuhan informasi yang dibutuhkan sebagai bahan visualisasi
- **Menentukan kriteria sukses** terhadap pendekatan data analitik yang dilakukan:
  - Menggunakan alat ukur seperti akurasi dari algoritma yang digunakan sebagai kriteria sukses
  - Contoh:
    - Pendekatan klasifikasi yang digunakan harus memenuhi tingkat akurasi minimal 70% dan sensitifitas minimal 90%
    - Visualisasi yang dihadirkan menjawab seluruh kebutuhan informasi yang dibutuhkan

Determining Data Mining Goals (2)

Template dan contoh output

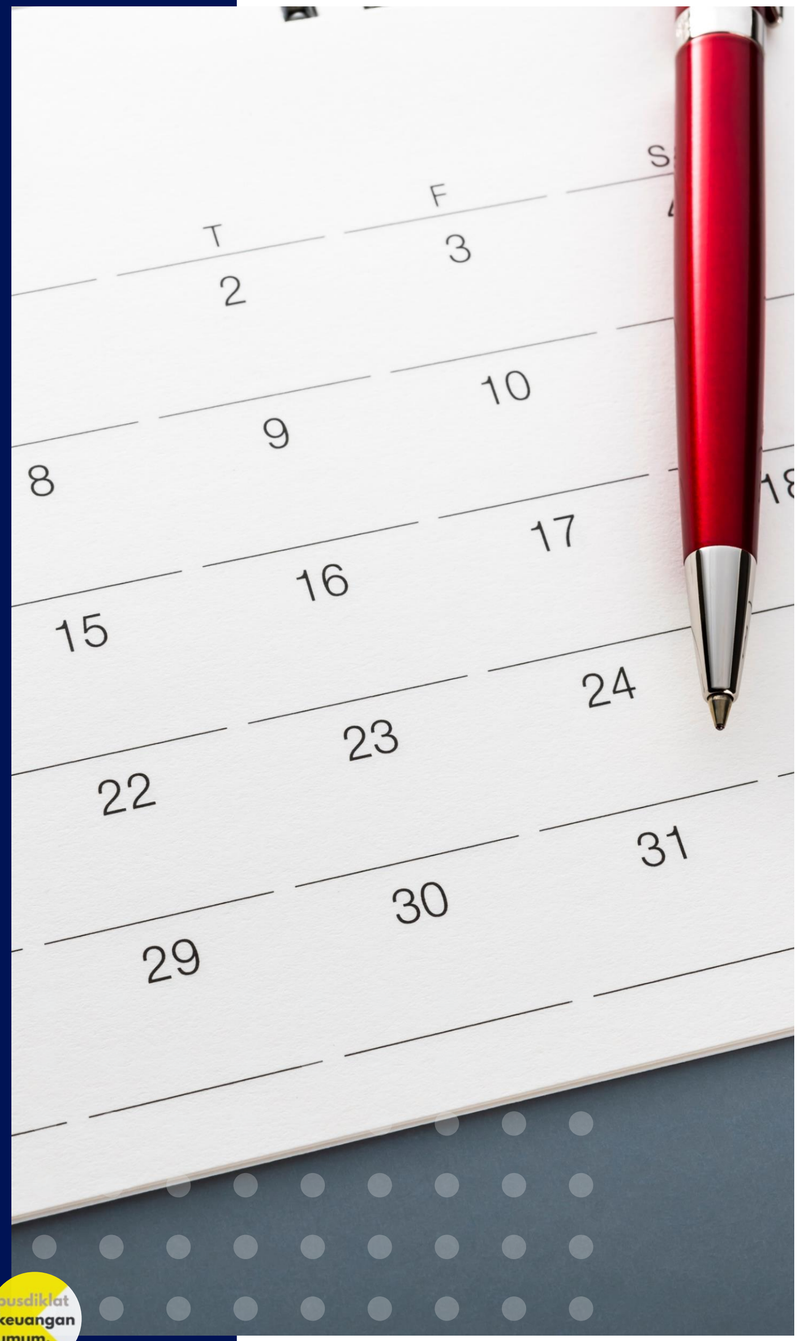
BU.3. Determine Data Mining Goals	
Translate business goals to data mining goals.	
Outputs	
Data mining goals	Describe the intended outputs of the project that achieve the business objectives.
	• Menggunakan pendekatan klasifikasi untuk menentukan dokumen barang kiriman yang benar-benar fraud atau terdapat temuan • Menggunakan pendekatan regresi untuk memperkirakan nilai fraud yang akan terjadi
Data mining success criteria	Define the criteria for a successful outcome of the project in technical terms
	• Pendekatan klasifikasi yang digunakan harus memenuhi tingkat akurasi minimal 70% dan sensitifitas minimal 90% • Pendekatan regresi yang digunakan harus memenuhi akurasi MAPE < 20



Produce a Project Plan

Project Plan and Initial Assessment of tools and Techniques

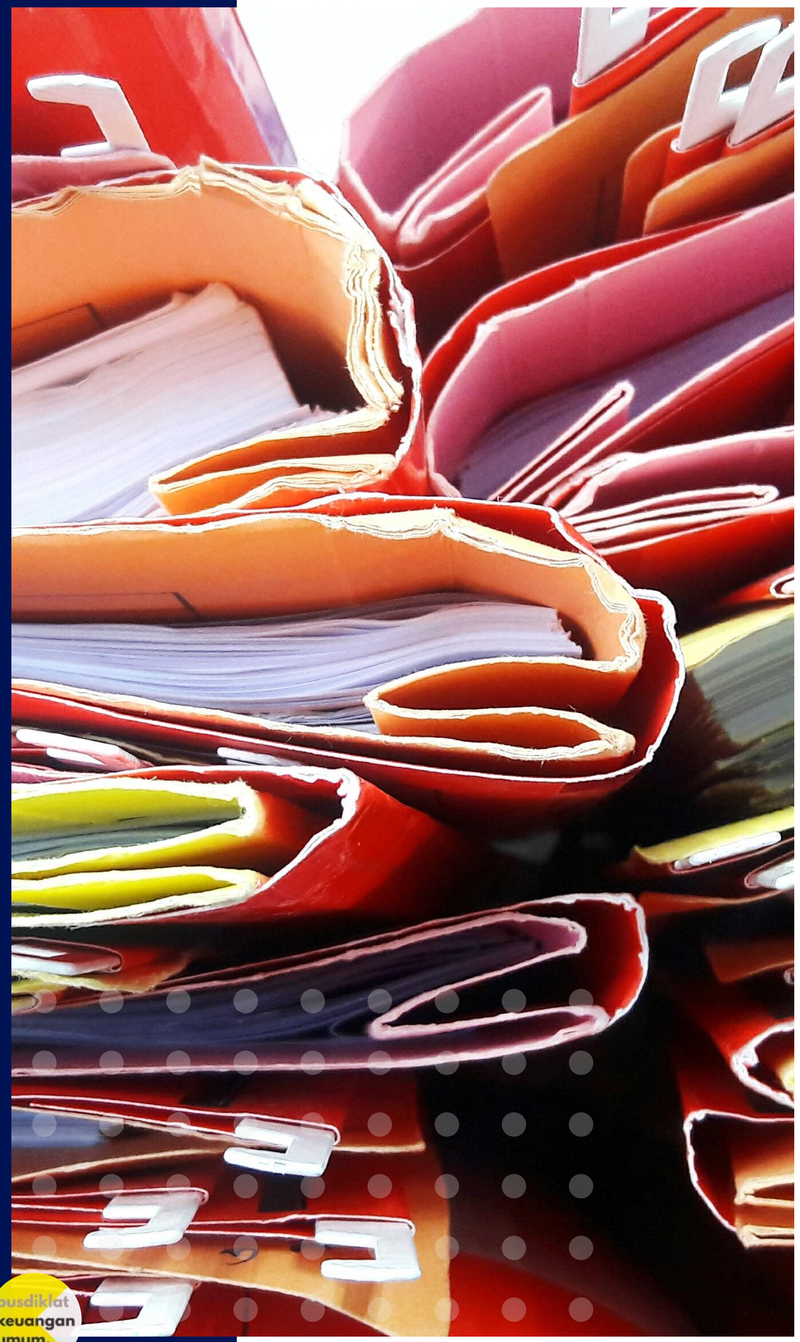
- Melakukan **rencana proyek**:
  - *Breakdown* kegiatan-kegiatan yang akan dieksekusi dalam proyek
  - Menentukan durasi setiap kegiatan
  - Menjelaskan kebutuhan yang telah didefinisikan pada bagian sebelumnya
  - Menentukan input dan output setiap kegiatan
- Melakukan **inventarisir terhadap *tools* dan pendekatan** yang akan digunakan:
  - Pendataan *tools* dan *software* yang akan digunakan
  - Membuat daftar pendekatan atau algoritma yang digunakan



Produce a Project Plan (2)

Template dan contoh output

BU.4. Produce a Project Plan	
Define a plan for achieving the data mining goals. The plan should specify the steps to be performed during the project, including the initial selection of tools and techniques.	
Outputs	
Project plan	List the stages to be executed in the project, including their duration, required resources, inputs, outputs and dependencies. Analyse dependencies between time schedule and risks.
	Business Understanding Phase Data Understanding Phase Data Preparation Phase Modeling Phase Deployment Phase Evaluasi Phase
Initial assessment of tools and techniques	This output performs an initial assessment of tools and techniques.
	Tools yang digunakan: Python 3.8 Jupiter Notebook **Pendekatan/Algoritma yang digunakan:** Klasifikasi Regresi



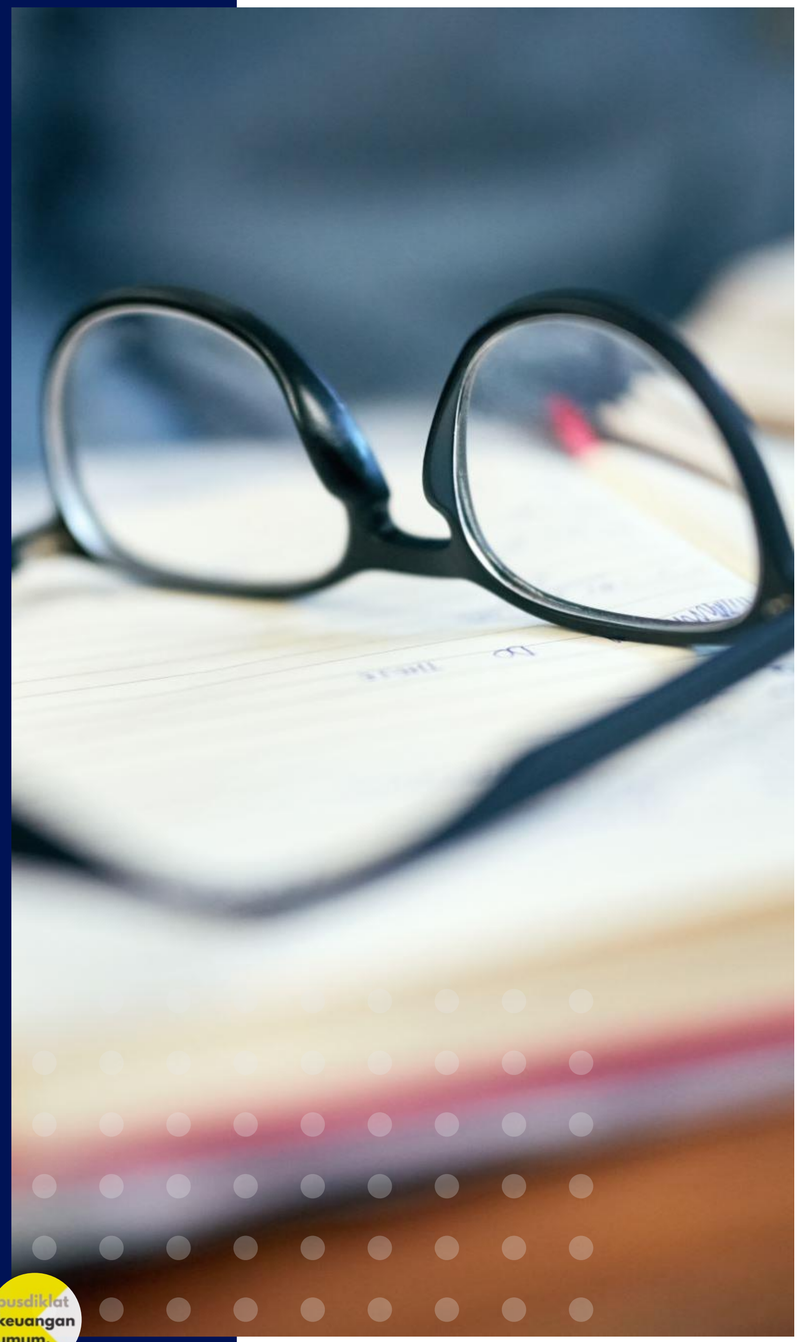
Collect Initial Data

Initial Data Collection Report

- **Mendefinisikan seluruh kebutuhan data** yang digunakan dalam proyek:
  - Buat perencanaan terhadap data yang dibutuhkan dalam proses analisis
  - Cek kembali jika data yang dibutuhkan tersedia atau tidak
  - Pilah kebutuhan kolom yang dibutuhkan dari suatu data
  - Tentukan periode waktu data yang dibutuhkan
  - Pengumpulan data bisa saja didapatkan melalui sumber data non-elektronik

	What	What	What	Where	Who	When	How
No.	Data Name	Number of Cols	Number of Rows	Data Source	Data Provider	Colelction Date	Acquired Method
1	Data Pengajuan Barang Kiriman Periode 2021	15 Columns	52.723.238 Rows	DB Report Barkir	Dit. IKC	1/2/22 10:00	Direct Query
2	Data Penetapan Barang Kiriman Periode 2021	15 Columns	52.723.238 Rows	DB Report Barkir	Dit. IKC	1/2/22 14:00	Direct Query
3	Data Barang Marketplace Hijau	N/A	N/A	Website marketplace Hijau	Marketplace Hijau	1/4/22 8:00	Webscrapping
4	Data Barang Marketplace Orange	N/A	N/A	Website marketplace Orange	Marketplace Orange	1/5/22 9:30	Webscrapping

pusdiklat keuangan umum.

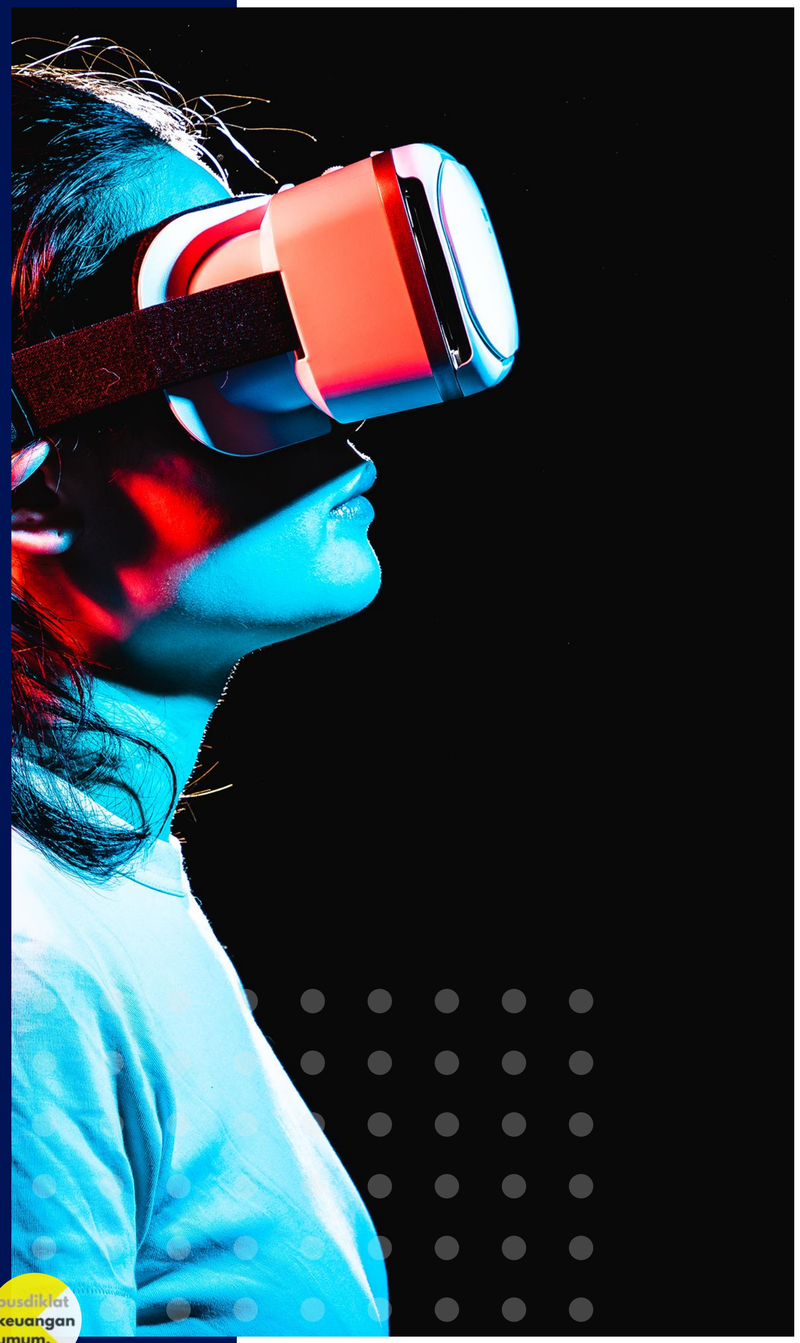


Describe Data

Data Description Report

- **Mendefinisikan detil data** yang telah ditentukan pada tahap *collect initial data*:
  - Menjelaskan format data yang digunakan
  - Kuantitas data (jumlah data)
  - Identifikasi kebutuhan kolom
  - Mendefinisikan atribut tipe data yang digunakan

Data Desctiption Report: Data Pengajuan Barang Kiriman Periode 2021 (Excel) – 1000 row						
No.	Name	Description	Data Type	Length	Accept Null Value	Note
1	Nomor AWB	Nomor Airway Bill Barang Kiriman	Varchar	15	No	-
2	Tanggal AWB	Tanggal airway Bill Barang Kiriman	Date	-	No	-
3	Nama Pemberitahu	Perusahaan yang mengirim data	Varchar	128	No	-
5	Consignee	Nama Pemilik Barang	Varchar	128	No	-
dst						



Explore Data

Data Exploration Report

- **Melakukan eksplorasi dari data** yang ada untuk **menemukan pengetahuan awal** terhadap masalah yang dihadapi**:**
  - Melakukan analisis terhadap atribut yang menarik untuk dilihat dengan pendekatan statistik dasar
  - Menentukan hipotesis awal terhadap analisis data yang akan dilakukan
  - Melakukan verifikasi terhadap keseluruhan atribut apakah masih perlu digunakan atau tidak

Verify Data Quality

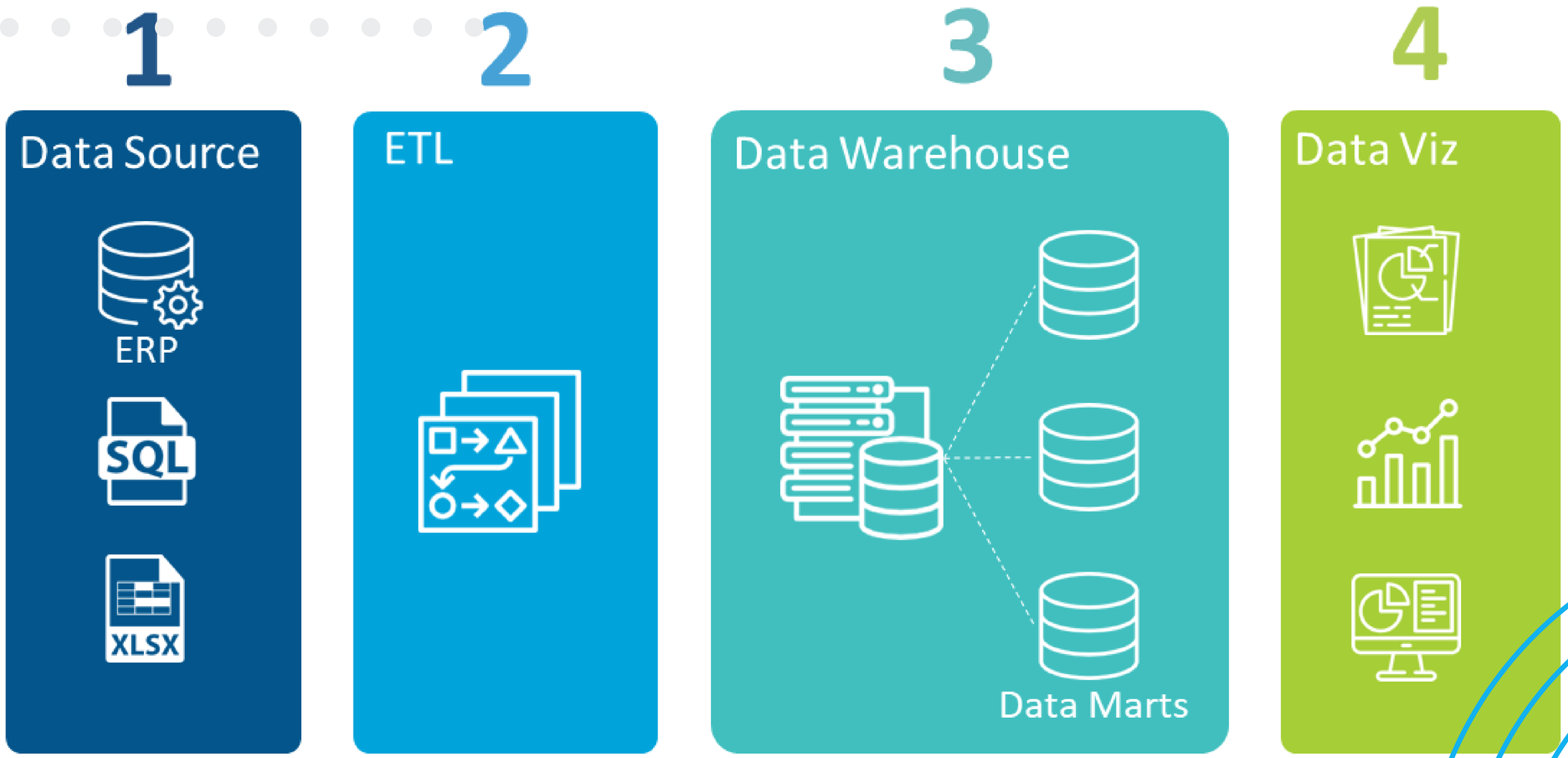
Data Quality Report

- **Melakukan pengecekan atribut dari data** yang ada untuk **menemukan *error*** pada data yang digunakan**:**
  - Membuat daftar atribut yang masih memiliki isu kualitas data (*missing, noise, outlier,* duplikasi) dan mitigasinya
  - Jika data berupa flat files, gunakan pengecekan delimiter dan pastikan jumlah kolom dan row nya sesuai
  - Lakukan pengecekan konsistensi dan redudansi yang terjadi dari sumber data yang berbeda

KEMENTERIAN KEUANGAN
REPUBLIK INDONESIA

Ekstrak, Transform, Load (ETL) sebagai Jembatan Data Operasional dengan Data yang akan Digunakan untuk Analisis

BI ARCHITECTURE



Sumber: https://t2client.com/bi/

KEMENTERIAN KEUANGAN
REPUBLIK INDONESIA

Sumber Data

- SLDK
- Open Data (Jabar, DKI, Satu Data Indonesia)
- Social Media
- Kaggle
- UCI Machine Learning Repository
- Google Dataset Search
- **Dan masih banyak lainnya...**

Thank you!

Any questions?

Credit:

- Presentation template by HiSlide.io
- Icons by Font Kiko
- Photos by Pexels

KEMENTERIAN KEUANGAN
REPUBLIK INDONESIA