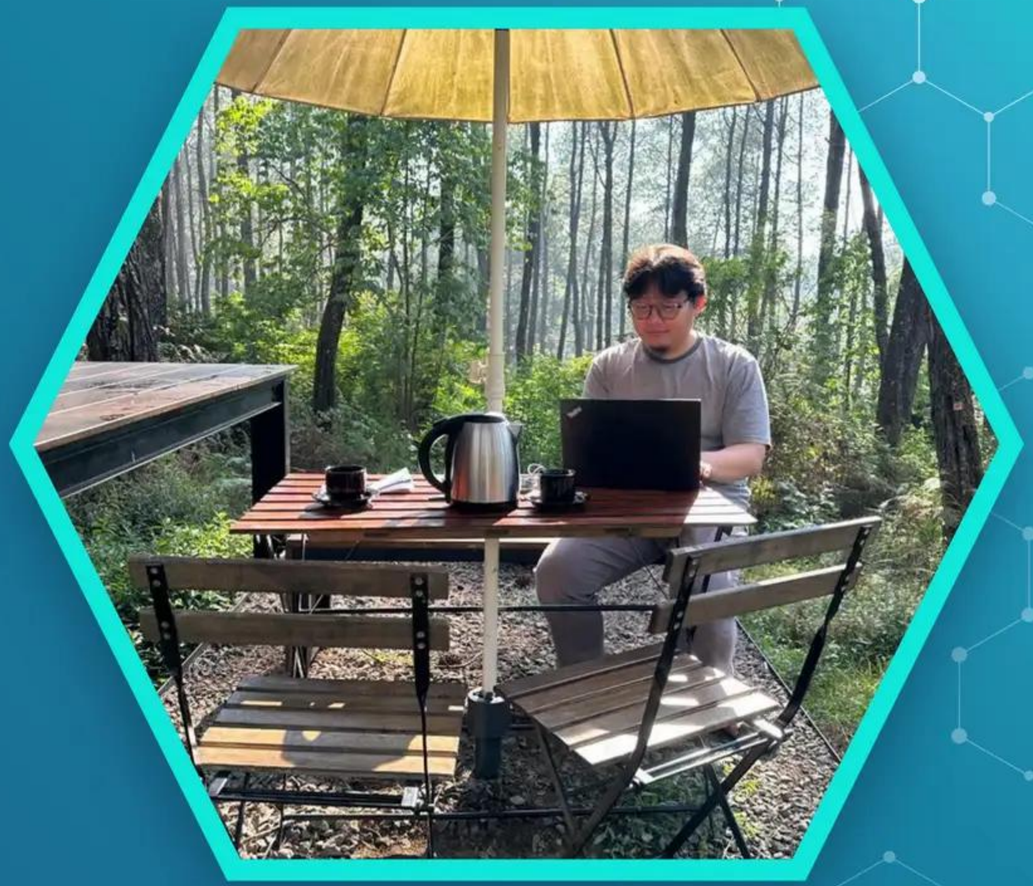
Analytics Today:
Data, Tech, and Regulation

Hendri Karisma

HELLO!

I AM HENDRI KARISMA, M.T.

- Principal R&D at **PT. Akar Intidata**
- Data Team Lead
- Working for Advanced Analytics Platform, Geo Performance, Fraud Detection, and Machine Learning Engine



You can find me at

@infoHendri (X/twitter)

@karism4_ (instagram)

“Data analytics converts raw data into actionable insights. It includes a range of tools, technologies, and processes used to find trends and solve problems by using data. Data analytics can shape business processes, improve decision-making, and foster business growth.”

Data Analytics

- Descriptive Analytics
- Diagnostics Analytics
- Predictive Analytics
- Prescriptive Analytics

Data Analytics vs AI vs Machine Learning

1 Data

Let's start with the Definition

“Data is Fact”

“Data is Fact”

“...Data just like fuel for Rocket.. And The Method is the Rocket... “

3 Important Statistics About How Much Data Is Created Every Day

FinancesOnline
REVIEWS FOR BUSINESS

1 How much data is generated every minute?

Source: Domo

41,666,667
messages shared by WhatsApp users

1,388,889
video / voice calls made by people worldwide

404,444
hours of video streamed by Netflix users

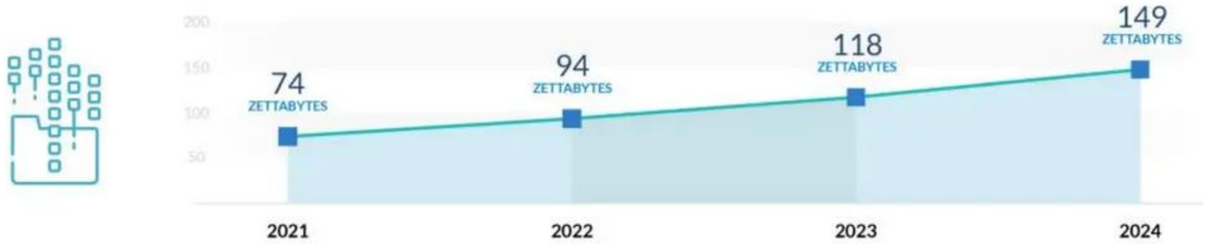
347,222
stories posted by Instagram users

150,000
messages shared by Facebook users

147,000
photos shared by Facebook users

2 Estimated Data Consumption from 2021 to 2024

Source: IDC / Statista



3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

2 TRILLION
searches on Google by the end of 2021

1.134 TRILLION MB
volume of data created every day

3,026,626
emails sent every second, 67% of which are spam

278,108 PETABYTES
global IP data per month by the end of 2021

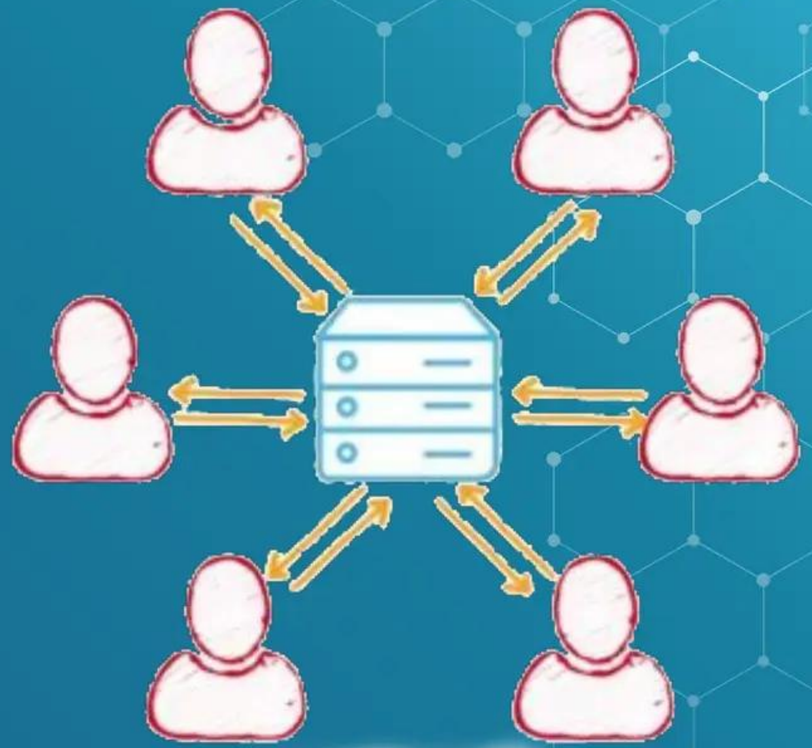
230,000
new malware versions created every day

82%
share of video in total global internet traffic at the end of 2021

The Type of data

first-party data

second-party data



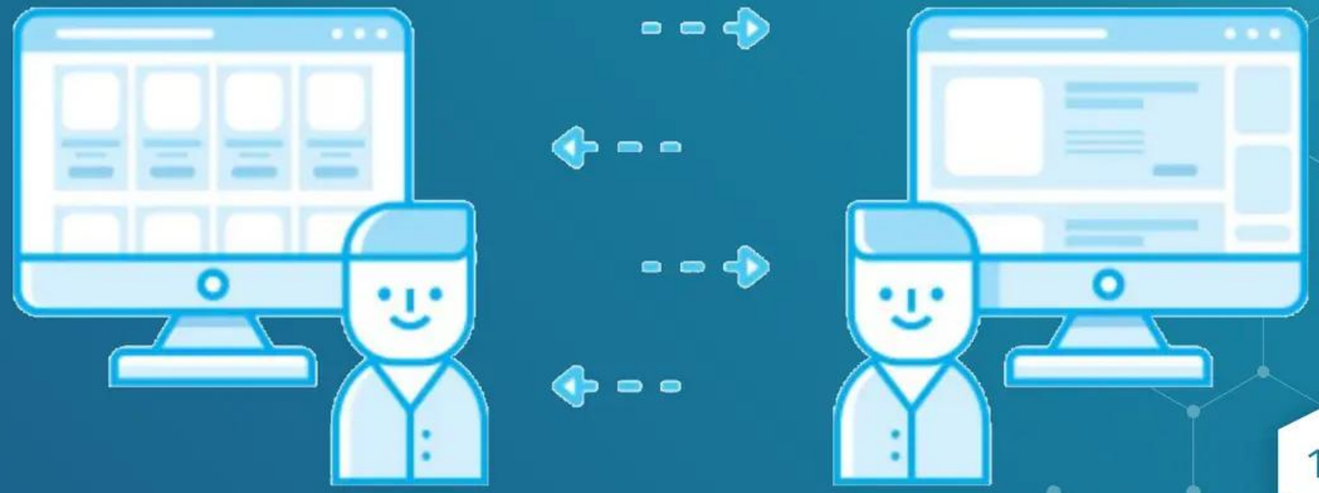
third-party data

First Party Data

- Data collected and owned by a brand
- “Free”, differentiated and highly relevant
- Accurate
- Ex : Transactional, Online Behaviour, CRM, Subscription, Social Data

Second Party Data

- From Strategic partner
- Another Company first party data
- Not broadly sold and available only to those within the agreement



Third Party Data

- Paid online and offline data made available
- Broadly sold and easly accessible
- Could less accurate

Why 3rd Party is matter?

- Efficient to enhance our first party data
- Came from outside our target market or customer
- Collected by multiple ways, from multiple sources
- Less reliable from first and second party data

How we get the 3rd party data?

To get more fuel for our rocket

How we get the 3rd party data?

Consumer

Single platform

Market Place

Data Provider

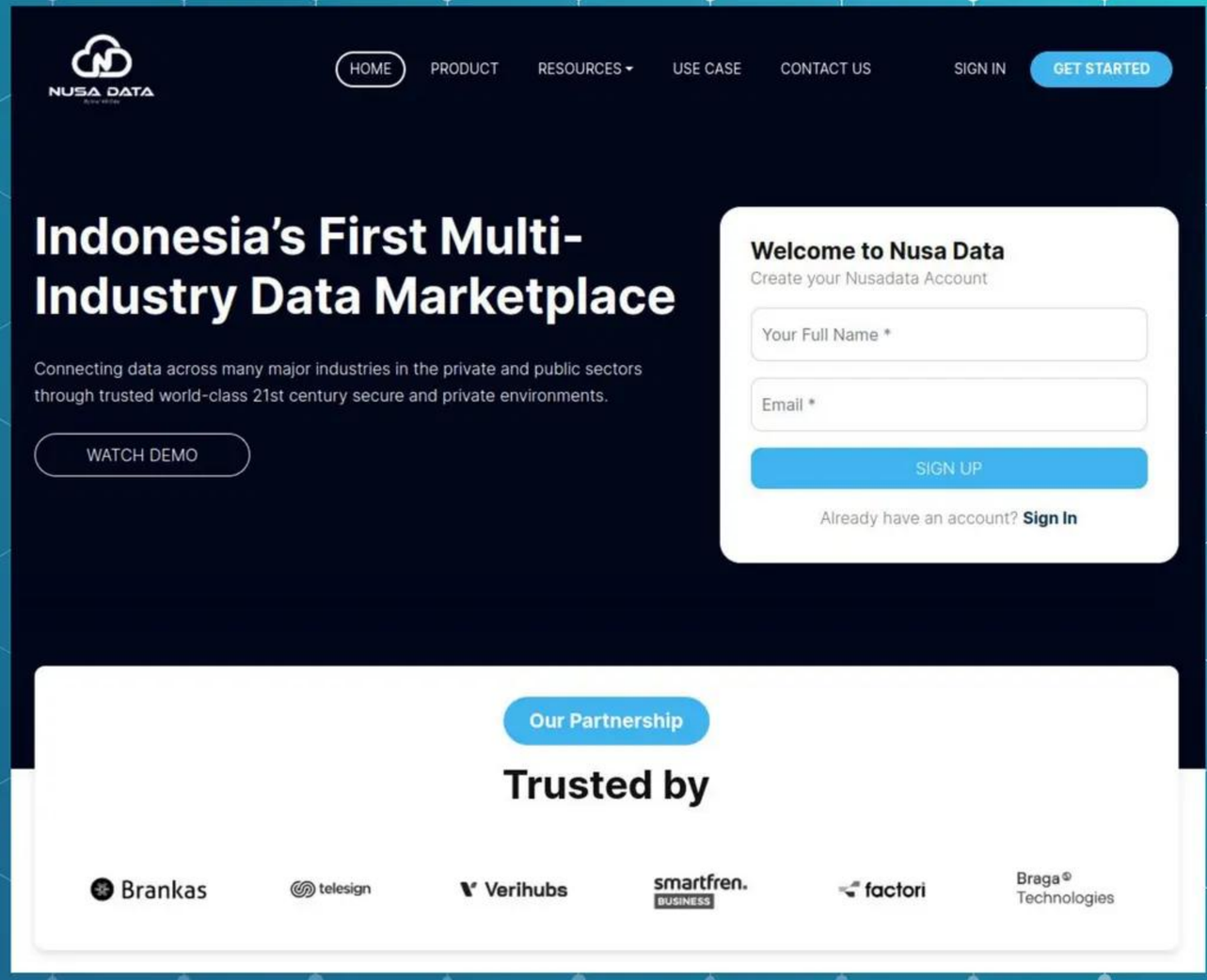
Open Finance

Telco

eKYC

Nusa Data

- Data Marketplace Multi-Industry
- Connecting between providers with the consumers



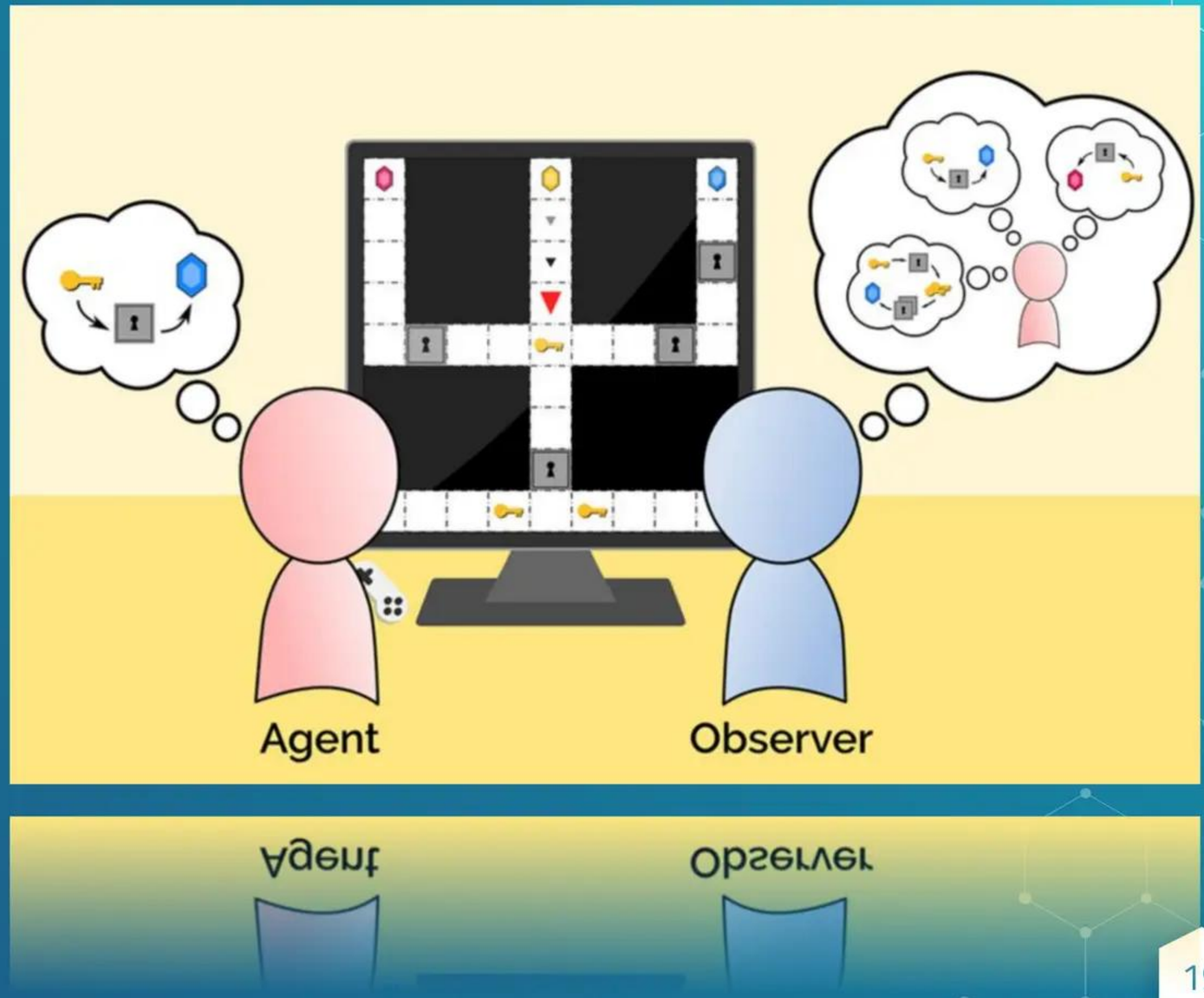
2

Technologies (methods or tools or apps)

Develop AI to Support The Process and Build The Solutions

Today we have AI,

Just use LLM!

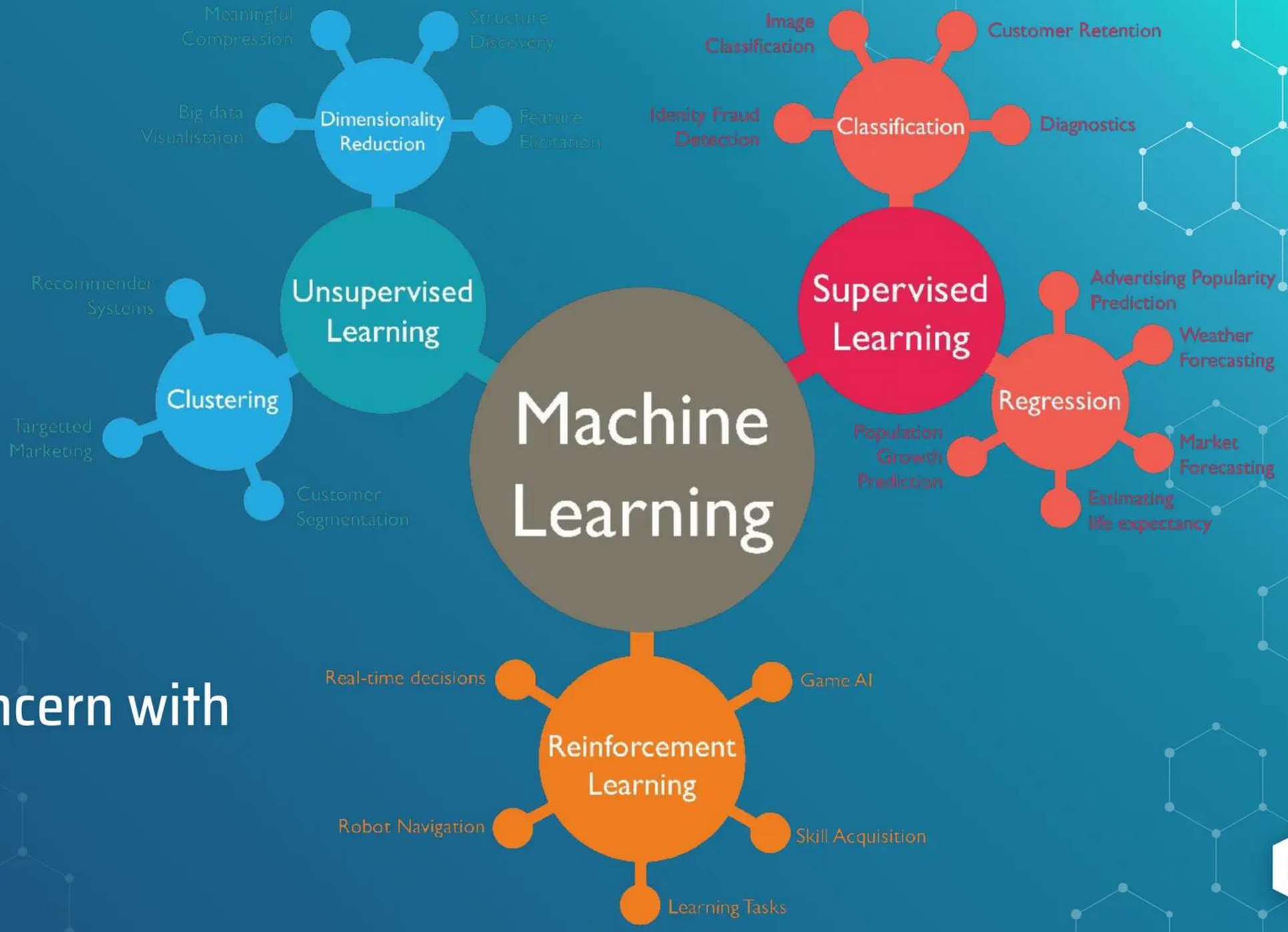


BARD AI

Large Language Model
Large Visualization Model
Generative ML
Deep Learning
Machine Learning

Is Industry directly need those kind of Techs?

Is it solve the problems for all industry areas?



They still concern with specific area

Issues in Industry

- Data
- Tech
- Experts
  - Data Scientist
  - AI/ML Engineer
  - Data Engineer
- Regulation
- Cost

Data Role

Scientist

The goal of a scientist is to answer questions and discover information about their chosen field of study.

Engineering

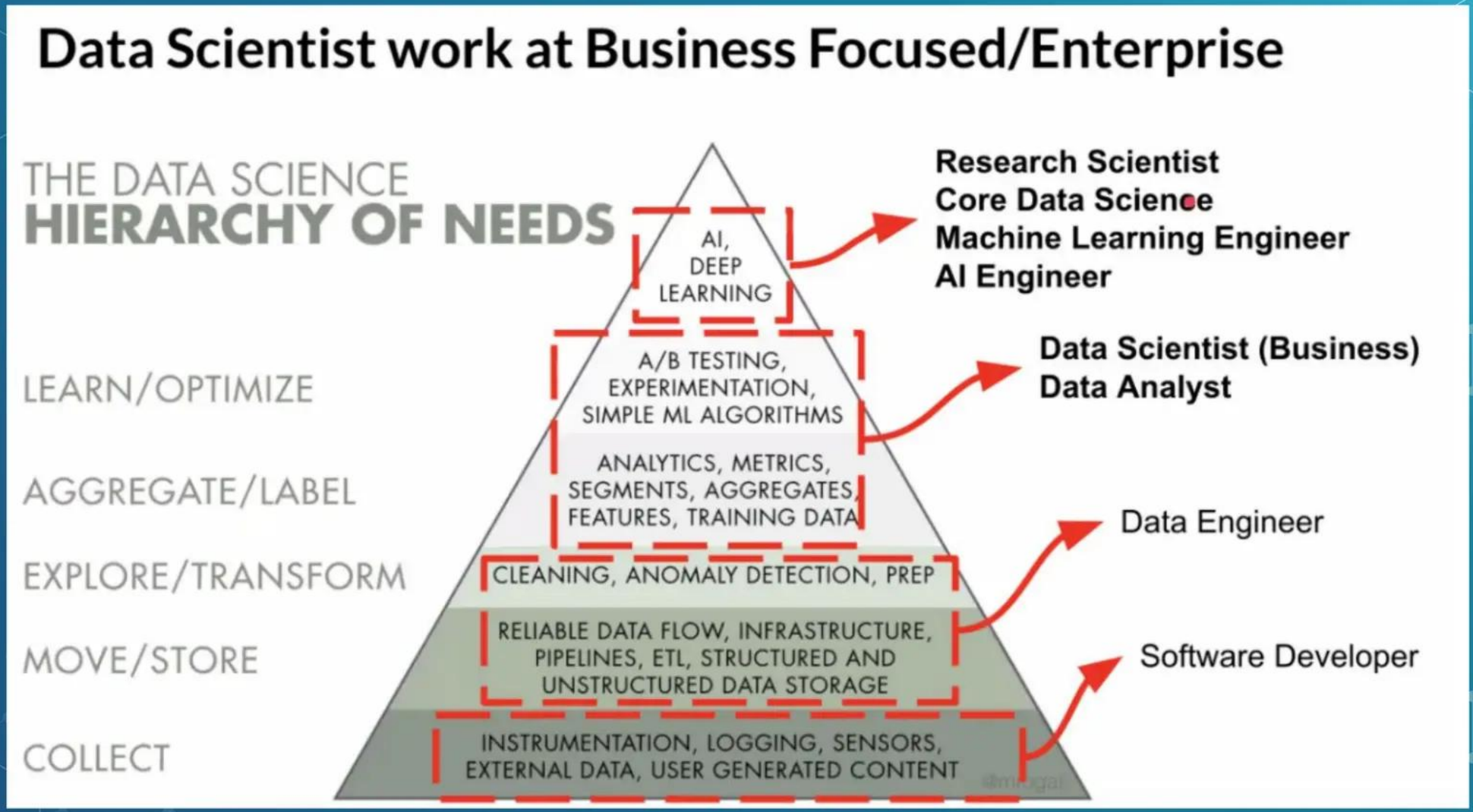
An engineer might produce physical item or blueprint for a new process.

.

Data Roles

- Data Scientist
- Data Analyst
- Business Intelligence
- Data Engineer
- AI/ML Engineer

Data Scientist work at Business Focused/Enterprise



Data Scientist

The term “data scientist” was coined as recently as 2008 when companies realized the need for data professionals who are skilled in organizing and analyzing massive amounts of data.1 In a 2009 McKinsey&Company article, Hal Varian, Google’s chief economist and UC Berkeley professor of information sciences, business, and economics, predicted the importance of adapting to technology’s influence and reconfiguration of different industries.

Data Engineer

Data engineers work in a variety of settings to build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret. Their ultimate goal is to make data accessible so that organizations can use it to evaluate and optimize their performance.

These are some common tasks you might perform when working with data:

- Acquire datasets that align with business needs
- Develop algorithms to transform data into useful, actionable information
- Build, test, and maintain database pipeline architectures
- Collaborate with management to understand company objectives
- Create new data validation methods and data analysis tools
- Ensure compliance with data governance and security policies

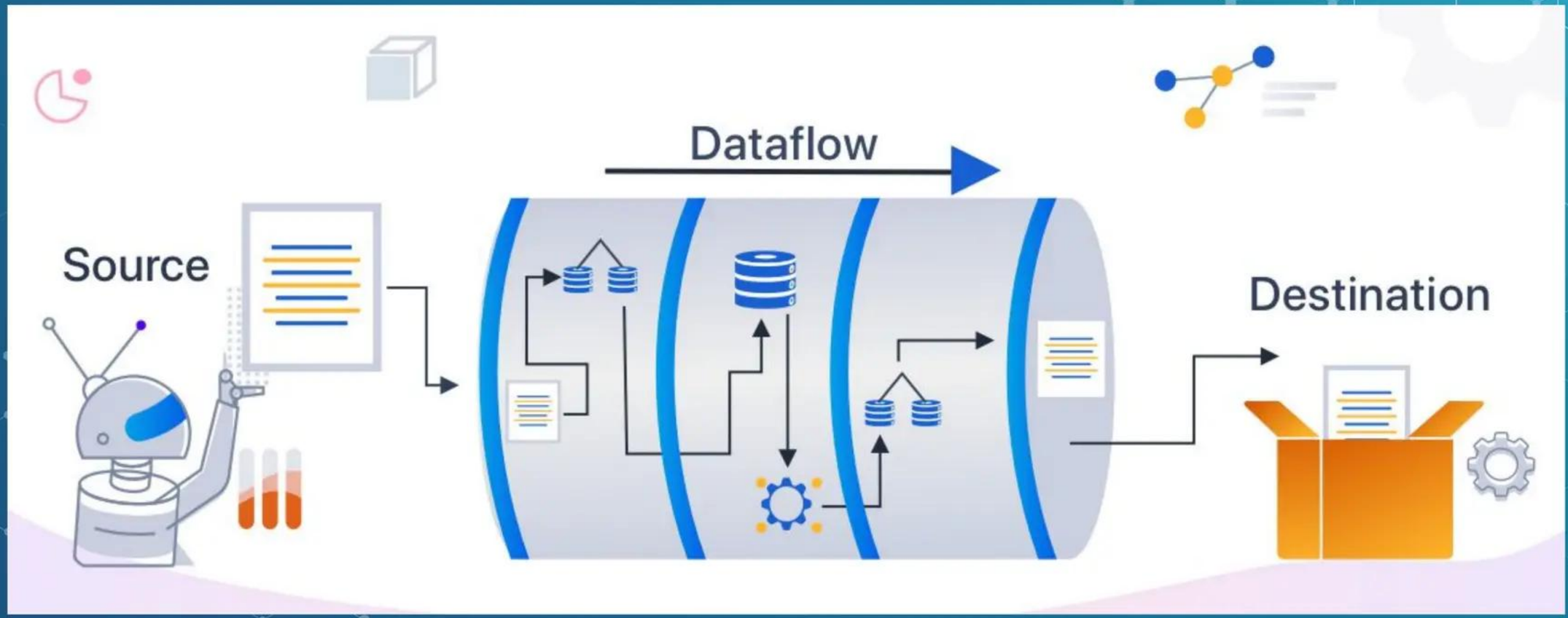
AI/ML Engineer

Artificial intelligence (AI) engineers are responsible for developing, programming and training the complex networks of algorithms that make up AI so that they can function like a human brain. This role requires combined expertise in software development, programming, data science and data engineering. Though this roleis related to data engineering, AI engineers are rarely required to write the code that develops scalable data sharing.

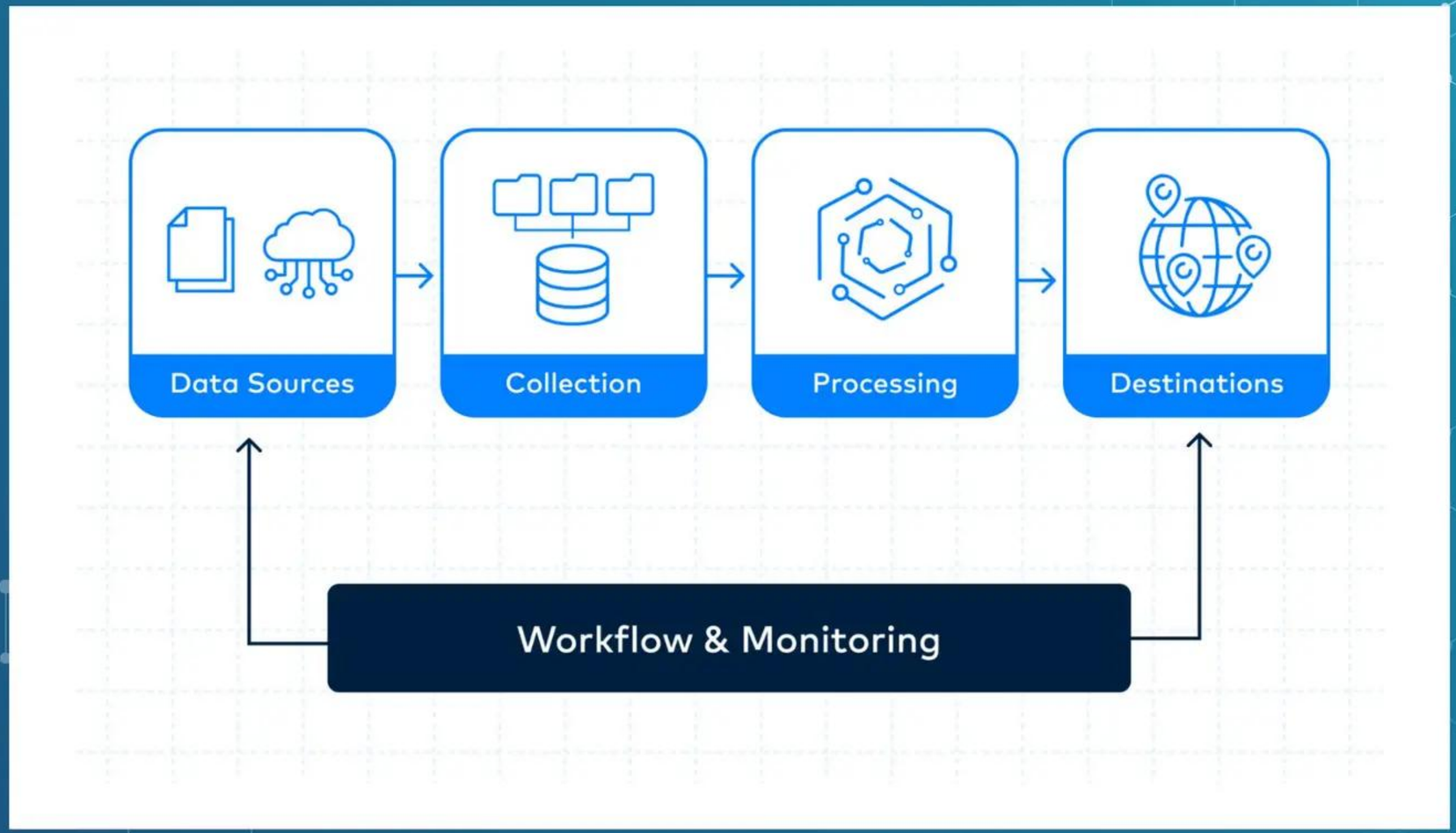
AI/ML Engineer Skill set

- Build AI models from the ground up and explain results to product managers and stakeholders
- Develop, test, and deploy AI models
- Convert machine learning models into APIs so other applications can utilize it
- Build data ingestion and data transformation infrastructure
- Work alongside data and business analysts
- Execute statistical analysis and tune results to extract better insights
- Automate infrastructure used by the data science team
- Create and manage AI development and production infrastructure

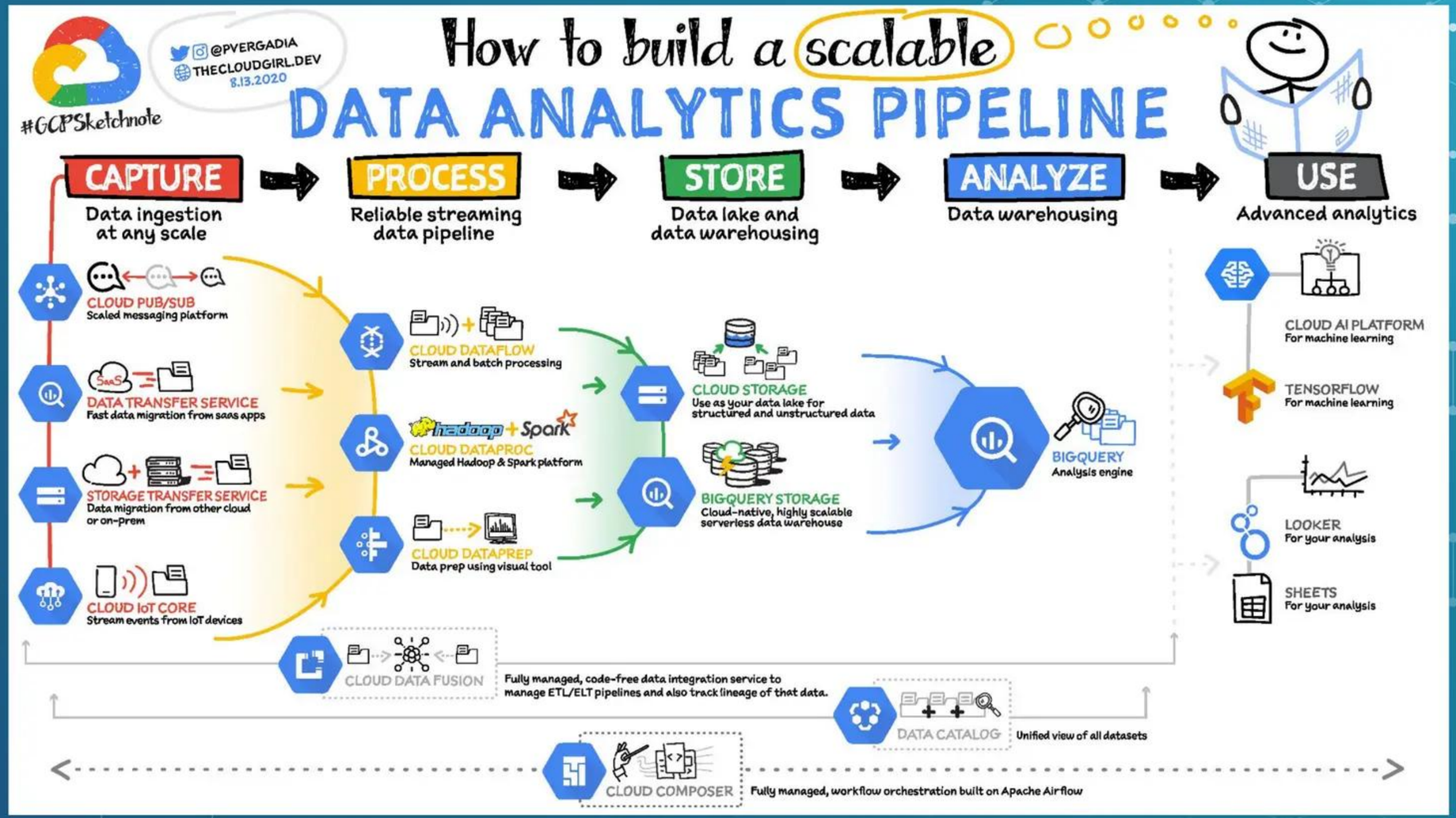
Tech that need to acquired to build the solution

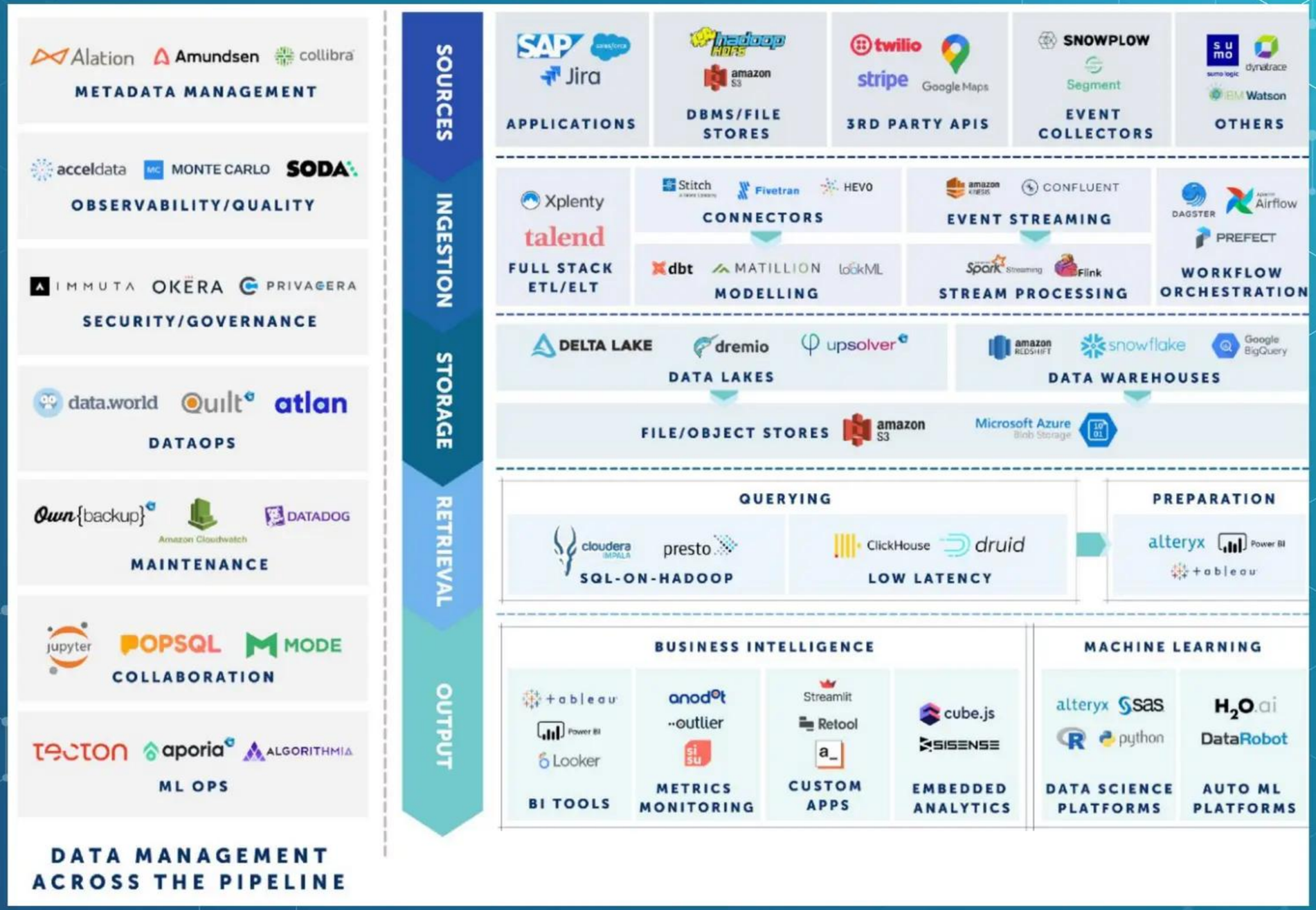


Zoom in


Data Sources
Collection
Processing
Destinations
Workflow & Monitoring

More detail


#GCPSketchnote
@PVERGADIA
THECLOUDGIRL.DEV
8.13.2020
How to build a scalable
DATA ANALYTICS PIPELINE
CAPTURE
Data ingestion at any scale
PROCESS
Reliable streaming data pipeline
STORE
Data lake and data warehousing
ANALYZE
Data warehousing
USE
Advanced analytics
CLOUD PUB/SUB
Scaled messaging platform
DATA TRANSFER SERVICE
Fast data migration from saas apps
STORAGE TRANSFER SERVICE
Data migration from other cloud or on-prem
CLOUD IoT CORE
Stream events from IoT devices
CLOUD DATAFLOW
Stream and batch processing
CLOUD DATAPROC
Managed Hadoop & Spark platform
CLOUD DATAPREP
Data prep using visual tool
CLOUD STORAGE
Use as your data lake for structured and unstructured data
BIGQUERY STORAGE
Cloud-native, highly scalable serverless data warehouse
BIGQUERY
Analysis engine
CLOUD AI PLATFORM
For machine learning
TENSORFLOW
For machine learning
LOOKER
For your analysis
SHEETS
For your analysis
CLOUD DATA FUSION
Fully managed, code-free data integration service to manage ETL/ELT pipelines and also track lineage of that data.
DATA CATALOG
Unified view of all datasets
CLOUD COMPOSER
Fully managed, workflow orchestration built on Apache Airflow


Alation
Amundsen
collibra
METADATA MANAGEMENT
acceldata
MONTE CARLO
SODA
OBSERVABILITY/QUALITY
IMMUTA
OKERA
PRIVACERA
SECURITY/GOVERNANCE
data.world
Quilt
atlan
DATAOPS
Own{backup}
Amazon Cloudwatch
DATADOG
MAINTENANCE
jupyter
POPSQL
MODE
COLLABORATION
tecton
aporia
ALGORITHMIA
ML OPS
DATA MANAGEMENT ACROSS THE PIPELINE
SOURCES
INGESTION
STORAGE
RETRIEVAL
OUTPUT
SAP
salesforce
Jira
APPLICATIONS
hadoop
amazon S3
DBMS/FILE STORES
twilio
stripe
Google Maps
3RD PARTY APIS
SNOWPLOW
Segment
EVENT COLLECTORS
sumo logic
dynatrace
IBM Watson
OTHERS
Xplenty
talend
FULL STACK ETL/ELT
Stitch
Fivetran
HEVO
CONNECTORS
dbt
MATILLION
lookML
MODELLING
amazon KINESIS
CONFLUENT
EVENT STREAMING
Spark Streaming
Flink
STREAM PROCESSING
DAGSTER
Airflow
PREFECT
WORKFLOW ORCHESTRATION
DELTA LAKE
dremio
upsolver
DATA LAKES
amazon REDSHIFT
snowflake
Google BigQuery
DATA WAREHOUSES
FILE/OBJECT STORES
amazon S3
Microsoft Azure
Blob Storage
QUERYING
cloudera IMPALA
presto
SQL-ON-HADOOP
ClickHouse
druid
LOW LATENCY
PREPARATION
alteryx
Power BI
tableau
BUSINESS INTELLIGENCE
tableau
Power BI
Looker
BI TOOLS
anodot
outlier
sisu
METRICS MONITORING
Streamlit
Retool
a_
CUSTOM APPS
cube.js
SISENSE
EMBEDDED ANALYTICS
MACHINE LEARNING
alteryx
sas
R
python
DATA SCIENCE PLATFORMS
H2O.ai
DataRobot
AUTO ML PLATFORMS

Engineering skill that need to be required

- SOLID, Design Pattern, Clean Architecture, etcs
- Software Engineering
- Tech Tools Characteristics, Behaviour
- Basic statistics, probilistics and random variable, basic machine learning.
- High performance computing
- Security and Regulation
- ML Ops

Why Engineering Important?

Sample Case:

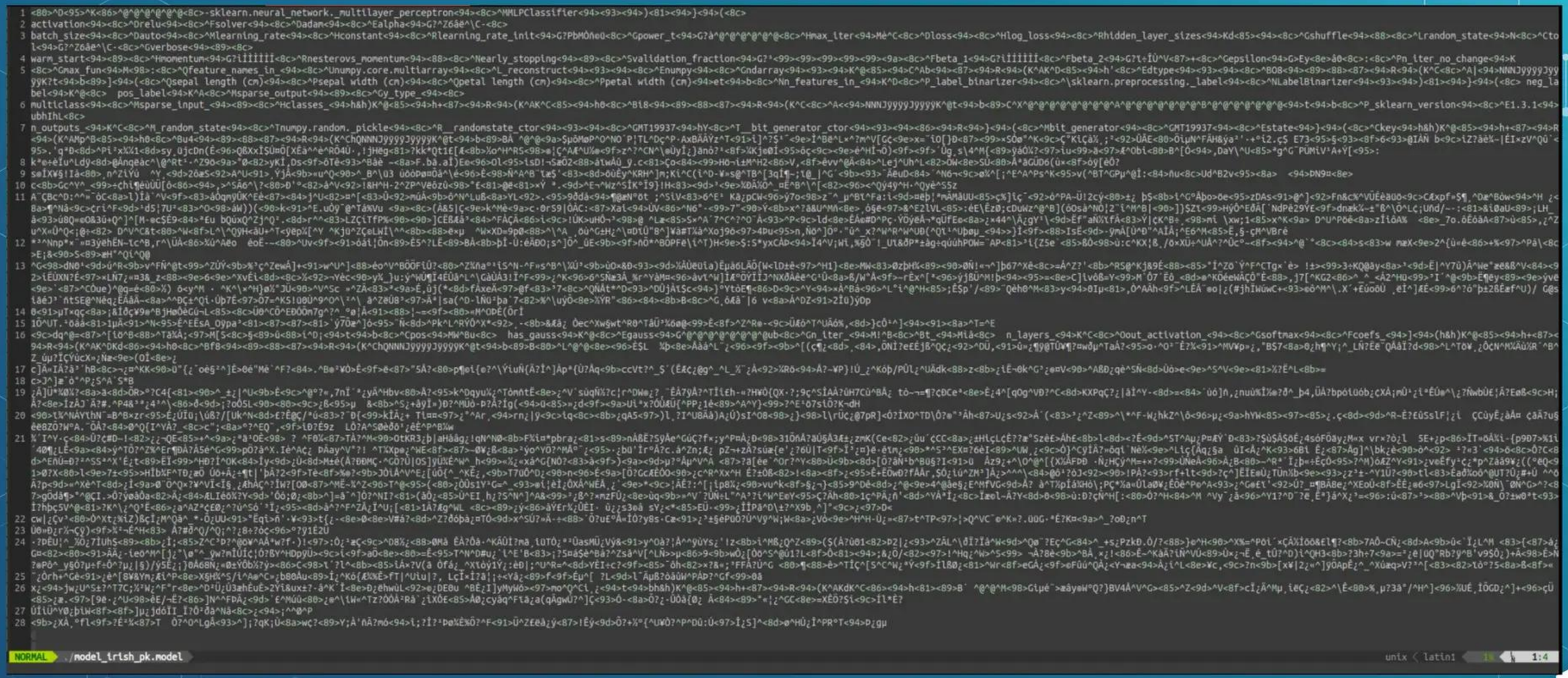
- We need to run ML service
  - Sometimes for specific cases
  - The classification process is about 500ms
- We run the process (preprocess to main analytics process) as a batch process or need to process data with Large Volume
- **The model is lightweight**

Build the Engine

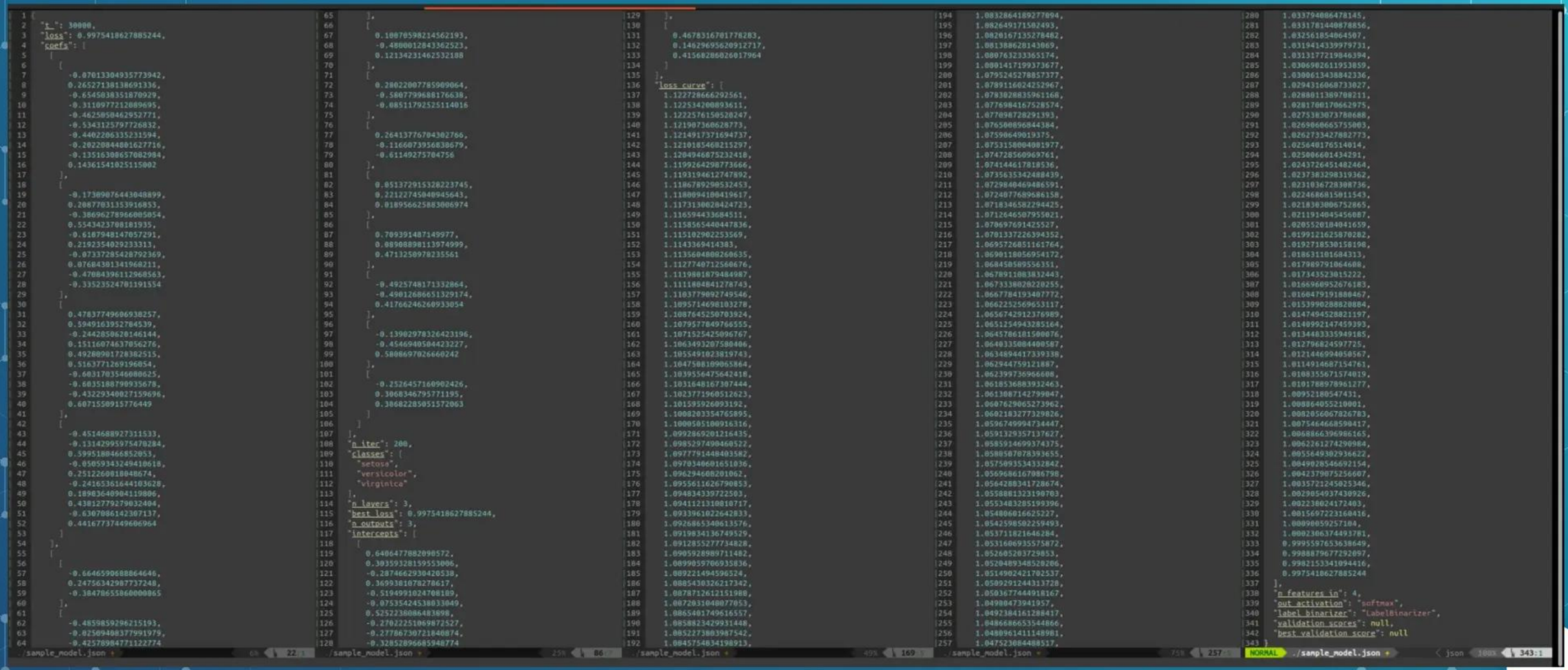
- We often use pickle or serializable file as model output
- The size bigger than it should be
- Impacted to the memory or the space complexity

```
total 44K
-rw-rw-r-- 1 hendri hendri 7,7K Okt 22 08:24 model_irish_json.json
-rw-rw-r-- 1 hendri hendri  25K Okt 22 08:41 model_irish_pk.model
-rw-rw-r-- 1 hendri hendri  778 Okt 22 08:41 sample_irish.py
drwxrwxr-x 5 hendri hendri 4,0K Okt 22 08:27 venv
```

Sample Model with Serializable File



Sample Model using Good Engineering



3

Regulation

Need to Comply with all standard especially security

GRC


Governance
Audits
Strategy
Controls
GRC
Policies
Risk
Processes
Performance
Risk
Compliance

Compliences

- GDPR (Europe)
- ISO 27001 (Information Security
- UU PDP
- PCI DSS (Payment)
- PA DSS (Payment)
- HIPAA (Healthcare)

What we need to protect?

PII

Personally identifiable information (PII) is any information connected to a specific individual that can be used to uncover that individual's identity, such as their social security number, full name, or email address

Consent

Consent means giving people genuine choice and control over how you use their data. If the individual has no real choice, consent is not freely given and it will be invalid. This means people must be able to refuse consent without detriment, and must be able to withdraw consent easily at any time.

Credentials

Credentials are a set of login or authentication data that verify a user's identity and grant them access to a particular system or service.

We need to know how handle the data and our limitation

Principles

Chapter 3 (Art. 12 – 23)
Rights of the data subject

- Art. 12 – Transparent information, communication and modalities for the exercise of the rights of the data subject
- Art. 13 – Information to be provided where personal data are collected from the data subject
- Art. 14 – Information to be provided where personal data have not been obtained from the data subject
- Art. 15 – Right of access by the data subject
- Art. 16 – Right to rectification
- Art. 17 – Right to erasure ('right to be forgotten')
- Art. 18 – Right to restriction of processing
- Art. 19 – Notification obligation regarding rectification or erasure of personal data or restriction of processing
- Art. 20 – Right to data portability
- Art. 21 – Right to object
- Art. 22 – Automated individual decision-making, including profiling
- Art. 23 – Restrictions

Chapter 4 (Art. 24 – 43)
Controller and processor

Chapter 5 (Art. 44 – 50)
Transfers of personal data to third countries

B. DESCRIPTION OF TRANSFER

Categories of personal data processed

Identification data	
Personal identification data	Title, name, first name, private address, phone numbers, email addresses (personal or professional at your choice).
Financial data	
Financial identification data	Bank account numbers, expenses and supporting documents.
Personal characteristics	
Personal details	Age, gender, year and place of birth, citizenship.
Training and studies data	
Academic background and professional qualification	Degree level, PhD thema, year of start of PhD thesis, date of award of PhD, participation in COST trainings.
Professional data	
Professional identification data	Position, institution of affiliation (name and address), place of work, scientific field of expertise, research area.
Career	Participation and role in COST Actions, your grants (dates, funding agency(ies)), CV, ORCID id, professional webpage if any

Fintech Regulation - Payment

Who Does PA DSS and PCI DSS Apply To?

PA DSS

Vendors that develop and sell payment applications

PCI DSS

Merchants and service providers that store, process, or transmit cardholder data

Your Level is:	Your business does:	You should:
4	- less than 20,000 eCommerce transactions per year - less than 1 million other transactions per year	- complete an annual risk assessment using an SAQ - conduct quarterly PCI scans
3	- 20,000 - 1 million transactions per year	- complete an annual risk assessment using an SAQ - conduct quarterly PCI scans
2	- 1-6 million transactions per year	- complete an annual risk assessment using an SAQ - conduct quarterly PCI scans
1	- 6 million+ transactions per year	- conduct an annual internal audit - conduct quarterly PCI scans

Our Standards (Indonesia)

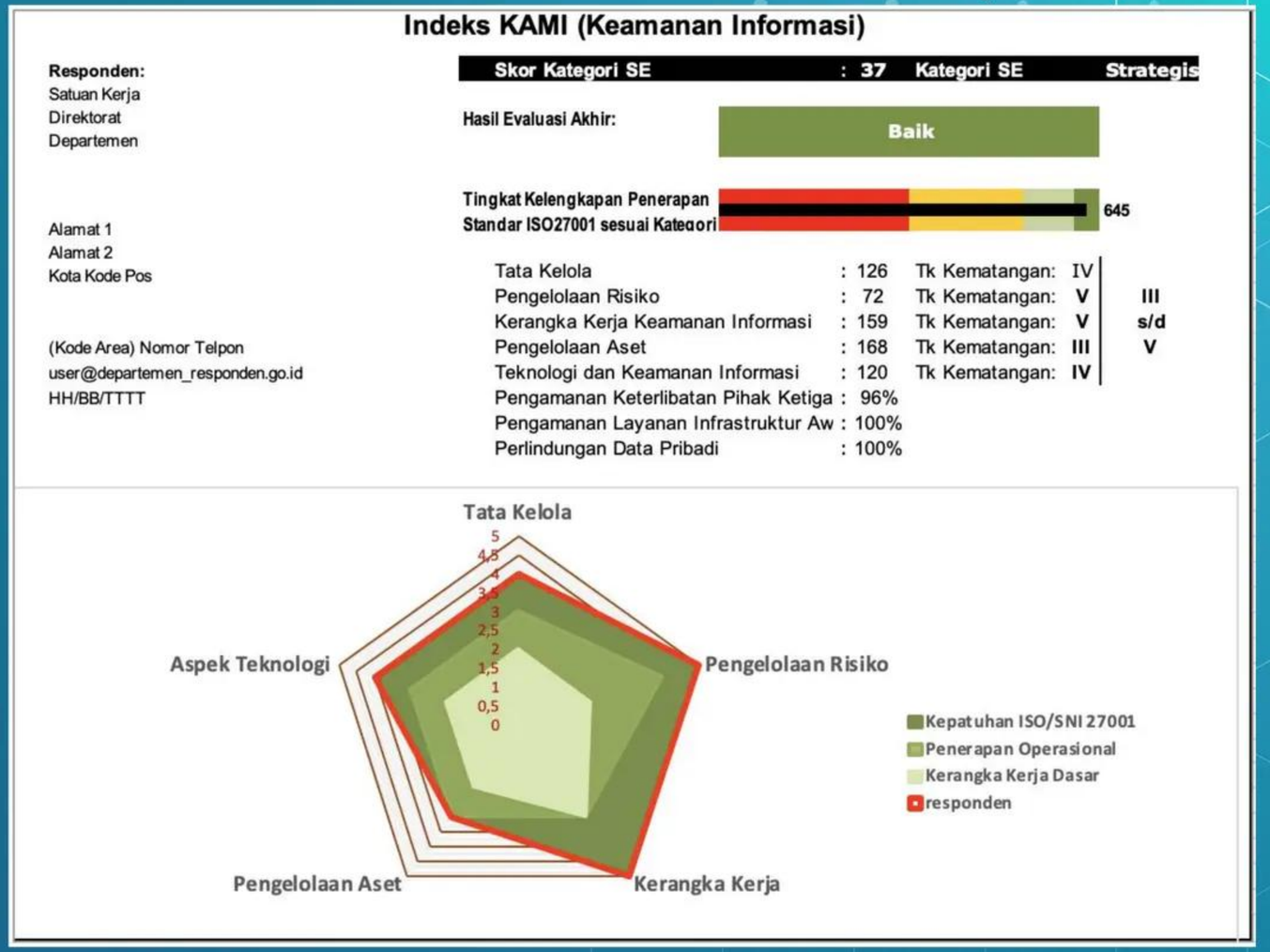
- UU PDP
- Index KAMI

Pasal 3

(1) Data Pribadi terdiri atas:
   a. Data Pribadi yang bersifat umum; dan
   b. Data Pribadi yang bersifat spesifik.

(2) Data Pribadi yang bersifat umum sebagaimana dimaksud pada ayat (1) huruf a meliputi:
   a. nama lengkap;
   b. jenis kelamin;
   c. kewarganegaraan;
   d. agama; dan/atau
   e. Data Pribadi yang dikombinasikan untuk mengidentifikasi seseorang.

(3) Data Pribadi yang bersifat spesifik sebagaimana dimaksud pada ayat (1) huruf b meliputi:
   a. data dan informasi kesehatan;



Conclusion

To build data or AI solution or engine or system or product is not just need the knowledge of the science of data or AI, even more than just engineering. Science will help us to understand the nature, the cases, engineering will help us to build the best tech solution, and regulation will keep it controlled that will make the system safe, secure, trusted, and relliable.

THANKS!

ANY QUESTIONS?

You can find me at:

- situkangsayur@gmail.com
- **hendri.karisma@akarintidata.ai**
- Telegram : siganteng