

## Guest speaker

What is a data scientist?

- A field that uses various tools and methods to extract knowledge and insights from data.
- Combines skills from statistics, computer science, business, and communication.

Key skills:

- Technical: Statistical analysis, data management, machine learning, data visualization, big data technologies.

Soft skills: Problem-solving, communication, business acumen.

Process:

### 1. Data Preparation:

- Collect data from various sources.
- Clean and transform the data.
- Store it in a data warehouse or mart.
- Use data visualization tools to explore the data.

### 2. Analyzing Data:

- Conduct exploratory data analysis (EDA) to understand patterns and trends.
- Build models to make predictions or classifications.
- Collaborate with stakeholders to define project goals and communicate findings.

### 3. Driving Innovation:

- Apply data insights to solve real-world problems across various industries, such as agriculture, manufacturing, and environmental management.
- Explore cutting-edge technologies like Generative AI for business applications.

Generative AI for Retail Business

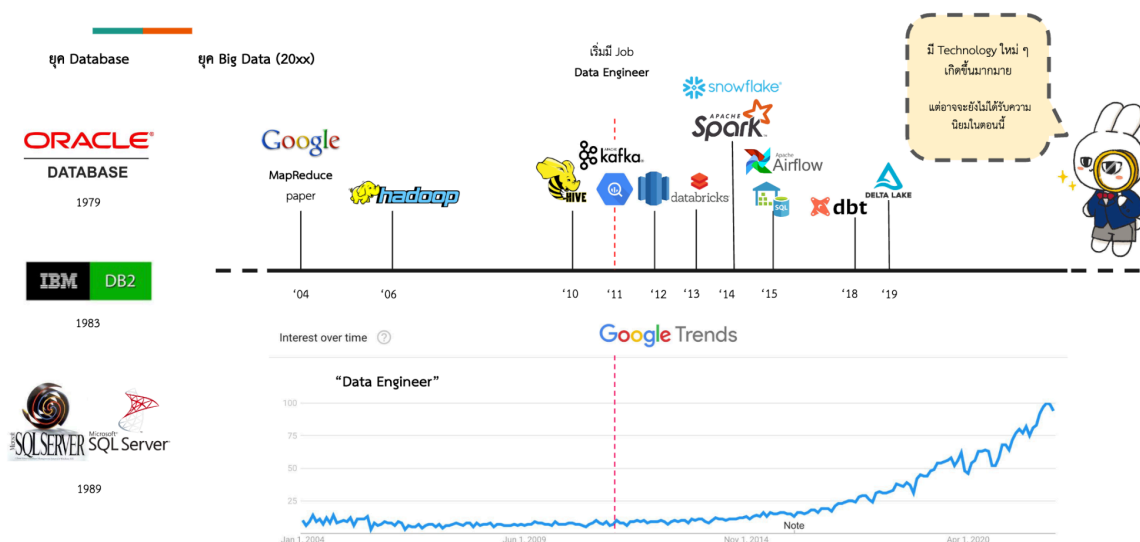
- ใช้ปิดการขาย เช่น Retail ขายน้ำตาล
- สามารถตอบได้ละเอียดมากกว่า
- เป็น Trend ที่สำคัญ จำเป็นต้องเก่งและฉลาดใช้



## Career Path in Data Science

- Data Analyst: Focuses on interpreting data, generating reports, and dashboards.
- Data Scientist: Engages in more complex problems and predictive analytics using advanced statistical methods and machine learning.
- Machine Learning Engineer: Specializes in building and deploying ML models and systems.
- Data Engineer: Concentrates on the architecture of data systems and managing data workflows.
- Must be always active all the time and learn to prompt and use it as a tool

## Data Engineer

- what is the difference between Model-centric and data-centric
  - model-centric: develop model development technique
  - data-centric: A lot of data
- Where does the data come from?
  - Applications, smart gadgets, IoT, etc.
- Application
  - The database can handle a limited amount of workload. Fix by separate workload
- Brief History of Data Engineer



- Create and look after Data Infrastructure
  - Database
    - MySQL
    - PostgreSQL
    - MongoDB
  - Data Lake
    - Hadoop HDFS
    - Amazon S3
    - Google Cloud Storage
    - Azure Blob
  - Data Warehouse
    - Amazon Redshift
    - Google BigQuery
    - Azure Synapse
    - Snowflake
    - Apache Hive
  - Cloud / on-premise Infrastructure
    - AWS
    - Google Cloud
  - Pokemon or Bigdata?
  - AI System = Software + Model + DATA
  - High-level of ML Systems
    1. Data
      - Data Engineering Pipelines
      - Data Quality is a key 
    2. Model
      - Machine Learning Pipelines
      - Train, Evaluate, Test 
    3. Software
      - Model Serving & Predictions
      - Deployment strategies & Infra
  - ML Engineer
    - Machine Learning Engineer คือ นักพัฒนาโปรแกรมที่วิจัย สร้าง และออกแบบ ซอฟต์แวร์ที่สามารถรันโมเดลทำนายผลได้ และสร้างระบบ AI โดยใช้ประโยชน์จากข้อมูลจำนวนมาก
  - DE vs DS vs MLE
    - DE concern about data
      - E.g. Data pipeline
    - DS concern about model
      - Stats,
      - Accuracy,

- Business value
- MLE concerned about the model of the product
  - Performance
  - Throughput
  - Automation
- MLE vs MLOps
  - MLOps: ideate from DevOps and the subset of MLE
  - Loop of Data → Model → Deploy → Data and go on
- “To make great products: do machine learning like the great engineer you are, not like the great machine learning expert you aren’t”

#### OpenThaiGPT LLM for Thai (ChatGPT in the Thai language)

- by NECTEC, AiAT, EAT
- ChatGPT translating Thai to English may cause meaning loss and the the sensitive data may leak (illegally)
- LLM
  - Generate the first word  $w_1$
  - And keep generating the next word  $w_k$  based on the previous words (a.k.a context)
- Transformer Model
  - Sequence-to-sequence generation
    - try to learn the order of word (subject + action + noun/ action + subject + noun / etc. )
    - focus the words. Try to compare between word and the place of the closet which is locked by a key. The key is the word trying to compare the other keyhole (query) (try to compare every word in the word list and return the value of the cosine similarity (dot product of the vector of the word) of the word vector) which is like an idiom.
  - Application with this LLM
    - try to compare Thai words and English words
    - and repeat this to get many layer idiom (idiom of idiom of idiom) to get the sentence
  - Dataset from all web 37.3 billion (from 2 trillion) and 20 billion words from Pantip
- Limitation
  - May cause AI hallucination:
    - The answer does not match and does not relate to the question

- Generate speech accurately in the first section, then the generated word is not related to the question