



# ***Arsitektur dan Teknologi Berbasis Software Bebas untuk Membangun Data Lake***



**Dr. Bambang Purnomosidi D. P.**

MTI - STMIK Akakom

PT. Wabi Teknologi Indonesia

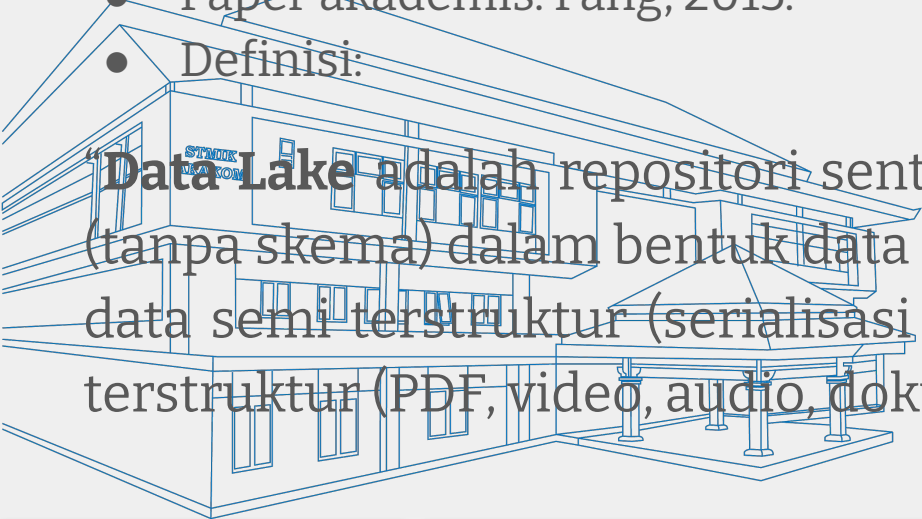
# Agenda

1. Memahami Data Lake
2. Arsitektur Data Lake
3. Teknologi Data Lake



# Memahami Data Lake

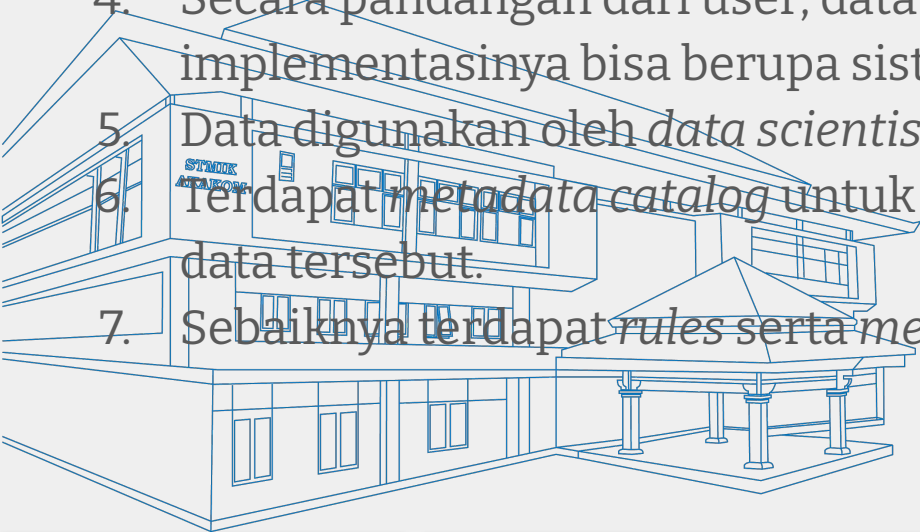
- Istilah **Data Lake** pertama kali dimunculkan oleh James Dixon - CTO Pentaho  
(<https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/>).
- Paper akademis: Fang, 2015.
- Definisi:



“**Data Lake** adalah repositori sentral untuk menyimpan data mentah (tanpa skema) dalam bentuk data terstruktur (tabel - baris dan kolom), data semi terstruktur (serialisasi JSON, XML, dll), maupun data tidak terstruktur (PDF, video, audio, dokumen, dll).”

(Madera C, et. al., 2017) mendefinisikan berbagai karakteristik Data Lake:

1. Tidak mempunyai skema
2. Memungkinkan untuk menyimpan semua format data
3. *Schema -on Read*: data merupakan data mentah - belum ditransformasikan
4. Secara pandangan dari user, data berada pada satu lokasi sentral, tetapi implementasinya bisa berupa sistem yang terdistribusi.
5. Data digunakan oleh *data scientist* maupun *data analyst*.
6. Terdapat *metadata catalog* untuk “menjelaskan” definisi data serta asosiasi data tersebut.
7. Sebaiknya terdapat *rules* serta *methods* untuk data governance.



Data Lake yang tidak dirancang dan diimplementasikan dengan baik akan menghasilkan:

1. Data Swamp
2. Data Hoarding



Data Lake berbeda dengan:

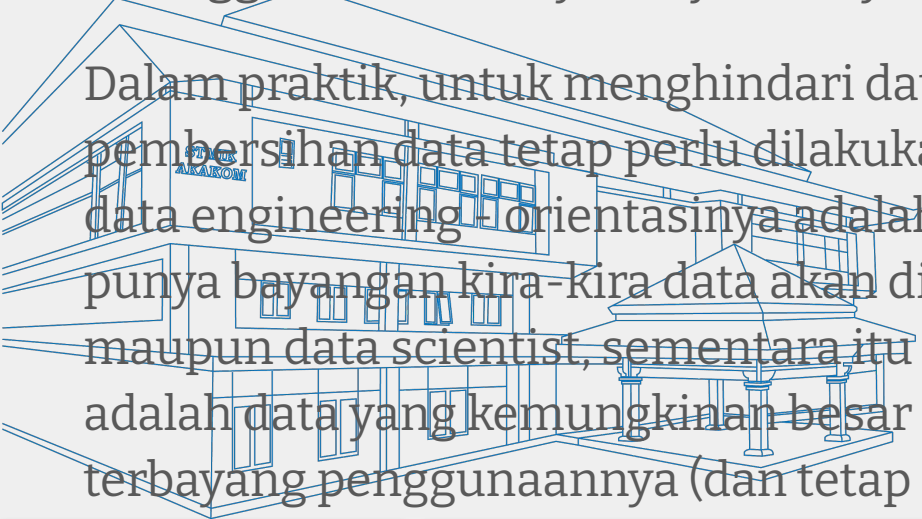
1. Data Warehouse
2. Data Mart



## Data Lake vs Data Engineering

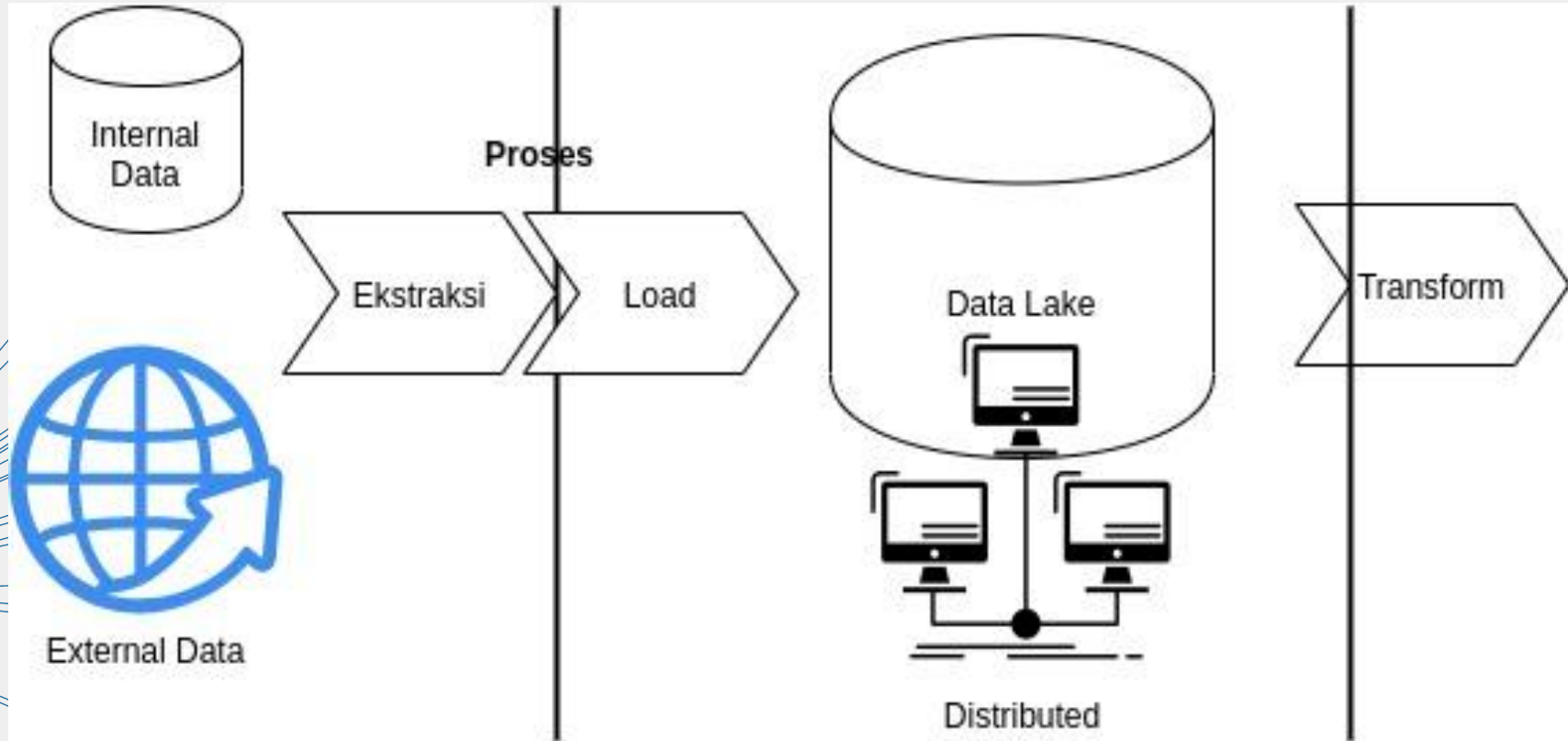
- Data Lake: ELT (Extract - Load - Transform)
- Data Engineering: ETL (Extract - Transform - Load)

Keduanya mempunyai user yang sama, yaitu data analyst dan data scientist sehingga data veracity menjadi isu yang sangat penting..



Dalam praktik, untuk menghindari data hoarding maupun data swamp, proses pembersihan data tetap perlu dilakukan untuk data lake. Perbedaannya, pada data engineering - orientasinya adalah tujuan penggunaan data, jadi sudah punya bayangan kira-kira data akan digunakan seperti apa oleh data analyst maupun data scientist, sementara itu pada data lake - data yang dimasukkan adalah data yang kemungkinan besar mempunyai value tetapi belum terbayang penggunaannya (dan tetap dibersihkan).

# Arsitektur Data Lake

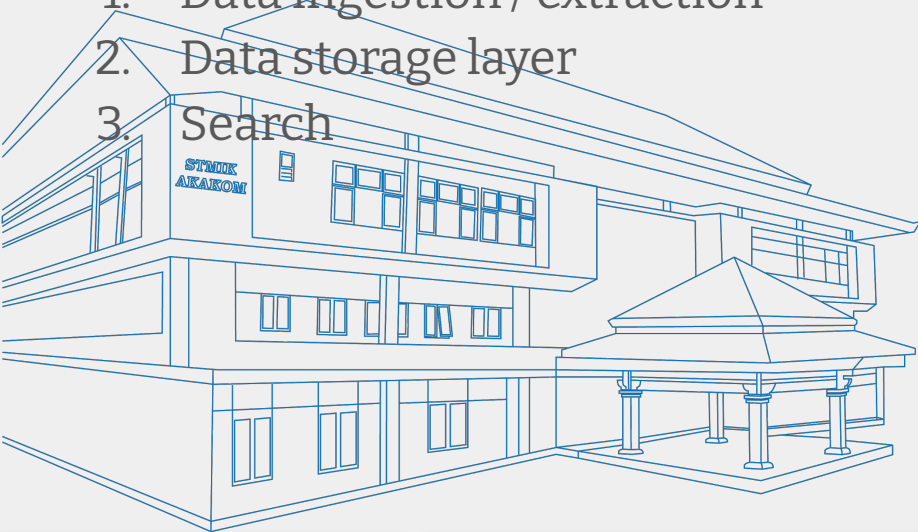




# Teknologi Data Lake

Ada beberapa komponen dari Data Lake, komponen ini akan menjadi acuan dari berbagai solusi software untuk implementasi Data Lake

1. Data ingestion / extraction
2. Data storage layer
3. Search

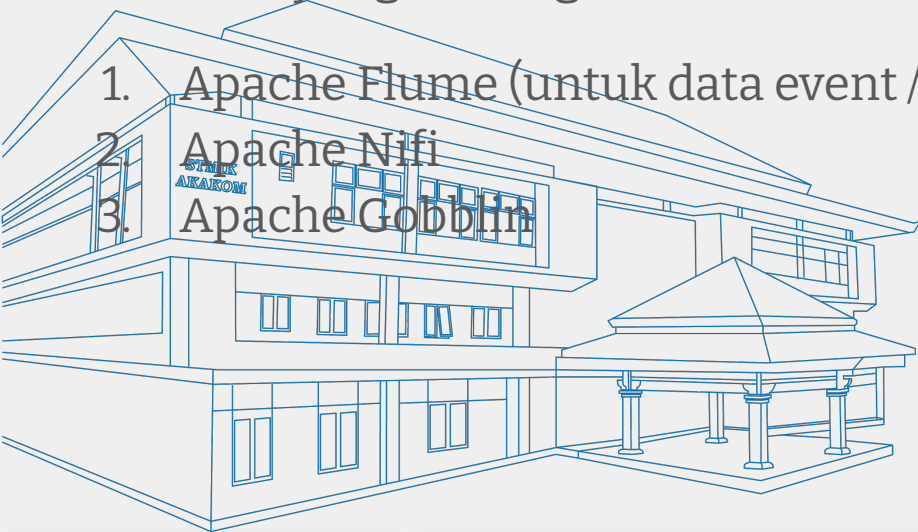


## Data Ingestion

Merupakan bagian yang digunakan untuk mengambil data dari berbagai sumber dengan berbagai format. Data bisa berupa data internal maupun eksternal. Data juga bisa dalam bentuk batch maupun realtime.

Software yang bisa digunakan:

1. Apache Flume (untuk data event / log)
2. Apache Nifi
3. Apache Gobbler



## Data Storage

Apache Hadoop menyediakan **HDFS** (Hadoop Distributed File System) sebagai basis dari media penyimpanan bagi:

- Apache Hive
- Apache HBase



## Data Search

Data yang sudah disimpan, diindex dengan menggunakan **Apache Solr**. Apache Solr menyediakan fasilitas untuk search data yang sudah tersimpan. Selain itu, Apache Solr juga menyediakan API yang bisa diakses menggunakan berbagai bahasa pemrograman



Terima kasih!

