Demystifying Data Engineering Terminologies

Unlock the secrets of data jargon

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA PIPELINE

A set of data processing elements connected in series, where the output of one element is the input of the next.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA WAREHOUSE

Structured repository optimized for analytics and reporting. Example: Snowflake storing transformed business data in dimensional models.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA MART

Subject-specific subset of a data warehouse. Example: Finance department's specialized analytical tables.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA ARCHIVING

Long-term storage of inactive data. Example: Moving 5+ year-old transactions to cold storage on tape or glacier storage.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA BACKUP

Copying data for recovery purposes. Example: Daily snapshots of database state for disaster recovery.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

OBJECT STORAGE

Storing data as objects with metadata and unique identifiers. Example: AWS S3, Azure Blob Storage, or Google Cloud Storage.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

METADATA MANAGEMENT

Organizing and maintaining data about other data. Example: Data catalog recording table schemas, lineage, and ownership.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

FULL LOAD

Replaces all target data with the entire source dataset. Example: Completely refreshing a data warehouse table with the latest data from source systems.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

INCREMENTAL LOAD

Adding only new or changed data to the target. Example: Loading only yesterday's transactions into a data warehouse.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

CDC (CHANGE DATA CAPTURE)

Tracking and capturing changes in source data. Example: Using database transaction logs to identify inserted, updated, or deleted records.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

ELT (EXTRACT, LOAD, TRANSFORM)

Pattern where data is loaded before transformation, leveraging target system processing. Example: Loading raw data into Snowflake and using SQL transformations within the platform.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

REAL-TIME ETL

Processing data as it arrives, in near real-time. Example: Streaming transactions from Kafka into a dashboard with sub-second latency.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

EVENT-DRIVEN ETL

Triggering ETL processes based on data events. Example: Initiating a data pipeline when a new file lands in S3.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

BATCH PROCESSING

Processing data in scheduled, discrete jobs. Example: Running daily aggregation jobs at midnight.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

MICRO-BATCH PROCESSING

Processing small batches of data at frequent intervals. Example: Running aggregations every 5 minutes on accumulated data.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA CLEANING

Removing or correcting inaccurate data. Example: Standardizing phone number formats or removing duplicate records.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA CONFORMING

Standardizing data to meet requirements. Example: Ensuring all date formats follow YYYY-MM-DD pattern.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA MASKING

Hiding sensitive data for privacy. Example: Replacing credit card numbers with XXX-XXX-XXXX format.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA VALIDATION

Verifying data accuracy. Example: Ensuring age values fall within a reasonable range (0-120).

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

DATA PROFILING

Analyzing data for structure and quality. Example: Examining the distribution of values in a column to identify outliers.

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/

Let's Chat!

What other data engineering terms do you find intriguing? Share in the comments!

POOJA JAIN
@https://www.linkedin.com/in/pooja-jain-898253106/