A black text with black letters

AI-generated content may be incorrect.

Extracting data from various data sources.

3 V ‘s of big data

1.Variety

2.Velocity

3.Volume

When we design data architecture it is critical to think about 3 V’s to be

Reliable, scalable, and maintainable systems.

Our collected data will be loaded into staging.

A black text on a white background

AI-generated content may be incorrect.

Duplicate data

In consistent values (ex: zip codes)

User left out field, we ended up missing field values.

Typing mistakes

The Goal of Transform phase is.

Converting data from an operational source format into data warehouse format.

A black and white logo

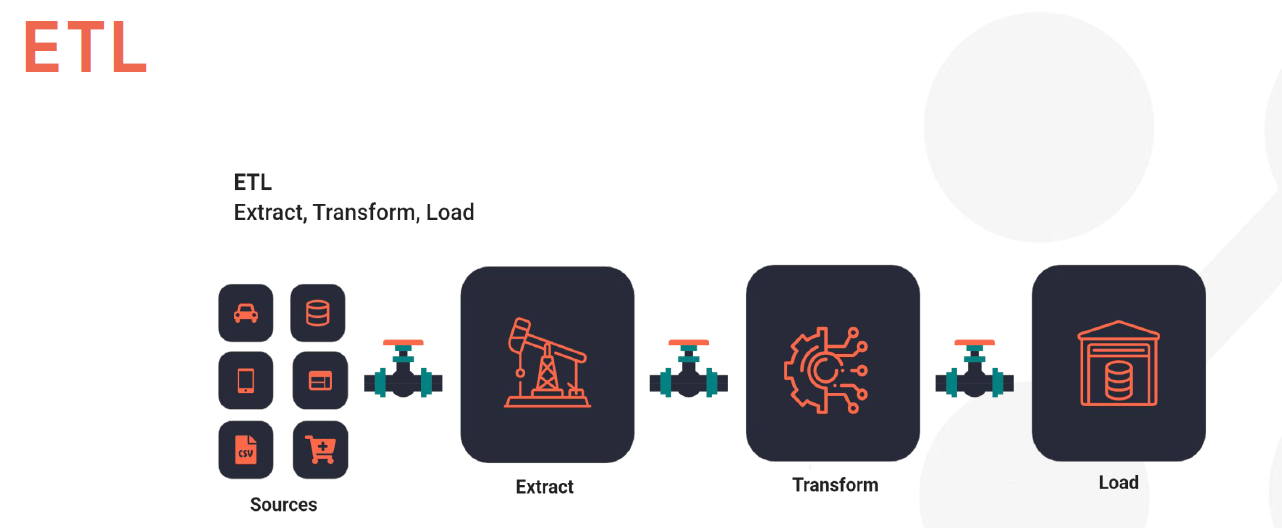
AI-generated content may be incorrect.

Data integration is nothing but writing out transformed data from staging to Target database.

We will load data into 2 ways.

1.Complete refresh

2.Incremental loading



**Issues in ETL**

If there is a change in schema the model Is broken.

If we think about data velocity changes over time.so scaling is not a easy task.

A diagram of a process

AI-generated content may be incorrect.

Datawarehouse is like snowflake, Redshift and Big query are extremely scalable and performant.

A close up of a logo

AI-generated content may be incorrect.

The Purpose of data warehouse is Data Analytics and Reporting.

A close-up of a data warehouse

AI-generated content may be incorrect.

If we have an Onprem data warehouse we have complete control over the setup.

So, we can decide the size of the storage, computing power and memory.

We can control both software and hardware.

On-prem data warehouse would be harder to scale as we need to purchase a new set of computers.

So these are reasons to move to the cloud data warehouse.

A screenshot of a computer

AI-generated content may be incorrect.

We have the option to store large files outside of the data warehouse (ex: S3,Blob storage)

A close-up of a book

AI-generated content may be incorrect.

It’s a repository where we can put all kinds of unstructured data, semi structured, and so on.

It is a very scalable file system.

On prem it is called HDFS (Hadoop distributed File System)

Cloud – Amazon S3, ADLS, BLOB containers, GCS.

A blue and white diagram

AI-generated content may be incorrect.

It’s combined with the best features of data lake and Lakehouse.

It sists on top of the low-cost cloud storage.

In Lakehouse cost efficiency storage provided by cloud provider and the ACID Transactional support to ensure consistency.

**Evaluation of Modern Data Stock**

When it comes to older days

Storage cost

**Year size Cost**

**1967 1 MB 1,00,000**

**1981 1 GB 3,00,000**

**1994 1 GB 1000**

**2000 1 GB 10**

**ETL Process was introduced in 1970’s**

**So Storage prices continued to drop.**

**Transistor prices continue to drop.**

**Faster data transmission over the network.**

**Shift from row-oriented data store to column-oriented data store.**

**SCD’s**

**SCD Type 0 – Not updating the DWH table when Dimension changes.**

**SCD Type 1 - updating the DWH table when Dimension changes, Overwriting the original data.**

**SCD Type 2 – Keeping full history – adding additional (historical) rows for each dimension changes.**

**SCD Type 3 – keeping limited data history – adding separate column for original and current each value.**

**A white text on a red background

AI-generated content may be incorrect.**

**dbt does not load or extract from the stage, but it transforms data.**

**dbt will deploy our analytics code following software engineering best practices like**

**modularity, portability and CI /CD, Testing and Documentation.**

**With dbt we can build great data pipelines.**

**We will write our code with dbt but dbt will compile with sql and executor data transformation jobs with snowflake or other data ware houses.**

**Transformations are Version controlled ,easily evaluated and at the top of the we automatically build DAGS.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A white text with orange and black text

AI-generated content may be incorrect.**

**A close-up of a logo

AI-generated content may be incorrect.**

**Preset is a BI tool to connect to snowflake Data ware house.**

**A white background with text

AI-generated content may be incorrect.**

**A white background with black text

AI-generated content may be incorrect.**

**Snowflake commands**

**USE ROLE ACCOUNTADMIN;**

We make sure we use super user for snowflake , so we have full access to all the permission management.

**Create the role for dbt**

**-- Create the `transform` role**

**CREATE ROLE IF NOT EXISTS TRANSFORM;**

**GRANT ROLE TRANSFORM TO ROLE ACCOUNTADMIN;**

**Note: we can think role is like a Group.**

**Snowflake User Creation**

**-- Use an admin role**

**USE ROLE ACCOUNTADMIN;**

**-- Create the `transform` role**

**CREATE ROLE IF NOT EXISTS TRANSFORM;**

**GRANT ROLE TRANSFORM TO ROLE ACCOUNTADMIN;**

**-- Create the default warehouse if necessary**

**CREATE WAREHOUSE IF NOT EXISTS COMPUTE\_WH;**

**GRANT OPERATE ON WAREHOUSE COMPUTE\_WH TO ROLE TRANSFORM;**

**-- Create the `dbt` user and assign to role**

**CREATE USER IF NOT EXISTS dbt**

**PASSWORD='dbtPassword123'**

**LOGIN\_NAME='dbt'**

**MUST\_CHANGE\_PASSWORD=FALSE**

**DEFAULT\_WAREHOUSE='COMPUTE\_WH'**

**DEFAULT\_ROLE=TRANSFORM**

**DEFAULT\_NAMESPACE='AIRBNB.RAW'**

**COMMENT='DBT user used for data transformation';**

**ALTER USER dbt SET TYPE = LEGACY\_SERVICE;**

**GRANT ROLE TRANSFORM to USER dbt;**

**-- Create our database and schemas**

**CREATE DATABASE IF NOT EXISTS AIRBNB;**

**CREATE SCHEMA IF NOT EXISTS AIRBNB.RAW;**

**-- Set up permissions to role `transform`**

**GRANT ALL ON WAREHOUSE COMPUTE\_WH TO ROLE TRANSFORM;**

**GRANT ALL ON DATABASE AIRBNB to ROLE TRANSFORM;**

**GRANT ALL ON ALL SCHEMAS IN DATABASE AIRBNB to ROLE TRANSFORM;**

**GRANT ALL ON FUTURE SCHEMAS IN DATABASE AIRBNB to ROLE TRANSFORM;**

**GRANT ALL ON ALL TABLES IN SCHEMA AIRBNB.RAW to ROLE TRANSFORM;**

**GRANT ALL ON FUTURE TABLES IN SCHEMA AIRBNB.RAW to ROLE TRANSFORM;**