EPAM’s Snowflake Hands-on Lab

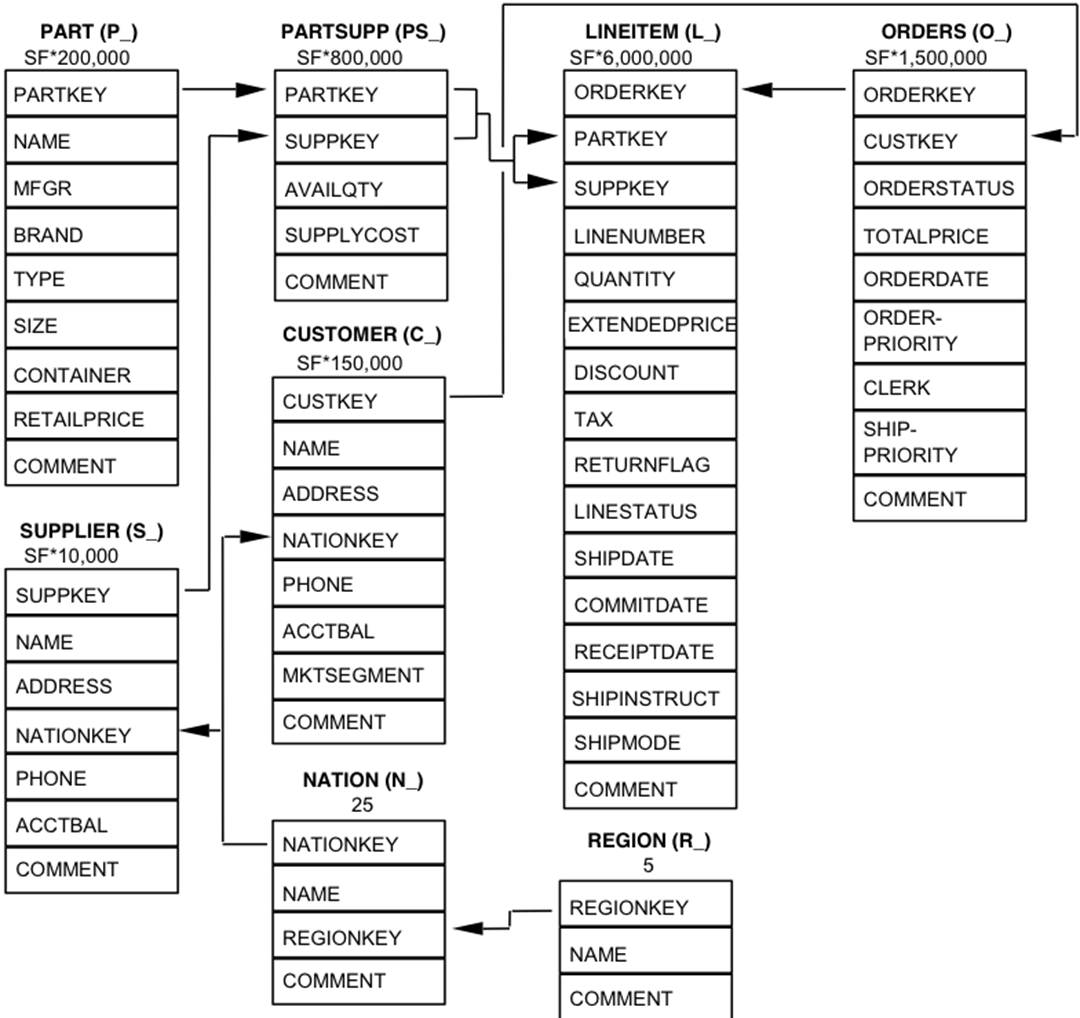
# Lab Overview

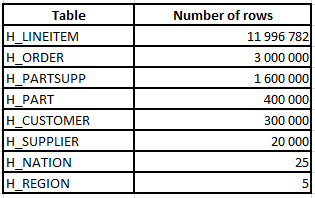
This Lab (prepared by your EPAM colleagues) offers a high-level description of the practical task for self-directed learning.

The target group for the Lab are DWBI engineers with experience in building Data Warehouses using other databases (Oracle, MS SQL, Teradata, etc.).

# Lab Data Set

Data set from [TPC-H benchmark](http://www.tpc.org/tpch/) is proposed for the Lab. TPC-H allows you to generate data for 8 tables. The data volume (in gigabytes) is defined by scale factor (SF). For the Lab purpose, you can [download](https://epam-my.sharepoint.com/:f:/p/maksim_krupenin/Ehu35G00GDVEt--YeQhrYc0BB63ueaqLIqU-zt4r9Dx4pQ?e=X12K0F) prepared in advance data set (2 GB of raw data, SF=2):





# In the shared folder, you can also find DDL script for the tables: *tpch\_ddl.sql*.

# Lab Description

Hands-on-lab is considered as completed if you score >= 60 points.

(Tasks 1, 8 – 5 points each, Tasks 2, 4, 5, 6 – 10 points each, Task 3 – 30 points, Task 7 – 20 points).

## Database creation

First, you need to create a separate database EPAM\_LAB in Snowflake.

## Data loading

In this step, you need to load Lab data set to an internal (Snowflake) or external stage. If you have an existing account in AWS/GCP/Azure cloud, an external stage would be preferable. Please note that you may need some data preparation steps before loading.

## ELT Data Workflow

Create two schemas in the DB you created before:

* CORE\_DWH
* DATA\_MART

Develop the following automated data workflow:

Stage -> CORE\_DWH -> DATA\_MART

Data in CORE\_DWH should be modeled according to 3NF (as is - no transformation). Star Schema is a target data model for DATA\_MART (data should be transformed accordingly).

The following Snowflake features should be used:

* Tasks
* Stored Procedures
* Tables Streams

*Note: No need to spend a lot of time on the modeling Star Schema implementing all nuances. You should rather focus on Snowflake’s capabilities.*

## Snowflake & 3rd party tools

When the data is loaded to DATA\_MART schema, connect Snowflake as a data source from any BI tool (Tableau, PowerBI, Qlik Sense, etc.) and create a simple dashboard.

Also, try connecting to Snowflake from any SQL editor (e.g. [DBeaver](https://dbeaver.io/)).

1. Snowflake SQL

From the shared folder you can also [download](https://epam-my.sharepoint.com/:f:/p/maksim_krupenin/Ehu35G00GDVEt--YeQhrYc0BB63ueaqLIqU-zt4r9Dx4pQ?e=X12K0F) the file with 22 TPC-H benchmarking queries (tpch\_benchmark\_queries.sql). Please note that the queries were modified to execute in AWS RedShift database, so some of them may require modifications for Snowflake. Use the queries to test how Snowflake works:

* Create several warehouses of different sizes and compare their performance.
* Test how Snowflake leverages different types of cash.
* Rewrite a couple of queries to execute on the Start Schema data model and compare performance (3NF vs Star Schema).
* Execute queries using SnowSQL (CLI Client).

## Other Snowflake features

Learn and test other interesting Snowflake features:

* Object Cloning
* Time Travel
* Data Sharing - share your DATA\_MART schema with a colleague who helps you with this Lab. Also, you may use a [Reader Account](https://docs.snowflake.com/en/user-guide/data-sharing-reader-create.html).

## Snowpipe

Automated incremental data loading using Snowpipe. Split lineitem & order files into several parts and simulate their sequential loading to stage buckets.

## Additional tasks

Connect your Snowflake account with partner applications available for a free trial (e.g. Fivetran, Periscope Data, Matillion in Partner Connect menu). Explore how selected tools work.