



Hands-on session preparation

For Windows Users: prepare WSL

- Install „Ubuntu 22.04 LTS“ from the Microsoft Store
- Search for „Windows-Features aktivieren oder deaktivieren“ in Windows Search Bar
 - Activate „Windows-Subsystem für Linux“ option
- The WSL data directory is: `\\wsl$`

Install required Packages

- On Linux and Windows with WSL (start Ubuntu 22.04 LTS - a terminal should open where you can run Linux commands)

```
sudo apt-get update && sudo apt-get upgrade -y  
sudo apt-get install git make gcc
```

- On MacOS

```
xcode-select --install
```

Download TPC-H data generator

- On Linux and Windows with WSL (start Ubuntu 22.04 LTS - a terminal should open where you can run Linux commands)

```
git clone https://github.com/gregrahn/tpch-kit.git  
cd tpch-kit/dbgen  
make MACHINE=LINUX DATABASE=POSTGRESQL
```

- On MacOS

```
git clone https://github.com/gregrahn/tpch-kit.git  
cd tpch-kit/dbgen  
make MACHINE=MACOS DATABASE=POSTGRESQL
```

Prepare TPC-H data

- Set DSS_PATH variable to specify where the generated data will be stored (directory has to exist already)

```
export DSS_PATH=/path/to/tpc-h/data/directory
```

- Generate TPC-H data with scale factor 1:

```
cd /path/to/tpch-kit/repo/dbgen  
./dbgen -s 1
```

- Check that your specified path contains 8 „*.tbl“ files

Install & build Spark

- Make sure you installed Java and the JAVA_HOME variable is set
- Download Spark from the official Spark website or clone the Github Repository

```
git clone https://github.com/apache/spark.git
```

- Navigate into the spark directory and run the build command:

```
./build/mvn -DskipTests clean package
```

- Try to start the Scala Spark Shell:

```
./bin/spark-shell
```

TPC-H Schema

