APACHE Spark™

Hands-on session preparation

Image-Source

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
```

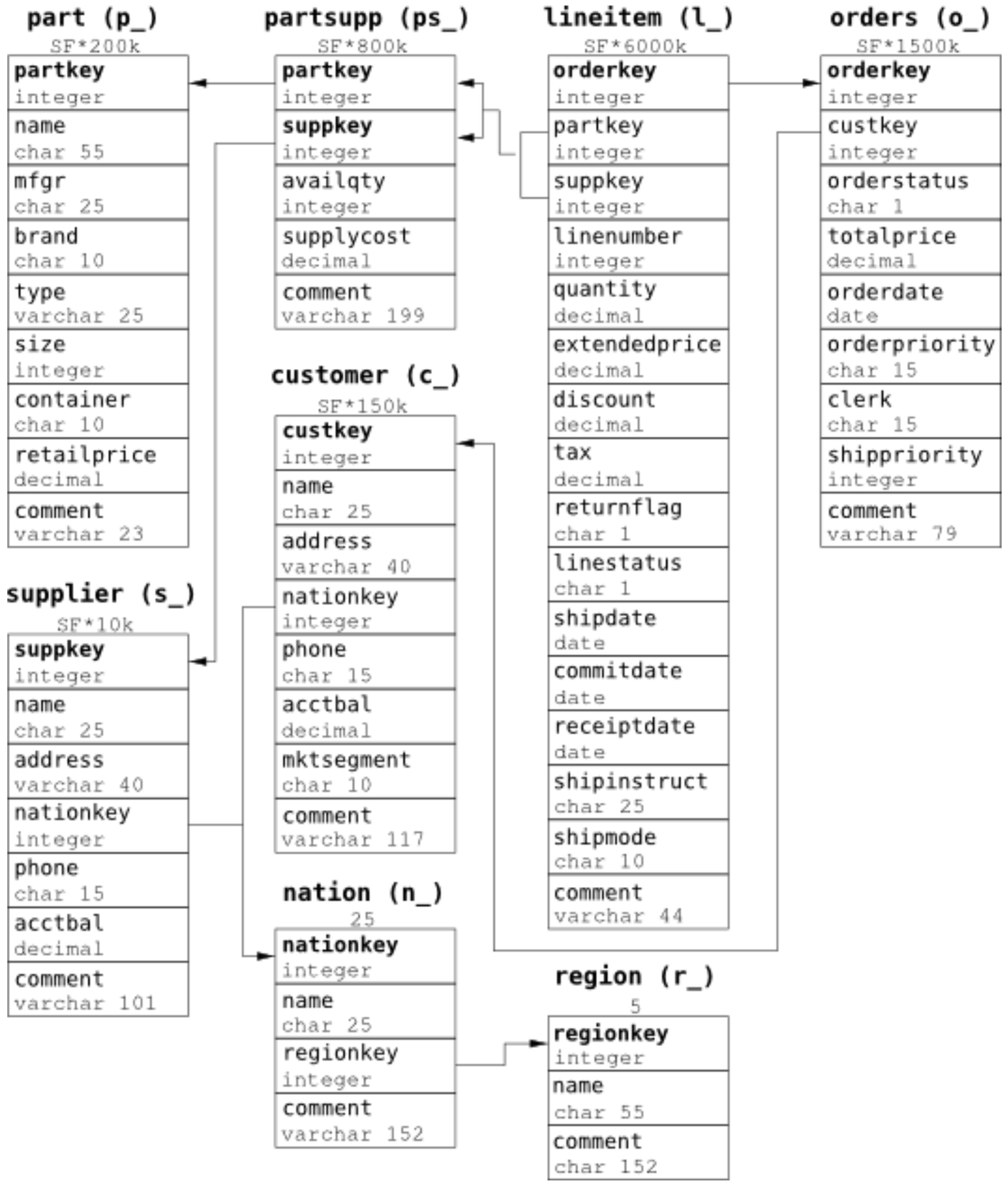
- Navigate into the spark directory and run the build command:

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

- Try to start the Scala Spark Shell:

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

TPC-H Schema


part (p_)
SF*200k
partkey integer
name char 55
mfgr char 25
brand char 10
type varchar 25
size integer
container char 10
retailprice decimal
comment varchar 23
partsupp (ps_)
SF*800k
partkey integer
suppkey integer
availqty integer
supplycost decimal
comment varchar 199
lineitem (l_)
SF*6000k
orderkey integer
partkey integer
suppkey integer
linenumber integer
quantity decimal
extendedprice decimal
discount decimal
tax decimal
returnflag char 1
linestatus char 1
shipdate date
commitdate date
receiptdate date
shipinstruct char 25
shipmode char 10
comment varchar 44
orders (o_)
SF*1500k
orderkey integer
custkey integer
orderstatus char 1
totalprice decimal
orderdate date
orderpriority char 15
clerk char 15
shippriority integer
comment varchar 79
customer (c_)
SF*150k
custkey integer
name char 25
address varchar 40
nationkey integer
phone char 15
acctbal decimal
mktsegment char 10
comment varchar 117
supplier (s_)
SF*10k
suppkey integer
name char 25
address varchar 40
nationkey integer
phone char 15
acctbal decimal
comment varchar 101
nation (n_)
25
nationkey integer
name char 25
regionkey integer
comment varchar 152
region (r_)
5
regionkey integer
name char 55
comment char 152