# Setting up the datasets :

* Goto the <https://github.com/dgadiraju/data>
* Clone or Download on virtual Machines created using Cloudera Quickstart or Hortonworks SandBox
* You can set up locally for practicing for Spark , but it is highly recommended to use HDFS which comes out of the box with CLoudera Quickstart or Hortonwords or our labs
* On lab they are already available
* Retail\_db

1. Master table

* Customers
* Products
* Categories
* Departments

1. Transaction tables

* Orders
* Orders\_items

# HIERARCHY OF TABLE:

Departments 🡨categories🡨producuts🡨orderitems🡨orders🡨customers.