# Description of Data

E-Commerce dataset is consisted of two txt files. First file is ‘os\_order.txt’ contains general information about orders between April 2010 to December 2011. The columns have buyer, date, location (ip) and status related information. Second file is ‘os\_order\_items.txt’ which contains more granular information about the details of every transaction. We can access individual item id’s in each transaction alongside with buying price and retail price with the quantities. Datasets are merged using order id’s when necessary through the analysis. Unofficial metadata is as follows:

**OS\_ORDER**

1. **ORDER\_ID:** Unique order id
2. **ORDER\_CODE:** 12 digit unique order code
3. **BUYER\_ID:** 5 digit unique number assigned to buyer
4. **PAY\_DT:** ‘yyyy-mm-dd HH:MM:SS’ formatted date for time of transaction
5. **CREATE\_IP:** IP Address where the order is placed from
6. **ORDER\_STATUS:** Status information takes values between 1-9, A, B, z

**OS\_ORDER\_ITEM**

1. **ITEM\_ID:** Unique index number for the purchased item
2. **ORDER\_ID:** Unique order id
3. **GOODS\_ID:** 7 digit unique item id of the purchased item
4. **GOODS\_NUMBER:** Quantity of the purchased item
5. **SHOP\_PRICE:** Retail price of the item
6. **GOODS\_PRICE:** Wholesale price of the item for the company
7. **GOODS\_AMOUNT:** Revenue generated from the transaction

What was difficult to understand?

In general, we had to work with lower level API’s in Hadoop Ecosystem. The complexity of the distribution system introduces additional challenges. We had difficulty understanding how map-reduce exactly works. Simple data transformation and ingestion tasks required significant effort from time to time. Lastly, integrating multiple technologies was a challenge but instructive task.

o What was too easy to understand?

Understanding working with the tools designed to provide conventional work flows was relatively easy. For example, Hive uses similar syntax to conventional SQL structure. Zeppelin uses the modern notebook style development interface, which is versatile and easy to follow. Accompanying, Hadoop Ecosystem and different tools with the Linux classes made it more comfortable working with several different tools using Terminal interface.

o Maybe something was not covered but you’d like it to be covered?

Managing Information Tech & Systems 5570 is in general a very successful hybrid course bringing management and IT aspects together. The Linux and Hadoop classes were instructive enough but still provided a lot of valuable hands-on experience.

o Any other comments are welcome