Data Processing :

Set up python

**Introduction**

* Python is an interpreter based programming language
* Adaptability of Python is very high in Data Engineering and Data  
  Science fields
* Spark APIs are well integrated with Python
* Highly relevant for Cloudera and Hortonworks Spark certifications

**Setup Python**

* Mac – pre-installed (if you need additional packages better to install  
  pip following instructions)
* Linux – pre-installed (if you need additional packages better to  
  install pip following instructions)
* Windows – Google and install using MSI
* [https://labs.itversity.com 4](https://labs.itversity.com/) – pre-installed
* For building applications, better to install IDE such as pycharm
* For certification preparation, practice using REPL

**REPL**

* REPL – Read, Evaluate, Print and Loop
* Use python command to launch REPL
* exit() or ctrl-D to come out
* Ctrl+l to clear the screen
* We can get the help using the help command
* Use up arrow to see the history

**Submitting as an Application**

* Let us create a directory structure for our application
* Launch web console by using your lab’s account
* Create a directory with the command mkdir **retail\_app**
* Go to **the retail\_app** folder using the command **cd retail\_app**
* Whenever we want to create an application we have to create in src/main folder so we have another subdirectory in a retail\_app directory. Use the command to create it mkdir **-p src/main/python**
* Go to the subdirectory using the command **cd src/main/python** . In this directory, we will create .py files to run the application
* Now create a file called **GetRevenueForOrderId** and paste the code.

C:\Users\lenovo\data\_science\Untitled Folder>python getRevenue\_parameters.py data\_folder/retail\_db/order\_items/part-00000 68880

999.77

data\_folder\retail\_db\order\_items

python getRevenue\_parameters.py data\_folder\retail\_db\order\_items/part-00000 2