Name: Sihang Wang

Course: DSCI 551

Assignment: Week 1 Summary

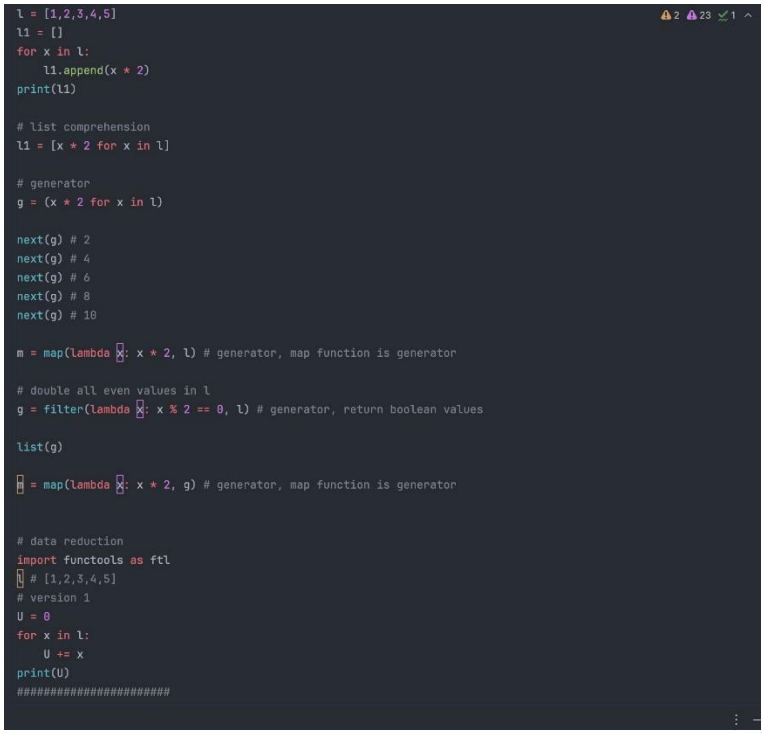
Date: January 18, 2025

Week 1 Summary

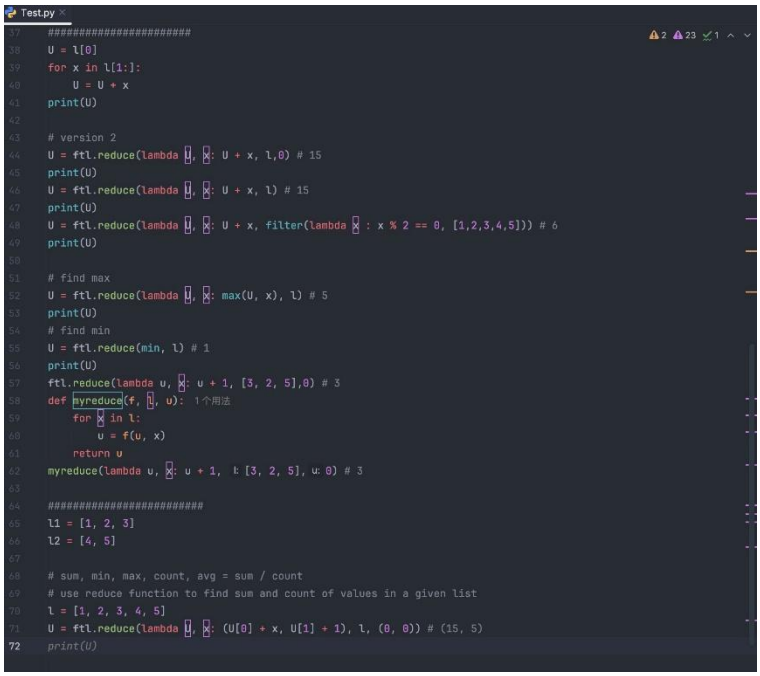
This is the first week of the semester, and the main content of the week is to discussing the syllabus and several introductions of topic. During the first lecture, the class basically talked about the syllabus and started with some basic knowledge of big data. During the second lecture, we mainly talked about four major topics which are storage system, file systems & file formats, database management systems (RDBMS), and big data solution stack. The main highlight that I made for storage systems was that the SSD, which contains 4KB block size for HDD and 128 MB block size in HDFS. For file systems & file formats, I listed highlights for inode which is identified by a number and can figure out location of inode from inumber. I understood that JSON, HTML, and XML are three mostly common used file formats and XML has lots of usage such as java archive and android app development. For database management systems, I have studied some of the knowledge in my undergraduate study, and it mainly talked about the query languages and query execution. The new things that I learned is that the goal of query optimization is to make declarative SQL query to imperative query execution plan. For big data solution stack, I learnt the knowledge mainly from the showcase of python in class. I could understand reducing function but still need to practice more because some python knowledge are not solid. I also learnt some terminal command such as nano py. I found that some topics like Hadoop and Spark were new to me, and I will try to study more about it.

There some technical knowledge that I learnt in this week, it mainly consists of python code map() and reduce() functions. I also know how to set up a virtual environment in AWS but it is not brand new to me because I used to use docker and AWS. For this week's progress, I think it is a good start to know what my weaknesses and what skills I need to study more. In the following week's summary, I will include my extra-study progress in coding and querying.

Python Showcase:



```python
l = [1,2,3,4,5]
l1 = []
for x in l:
    l1.append(x * 2)
print(l1)

# list comprehension
l1 = [x * 2 for x in l]

# generator
g = (x * 2 for x in l)

next(g) # 2
next(g) # 4
next(g) # 6
next(g) # 8
next(g) # 10

m = map(lambda x: x * 2, l) # generator, map function is generator

# double all even values in l
g = filter(lambda x: x % 2 == 0, l) # generator, return boolean values

list(g)

n = map(lambda x: x * 2, g) # generator, map function is generator


# data reduction
import functools as ftl
l # [1,2,3,4,5]
# version 1
U = 0
for x in l:
    U += x
print(U)
########################
```



```python
Test.py
    ######################
    U = l[0]
    for x in l[1:]:
        U = U + x
    print(U)

    # version 2
    U = ftl.reduce(lambda U, x: U + x, l,0) # 15
    print(U)
    U = ftl.reduce(lambda U, x: U + x, l) # 15
    print(U)
    U = ftl.reduce(lambda U, x: U + x, filter(lambda x : x % 2 == 0, [1,2,3,4,5])) # 6
    print(U)

    # find max
    U = ftl.reduce(lambda U, x: max(U, x), l) # 5
    print(U)
    # find min
    U = ftl.reduce(min, l) # 1
    print(U)
    ftl.reduce(lambda u, x: u + 1, [3, 2, 5],0) # 3
    def myreduce(f, l, u):  1个用法
        for x in l:
            u = f(u, x)
        return u
    myreduce(lambda u, x: u + 1,  l: [3, 2, 5], u: 0) # 3

    ########################
    l1 = [1, 2, 3]
    l2 = [4, 5]

    # sum, min, max, count, avg = sum / count
    # use reduce function to find sum and count of values in a given list
    l = [1, 2, 3, 4, 5]
    U = ftl.reduce(lambda U, x: (U[0] + x, U[1] + 1), l, (0, 0)) # (15, 5)
    print(U)
```

AWS showcase: