Exploring Data Science with Microsoft Materials

Zahra Ghausi, Brian Flores, and Nicholas Davies
(Advisor: Dr. Daehee Kim)
Computer Science, California State University Stanislaus

Abstract

Data Science is the process of collecting, storing, and analyzing data. It is important to learn about how files and data are stored, retrieved, and visualized in order to be able to extend our knowledge to the present Medical Big Data Analysis research project. Microsoft provides efficient materials, Data Science for beginners, that will focus on foundational concepts and practical applications of Data Science. This semester, we have been using the materials for learning Data Science with Github, VS Code, Jupyter Notebook, Cosmos DB emulator, Putty, Relational DB (MySQL), and NoSQL DB (MongoDB). The audience could get a better understanding of Data Science.

Introduction

Data Science is the process of collecting, storing, and analyzing data. Data Analysts use data to tell compelling stories to inform business decisions. We discovered how Data Science is defined and learned about Data Science. We have been using Microsoft materials [1] consisting of 20 lessons with tools such as VS Code and Jupyter Notebook this semester. The Microsoft materials provide resources like videos, pre and post quizzes, assignments, written instructions, software and directions to be able to work with different type of data like structured (table with rows and columns), unstructured, and semi-structure (JSON and XML), and libraries such as Pandas, Numpy, Scipy, and Matplotlib to manipulate data frame and to visualize data, and machining tools like Tensorflow.

Overview of Materials

- Introduction (including Probability): Lessons 1-4
- Working With Data (relational DB, NoSQL DB, Python): Lessons 5-8
- Data Visualization (Quantities, Distributions, Proportions, Relationships): Lessons 9-13
- Data Science Life Cycle (Data Acquisition, Data Analysis, Data Presentation): Lessons 14-16
- Data Science in the Cloud (ML model): Lessons 17-19
- Data Science in the Real World (Health Care, Transport, Banking & Financing, Sports, Research, Ecommerce): Lesson 20

Figures copied from [1]

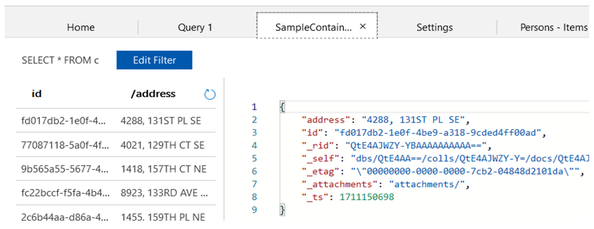
Lessons 1-4: Data

Lessons 1-4 covered over different data, like structured (e.g., table), semi-structured (e.g, JSON or XML), and unstructured (e.g., log file). We learned about different structures, like raw data, quantitative data, structured data, unstructured data and semi-structured data. It also went over how to create meaningful visualizations from data. The figure below shows an example of JSON (JavaScript Object Notation) format data. The example shows a JSON object (document) that consists of a key ("weather") and a value (an array with two JSON objects). An object in the array has four pairs each of which has a key (e.g., "degree") and a value (e.g., 75).

```
key                          value (array)
{"weather":[
    { "zip":95382, "degree":82, "city":"Turlock",  "state":"CA" },
    { "zip":94089, "degree":75, "city":"San Jose", "state":"CA" },
]                          pair (key:value)
}
```

Lessons 5-6: Database

Lessons 5-6 discussed about non-relational and relational database. An example of a non relational database we learned was MongoDB, which is a document NoSQL database. We used Azure Cosmos DB emulator to practice document NoSQL database by importing JSON files. We went over a relational database using SQL (Structure Query Language). We practiced tables through MySQL. We inserted data in tables, but also retrieved data from tables by using join base on a primary and a foreign key. The figure below shows the dashboard of the Cosmos DB emulator [2], which shows a document with multiple key value pairs.



Lessons 7-8: Python

Lessons 7-8 covered multiple libraries like Pandas, SciPy, Numpy, and Matplotlib. We manipulated DataFrames through Pandas and learned about image processing through Microsoft Azure Custom Vision. We also explored image classification where multiple pictures for an object were taken and analyzed. The figure below shows how to visualize the series (arrays) data with plot function.

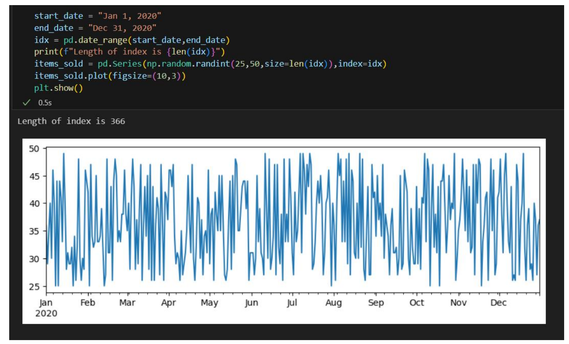


Figure copied from [1]

Present Research

Elastic High Speed Medical Big Data Analysis System to Discover Associations between Genetic Variants and Diseases: 10 Gbps network, 60TB disks, 10 TB medical dataset from MCRI (Marshfield Clinic Research Institute)

Acknowledgment

This research activity is funded all or in part by the Stanislaus State ASPIRE program through a U.S. Department of Education Title III Grant # P031C210159.

References

[1] Microsoft, "Data-Science -For-Beginners", https://github.com/microsoft/Data-Science-For-Beginners/tree/main
[2]Azure Cosmos DB emulator https://learn.microsoft.com/en-us/azure/cosmos-db/emulator