Analysis of the Large Volumed Data Files UsingPython Pandas and Snowflake Query

Dr. Anita Mahajan[1], Ajay Varma[2]
[1,2]Acropolis Institute of Technology & Research, Indore

Abstract-**Due to increasing volume of the data that is being utilized every day, the efforts required to make the data into a proper presentable format is very tedious and time taking task. Since, the latest databases that are currently widely used by the various organizations do provide the integration facilities with the various cloud-based AWS technologies. However, since this kind of operations requires continuouscommunication with the database server and the AWS server. Also, in case of the data fetching from the staging area through the cloud-based databases, applying the file formatting options and the filtering the rows from the CSV file using various conditions on various columns needs time. That's why querying the data from the staging files somehow fails in this scenario.**

Keywords: **AWS, cloud-based databases, CSV, Database server.**

INTRODUCTION

In current scenario, the analysis of data has been a most important and the effective way of decision making for small and big size organizations. Infect, currently its like the data is being coming from so many different heterogenous kinds of sources, that combining them into a single target and make insights is a tedious task. We do have many databases like Postgres, MongoDB and so on for these purposes, but to check the validity of the data and the correctness of the data from source to target, it becomes very necessary to go deep-dive into the source files (either large volume CSV or database exports). Also, for the business perspective, if they want to do the analysis of the data on a higher level, then it becomes a task that takes a lot of time for the comparison and analysis to be done. Thisdelay caused due to the analysis and checking of the data may result in good amounts of losses for the companies. Due to this reason, it is very important to get the data analysed and verified in the best possible optimized way in order to save time and make more and more profit. Also, one more factor that needs to be considered is that the language or method we are using for analysis of large business files should be as simple as normal language that we speak. So, that if any new person gets to understand the working of the entire validation, he/she could easily understand and implement any changes if required.

WHY PYTHON?

Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
Python has a simple syntax similar to the English language.
Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
Python runs on an interpreter system, meaning that code can be executed as soon as it iswritten. This means that prototyping can be very quick.
Python can be treated in a procedural way, an object-oriented way or a functional way.

DATA ANALYSIS USING PYTHON

For many people (myself among them), the Python language is easy to fall in love with. Since its first
For people including myself, the easy language to fall in love with is Python. Since its first Look in 1991, The most dynamic , programming languages, along with Perl, Ruby, and others is the Python. Python and Ruby have become most famous in current year. It is used for building websites by using multipleweb frameworks, like Django(Python) and Rails (Ruby). Those languages are called scripting languages.This helps in write quick-and-dirty small programs, or scripts. I hate term "scripting language " because it has connotation that we cannot use them for building mission-critical software. Python is different from other interpreted languages because of its large and active scientific

computing community. Since the early 2000s from the adoption of Python for scientific computing in both industry applications and academic research has gradually increased.

ESSENTIAL PYTHON LIBRARIES

Those who are less familiar with the Python language ecosystem and throughout the book, which is used in libraries, Here is the overview of every library.

NumPy

NumPy, short for Numerical Python, for the scientific computing of Python NumPy is the foundational package. The majority of this book will be based on NumPy and libraries built on top of NumPy. It provides, along with other things: It has a quick and structured multidimensional array object ndarray. Functions for mathematical operations between arrays or performing element-wise computations with arrays Tools for reading and writing array-based data sets to disk Linear algebraoperations, Fourier transform, and random number generation Tools for integrating connecting C, C++, and Fortran code to Python Beyond the fast array-processing capabilities that NumPy adds to Fortran code to Python Beyond the fast array-processing capabilities that NumPy adds to Python, Theprimary container for data to be passed between algorithms is one of its primary purposes with regards to data analysis .For numerical data, the much more efficient way of storing and manipulating data than the other built-in Python data structures are NumPy arrays. The data stored in a NumPy array can be operate by without copying any data in libraries written in a lower-level language, such as C or Fortran.

PANDAS

Structured data fast, easy and expressive to make working with Pandas which provides rich data structures and functions designed. A powerful and productive data analysis environment as you will see, is one of the critical ingredients enabling Python.

DATA FRAMES

A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. A Data frame is 2-D data structure, i.e., when the data is aligned in a tabular fashion in columns and rows.

Features of DataFrame

There are different types of Potentially columns

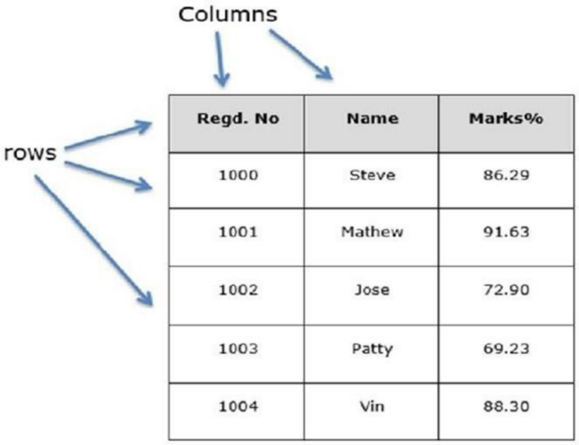
Size – Mutable

Labelled axes (columns and rows)

Arithmetic operations can be performed on columns and rows

STRUCTURE OF DATAFRAMES

Let us assume that we are creating a data frame with student’s data.



Regd. No	Name	Marks%
1000	Steve	86.29
1001	Mathew	91.63
1002	Jose	72.90
1003	Patty	69.23
1004	Vin	88.30

Creation of Panda’s Data frame

Creation of a dataframe in python is being created by using a constructor: pandas.DataFrame(data, index, columns, dtype, copy)

Following is the parameters of conductors-

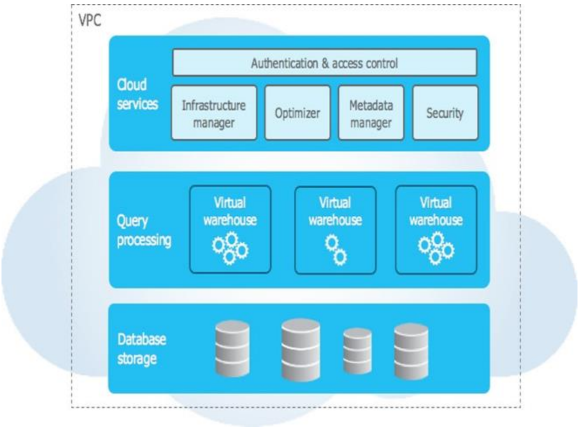
Sr.No	Parameter and discription
1	DATA-data takes various forms like ndarray, series, map, lists, dict, constants and alsoanother DataFrame.
2	INDEX-For the row labels, the Index to be used for the resulting frame is OptionalDefault np.arange(n) if no index is passed.
3	COLUMNS-For column labels, the optional default syntax is - np.arange(n). This is only true if no index is passed.
4	DTYPE-Data type of each column.
5	COPY-This command (or whatever it is) is used for copying of data, if the default is False.

SNOWFLAKE DATABASE

Snowflake is the one of the most prominent cloud-based databases that has been used in the current scenarios. The main concept of the snowflake database is the Software-As-A-Service.

Data Warehouse as a Service
No need of any software/hardware for installation, conifugration or selection.
There is no software required for installation.
Ongoing maintained, management and tuning is managed by Snowflake.

Snowflake Architecture



The main components of the Snowflake Architecture:
Data storage
Query Processing
Cloud Services

DATA STORAGE

When data is loaded into Snowflake, the data into its internal optimized, compressed, columnar format is reorganizes by Snowflake. This optimized data is stored in cloud storage by Snowflake.
All aspects of how this data is stored is managed by Snowflake - the organization, structure,compression, metadata, statistics, file size, and others aspects of data storage are managed by Snowflake. The data objects are not directly visible nor accessible by customers which is stored bySnowflake; they are only accessible through SQL query operations run using Snowflake.

QUERY PROCESSING

Query execution is performed in the processing layer. Queries using “virtual warehouses “is processed by Snowflake. The MPP compute cluster composed of multiple compute nodes allocated by Snowflake from a cloud provider by each virtual warehouse. An independent compute cluster that does not share compute resources with other virtual warehouses is the virtual warehouses. As a result, their is no impact on virtual warehouses and onperformance of other virtual warehouses

CLOUD SERVICES

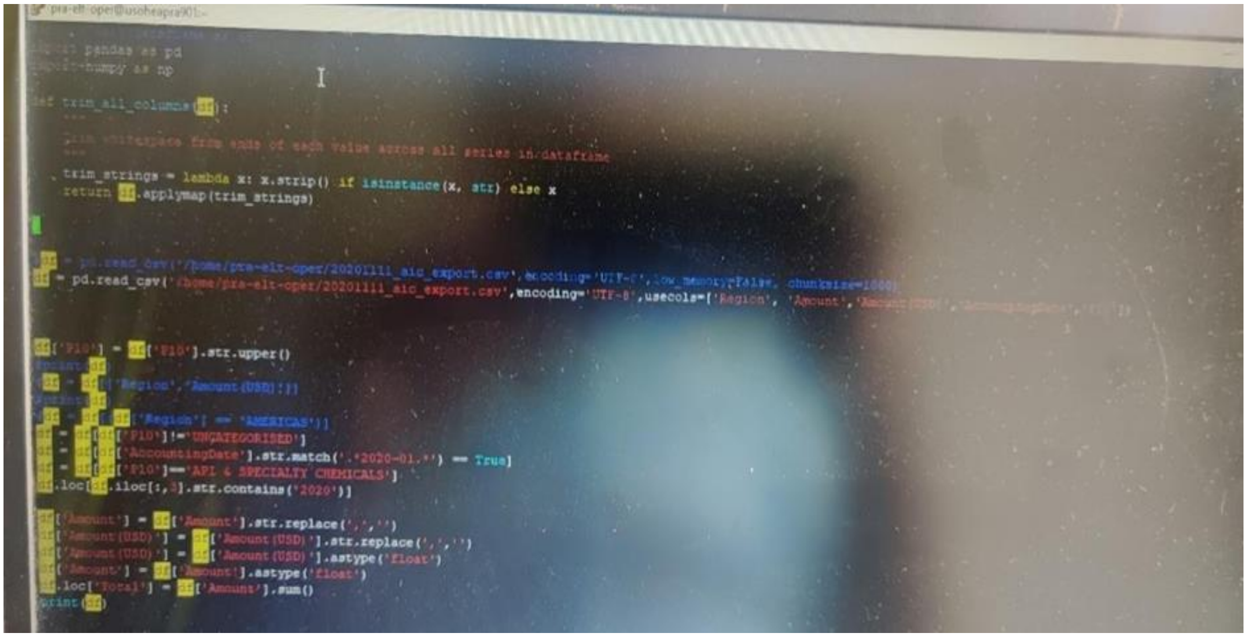
A collection of services that coordinate activities across Snowflake is the cloud services layer. The different components of Snowflake in order to process user requests, from login to query dispatch is all this services tie together.
From the cloud provider the cloud services layer also runs on compute instances provisioned bySnowflake.

EXPERIMENTAL SETUP

For the experimental purpose, we took a CSV file uploaded on S3 bucket of around 8GB, with almost 3 crores records, along with 150 columns. The main purpose of the experiment was to check the performance of the data filtering queries that were being executed, using Snowflake database and the same query using the Python Pandas code that has been written.
We first determined the various test cases that we need to check with both scenarios using the Snowflake database and the Pandas code.
Setup:
Number of records: 8 crores
Number of columns: 133
Conditions: 5

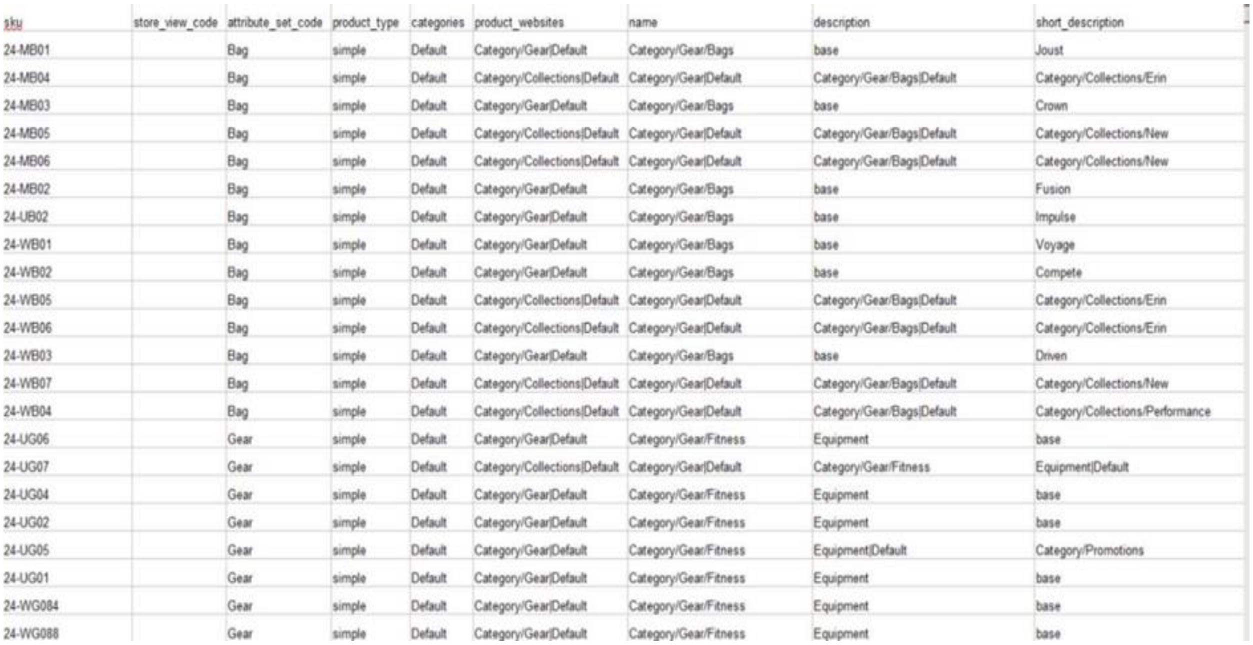
Query	Execution Time (Using Snowflake)	Execution Time (Using Pandas Code)
Data filtering using the regular expressions and applying the aggregate functions.	8-10 mins	1 mins

Code Snippet:



```
pra-elt-oper@usoheapra901:~

import pandas as pd
import numpy as np

def trim_all_columns(df):
    """
    Trim whitespace from ends of each value across all series in dataframe
    """
    trim_strings = lambda x: x.strip() if isinstance(x, str) else x
    return df.applymap(trim_strings)


#df = pd.read_csv('/home/pra-elt-oper/20201111_aic_export.csv',encoding='UTF-8',low_memory=False, chunksize=1000)
df = pd.read_csv('/home/pra-elt-oper/20201111_aic_export.csv',encoding='UTF-8',usecols=['Region', 'Amount','Amount(USD)', [illegible]])


df['F10'] = df['F10'].str.upper()
#print(df)
#df = df[['Region','Amount(USD)']]
#print(df)
#df = df[df['Region'] == 'AMERICAS']]
df = df[df['F10']!='UNCATEGORISED']
df = df[df['AccountingDate'].str.match('.*2020-01.*') == True]
df = df[df['F10']=='API & SPECIALTY CHEMICALS']
df.loc[df.iloc[:,3].str.contains('2020')]

df['Amount'] = df['Amount'].str.replace(',','')
df['Amount(USD)'] = df['Amount(USD)'].str.replace(',','')
df['Amount(USD)'] = df['Amount(USD)'].astype('float')
df['Amount'] = df['Amount'].astype('float')
df.loc['Total'] = df['Amount'].sum()
print(df)
```

Sample of the CSV file that is being used for validation



sku	store_view_code	attribute_set_code	product_type	categories	product_websites	name	description	short_description
24-MB01		Bag	simple	Default	Category/Gear/Default	Category/Gear/Bags	base	Joust
24-MB04		Bag	simple	Default	Category/Collections/Default	Category/Gear/Default	Category/Gear/Bags/Default	Category/Collections/Erin
24-MB03		Bag	simple	Default	Category/Gear/Default	Category/Gear/Bags	base	Crown
24-MB05		Bag	simple	Default	Category/Collections/Default	Category/Gear/Default	Category/Gear/Bags/Default	Category/Collections/New
24-MB06		Bag	simple	Default	Category/Collections/Default	Category/Gear/Default	Category/Gear/Bags/Default	Category/Collections/New
24-MB02		Bag	simple	Default	Category/Gear/Default	Category/Gear/Bags	base	Fusion
24-UB02		Bag	simple	Default	Category/Gear/Default	Category/Gear/Bags	base	Impulse
24-WB01		Bag	simple	Default	Category/Gear/Default	Category/Gear/Bags	base	Voyage
24-WB02		Bag	simple	Default	Category/Gear/Default	Category/Gear/Bags	base	Compete
24-WB05		Bag	simple	Default	Category/Collections/Default	Category/Gear/Default	Category/Gear/Bags/Default	Category/Collections/Erin
24-WB06		Bag	simple	Default	Category/Collections/Default	Category/Gear/Default	Category/Gear/Bags/Default	Category/Collections/Erin
24-WB03		Bag	simple	Default	Category/Gear/Default	Category/Gear/Bags	base	Driven
24-WB07		Bag	simple	Default	Category/Collections/Default	Category/Gear/Default	Category/Gear/Bags/Default	Category/Collections/New
24-WB04		Bag	simple	Default	Category/Collections/Default	Category/Gear/Default	Category/Gear/Bags/Default	Category/Collections/Performance
24-UG06		Gear	simple	Default	Category/Gear/Default	Category/Gear/Fitness	Equipment	base
24-UG07		Gear	simple	Default	Category/Collections/Default	Category/Gear/Default	Category/Gear/Fitness	Equipment/Default
24-UG04		Gear	simple	Default	Category/Gear/Default	Category/Gear/Fitness	Equipment	base
24-UG02		Gear	simple	Default	Category/Gear/Default	Category/Gear/Fitness	Equipment	base
24-UG05		Gear	simple	Default	Category/Gear/Default	Category/Gear/Fitness	Equipment/Default	Category/Promotions
24-UG01		Gear	simple	Default	Category/Gear/Default	Category/Gear/Fitness	Equipment	base
24-WG084		Gear	simple	Default	Category/Gear/Default	Category/Gear/Fitness	Equipment	base
24-WG088		Gear	simple	Default	Category/Gear/Default	Category/Gear/Fitness	Equipment	base

ADVANTAGES OF THE PANDA'S APPROACH

Any requirement changes in case. More conditions to be applied for data filtering are very easy to implement in the this approach.

The code complexity is minimum, so a person having just the basic knowledge of the pythonlanguage can easily understand the code.

Since, the python can work on any platform, so that feature makes this validation code platform specific.

In case of this approach, its very easy to filter out only the required columns out of the somany columns from CSV file.

Along with the column filteration, basic operations can also be performed on the columns inthe python code itself.

Columns need not to be remembered in this case with their occurrence index, only the nameof the columns works well. On the other hand, in case of querying using Snowflake, we need to remember the column indices as well, since in the Snowflake case, the columns are identified by their indices.

CONCLUSION

The above experimental setup proves that applying extraction and transformation on a large sized CSV file is very easy through the use of Pandas' library in Python then in comparison to the snowflake'spre-built functions to extract and check the data using SQL query.

Also, the simplicity of the code to be used for determining the structure and the various components out of the CSV file is very easy in case of Panda's execution.

REFERENCE

[1] Jitendra Pramanik and Abhay Kumar Samal and Kabita Sahoo, Exploratory Data Analysis using Python, International Journal of Innovative Technology and Exploring Engineering, October-2019.

[2] Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, and Chad Marston, "A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets," Visual Informatics, Volume 2, Issue 4, December 2018, pp. 235-253.

[3] John T. Behrens, "Principles and Procedures of Exploratory Data Analysis," Psychological Methods, 1997, Vol. 2, No. 2, pp.131-160.

[4] Chokey Wangmo, "An Exploratory Study on Bank Lending To SME Sector In Bhutan," International Journal of Scientific & Technology Research, volume 6, issue 11, November 2017, pp. 47-51.

[5] Matthew Ntow-Gyamfi and Sarah Serwaa Boateng, "Credit Risk and Loan Default among Ghanaian Banks: An Exploratory Study," Management Science Letters, Vol. 3, 2013, pp.753–762.

[6] X. Francis Jency, V. P. Sumathi, Janani Shiva Sri, "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients," International Journal of Recent Technology and Engineering (IJRTE), Volume-7 Issue-4S, November 2018, pp.176-179.

[7] K. Ulaga Priya, S. Pushp, K. Kalaivani, A. Sartiha, "Exploratory Analysis on Prediction of Loan Privilege for Customers using Random Forest," International Journal of Engineering & Technology, Vol. 7, Issue 2.21, 2018, pp. 339-341.

[8] Bogumil M. Konopka, FelicjaLwow, Magdalena Owczarz, ŁukaszŁaczmański, "Exploratory Data Analysis of a Clinical Study Group: Development of a Procedure for Exploring Multidimensional Data," Introduction To Machine Learning using Python [Online], Available:https://www.geeksforgeeks.org/introduction-machine-learning-using-python/

[9] Exploratory data analysis – From Wikipedia, the free encyclopaedia [Online], Available: https://en.wikipedia.org/wiki/Exploratory_data_analy