# Data Management using Python

**Anant Prakash Awasthi |** Lead Data Scientist | Optum

## Quick Disclaimer

This discussion is under sole knowledge of mine and not a part of any statement, policy or no where related to my employer.

# Data Management

**Organization**

Effective data management helps to organize data in a way that is easy to access and understand. It allows data to be stored, sorted, and retrieved efficiently, which saves time and improves productivity.

**Accuracy**

Proper data management ensures that data is accurate, complete, and up-to-date. This is important because inaccurate or incomplete data can lead to incorrect decisions and poor outcomes.

**Security**

Data management includes measures to protect data from unauthorized access, theft, or loss. This is particularly important for sensitive or confidential data that could cause harm if it falls into the wrong hands.

**Compliance**

Many industries and organizations have legal or regulatory requirements for data management. Failure to comply with these requirements can result in penalties, fines, or legal action.

**Analysis**

Effective data management is essential for data analysis. Data needs to be properly structured, cleaned, and transformed before analysis can be performed. This ensures that analysis is accurate and produces meaningful insights.

# Data Management in Python

Data management in Python involves using different libraries and tools to handle data in various formats and from various sources. Here are some key aspects of data management in Python:

### Data Manipulation

Python libraries such as **Pandas** and NumPy provide powerful tools for manipulating and analyzing data. These libraries enable you to clean, filter, sort, reshape, merge, and aggregate data to prepare it for analysis.

### Data Visualization

Python libraries such as Matplotlib, Seaborn, and Plotly enable you to create visualizations to gain insights from data. These libraries provide various chart types such as bar charts, line charts, scatter plots, and heat maps.

### Data Storage

Python provides several libraries for working with databases such as SQLite3, MySQL, PostgreSQL, and MongoDB. These libraries enable you to connect to databases, execute SQL queries, and manage transactions.

### Data Cleaning

Python provides libraries such as Pyjanitor and Dora for cleaning and transforming data. These libraries provide functions for handling missing values, converting data types, and dealing with outliers.

### Data Integration

Python provides libraries such as Apache Airflow and PySpark for integrating data from multiple sources. These libraries enable you to extract data from different sources, transform it, and load it into a destination system.

# Logistics for this Discussions

Data Management consist of many activities (sometimes called Tasks). We can classify data management tasks majorly into four categories for Pandas: