

Numpy And Pandas

Numpy (Numerical Python) (created in 2005 by Travis Oliphant)

- library used for working with arrays,
- Pandas, Matplotlib and scikit-learn are built using this.
- open source
- stands for Numerical Python
- Numpy arrays are called ndarray or N-dimensional arrays and they store elements of the same type and size
- import numpy as np
- used for multi-dimensional array operations
- used in machine learning and deep learning

→ Arrays - collection of elements/values that can have one or more dimensions.

→ An array of one dimension - vector
two dimensions - matrix

→ list has multiple data type but array has a single type. Handling is easy and more efficient at storing and manipulating the array

→ Numpy arrays - perform element-wise operation but list not possible.

→ It is known for its high performance and provides efficient storage and data operations as array grows in size

Functions of Numpy (Numpy Applied)

- 1) Data Manipulation and analysis - for data cleaning, transform the data, is processed through various numpy mathematical operation.
- 2) Machine Learning

Fastest Way to

- 2) Machine Learning: ML tasks math intensive and ML libraries (Tensorflow and scikit-learn) use Numpy for mathematical computations to support ML algorithm and model training.
- 3) Scientific Computing: handles matrix multiplication, eigenvalue calculation and differential equation. Valuable in modeling, simulation and other process.
- 4) Signal and Image processing: Image and signal, a data array provide tools to perform processing tasks.

Pandas

→ size mutability: columns can be inserted or deleted

→ Dataset merging and joining.

→ cleaning

series → Excel sheet

Dataframe (2D) → tabular structure.

Database

structured collection of data

- improve business processes
- keep track of customers (Amazon, Flipkart)
- store personal data
- keep a record of music or book collection
- Finances
- government organization
- Healthcare
- weather

Types of Database

Relational database - tabular structure

Object oriented database

Distributed database

Data warehouses - cloud storage

NoSQL databases - No structured (information - computer by books are different no structural defined)

SQL

Structured Query language

Access and manipulate database

eg MySQL, Oracle, Sybase

Importance of SQL

- Foundation of Data management
- Data extraction
- Data Manipulation
- Integration with Data tools
- Data security and integrity - (aggregat)
- Handling large Datasets
- Support for Advanced analytics
- Real time Data processing

→ Data warehousing
→ ETL Processes → (Extract Transform Load)

Terminology

- Rows of Table are referred to as relations
- columns of a table - attribute, or field,
- Rows - tuples or records
- no. of attributes

Different types of key

(individually taken separate data from a table)
(customer ID)

- Primary key :- a unique column to identify a row
- Foreign key :- Primary key of one table is the foreign key of another table - (to know relationship between tables)
- Composite key
- unique key :- The only difference between unique and primary key is that unique key can have null values

Datatypes

- char - string varchar

Database languages

- Data Definition language → how database should reside in database
(create, alter, drop)
- Data Manipulation → close, modify, delete and update data in database (select, insert, delete, update)
- Data Control → rights, permissions and other controls of the database (grant, revoke)
- Transaction control → Deal with transaction with a database (commit, Rollback, savepoint)

TRUNCATE TABLE - Keeping the structure and

Data Warehousing :- SQL is fundamental in building and querying data warehouses, where large amount of historical data are stored for analysis.

ETL Processes :- SQL is used in ETL process to extract data from various sources, transform it into a suitable format and load it into data storage system.

Business Intelligence :- Business intelligence tool uses SQL to create dashboards, reports and visualization that aid in decision-making.

Data Engineering

→ preparing, processing and moving data for analysis.

→ Data engineering build system, pipeline
Data science provides the data
(engineer - making kitchen
stove → work with ingredients)

Stages of a Data Pipeline (ETL/ELT)

→ Extract: collect raw data from source

→ Transform: clean, format, validate and enrich the data (handling missing values, joining multiple data sources)

→ Load
Store data into a target system like a database (cloud/data warehouse)

eg: E-commerce

+ extract → order data from website

+ Transform → clean duplicate orders, fix missing info (while system is hanging, when we click the order multiple times)

+ Load → store into central database for ordering

Pipeline Components:

Ingestion - collect data

Processing - clean, transform, enrich

Storage - save

Output - visualization, dashboards, report

Python for pipelines

→ libraries - pandas, numpy, os, schedule

Workflow Automation

Data is always changing / to get real time information and all we use automation

Saves time, reduce mistake

Bank sending message when amount is withdrawn

Data Ingestion

• gathering, managing and utilizing data efficiently
• plays foundational role in data processing pipeline system and taking diverse data and storing it in a centralized processing and analyzing place
Key steps in Ingestion Process

- 1) Data collection - Gather raw data from various sources
- 2) Data transformation - Clean, normalize and enrich the data
- 3) Data loading - move the transformed data into a target system

Batch - Data collected and processed at scheduled intervals - Daily report, payroll backups

Realtime - Data is ingested as it is generated - Fraud detect, live dashboard enabling instant processing

Micro-Batching - A hybrid approach - To T data, semi-line analysis
small batches processed frequently

Data Ingestion Workflow

→ Data source identification - register data source

→ Data Extraction - Extract data

→ Data staging - store the data temporarily

→ Data validation - validate the data for completeness and accuracy

→ Data validation - validate the data for completeness and accuracy

→ Data transformation - clean, normalize and enrich the data for loading

→ Data loading - load the transformed data into target storage

→ Data monitoring - monitor performance and make necessary adjustments

Characteristics

- latency -
 - Data size - Handle large volumes
 - Processing - Done in groups
 - Scheduling - usually triggered by time
 - Fault tolerance - easier to retry failed batches
 - Cost - generally cheap
- Eg. ^(batch mode) Payroll systems, Bank Statements generation, E-commerce -

Limitations

delayed results, Resource spikes, complex Debugs

Stream Processing

continuous and real-time processing of data
no delay.

Kafka - real time data processing (distributed system in database)

- Producer → that ~~data~~ generates data - (maybe system)
 - Topic → (type, failing, success)
 - partition
 - Consumers / consumer group - group of people who handle all the intent.
 - Broker → manages and manages the data
 - Zookeeper → a person who monitors all the Broker
 - offset - used to track the handling of data protecting the data without duplicating it on id.
- events contain all the details of the data about who created and handled it


Feature Engineering

process of using domain knowledge to extract and transform raw data into features - involves selecting, modifying or creating new features

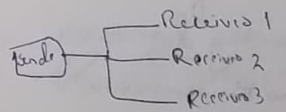
Apache Kafka

Publish / subscribe with Message queue

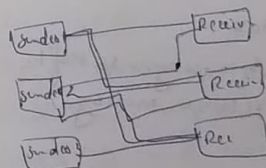
communication → two



→ sender needs to know the address of the receiver
→ Both need to agree upon a protocol / format
→ provides static bid

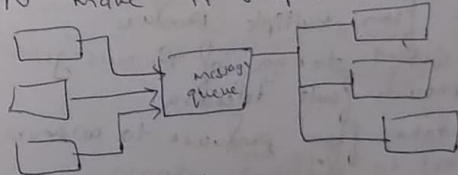


→ each responsible for a different type of process
strong coupling between sender and receiver



management is more troublesome

To make it simple use message queue



senders publish to message queue and receiver subscribe to the message queue / oblivious of other receiver or sender

this is known as pub-sub

Publish/subscribe Advantages

- Decoupling of publishers and subscribers
 - easy management and changes (can move / added without any change)
- Scaling (a-)
- Back-pressure handling - if publisher generates a spike, message queue act as a buffer zone to cache data til subscribers pulls it down
- Reliability - tracking of data.

Apache Kafka

- event / messages streaming platform
event / message → data transferred
- Critical piece of the big data puzzle
- open-source technology with commercial options
- one of the most popular
- producers (publishers) push messages to Kafka, subscribers / consumers receive messages

Kafka Data Function

- * collects data from multiple producers
- * provides persistent storage of the messages received - provide fault tolerance
- * transports data from producers to consumers
- * can transport to different networks
- * Distribute data to concurrent customers
- * Tracking of message consumption by one consumer

Even if the consumers go down, the message will return.

Benefits of Kafka

- handle large data (high throughput)
- supports multiple concurrent producers and multiple subscribers for the same data
- low latency
- fault tolerance (reliable stored and distributed)
- Decoupling (didn't know about other producers / software development become easier)
- Back pressure handling - consumers can use it even after the producers provide it after long time
- Horizontal scalability (additional nodes can be added)
↓
- low latency - provide stream processing store and forward capability - batch application

Use Cases

- * Asynchronous messaging between services
- * Real-time stream processing
- * Logging and monitoring (log messages and alerts)
- * Event sourcing
- * real-time analyzing (Apache spark or flink)

Kafka Messages - unit of data that is collected, stored and distributed by kafka
row/record in a database, it can have attributes and values like a map, obj contains images or videos/audio.

kafka treat all events as byte array

Producers and consumers have to agree upon the content and format of the message and be able to serialize and deserialize them

→ size limits exist in kafka

(max.message.bytes) → configurable by parameter
default is 1MB.

request batch request for efficiency.

Key Contents of A Message

Key - defined by the producer of the message
not mandatory and need not be unique
used for partitioning data

Value - Actual content - byte array - user defined
not infer anything from the message content

Timestamp - automatically timestamped

two types -
- Event time - when the producer creates the timestamp
- ingestion time - broker stamps it when it stores the record

if key is not provided by the producer kafka assign a random key.

Topics - Hold and manage messages

- Holder of messages (pile with papers)
↳ database table entries/records
- Queue for the similar content (same set of concepts already)
(Has no limitation on what the content will be)
sales transaction, audit logs, video files
- supports multiple topics (based on use case)
- support multiple producers and consumers concurrently
- multiple partition that physically split data across multiple files.

Brokers

Central brain in everything that kafka does

→ A running kafka instance

(it is a physical process that runs on base operating system and executes all kafka functions)

- listen on a specific port (usually 9092 but is configurable)
- receives messages from producers and stores it in logs, consumers subscribe to specific topic within the broker. the broker sends only the new message (keeps track of the last message it send). If a consumer dies the server everything

- subscription management
- manage topics, partitions and logs (lifecycle of topics)
- clustering capabilities (multiple kafka brokers act as a cluster and one among them will act as the active controller for the cluster)
- Each partition will have a leader
- provide fault tolerance

logs - physical files where data is stored before giving to consumers. Managed by kafka brokers. Multiple files. When a file is filled it is rolled over and a new file will continue the process. The data is kept for a configured time. Normally it is seven. Physical space management is required. log.dirs is used to place path for directory

Producers

eg:- web server that push data
a log/scribbles pushing in log messages
server pushing elementary data
Producer is formed from client libraries within their code and publish data
multiple concurrent producers
producers have to identify the key of a message

message serialized to bytes.
synchronous and non-synchronous (doesn't wait for brokers to acknowledge and publish)

Consumers

eg:- an archiver, a real-time data aggregator
consumer can consume a file as long as the message is stored in the log file
client libraries to consume data
deserialize bytes to data structure
offset management

Kafka client scripts

docker exec -it kafka-brokers /bin/bash - enter kafka container
to get into root directory - /opt/bitnami/kafka
logs directory - contains logs which are useful for troubleshooting
bin directory - contains shells for kafka management, publishing and subscribing to interact with kafka

script to create topic is kafka-topics.sh
replication - 1 means only one kafka broker created in system
list of topics - list action / describe (more details)

To publish data - kafka-console-producer.sh.
gets list of all parameters.
end the program with control+c or control+d

consume the message - kafka-console-consumer.sh
all message \rightarrow from-beginning

kafka-topics.sh shell script can be used to
modify and delete topics
delete command should be added

log.dirs - which all message publish to kafka
are stored.

Partitions

- each topic can 1-n partition
 - partitions allow kafka to scale
 - have separate log files
 - has a leader broker
 - Enable consumers to handle better
- } too many partition

Note: \rightarrow Each message goes to only one partition (explicitly it will all get it)
 \rightarrow message ordering guaranteed within a partition only
 \rightarrow same message key = same partition
 \rightarrow partition count cannot be decreased after a topic is created.

To publish data with keys we use colon operator
to separate data with keys

pass.key = True - tells the producer to read key as
key separator - specify the separator character

Consumer Group

- + A group of consumers who share a topic workload
- + each message goes to only one consumer in a group
- + consumers split workload through partition.
no. of partition \geq no. of consumers

Offsets

- no. to track message consumption by consumer and partition
a message ID is given to message
- Broker keeps track of what is sent and acknowledge
- current offset: last message sent to a given consumer
- committed offset: last message acknowledge by consumer
- + Broker resend uncommitted message in case of failure

A properties object is created to specify the kafka connection properties: a mandatory properties required

BOOTSTRAP_SERVERS property - specify list of kafka brokers
external port number

specify the key and value serializers class
a string serializer is used as both are string.

ConsumerConfig parameter is used to
auto reset config parameter is used to
specify all message or new messages
subscribe method is used to define subscribe to a topic

Stream Processing Architecture

Continuous Data Sources
(sensors, web clicks, payment
Kafka et)



Data Ingestion
(Kafka, Istio et)



Stream Processing Engine
(Flink, spark streaming,
storm, Dataflow et)



Alerting system
(slack, Email)



Downstream DB
(NoSQL, DW)



Real-time Dashboard
(Grafana, kibana)

Pandas Advantages

- Data from different resources can be loaded
- Easy handling of missing data (represented as NaN)
- size mutability: columns can be inserted/deleted from Data frame
- Data set merging and joining.
- Allow us to analyze big data and make conclusion efficiently based on statistical theories
- pandas can clean messy data sets and make them readable and relevant
- Relevant data is very important in data science

Python lists

- * contain only ~~heterogeneous~~ elements
- * can't perform element wise operation
- * suitable for data storage ~~but~~ not ideally intended for mathematical

Numpy Arrays

contain only homogeneous data which make it more efficient at storing and manipulating. perform element wise operations easily support math intensive tasks.

A numpy array can specify and support various data types. any array created in numpy should use only one desired data type - a different array for different data type. This requires less memory and more efficient system performance in mathematical operations.

Tableau

www.tableau.com - Tableau Desktop

Connect panel contains and provides data, we can collect data from Tableau server, or file - (Excel, JSON, PDF) or from a server (MySQL, Amazon Redshift). we could also download additional connectors.

Different file types

→ Tableau workbook (.twb) - stores a visualization without a source data. But contains the link of the data source that provides the data.

→ Tableau datasource (.tde) stores the server address, password and other information required to access a datasource (Oracle database).

→ Tableau bookmark (.tbn) - stores a connection to a worksheet in another Tableau workbook. This is useful if we want to reference another workbook but without bringing any other data.

→ Tableau Data Extract (.tde) - stores Tableau data as a filtered and aggregated extract rather than full dataset this would be a subset of it.

→ Tableau Packaged workbook (.twbx) - stores extracted data and visualization for viewing in any full version of Tableau - good for including visualization and data source.

Note: - help menu is available in start screen

help menu can also be accessed by F1 key

- Content pane - we could choose the file needed
- Tableau desktop - create dashboard on computer
- Tableau prep - prepare and clean data

- To connect to one's system is easy but to connect to a server like SQL we need credentials.

Join related data sources

drag and drop. then the bottom will show these relationship if any. such that common columns. (Tableau can identify key). this happens because of the key field - (order/primary key)

by relation ^(field) → a foreign key - indicates a particular column is a primary key in another set

[Show me]

Join tables with different names in common fields.

Tableau may not be able to identify auto. ∴ we need to specify the field

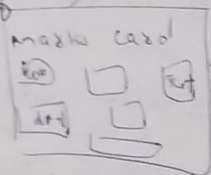
Clean up in Data

- 1) To do it without changing data source click the cleared with data interpreter check box ☒
- 2) by unchecking we get table / Excel table drag both sheet 1 and sheet 2 → table 1

Reorder fields in a visualization

meaning to create visualization by dragging,
order creates difference in visualization
(emphasize of data) - order matters.

Change Summary operation (Avg/Count)
Default, it is sum. click down arrow and



click on measure and make
the change

Split text into multiple columns

split to multiple columns - two information in
a single column can be split to
different columns.

+ look for a character test split then
(comma maybe used to split it into two)

+ click the arrow in column header and
click split

+ original column remains and new columns are
also added.

+ If I don't want automatic split, I
would do custom split

+ rename the fields - click rename on dropdown
button.

Display the data underlying a workbook

To know the data source of visualization,
click data source on the bottom left corner
click, takes us to data source package.

Here it shows the name of the current table
a person is viewing and the no. of rows
displayed out of the total

Initially each column will be in data source order
By clicking on the control of each column
header we can change it.

Add, Duplicate and Rename Worksheet

→ To add a worksheet → go to bottom panel
and next to sheet 1 we have new worksheet
or (control + M) - (another way)

→ To rename double click of sheet name and
write it

→ To duplicate → right click on the sheet name
and click duplicate

Reorder, clear and delete worksheet

+ To reorder - we can drag a particular one
in the sheet name and place where it is required
black line representing where it will be placed

+ To clear - go to worksheet menu on top
and click clear and this sheet. (but sheet
name will not change)

+ select the worksheet, right click it and
click delete.

Change worksheets tab colour

To bring attention, Right click on sheet name we get a dialog box and click colour.
To undo - right click and click colour and then none

Show or Hide Worksheet Element,
down arrow and hide option
side bar or navigation box
to bring them back → go to worksheet and
choose title / caption / card and reset.

Save them

Written to tool bar and choose save icon or
File menu and save as.

to revert all change use File or File menu
and revert to save (go to last saved one)

Save as PDF

File menu and print to PDF (In options as default.)

Create a Packaged Workbook

To share with someone who doesn't have
Tableau use packaged workbook.

go to file → export packaged dialog box.
to get tableau reader from website.

Export data to a CSV File

go to data source.
go to data menu → to the data source and
export to CSV.

Tableau Operators and built-in functions

- + for addition and string concatenation.
- - for subtraction and negation
- * for multi
- / - division
- % modular division - remainder
- ^ exponent
- comparison operator - < > - not equal to
- AND, OR, NOT logical operators

Functions

- SUM - sum
- AVG - finds average
- MIN and MAX
- STDEV - standard deviation
- ROUND - rounds to a specified number of points
- ABS - Absolute value
- CEILING - rounds number up to nearest integer
- FLOOR - rounds number down to nearest integer

Create Calculated Field

From Analysis menu → choose create calculated field
→ calculation dialog box appears.

write it in square brackets.
for calculated field equal sign before hashtag
We can rename the calculated field

Create Calculated Field in shelf

double click here.
means typing in the marks card, the same statement
→ type here in square bracket

apply it by control + enter.
it won't be added automatically to pills area. so drag

Add a table calculation

To create table calculation like difference of two columns, click the down arrow in the ~~chart~~ marks card part that show the calculation and click create table calculation and choose from the option,

To clear it click the clear table calculation at the same place

For it also follow same step and click edit table calculation and a dialog box appears

Use level of detail expressions (LOD) $\frac{1}{\frac{110}{4}}$

Analysis \rightarrow create calculated field

To indicate level of detail use curly bracket,

{INCLUDE (customer state); sum(orders Total)}

To exclude use EXCLUDE-keyword

Calculate measures (average, median, min, max)

If you already make orders total as sum it's not work if you want to find its average. So, put it in mark card (appear as detail) and click test then it's added.

to make separate columns drag it from the marks card to measure value.

Count Occurrences and Distinct items

Display summary card

Worksheet - show summary (about writing of the sheet)
change summary card - by clicking the down arrow in summary card.

Calculate Percentages within a Table

Analysis menu \rightarrow percentage of \rightarrow columns to get percentage of the column

Analysis \rightarrow percentage of \rightarrow column pane

Forecast future values

What will happen in future - go to side bar in the side pane and click Analytics and under model choose forecast
Analysis menu ~~forecast~~ pick forecast and forecast options is automatic

Sort based on a field's values

\rightarrow Sort by click sort on tool bar (ascending/descending)
or In the heading of columns we can sort
 \rightarrow otherwise in rows \rightarrow click region and click sort then data source order / Alphabetic etc.
or sort manually (sort by)
 \rightarrow can also sort by the basis of another field not in visualization by clicking field and choosing which field as basis

Create a Nested Sort

Click the down arrow of a field in rows and sort by and nested.

Create a Selection filter

drag the field to filter dialog box

→ In general: we could choose the data to display

→ wildcard filter: checks for a string that matches the given string.

"batt" → gives everything having "batt"

It has (contains, starts with, ends with, exactly matches) options

→ Condition filter: - by field → def. the condition

→ Top filter: - To get maximum 10 datas by field - (Top/bottom) - select the number

Edit, clear and Delete Filters

In the filter dialog box - click down arrow and click edit
Go clear in worksheet menu - click clear - click filters

Filter data using Parameters

click the arrow with the field in sid pane

→ click create → click parameter and give necessary instructions. → side pane will have the new data

later drag the field to filter dialog box and click top, by field and click the new change option

→ to edit - click arrow in the pane - click edit
similarly we can do rename and delete.

Define a Group

Going to sidebar and double clicking it → click create

→ click group

group the data accordingly.

grouping helps to group regions on the same direction

Define a group for other values

other - checkbox group all other into a single group → it has a (paperclip icon)


Find members of a group

click find button in group → ~~check~~ with a condition box (case insensitive) and click group

Edit and delete group

right click on the group to be changed and click edit Alias.
right click on table name and click edit group then click it and ungroup.

Define a Set

click the required states name directly from the worksheet and go to tool box that appears and click create set button (window icon)
give new name, then ok. (it has  icon)
Analysis → total → add subtotal.

Combine Sets

If we add 2 sets to the filter dialog box then nothing appears as it takes the intersection. To combine it select both the sets of side pane and right click and create combine set.

(All members), intersection, right join, left join

To add new members to existing set - right click - click set icon - add to the required set

Create column and stacked column chart.

Stacked column charts are best for summarizing categorical data.
Scatter plot - need two numerical variables to dimension

Identify clusters:- Go to analytic tab on side pane take the columns to details → we get clusters drag to visualization -

size of bins in a histogram can be changed by right clicking on field and click create and click bins


Area fill charts show how categories of data contribute to a total over time.

Change Appearance of Data

Go to format menu in tool box. The format box appears in the side. The style of data can be changed (worksheet, pane, Header, Tooltip, title)

Shading can be done for rows. (pane, header colour can be changed)

The format of a field can be changed by going to sidebar and click down arrow of a field click default properties - the format remains same across only in the particular workbook

To make visualization to presentation mode, go to tool bar and press  → this icon or F7 (press F7 not return back)

The tooltip can be formatted by clicks worksheet in menu and click tooltip