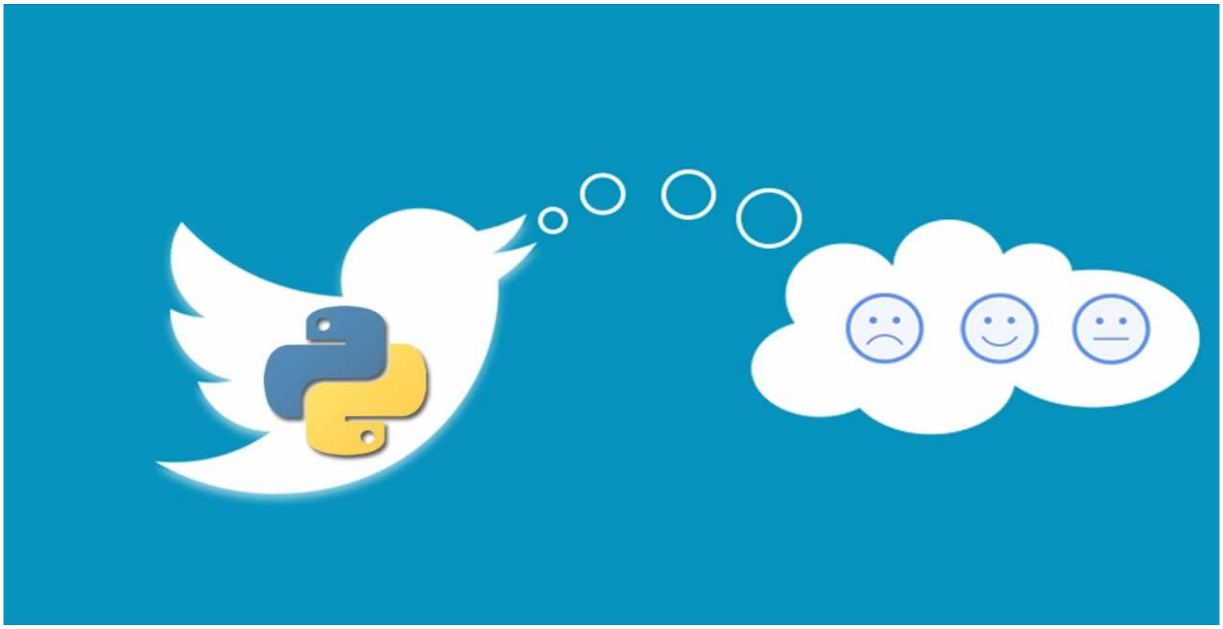
BIG DATA ANALYSIS WITH IBM CLOUD DATABASE

[Phase 5]



SUBMITTED BY:

NAME: Vasundharaa R N

NM-ID: CF2B68181F8095C8E024EDEF64F189C1

REG-NO: 810721104056

Project Description:

In the era of information abundance, organizations face the challenge of managing, processing, and extracting valuable insights from massive datasets. The "Big Data Analysis with IBM Cloud Database" project is designed to address this challenge by leveraging the power of IBM Cloud Database services to unlock the potential hidden within vast amounts of data.

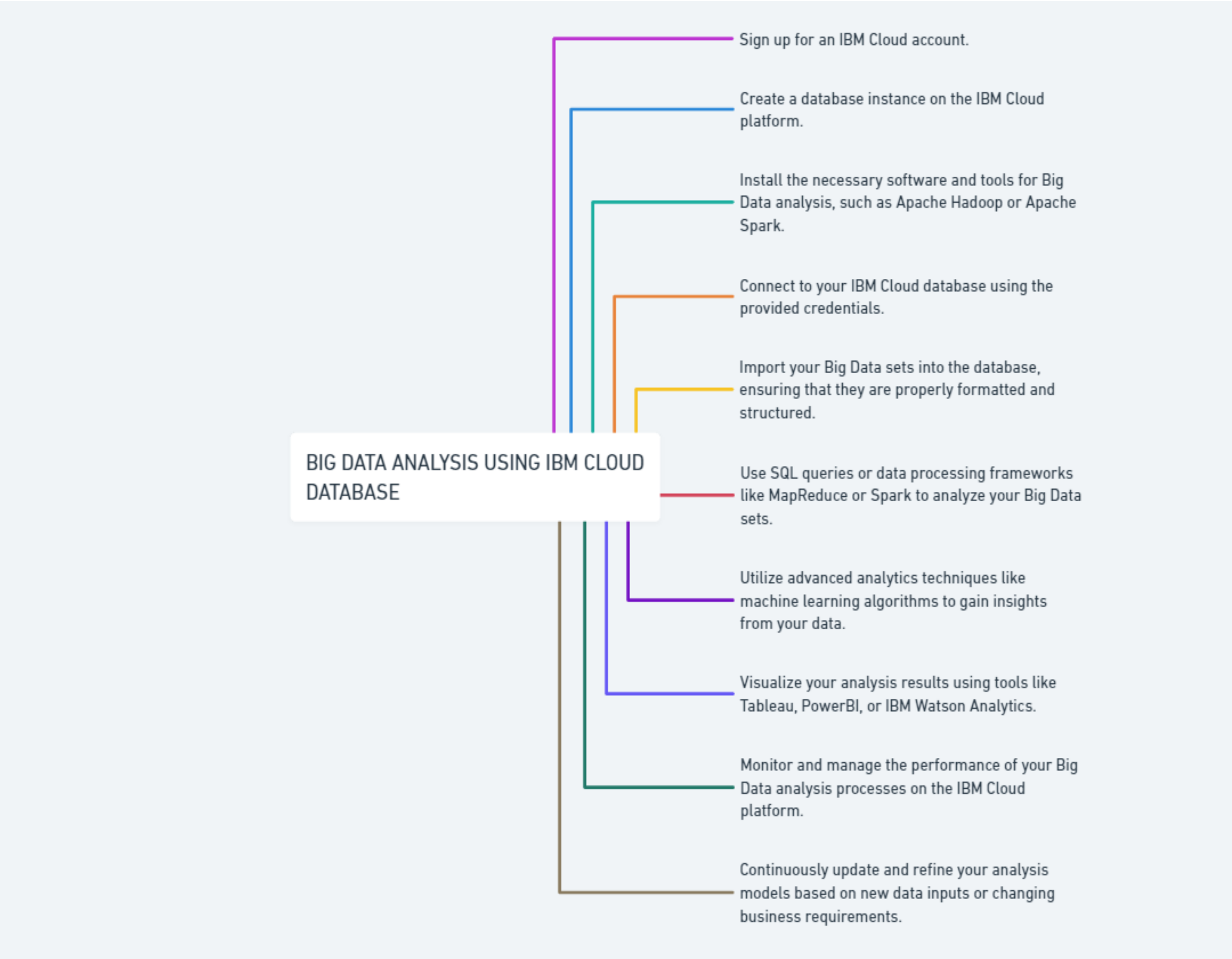
Objective:

The objective of the project "Big Data Analysis with IBM Cloud Database" is to leverage the power of IBM Cloud Database services to analyze and derive valuable insights from large datasets. This project aims to

1. Implement a scalable and efficient data storage and management system using IBM Cloud Database technologies.
2. Explore and apply advanced data analytics and machine learning techniques to extract meaningful patterns and trends from the stored data.
3. Develop data visualization tools and dashboards to effectively communicate the analysis results.
4. Demonstrate the potential of IBM Cloud Database in handling and processing big data, showcasing its performance, scalability, and cost-effectiveness.
5. Provide recommendations and insights that can inform data-driven decision-making for businesses and organizations.

By achieving these objectives, the project seeks to demonstrate the capabilities of IBM Cloud Database for big data analytics and contribute to the growing field of data-driven insights and decision-making.

Design Thinking



BIG DATA ANALYSIS USING IBM CLOUD DATABASE

Sign up for an IBM Cloud account.

Create a database instance on the IBM Cloud platform.

Install the necessary software and tools for Big Data analysis, such as Apache Hadoop or Apache Spark.

Connect to your IBM Cloud database using the provided credentials.

Import your Big Data sets into the database, ensuring that they are properly formatted and structured.

Use SQL queries or data processing frameworks like MapReduce or Spark to analyze your Big Data sets.

Utilize advanced analytics techniques like machine learning algorithms to gain insights from your data.

Visualize your analysis results using tools like Tableau, PowerBI, or IBM Watson Analytics.

Monitor and manage the performance of your Big Data analysis processes on the IBM Cloud platform.

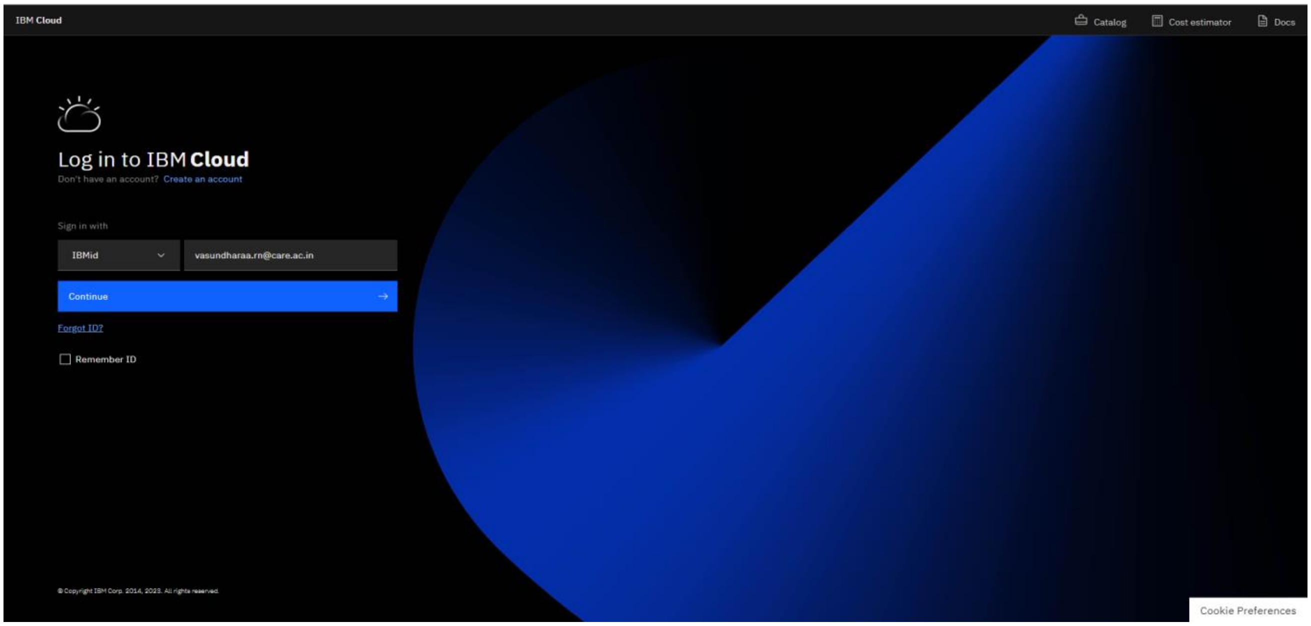
Continuously update and refine your analysis models based on new data inputs or changing business requirements.

Development phase:

1. **Dataset Collection:**
   - As per the project, We’ve to select social trends or climate dataset.
   - So, We took twitter data set which is in the CSV format for our big data analysis.

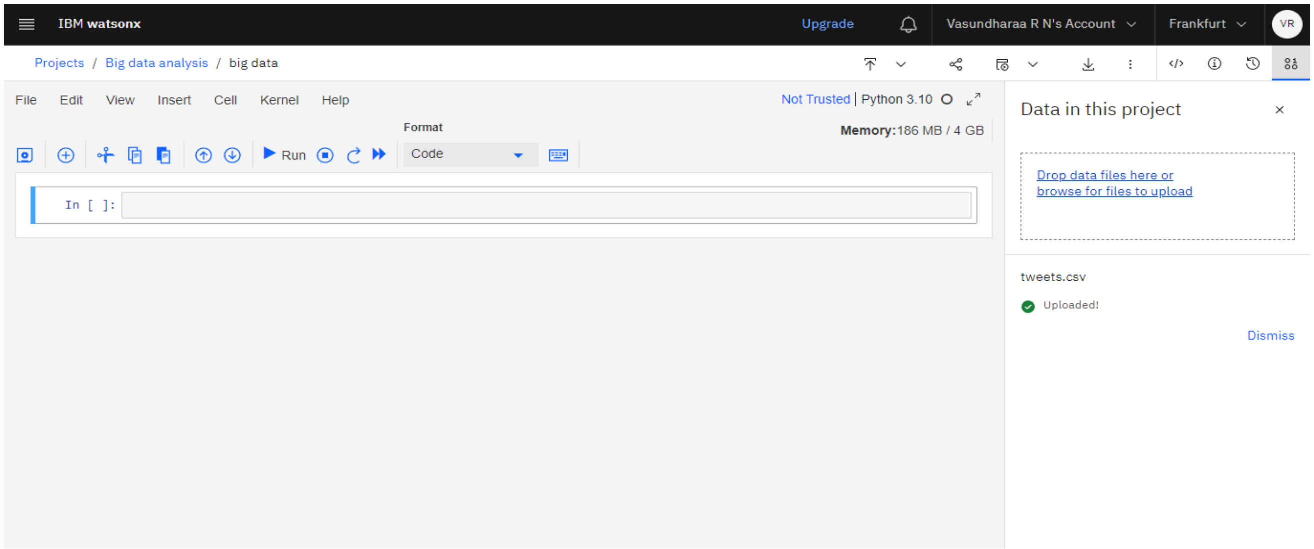
2. **Data Upload to Cloud Database**

Step-1: Login to the IBM cloud



vasundharaa.rn@care.ac.in

Step-2: Open Watsonx jupyter notebook, Go to the upload asset to project, Then browse the dataset & upload

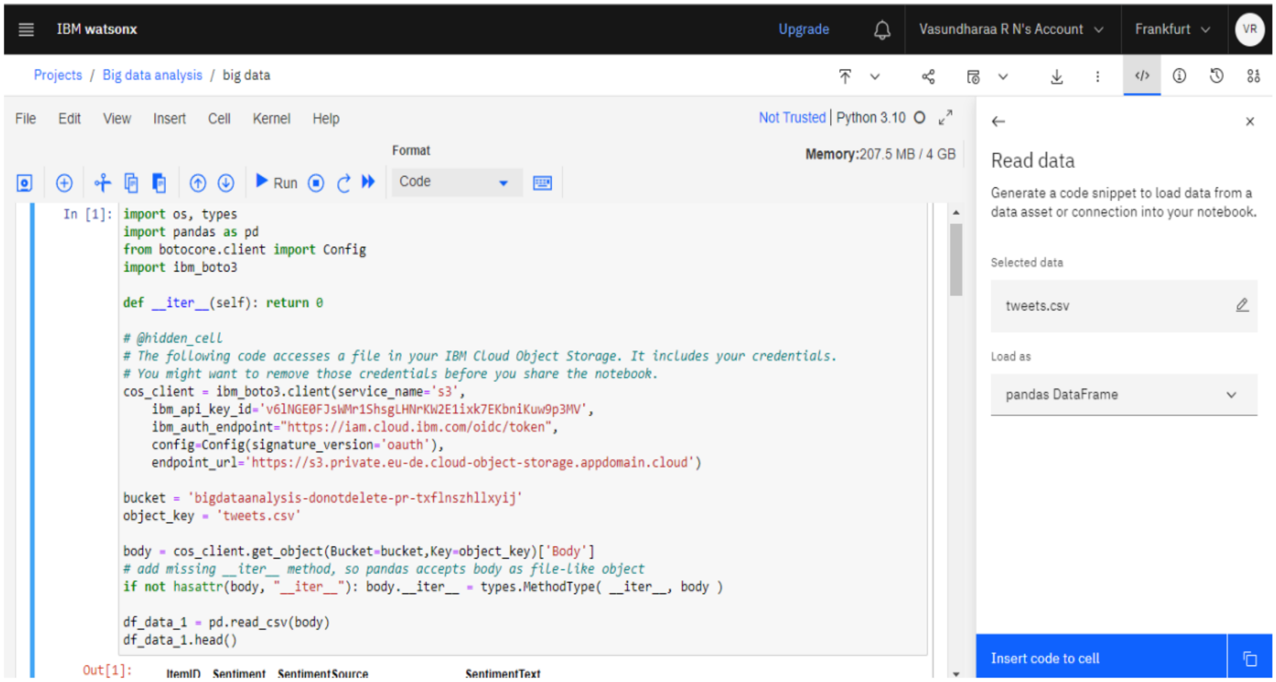

IBM watsonx
Upgrade
Vasundharaa R N's Account
Frankfurt
VR
Projects / Big data analysis / big data
File
Edit
View
Insert
Cell
Kernel
Help
Not Trusted | Python 3.10
Format
Memory:186 MB / 4 GB
Run
Code
Data in this project
Drop data files here or browse for files to upload
Dismiss

```
In [ ]:
```

tweets.csv

Uploaded!

Step-3: After step-1, go to code snippets, then click read data & select the data from project. Then click load as panda frame, the above codes will be display



Memory:207.5 MB / 4 GB

Read data

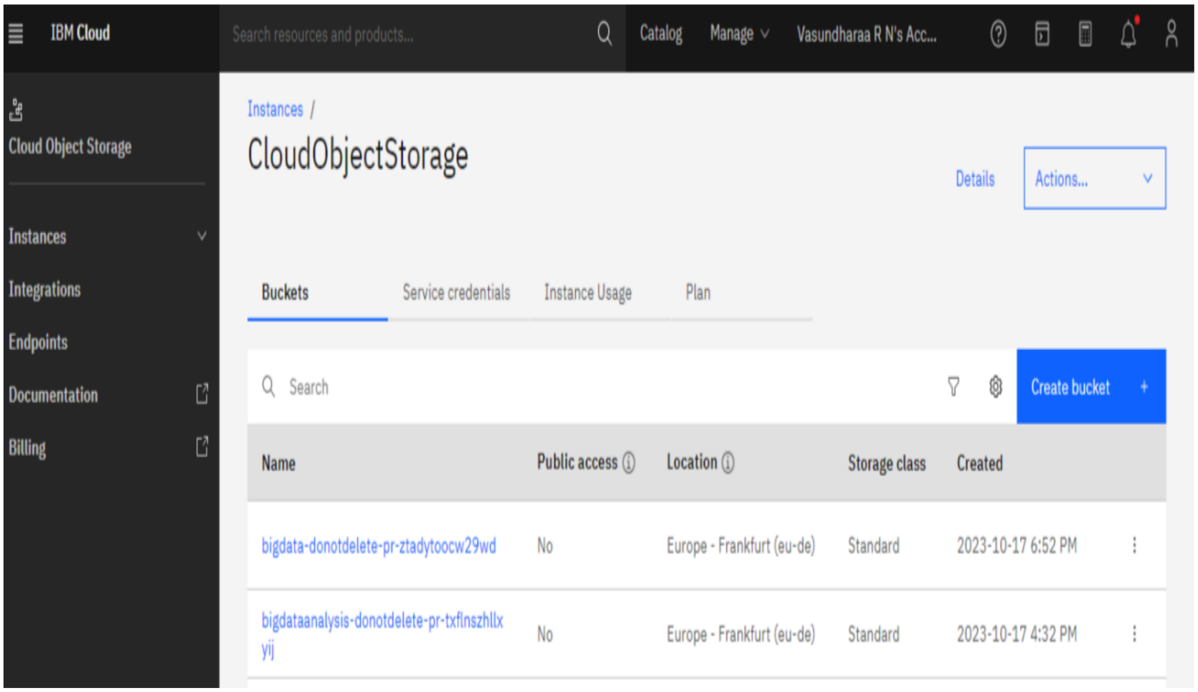
Generate a code snippet to load data from a data asset or connection into your notebook.

Selected data: tweets.csv

Load as: pandas DataFrame

```
In [1]: import os, types
        import pandas as pd
        from botocore.client import Config
        import ibm_boto3

        def __iter__(self): return 0

        # @hidden_cell
        # The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.
        # You might want to remove those credentials before you share the notebook.
        cos_client = ibm_boto3.client(service_name='s3',
            ibm_api_key_id='v6lNGE0FJsWMr1ShsgLHNrKW2E1ixk7EKbniKuw9p3MV',
            ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",
            config=Config(signature_version='oauth'),
            endpoint_url='https://s3.private.eu-de.cloud-object-storage.appdomain.cloud')

        bucket = 'bigdataanalysis-donotdelete-pr-txflnszhllxyij'
        object_key = 'tweets.csv'

        body = cos_client.get_object(Bucket=bucket,Key=object_key)['Body']
        # add missing __iter__ method, so pandas accepts body as file-like object
        if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

        df_data_1 = pd.read_csv(body)
        df_data_1.head()

Out[1]:     ItemID  Sentiment  SentimentSource            SentimentText
```

Step-4: Buckets are created in IBM cloud object storage



CloudObjectStorage

Name	Public access	Location	Storage class	Created
bigdata-donotdelete-pr-ztadytoocw29wd	No	Europe - Frankfurt (eu-de)	Standard	2023-10-17 6:52 PM
bigdataanalysis-donotdelete-pr-txflnszhllxyij	No	Europe - Frankfurt (eu-de)	Standard	2023-10-17 4:32 PM

3. **Extracting the data from the cloud**
   - As we already known, Data are loaded in cloud object storage as pandas data frame.
   - Convert pandas data frame into pyspark dataframe.

4. **Validate the data**
   - Validating a dataset refers to the process of assessing and verifying the quality, accuracy, completeness, and reliability of the data before using it for analysis or other purposes. Dataset validation is a critical step in data preparation to ensure that the data is fit for its intended use
   - Check for missing values
   - Check for duplicates

5. **Data Preprocessing**

   a) **Removing Unnecessary word**
      - Tokenization involves breaking down text into smaller units, typically words or subwords, to facilitate furtherprocessing
      - Stop word removal is a basic preprocessing step and may not always be appropriate for sentiment analysis..
      - The decision to remove stop words or not should be made based on the nature of text data.

   b) **Converting words feature into numerical feature**
      - In machine learning , converting words or text features into numerical features using hashing is a technique known as "hashing trick."
      - Hashing reduces the dimensionality of the text data, allowing you to use it in machine learning models without explicitly storing a vocabulary.

   c) **Text Preprocessing:**
      - Reducing words to their base form can help reduce the dimensionality of your data and group related words together.
      - Stemming and lemmatization are two techniques for this.
      - Stemming often results in the removal of prefixes or suffixes, while lemmatization returns the base or dictionary form of a word.

6. **Building the data model**

   a) **Splitting the data**
      - Divide your data into two subsets:
        a. The training set
        b. The testing set.

   b) **Preparing the train data**
      - Choose an appropriate model or algorithm based on your objectives.
      - For big data, distributed machine learning frameworks like Spark MLlib, TensorFlow, or scikit-learn are commonly used.
      - Here, We choose spark MLlib in that we used Logistic regression model.

c)Train our classifier model using training data

- Training a classifier model using training data involves using a machine learning or deep learning algorithm to build a model that can predict the labels or categories of data points based on the features you provide.
- So,We use logistic regression model to train the model.

d) Prepare testing data

- The testing data can be used for the metrics and techniques appropriate for your specific task, such as accuracy, precision, recall, F1-score, or mean squared error, among others.

e) Model Evaluation

- Evaluate the model's performance using appropriate metrics
- Common metrics include accuracy, precision .

7. Data Visualization

a) Word Cloud

- Creating a word cloud is relatively simple using various libraries and tools, such as Python's word cloud library or online word cloud generators.
- So, We're creating two word cloud
  1. Positive word cloud
  2. Negative word cloud

b) Pie Chart:

- Creating a pie chart in PySpark to represent positive and negative words is not a direct task for PySpark. PySpark is typically used for distributed data processing and analysis, while data visualization, including creating pie charts, is usually done using libraries like Matplotlib, Seaborn, or Plotly.
- However, you can use PySpark to process and analyze the data and then create a pie chart using a data visualization library.