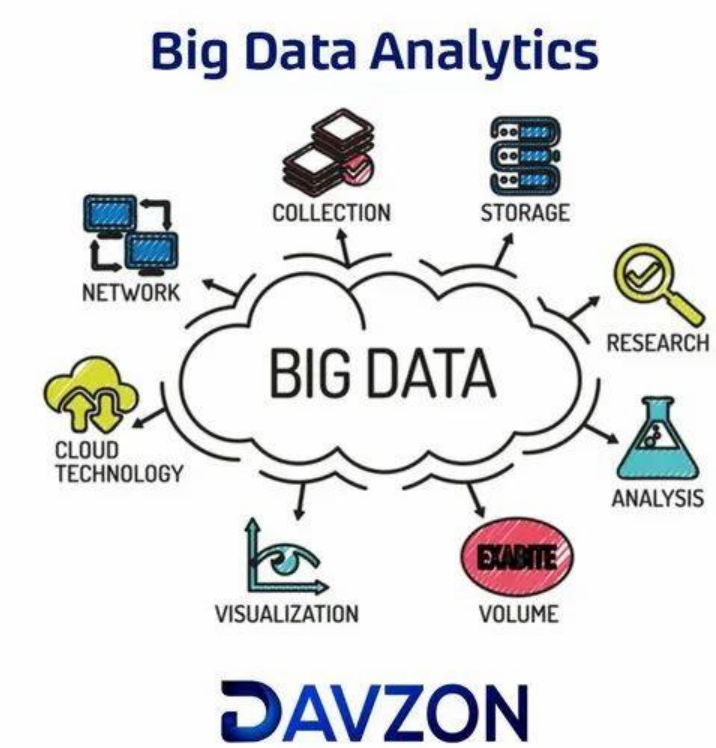
shutterstock.com · 370904066

BIG DATA ANALYSIS ON IBM CLOUD

Phase 2

SUBMITTED BY
Vasundharaa R N
Reena A R
Sindhuja S
Vaishnavi A S
Vidhya R



INTRODUCTION:

- ✓ Big data analysis is the process of collecting, examining, and analysing large amounts of data to discover market trends, insights, and patterns that can help companies make better business decisions.
- ✓ It involves usage of advanced analytics,technique to process and analyze vast amounts of data to extract meaningful and actionable information.
- ✓ Big Data Analysis empowers organizations to make data-driven decisions, enhance operational efficiency, predict future trends, improve customer experiences, and gain a competitive advantage in today's data-driven world.

DESCRIPTION

Data Collection:

- ✓ Gather a vast amount of textual data from various sources such as social media platforms, customer reviews, forums, blogs, news articles, and other relevant sources.
- ✓ So, Here we using “twitter” dataset for this project.

Data Preprocessing:

- ✓ Clean and preprocess the collected textual data.
- ✓ This involves tasks such as removing irrelevant information, handling missing data, standardizing text (lowercasing, removing special characters, etc.), tokenization (splitting text into words or phrases), and removing stop words.

Data Splitting:

- ✓ Split dataset into training, validation, and testing sets.
- ✓ The training set is used to train the model, the validation set is used to tune hyperparameters, and the testing set is used to evaluate model performance.

Sentiment Lexicon and Annotation:

- ✓ Utilize sentiment lexicons, which are dictionaries or databases containing words or phrases associated with their sentiment scores (e.g., positive, negative, neutral).
- ✓ These lexicons can help in classifying the sentiment of the preprocessed text.

Model Training:

- ✓ Train the selected model(s) using the training data, optimizing the chosen objective function.
- ✓ Here, We using logistic regression model for training the data.

Model Evaluation:

- ✓ Evaluate the model's performance on the validation and testing sets using appropriate evaluation metrics.

Analysis and Visualization:

- ✓ Analyze the sentiment analysis results to derive insights, trends, and patterns related to the sentiments expressed in the data.
- ✓ Visualize the sentiment distribution and trends using charts, graphs, or other visualization techniques.

PROGRAM

Import modules and create spark session

```
#import modules
from pyspark.sql.types import *
from pyspark.sql.functions import *
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.feature import HashingTF, Tokenizer, StopWordsRemover

#create Spark session
appName = "Sentiment Analysis in Spark"
spark = SparkSession \
    .builder \
    .appName(appName) \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

Read data file into Spark dataFrame

```
#read csv file into dataFrame with automatically inferred schema
tweets_csv = spark.read.csv('dataset/tweets.csv', inferSchema=True, header=True)
tweets_csv.show(truncate=False, n=3)
```

Divide data into training and testing data

```
#divide data, 70% for training, 30% for testing
dividedData = data.randomSplit([0.7, 0.3])
trainingData = dividedData[0] #index 0 = data training
testingData = dividedData[1] #index 1 = data testing
train_rows = trainingData.count()
test_rows = testingData.count()
print ("Training data rows:", train_rows, "; Testing data rows:", test_rows)
```

Prepare training data

```
tokenizer = Tokenizer(inputCol="SentimentText", outputCol="SentimentWords")
tokenizedTrain = tokenizer.transform(trainingData)
tokenizedTrain.show(truncate=False, n=5)
```

Removing stop words (unimportant words to be features)

```
swr = StopWordsRemover(inputCol=tokenizer.getOutputCol(),
                       outputCol="MeaningfulWords")
SwRemovedTrain = swr.transform(tokenizedTrain)
SwRemovedTrain.show(truncate=False, n=5)
```

Converting words feature into numerical feature. In Spark 2.2.1,it is implemented in HashingTF funtion using Austin Appleby's MurmurHash 3 algorithm

```
hashTF = HashingTF(inputCol=swr.getOutputCol(), outputCol="features")
numericTrainData = hashTF.transform(SwRemovedTrain).select(
    'label', 'MeaningfulWords', 'features')
numericTrainData.show(truncate=False, n=3)
```

Train our classifier model using training data

```
lr = LogisticRegression(labelCol="label", featuresCol="features",
                        maxIter=10, regParam=0.01)
model = lr.fit(numericTrainData)
```

Prepare testing data

```
tokenizedTest = tokenizer.transform(testingData)
SwRemovedTest = swr.transform(tokenizedTest)
numericTest = hashTF.transform(SwRemovedTest).select(
   'Label', 'MeaningfulWords', 'features')
numericTest.show(truncate=False, n=2)
```

Predict testing data and calculate the accuracy model

```
prediction = model.transform(numericTest)
predictionFinal = prediction.select(
   "MeaningfulWords", "prediction", "Label")
predictionFinal.show(n=4, truncate = False)
correctPrediction = predictionFinal.filter(
   predictionFinal['prediction'] == predictionFinal['Label']).count()
totalData = predictionFinal.count()
print("correct prediction:", correctPrediction, ", total data:", totalData,
     ", accuracy:", correctPrediction/totalData)
```

Technologies used:

- ✓ Python
- ✓ Pyspark
- ✓ Machine learning
- ✓ Hadoop
- ✓ IBM DB2
- ✓ Tableau

Conclusion:

- ✓ Big Data Analytics on the IBM Cloud offers a powerful and comprehensive solution for processing, analyzing, and deriving actionable insights from large and complex datasets.
- ✓ The integration of IBM's cloud infrastructure with cutting-edge analytics tools and technologies provides organizations with the scalability, performance, and agility needed to effectively manage and extract value from their data resources.
- ✓ In conclusion, Big Data Analytics on the IBM Cloud stands as a robust solution, combining the power of scalable infrastructure, advanced analytics capabilities, and security features.
- ✓ This integration equips organizations to harness the full potential of their data, drive innovation, gain competitive advantages, and make data-driven decisions critical for their success and growth in today's data-centric landscape.