ProjectPro

End-to-end solved Projects in Data Science and Big Data

ProjectPro

Table of Contents

ProjectPro

Data Science Projects

ProjectPro

Sales Forecasting

Overview of the Project

- For any departmental store, it is important to have a rough idea of their sales, so that they can plan their inventory accordingly.

- Determining the sale of high-selling and low-selling products of each category is also key for inventory planning.

Problem Statement:

- In this data science project, you will use R programming language to predict the sales of each department of a store using the Walmart dataset.

Data Description

- Walmart Dataset has sales data of 45 stores based on store, department and week.
- Size and type of each store has been mentioned.
- Holidays weeks have been provided.
- Price markdown data (almost like discount data) has been mentioned.
- A few macro-indicators like CPI, Unemployment rate, Fuel price etc. are also provided.

Learnings from the Project

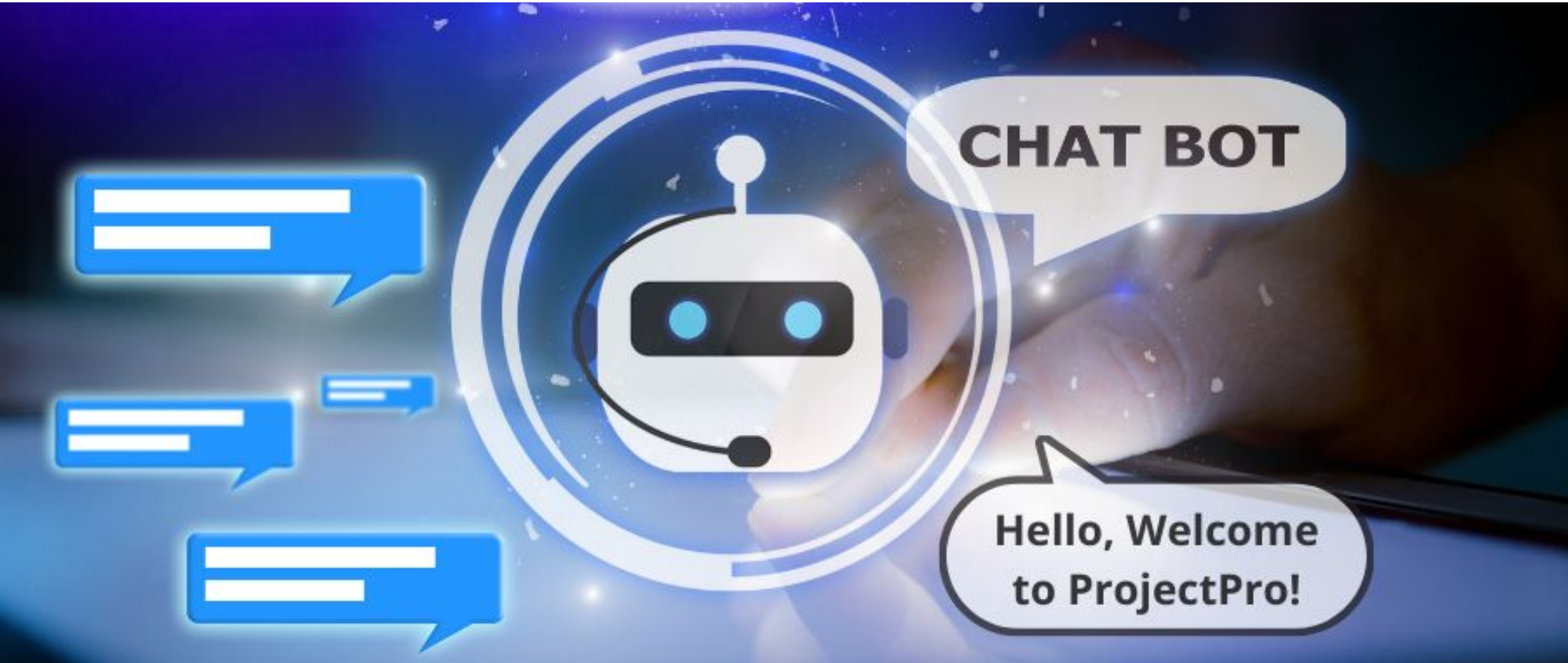
- Exploratory Data Analysis (EDA) techniques.
- Handling Missing values in a dataset.
- Using Univariate Analysis.
- Performing Bi-variate Analysis.
- Time Series ARIMA models Implementation.
- Using multiple metrics to compare the performance of different models.

Find the full solution of this project:
Walmart Sales Forecasting Data Science Project

ProjectPro

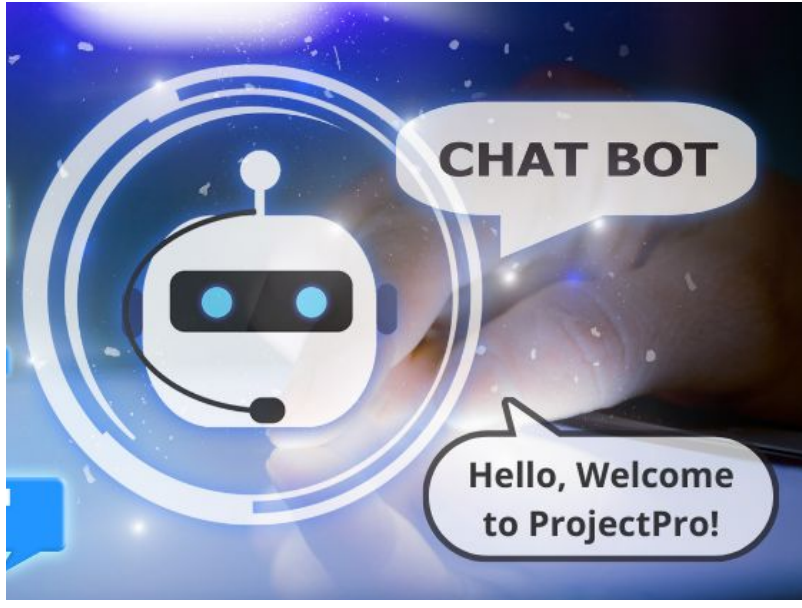
ProjectPro

Building a Chatbot Application



Overview of the Project

- For many small businesses, it is difficult to invest in human resources that can interact with their customers instantly and solve their queries.
- As an alternative to customer care team, they can make a chatbot to resolve their customers' queries automatically.



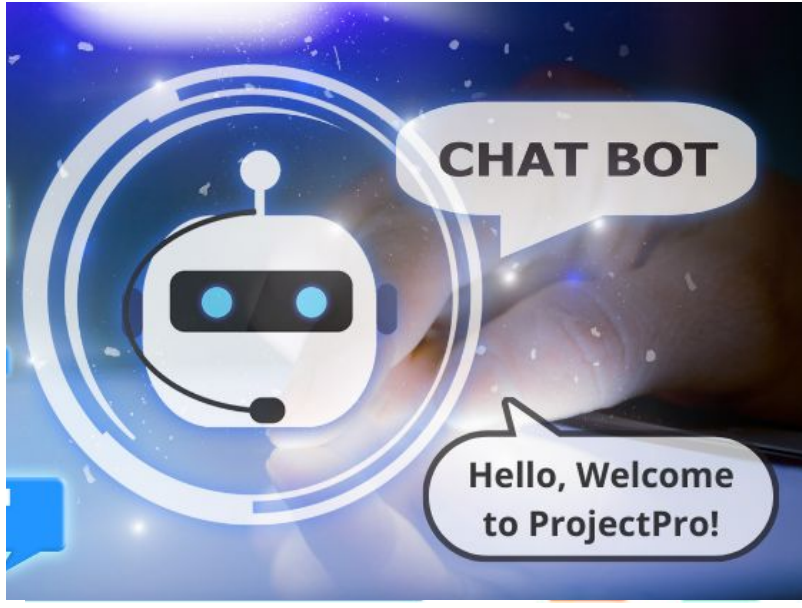
Problem Statement:

- In this data science project, you will apply Natural Language Processing techniques in Python programming language to build a Chatbot system.

ProjectPro

Data Description

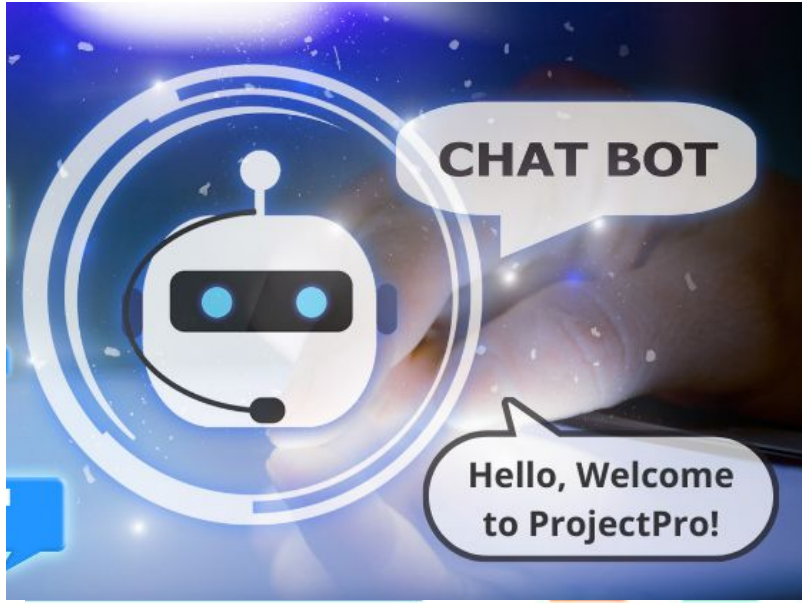
- The data is contained in a text file, named, 'leaves.txt'
- It has three types of values: text, category, answer.
- Text represents the input the user will pass to interact with the bot.
- Category is label used to differentiate different types of user queries.
- Answer represents the reply of the bot to the users' queries.
- The file has about 140 rows for each of the three data values.



ProjectPro

Learnings from the Project

- Introduction to NLTK library in Python
- NLP Techniques:
  - Tokenization
  - Lemmatization
  - Stemming
  - Removing Stopwords
  - Parts-of-Speech Tagging
- Bag-of-Words Model
- Decision Tree and Naive Bayes Classifier.
- Building an NLP-based chatbot engine that any UI can utilise.

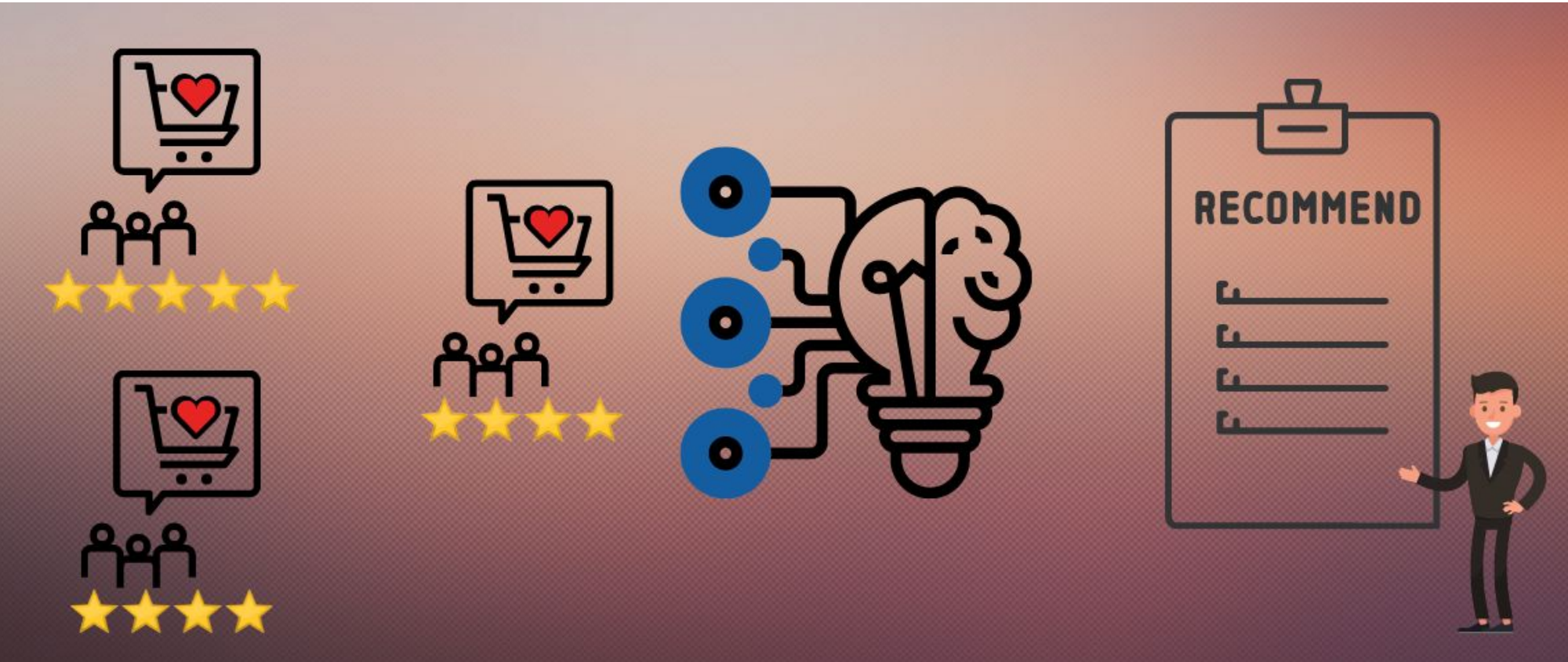


Find the full solution of this project:
Natural language processing Chatbot application using NLTK for text classification

ProjectPro

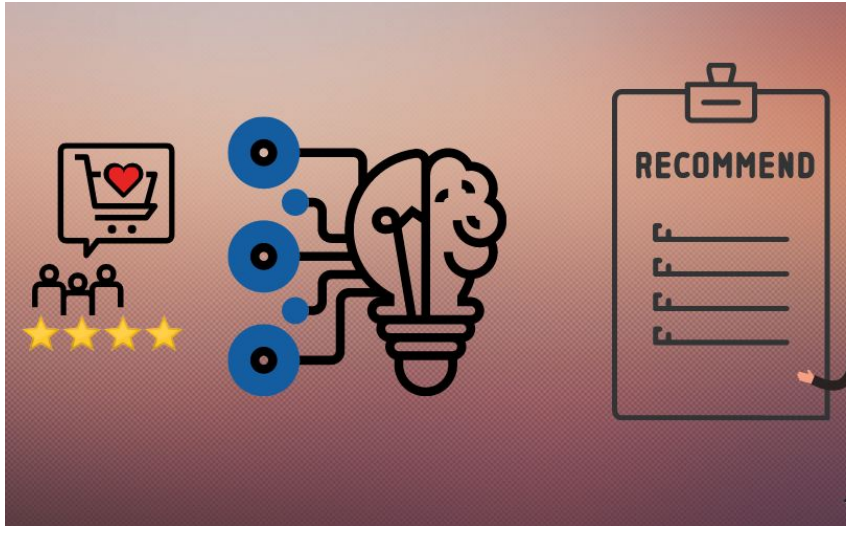
ProjectPro

Recommendation System



Overview of the Project

- While visiting a shopping mall, many salesmen often try to recommend the customers exciting deals and offers that might of interest them.
- Similarly, e-commerce sites use recommender systems to suggest products to their customers that they are highly likely to buy.

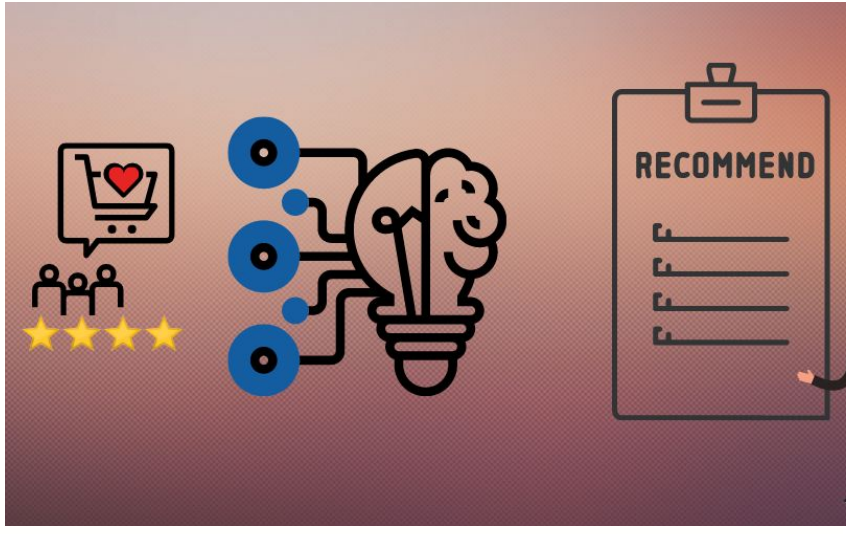


Problem Statement:

- In this data science project, you will build collaborative filtering algorithm based recommender system.

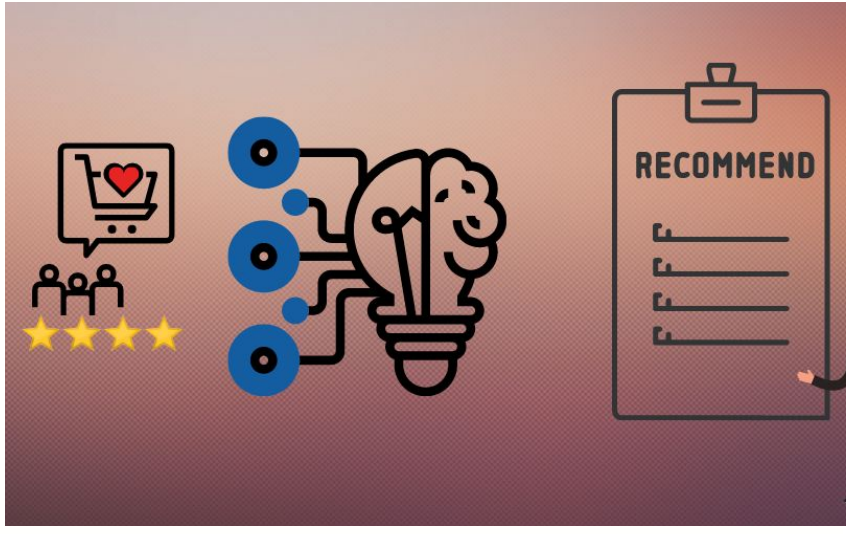
Data Description

- The data is contained in a csv file, named 'ratings_beauty.csv'.
- It has four types of values: userid, productid, ratings, and timestamp.
- UserId and Productid are used for identification purposes.
- Ratings are the feedback provided by the user corresponding to a product in that row
- Timestamp represents the time at which user submitted the rating.
- The file has about 2 million reviews and ratings for beauty products available on Amazon.



Learnings from the Project

- Introduction to Recommender Systems
- Popular Exploratory Data Analysis (EDA) Techniques
- Data Visualization
- Data Encoding Methods
- Cosine Similarity and Centres Cosine Similarity
- User-Item Matrix
- Ways to identify similar customers



Find the full solution of this project:
Build a Collaborative Filtering Recommender System in Python

ProjectPro

Market Basket Analysis

Overview of the Project

- Whenever customers purchase certain products from a store, it is important for the store to understand their buying patterns. This can help stores in better placement of specific products.
- The way to understand these patterns is called Market Basket Analysis/

Problem Statement:

- In this data science project, you will use Apriori and FP growth algorithms to understand Market Basket Analysis.

ProjectPro

Data Description

- The data is contained in 4 csv files,
- ‘customer.csv’ has the customer details
- ‘product.csv has the product information
- ‘product_class.csv’ has the details of product department
- ‘region.csv’ has information about the location where the store is located.
- ‘sales.csv’ contains the details of sales of each product.
- ‘stores.csv’ has more information about the stores.
- ‘time_by_day.csv’ has time of the day when an item was purchased from the store.

Learnings from the Project

- Introduction to Market Basket Analysis/ Product Association Analysis
- Apriori Algorithm
- Fpgrowth Algorithm
- Bivariate Analysis
- Feature Analysis
- One Hot Encoding
- User-Item Matrix

Find the full solution of this project:
[Customer Market Basket Analysis using Apriori and Fpgrowth algorithms]

ProjectPro

ProjectPro

Resume Parsing Application

Overview of the Project

- Most companies these days wants to efficiently utilise their human resources by automating tasks wherever possible.
- The efficiency of HR Management team members can be improved by providing them a Resume Parsing system which can shortlist resumes automatically.

Problem Statement:

- In this data science project, you will apply Natural Language Processing techniques in Python programming language to build a Resume Parsing system.

Data Description

- The data for this project is available in JSON file format.
- The file has resume content in the form of text.
- The content from the resume has been labelled (skills, location, etc. of the applicant) for your convenience.
- This format is not what Spacy is served with.
- You will learn how to make the given data Spacy-friendly.

Learnings from the Project

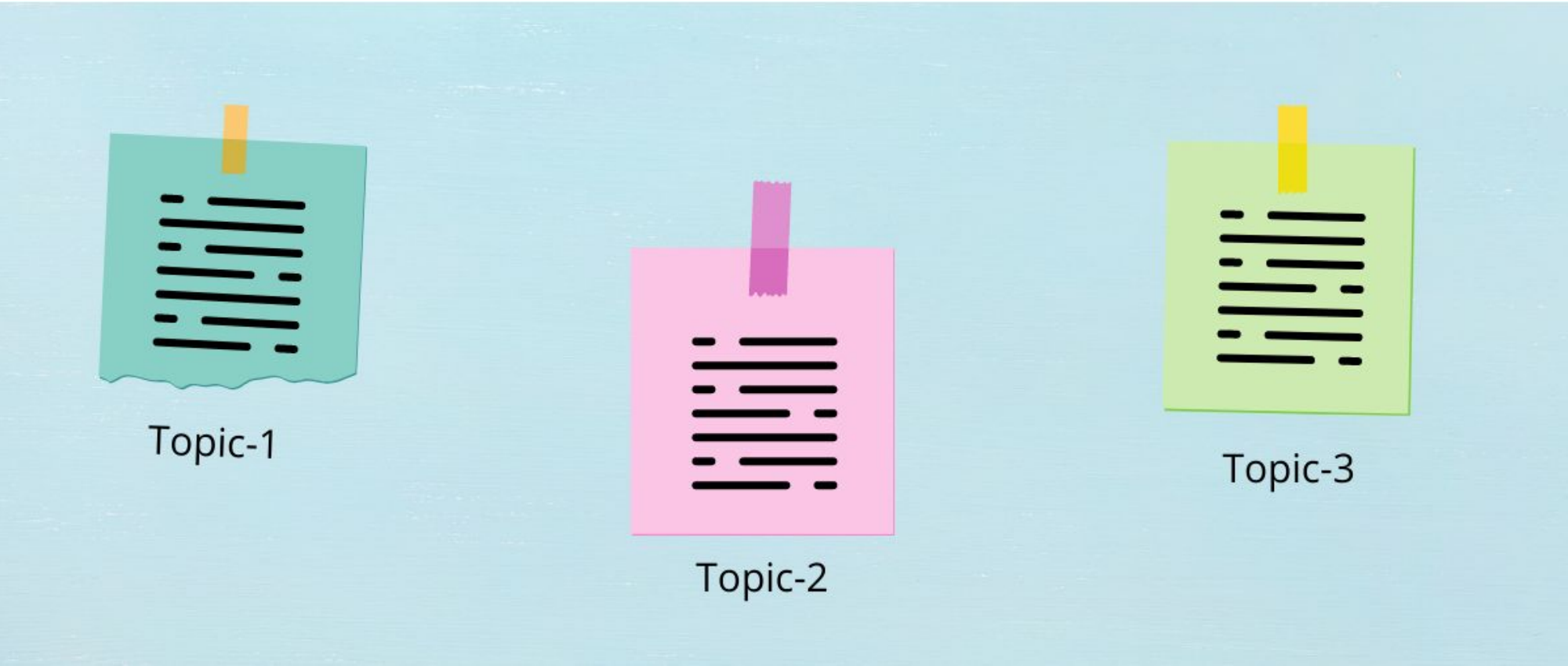
- Introduction to Natural Language Processing (NLP) and generic Machine learning workflow
- Introduction to Spacy NLP Library in Python
- Utilising Annotations and Entities in Spacy
- Text Classification
- Optical Character Recognition
- Text extraction from PDFs
- TIKA OCR Procedure

Find the full solution of this project:
Resume parsing with Machine learning - NLP with Python OCR and Spacy

ProjectPro

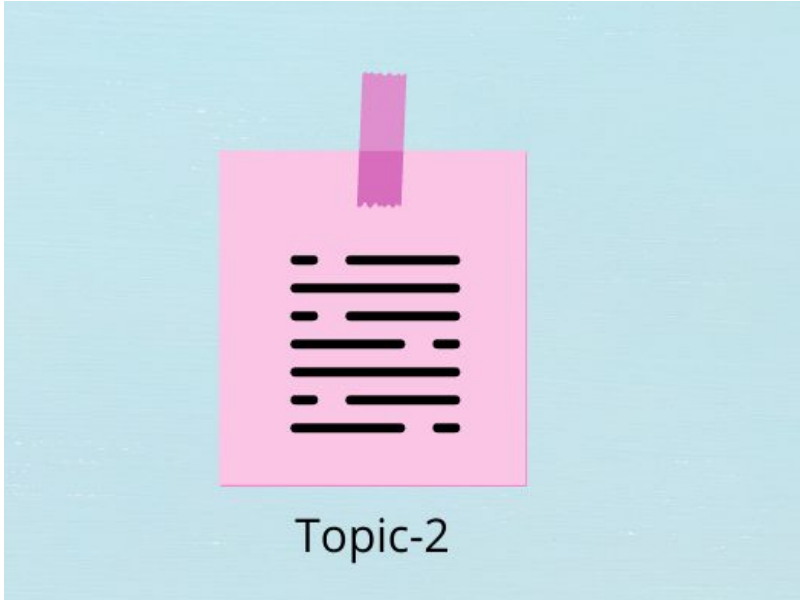
ProjectPro

Topic Identification



Overview of the Project

- When exploring NLP machine learning algorithms, an interesting application is found in projects titled Topic Identification.
- Here you are a given a document with a certain set of words and the task is to label that document with a title that best describes its content.

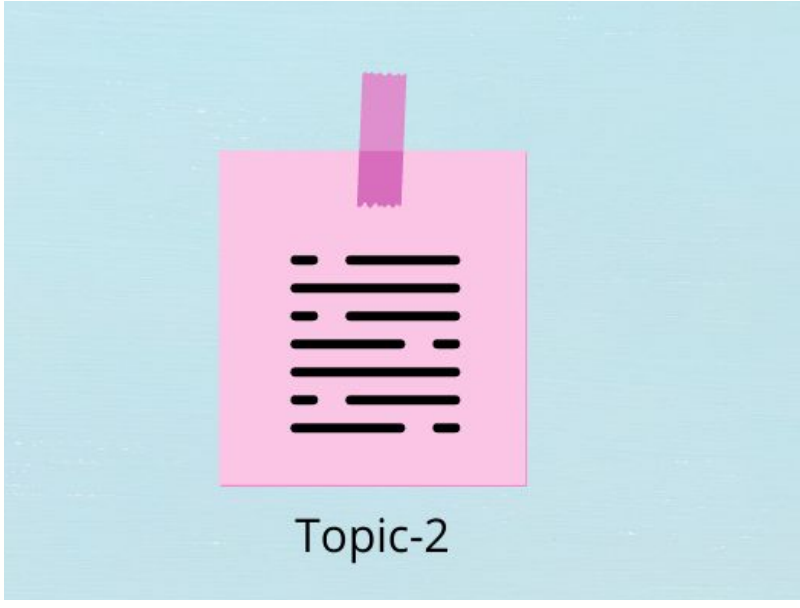


Problem Statement:

- In this data science project, you will apply Natural Language Processing techniques in Python programming language to build a Topic Modelling system.

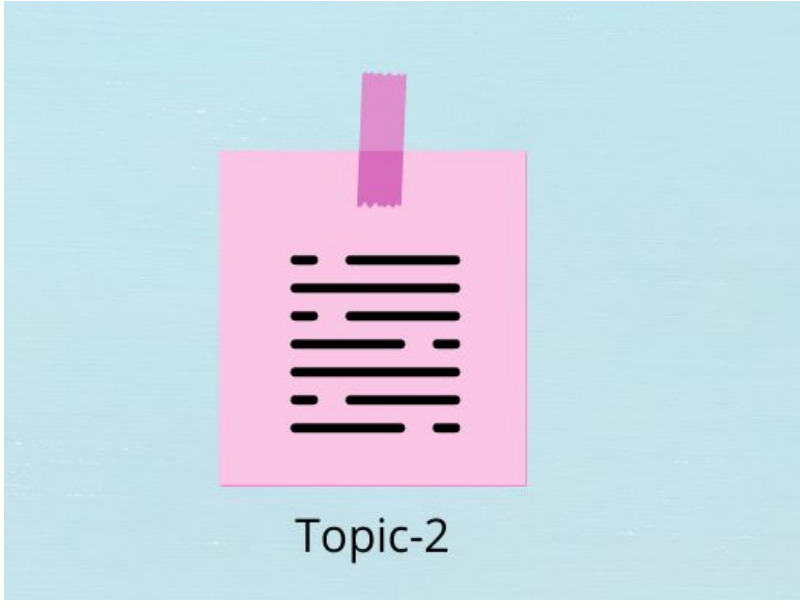
Data Description

- This project will utilise tweets from twitter to explain Topic Modeling in Python
- The data will be provided to you in csv format.
- It has four types of data values: username, tweets, date, mentions.
- The username is the unique twitter handle of a user whose tweet we will be using.
- Tweets has the content of a tweet by the corresponding user.
- Date specifies the date of the tweet creation.
- Mentions has details of the other twitter handles that have been referred to in a tweet.



Learnings from the Project

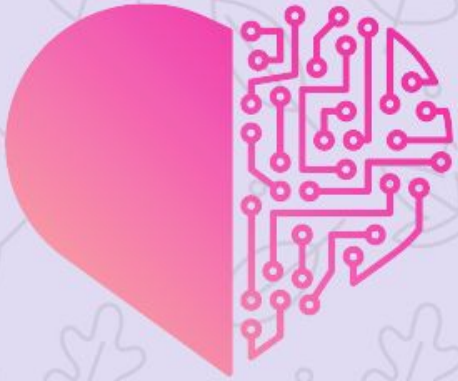
- Exploring various methods of Natural Language Toolkit (NLTK) library in Python
- Cleaning and Analysing Textual data
- Converting unstructured data to structured data
- K-Means Clustering Machine Learning Algorithm
- Clustering text from Twitter
- Tokenization of text into words
- Identifying topic of the given text



Find the full solution of this project:
Topic modelling using Kmeans clustering to group customer reviews

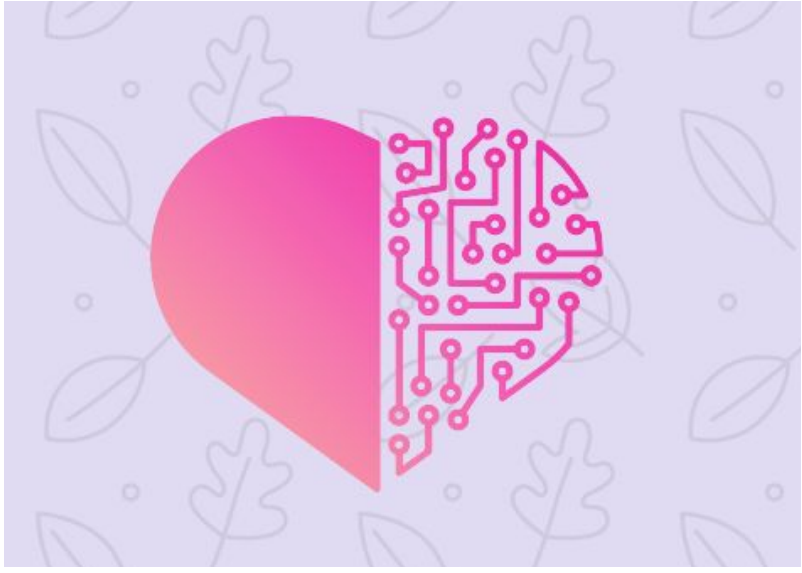
ProjectPro

Sentiment Analysis & Ranking



Overview of the Project

- For businesses to grow and evolve, it is crucial to analyse their customer views in order to understand their needs.
- The analysis can help them in taking decisions that can lead to an exponential growth in their sales.



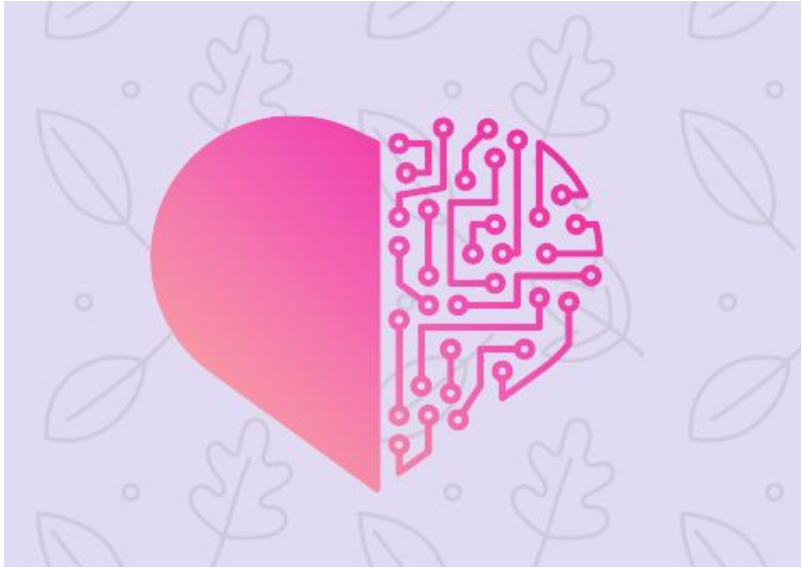
Problem Statement:

- In this data science project, you will apply Natural Language Processing techniques in Python programming language to build a Customer Reviews Sentiment Analysis and Ranking system.

ProjectPro

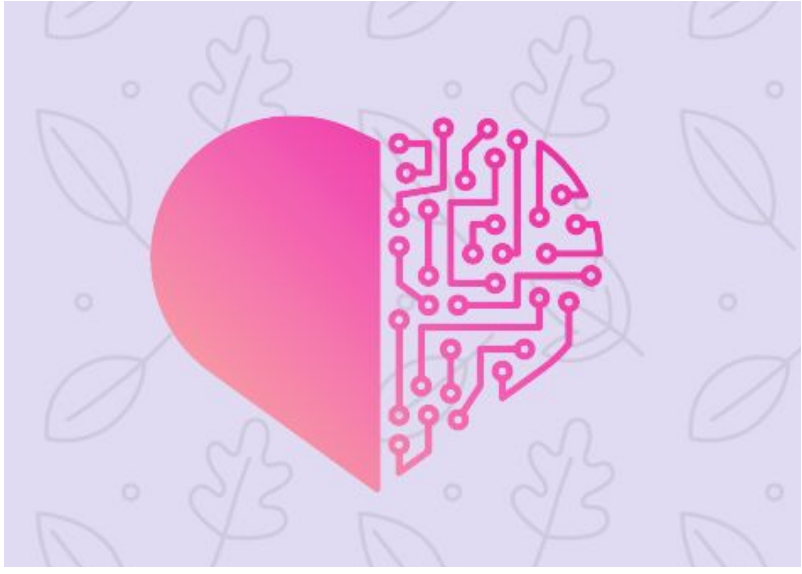
Data Description

- The data for this project will be provided to you in
- The ‘train.csv’ file has 1676 rows and 3 columns.
- The columns contain information about specific products, their reviews and a label for each review.
- There are about 8 different products whose reviews have been listed.
- The label 0 is used to indicate non-informative reviews and 1 indicates the review are informative.



Learnings from the Project

- Implementing Data Preprocessing techniques of on customer reviews data
- Using Markov Chain method to extract gibberish
- Gathering text for features from reviews
- Performing Sentiment Analysis over the reviews
- Implementation of TF-IDF Text Vectorizer
- Using Random Forest Machine Learning Algorithm for pairwise ranking of reviews



Find the full solution of this project:
Ecommerce product reviews - Pairwise ranking and sentiment analysis

ProjectPro

ProjectPro

Loan Eligibility Prediction

Overview of the Project

- Various banks receive a large number of loan applications everyday. Thus, a lot of time is invested in analysing these applications.
- The task of analysing applications can be automated using Machine learning algorithms.

Problem Statement:

- In this data science project, you will apply machine learning algorithms in Python to predict the eligibility of a loan applicant to repay it.

ProjectPro

Data Description

- The data is contained in a CSV file.
- It has 1111107 rows and 19 columns.
- It contains the following details:

  Loan ID, Customer ID, Loan Status, Current Loan Amount, Term, Credit Score, Years in current job, Home Ownership, Annual Income, Purpose, Monthly Debt, Years of Credit History, Months since last delinquency, Number of open accounts, Number of credit problem, Current Credit Balance, Maximum Open Credit, Bankruptcies, Tax Liens.

Learnings from the Project

- Utilising different libraries in Python and their significance
- Building custom functions for utilising ML algorithms
- Data Procurement
- Dealing with missing values in data
- Preprocessing data before the application of ML algorithms
- Using Gradient Boosting and XGBoost
- Calculating various metrics to identify the best model

Find the full solution of this project:
Loan Eligibility Prediction using Gradient Boosting Classifier

ProjectPro

ProjectPro

Retail Price Optimization

Overview of the Project

- A significant amount of time is spent by retail store owners in deciding the price of an item. There are many aspects of the items they have to keep in mind while deciding the prices.
- A good analysis of various characteristics of all the items can assist in optimizing the retail prices.

Problem Statement:

- In this data science project, you will analyse sales data of Cafe and predict prices of their items using ML algorthims.

ProjectPro

Data Description

- The data is contained in three CSV files.
- First is ‘Cafe - Sell Meta Data.csv’. This file has details about sales made by the cafe.
  Columns: Sell ID, Sell Category, Item ID, Item Name
- Next is ‘Cafe - Transaction - Store.csv’. This file contains information about transactions and sale receipts of the cafe.
  Columns: Calendar Date, Price, Quantity, Sell ID, Sell Category
- And, the last is ‘Cafe - DateInfo.csv’. This has date information corresponding to the transactions performed.
  Columns: Date, Year, Holiday, Weekend, School Break, Temperature, Outdoor

Learnings from the Project

- Introduction to Retail Price Optimization problem in Machine Learning
- Understanding price elasticity of demand
- Working with Jupyter Notebooks
- Making generic codes for price optimisation for different items
- Methods of choosing the best prediction model
- Predicting Price elasticity of demand for all items

Find the full solution of this project:
Machine Learning project for Retail Price Optimization

ProjectPro

ProjectPro

Driver Availability Prediction

Overview of the Project

- Covid-19 forced people to stay indoors and the food delivery apps thus noticed a surge in orders.
- Estimating delivery charges is not an easy task for the delivery apps companies as it depends on complicated features.

Problem Statement:

- In this data science project, you will machine learning techniques in Python programming language to efficiently allocate drivers the orders for delivery.

Data Description

- The data is contained in three CSV files: ‘pings.csv’, ‘drivers.csv’, ‘test.csv’
- ‘pings.csv’ has information about when a certain driver received a ping. So, it has two columns driver_id and timestamp.
- ‘drivers.csv’ has biodata (driver ID, gender, age, number of kids) of about 2500 drivers.
- ‘test.csv’ will be used for testing the model that we will help you build in this project.

Learnings from the Project

- Transforming a time series problem to a supervised learning problem
- Introduction to Multi-step time series forecasting
- Concept of Lead-Lag and Rolling Mean
- Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) in Time Series
- Understanding recursive multi-step prediction strategy
- Using Random Forest and XGBoost for predicting online hours of a driver

Find the full solution of this project:
Demand prediction of driver availability using multistep time series analysis

ProjectPro

ProjectPro

Big Data Projects

ProjectPro

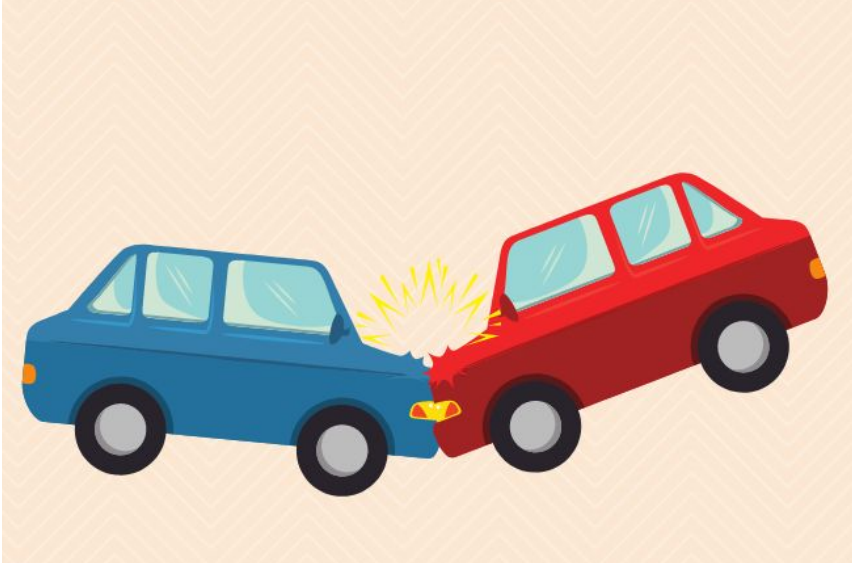
Event Data Analysis

Overview of the Project

- When aiming for the role of a Big Data Engineer, it is important to have experience with real-time streaming data.

Problem Statement:

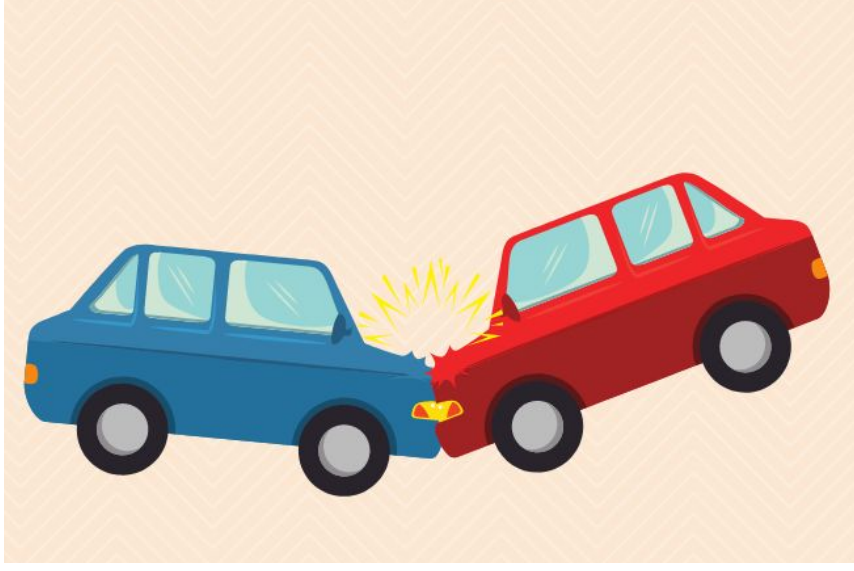
- In this Big Data project, you will learn how to extract real time streaming event data from the given dataset API. The dataset that will be provided is about accidents in New York City.



ProjectPro

Data Description

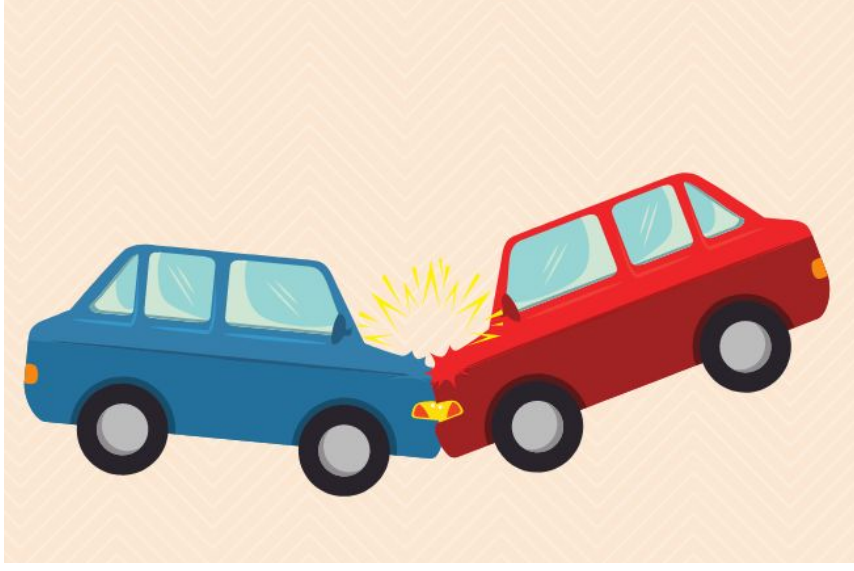
- The Dataset that will be used for this Big Data project is Motor Vehicles Collisions Dataset that is updated regularly.
- The dataset is open source using the API the host website provides.
- There are about 29 columns in the dataset containing information about the accidents like, location latitude and longitude, time of the crash, etc.
- The dataset will be available to your in the JSON format when you will use its API.



ProjectPro

Learnings from the Project

- Introduction to the following:
  - i. Apache Nifi
  - ii. Apache Spark
  - iii. AWS Elk Stack Elasticsearch
  - iv. Apache Kibana
  - v. Logstash
- Storing data in Apache Hadoop Distributed File system (HDFS)
- Using PySpark for Data Exploration and Data Analysis

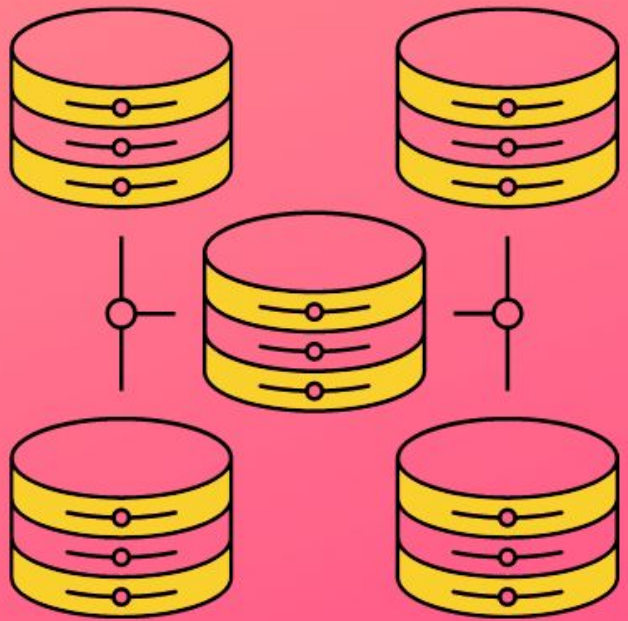


Find the full solution of this project:
Event Data Analysis using AWS ELK Stack

ProjectPro

ProjectPro

Build a Big Data Pipeline

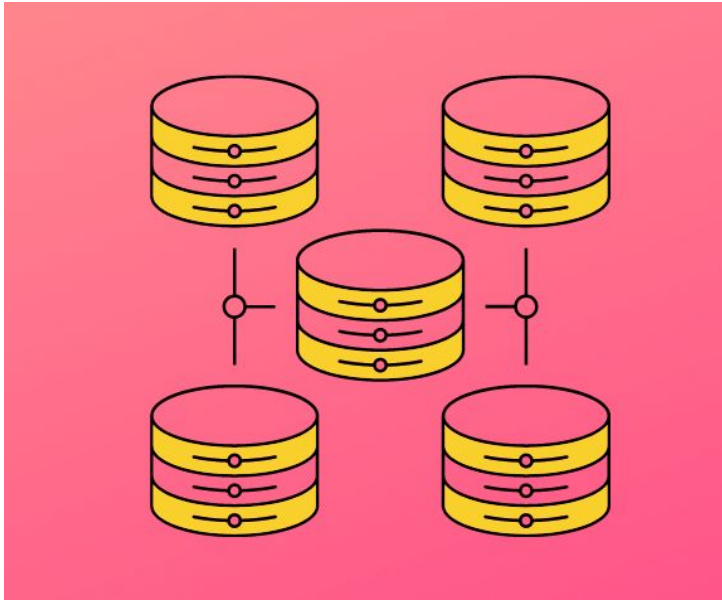


Overview of the Project

- Decisions in the aviation industry are highly data-driven. It is not only the flight timings and locations that are relevant, but also the load of passengers, the weight of their luggage, weather, security, time of the year, specifications of the aircraft, etc.

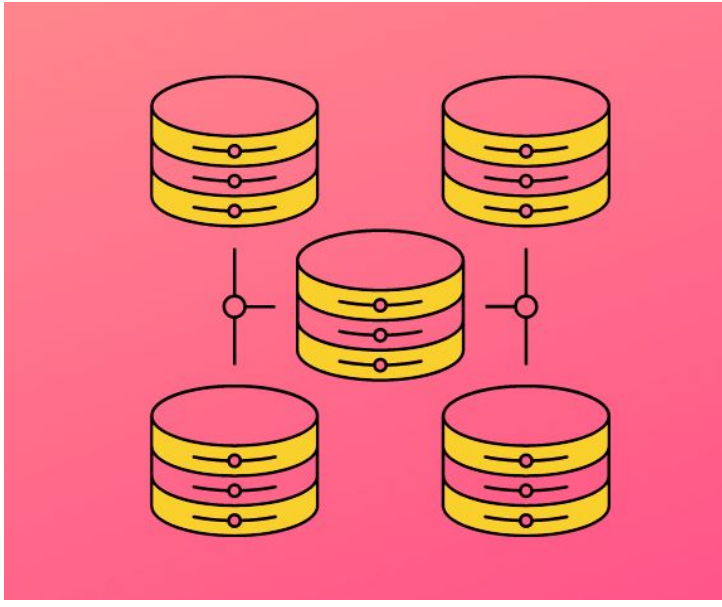
Problem Statement:

- In this Big Data project, you will learn how will work on an aviation dataset to create a big data pipeline at scale on AWS.



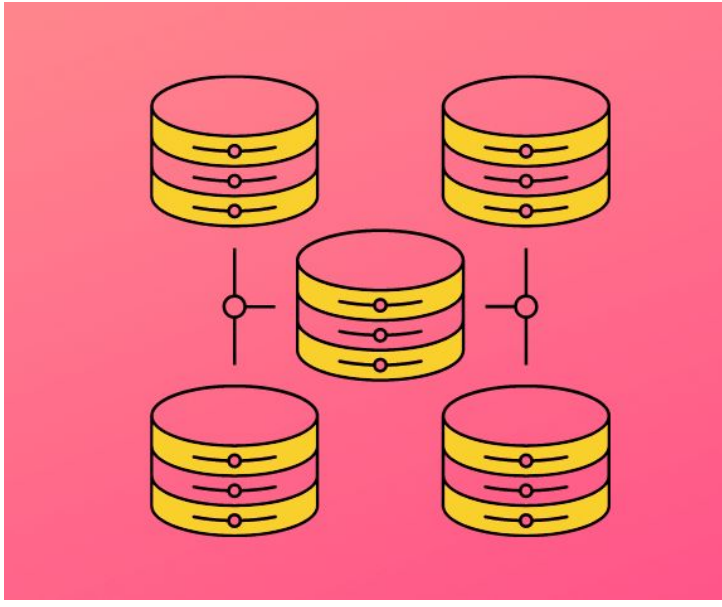
Data Description

- In this Big data project you will work with a dataset provided by an API from ICAO.
- The dataset you will be working with is called Incidents.
- The data format will be JSON.
- It has 20+ features which include State of Occurence, Time, Location, Operator, Model, State of Registry, Flight Phase, Fatalities, etc.
- After hitting the URL of the website, you will be able to access this data.



Learnings from the Project

- Introduction to Big Data and Big Data Pipeline
- Explore Data Architecture using Nifi, Kafka, and Hive
- Apache Kafka Vs Apache Flume
- Optimisation Techniques in Apache Hive
- Understanding HDFS and Druid
- Transferring Data from NiFi to Kafka
- Using MySQL and AWS Quicksight

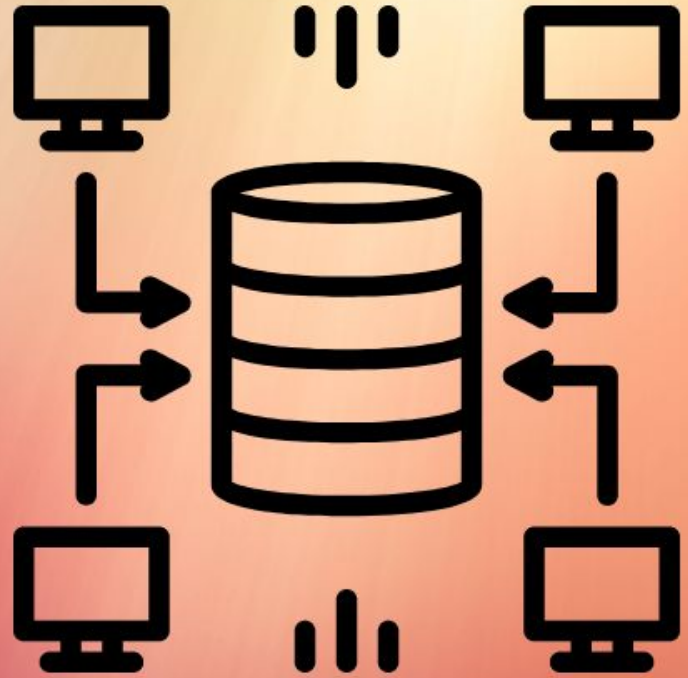


Find the full solution of this project:
Build a big data pipeline with AWS Quicksight, Druid, and Hive

ProjectPro

ProjectPro

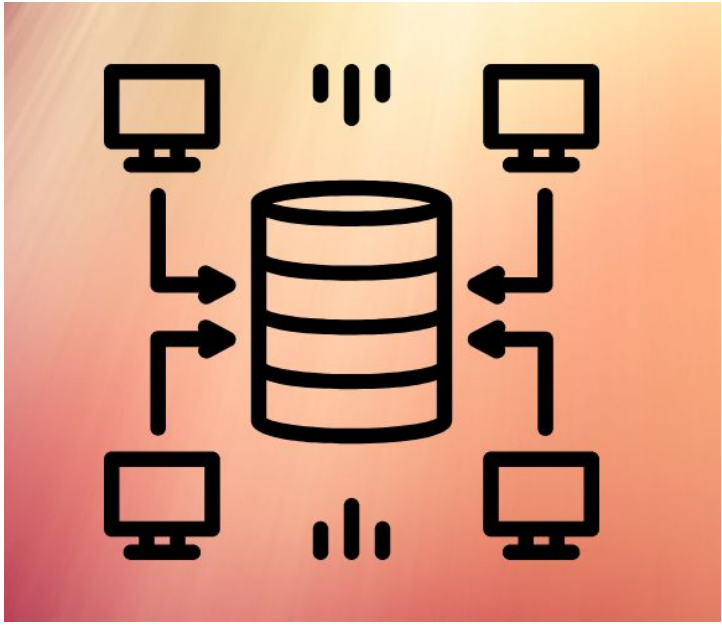
Build an ETL Data Pipeline



- EXTRACT
- TRANSFORM
- LOAD

Overview of the Project

- This project is a follow up of the previous project. However, it is not same. You will be introduced to other technologies widely used by Big Data Engineers, that were not covered in the previous project.



Problem Statement:

- In this Big Data project, you will learn how will work on a sales dataset to create a big data pipeline at scale on AWS.

Data Description

- In this project, you will not retrieve data from a website. Rather, the dataset will be provided to you in a CSV format.
- The filename is ‘sales_data.csv’.
- There are 14 columns in the dataset which include geographical information about the purchase of products, item type, mode of purchasing (online or offline), unique IDs of the products, shipping date of the order, quantity of the sold products in a particular order, a label for priority of the order, etc.

ProjectPro

Learnings from the Project

- Implementing a big data pipeline on AWS via Software As A Service (SAAS) method
- Building scalable and reliable architectures.
- Comparison of utilisation of IAAS, SAAS, and PAAS wrt Big Data
- Extracting raw data of sales int AWS S3
- Using EMR Hive and Tableau Desktop together
- Visualization of data using Tableau

Find the full solution of this project:
AWS Project - Build an ETL Data Pipeline on AWS EMR Cluster

ProjectPro

ProjectPro

Market Basket Analysis

Overview of the Project

- Bitcoins are gradually becoming more popular. So much that even of the world's richest person, Elon Musk, has started tweeting about it.
- People are considering bitcoin as a currency of the future.

Problem Statement:

- In this Big Data project, you will implement Bitcoin Mining on Amazon Web Services using exciting Big Data tools.

ProjectPro

Data Description

- In this Big Data project, you will work with Bitcoin Dataset.
- You will use Python API to extract the data from a URL.
- The dataset on the URL has information about various cryptocurrencies available like Dogecoin, Litecoin, Bitcoin, etc.
- For each cryptocurrency, you will be having its ID, name, name_id, its price in US Dollars, percentage change in its price over the past hour and past 24 hours price in bitcoin, how much market it is capturing, etc.

ProjectPro

Learnings from the Project

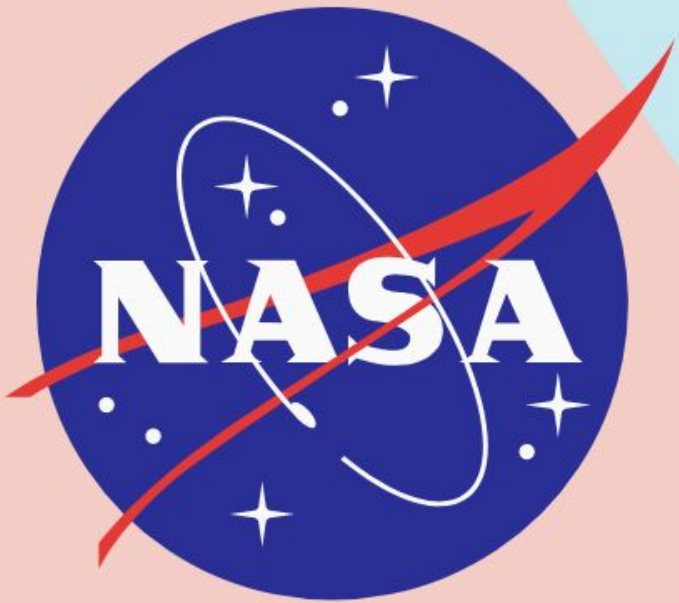
- Introduction to Data Warehousing in Hive and Spark
- Using AWS Quicksight for visualising data
- Using Python API for extracting data
- Uploading data from Ec2 instance to HDFS
- Understanding PySpark
- Creating Tables in Apache Hive
- Working with Apache Hadoop and Apache Spark

Find the full solution of this project:
Bitcoin Data Mining on AWS Free Tier

ProjectPro

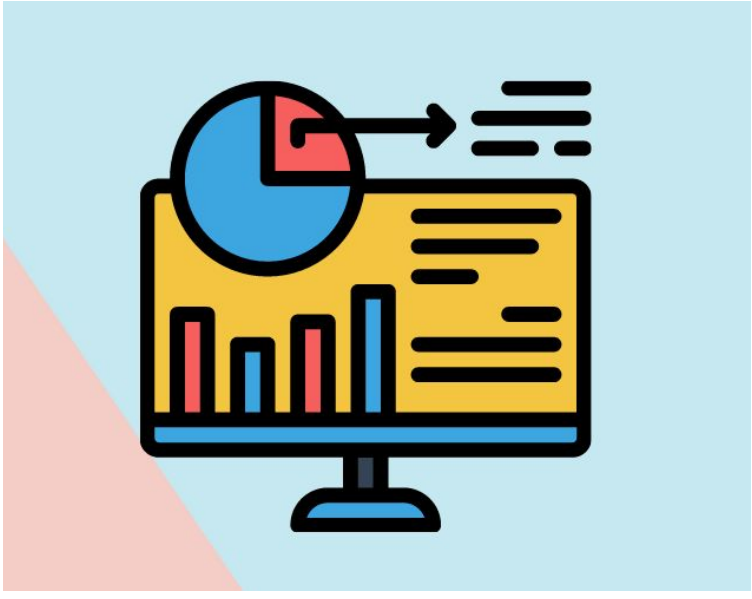
ProjectPro

Analysing NASA Log Files



Overview of the Project

- You have must have heard of logbooks when you were in school. Teachers usually maintain a logbook to keep a track of daily events in the classroom.
- Similar to that, a computer also stores information about events in log files.



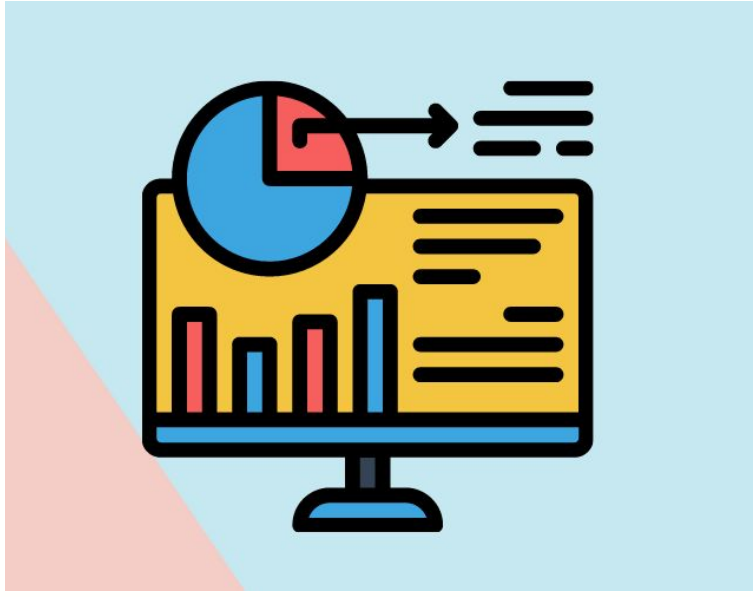
Problem Statement:

- In this Big Data project, you will work with practical data of logs obtained from NASA Kennedy Space Center WWW Server. techniques in Python programming language to build a Resume Parsing system. You will use Big Data tools to perform scalable analytics over the data.

ProjectPro

Data Description

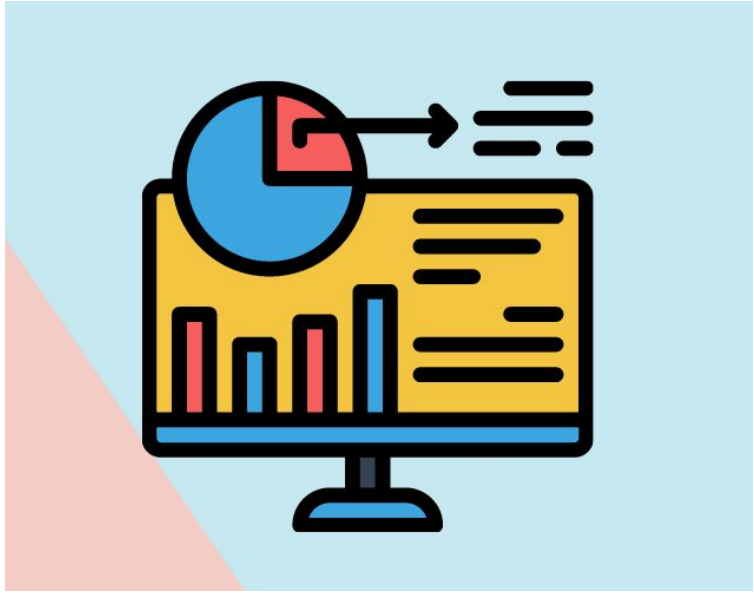
- In this Big Data project, you will work on a dataset that is in CSV file format.
- The file has information about NASA web logs in 1995.
- It has remote host name, the time stamp for access, then the request type, the request path. In addition, it contains the details of request status and the size of the data that was requested.
- The dataset does not contain information about user authentication name unlike usual log files.



ProjectPro

Learnings from the Project

- Using NiFi to extract dataset from servers
- Using Apache Kafka for preparing topics and generating logs
- A detailed description of logs and their analytics
- Introduction to Lambda Architecture
- Understanding Docker, Port Forwarding
- Working together with Plotly, Dash, and Cassandra to display metrics live
- Loading Data in Cassandra



Find the full solution of this project:
Log Analytics Project with Spark Streaming and Kafka

ProjectPro

ProjectPro

Movie Recommendation

Overview of the Project

- Covid-19 has changed the world of cinema. Netflix and Amazon Prime are the popular OTT platforms that have replaced cinema halls.
- While using these platforms, one doesn't need to search for movies to watch, because the app recommends on its own.

Problem Statement:

- In this Big Data project, you will use Big Data tools to perform analysis over a movies dataset and build a recommendation system using it.

Data Description

- In this Big Data project, you will work on a dataset by Movielens.
- To implement the solution of this project, you can work with any dataset from the MovieLens.
- Their 25M dataset has movies reviews from about 1,62,000 users for 60K+ movies.
- The dataset will be provided to you in zip that contains several CSV files each having information about movies, ratings, tags, etc.

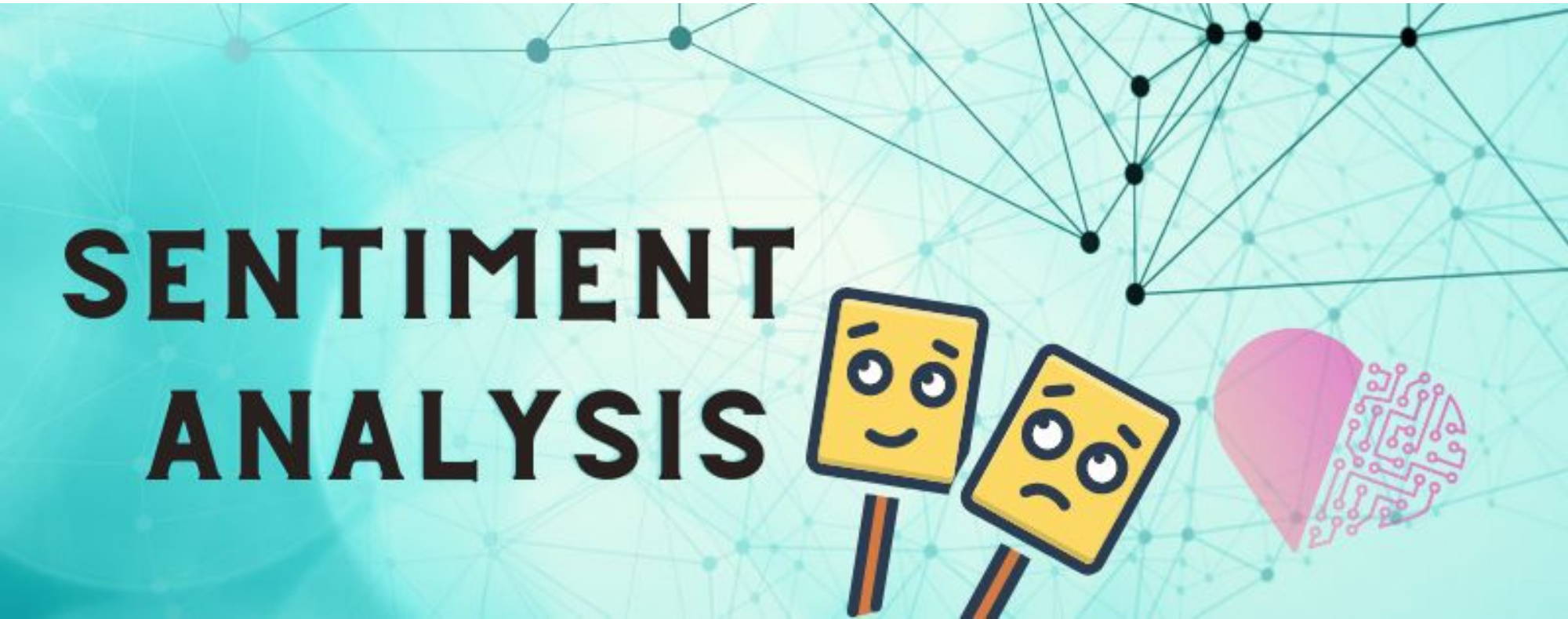
Learnings from the Project

- Implementing a big data pipeline on Microsoft Azure and Databricks Spark
- Obtaining subscription of MS Azure
- Preparing Resource Group
- Introduction to Azure Data Factory, Azure Databricks and Storage Account on Azure
- Building and executing ADF Pipelines
- Using Spark SQL on Databricks
- Deploying models using Flask API

Find the full solution of this project:
Movielens dataset analysis for movie recommendations using Spark in Azure

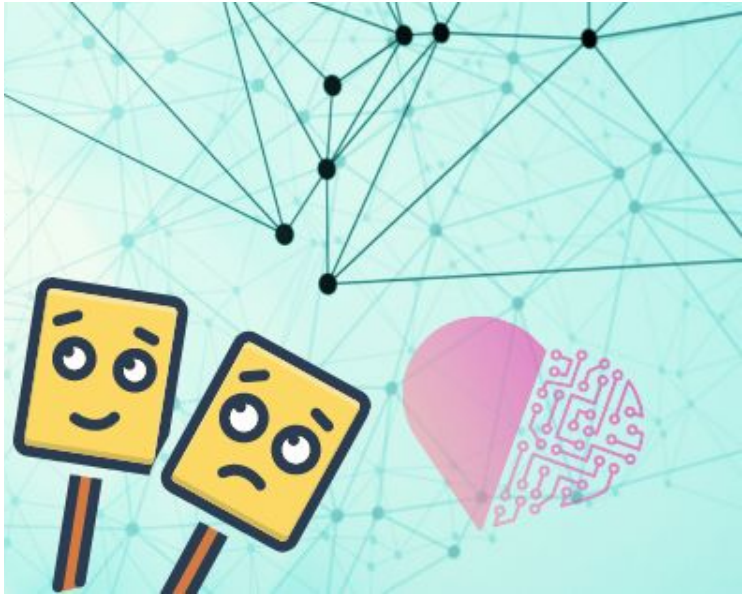
ProjectPro

Sentiment Analysis



Overview of the Project

- As discussed before in this document, understanding sentiments of the customers is important if the businesses intend to seek growth.
- Manually reading and analyzing the sentiments is a difficult task, especially when they are large in number.

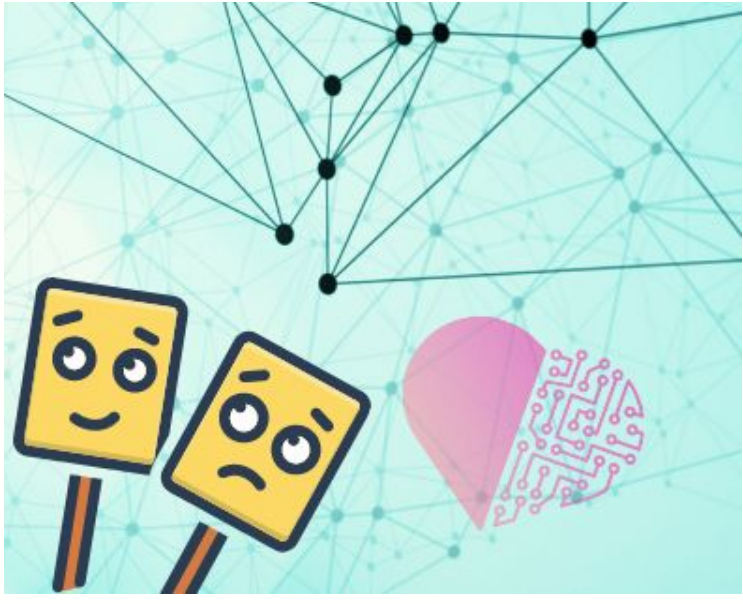


Problem Statement:

- In this Big Data project, you will work with a dataset and tweets to perform sentiment analysis over customer reviews.

Data Description

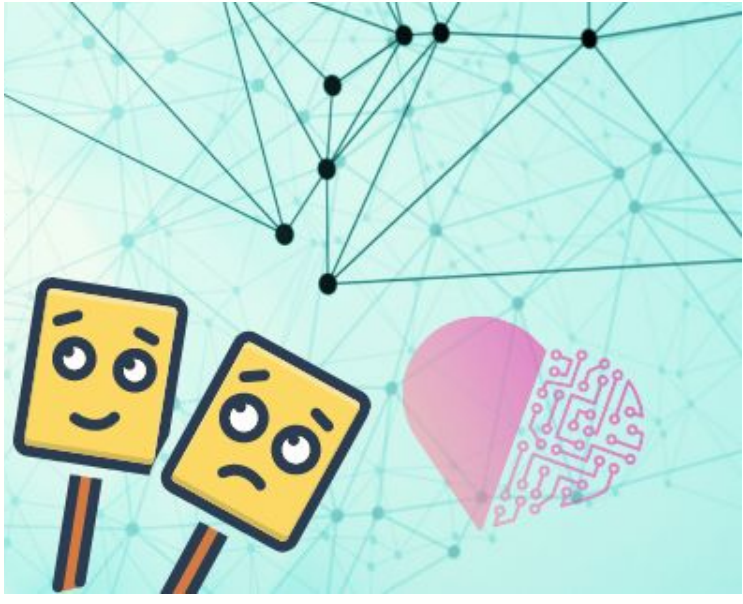
- The dataset you will be working on this project has product reviews from Amazon
- The dataset has 100M+ product reviews submitted from May 1996-July 2014
- You will learn how to work with dataset of a particular category of products: Sports and Outdoors.
- You will have following details of the reviews:
  whether the review is helpful or not, overall rating out of 5, content of the review, date and time on which the review was submitted, etc.



ProjectPro

Learnings from the Project

- Introduction to Containers and Sentiment Analysis
- Understanding complete architecture of the project
- Using Docker-composer
- Dividing data into buckets for labelling
- Data Ingestion using NiFi
- Using NiFi for producing tweets
- Using Kafka to read data
- Performing Sentiment Analysis using Spark
- Using MongoDB for storing data

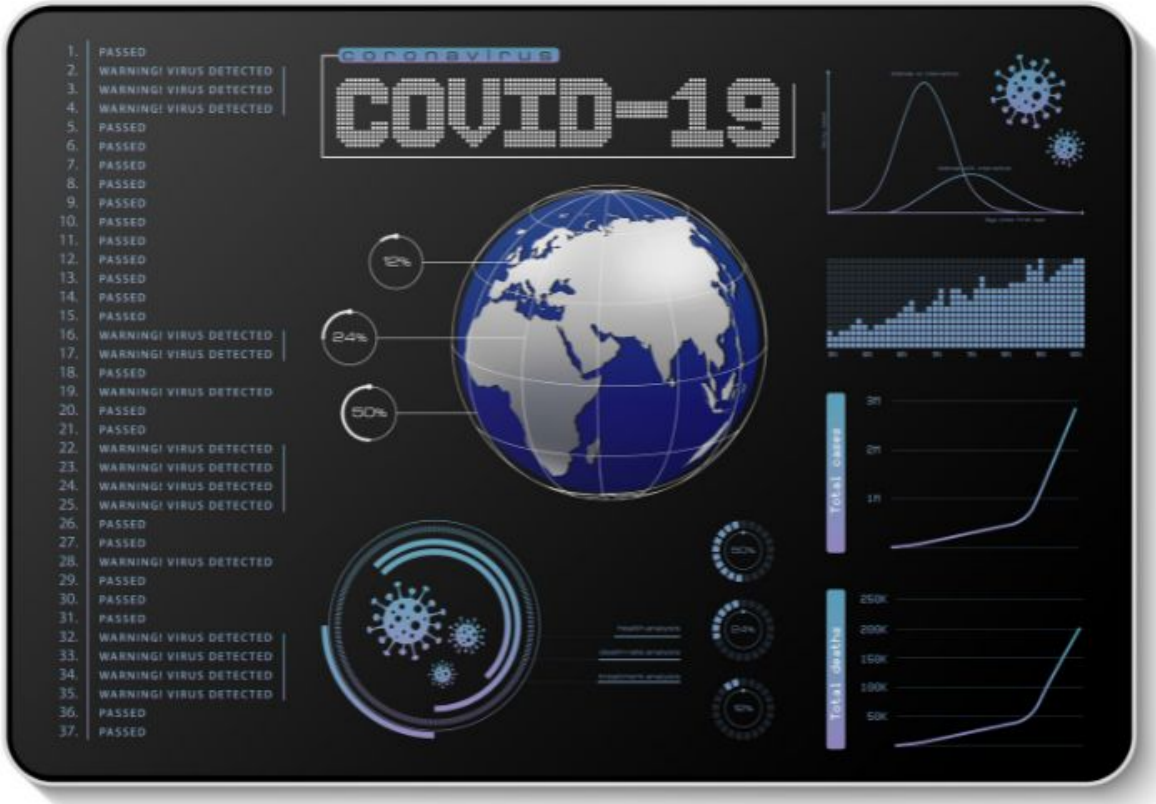


Find the full solution of this project:
Real-Time Streaming of Twitter Sentiments AWS EC2 NiFi

ProjectPro

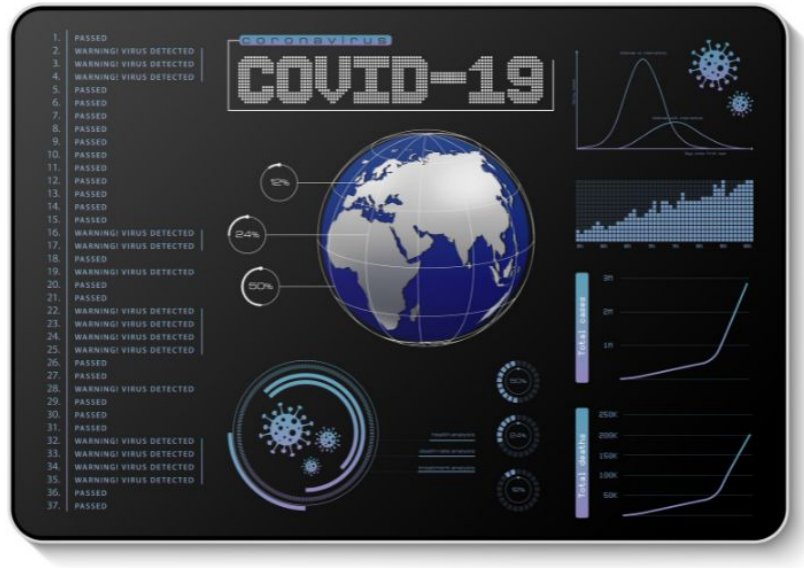
ProjectPro

Big Data Project using Covid Data



Overview of the Project

- During the mid-2020, most active users of LinkedIn who followed Big Data and Data Science projects, noticed that a large number of project ideas related to COVID-19 datasets were being implemented and shared on the website.



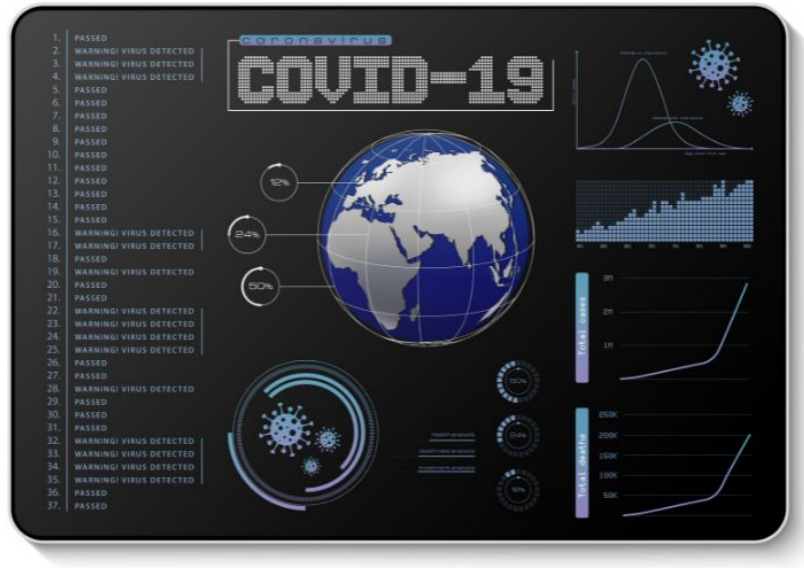
Problem Statement:

- In this Big Data project, you will build a Big data pipeline based on messaging. You will explore exciting Big Data tools by working on a COVID-19 dataset.

ProjectPro

Data Description

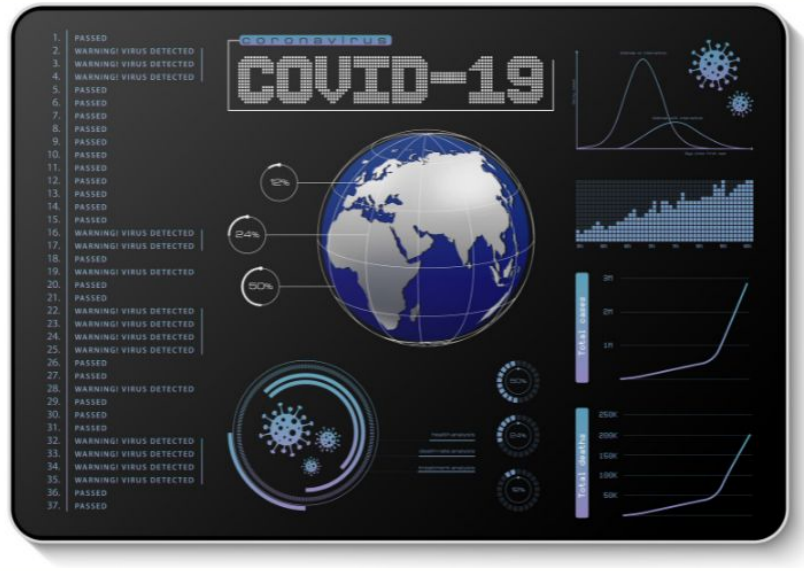
- In this Big Data project, you will use COVID-19 dataset obtained from an API.
- It is a summary dataset for COVID-19 patients worldwide.
- It contains information about active covid cases, total number of deaths, total number of people who have tested positive for the virus so far, total number of people who have recovered from the virus. This data is available for all countries and for the whole globe as well.
- Each country has a special variable called country code for easier representation.



ProjectPro

Learnings from the Project

- Implementing a Big Data pipeline on AWS
- Understanding how to use big data tools to automate tasks
- Using NiFi to import real time streaming data from an external API
- Transforming JSON file data into CSV using NiFi and storing them in HDFS
- Creating tables using Hive
- Analysing Data using Tableau and AWS Quicksight



Find the full solution of this project:
[Create A Data Pipeline Based On Messaging Using PySpark And Hive - Covid-19 Analysis]

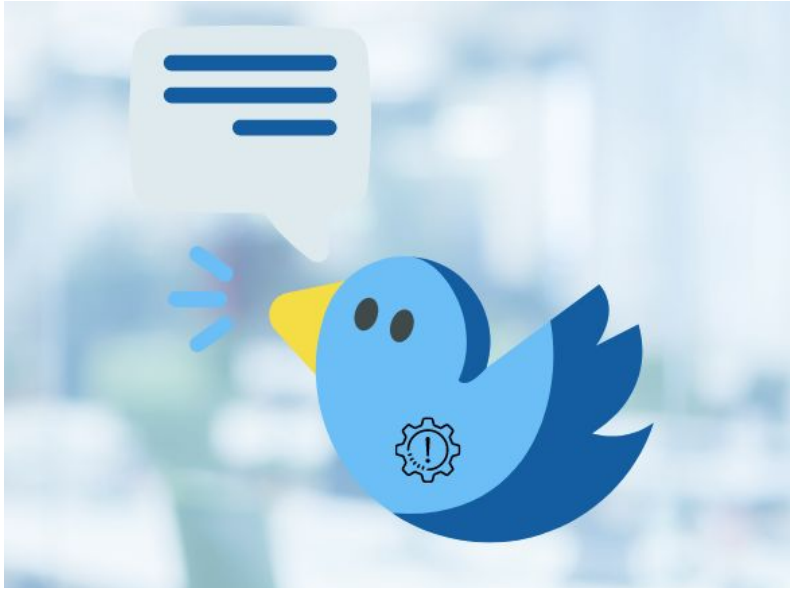
ProjectPro

ProjectPro

Auto-Reply Twitter Handle

Overview of the Project

- Over the years, social media platforms have evolved their position as a communication tool. They not only connect friends, but also businesses and customers.
- Twitter is one such platforms that has marked a distinct space in the domain of social media.



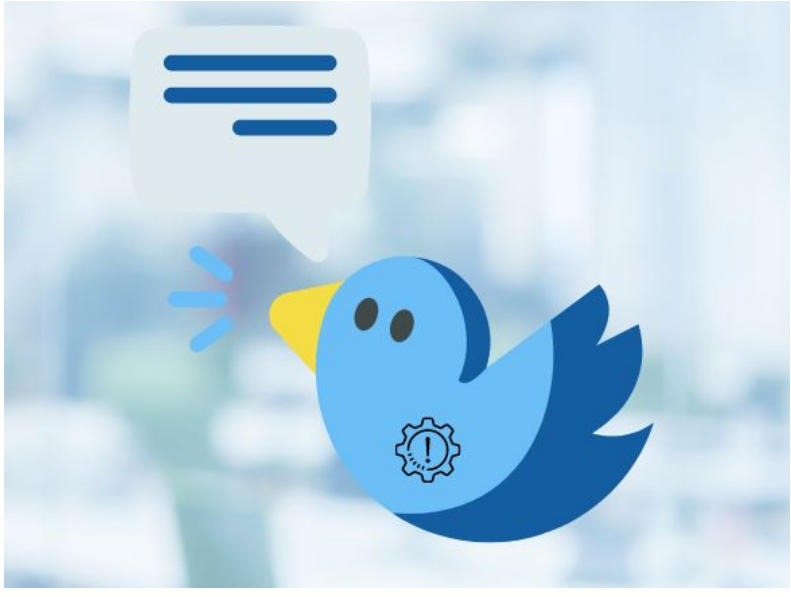
Problem Statement:

- In this Big Data project, you will fetch data from twitter and build an automatic system for replying to tweets for a business.

ProjectPro

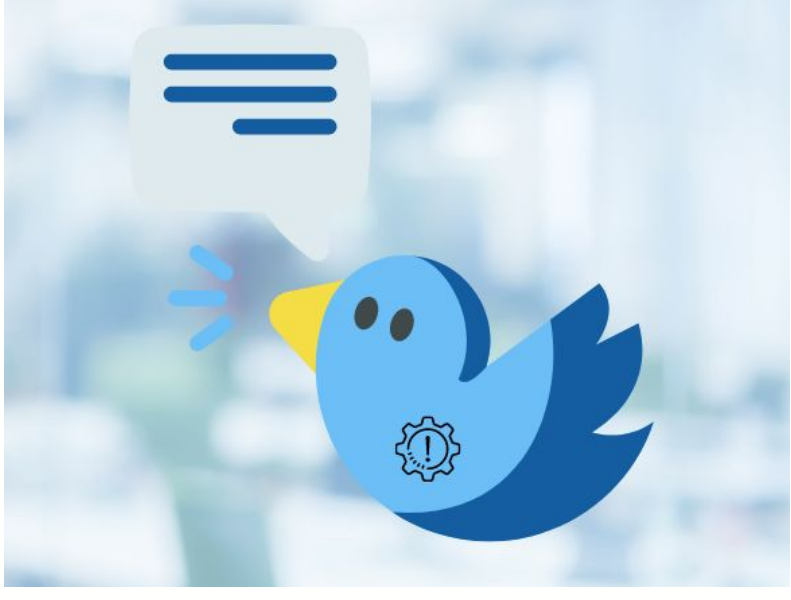
Data Description

- In this project, you will work with the dataset of tweets of an airline.
- The dataset has relevant details like airline names as tags, tweet content, sentiment of the tweet (positive or negative or neutral).
- It also has a topic for each tweet which can be either of the following: Baggage Issue, Customer Experience, Delay and Customer Service, Extra charges, Online Booking, Reschedule and Refund, Reservation Issue, Seating Preferences.



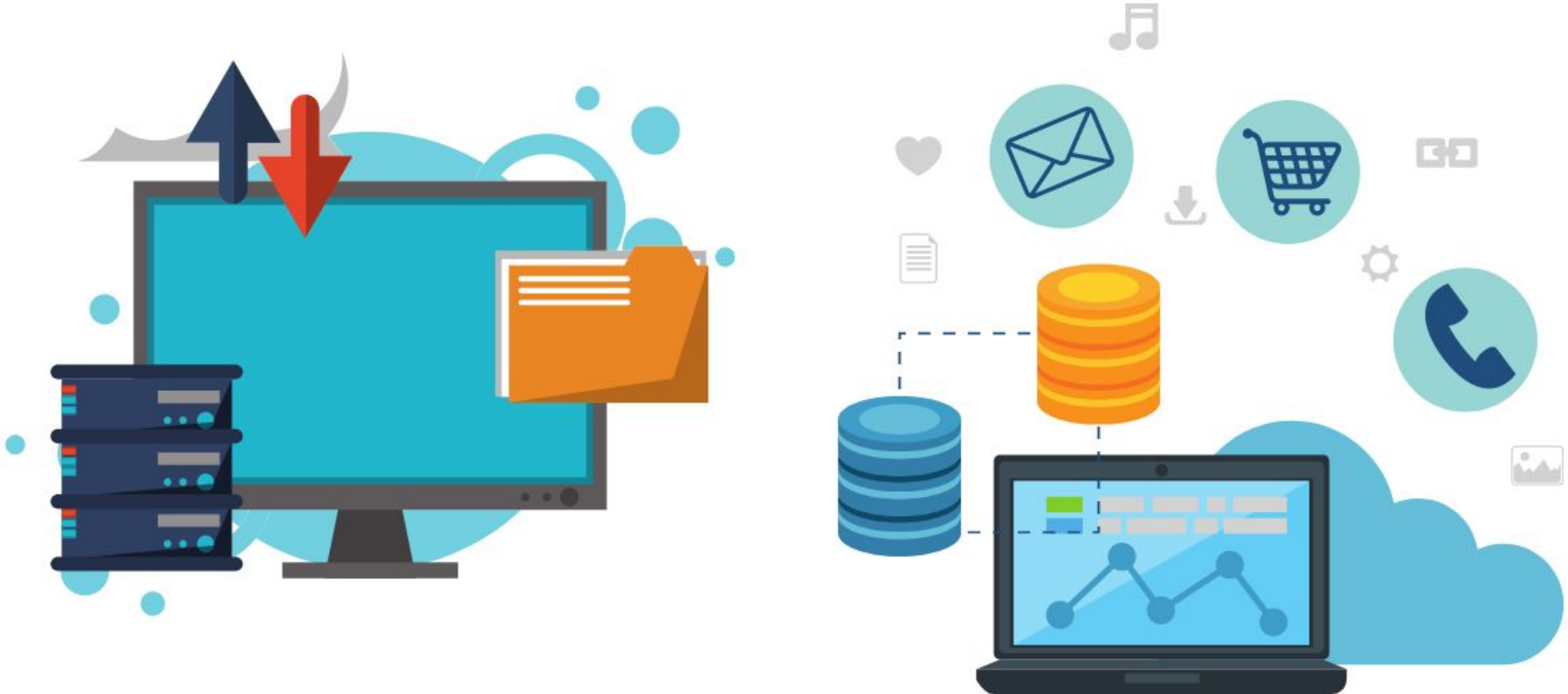
Learnings from the Project

- Using Tweepy for extracting data from tweets and replying to a tweet
- Using Flask API for building an API that will generate replies.
- Data Ingestion using Apache Kafka
- Utilizing Spacy for Named Entity Recognition
- Exploring Python data science libraries like Pandas, NumPy, and Matplotlib.
- Using Tensorflow and Keras frameworks in Python



ProjectPro

Setting up a Redshift ETL Pipeline



Overview of the Project

- ETL: Extract, Transform, and Load is a method of collecting data from various sources and transferring it to a data warehouse.
- ETL is one of the most widely used methods for data warehousing in various top companies.

Problem Statement:

- In this Big Data project, you will prepare a Redshift ETL Big data pipeline using various Big Data tools by AWS.

Data Description

- You will work with a dataset of customer reviews of certain products available on Amazon.
- If you want to target a specific category of products, then you can experiment with small subsets as well.
- Each small subset contains the following information:
  whether the review is helpful or not, overall rating out of 5, content of the review, date and time on which the review was submitted, etc.

Learnings from the Project

- Understanding the Architecture of the whole project
- Building a Virtual Private Cloud (VPC)
- Creating a Redshift cluster and exploring its usage.
- Utilising the AWS CLI tool for building S3 buckets
- Designing and running Glue jobs.
- Building an Amazon Simple Notification Service (SNS).

ProjectPro

The learning path helps you build a successful career in big data or data science every step of the way. ProjectPro many more fantastic and interesting projects added every month.

Explore solved end-to-end Big Data, Data Science, and Machine Learning Projects.

Don’t just stop here. Explore more!