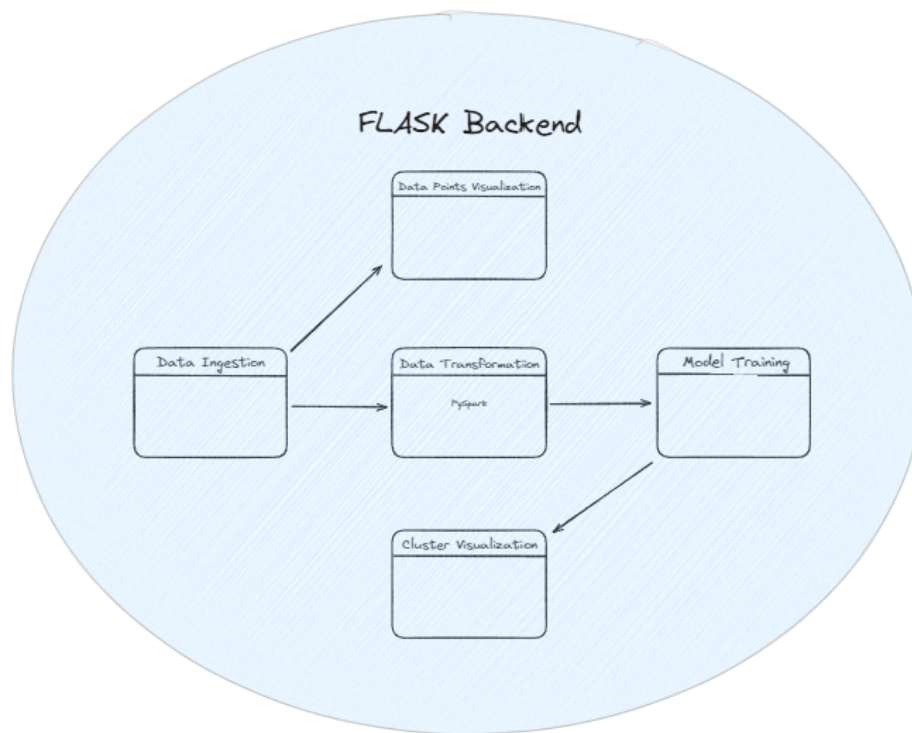# iNeuron.ai
Architecture
Big Data Internship Report

Social media community using optimized clustering algorithm

Submitted by
Arbash Hussain

Architecture

Architecture Details

The project aims to cluster users in a social network based on their activity data using a Big Data approach. This involves data ingestion, transformation, modeling, and visualization. Then for interaction with users, the project has an api developed in flask.

System Overview

- Data Source: Synthetic user and post data generated using Faker.
- Data Processing Framework: Apache Spark for handling large datasets.
- Machine Learning Algorithm: K-Means Clustering using Scikit-learn.
- Visualization: Plotly for visualizing clustered data.
- Web Interface: Flask for pipeline interaction.

Key Components

- Data Ingestion: Generates synthetic data using Faker and processes it with Apache Spark.
- Data Transformation: Aggregates user data and applies PCA for dimensionality reduction.
- Model Training: Utilizes K-Means clustering to group users based on activity.
- Visualization: Displays clusters using an interactive Plotly graph.
- Web Application: Provides a user interface to run the pipeline and view results.

Technologies Used

- Python
- Apache Spark
- Scikit-learn
- Plotly
- Flask