Python Spark Certification Training Using PySpark

Certification Project

edureka!

Certification Project

Bicycle Sharing Demand

Domain – Transportation Industry

Business challenge/requirement

With the spike in pollution levels and the fuel prices, many Bicycle Sharing Programs are running around the world. Bicycle sharing systems are a means of renting bicycles where the process of obtaining membership, rental and bike return is automated via a network of joint locations throughout the city. Using this system people can rent a bike from one location and return it to a different place as and when needed.

Data Set

Data contains hourly rental data spanning two years. Training set comprised of the first 19 days of each month while the test set is the 20th to the end of month.

Considerations

You are building a Bicycle Sharing demand forecasting service that combines historical usage patterns with weather data to forecast the Bicycle rental demand in real-time. To develop this system, you must first explore the dataset and build a model. Once it's done you must persist the model and then on each request run a Spark job to load the model and make predictions on each Spark Streaming request.

Data Exploration and Transformation

Explore the data and develop the model in Spark Shell

1. Read dataset in Spark
2. Get summary of data and variable types
3. Decide which columns should be categorical and then convert them accordingly
4. Check for any missing value in dataset and treat it
5. Explode season column into separate columns such as season_<val> and drop season
6. Execute the same for weather as weather_<val> and drop weather
7. Split datetime into meaning columns such as hour, day, month, year, etc.
8. Explore how count varies with different features such as hour, month, etc

Model Development

1. Split the dataset into train and train_test.
2. Try different regression algorithms such as linear regression, random forest, etc. and note accuracy.
3. Select the best model and persist it

Model Implementation and Prediction

Application Development for Model Generation

For the above steps wrote write an application to:

1. Clean and Transform the data
2. Develop the model and persist it.

Application Development for Demand Prediction

Model Prediction Application – Write an application to predict the bike demand based on the input dataset from HDFS:

1. Load the persisted model.
2. Predict bike demand
3. Persist the result to RDBMS

Application for Streaming Data

Write an application to predict demand on streaming data:

1. Setup flume to push data into spark flume sink.
2. Configure spark streaming to pull data from spark flume sink using receivers and predict the demand using model and persist the result to RDBMS.
3. Push messages from flume to test the application. Here application should process and persist the result to RDBMS

Dataset: You can download the dataset from here