BD_24_Summative-Spark

July 20, 2018

1 Spark

In [1]:
```
import findspark
findspark.init()

#from pyspark.sql.session import SparkSession
from pyspark.sql import SparkSession

spark = SparkSession.builder.master("local[*]").getOrCreate()
```

In [2]:
```
# read csv file
data = spark.read.csv('sensor_data.csv',header=True)
```

In [3]:
```
from pyspark.sql.types import DoubleType, IntegerType

#convert all columns
for col_name in data.columns:
    data = data.withColumn(col_name, data[col_name].cast(DoubleType()))
```

In [4]:
```
# inspect the first 10 rows
data.show(100)

# the printSchema() method tells you the data type of each column
data.printSchema()
```

```
+----+---------+-----------------+-----------------+-----------------+-----------------+----
|Date|Device_ID|            Temp1|         Temp1_ave|          Temp1_sd|            Temp2|
+----+---------+-----------------+-----------------+-----------------+-----------------+----
|null|      4.0|26.809322978877425| 6.702330744719356|              0.0| 39.49025527676517| 9.8
|null|      7.0| 47.11805302640779|11.754296029930412| 5.051965285211055| 61.82985832904858|16.3
|null|      6.0| 42.90664203147708|15.613316184695398| 6.840977654674061| 57.19730623462479| 21.
|null|      4.0| 41.08917584805568|18.652689971692283| 7.925440713617367|55.198093432861256|25.3
|null|      8.0| 58.21307761699691|21.998427558106112| 9.748124352541879| 74.03438537869661| 29.
|null|      4.0|29.955361583677057| 24.00283526904515| 9.963760586188132|42.950897742044766| 32.
|null|      7.0|61.154504228264315|26.403470434163175| 10.93949082352105| 77.26995465109076|35.3
|null|      9.0| 23.95729066993767|27.677384596910024|10.773736728799927|36.353019736931444| 37.
|null|      2.0| 38.48319240889367| 28.72066251995634|10.577513361379625| 52.33151164978304|38.5
```

```
|null|      4.0|28.578187897659937|29.343067408226716|10.206953542739077| 41.43600668742593|39.4
|null|      0.0| 61.14964542885435| 32.82242882467565| 7.455527630296799| 77.26460997173979| 43.
|null|      3.0| 24.94857562201685| 34.87771295190258| 5.269447610147681|37.443433184218534| 46.
|null|      1.0| 85.13894576791235|37.474983640733655| 5.338669030292077|103.65284034470359|50.0
|null|      2.0| 36.66560703813738|39.312347437109004| 4.819164735913903|50.332167741951125| 52.
|null|      3.0| 24.74365043243004|39.853634604776346|4.6481940797500565| 37.21801547567305| 53.
|null|      5.0| 41.00262030042762| 40.53578144844561| 4.223918265236809|55.102882330470386|54.0
|null|      7.0|45.660867553058225| 40.66005066431742|4.2482878000296695| 60.22695430836405| 54.
|null|      1.0|   93.188165665542|42.484024451111374| 5.758471323141652| 112.5069822320962| 56.
|null|      9.0|25.137207454510232| 43.51835516734725| 5.619815536856776| 37.65092819996126| 57.
|null|      1.0|101.06531426679787|46.106281363752615| 6.890767015079256|121.17184569347766| 60.
|null|      2.0| 34.05561333623302|  47.3698752087411| 7.085302334224268| 47.46117466985633| 61.
|null|      7.0|48.181561106896126|48.898395229604446| 6.373816557052487| 62.99971721758574|  63
|null|     10.0| 59.88844365535325| 49.41352269243798| 6.590272061333156| 75.87728802088859|  64
|null|      5.0| 36.63654167895332|49.799141655583526| 6.499905408404016| 50.30019584684866|  64
|null|      6.0| 35.70393936543255| 50.36236310126775| 5.912718429328097|  49.2743333019758| 65.
|null|      8.0| 56.49802960366512| 51.17216441811185| 5.044806276095282| 72.14783256403163| 66.
|null|      8.0| 51.87968353752222| 51.93498580535653| 4.095110993449273| 67.06765189127444| 67.
|null|      8.0| 58.71783859211923| 51.64534366895448| 3.980616995724839| 74.58962245133115| 66.
|null|      3.0| 29.64880566565567| 51.54090202193156| 4.102211159110559| 42.61368623222124| 66.
|null|     10.0|  59.6606929201749| 50.42745525299882|2.7043665006637756|  75.6267622121924|  65
|null|      5.0| 39.14293311477787|49.719553170762865| 2.564952563425427| 53.05722642625566| 64.
|null|      0.0|  74.3327995004876| 49.84240756892569|2.7264367205082918| 91.76607945053637| 64.
|null|      3.0|23.941554596282813| 49.03587614107106| 2.418046098840527| 36.33571005591109|63.9
|null|      5.0| 32.74871143582221| 48.33378181410179| 2.988087343566726| 46.02358257940443| 63.
|null|      7.0| 48.77922653757037|  48.1340932481783| 3.170060691257809| 63.65714919132741|62.9
|null|      0.0| 67.00524804730043| 48.24700728482655|  3.21667700436947| 83.70577285203048| 63.
|null|      7.0|49.328613179170816|48.267916053353964|3.2265761992695943|  64.2614744970879|  63
|null|      6.0| 46.12709275200149|47.968828983746576| 3.003408460384522| 60.73980202720164| 62.
|null|      8.0| 53.89706524372359| 48.35072017099274|3.0180535335172443| 69.28677176809595|63.1
|null|      1.0| 89.63568339030601| 49.38651332318063| 4.640890425100705|108.59925172933663| 64.
|null|      9.0| 23.34500579757141|49.768410004391384| 4.589993750864147| 35.67950637732855| 64.
|null|      3.0| 25.31695832329066| 48.82943648586954| 4.614668912524848| 37.84865415561973| 63.
|null|      5.0| 37.33061400865023|48.459932832287336|4.9322114281910165|51.063675409515255|63.3
|null|      2.0| 39.86145925583612| 48.36062378760103| 5.050143006530877| 53.84760518141973|63.1
|null|      1.0| 87.64926845250916| 49.25789305195976| 5.069059196309225|106.41419529776009| 64.
|null|      6.0| 40.98846788367812|49.280425496138264|5.0731337042924345|55.087314672045935| 64.
|null|      3.0| 27.32364087418675| 48.74720052164754| 5.258181308140732|40.056004961605424| 63.
|null|      5.0| 39.70371537957944| 48.18669735646894| 5.500155548033561| 53.67408691753739| 63.
|null|      8.0| 74.18864708748593| 48.27360952867905| 5.537778803721562| 91.60751179623453| 63.
|null|      5.0| 43.63201980658634| 47.18870206824364|3.9639859867214606| 57.99522178724498|61.9
|null|      7.0|  49.6091067472683|47.031623996659505| 3.844238484642558| 64.57001742199513| 61.
|null|     10.0| 64.18951380469294|47.885629330006466| 4.121796833989696| 80.60846518516223| 62.
|null|      1.0| 92.27170940626964|49.899660714957164| 5.708579384307329|111.49888034689661| 64.
|null|      4.0| 39.52103488558373| 51.40167364441388| 5.396409991220145| 53.47313837414211| 66.
|null|      3.0|31.598967396131634| 51.12692581509698|5.3573592886155765|  44.7588641357448| 66.
|null|      1.0| 88.95171063281808|52.119946011837804| 5.994284729610825| 107.8468816960999| 67.
|null|      5.0| 42.16802270091739| 53.23582070506169| 5.500457204307668| 56.38482497100913|  68
```

```
|null|      7.0|45.352970246734266| 54.21395809665847|  4.41714381073814|  59.8882672714077| 69.
|null|      3.0| 25.31531252820236| 53.72572777637397|  4.97415712495911|  37.8468437810226|   6
|null|      3.0| 30.44266302775148| 53.02982111668973| 5.949963107106021| 43.48692933052663| 68.
|null|      9.0| 19.58452720363539| 51.75727663333572|7.7114488808885655|31.542979923998928|  66
|null|      1.0| 89.46538189387785| 51.43476497304984| 7.719199460551777|108.41192008326564| 66.
|null|      5.0| 41.63532819575759| 49.92697169757262| 6.757788255832079| 55.79886101533335| 64.
|null|      4.0|  32.3333472877084|48.616434551017825|6.5244682790961965| 45.56668201647924| 63.
|null|      1.0| 89.43195302862108|49.079356331913964| 6.810620845209115| 108.3751483314832| 63.
|null|      5.0| 46.35799753081059|48.361704840035884|5.9174843114681615| 60.99379728389166| 63.
|null|      1.0| 92.63443410202945| 49.08512650615512|  7.09675704555434| 111.8978775122324| 63.
|null|      4.0|  37.9415845590416| 49.44240811355224| 7.375536342819978| 51.73574301494576| 64.
|null|      1.0| 83.48858200973001|51.164701056138256| 8.356331331195683|101.83744021070302| 66.
|null|      8.0| 61.05560029594513|53.221744194782616| 8.437489266142466| 77.16116032553965| 68.
|null|      7.0|43.082287582620715|55.351970558240524| 6.411856994080157| 57.39051634088279|  70
|null|      9.0|25.889084604345662|55.360232898595655|6.4047441570543695| 38.47799306478023| 70.
|null|     10.0|60.346532090682494| 55.83420975123511| 5.986273511424731| 76.38118529975075| 71.
|null|      9.0|20.230185181647883|55.887113338063195| 5.882013165144085| 32.25320369981267| 71.
|null|     10.0| 68.88371226368885| 55.41308500906096| 6.098539186625547| 85.77208349005774| 70.
|null|      9.0|15.100009202836269| 54.27611405410992| 7.361679387137321|26.610010123119896| 69.
|null|      7.0| 52.51638232900415|52.420434543571005| 7.280233522295084| 67.76802056190456| 67.
|null|      4.0|37.426595839100365|51.015800192668294| 7.664625052301054|51.169255423010405| 66.
|null|      8.0| 69.67400960631642| 49.61696011940592| 6.459550630645657| 86.64141056694807| 64.
|null|      1.0| 95.52900417489109| 49.42966516143279|   6.0909466735869| 115.0819045923802|  64
|null|      2.0|36.583365414413045| 49.12672088874774| 5.727734236651658| 50.24170195585435| 64.
|null|      6.0| 35.35833759882016| 49.13624400909582| 5.729058877619703|48.894171358702174| 64.
|null|      8.0| 64.42337748410608| 49.24530748419248| 5.801919552690064| 80.86571523251669| 64.
|null|      8.0| 62.12272779462968| 50.37441865583951| 5.853474584474721| 78.33500057409266| 65.
|null|      3.0|20.477134202338917| 50.01108758304103| 5.943754687302341| 32.52484762257281|  65
|null|      2.0|  43.2387899088701|50.442058796093036| 5.460891786632595| 57.56266889975712| 65.
|null|     10.0|  59.5857989596201| 50.94202262164742| 5.101037777628926| 75.54437885558211| 66.
|null|      7.0| 55.42566055454816|51.766972108699406| 4.297037922677349| 70.96822661000297| 66.
|null|      6.0|32.739595448549224| 51.46232387004422| 4.553936769551084| 46.01355499340415| 66.
|null|     10.0| 57.89260869109902| 50.29292780395803| 3.299537273054879| 73.68186956020892| 65.
|null|     10.0| 59.90249018914247| 49.99885887376162| 2.997661561275365| 75.89273920805672| 64.
|null|      9.0| 28.61467317249583|  49.6097155654562|3.2197412581236837| 41.47614048974542| 64.
|null|      9.0|28.237294471391884|48.413206008909285| 3.705655919126634| 41.06102391853108| 63.
|null|      0.0| 72.35572900268393|47.771648871700464| 2.838812402899347| 89.59130190295232| 62.
|null|      1.0| 85.41416668849033|48.913906830947624| 4.223048194101935|103.95558335733936| 63.
|null|      7.0| 47.00343819991435|  49.8647165076591|4.5183074724465575| 61.70378201990579| 64.
|null|      2.0|35.571165760447464|49.977457935213394| 4.519142095656611| 49.12828233649221| 64.
|null|      1.0| 87.24738962088948|50.857557232537644|  5.38673040557218|105.97212858297843| 65.
|null|      5.0| 39.77175008814805| 51.69343557152081| 5.271198129478133|53.748925096962864|  66
|null|      5.0|36.992251983376434| 51.79783540806511| 5.231819504388494| 50.69147718171408| 66.
+----+---------+------------------+------------------+------------------+------------------+----
only showing top 100 rows

root
 |-- Date: double (nullable = true)
```

```
 |-- Device_ID: double (nullable = true)
 |-- Temp1: double (nullable = true)
 |-- Temp1_ave: double (nullable = true)
 |-- Temp1_sd: double (nullable = true)
 |-- Temp2: double (nullable = true)
 |-- Temp2_ave: double (nullable = true)
 |-- Temp2_sd: double (nullable = true)
 |-- Temp3: double (nullable = true)
 |-- Temp3_ave: double (nullable = true)
 |-- Temp3_sd: double (nullable = true)
 |-- Temp_Ambient: double (nullable = true)
 |-- Temp_Am_ave: double (nullable = true)
 |-- Temp_Am_sd: double (nullable = true)
```

1.0.1 Prepare data for model

```
In [5]: # Split the data into training and test sets (30% held out for testing)
        (trainingData, testData) = data.randomSplit([0.7, 0.3])

In [6]: from pyspark.ml import Pipeline
        from pyspark.ml.feature import OneHotEncoder, VectorAssembler
        from pyspark.ml.regression import DecisionTreeRegressor
        from pyspark.ml.evaluation import RegressionEvaluator

In [7]: # assemble variables to one feature column
        assembler = VectorAssembler(
            inputCols = ['Device_ID', 'Temp1', 'Temp2','Temp3','Temp_Ambient'],
            outputCol = "features")

        #define the estimator - decision tree
        dt = DecisionTreeRegressor(labelCol="Device_ID", featuresCol="features")

        # Chain indexers and tree in a Pipeline
        pipeline = Pipeline(stages=[assembler, dt])
```

1.0.2 Fit pipeline and transform data

```
In [8]: #fit the pipeline
        PipelineModel = pipeline.fit(trainingData)

        # transform using the pipeline
        predictions = PipelineModel.transform(testData)

        # evaluate model fit
        predictions.select("prediction", "Device_ID")
        evaluator = RegressionEvaluator(
```

```
            labelCol="Device_ID", predictionCol="prediction", metricName="rmse")
        rmse = evaluator.evaluate(predictions)
```

In [9]:
```
predictions.show()
```

```
+----+---------+------------------+------------------+------------------+-----------------+-----
|Date|Device_ID|             Temp1|         Temp1_ave|          Temp1_sd|            Temp2|
+----+---------+------------------+------------------+------------------+-----------------+-----
|null|      0.0| 61.14964542885435| 32.82242882467565| 7.455527630296799|77.26460997173979| 43.9
|null|      0.0| 61.40414948957228| 36.01476215770229| 4.091746723223278|77.54456443852952|49.61
|null|      0.0| 62.71132194066853|  40.7814903289155| 3.855135636520485|78.98245413473539| 54.8
|null|      0.0| 62.75904439465696|52.956074380882754|2.6852696762147925|79.03494883412266| 68.2
|null|      0.0| 63.03687215832894| 41.54623193006443| 3.312626890133367|79.34055937416184| 55.7
|null|      0.0| 63.31502102559877|  48.1733807358464| 7.099997451785536|79.64652312815865|62.99
|null|      0.0|63.430181435504004| 53.37820310837903|2.2635303796313657| 79.7731995790544| 68.7
|null|      0.0| 63.44980651960239| 46.13079958431568| 6.712322655623347|79.79478717156263|60.74
|null|      0.0| 64.12564567369651| 44.40435205900481| 3.362210187323846|80.53821024106617| 58.8
|null|      0.0| 64.53018603256234|   47.190289494113| 5.035759648164062|80.98320463581858|61.90
|null|      0.0| 65.17404507459204|51.448235322453925| 5.160542925094705|81.69144958205125| 66.5
|null|      0.0| 65.72958680736592| 56.19231743756139|4.6612773067474516|82.30254548810252| 71.8
|null|      0.0| 67.26676882422221| 48.97793078394485| 3.319671384079394|83.99344570664444| 63.8
|null|      0.0|  67.7625291843144| 50.60791645934563|3.4195457417093844|84.53878210274586| 65.6
|null|      0.0| 67.91851161062704| 45.67684399567627| 4.583746841280199|84.71036277168974| 60.2
|null|      0.0| 68.18085502690612|46.761002823756826|4.4336419725855585|84.99894052959674| 61.4
|null|      0.0| 68.36905540607681|53.738525915599254| 8.206394614752899| 85.2059609466845| 69.1
|null|      0.0| 68.54833033126368|   57.139803753407| 4.634278884796937|85.40316336439005|  72.
|null|      0.0| 68.60788867420946|44.708384891329295| 4.912916224261732|85.46867754163041| 59.1
|null|      0.0| 68.77870074809725| 52.70727030377573| 5.094252003473131|85.65657082290699|  67.
+----+---------+------------------+------------------+------------------+-----------------+-----
only showing top 20 rows
```

In [10]:
```
##Root mean square error
print(rmse)
```

```
0.0
```

1.0.3 Kmeans clustering

In [11]:
```
from pyspark.ml.clustering import KMeans

# Trains a k-means model with 4 clusters.
kmeans = KMeans(featuresCol='features', predictionCol='prediction',k=4)

#transform data using pipeline
pipeline = Pipeline(stages=[assembler, kmeans])
```

```
        #fir pipeline
        PipelineModel = pipeline.fit(data)

        # transform using the pipeline
        predictions = PipelineModel.transform(data)
```

In [12]:
```
        #view result
        predictions.show()
```

```
+----+---------+-----------------+-----------------+-----------------+-----------------+----
|Date|Device_ID|            Temp1|         Temp1_ave|          Temp1_sd|            Temp2|
+----+---------+-----------------+-----------------+-----------------+-----------------+----
|null|      4.0|26.809322978877425| 6.702330744719356|              0.0| 39.49025527676517| 9.8
|null|      7.0| 47.11805302640779|11.754296029930412| 5.051965285211055| 61.82985832904858|16.3
|null|      6.0| 42.90664203147708|15.613316184695398| 6.840977654674061| 57.19730623462479| 21.
|null|      4.0| 41.08917584805568|18.652689971692283| 7.925440713617367|55.198093432861256|25.3
|null|      8.0| 58.21307761699691|21.998427558106112| 9.748124352541879| 74.03438537869661| 29.
|null|      4.0|29.955361583677057| 24.00283526904515| 9.963760586188132|42.950897742044766| 32.
|null|      7.0|61.154504228264315|26.403470434163175| 10.93949082352105| 77.26995465109076|35.3
|null|      9.0| 23.95729066993767|27.677384596910024|10.773736728799927|36.353019736931444| 37.
|null|      2.0| 38.48319240889367| 28.72066251995634|10.577513361379625| 52.33151164978304|38.5
|null|      4.0|28.578187897659937|29.343067408226716|10.206953542739077| 41.43600668742593|39.4
|null|      0.0| 61.14964542885435| 32.82242882467565| 7.455527630296799| 77.26460997173979| 43.
|null|      3.0| 24.94857562201685| 34.87771295190258| 5.269447610147681|37.443433184218534| 46.
|null|      1.0| 85.13894576791235|37.474983640733655| 5.338669030292077|103.65284034470359|50.0
|null|      2.0| 36.66560703813738|39.312347437109004| 4.819164735913903|50.332167741951125| 52.
|null|      3.0| 24.74365043243004|39.853634604776346|4.6481940797500565| 37.21801547567305| 53.
|null|      5.0| 41.00262030042762| 40.53578144844561| 4.223918265236809|55.102882330470386|54.0
|null|      7.0|45.660867553058225| 40.66005066431742|4.2482878000296695| 60.22695430836405| 54.
|null|      1.0|   93.188165665542|42.484024451111374| 5.758471323141652| 112.5069822320962| 56.
|null|      9.0|25.137207454510232| 43.51835516734725| 5.619815536856776| 37.65092819996126| 57.
|null|      1.0|101.06531426679787|46.106281363752615| 6.890767015079256|121.17184569347766| 60.
+----+---------+-----------------+-----------------+-----------------+-----------------+----
only showing top 20 rows
```

1.0.4 END

In [13]:
```
spark.stop()
```