Spark Pipeline Exercise

Steve Mortensen

DS 6003

Motivation

- Housing and pricing data from King's County, Washington (near Seattle)
- I will be moving to Texas in the summer
- Curious about patterns that affect housing prices

Code Snippet

In [15]:
```
from pyspark.ml.linalg import Vectors, VectorUDT # nb: bad form, done for pedagogy
from pyspark.ml.feature import VectorAssembler
```

In [16]:
```
assembler = VectorAssembler(
    inputCols=['sqft_living','sqft_lot','sqft_above','sqft_basement'],
    outputCol='features')
df = df.withColumnRenamed('price','label')
lr_df = assembler.transform(df)
lr_df = lr_df.select(['features','label'])
lr_df.take(5)
```

Out[16]:
```
[Row(features=DenseVector([1180.0, 5650.0, 1180.0, 0.0]), label=221900.0),
 Row(features=DenseVector([2570.0, 7242.0, 2170.0, 400.0]), label=538000.0),
 Row(features=DenseVector([770.0, 10000.0, 770.0, 0.0]), label=180000.0),
 Row(features=DenseVector([1960.0, 5000.0, 1050.0, 910.0]), label=604000.0),
 Row(features=DenseVector([1680.0, 8080.0, 1680.0, 0.0]), label=510000.0)]
```

- Here, I created a multi-element vector for input into my linear model

Visualization – Residual Plot

In [25]:
```
from matplotlib import pyplot
```

In [26]:
```
trainingSummary = lrModel.summary
pd_res = trainingSummary.residuals.toPandas()
pd_pred = trainingSummary.predictions.toPandas()
```

In [36]:
```
fig = pyplot.figure()
fig.suptitle('Training Model Residual Plot')
pyplot.rcParams["figure.figsize"] = [9.0,6.0]
pyplot.xlabel('Fitted Values')
pyplot.ylabel('Residuals')
pyplot.scatter(pd_pred.prediction,pd_res)
```

Out[36]: <matplotlib.collections.PathCollection at 0x7f14654b3c18>