In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv('results/CLOUD_COMPARE_32_CORES.csv')
```

In [3]:
```
df = df.drop(columns=['Unnamed: 0', 'dataset_size_num'])
```

In [4]:
```
df['dataset_size'] = df['dataset_size']/100000000
```

In [5]:
```
df = df.round(4)
```

In [ ]:
```

```

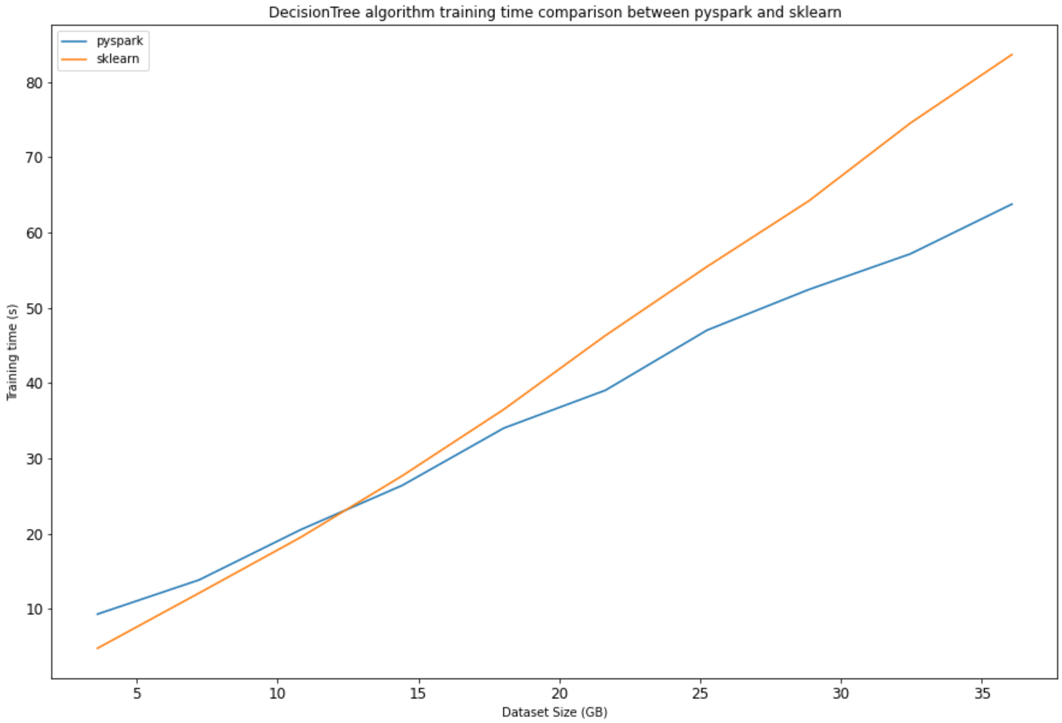
In [6]:
```
df
```

Out[6]:

	dataset_size	pyspark_time	pyspark_train_time	pyspark_predict_time	sklearn_time	sklearn_train_time	sklearn_predict_time
0	3.6064	9.3077	9.2882	0.0195	4.8605	4.7474	0.1131
1	7.2129	13.8471	13.8299	0.0172	12.3286	12.0859	0.2427
2	10.8193	20.5202	20.5041	0.0160	19.8822	19.4906	0.3915
3	14.4257	26.4062	26.3913	0.0149	28.1720	27.6649	0.5071
4	18.0322	34.0198	34.0047	0.0151	37.1242	36.4931	0.6311
5	21.6386	39.0455	39.0290	0.0165	47.0924	46.3083	0.7841
6	25.2450	47.0292	47.0107	0.0185	56.4024	55.4575	0.9449
7	28.8514	52.4215	52.4038	0.0177	65.1727	64.1710	1.0017
8	32.4579	57.1572	57.1434	0.0138	75.7063	74.5154	1.1908
9	36.0643	63.7693	63.7549	0.0145	84.9322	83.6269	1.3053

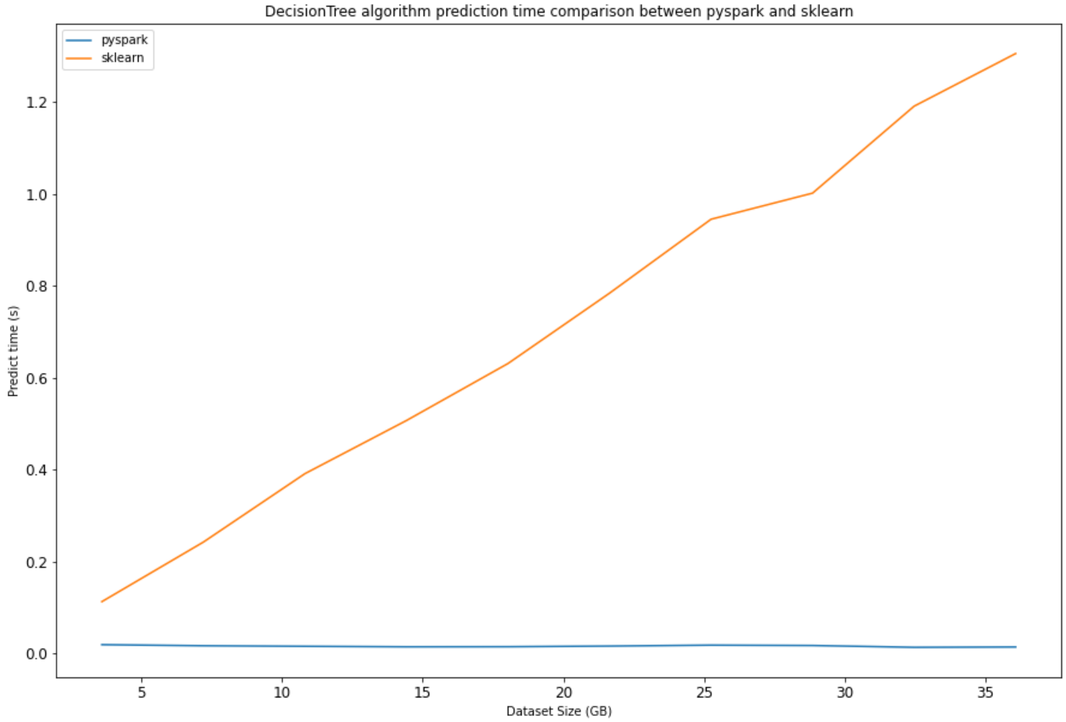
In [7]:
```
dftemp = df[['pyspark_train_time', 'sklearn_train_time', 'dataset_size']]
dftemp.columns=['pyspark', 'sklearn', "Dataset Size (GB)"]
dftemp.plot.line(
    x="Dataset Size (GB)",
    xlabel="Dataset Size (GB)",
    ylabel="Training time (s)",
    rot=0,
    title='DecisionTree algorithm training time comparison between pyspark and sklearn',
    figsize=(15,10),
    fontsize=12)
```

Out[7]:
```
<AxesSubplot:title={'center':'DecisionTree algorithm training time comparison between pyspark and sklearn'}, xlabel='Dataset Size (GB)', ylabe
l='Training time (s)'>
```



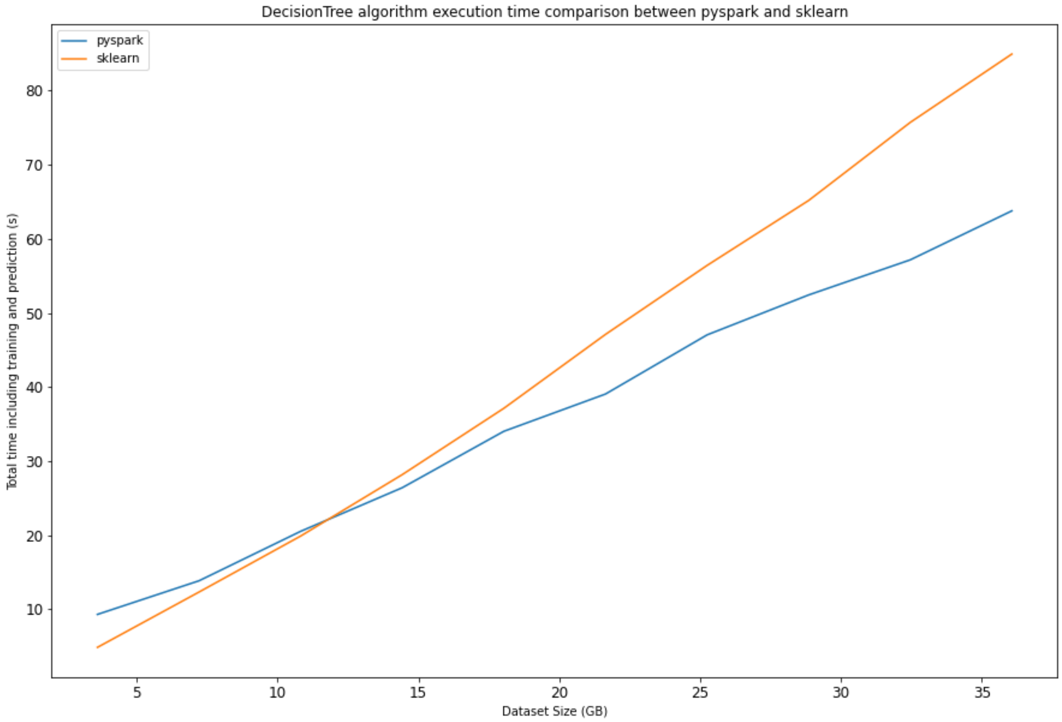
In [8]:
```
dftemp = df[['pyspark_predict_time', 'sklearn_predict_time', 'dataset_size']]
dftemp.columns=['pyspark', 'sklearn', 'dataset_size']
dftemp.plot.line(
    x='dataset_size',
    xlabel="Dataset Size (GB)",
    ylabel="Predict time (s)",
    rot=0,
    title='DecisionTree algorithm prediction time comparison between pyspark and sklearn',
    figsize=(15,10),
    fontsize=12)
```

```
Out[8]: <AxesSubplot:title={'center':'DecisionTree algorithm prediction time comparison between pyspark and sklearn'}, xlabel='Dataset Size (GB)', yla
bel='Predict time (s)'>
```



```
In [9]: dftemp = df[['pyspark_time', 'sklearn_time', 'dataset_size']]
        dftemp.columns=['pyspark', 'sklearn', 'dataset_size']
        dftemp.plot.line(
            x='dataset_size',
            xlabel="Dataset Size (GB)",
            ylabel="Total time including training and prediction (s)",
            rot=0,
            title='DecisionTree algorithm execution time comparison between pyspark and sklearn',
            figsize=(15,10),
            fontsize=12)
```

```
Out[9]: <AxesSubplot:title={'center':'DecisionTree algorithm execution time comparison between pyspark and sklearn'}, xlabel='Dataset Size (GB)', ylab
el='Total time including training and prediction (s)'>
```



```
In [10]: dftemp1 = df
         dftemp1.columns = ['Dataset Size (GB)', 'PySpark Total time (s)', 'PySpark Training time (s)',
                 'PySpark Predict time (s)', 'Sklearn Total time (s)', 'Sklearn Training time (s)',
                 'Sklearn Predict time (s)']
```

```
In [11]: dftemp1
```

Out[11]:

	Dataset Size (GB)	PySpark Total time (s)	PySpark Training time (s)	PySpark Predict time (s)	Sklearn Total time (s)	Sklearn Training time (s)	Sklearn Predict time (s)
0	3.6064	9.3077	9.2882	0.0195	4.8605	4.7474	0.1131
1	7.2129	13.8471	13.8299	0.0172	12.3286	12.0859	0.2427
2	10.8193	20.5202	20.5041	0.0160	19.8822	19.4906	0.3915
3	14.4257	26.4062	26.3913	0.0149	28.1720	27.6649	0.5071
4	18.0322	34.0198	34.0047	0.0151	37.1242	36.4931	0.6311
5	21.6386	39.0455	39.0290	0.0165	47.0924	46.3083	0.7841
6	25.2450	47.0292	47.0107	0.0185	56.4024	55.4575	0.9449
7	28.8514	52.4215	52.4038	0.0177	65.1727	64.1710	1.0017
8	32.4579	57.1572	57.1434	0.0138	75.7063	74.5154	1.1908
9	36.0643	63.7693	63.7549	0.0145	84.9322	83.6269	1.3053

In [ ]: