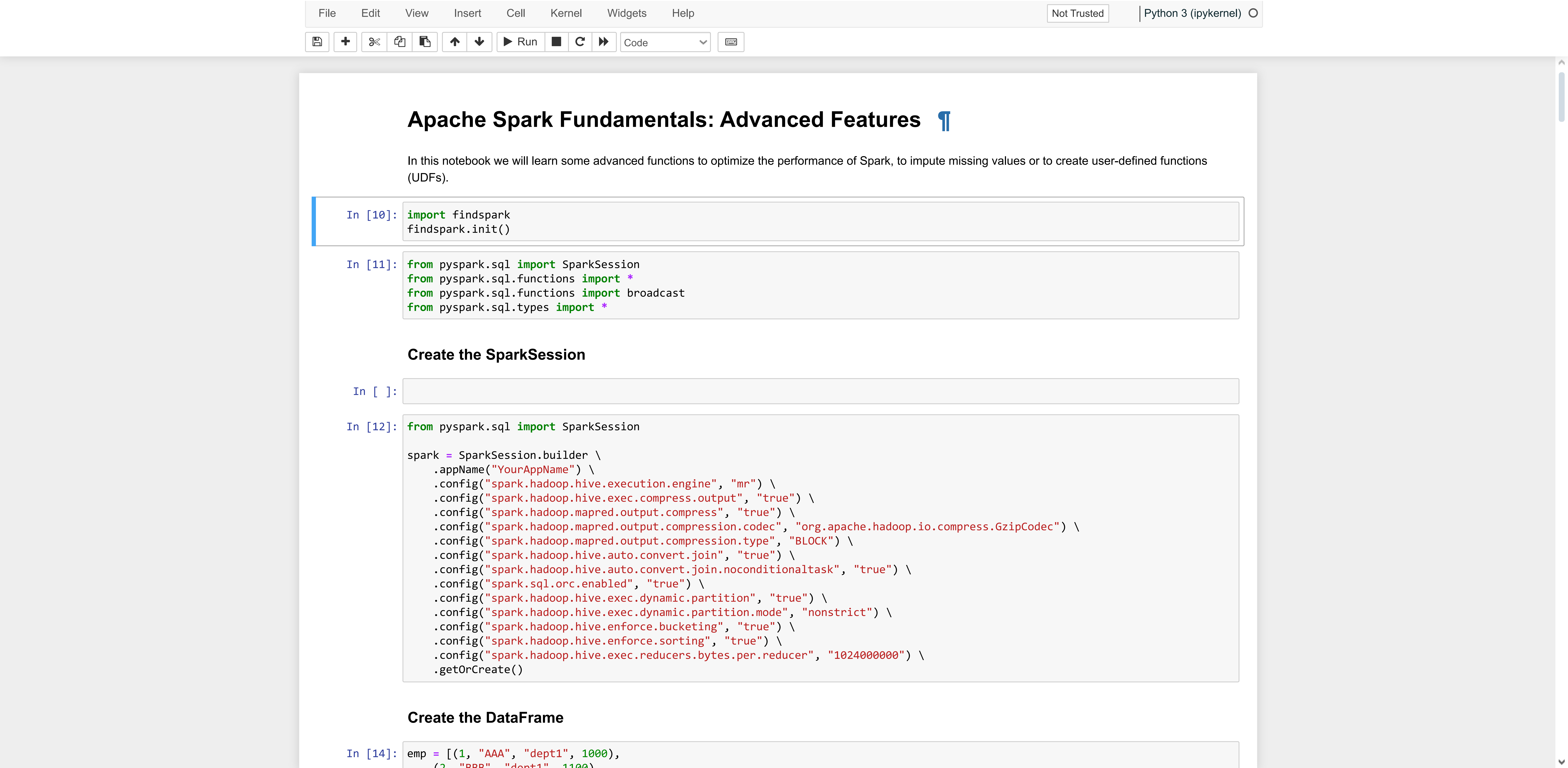

Apache Spark Fundamentals: Advanced Features ¶

In this notebook we will learn some advanced functions to optimize the performance of Spark, to impute missing values or to create user-defined functions (UDFs).

In [10]:

```
import findspark
findspark.init()
```

In [11]:

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.functions import broadcast
from pyspark.sql.types import *
```

Create the SparkSession

In [ ]:

In [12]:

```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("YourAppName") \
    .config("spark.hadoop.hive.execution.engine", "mr") \
    .config("spark.hadoop.hive.exec.compress.output", "true") \
    .config("spark.hadoop.mapred.output.compress", "true") \
    .config("spark.hadoop.mapred.output.compression.codec", "org.apache.hadoop.io.compress.GzipCodec") \
    .config("spark.hadoop.mapred.output.compression.type", "BLOCK") \
    .config("spark.hadoop.hive.auto.convert.join", "true") \
    .config("spark.hadoop.hive.auto.convert.join.noconditionaltask", "true") \
    .config("spark.sql.orc.enabled", "true") \
    .config("spark.hadoop.hive.exec.dynamic.partition", "true") \
    .config("spark.hadoop.hive.exec.dynamic.partition.mode", "nonstrict") \
    .config("spark.hadoop.hive.enforce.bucketing", "true") \
    .config("spark.hadoop.hive.enforce.sorting", "true") \
    .config("spark.hadoop.hive.exec.reducers.bytes.per.reducer", "1024000000") \
    .getOrCreate()
```

Create the DataFrame

In [14]:

```
emp = [(1, "AAA", "dept1", 1000),
    (2, "BBB", "dept1", 1100),
```