누구나 빅데이터 처리 할 수 있는

APACHE PySpark™

Part 2

groupBy

```
from pyspark.sql import SparkSession
# 데이터 그룹하기
grouped_df =df.groupBy("column1").count()
grouped_df.show()
```

***count()를 붙이면 Pandas의 value_counts()와 비슷합니다.**

Sort & orderBy

```
from pyspark.sql import SparkSession
# 열에 있는 데이터 정렬하기 (sort)
sort_df = df.sort("column1")
sorted_df.show()
# 열에 있는 데이터 정렬하기 (orderby)
ordered_df = df.orderBy("column1")
ordered_df.show()
```

df.na.drop

```
from pyspark.sql import SparkSession

# 결측치 없애기

df_na_drop = df.na.drop()

df_na_drop
```

df.na.fill

```
from pyspark.sql import SparkSession
# 데이터 채우기
df_na_filled = df.na.fill(0)
df_na_filed.show()
```

df.na.replace

```
from pyspark.sql import SparkSession

# 데이터 A 대신에 넣는 데이터B

df_na_replace = df.na.replace("old_value", "new_value")

df_na_replace.show()
```

describe & summary

```
from pyspark.sql import SparkSession
# 숫자 열에 대한 기본 통계를 계산
df.describe().show()


#데이터프레임 요약
df.summary().show()
```

withColumnRenamed

```
from pyspark.sql import SparkSession
# 데이터 열 이름 바꾸기
renamed_df = df.withColumnRenamed (“column1”, “new_column1”)
```

Head

```
from pyspark.sql import SparkSession
# 데이터 헤드
head_row = df.head()
print(head_row)
```

toPandas

```
from pyspark.sql import SparkSession
# 데이터프레임에서 판다스로 변환
pandas_df = df.toPandas()
print(pandas_df.head())
```

Dataframe 만들기 (1)

from pyspark.sql import Row

```
df = spark.createDataFrame([
    Row(a=1, b=2., c='string1', d=date(2000, 1, 1), e=datetime(2000, 1, 1, 12, 0)),
    Row(a=2, b=3., c='string2', d=date(2000, 2, 1), e=datetime(2000, 1, 2, 12, 0)),
    Row(a=4, b=5., c='string3', d=date(2000, 3, 1), e=datetime(2000, 1, 3, 12, 0))
])
df
```

Dataframe 만들기 (2)

```
pandas_df = pd.DataFrame({
    'a': [1, 2, 3],
    'b': [2., 3., 4.],
    'c': ['string1', 'string2', 'string3'],
    'd': [date(2000, 1, 1), date(2000, 2, 1), date(2000, 3, 1)],
    'e': [datetime(2000, 1, 1, 12, 0), datetime(2000, 1, 2, 12, 0), datetime(2000, 1, 3, 12, 0)]
})
df = spark.createDataFrame(pandas_df)
df
```