Spark

Apache Spark merupakan open-source cluster framework computing, yang dibangun untuk pemrosesan big data dengan cepat

Berikut adalah **command spark** yang sering di gunakan untuk bigdata :

NO	METHOD	DESCRIPTION
1	baseRelationToDataFrame(BaseRelation baseRelation)	Mengubah Base Relation dari sumber data external menjadi dataframe
2	builder()	Digunakan untuk membangun sparkSession
3	catalog()	Interface yang dimana user bisa membuat,menghapus atau mengubah database,table ataupun fungsi
4	clearActiveSession()	Menghapus semua session aktif di waktu itu
5	clearDefaultSession()	Menghapus default session yang dibuat user itu sendiri
6	close()	Persamaan stop()
7	conf()	Configurasi runtime di spark
8	createDataFrame(JavaRDD<?> rdd, Class<?> beanClass)	Menerapkan skema ke RDD Jaba Beans
9	createDataFrame(JavaRDD<Row> rowRDD, StructType schema)	Membuat DataFrame ddari JavaRDD yang berisi baris menggunakan skema yang di berikan.
10	createDataFrame(java.util.List<?> data, Class<?> beanClass)	Menerapkan skema ke list java beans
11	createDataFrame(java.util.List<Row> rows, StructType schema)	Membuat DataFrame dari java.util.List yang berisi baris menggunakan skema yang diberikan.
12	createDataFrame(RDD<?> rdd, Class<?> beanClass)	Menerapan skema ke rdd java beans
13	createDataFrame(RDD<A> rdd, scala.reflect.api.TypeTags.TypeTag<A> evidence$2)	Membuat dataframe dari Produk RDD
14	createDataFrame(RDD<Row> rowRDD,	Membuat data Frame RDD yang berisi baris

	StructType schema)	menggunakan skema yang di berikan
15	createDataFrame(scala.collection.Seq<A> data, scala.reflect.api.TypeTags.TypeTag<A> evidence$3)	Membuat Dataframe dari seq Produk Lokal
16	createDataset(java.util.List<T> data, Encoder<T> evidence$6)	Membuat data set dari type Java.util.List
17	createDataset(RDD<T> data, Encoder<T> evidence$5)	Membuat Dataset dari RDD
18	createDataset(scala.collection.Seq<T> data, Encoder<T> evidence$4)	Membuat dataset dari seq Produk Lokal
19	emptyDataFrame()	Membalikn Dataframe tanpa isi
20	emptyDataset(Encoder<T> evidence$1)	Membuat dataset tanpa isi
21	experimental()	Kumpulan metode yang dianggap eksperimetal, tetapi dapat di gunakan untuk menghubungkan perencana query untuk fungsi tingkat lanjut
22	getActiveSession()	Melihat semua session yang aktif
23	getDefaultSession()	Melihat Sessuin Yang aktif yang di buat user
24	implicits()	Accessor untuk objek Scala Bersarang
25	listenerManager()	Antarmuka untuk mendaftarkan QueryExecutionList eners kustom yang mendengarkan metrik eksekusi
26	newSession()	Membuat Session Baru
27	range()	Mengeluarkan Panjang dataset
28	read()	Membaca data dari sumber yang tersedia
29	readStream()	Membaca data dari sumber data Stream
30	sql()	Eksekusi Query Di Spark
31	stop()	Memberhentikan Spark konteks yang sedang berjalan

Menggunakan Spark DI CDSW

1. Import Yang Diperlukan Di code

```
1  from pyspark.sql import SparkSession
2  from pyspark.sql.functions import udf
3  from pyspark.sql.types import StringType,MapType
4  import datetime
5  from datetime import datetime, timedelta
6  import time
7  import math
8  import pandas as pd
9  pd.options.display.html.table_schema=True
10 pd.options.display.max_columns=999
11 pd.options.display.max_rows=999
```

2. Membuat Spark Session

```
spark = SparkSession \
  .builder \
  .appName("Iqbal - list pinjaman") \
  .config('spark.dynamicAllocation.enabled', 'false') \
  .config('spark.executor.instances', '2') \
  .config('spark.executor.cores', '4') \
  .config('spark.executor.memory', '16g') \
  .config('spark.yarn.executor.memoryOverhead', '8g') \
  .enableHiveSupport() \
  .getOrCreate()

dataFrame = spark.read.table('temp.bafa_transaksi_1_nasabah')
```

3. Membaca Tabe HDFS

```
dataFrame = spark.read.table('temp.bafa_transaksi_1_nasabah')
```

Untuk data dari luar hdfs bisa menggunakan driver seperti di bawah

```
ingest = spark.read.format("jdbc")\
    .option("driver","com.microsoft.sqlserver.jdbc.SQLServerDriver")
```

4. Membuat ETL data

```
acctno = dataFrame.select('acctno')\
         .limit(10)
```

5. Save hasil ETL ke HDFS

```
acctno.write.mode('overwrite')\
        .saveAsTable("temp.tabel_baru")
```