Spark DataFrame
Spark SQL

Week 2

WHY

Spark DF/SQL?

	RDD	Spark DF	Spark SQL
定義 schema	無 （index 值）	有 （欄位名稱）	有 （欄位名稱）
易使用性	低 （困難）	中	高 （簡單）
執行速度	較慢	較快	
功能性	最完整		

操作進入點

- **SQLContext**
- **HiveContext**

  繼承自 SQLContext

  較完善的 HiveSQL 解析器，建議使用

建立 Spark DF

1. Python list to Spark DF
2. Pandas DF to Spark DF
3. text, csv, parquet file...
4. Spark RDD to Spark DF

轉換 Spark DF

- Select
- Add Column
- Filter
- Order By
- Distinct
- Group By
- Join

執行 Spark DF

1. show
2. take
3. collect
4. count

create, transform 不會執行

只有 action 會實際執行

Persist Spark DF

- df.persist() or df.cache()
- df.unpersist()

Default storage level: MEMORY_AND_DISK

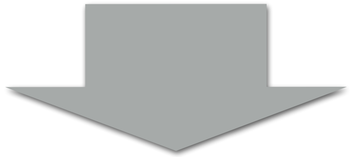
Save Spark DF

1. Save as parquet file
2. Save as csv file..., etc.
3. saveAsTable

Register Spark DF

- 「註冊」為一張暫時表，透過 SQL 分析其中的資料

df.registerTempTable(“table_name”)



Spark SQL

Python to Teradata

Spark to Teradata

Spark to Hive

程式碼放在公槽

\88.88.11.57\br000960_share\99.Team_DS\pyspark_training_week2\ pyspark_connectDB.ipynb