The SparkSQL things you maybe confuse

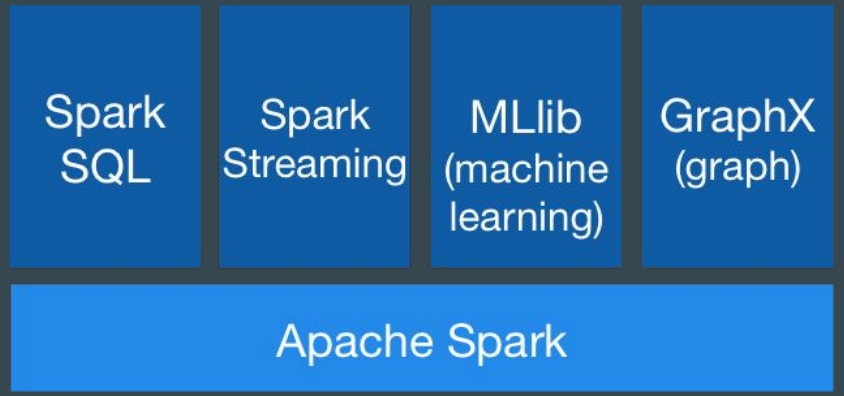
...

vito@is-land.com.tw
is-land System Inc.
2016/08/24

Do you ever confuse with : SparkSQL, SchemaRDD, DataFrame, and Dataset ?

About SparkSQL's history

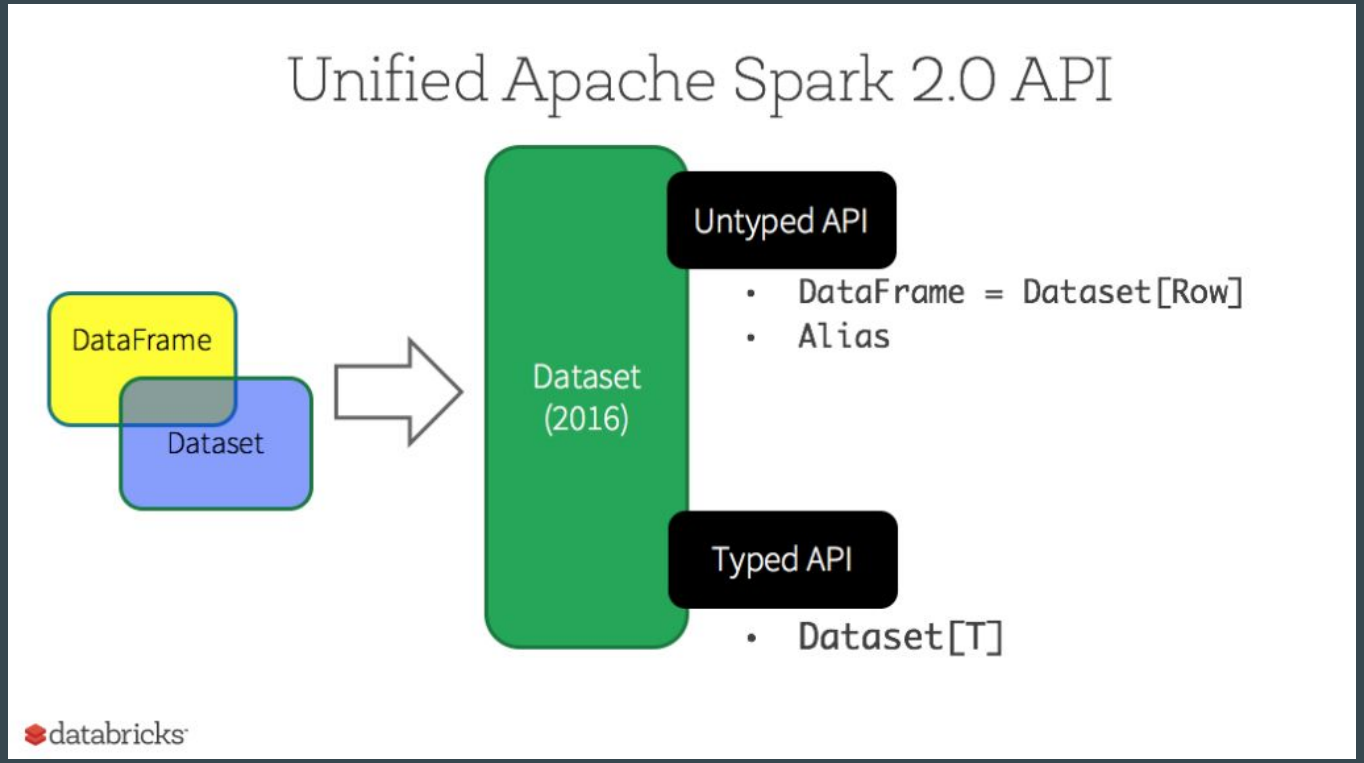
- SparkSQL - a high level module, new from Spark 1.0.0
- SchemaRDD - new from Spark 1.0.0
  - from Shark project
- DataFrame - new from Spark 1.3.0
- Dataset - new from Spark 1.6.0(experimental)



Breif explanation

- **SparkSQL** is a high level module name
- **SchemaRDD** - 讓 RDD 操作類似 RDB 中 table 概念的資料
  - RDD[Row]
  - 以 row / column 角度看資料
- **DataFrame** = SchemaRDD
- **Dataset** - 讓操作 RDD 的同時又能享有 SparkSQL 裡 Catalyst Optimzer 帶來的效能提昇

Unified DataFrame & DataSet



https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html

Apache Spark 1.0.x

- **In SQLContext**
  - **def** sql(sqlText: String): SchemaRDD = **new** SchemaRDD(**this**, parseSql(sqlText))
- **class** SchemaRDD(**val** sqlContext: SQLContext, **val** baseLogicalPlan: LogicalPlan) **extends** RDD[Row](sqlContext.sparkContext, *Nil*) **with** SchemaRDDLike

Example

- RDDRelation.scala
- SQLQuery.scala

Apache Spark 1.3.x

- In SQLContext
  - **def** sql(sqlText: String): DataFrame
- **type** SchemaRDD = DataFrame
- **class** DataFrame **private**[sql](**val** sqlContext: SQLContext, **val** queryExecution: SQLContext#QueryExecution) **extends** RDDApi[Row] **with** Serializable
-

Example

- RDDRelation.scala
- DataFrameSuite.scala

Apache Spark 1.6.x

- **class** DataFrame **private**[sql](**override val** sqlContext: SQLContext, **override val** queryExecution: QueryExecution) **extends** Queryable **with** Serializable
- **class** Dataset[T] **private**[sql](**override val** sqlContext: SQLContext, **override val** queryExecution: QueryExecution, tEncoder: Encoder[T]) **extends** Queryable **with** Serializable **with** Logging
-
- DataFrame 與 Dataset 兩者在繼承關係: 不相關
- 使用 DataFrame.as 轉換成 dataset

Example

- RDDRelation.scala
- DataFrameSuite.scala
- DatasetSuite.scala

Apache Spark 2.0

- ```
  class Dataset[T] private[sql](val sparkSession:
  SparkSession, val queryExecution: QueryExecution,encoder:
  Encoder[T]) extends Serializable
  ```
- **type** DataFrame = Dataset[Row]
-
  ```

Example

- DataFrameSuite.scala
- DatasetSuite.scala

Reference

- Spark 1.0.2 - Spark SQL Programming Guide
- Spark 1.3.1 - Spark SQL Programming Guide
- Spark 1.6.2 - Spark SQL Programming Guide
- Spark 2.0.0 - Spark SQL Programming Guide
- A Tale of Three Apache Spark APIs: RDDs, DataFrames, and Datasets When to use them and why
- Apache Spark 2.0: Faster, Easier, and Smarter