Lesson 7

Extract, Transform and Load Process

ETL process three functions

- **Extract** which does the acquisition of data from Data Store querying or from another program,
- **Transform** which does the change of data into a desired file, columnar, tabular or other.
- **Load** which does the process of placing transformed data into another Data Store or data warehouse

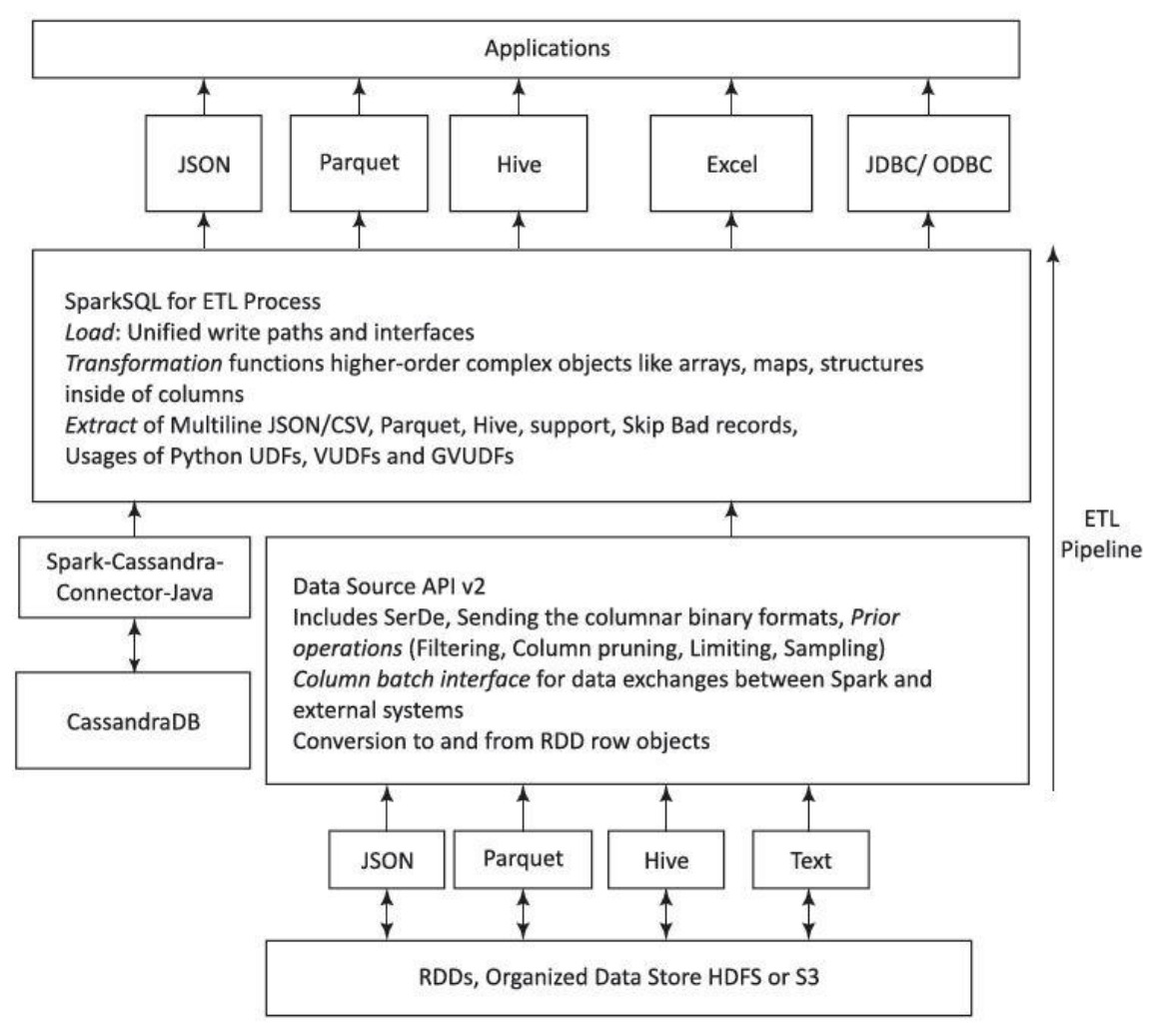
Transform Functions

- join(), groupBy(), cogroup(), filter(), map(), mapValues(), flatMap(), sort(), pratitionBy(), groupByKey(), reduceByKey(), aggregateByKey(), pipe(), coalesce(), sample(), union(), crossProduct()

Spark 2.3 with Pandas

- Includes transformation functions on complex objects like arrays, maps and set of columns
- Pandas provide powerful transformation UDFs, VUDFs and GVUDFs

Figure 5.9: An ETL pipeline using Spark SQL for ETL Process and Data Source API v2 in Spark 2.3.



Extract

- **Skipping Corrupt or Bad Records or Files**

Extract and Load

- Multi-line JSON/CSV Support
- Load and Save files: SerDe uses codes for obtaining records from unstructured data
- Save process uses serializer codes
- Loading (extracting) process uses deserializer.

Example for Load and Save

- Example 5.13 explains the codes for sequence File, JSON and CSV file load and save functions for obtaining records/rows/files

Example

- Example 5.14 explains Spark SQL transformations in Spark 2.3
- Complex objects, nested tables (one column rows) and array transformations
- Using the DataframeWriter API.

Summary

- Extract, Transform and Load
- Transform functions
- Load and Save
- Spark 2.3 includes transformation functions on complex objects like arrays, maps and set of columns
- Pandas provide powerful transformation UDFs, VUDFs and GVUDFs

End of Lesson 7 on

Applications and Big Data analytics using Spark