Spark SQL

Igor Yakushin
ivy2@uchicago.edu

Spark SQL

- Spark SQL is a component on top of Spark that introduced a data abstraction called DataFrames, which provides support for structured and semi-structured data.
- DataFrame is like a table distributed over the cluster similar to RDD
- One can query DataFrame using
  - Spark language
  - SQL - language for queries on relational databases
- In both cases the query optimization is used that typically results in a better performance over RDDs
- One can create DataFrame by applying transformations to other DataFrames, from the same sources as RDD: RDD, text files, json files, arrays, files in various Hadoop formats like parquet.
- Spark SQL can be interfaced with relational databases via ODBC/JDBC
- Spark SQL can use distributed Thrift store via ODBC/JDBC
- Spark SQL can use Hive store

Spark SQL

- Spark frames are similar to Pandas frames but contrary to Pandas frames they are divided and distributed accross multiple nodes, do not have to fit into a memory of a single machine, can be operated in parallel
- If you have sufficiently small Spark frame, you can convert it to Pandas with toPandas() method. But be careful: it works like collect() for RDD, bringing the whole data frame to a login node. If the data frame is too large, you can crash the login node.