CLOUD COMPUTING APPLICATIONS

Roy Campbell & Reza Farivar

Spark SQL

Spark SQL

- Structured Data Processing in Apache Spark
- Built on top of RDD data abstraction
- Need more information about the columns (Schema)
  - Can use this for optimizations
- Spark can read data from HDFS, Hive tables, JSON, etc.
- Can use SQL to query the data
- When needed, switch between SQL and python/java/scala
- Strong query engine

DataSet

- A Dataset is a distributed collection of data
- Dataset provides the benefits of RDDs (strong typing, ability to use powerful lambda functions) with the benefits of Spark SQL's optimized execution engine
- A Dataset can be constructed from JVM objects and then manipulated using functional transformations (map, flatMap, filter, etc.)
- The Dataset API is available in Scala and Java
  - Python does not currently (as of 2.0.0) have the support for the Dataset API.
  - But due to Python's dynamic nature, many of the benefits of the Dataset API are already available (row.columnName).

DataFrame

- A DataFrame is a *Dataset* organized into named columns
- Equivalent to a table in a relational database or a data frame in R/Python
- DataFrames can be constructed from:
  - structured data files
  - tables in Hive
  - external databases
  - existing RDDs
- The DataFrame API is available in Scala, Java, Python, and R