A Gentle Introduction to Spark 2.0.

Based on Madhukara Phatak posts at
`http://blog.madhukaraphatak.com/categories/spark-two/`.

Andres Calderon

May 8, 2017

Datasets

- `Dataset` - the new abstraction of Spark.
  - Replace `RDD` as standard abstraction layer.
  - `Dataframe API` becomes its subset.
  - [*LowLevel*] `RDD API` $\longrightarrow$ `Dataframe API` $\longrightarrow$ `Dataset` [*HighLevel*]

SparkSession

- `SparkSession` - New entry point of Spark
  - Replace `SparkContext` as standard entry point.
  - Combine `SQLContext`, `HiveContext` and future `StreamingContext`.

Creating SparkSession

Demo at `https://tinyurl.com/demospark`

Introduction to Dataset

- A `Dataset` is a **strongly typed collection of domain-specific objects** that can be transformed in parallel using functional or relational operations.
- Each `Dataset` also has an untyped view called a `DataFrame`, which is a `Dataset` of `Row`.

Introduction to Dataset

- RDD represents an immutable,partitioned collection of elements that can be operated on in parallel
- The major difference is, Dataset is collection of domain specific objects where as RDD is collection of any object.