A Gentle Introduction to Spark 2.0.

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

Andres Calderon

May 9, 2017

Outline

1. Spark Overview
2. Spark Session API
3. Wordcount in Dataset API

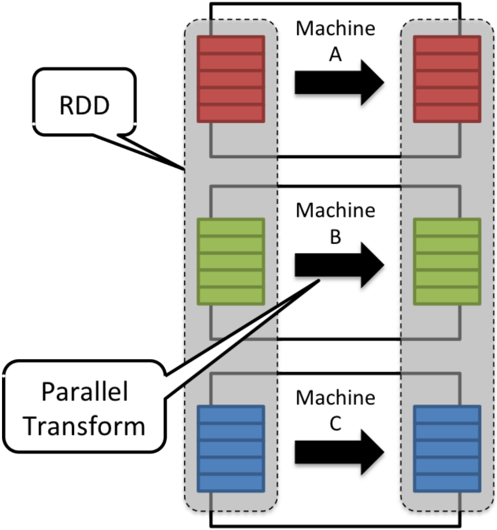
Overview

- Apache Spark provides an API centered on a data structure called the resilient distributed dataset (RDD).
- RDD: a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way.

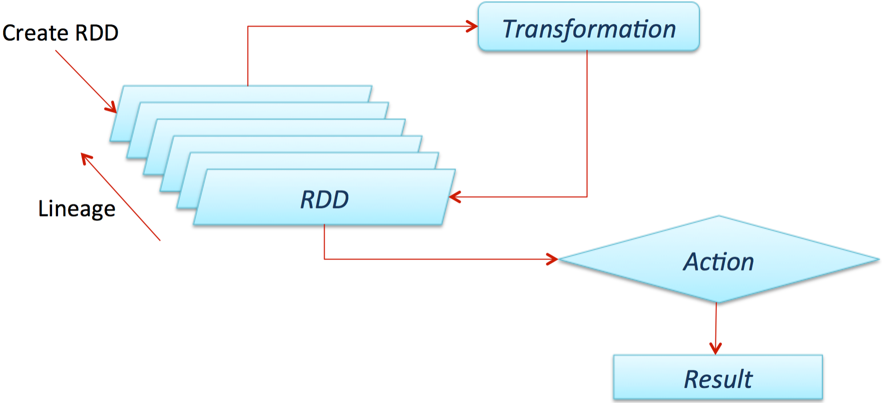
Overview

- Response to limitations in the MapReduce cluster computing paradigm, which forces a linear dataflow structure ...
- Read from disk $\longrightarrow$ Map across the data $\longrightarrow$ Reduce results $\longrightarrow$ Store to disk..
- Spark's RDDs function as a working set for distributed programs that offers a form of distributed shared memory.

Overview



Overview



Spark APIs

- APIs in different languages:
    - Scala
    - Python
    - R
    - Java

Spark APIs

```
// create a spark config object
val conf = new SparkConf().setAppName("wiki_test")
// Create a spark context
val sc = new SparkContext(conf)
// Read files from "somedir" into an RDD
// of (filename, content) pairs.
val data = sc.textFile("/path/to/somedir")
// Split each file into a list of tokens (words).
val tokens = data.flatMap(_.split(" "))
// Add a count of one to each token,
// then sum the counts per word type.
val wordFreq = tokens.map((_, 1)).reduceByKey(_ + _)
// Get the top 10 words. Swap word and count to sort by count.
wordFreq.sortBy(s => -s._2).map(x => (x._2, x._1)).top(10)
```

Spark APIs

```scala
import org.apache.spark.sql.SQLContext

// URL for your database server.
val url = "jdbc:mysql://IP:Port/db?user=username;password=passwd"
// Create a sql context object
val sqlContext = new org.apache.spark.sql.SQLContext(sc)

val df = sqlContext
  .read
  .format("jdbc")
  .option("url", url)
  .option("dbtable", "people")
  .load()

// Looks the schema of this DataFrame.
df.printSchema()
// Counts people by age
val countsByAge = df.groupBy("age").count()
```

Spark APIs

- Other Spark's Frameworks:
    - Spark Streaming
    - MLlib Machine Learning Library
    - GraphX

Outline

1. Spark Overview

2. Spark Session API

3. Wordcount in Dataset API

Datasets

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

SparkSession

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

Outline

1. Spark Overview

2. Spark Session API

3. Wordcount in Dataset API

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.

Creating SparkSession

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