A Gentle Introduction to Spark 2.0.

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

Andres Calderon

April 24, 2018

Outline

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

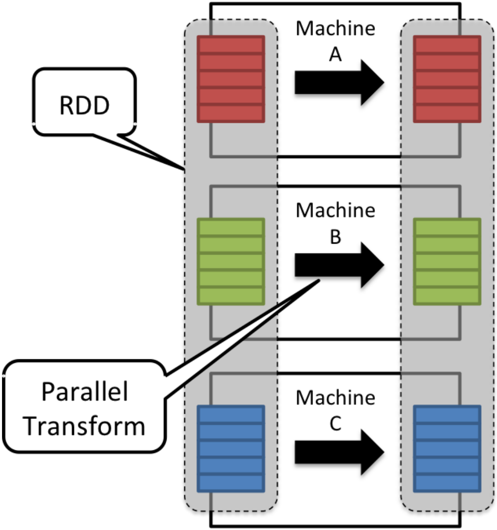
Overview

- 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.

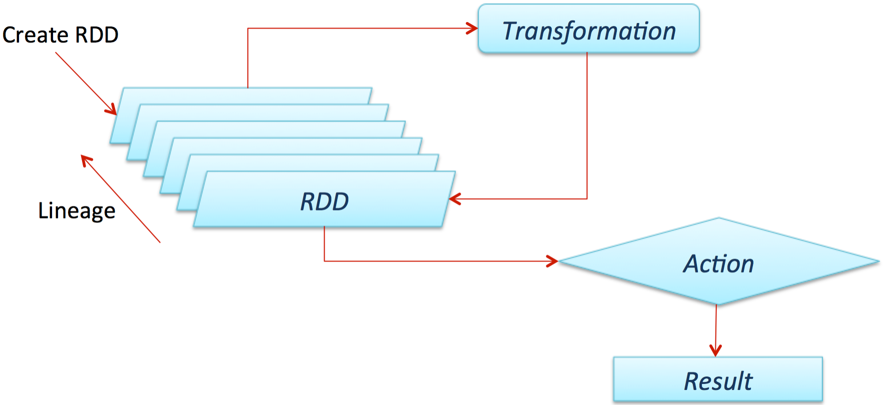
Overview

- 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

```
df = spark.read.json("logs.json")
df.where("age > 21")
        .select("name.first").show()
```

Plenty of examples at
http://spark.apache.org/docs/latest/quick-start.html

Spark APIs

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

Outline

1 Spark Overview

2 Spark Session API

3 Wordcount in the 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.

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` is also an immutable, partitioned collection of elements that can be operated on in parallel, but...
- `Dataset` is collection of **domain specific** objects where as `RDD` is collection of any object.
- A `Dataset` will require you a *schema* of your data...

Outline

1 Spark Overview

2 Spark Session API

3 Wordcount in the Dataset API

Directory layout

Your directory layout should look like this...

```
and@and-laptop:~/Dropbox/Academic/CS236_Spring_2018/WordCount$ find .
.
./build.sbt
./src
./src/main
./src/main/resources
./src/main/resources/data.txt
./src/main/scala
./src/main/scala/WordCount.scala
```

built.sbt

```
name := "WorkCount"
organization := "UCR-DBLab"
version := "1.0"

scalaVersion := "2.11.8"

libraryDependencies += "org.apache.spark" %% "spark-sql" % "2.1.0"

mainClass in (Compile, run) := Some("WordCount")
mainClass in (Compile, packageBin) := Some("WordCount")
```

WordCount.scala

```scala
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._

object WordCount {
  def main(args: Array[String]) {
    val spark = SparkSession.builder
      .master("local[*]")
      .appName("WordCount")
      .getOrCreate()
    import spark.implicits._

    val data = spark.read.text("src/main/resources/data.txt").as[String]
    val words = data.flatMap(value => value.split("\\s+"))
    val groupedWords = words.groupByKey(_.toLowerCase)
    val counts = groupedWords.count().toDF("word", "count")
    counts.orderBy(desc("count")).show()

    spark.close()
  }
}
```

Package

```
and@and-laptop:~/Dropbox/Academic/CS236_Spring_2018/WordCount$ sbt
[info] Loading project definition from /home/and/Dropbox/Academic/CS236_Spring_2018/WordCount/project
[info] Loading settings from build.sbt ...
[info] Set current project to PFlock (in build
↪  file:/home/and/Dropbox/Academic/CS236_Spring_2018/WordCount/)
[info] sbt server started at local:///home/and/.sbt/1.0/server/13280965ff4c9d1c2ea6/sock
sbt:PFlock> package
[info] Compiling 2 Scala sources to
↪  /home/and/Dropbox/Academic/CS236_Spring_2018/WordCount/target/scala-2.11/classes ...
[info] Done compiling.
[info] Packaging
↪  /home/and/Dropbox/Academic/CS236_Spring_2018/WordCount/target/scala-2.11/pflock_2.11-2.0.jar ...
[info] Done packaging.
[success] Total time: 31 s, completed Apr 24, 2018 2:29:35 PM
sbt:PFlock>
```

Package

```
and@and-laptop:~/Dropbox/Academic/CS236_Spring_2018/WordCount$ spark-submit
↪  /home/and/Dropbox/Academic/CS236_Spring_2018/WordCount/target/scala-2.11/pflock_2.11-2.0.jar
+-----------+-----+
|       word|count|
+-----------+-----+
|         of|   14|
|        the|   12|
|    dataset|   12|
|         in|   12|
|         is|   12|
|           |   10|
|       this|    7|
|         to|    6|
|        rdd|    5|
|      spark|    5|
|          a|    4|
| collection|    4|
|      where|    4|
|        all|    4|
|         be|    4|
|abstraction|    4|
|        new|    4|
|     spark.|    4|
|         so|    4|
|        for|    4|
+-----------+-----+
only showing top 20 rows
```