Bookmarks

- Week 1 - Apache Spark Programming Model
- **Week 2 - The Structured Query Language and Spark SQL**
  - **Lecture 2: The Structured Query Language and Spark SQL**
    Quizzes
  - **Lab 1A/1B - Learning Apache Spark (Due September 10, 2016 at 23:59 UTC)**
    Lab
- Week 3 - Analyzing Semi-Structured Data with Apache Spark

Bookmark

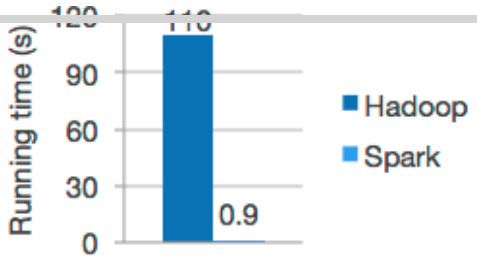
Apache Spark: Technology Trends, Opportunity, and Advantages

BERCS1052016-V001100

Using memory instead of disks offers two huge benefits. The first benefit is that memory is much faster than disks. The time to read or write a value to memory is only a few hundred nanoseconds (100/1,000,000,000th of a second!), while the time to read or write is tens of milliseconds (10/1,000th of a second) -- that means memory is a 100,000 times faster than disks. The second benefit is that keeping intermediate results in memory means that they do not have to be converted into a format that can be stored on disks. The process of converting a memory object to a disk object is called serialization and the process of converting a disk object to a memory object is called deserialization. Serializing and deserializing objects is a very expensive and time consuming process. Keeping intermediate results in memory avoids this significant overhead.

Taken together, the faster access times and avoidance of serialization/deserialization overhead make Apache Spark much faster than Apache Hadoop/Map Reduce - up to 100 times faster!

Here is an example of the performance difference between Apache Hadoop and Apache Spark when performing logistic regression, a common prediction technique:



Logistic regression in Hadoop and Spark

Hadoop/Map Reduce and Apache Spark Differences

(1 point possible)

Apache Spark is often faster than a traditional Hadoop/MapReduce implementation because:

- [x] It sends less data over the network
- [x] Results do not need to be written to disk ✔
- [x] It detects machine failures more quickly
- [x] It replicates the output of each task to recover from failures quickly
- [x] Results do not need to be serialized ✔

✖

Note: Make sure you select all of the correct options—there may be more than one!

EXPLANATION

Apache Spark keeps results in memory so they do not need to be serialized (converted into a format that can be stored on disk) and they do not need to be written to disk.

You have used 4 of 4 submissions