Apache Hadoop vs Spark

248201V
A.M.K.H.K Abeykoon
Big Data Analytics Technologies

About me

Hey I’m a Hasintha!

BSc (Hons) in IT (UoM)

Software Engineer @ Eight25Media

Apache Hadoop

- A framework (2006) for distributed processing
- Large Datasets
- High availability & fault tolerant
- HDFS
- MapReduce
- Hadoop Common
- Works with HDFS

Apache Spark

- Newer OS framework (2012)
- In memory computations
- Perfect fit for real time data processing - streaming data
- Spark Core
- Spark SQL
- Spark Streaming
- Complements Hadoop & other tools in ecosystem
- Relatively faster

Comparison

Hadoop

- Works on top of Disk - slower
- More suitable for batch processing
- High latency non-interactive
- Less memory usage - low cost
- Java or Python applications
- No native/library support for ML

Spark

- Works in memory - save Disk IO - faster
- Fit for real-time streams of data
- Low latency & interactive processing
- High memory usage - higher cost
- Java, Python, R, Scala, Spark SQL
- ML Library built in

Hadoop & Spark are not mutually exclusive. Optimal benefit can be obtained by using both in correct use-case.
Many organizations use both for better gains.