What are use cases for Spark vs Hadoop?

Re-Ask Follow 135

4 Answers

Radha Krishna Kanth Popuri, I work on Cloud services - end to end
Upvoted by Igor Markov, EECS Prof at Michigan - currently at Google • Sean Owen, Director, Data Science @ Cloudera

The main use cases for Spark are iterative Machine Learning algorithms and Interactive analytics.

From the ML side

Most ML algorithms run on the same data set iteratively and in MapReduce , there was no easy way to communicate a shared state from iteration to iteration.
Some attempts to run ML on MapReduce is documented here on the work done at Stanford.
www.cs.stanford.edu/~ang/papers/nips06-**mapreducemulticore**.pdf

Search for questions, people, and topics

e form of MapReduce Design Patterns for various use cases in this book - Data-Intensive Text Processing with MapReduce (Synthesis Lectures on Human Language Technologies): Jimmy Lin, Chris Dyer, Graeme Hirst: 9781608453429: Amazon.com: Books
Most of these techniques attempted to use Java specific features in Hadoop(ThreadLocal etc) even though MapReduce in theory did not offer any shared state communication model.

Spark is the next stage in the evolution of this. The fundamental thinking is that fine grained mutable state is a very low level abstraction and building block for ML algorithms ; Hence Spark was an attempt to raise this abstraction to coarse grained immutable data called RDD's (Resilient DIstributed DataSets) ;

Since HDFS never really supported multiple writer concurrent appends anyway , it follows that RDD's are not giving up much by being immutable - whereas you gain a lot by having both immutability and a higher level of abstraction to begin with for big data.

Interactive Analytics

If communicating shared state was one problem, the other problem was that MapReduce was initially created for batch analytics - with only two operators map/reduce. However it was becoming very clear that most interactive analytics queries required many more map/reduce jobs to achieve their purpose.

Cascading etc was one way to approach this. Another way to approach this was to create a high level SQL like language and compile the language to generate these MapReduce queries(Hive/Pig) . However since all these jobs did multiple passes on data (each time loading from HDFS) - they could not achieve the latencies expected of Interactive analytics.

Hadoop ecosystem quickly realized that the generation of mapReduce jobs and running them sequentially was not the right approach for Interactive analytics and there needed a way to directly operate on HDFS. Google Dremel/Cloudera Impala and others (I call it "Data Center SQL" - with SQL and their Multi-Level serving trees directly operating on HDFS) was one approach.

Another point to note is that main memory became even more cheaper during this time.

Upvote Downvote Comment 34,849

approaches. So to be precise Spark is a batch analytics system that can

SHARE QUESTION

Twitter • 6
Facebook • 11
Google+

RELATED TOPICS

Technology
Software and Applications
Computer Programming
Programming Languages
Python (programming language)

RELATED QUESTIONS

When is Hadoop MapReduce better than Spark?
Are there cases where traditional DBMS is replaced with Hadoop?
Which is easier to configure, Spark, Hadoop, or something else?
What are less known Hadoop use cases you have recently come across?
What are use some cases of Apache Hadoop in trading?
Do I need to learn Hadoop first to learn Apache Spark?
What are the implications for NoSQL databases given the introduction of Apache Spark? Will Spark replace NoSQL use cases b...
h Sign In
Hadoop/Hive stack? Anybody using it in production?
What are some good books on Apache Spark and real-time analysis with Hadoop?
What are the best practices for migrating a project from plain Hadoop to Spark on Hadoop?

masquerade as an interactive analytics system because of operating on in-memory RDD's and the caching hence possible.

81 upvotes • Updated 6 May, 2014

More Answers Below. **Related Questions**

When is Hadoop MapReduce better than Spark?

Are there cases where traditional DBMS is replaced with Hadoop?

Which is easier to configure, Spark, Hadoop, or something else?

What are less known Hadoop use cases you have recently come across?

What are use some cases of Apache Hadoop in trading?

Denny Lee, data dork

In the case of Spark in relationship to Hadoop, I would say it is that Spark and Hadoop complement each other instead of competing with each other. Hadoop is your uber-store of all of your semi-structured data within HDFS and it has the flexibility of Map Reduce so you can query the data. It is the iron horse that is implemented first so it is possible to go to the next level in Big Data Analytics.

Spark starts with the same concept of being able to run MR jobs except that it first places the data into RDDs (Resilient Distributed Datasets) so that this data is now stored in memory so its more quickly accessible. That is the same MR jobs can orders of magnitude faster because the data is accessed in memory. It adds to that flexibility and speed with ability to write queries in Scala (as William noted), Java (like Hadoop), and Python.

Spark is part of the Berkeley Data Analytics Stack (BDAS) which also include Tachyon (an in-memory file system), Spark Streaming, and to be released as part of Spark 0.8 the ability to run graph algorithms. As quoted by Reynold Xin in Spark: Open Source Superstar Rewrites Future of Big Data | Wired Enterprise | Wired.com , "Spark is the swiss army knife of Big Data Analytics"

49 upvotes • Written 14 Aug, 2013

Upvote Downvote Comments 1 25,590

Mayur Rustagi, Consultant, sigmoid.com

Spark is amazing for in-memory and more importantly iterative computing.
The key benefit it offers is caching intermediate data in-memory for better access times.
Some use cases where Shark outperforms Hadoop
1. Real Time querying of data: Querying in secs rather than minutes using Shark
2. Stream processing: Fraud detection and log processing in live streams for alerts, aggregates and analysis
3. Sensor data processing: Where data is fetched and joined from multiple sources, in-memory dataset really shine as they are easy and fast to process.
We have dealt with a lot more use cases with several companies using Shark & Spark. Contact us for more !!!!

13 upvotes • Updated 18 Dec, 2013

Upvote Downvote Comment 12,230

William Emmanuel Yu, you know it? Join us!
William has 1,030+ answers in Computer Science.

Spark together with **Scala** is a lot easier to for executing little programs than just plain **Hadoop**. This is especially convenient for people with development background who like to run "stuff" (ad-hoc queries) on data in hadoop/hdfs. This remove the need to know about the underlying hadoop layer and just think of it as data.

Update: For those with a database background, there is a tool called **Shark** that provides a Hive-like SQL-like interface but with access to Scala MLlib capabilities. with Spark performance.

14 upvotes • Updated 3 Aug, 2013

Upvote Downvote Comments 1 14,546

1 Answer Collapsed

Write an answer

Related Questions

Do I need to learn Hadoop first to learn Apache Spark?

What are the implications for NoSQL databases given the introduction of Apache Spark? Will Spark replace NoSQL use cases by supporting real-ti...

How do you compare Spark with Hadoop/Hive stack? Anybody using it in production?

What are some good books on Apache Spark and real-time analysis with Hadoop?

What are the best practices for migrating a project from plain Hadoop to Spark on Hadoop?

What are some unusual uses for Hadoop?

Is Hadoop dead and is it time to move to Spark?

What are the use cases for low-latency analytical queries on Hadoop?

Why do people use Hadoop or Spark when there is ElasticSearch?

Do I need Hadoop for Spark usage?

Will Spark overtake Hadoop? Will Hadoop be replaced by Spark?

For analysis use cases in "big data", what are the relative pros and cons of MPP DBMSs versus Hadoop & other solutions?

What is the difference between Apache Spark and Apache Hadoop (Map-Reduce) ?

What are some real life use cases of Apache Hadoop for oil and gas downstream and the automobile industry?

How does hardware configuration for a YARN/Spark cluster change vs. Hadoop/MapReduce?

Top Stories

Sitemap # A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
About - Careers - Privacy - Terms