Introduction to Spark

Joe

galvanize

Introduction

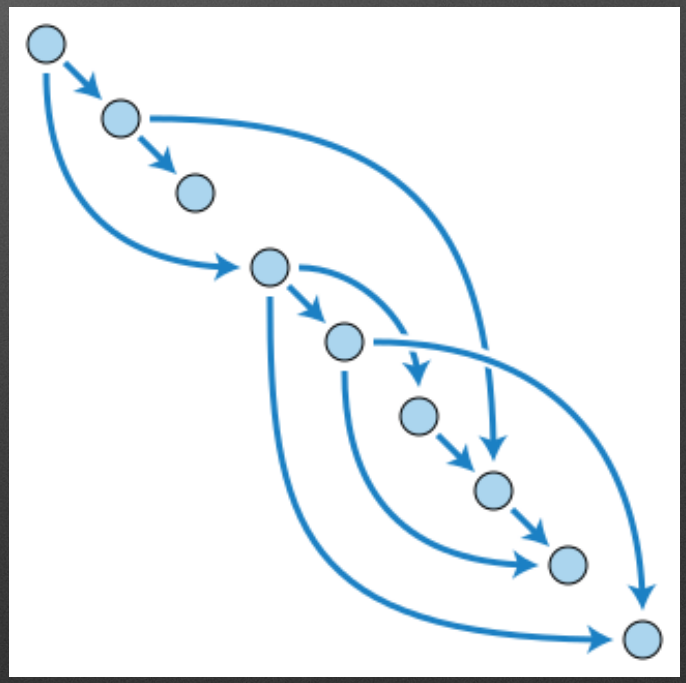
Morning Objectives

1. Understand why Spark *can* be much faster, and discuss important tips to optimize performance.

2. Do the ‘hello world’ of distributed computing - perform a word from a large corpus of documents.

Apache Spark

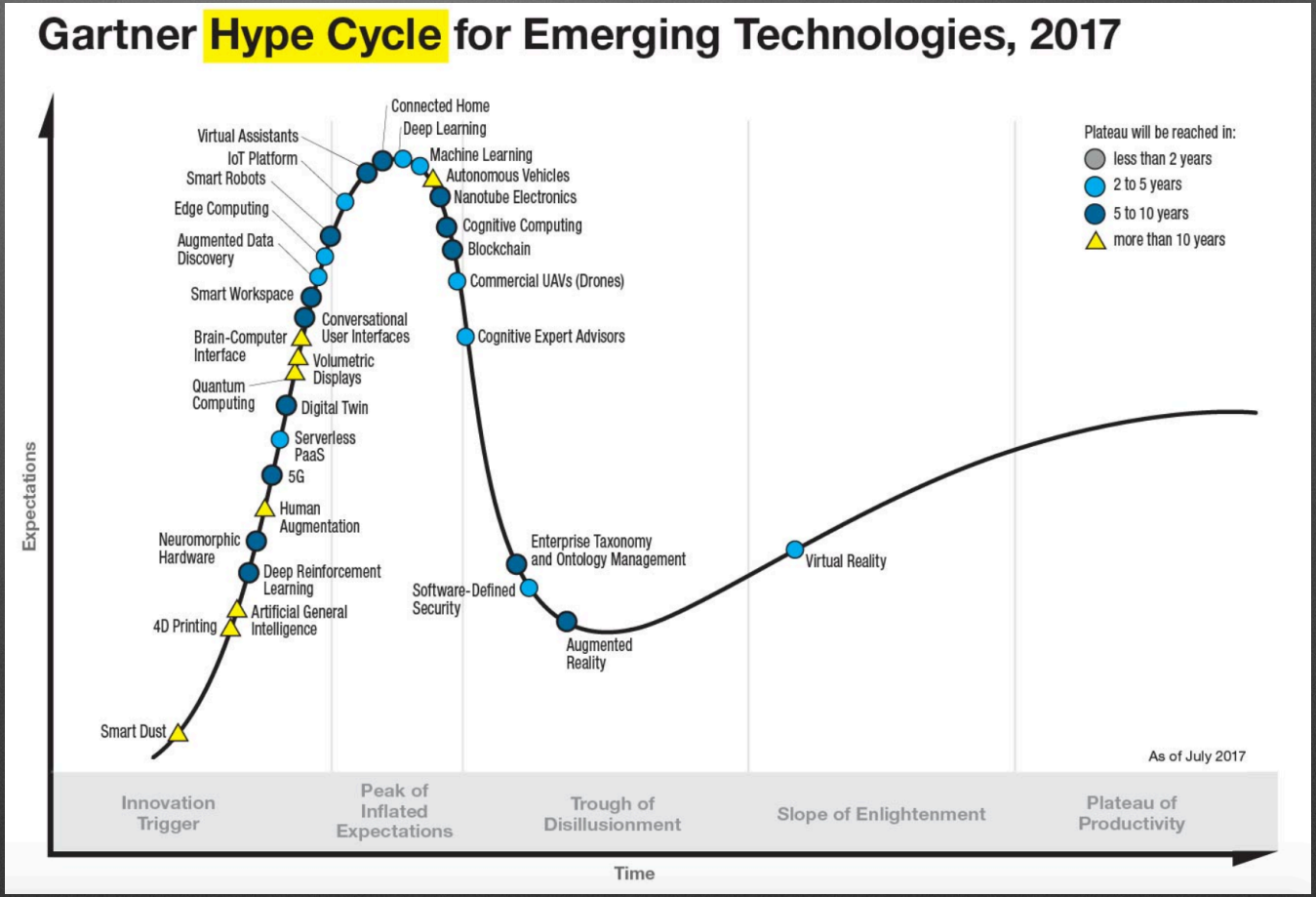
- Apache Spark is a fast and general engine for large-scale data processing
- Spark runs extremely fast because it performs in memory computations
- Spark has interfaces for Scala, Java, R, and Python
- Spark is an engine for building a directed analytic graph



A Note of Caution

I love spark.

Strong fluency with pandas/numpy/ect are useful so do not become overly reliant on using spark for data manipulation!



galvanize

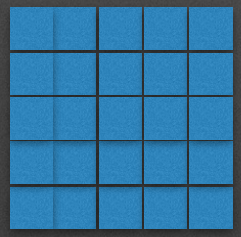
Mechanics

In Memory Computation

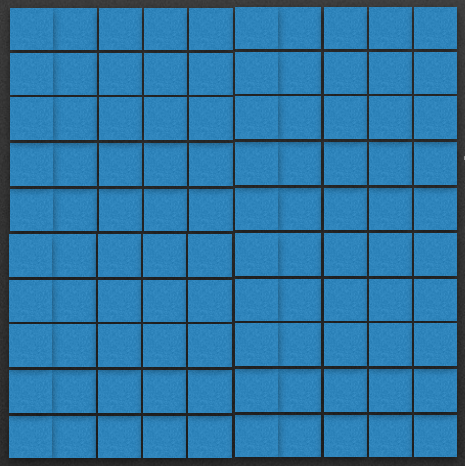
1 ns

1 floating point operation

mutex lock

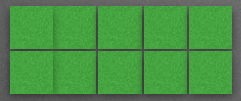


main memory ref

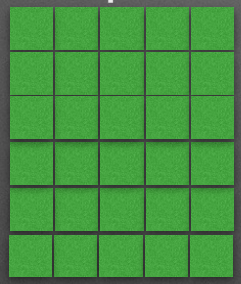


100 ns
=

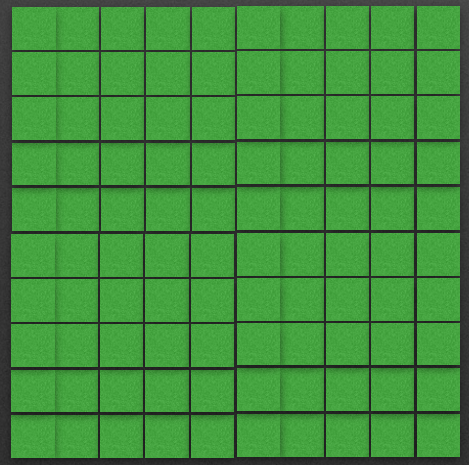
1µs



compress 1 KB

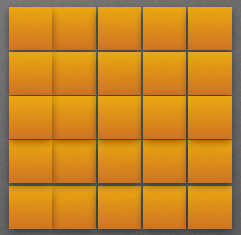


Send 1 KB over
1 Gbps network

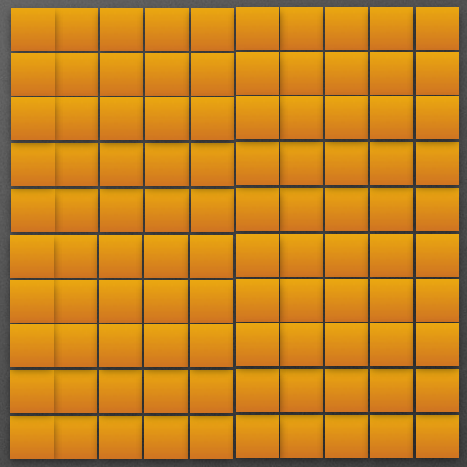


10 µs
=

Read 1 MB
sequentially from
memory

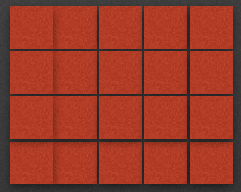


Read 1 MB
sequentially from SSD



1 ms
=

Read 1 MB
sequentially from disk



OK, time to get our hands
dirty

Morning Objectives

1. Understand why Spark *can* be much faster, and discuss important tips to optimize performance.

2. Do the ‘hello world’ of distributed computing - perform a word from a large corpus of documents.