The Unix operating system provides a simple model for accessing data. At the programmer level, have simple operations: read and write sequences of bytes. On top of this is built the reading of clusters of bytes (numerals, lines, words separated by whitespace, etc.). Programs typically have a standard input, standard output, and standard error: three streams. Hence, programs can be fitted together into series of programs—pipelines:

“The should have some words of caution to programmers”

A Resilient Distributed Dataset (RDD) is a collection of values or key-value pairs (as in a Python dictionary). Provides a variety of familiar operations on them:

- Unix-like: sort, count, distinct (uniq), pipe.
- SQL-like: union, intersection, join (multiple tables).
- General sequence operations: map, filter, reduce (accumulate).

Apache Spark

The Apache Spark framework (first developed here) attempts to provide a simple view of large datasets, permitting intuitive and flexible constructions analogous to Unix pipelines.

Map-Reduce

One type of such distributed computation is the Spark framework. It is still experimental and subject to implementation on the Spark framework.

Big Data

So what happens with huge amounts of data?

Examples (from Anthony Joseph, although a bit dated):

- Facebook’s daily logs: 60 Terabytes (60,000 Gigabytes)
- 1,000 genomes project: 200 Terabytes
- Google web index: 10+ Petabytes (10,000,000 Gigabytes)
- Disk speeds: order of 100 MBytes/sec, maybe a few times that for SSDs (solid-state disks).

Bottom line: need to break up—distribute—data and attach computation to each chunk.

Apache Spark

The Apache Spark framework provides a simple view of large datasets, permitting intuitive and flexible constructions analogous to Unix pipelines.

Map-Reduce

One type of such distributed computation is the Spark framework. It is still experimental and subject to implementation on the Spark framework.

Big Data

So what happens with huge amounts of data?

Examples (from Anthony Joseph, although a bit dated):

- Facebook’s daily logs: 60 Terabytes (60,000 Gigabytes)
- 1,000 genomes project: 200 Terabytes
- Google web index: 10+ Petabytes (10,000,000 Gigabytes)
- Disk speeds: order of 100 MBytes/sec, maybe a few times that for SSDs (solid-state disks).

Bottom line: need to break up—distribute—data and attach computation to each chunk.