Big Data and Security

Jeff Borowitz, PhD

Lecturer

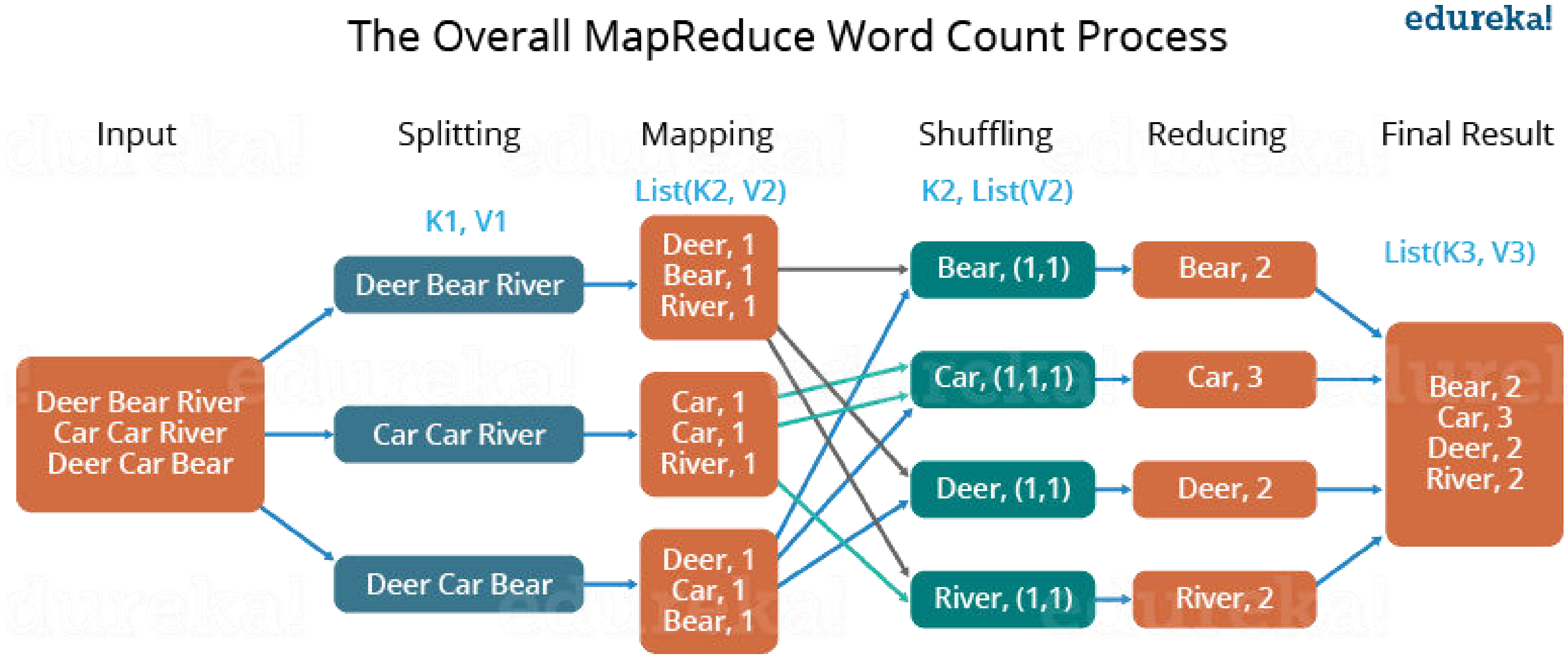
Sam Nunn School of International Affairs

Software Architecture for Big Data

MapReduce

- MapReduce is a programming framework developed by Google, and released in a paper in 2003
- It allows the splitting of work among an arbitrary number of computers
- Programming model
  - Start with a list of (key, value) pairs
  - Apply a “map” function
  - The result is a new list of different (key, value) pairs
  - Then, collect all the pairs with the same key together, so you have (key, list of values)
  - Then call the “reduce” function on this list, which produces a list of values corresponding to each key
- The Master node role (called “TaskTracker”) Dole out list to follower (“JobTracker”) nodes which do “map”
  - Collect and aggregate by second key
  - Collect final list

MapReduce Picture



The Overall MapReduce Word Count Process

edureka!

Input

Deer Bear River
Car Car River
Deer Car Bear

Splitting

K1, V1

Deer Bear River

Car Car River

Deer Car Bear

Mapping

List(K2, V2)

Deer, 1
Bear, 1
River, 1

Car, 1
Car, 1
River, 1

Deer, 1
Car, 1
Bear, 1

Shuffling

K2, List(V2)

Bear, (1,1)

Car, (1,1,1)

Deer, (1,1)

River, (1,1)

Reducing

Bear, 2

Car, 3

Deer, 2

River, 2

Final Result

List(K3, V3)

Bear, 2
Car, 3
Deer, 2
River, 2

Source: https://www.edureka.co/blog/mapreduce-tutorial/

Hadoop

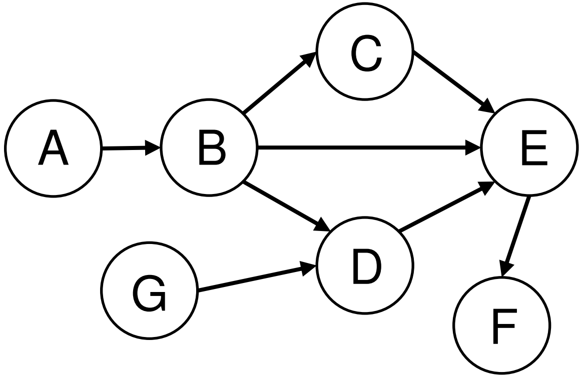
- Hadoop is an open source, commonly used system which can be used for MapReduce.
- Visit this module’s resource list to find related web links – there you will find more than you likely wanted to know
- Strengths
  - Parallelizes very parallel tasks very cheaply
  - Can be very redundant

Hadoop’s Weaknesses

- No partial data reuse
  - What if you want the word counts again, on a slightly larger bit of text?
- Fundamentally, it parallelizes very parallel tasks
- There’s a fair amount of overhead with every map reduce job.
- MapReduce is very low level

Spark: The Big Data Platform

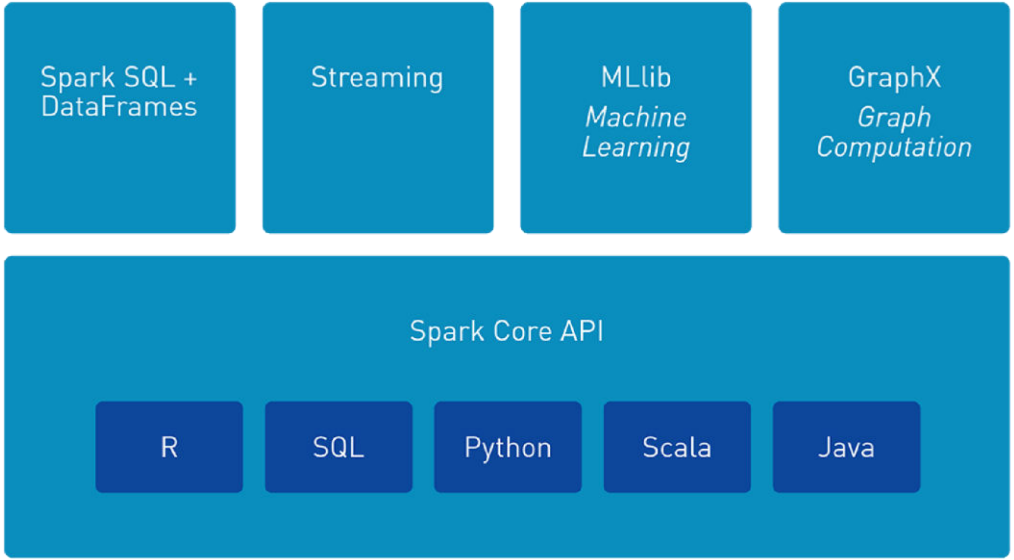
- Visit module 2 resources for a web link to Spark
- “Directed acyclic graph” computations
- Resilient distributed data stores
  - (Checks if computation is complete, if not, knows how to repeat)



What’s the Next Step?

- What do social scientists/data scientists/data analysts do with “big data”?
    - Data transformation (Extract, transform, load)
    - SQL-like queries, one off for an analysis or for a dashboard
    - Several versions of regressions

Spark has Higher Level Tools and Is Accessible in Many Languages



Source: https://mapr.com/products/apache-spark/

Spark and Other Languages

- Spark supports Java, Scala, and Python already
- R users can access Spark using sparklyr.
- Dataframes are key data constructs for analysis
  - It’s like a spreadsheet, and exists in R and Python.
- In sum: tooling is getting good so more and more people could use Spark

Who Needs Spark?

- One important question: Which sorts of questions really need to be answered on such large data?
- Let’s think about our biggest data sets we need to analyze?
  - Census bureau data: about 1,000 questions about 15M households
  - Micro-data on all health insurance customers for a company: about 50M people, with interactions
- People who do need Spark:
  - Drone service checking for broken infrastructure
  - E.g. Google, other web-scale companies.
  - Cybersecurity folks!

Spark and Cybersecurity

- IoT devices create a lot of data, but could be compromised. Is a given IoT device being maliciously controlled?
- For moderately sized computer systems, computers do tons of actions and create all kinds of technical logs.
- Is there evidence of malicious stuff in there? For a moderately sized computer system, there are likely tons of external requests for information - which of these is malicious?

In Summary

- MapReduce was created to distribute simple, parallel work amongst computers
  - Core example is processing web text
  - Hadoop is an open source platform one can use MapReduce on
- Spark is more practical for data analytics
  - It has higher level tools and works in many languages for those who need it