R and Hadoop

Ram Venkat
Dawn Analytics

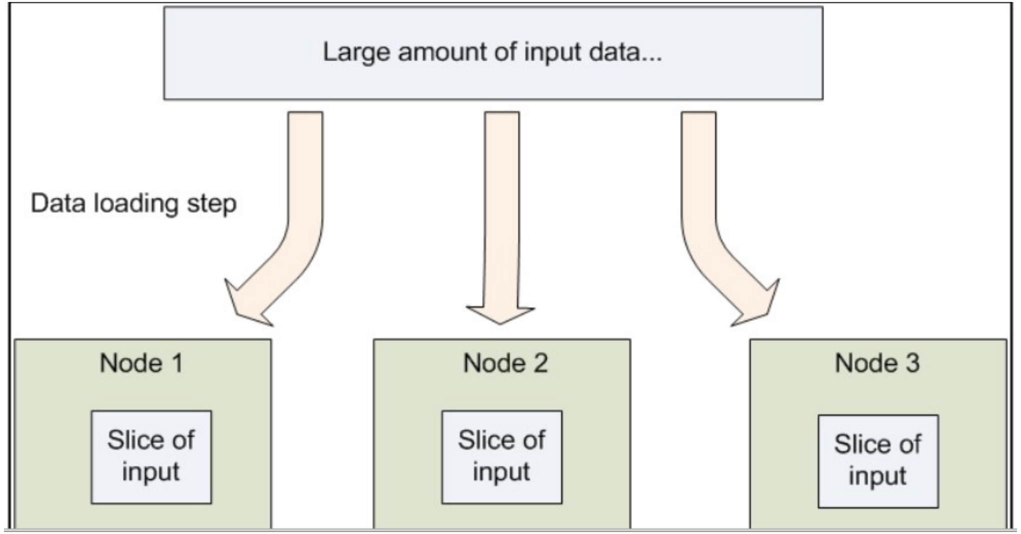
What is Hadoop?

- Hadoop is an open source Apache software for running distributed applications on 'big data'
- It contains a distributed file system (HDFS) and a parallel processing 'batch' framework
- Hadoop is written in java, runs on unix/linux for development and production
- Windows and Mac can be used as development platform
- Yahoo has > 43000 nodes hadoop cluster and Facebook has over 100 PB(PB= 1 M GB) of data in hadoop clusters

Hadoop overview (1/2)

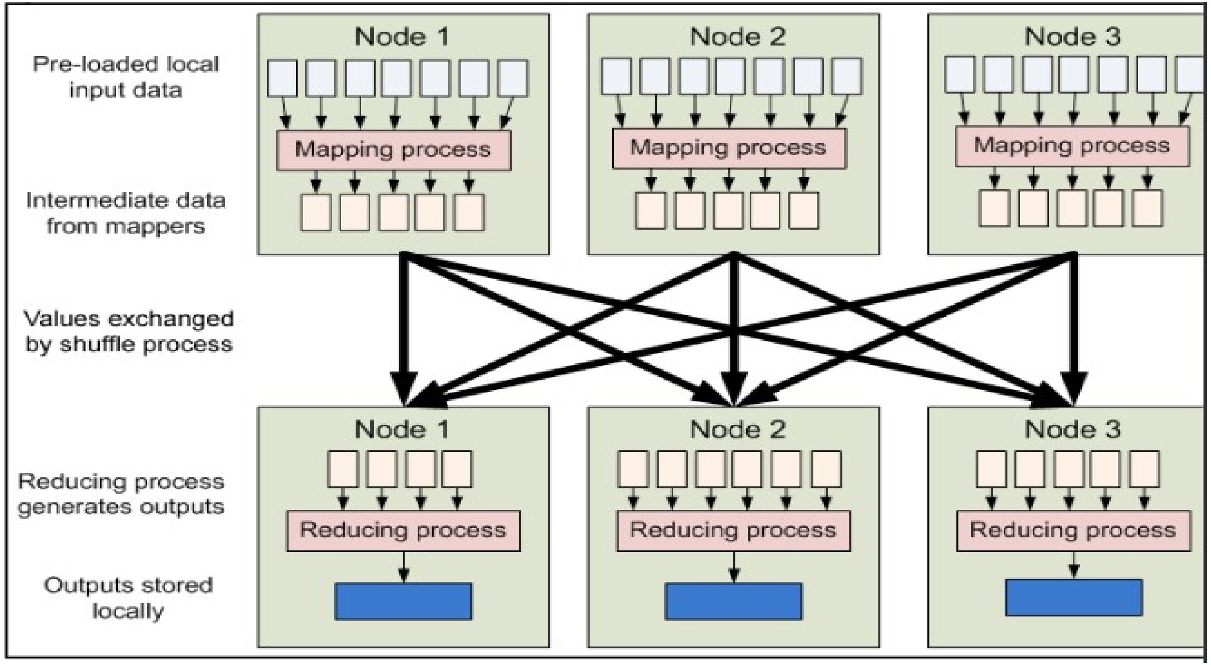
Central Idea: Moving computation to data and compute across nodes in parallel

- Data Loading



Hadoop overview (2/2)

Parallel Computation: MapReduce



Map Reduce : Example 'Hello word'

- Mathematically, this is what MapReduce is about:

  —map (k1, v1) ➔ list(k2, v2)

  —reduce(k2, list(v2)) ➔ list(<k3, v3>)

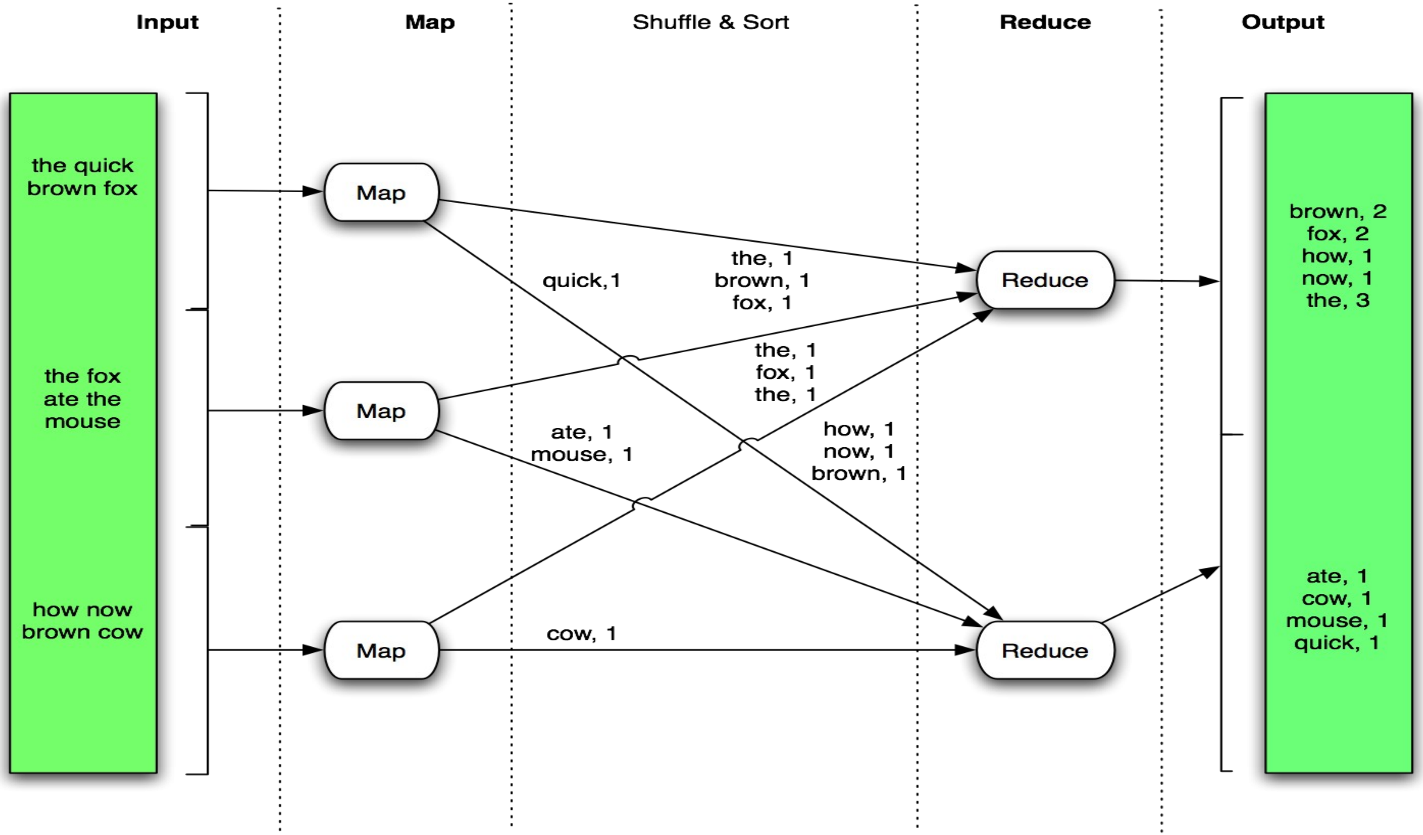
- Implementation of the 'hello word' (word count):

  Mapper: K1 -> file name, v1 -> text of the file

  K2 -> word, V2 -> “1”

  Reducer: Sums up the '1' s and produces a list of words and their counts

Word Count (slide from Yahoo)



R libraries to work with Hadoop

- 'Hadoop Streaming' - An alternative to the Java MapReduce API
- Hadoop Streaming allows you to write jobs in any language supporting stdin/stdout.
- R has several libraries/ways that help you to work with Hadoop:
  - Write your mapper.R and reducer.R and run a shell script
  - 'rmr' and 'rhadoop' from revolution analytics
  - 'rhipe' from Purdue University statistical computing
  - 'RHive' interacts with Hadoop via Hive query

Word Count Demo with R(rmr)

```
mapper.wordcount = function(key, val) {

               lapply(
                 strsplit( x = val, split = " ")[[1]],
                 function(w) keyval(w,1)
               )
}


 reducer.wordcount = function(key, val.list) {
             output.key = key
             output.val = sum(unlist(val.list))
                return( keyval(output.key, output.val) )
 }
```

More advanced example – Sentiment Analysis in R(rmr)

- One area where Hadoop could help out traders is in sentiment analysis
- Oreilly Strata blog 'Trading on sentiment' : http://strata.oreilly.com/2011/05/sentiment-analysis-finance.html
- Demo2 is modified code from this example from Jeffrey Breen on airlines sentiment analysis : http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/
- Jeffrey has been very active in R groups in Chicago area, This is another tutorial last month on R and Hadoop by Jeffrey : http://www.slideshare.net/jeffreybreen/getting-started-with-r-hadoop

Demo2 Sentiment Analysis with rmr

```
mapper.score = function(key, val) {

# clean up tweets with R's regex-driven global substitute, gsub():
    val = gsub('[[:punct:]]', '', val)
    val = gsub('[[:cntrl:]]', '', val)
    val = gsub('\\d+', '', val)

# Key is the Airline we added as tag to the tweets
    airline = substr(val,1,2)

# Run the sentiment analysis
    output.key = c(as.character(airline), score.sentiment(val,pos.words,neg.words))

# our interest is in computing the counts by airlines and scores, so 'this' count is 1
    output.val = 1
    return( keyval(output.key, output.val) )
}
```

Demo3 - Hive

- Hive is a data warehousing infrastructure for Hadoop
- Provides a familiar SQL like interface to create tables, insert and query data
- Behind the scene , it implements map-reduce
- Hive is an alternative to our hadoop streaming we covered before
- Demo3 – stock query with Hive

Use cases for Traders

- Stock sentiment analysis
- Stock trading pattern analysis
- Default prediction
- Fraud/anomoly detection
- NextGen data warehousing

Hadoop support - Cloudera

- Cloudera distribution of hadoop is one of the most popular distribution (I used their CDH3v5 in my 2 demos above)
- Doug Cutting, the creator of Hadoop is the architect with Cloudera
- Adam Muise , a Cloudera engineer at Toronto is the organizer of Toronto Hadoop user Group (TOHUG)
- Upcoming meeting organized by TOHUG on the 30th October - “PIG-fest”

Hadoop Tutorials and Books

- http://hadoop.apache.org/docs/r0.20.2/quickstart.html
- Cloudera: http://university.cloudera.com/
- Book: “Hadoop in Action” – Manning
- Book: “Hadoop - The Definitive Guide” – Oreilly
- Hadoop Streaming: http://hadoop.apache.org/docs/mapreduce/r0.21.0/streaming.html
- Google Code University: http://code.google.com/edu/parallel/mapreduce-tutorial.html
- Yahoo's Tutorial : http://developer.yahoo.com/hadoop/tutorial/module1.html

Thank You

For any clarification, send e-mail to ram@dawnanalytics.com