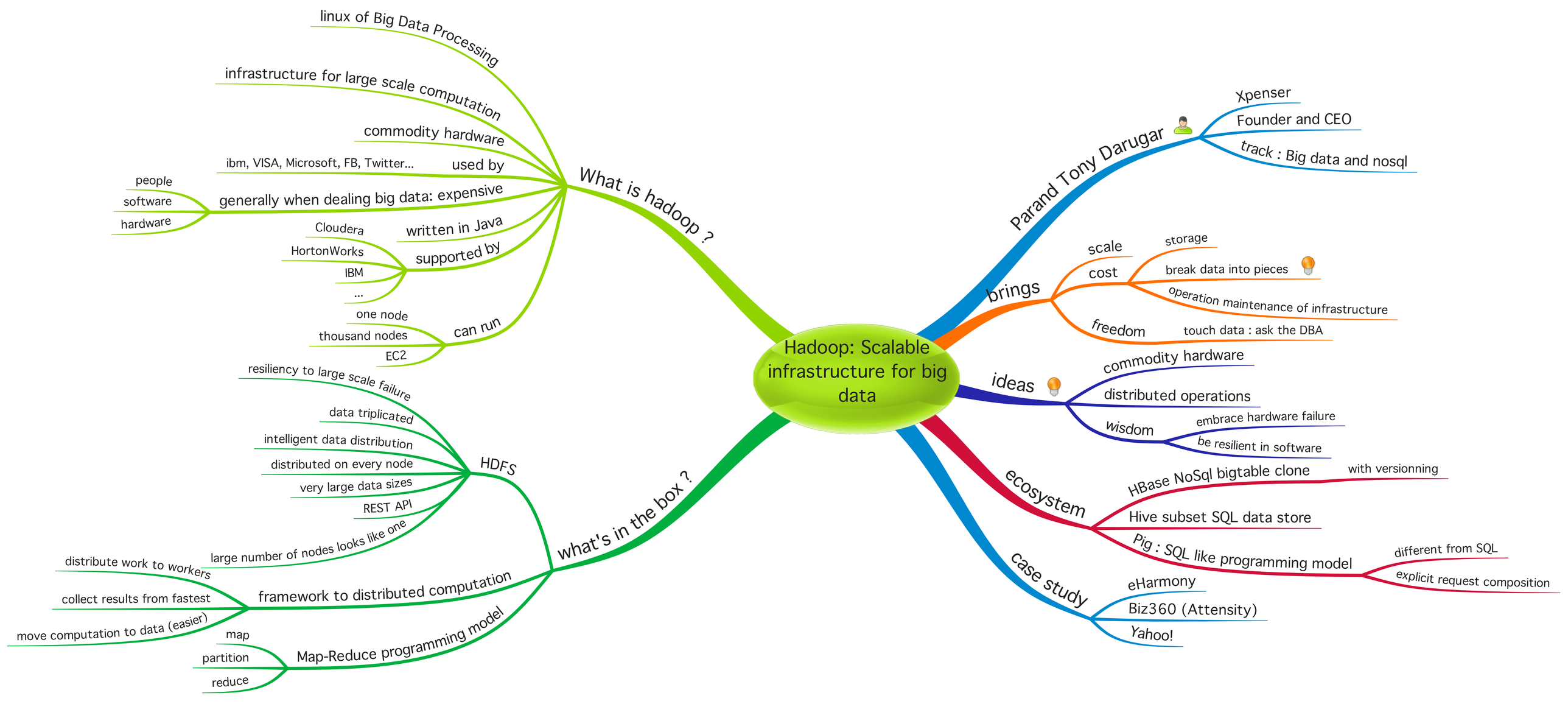
Hadoop: Scalable infrastructure for big data
What is hadoop ?
linux of Big Data Processing
infrastructure for large scale computation
commodity hardware
used by
ibm, VISA, Microsoft, FB, Twitter...
generally when dealing big data: expensive
people
software
hardware
written in Java
supported by
Cloudera
HortonWorks
IBM
...
can run
one node
thousand nodes
EC2
what's in the box ?
HDFS
resiliency to large scale failure
data triplicated
intelligent data distribution
distributed on every node
very large data sizes
REST API
large number of nodes looks like one
framework to distributed computation
distribute work to workers
collect results from fastest
move computation to data (easier)
Map-Reduce programming model
map
partition
reduce
Parand Tony Darugar
Xpenser
Founder and CEO
track : Big data and nosql
brings
scale
cost
storage
break data into pieces
operation maintenance of infrastructure
freedom
touch data : ask the DBA
ideas
commodity hardware
distributed operations
wisdom
embrace hardware failure
be resilient in software
ecosystem
HBase NoSql bigtable clone
with versionning
Hive subset SQL data store
Pig : SQL like programming model
different from SQL
explicit request composition
case study
eHarmony
Biz360 (Attensity)
Yahoo!