2013

What’s Behind the Curtain? The Infrastructure Supporting Big Data

Greg Wolffe
Grand Valley State University, wolffe@gvsu.edu

What’s Behind the Curtain?

the infrastructure supporting Big Data

greg wolffe
School of Computing and Information Systems

MGM

Big, really big, data

practical definition of Big Data:

“current methods won’t work”

- IBM Systems and Technology Group

CERN: Higgs boson

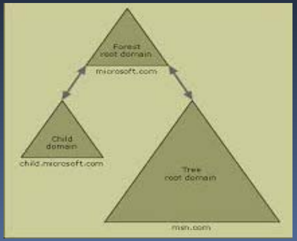
- 200 Petabytes

Networking

metagenomics :
human gut biome = 500 GBase
⇨ ~12 hrs

NOAA
Integrated Surface Weather Data
(1903-present)

Ethernet ⇨ Infiniband



Storage

Facebook: Instagram

commercial storage

- I/O bottleneck

NFS ⇨ HDFS

Exa	1,152,921,504,606,846,976 bytes
Peta	1,125,899,906,842,624 bytes
Tera	1,099,511,627,776 bytes
Giga	1,073,741,824 bytes
Mega	1,048,576 bytes
Kilo	1,024 bytes
Byte	1 byte

High-performance computing

real-time analysis

- Smart grid
- financials

large-scale analysis

- genomics
- retail data mining

Google

not a "traditional" supercomputer ⇨ HTC

Google

the “Cloud”

concept:

- 2003 – “The Google File System”
- 2004 – “MapReduce: Simplified Data Processing on Large Clusters”

⇨ Hadoop

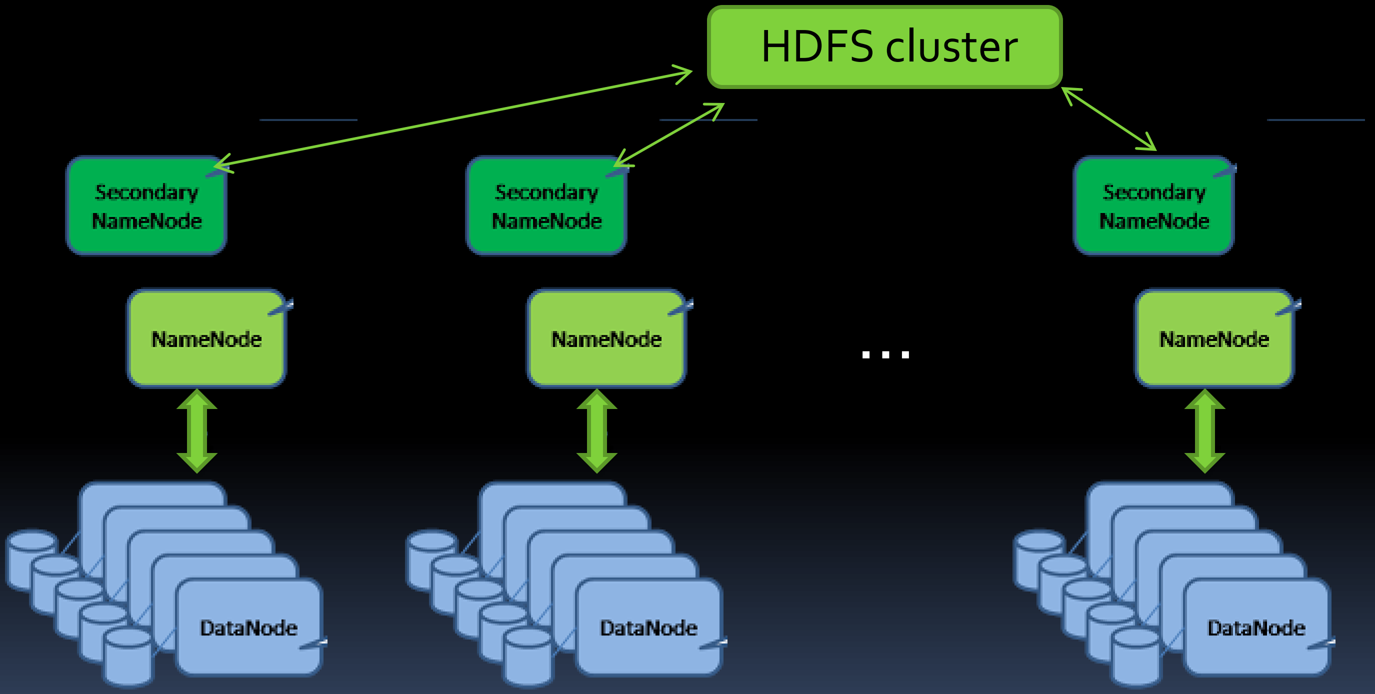
infrastructure:

- Amazon Web Services

⇨ the Cloud

Hadoop Distributed File System

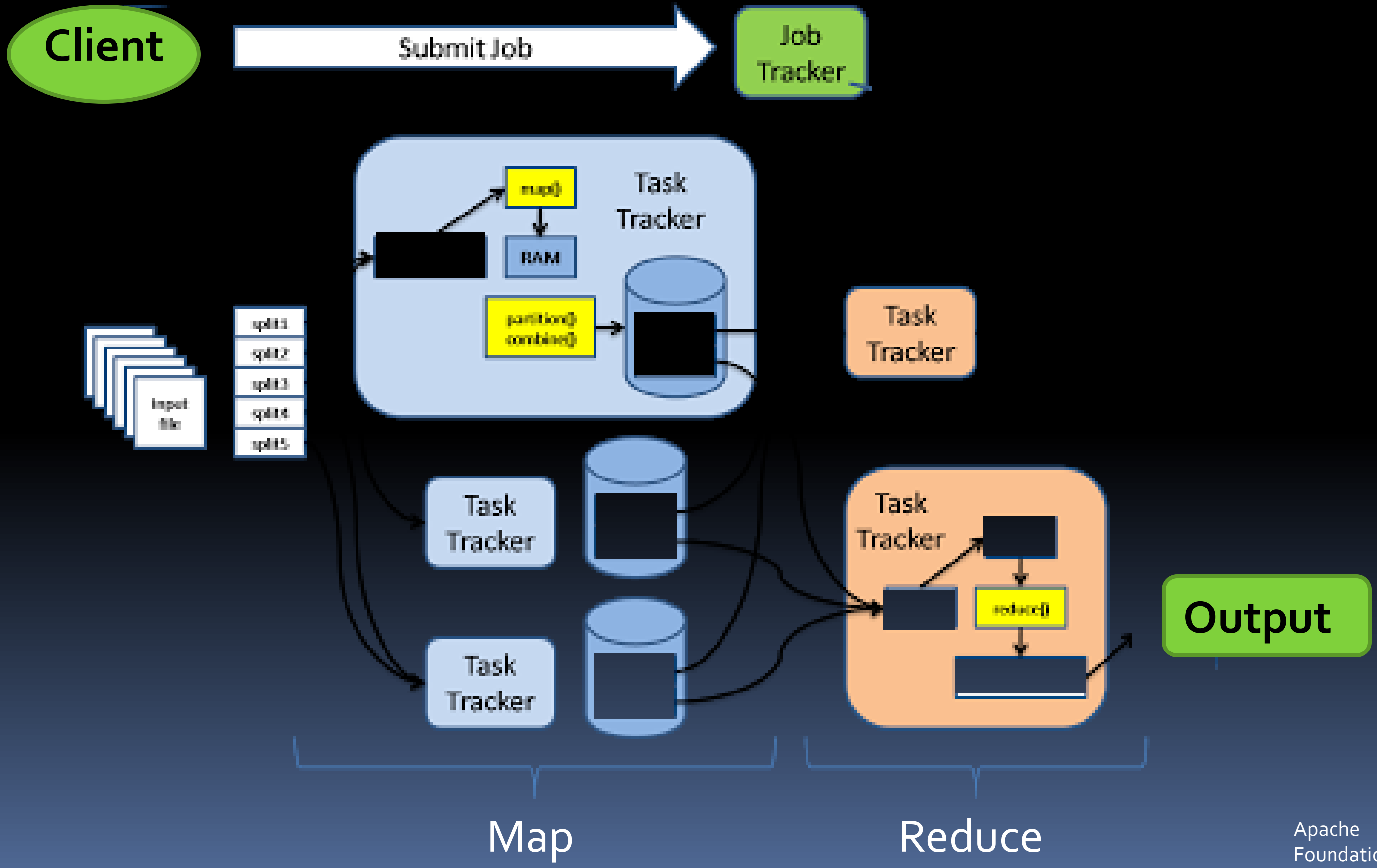
- flexible, redundant, optimized



Apache
Foundation

MapReduce

- massively distributed execution



Apache Foundation

Brave new world

"It is very sad that nowadays there is so little useless information."
– Oscar Wilde.