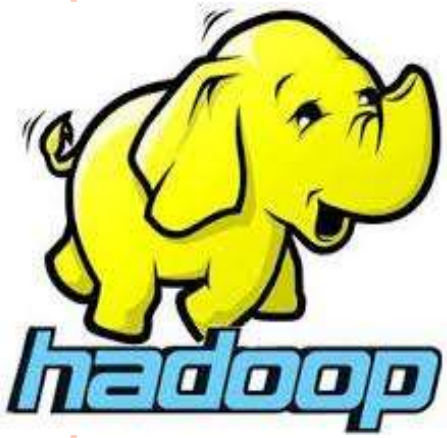


HADOOP TECHNOLOGY

Contents

Introduction to Hadoop

- HADOOP WAS CREATED BY DOUGH CUTTING AND MIKE CAFARELLA IN 2005.
- HADOOP USES A CLUSTER OF COMMODITY SERVERS IN TIGHTLY CONNECTED NETWORK
- HADOOP IS A OPEN SOURCE FRAME WORK WRITTEN IN JAVA.
- DISTRIBUTED DATA STORGE.
- PARALLEL PROCESSING OF DATA.

Cluster of machines running Hadoop at Yahoo!

Why Hadoop ?

SCENARIO 1ST :-

Processing Vcards:

Example of VCARD

Process VCards and extract,

Email addresses

Twitter & Facebook URLs

(Assuming 100 vcards per second)

1 million VCards takes ~= 160 minutes

100 million VCards takes ~= 11 days

SCENARIO 2:-

- 1 GB – 10 GB – 100 GB --- limits
- More Investments
- -- 10 TB – 100 TB --- again limits
- Data from Facebook, Twitter, RFID readers, sensors.
- Structured / Unstructured

What are the basic problems with :-

Processing large/unstructured/complex data ?

Storage problem

Processing problem

Solution is ?

Requirements

We need a system that should support :-

Multiple machines

Built in backup

Built in fail-over

Distributed Parallel processing

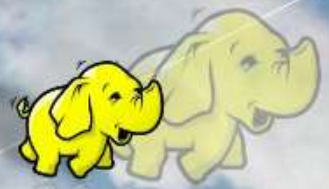
Ability to distribute work evenly on all machines

A single System

Easy to use

Hadoop….

WHAT HADOOP ISN'T



- A replacement for relational and data warehouse systems
- A transactional / online / serving system
- A low latency or streaming solution

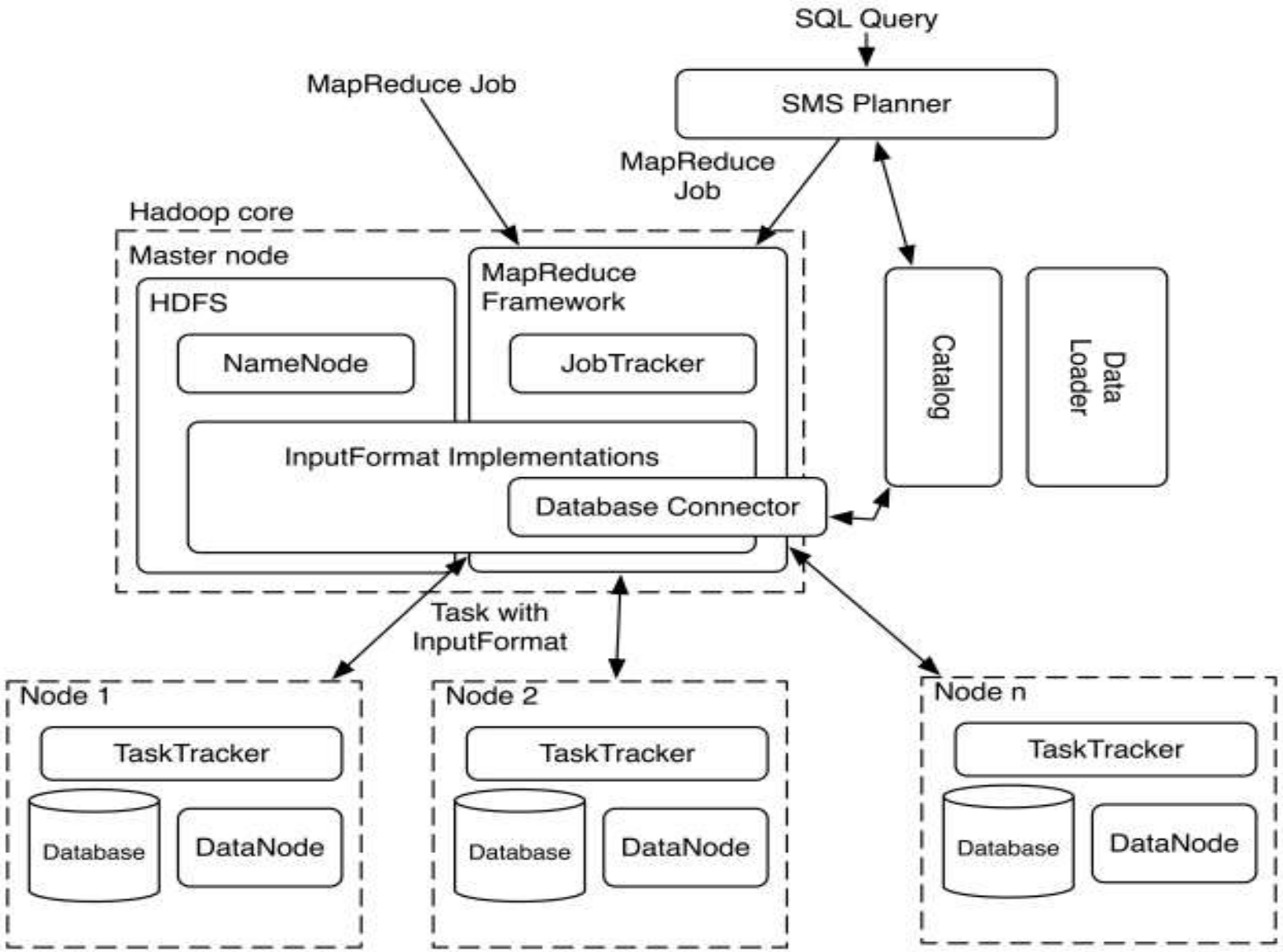
Pillars of Hadoop

- **Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.**

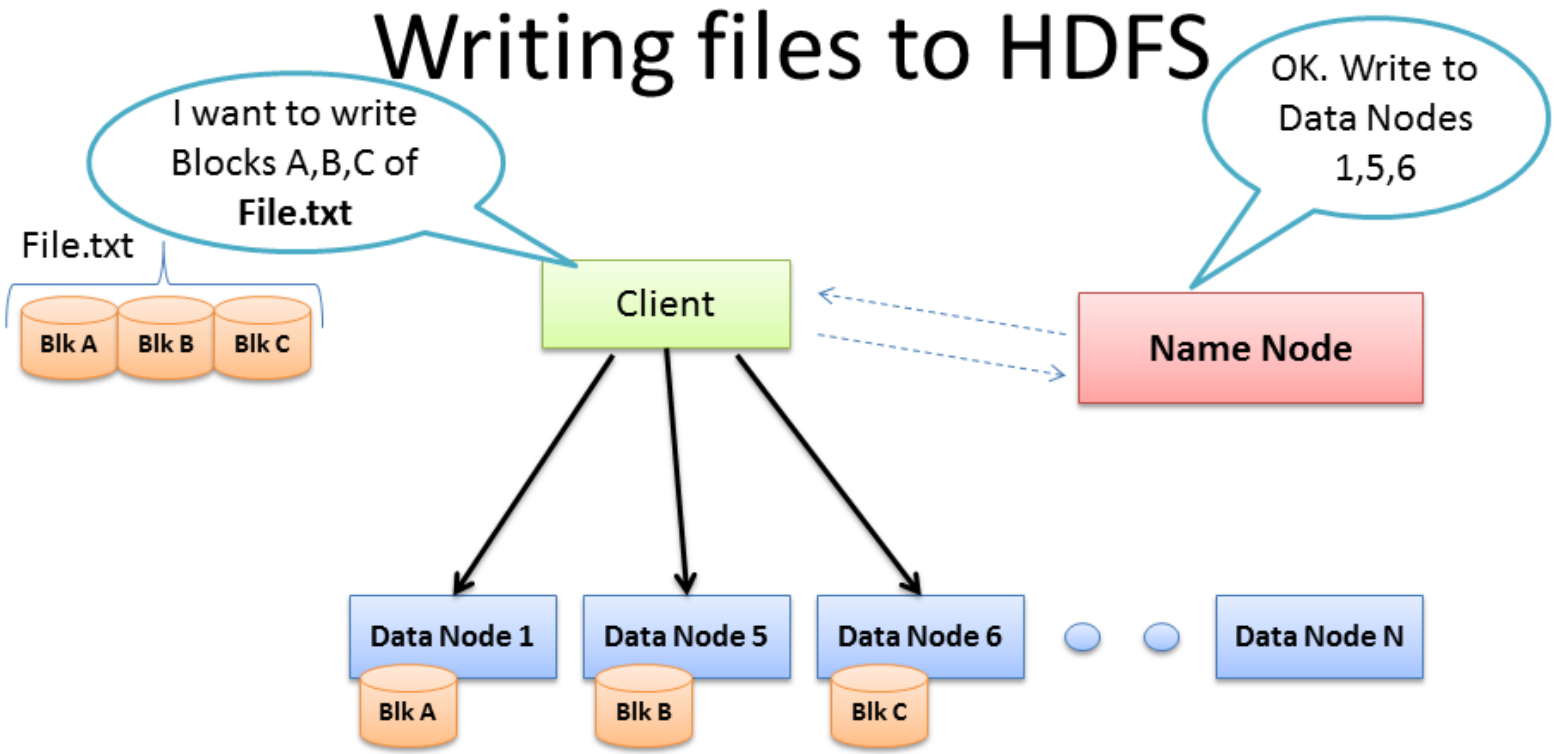
- **Hadoop YARN – a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.**

- **Hadoop MapReduce – a programming model for large scale data processing.**

Architecture of Hadoop



HDFS Architecture



- Client consults Name Node
- Client writes block directly to one Data Node
- Data Nodes replicates block
- Cycle repeats for next block

- **Name node:-** The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the NameNode by inodes.

- **Data Node:-** Each block replica on a DataNode is represented by two files in the local native filesystem. The first file contains the data itself and the second file records the block's metadata

- **HDFS Client:-** User applications access the filesystem using the HDFS client, a library that exports the HDFS filesystem interface.

MapReduce in Hadoop

• **Mapreduce** is a progrmming model for processing and generating large data sets with a parallel, distributed algorithms on a cluster

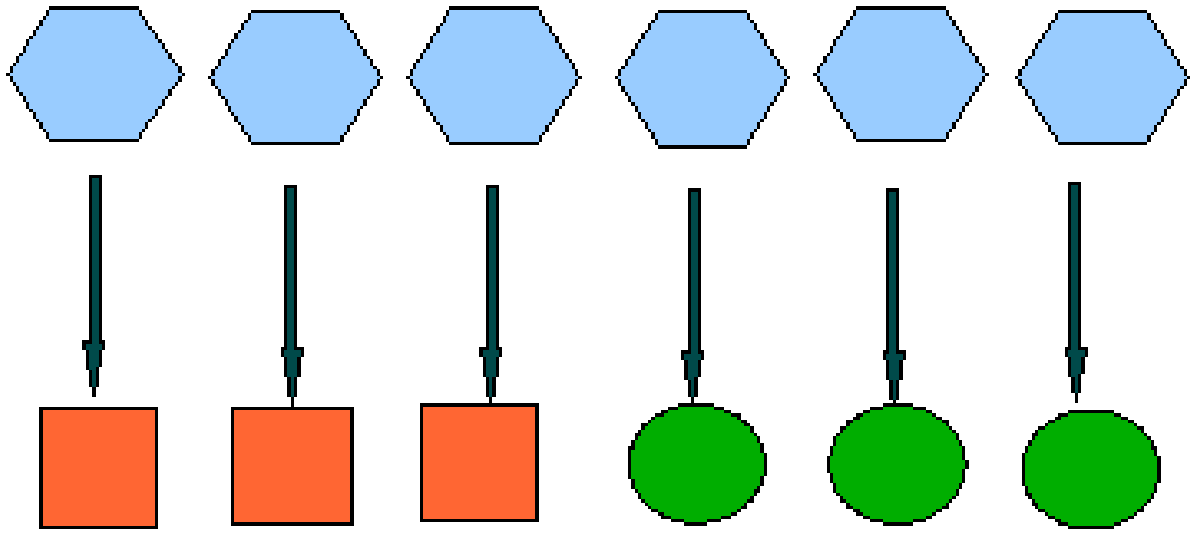
• **MapReduce** is an associated implementation for processing and generating large data sets.

• A **Map-Reduce** job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner.

• A **MapReduce** job is a unit of work that the client wants to be performed: it consists of the input data, the MapReduce program, and configuration information. Hadoop runs the job by dividing it into tasks, of which there are two types: map tasks and reduce tasks

The Programming Model Of MapReduce

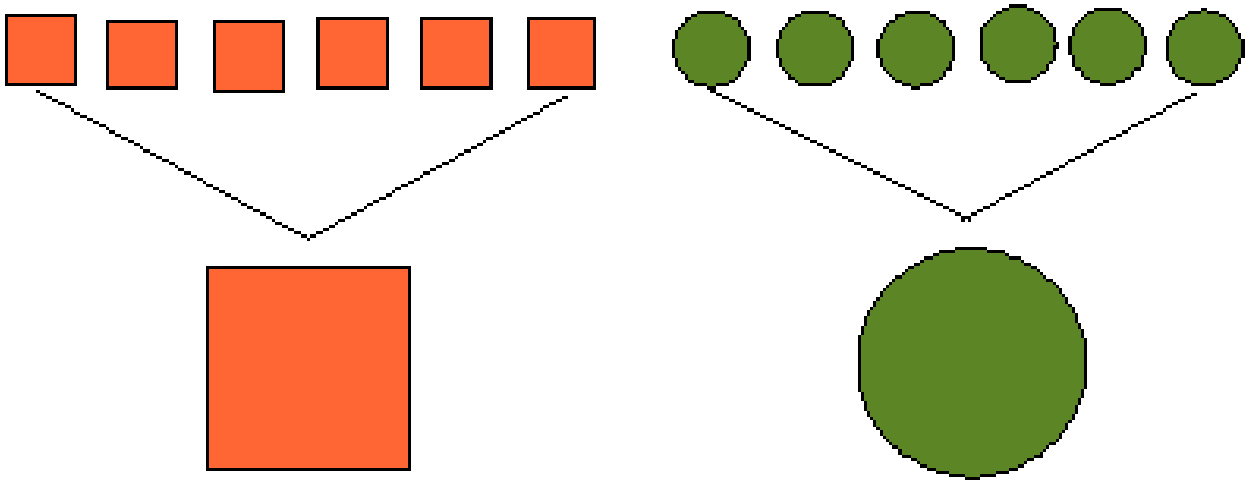
- *Map*, written by the user, takes an input pair and produces a set of *intermediate* key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key I and passes them to the *Reduce*



MAP

map (in_key, in_value) -> (out_key, intermediate_value) list

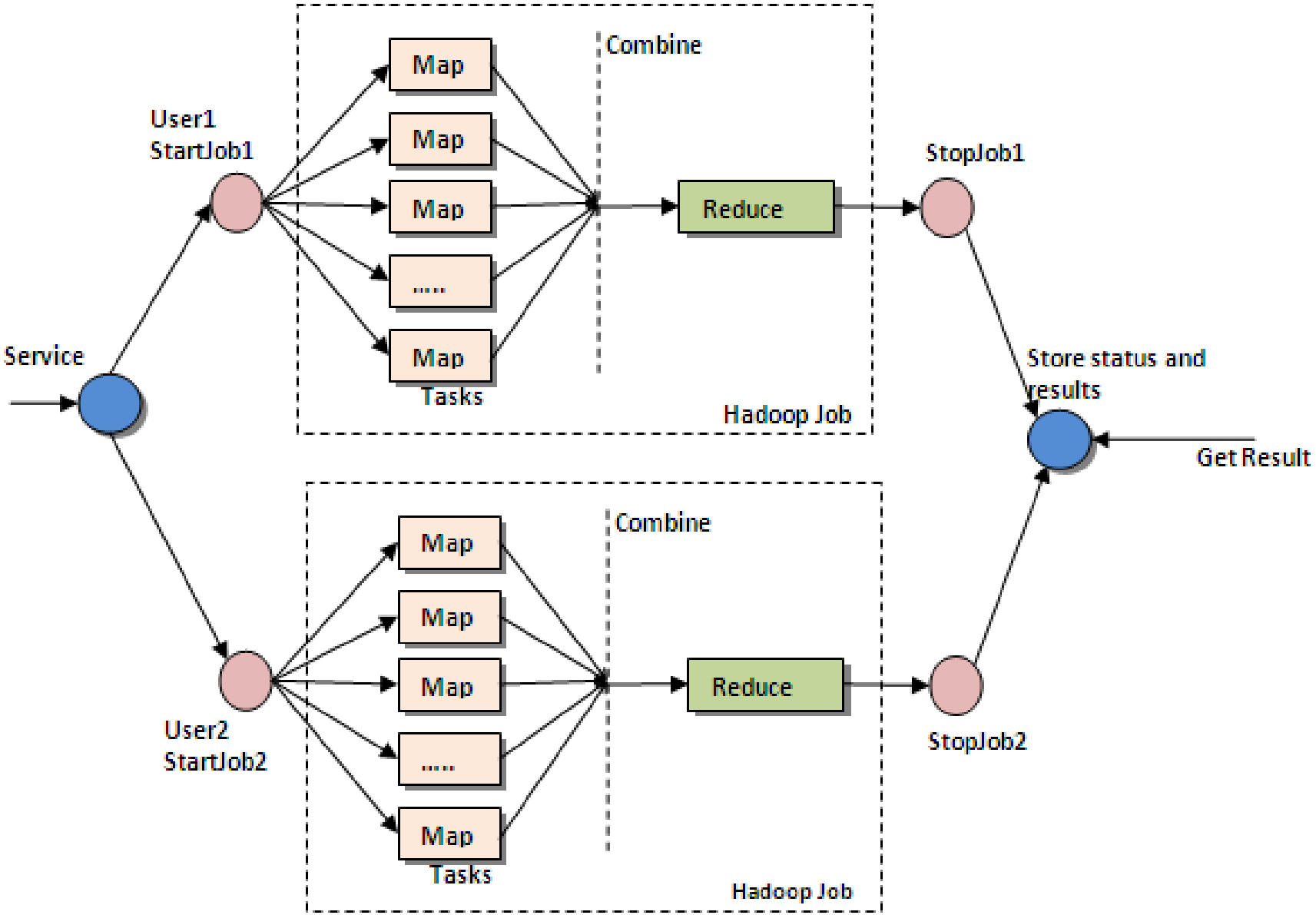
- The *Reduce* function, also written by the user, accepts an intermediate key I and a set of values for that key. It merges together these values to form a possibly smaller set of values



REDUCE

reduce (out_key, intermediate_value list) ->out_value list

MapReduce Architecture



Hadoop Projects

- Pig
- Mahout
- Hive
- Avro
- Strom

Distributers of Hadoop

- Amazon web
- Services
- Apache Bigtop
- Cascading
- Cloudera
- Cloudspace
- Datameter

Users Of Hadoop

2007

YAHOO!
Powerset
last·fm

2008

Google
ao e grape
ImageShack
Cascading
IBM
facebook
ENORMO
A9
krugle
rackspace HOSTING
Lookery Control freaks welcome
The New York Times
Joost
Zvents
FORMATION CIENCES INSTITUTE
News Corporation
Cornell University Computing and Information Science
Visible MEASURES
LOTAME
NetSeer
parc Palo Alto Research Center
veoh

2009

AOL
cloudera
deepdyve
cooliris
eyealike
TEXTMAP THE ENTITY SEARCH ENGINE
iterend
tailsweep
hulu
RapLeaf
USCMS
Ning
quantcast
amazon web services
pressflip
detikSearch
WorldLingo
Systems@ETH Zürich
VK SOLUTIONS Global Solutions Provider
TARAGANA
HOSTING HABITAT
HOLA
Terrier
adknowledge
stampede beta

2010

SAMSUNG
rubicon
BERKELEY LAB
VISIBLE
APOLLO GROUP
ADSDAQ
rackspace HOSTING
RapLeaf
wordnik
MOBIGEN
comScore
trulia real estate search
Accela COMMUNICATIONS
Forward3D
Linked in
Microsoft
Infochimps Find the world's data
Pharm2Phork
ADMELD
gumgum
BrainPad
Pronux
The Datagraph Blog
NETFLIX
mobileanalytics.tv
markt24.de
twitter
media6degrees
BEEBLER
SLC Security
ebay

Conclusion

- As the amount of data being stored around the globe continues to rise and the cost of technologies that enable the extraction of meaningful patterns . As the amount of data and cost of handling it increases this make difficult to organization to affort the cost and store the high amount of data and to process it.Then the hadoop is the best choice for the growing world by its easy handling and large storing of data.