Hadoop_Big Data
• Apache top level project, open-source implementation of frameworks for reliable, scalable, distributed computing and data storage.

• It is a flexible and highly-available architecture for large scale computation and data processing on a network of commodity hardware.
• Designed to answer the question: “How to process big data with reasonable cost and time?”
Search engines in 1990s

MetaCrawler Parallel Web Search Service
by Erik Selberg and Oren Etzioni

1996

MetaCrawler Parallel Web Search Service
by Erik Selberg and Oren Etzioni

Try the new MetaCrawler Beta!
If you're searching for a person's home page, try Abot!

- Examples
- Beta Site
- Add Site
- About

Search for:

- as a Phrase
- All of these words
- Any of these words
- Fast Search
- Comprehensive Search

For better results, please specify:

- Search Region: The World
- Search Sites: Any

Performance parameters:

- Max wait: 15 minutes
- Match type: Loose

webmaster@metacrawler.com

1996

Lycos
It's amazing where Go Get it will get you.

Find: Go Get it

Enhance your search

1996

The Wired Search Center

Search: The Web

Find: 10 full descriptions

1997

Search engines in 1990s

1996

Excite

Researching stocks?
Buying a car?
Planning a wedding?
Check out ExciteSeeing Tours.

Excite Reviews: site reviews by the web's best editorial team.

Arts
Business
Computing
Hobbies
Education
Health
Life & Style
Politics
Regional
Science
Shopping
Sports

Bill Mitchell: Satire that clicks!

Wired

The Wired Search Center

Search: The Web

Find: 10 full descriptions

1997
Google search engines

1998

2013
The Google File System
Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google

MapReduce: Simplified Data Processing on Large Clusters
Jeffrey Dean and Sanjay Ghemawat
jeff@google.com, sanjay@google.com
Google, Inc.

Bigtable: A Distributed Storage System for Structured Data
Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tahina Chandra, Andrew Fikes, Robert E. Gruber
info@bigtable.com
Google, Inc.
What is Hadoop?

- **Hadoop:**
  
  - an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license.

- **Goals / Requirements:**
  
  - Abstract and facilitate the storage and processing of large and/or rapidly growing data sets
    - Structured and non-structured data
    - Simple programming models
  
  - High scalability and availability
  
  - Use commodity (cheap!) hardware with little redundancy
  
  - Fault-tolerance
  
  - Move computation rather than data
Hadoop Framework Tools

- BI Applications (query, analytics, reporting, statistics)
  - Orchestration Framework
    - HBase
    - Chukwa
    - Flume
    - ZooKeeper
  - Data Access Framework
    - Pig
    - Hive
    - Sqoop
    - Avro
  - Network
    - Data Storage Framework (HDFS)
    - Data Processing Framework (MapReduce)
  - JVM
  - Operating System (Linux)
  - Dell PE-R, PE-C Servers
Hadoop in the Wild

**Three main applications of Hadoop:**

- Advertisement (Mining user behavior to generate recommendations)
- Searches (group related documents)
- Security (search for uncommon patterns)