

```
/*  
 * @package WordPress  
 * @subpackage Default_Theme  
 */  
?>  
<!DOCTYPE html PUBLIC "-//W3C/  
html xmlns="http://www.w3.  
<head profile="http://  
<meta http-equiv="Co  
<title><?php vp  
<link rel="*  
<link rel.  
<caty!
```

# Hadoop\_Big Data



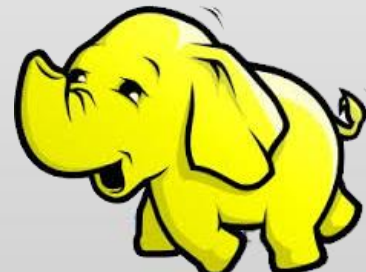
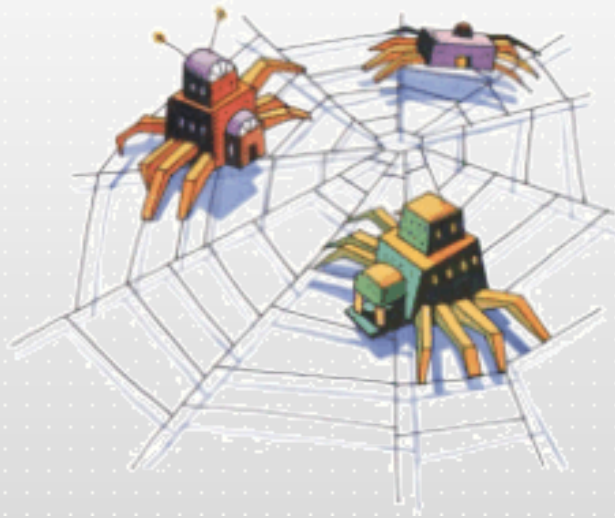
```
/*
 * @package WordPress
 * @subpackage Default_Theme
 */
?>
<!DOCTYPE html PUBLIC "-//W3C
html xmlns="http://www.w3.
<head profile="http://
meta http-equiv="Co
<title><?php vp
<link rel="
<link rel="
<caty!
```



- 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.

# Brief History of Hadoop

- Designed to answer the question:  
**“How to process big data with reasonable cost and time?”**



# Search engines in 1990s

```

<?package WordPress
<?package Default_Theme
<?
<!DOCTYPE html PUBLIC "-//W3C
html xmlns="http://www.w3.
head profile="http://
meta http-equiv="Co
<title><?php vp
<link rel="
<link rel="
<?php
  
```



## 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 [Abov!](#)

[Examples](#) • [Beta Site](#) • [Add Site](#) • [About](#)

Search for:

as a Phrase  All of these words  Any of these words

For better results, please specify:

Search Region:  Search Sites:

Performance parameters:

Max wait:  minutes Match type:

[About](#) | [Help](#) | [Problems](#) | [Add Site](#) | [Search](#)  
[webmaster@metacrawler.com](mailto:webmaster@metacrawler.com)  
 © Copyright 1995, 1996 Erik Selberg and Oren Etzioni

1996

**excite** search reviews city.net NEW live! reference?

excite home maps news people finder

**Excite Search:** twice the power of the competition.

What:

Where: World Wide Web

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

- Arts
- Business
- Computing
- Education
- Entertainment
- Health
- Hobbies
- Life & Style
- Money
- News & Reference
- Personal Pages
- Politics & Law
- Regional
- Science
- Shopping
- Sports

[Bill Mitchell](#)  
Satire that clicks!

1996

Serious Sports Fans Only \$1,000,000 in Cash and Prizes!  
For serious sports fans only! Play Fantasy Football!

**LYCOS** It's amazing where  
Go Get It will get you.

Find:

[Enhance your search.](#)



[New Search](#) • [Top News](#) • [Sites by Subject](#) • [Top 5% Sites](#) • [City Guide](#) • [Pictures & Sounds](#)  
[PeopleFind](#) • [Point Review](#) • [Road Maps](#) • [Software](#) • [About Lycos](#) • [Club Lycos](#) • [Help](#)

[Add Your Site to Lycos](#)

Copyright © 1996 Lycos™, Inc. All Rights Reserved.  
Lycos is a trademark of Carnegie Mellon University.  
[Questions & Comments](#)

1996

HELP WIRED NEWS HOTWIRED WIRED MAGAZINE SUCK.COM

**The WRED Search Center**

look for:

for more options use [SuperSearch](#)

Date:

Country:

Include media type:

Image  Audio  Video  Downloads

Return Results:

Sandbox Entertainment

Shop WIRED Holiday Gift Guide

**SOMETHING HAS SURVIVED.**

Find more deals

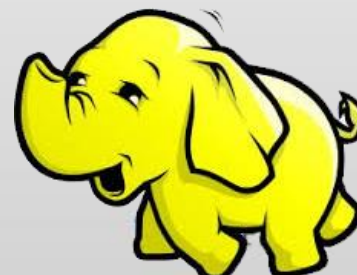
Log

Cyberlan Outpost

Microsoft® Expedia™ Travel

ONSAFE

1997

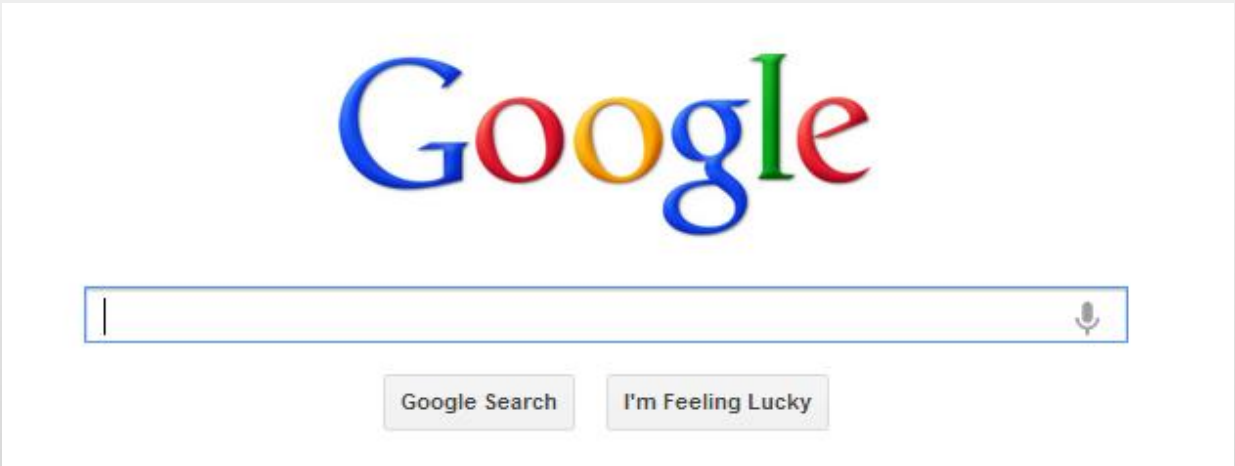


# Google search engines

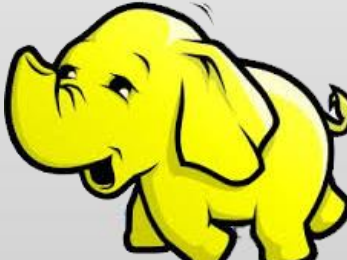
```
/*  
 * @package WordPress  
 * @subpackage Default_Theme  
 */  
  
<?php  
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">  
<html xmlns="http://www.w3.org/1999/xhtml" <!-- [document profile] -->  
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />  
<title><?php wp_title()></title>  
<link rel="stylesheet" type="text/css" href="http://www.wordpress.com/wp-content/themes/default/css/style.css" />  
<link rel="stylesheet" type="text/css" href="http://www.wordpress.com/wp-content/themes/default/css/print.css" />  
</html>
```



1998



2013







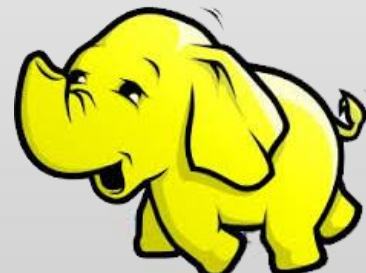
# What is Hadoop?

- **Hadoop:**

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

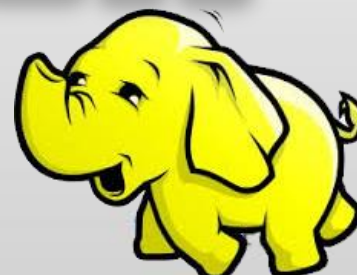
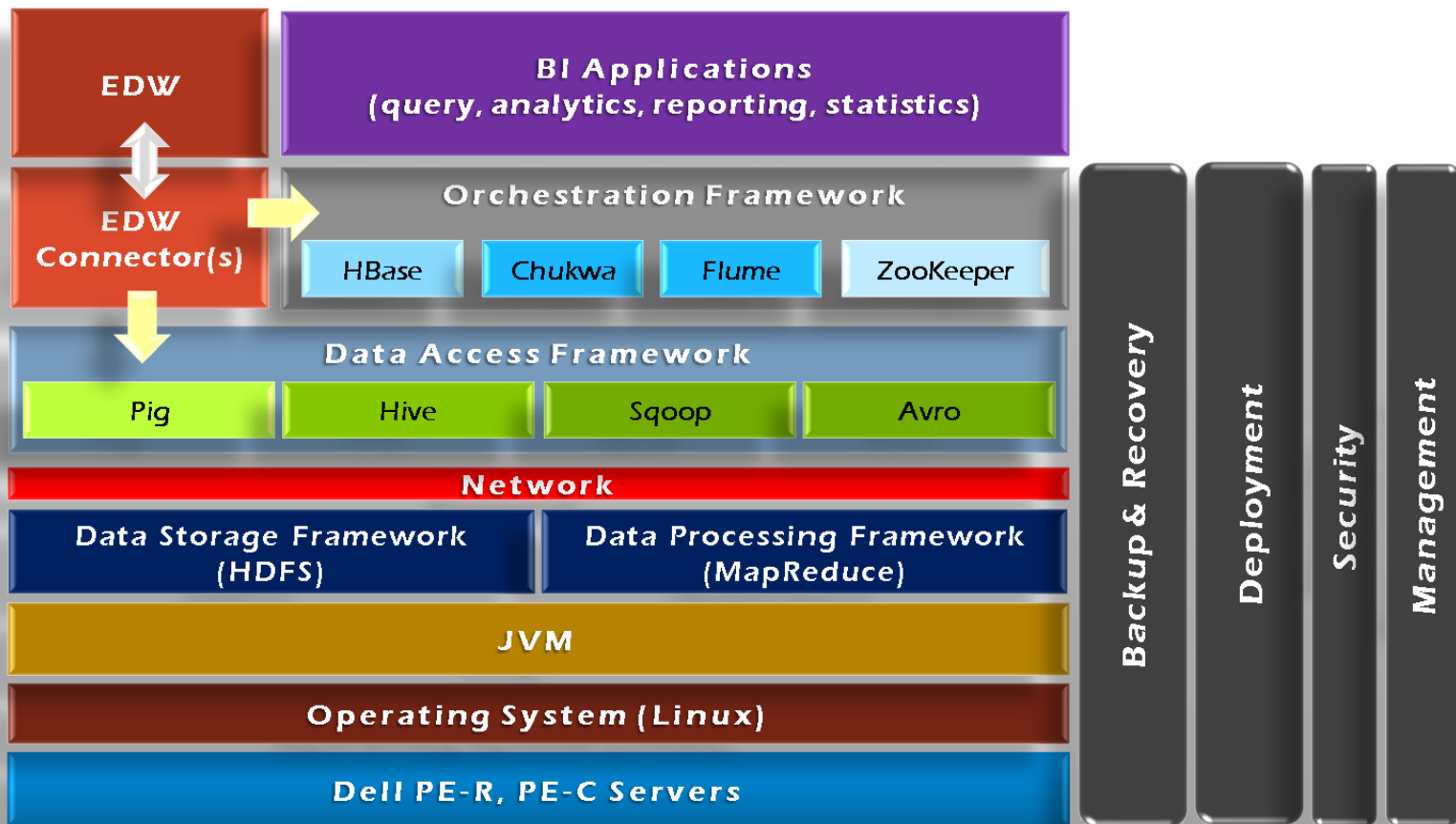
- **Goals / Requirements:**

- 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

```
<!-- Package WordPress -->
<!-- Package Default_Theme -->
?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" >
<head profile="http://gmpg.org/xfn/1.0">
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
<title><?php wp_title() ?></title>
<link rel="stylesheet" type="text/css" href="http://www.wordpress.com/wp-content/themes/default/css/style.css" />
</head>
<body >
</body>
</html>
```





# 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)

