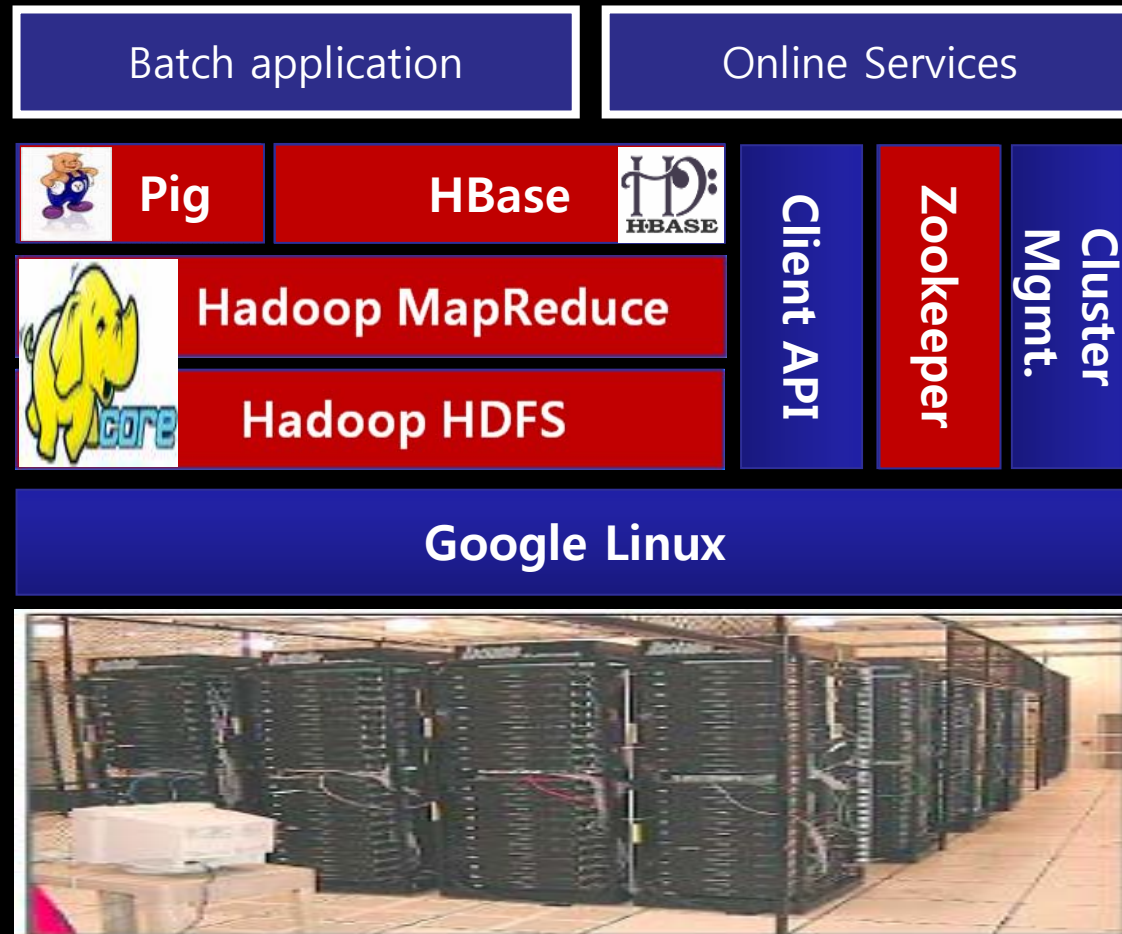


Hadoop

open-source software for reliable,
scalable, distributed computing

김형준, NHN

Hadoop overview



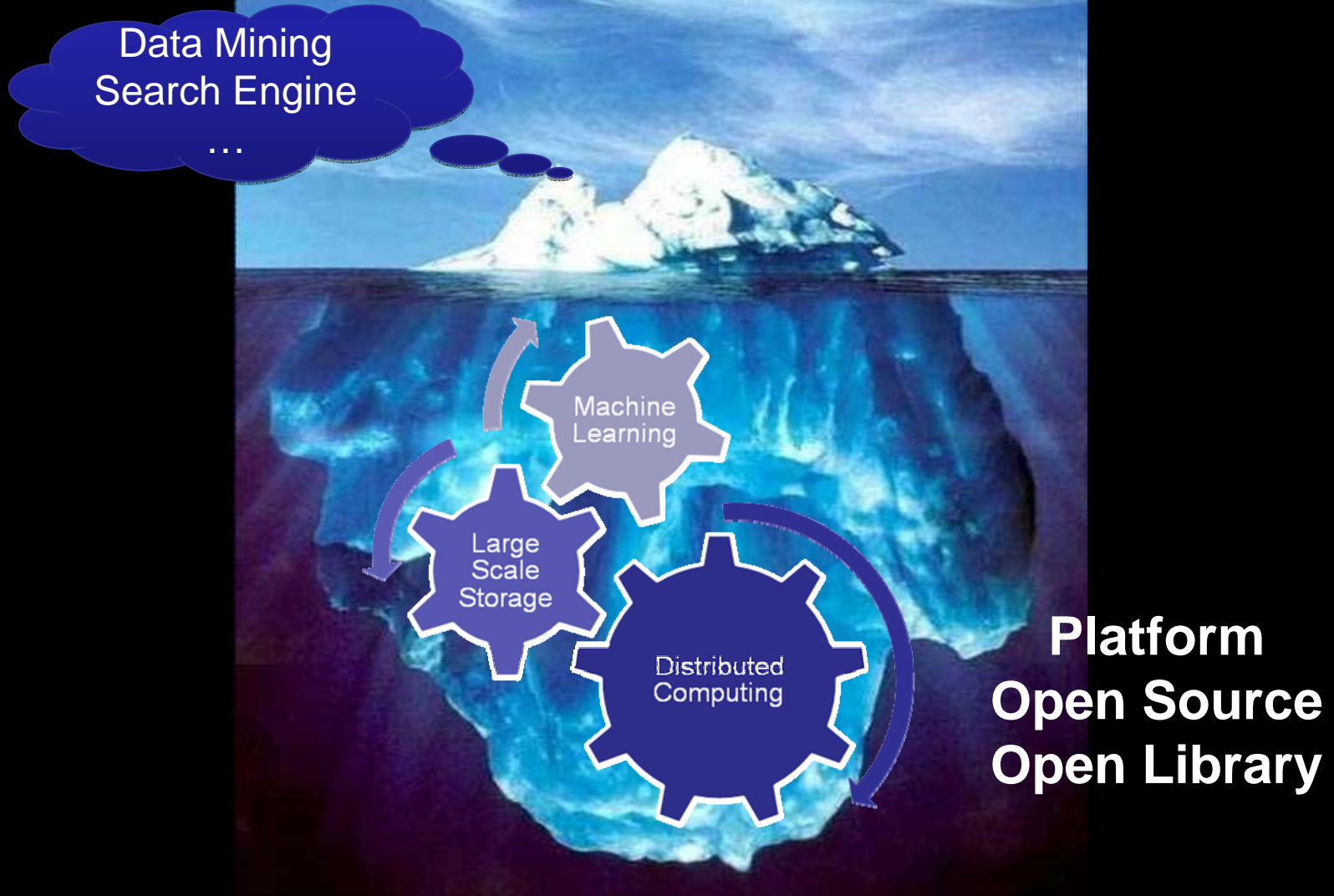
A Brief History of Hadoop

- Feb 2003: First MapReduce library written at Google
- Oct 2003: Google File System paper published
- Dec 2004: Google MapReduce paper published
- Jul 2005: Nutch uses new MapReduce implementation
- Nov 2006: Google Bigtable paper published
- **Feb 2006: Hadoop moves into new Lucene sub-project**
- Feb 2007: First HBase code appeared
- Apr 2007: Yahoo! running Hadoop on 1000-node cluster
- **May 2007: NHN running Hadoop on 20-node**
- Jan 2008: Hadoop made an Apache Top Level Project
- March 2008: HBase 0.1.0 released
- Jul 2008: Zookeeper joined as new Hadoop sub-project

Hadoop Component

- Hadoop Distributed File System - HDFS
 - Java interface
 - Shell interface
 - C - libhdfs
 - HTTP FileSystem
 - Web interface
 - Eclipse plugin
- Hadoop MapReduce
 - Java api
 - Streaming api . via stdin/stdout
 - Pipes C++ api - via sockets

Open platform



Hadoop Users

- Over 50 reference sites
 - <http://wiki.apache.org/hadoop/PoweredBy>
 - Includes Yahoo!, IBM, Google, Facebook

The screenshot displays the Hadoop Source website interface. At the top, the "YAHOO! DEVELOPER NETWORK" logo is visible on the left, and "HADOOP SOURCE" with a blue elephant logo is in the center. Navigation links for "Home", "Sources", "Forums", and "Developers" are present. A search bar is located on the right side of the header.

The main content area is divided into several sections:

- Category:** A list of categories including Example, Core, Utility, Machine Learning (ML), Information Retrieval (IR), Natural Language Processing (NLP), Bioinformatics (Bio), and Unclassified.
- Tags:** A list of tags including WordCount, Example, WordCount, Example, Sort, Example, PageRank, and IR.
- Browse Sources:** A list of code examples, ordered by date/view. Two examples are visible:
 - Random Writer Example** by hadoop on 2008-09-26 16:36:49. Description: "RandomWriter uses map/reduce to just run a distributed job where there is no interaction between the tasks and each task write a large unsorted random binary sequence file of BytesWritable." Tags: RandomWriter, Example. Category: Example. View: 20, Download: 84.
 - Sort Example** by hadoop on 2008-09-26 16:33:06. Description: "Sort is the trivial map/reduce program that does absolutely nothing other than use the framework to fragment and sort the input values." Tags: Sort, Example, Sort, Example, Sort, Example, Sort, Example. Category: Example. View: 6, Download: 100.

Overlaid on the screenshot are several other elements:

- A "Back Doc" search box with the text "Search for chiropractic medicine, books, products, lo".
- A "krugle" logo.
- A "Code Search for Hadoop" search box showing "Code Files 1-10 (out of about 2050 matching files)" and a list of results including "HadoopJob.java", "svn.apache.org", "Hadoop", and "Apache-2.0".
- A "Microsoft" logo with links for "Feedback", "Need Help?", and "Powerlabs".

Related Project

- Sub-project
 - HBase, Zookeeper
- Contributed Project
 - capacity-scheduler, chukwa, data_join, ec2, Eclipse-plugin, failmon, fairscheduler, fuse-dfs, hive, hod, index, streaming, thriftfs
- Others
 - Mahout, Katta, Cascading, Parhely, Happy, SmartFrog, CloudBase, Greenplum
 - Hadoop Source
- Korea hadoop user group: <http://www.hadoop.or.kr>

Why Hadoop

- This is my direct-client opening for a **Hadoop Engineer** - NYC, NY. This is permanent position. The **salary range is 130-190K**. Relocation assistance is provided for this position.
- Responsibilities
 - Develop and support a **secure and flexible large-scale data processing infrastructure for research and development** within the company. As a core member of a small and deeply talented team, you will be responsible across many technical aspects of helping to deliver the results of our R&D as a world-class platform for partners and customers.
- Qualifications
 - Bachelor's Degree in Engineering, Computer Science, or related technical field.
 - **Required: real world experience building data solutions using Hadoop.**
 - Strong design/admin experience with relational database systems, esp. MySQL and/or PostgreSQL.
 - At least 4 years software engineering experience designing and developing modern web-based consumer-facing server solutions in rapid development cycles
 - Expert in Java (C++, Python, a plus) development and debugging on a Linux platform.
 - A deep and powerful need to create useful, readable and accurate documentation as you work.