

A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Data Management

CT051-3-M

Topic 7 – Hadoop

Topic & Structure of Lesson

- What is Hadoop?
- Hadoop Framework
- Hadoop's Architecture
- Hadoop in the Wild
- Data warehouse to Hadoop

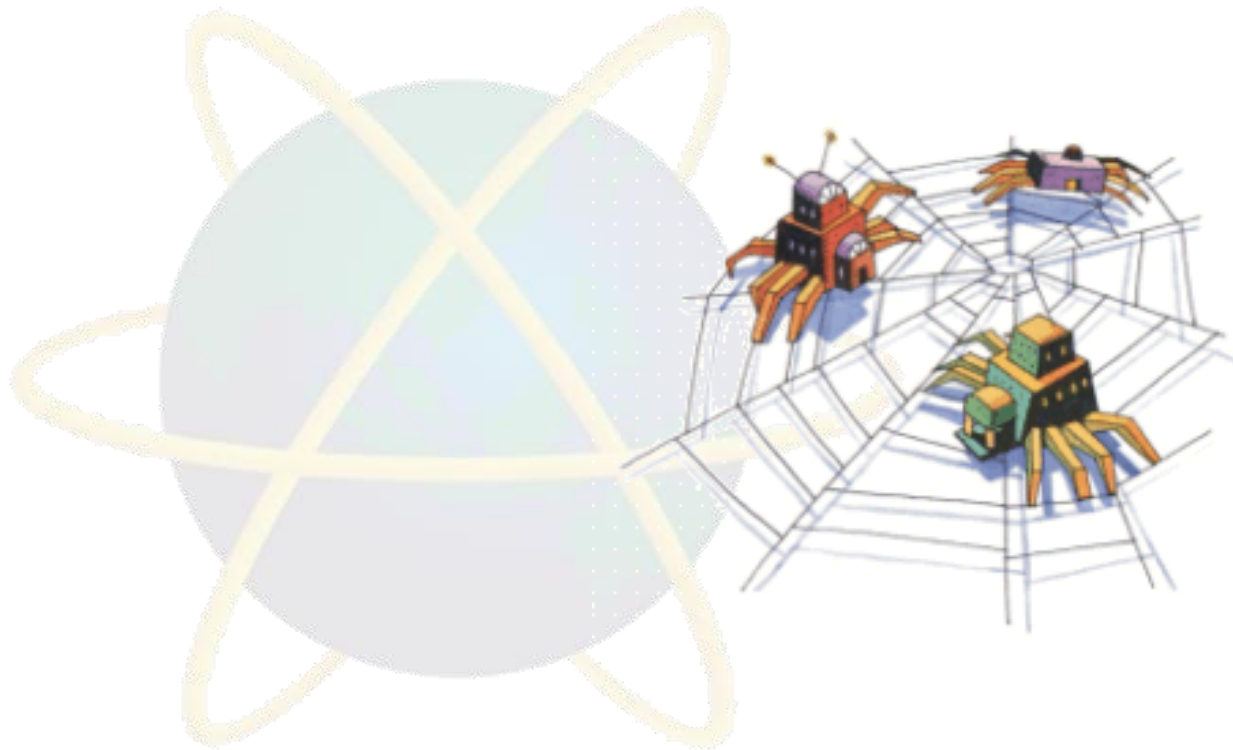
What is



- 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



A.P.U.
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

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 [Abow!](#)

• [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
© 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:

Researching stocks?
Buying a car?
Planning a wedding?
[Check out ExciteSooing Tours.](#)

[Bill Mitchell](#)
[Satire that clicks!](#)

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

- [Arts](#) • [Entertainment](#) • [Money](#) • [Regional](#)
- [Business](#) • [Health](#) • [News & Reference](#) • [Science](#)
- [Computing](#) • [Hobbies](#) • [Personal Pages](#) • [Shopping](#)
- [Education](#) • [Life & Style](#) • [Politics & Law](#) • [Sports](#)

1996

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

Find:

[Enhance your search.](#)

[New Search](#) • [TopNews](#) • [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 LycosTM, Inc. All Rights Reserved.
Lycos is a trademark of Carnegie Mellon University.
[Questions & Comments](#)

1996

HELP WIRED NEWS NOTWIRED WIRED MAGAZINE SUCK.COM

The WRED Search Center

look for:

for more options use [SuperSearch](#)

Date: in the last week

Geotarget: North America (.com)

Include media type:
☐ Image ☐ Audio ☐ Video ☐ Shockwaves

Return Results:
 10 full descriptions

Find:
[Discussion](#)
[People](#)
[Email Addresses](#)

[Sandbox Entertainment](#)
[Shop WIRED Holiday Gift Guide](#)
SOMETHING HAS SURVIVED.
[Find more deals](#)

[Log](#)
[Cybertan Outpost](#)
[Microsoft® Expedia™ Travel](#)
[GSAF](#)

1997

Google search engine



A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION



1998



Google

A screenshot of the 2013 Google search engine interface. It features the modern multi-colored "Google" logo. Below the logo is a large, empty search bar with a microphone icon on the right side. Underneath the search bar are two buttons: "Google Search" and "I'm Feeling Lucky".

2013

Google Origins

2003

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google*



2004

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.



2006

Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber

{fay,jeff,sanjay,wilson,h,kerr,m,b,tushar,fikes,gruber}@google.com

Google, Inc.



Abstract

Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large number of nodes. It is designed to store petabytes of data across thousands of commodity servers. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google File. These applications place very different demands on Bigtable, both in terms of data size (from URLs to

achieved scalability and high performance, but Bigtable provides a different interface than such systems. Bigtable does not support a full relational data model; instead, it provides clients with a simple data model that supports dynamic control over data layout and format, and allows clients to reason about the locality properties of data represented in the underlying storage. Data is indexed using row and column names that can be arbitrary strings. Bigtable also treats data as uninterpreted strings.

Hadoop's Developers



Doug Cutting



Michael J. Cafarella

2005: Doug Cutting and Michael J. Cafarella developed Hadoop to support distribution for the [Nutch](#) search engine project. The project was funded by Yahoo.

2006: Yahoo gave the project to Apache Software Foundation.

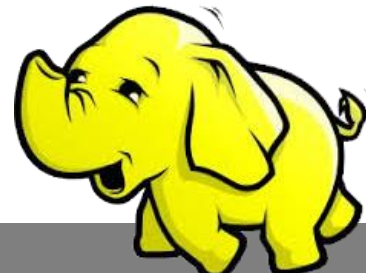


Some Hadoop Milestones

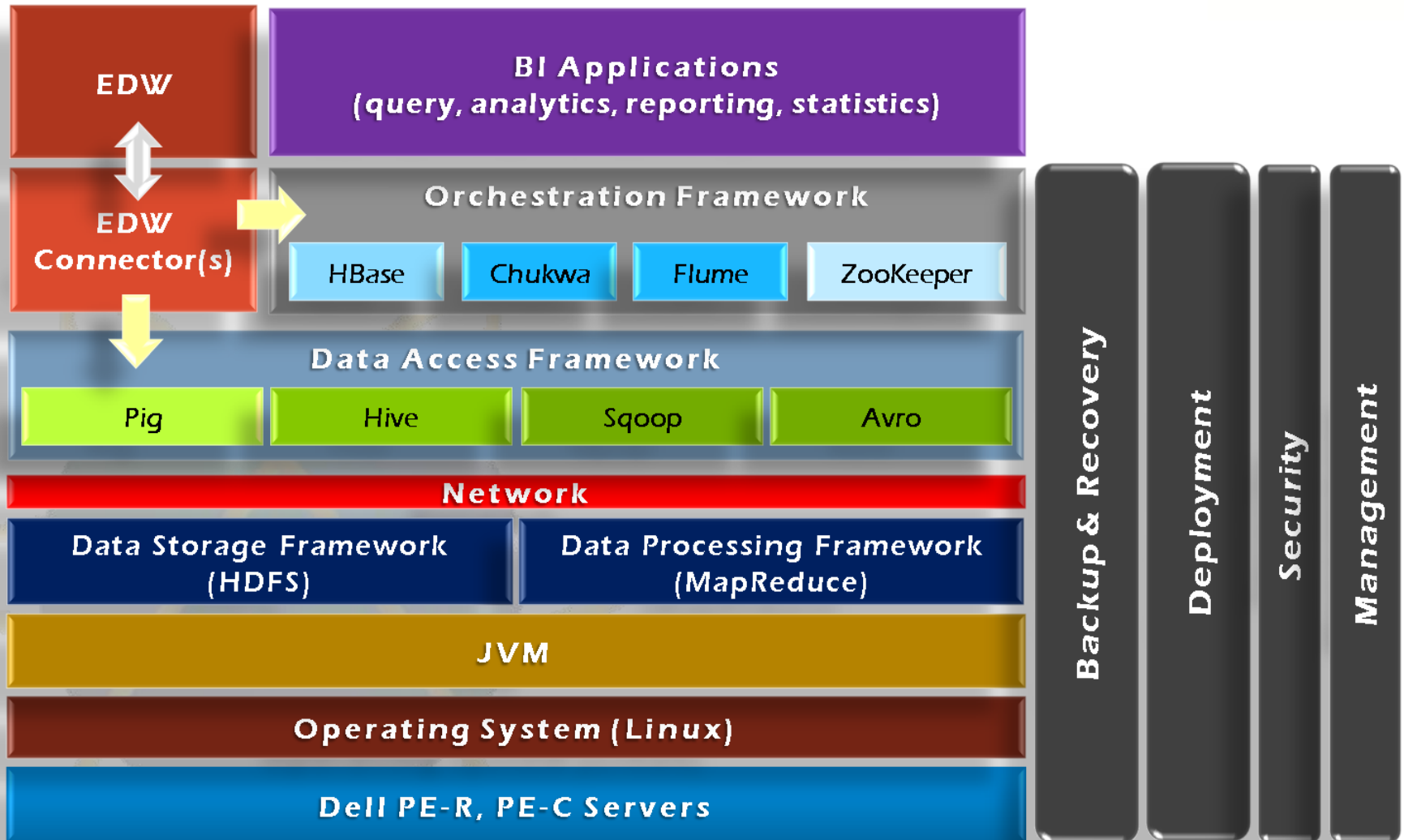
- 2008 - **Hadoop Wins Terabyte Sort Benchmark** (sorted 1 terabyte of data in 209 seconds, compared to previous record of 297 seconds)
- 2009 - Avro and Chukwa became new members of Hadoop Framework family
- 2010 - Hadoop's Hbase, Hive and Pig subprojects completed, adding more computational power to Hadoop framework
- 2011 - ZooKeeper Completed
- 2013 - Hadoop 1.1.2 and Hadoop 2.0.3 alpha.
 - Ambari, Cassandra, Mahout have been added
- 2015 – Hadoop Yarn
- 2017 – Hadoop 3.0 (Alpha)

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



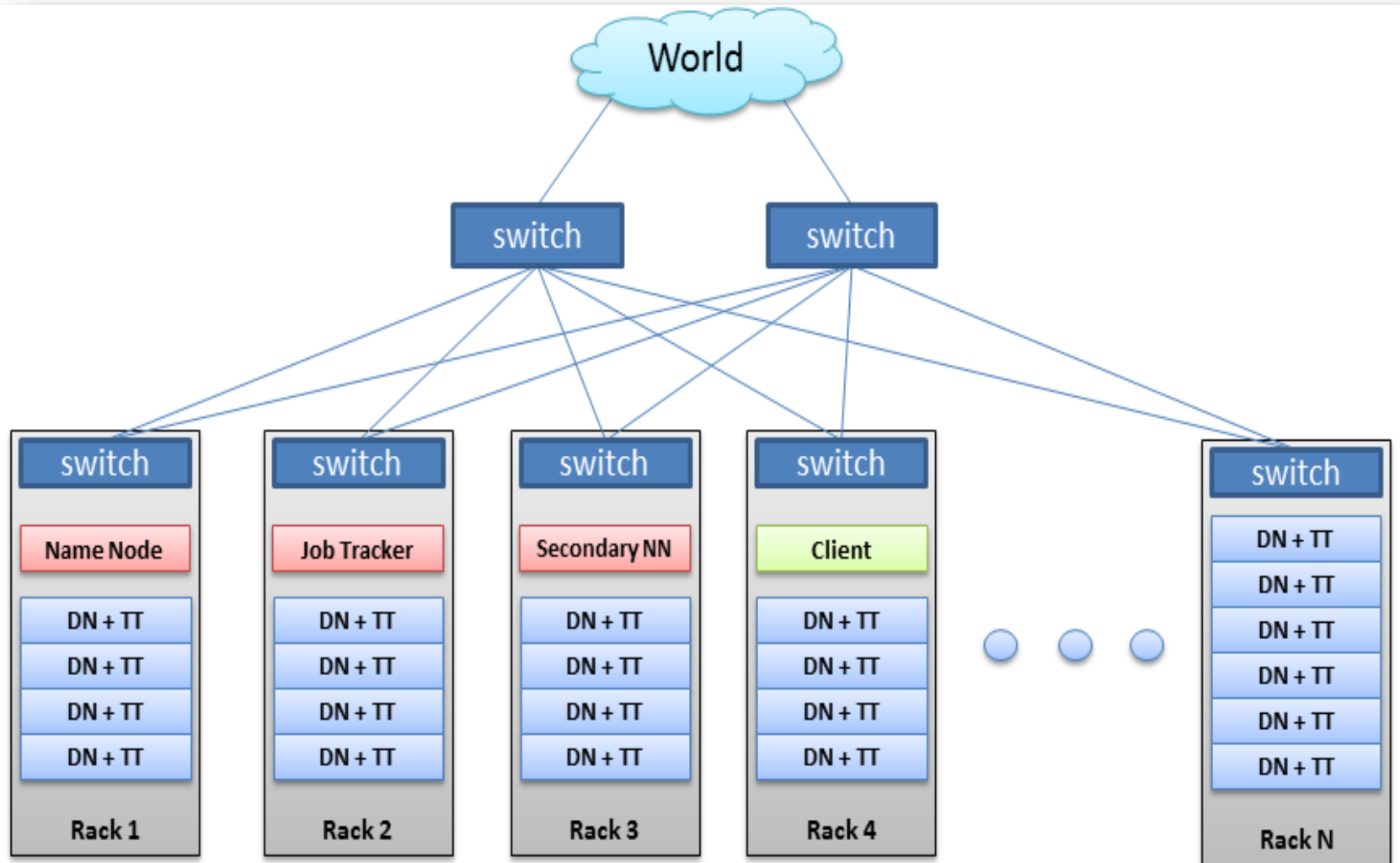
Hadoop's Architecture

- Distributed, with some centralization
- Main nodes of cluster are where most of the computational power and storage of the system lies
- Main nodes run TaskTracker to accept and reply to MapReduce tasks, and also DataNode to store needed blocks closely as possible
- Central control node runs NameNode to keep track of HDFS directories & files, and JobTracker to dispatch compute tasks to TaskTracker
- Written in Java, also supports Python and Ruby

Hadoop Architecture



A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION



Hadoop's Architecture

- Hadoop Distributed Filesystem
- Tailored to needs of MapReduce
- Targeted towards many reads of filestreams
- Writes are more costly
- High degree of data replication (3x by default)
- No need for RAID on normal nodes
- Large blocksize (64MB)
- Location awareness of DataNodes in network

Hadoop's Architecture

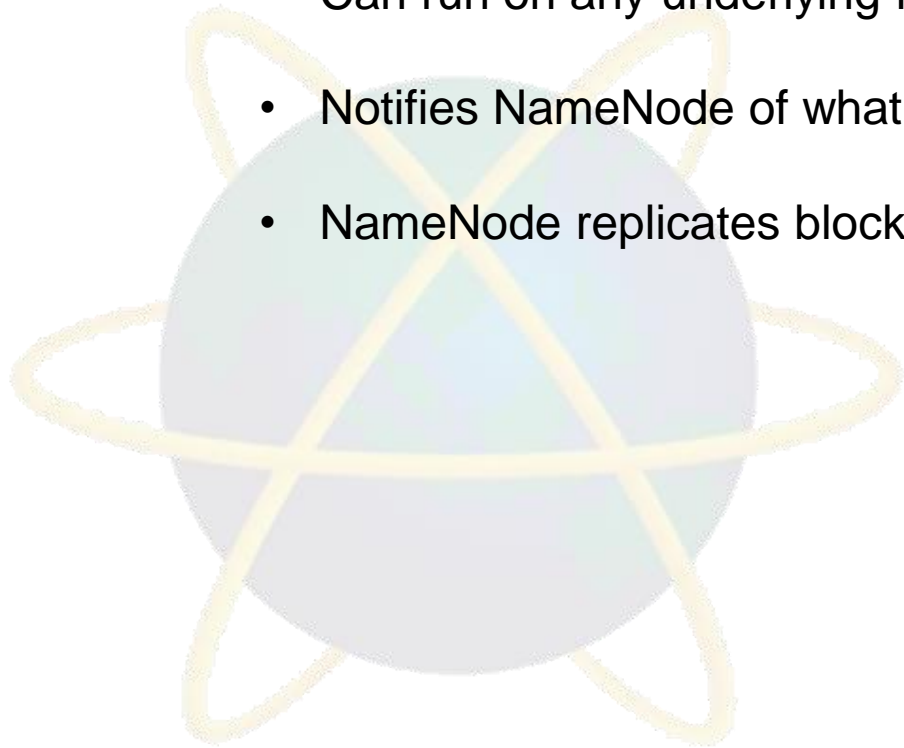
NameNode:

- Stores metadata for the files, like the directory structure of a typical FS.
- The server holding the NameNode instance is quite crucial, as there is only one.
- Transaction log for file deletes/adds, etc. Does not use transactions for whole blocks or file-streams, only metadata.
- Handles creation of more replica blocks when necessary after a DataNode failure

Hadoop's Architecture

DataNode:

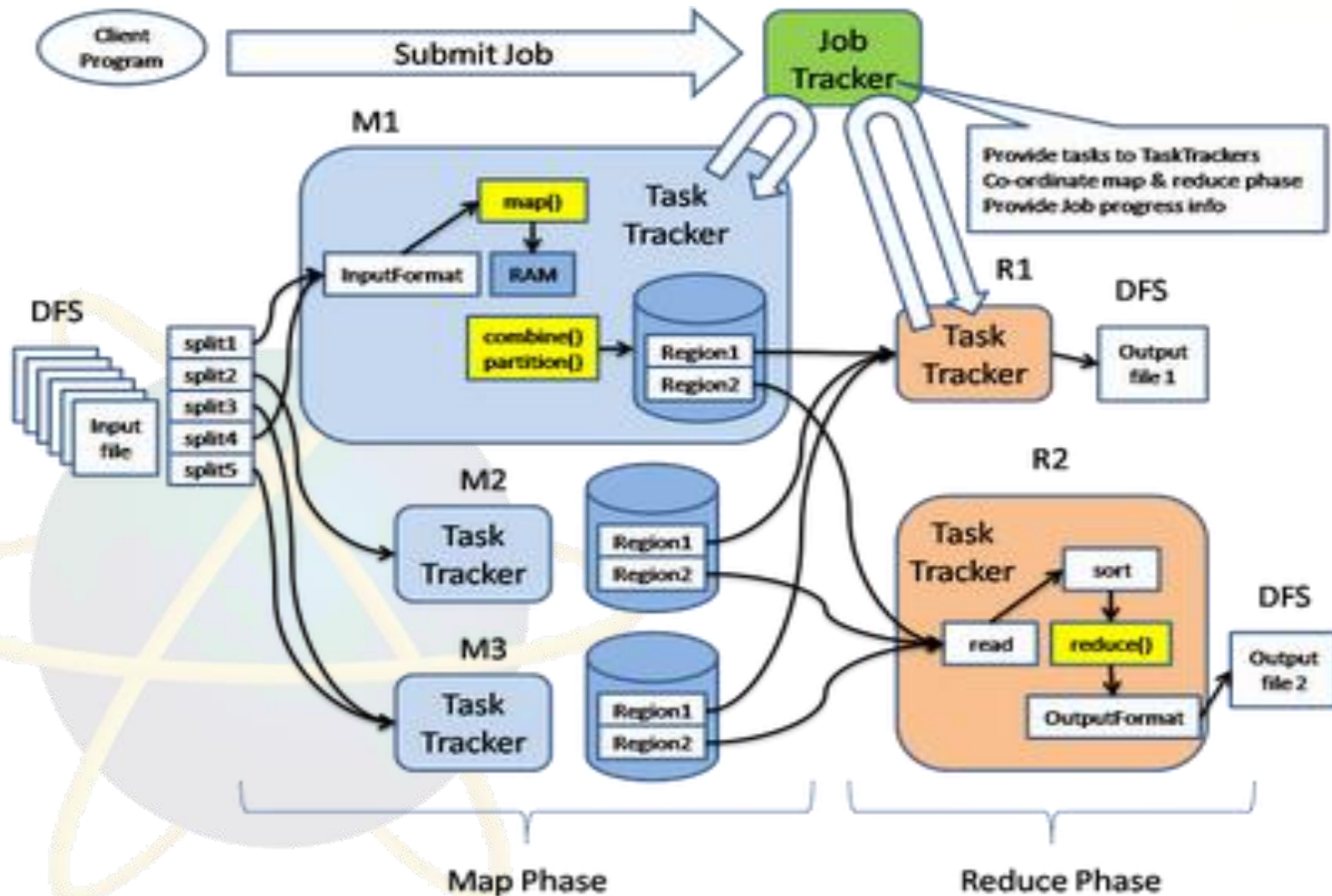
- Stores the actual data in HDFS
- Can run on any underlying filesystem (ext3/4, NTFS, etc)
- Notifies NameNode of what blocks it has
- NameNode replicates blocks 2x in local rack, 1x elsewhere

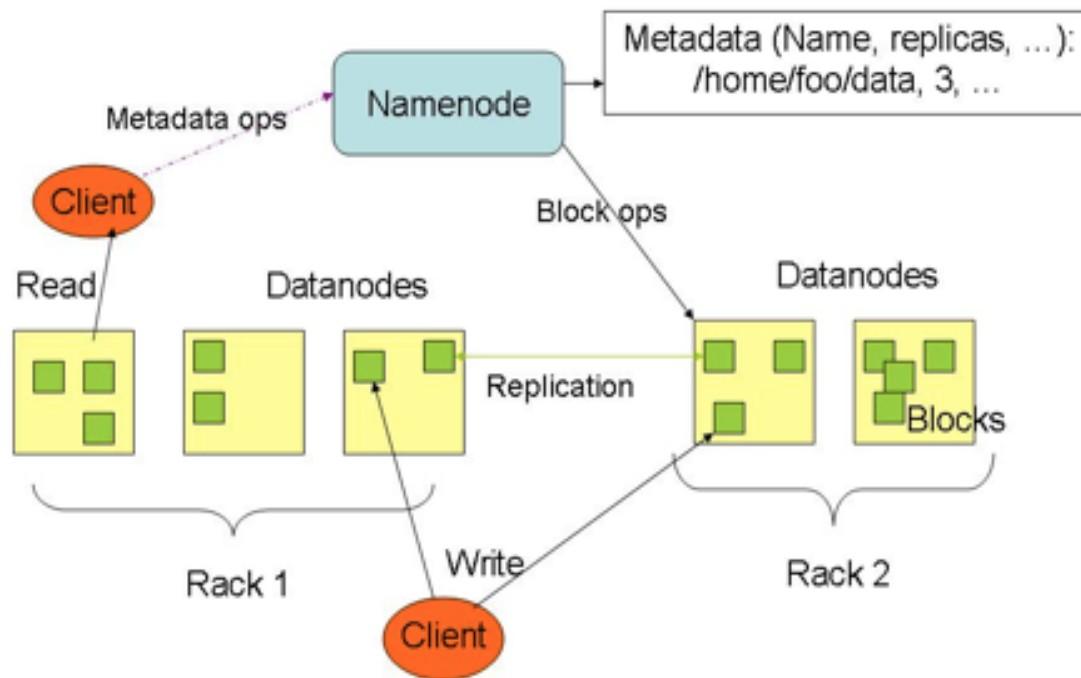


Hadoop's Architecture: MapReduce Engine



A . P . U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

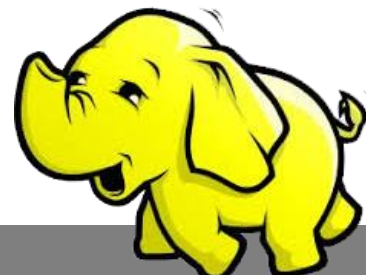




Hadoop's Architecture

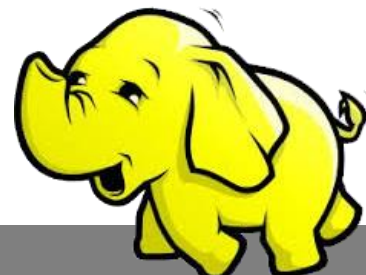
MapReduce Engine:

- JobTracker & TaskTracker
- JobTracker splits up data into smaller tasks("Map") and sends it to the TaskTracker process in each node
- TaskTracker reports back to the JobTracker node and reports on job progress, sends data ("Reduce") or requests new jobs



Hadoop's Architecture

- None of these components are necessarily limited to using HDFS
- Many other distributed file-systems with quite different architectures work
- Many other software packages besides Hadoop's MapReduce platform make use of HDFS



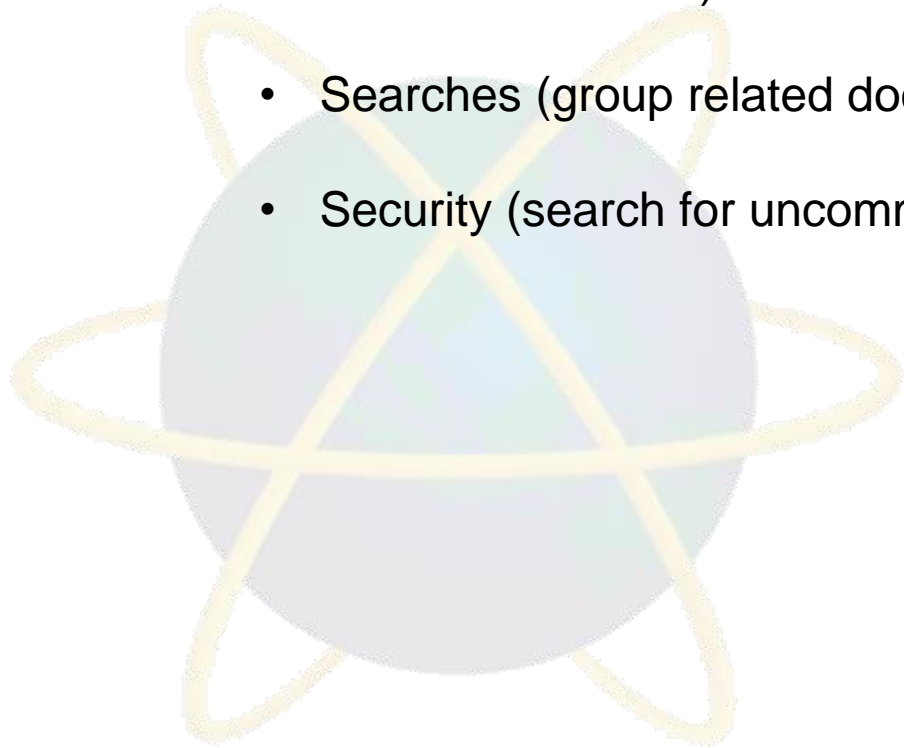
Hadoop in the Wild

- Hadoop is in use at most organizations that handle big data:
 - Yahoo!
 - Facebook
 - Amazon
 - Netflix
 - Etc...
- Some examples of scale:
 - Yahoo!'s Search Webmap runs on 10,000 core Linux cluster and powers Yahoo! Web search
 - FB's Hadoop cluster hosts 100+ PB of data (July, 2012) & growing at ½ PB/day (Nov, 2012)

Hadoop in the Wild

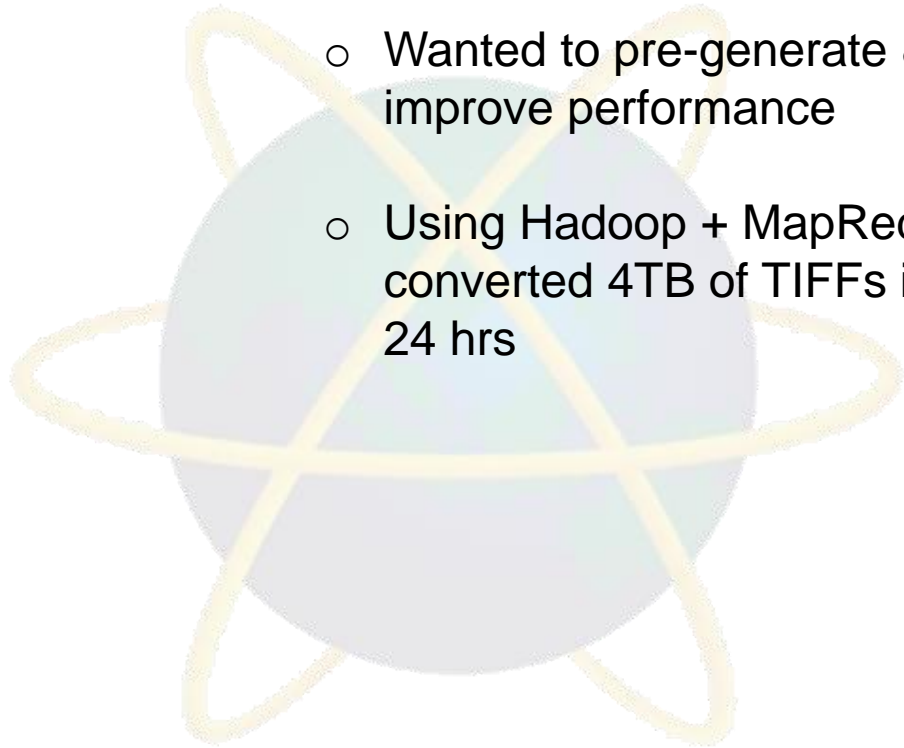
Three main applications of Hadoop:

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



Hadoop in the Wild

- Non-realtime large dataset computing:
 - NY Times was dynamically generating PDFs of articles from 1851-1922
 - Wanted to pre-generate & statically serve articles to improve performance
 - Using Hadoop + MapReduce running on EC2 / S3, converted 4TB of TIFFs into 11 million PDF articles in 24 hrs



Hadoop in the Wild: Facebook

Messages



A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

- Design requirements:
 - Integrate display of email, SMS and chat messages between pairs and groups of users
 - Strong control over who users receive messages from
 - Suited for production use between 500 million people immediately after launch
 - Stringent latency & uptime requirements



Hadoop in the Wild



A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

- System requirements
 - High write throughput
 - Cheap, elastic storage
 - Low latency
 - High consistency (within a single data center good enough)
 - Disk-efficient sequential and random read performance



Hadoop in the Wild

- Classic alternatives
 - These requirements typically met using large MySQL cluster & caching tiers using Memcached
 - Content on HDFS could be loaded into MySQL or Memcached if needed by web tier
- Problems with previous solutions
 - MySQL has low random write throughput... BIG problem for messaging!
 - Difficult to scale MySQL clusters rapidly while maintaining performance
 - MySQL clusters have high management overhead, require more expensive hardware

Hadoop in the Wild

- Facebook's solution
 - Hadoop + HBase as foundations
 - Improve & adapt HDFS and HBase to scale to FB's workload and operational considerations
 - Major concern was availability: NameNode is SPOF & failover times are at least 20 minutes
 - Proprietary "AvatarNode": eliminates SPOF, makes HDFS safe to deploy even with 24/7 uptime requirement
 - Performance improvements for realtime workload: RPC timeout. Rather fail fast and try a different DataNode

Hadoop Highlights

- Distributed File System
- Fault Tolerance
- Open Data Format
- Flexible Schema
- Queryable Database

Why use Hadoop?

- Need to process Multi Petabyte Datasets
- Data may not have strict schema
- Expensive to build reliability in each application
- Nodes fails everyday
- Need common infrastructure
- Very Large Distributed File System
- Assumes Commodity Hardware
- Optimized for Batch Processing
- Runs on heterogeneous OS

DataNode

- A Block Server
 - Stores data in local file system
 - Stores meta-data of a block - checksum
 - Serves data and meta-data to clients
- Block Report
 - Periodically sends a report of all existing blocks to NameNode
- Facilitate Pipelining of Data
 - Forwards data to other specified DataNodes

Block Placement

- Replication Strategy
 - One replica on local node
 - Second replica on a remote rack
 - Third replica on same remote rack
 - Additional replicas are randomly placed
- Clients read from nearest replica

Data Correctness

- Use Checksums to validate data – CRC32
- File Creation
 - Client computes checksum per 512 byte
 - DataNode stores the checksum
- File Access
 - Client retrieves the data and checksum from DataNode
 - If validation fails, client tries other replicas

Data Pipelining

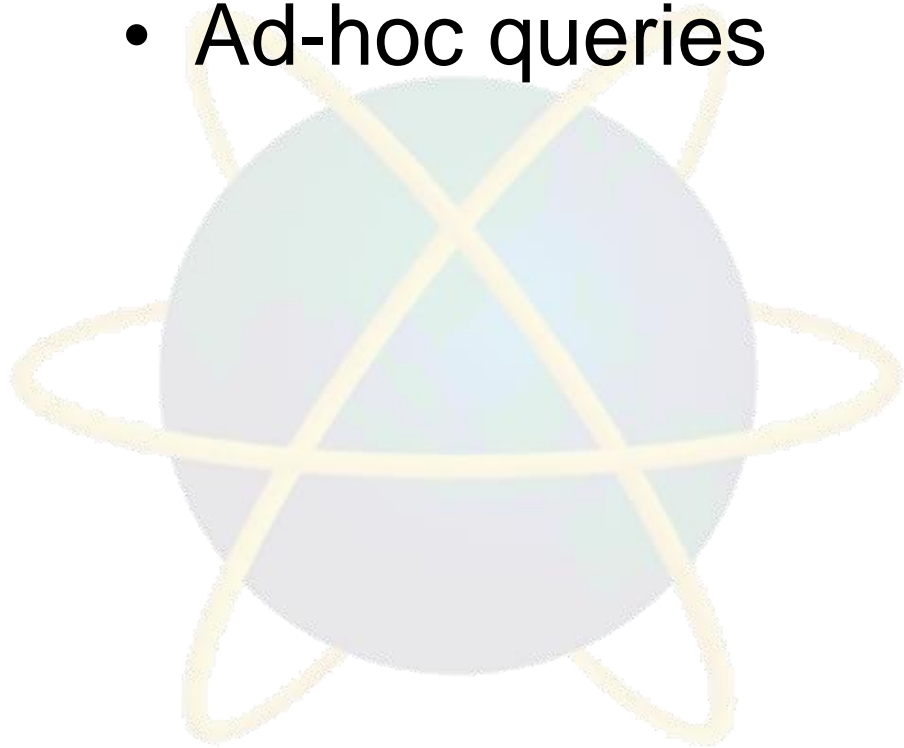
- Client retrieves a list of DataNodes on which to place replicas of a block
- Client writes block to the first DataNode
- The first DataNode forwards the data to the next DataNode in the Pipeline
- When all replicas are written, the client moves on to write the next block in file

Hadoop MapReduce

- MapReduce programming model
 - Framework for distributed processing of large data sets
 - Pluggable user code runs in generic framework
- Common design pattern in data processing
 - `cat * | grep | sort | uniq -c | cat > file`
 - `input | map | shuffle | reduce | output`

MapReduce Usage

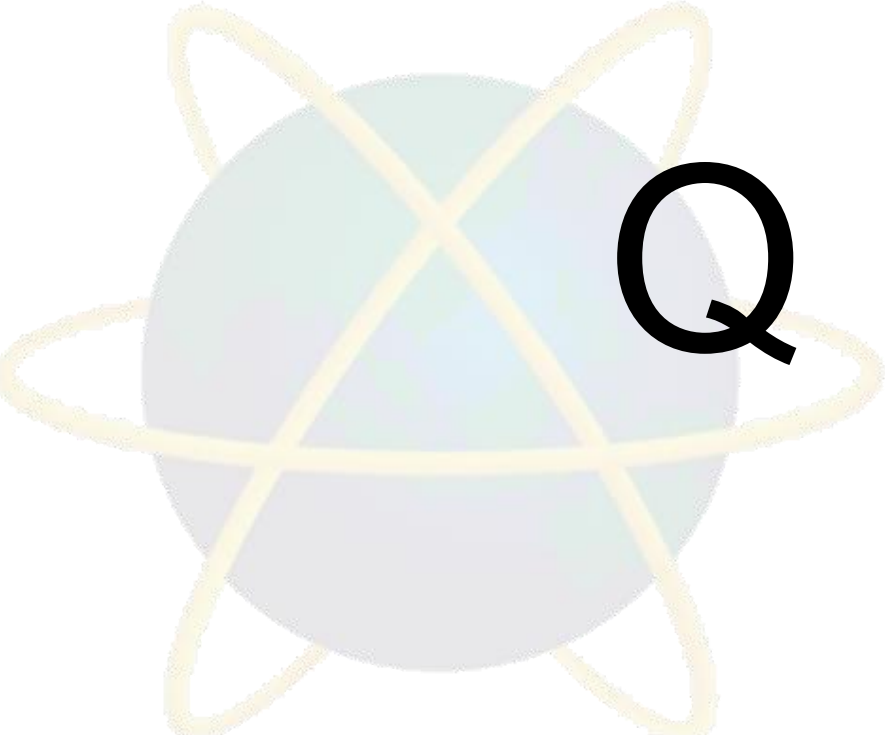
- Log processing
- Web search indexing
- Ad-hoc queries



Offloading The Data Warehouse to Hadoop



Question & Answer Session



Q & A