

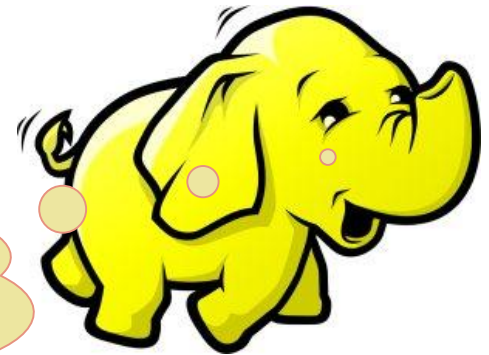
Introduction To Hadoop

- Hadoop Overview
- Hadoop History
- Characteristics of Hadoop
- What does Hadoop Offers
- Why Hadoop?
- Hadoop Ecosystem
- Use Cases

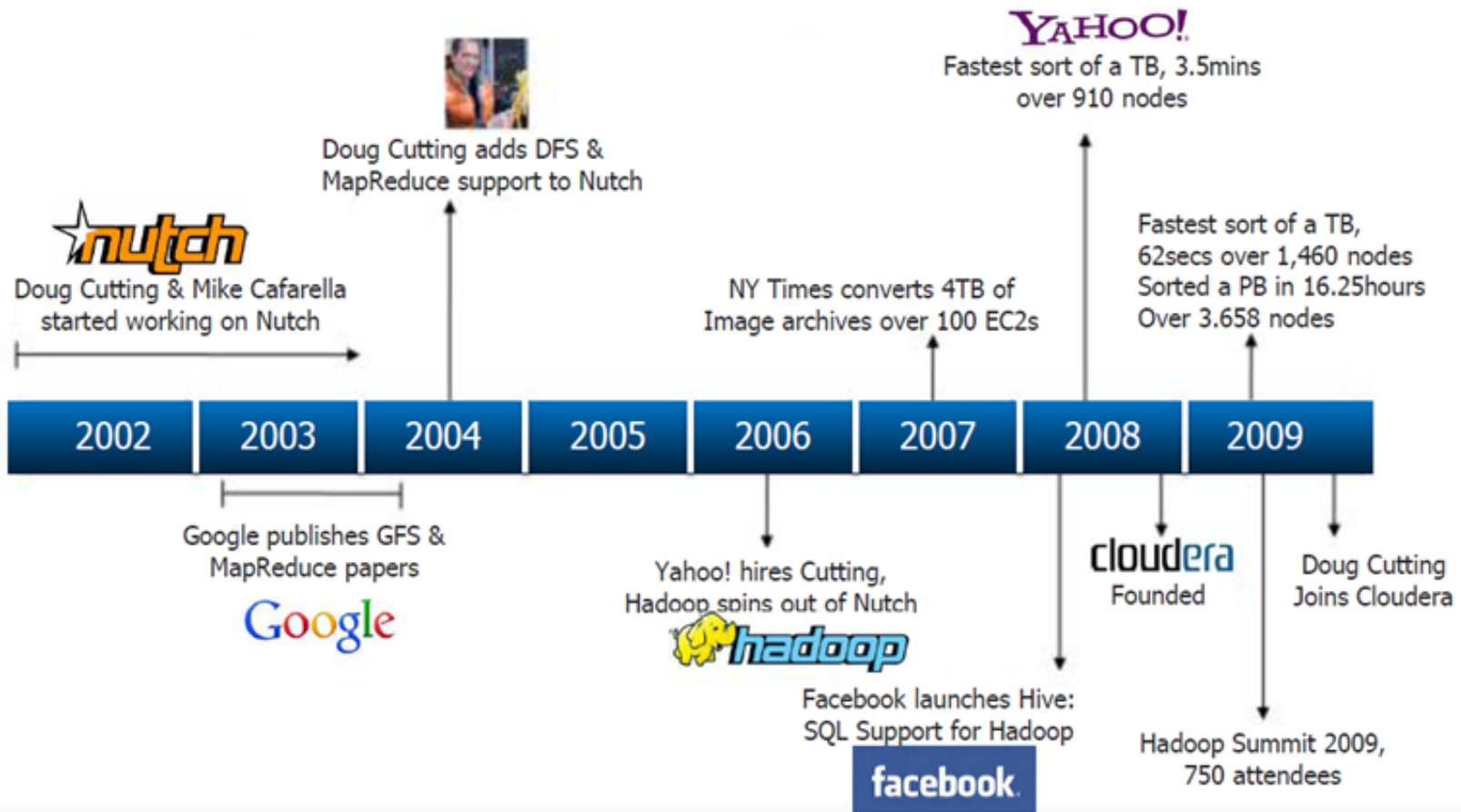
Hadoop Overview

- An Open Source software framework for processing vast amounts of data in parallel on large clusters of commodity hardware (potentially scaling to thousands of nodes) in a reliable, fault-tolerant manner
- Top level Apache project was developed by Doug Cutting
- Inspired from “Google's MapReduce” algorithm for running distributed applications that process large amounts of data
- Yahoo is the largest contributor to the project and uses it extensively

Hadoop was named after
Doug Cutting son's toy
elephant



Hadoop History



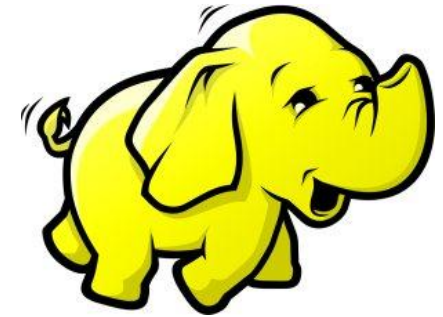
Hadoop 2.0 was released in 2012. All the upcoming pages and PDFs Information would be with respect to Hadoop 2.0 architecture

- Robust: Can handle hardware failures as data is stored in multiple nodes
- Scalable: Hadoop scales linearly to handle larger data by adding more nodes to the cluster
- Simple: Hadoop allows users to quickly write efficient parallel code
- Emphasis on high throughput as opposed to low latency
- Hadoop focuses on moving code rather than data thereby increasing the overall throughput
- Portable: HDFS(Hadoop Distributed File System) is written in Java and hence it is easily portable in Hadoop framework
- Handles data in key/value format instead of using relational tables

Hadoop and MapReduce offer



Scalability and a high degree of fault tolerance



Ability to quickly analyze massive collections of records without forcing data to first be modeled, cleansed, and loaded



**Easy to use
programming
paradigm for writing
and executing
analysis programs
that scale to 1000s of
nodes and PBs of data**



Low, up front software and hardware costs

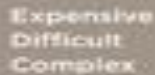


Think data warehousing for Big Data

1

VS.

- Commodity hardware
- Open Source software
- No data models required
- Any data types



VS.

Cheap
Simple
Easy.



2

3



Hadoop



Hadoop

Proprietary OS
Database
Storage Area Network

VS.

Hadoop

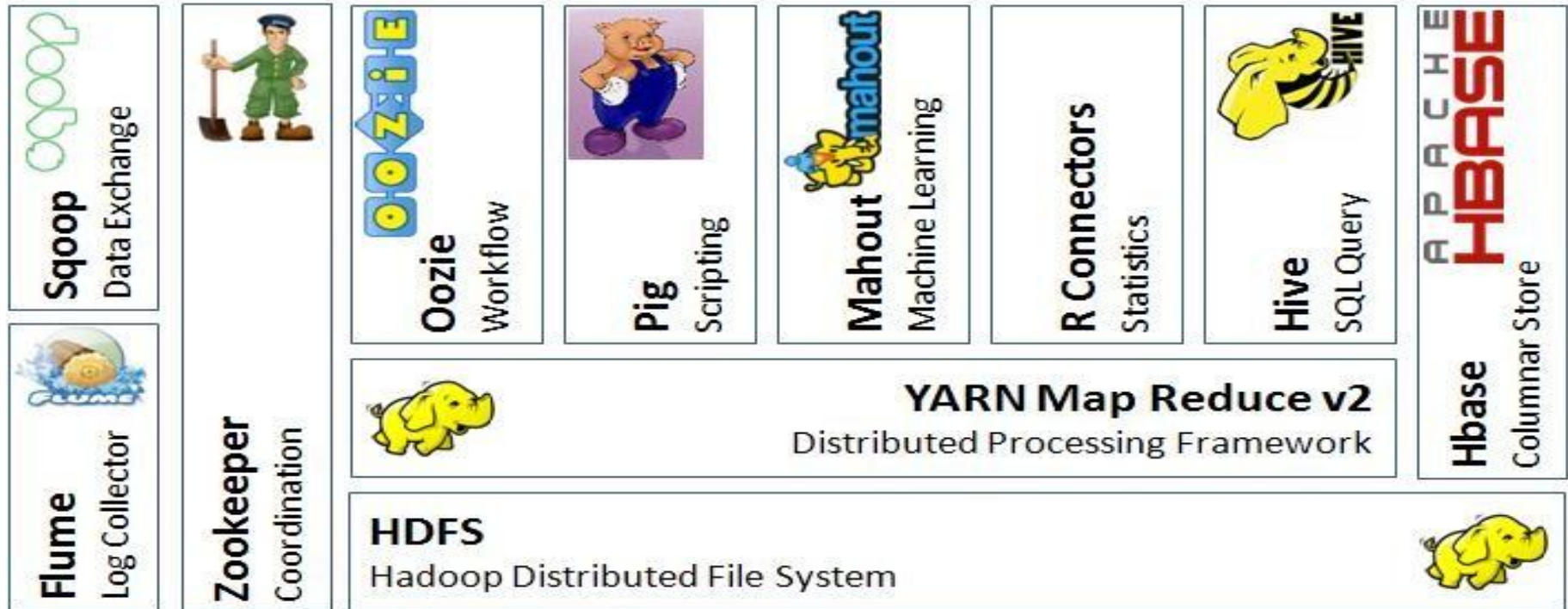


4

Hadoop Ecosystem



Apache Hadoop Ecosystem



Retail

- Retail Giants like WalMart uses Bill information collected from their various stores to find the products that are in great demand and also the products that is always bought with another (eg: Bread and Jam)
- So these data can be used to stock these products together to ensure maximum visibility for the other product.

Email


- Hadoop can be effectively used to filter spam mails. A sample set of spam mails are examined and the frequently occurring words are found out.
- The commonly occurring words in every mail that is sent is checked against the “spam words” and is decided whether it is a spam or not.

Mobile Service Provider

- Every call that is made by a customer is logged.
- The Service Provider can obtain valuable information from these log files.
- The offer that is most preferred by the customers and the type of offer which each demo graph uses can be found out.

- Hadoop In Action by Chuck Lam
- Hadoop: The Definitive Guide by O'Reilly
- <http://www.lecturemaker.com/2011/02/rhipe/>
- <http://datameer.com/solutions/why-hadoop.html>
- www.trend.org/event/2010contest/topic2.ppt
- “*Apache Hadoop Goes Realtime at Facebook*”
RealtimeHadoopSigmod2011.pdf

Thank You



TATA

Promise what we deliver.
Deliver what we promise. That's
certainty

Critical situations. Ruthless competition. Unforgiving customers. Thankfully you can be absolutely sure of your IT solutions with Tata Consultancy Services (TCS). As one of the world's fastest growing technology and business solutions providers, TCS has built a reputation of delivery excellence based on world class IT solutions that are on time, within budget and consistently deliver superior quality. So it comes as no surprise that we pioneered the concept of the Global Network Delivery Model, Developed Innovation Labs and Solution Accelerators. Achieving a level of delivery excellence that provides greater value to our customers and is the industry benchmark. Enabling our clients to experience certainty.

TATA CONSULTANCY SERVICES
Experience certainty.
IT Services • Business Solutions • Outsourcing

To learn how your business can experience certainty, visit www.tcs.com