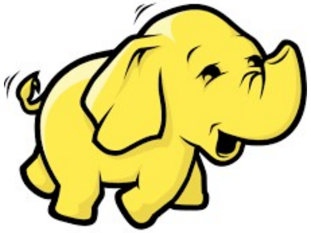
Hadoop - History and Evolution

Tushar B. Kute,
http://tusharkute.com



Hadoop

- Hadoop is an open source framework overseen by Apache Software Foundation which is written in Java for storing and processing of huge datasets with the cluster of commodity hardware.
- There are mainly two problems with the big data. First one is to store such a huge amount of data and the second one is to process that stored data.
- The traditional approach like RDBMS is not sufficient due to the heterogeneity of the data.

Hadoop

- So Hadoop comes as the solution to the problem of big data i.e. storing and processing the big data with some extra capabilities.
- There are mainly two components of Hadoop which are Hadoop Distributed File System (HDFS) and Yet Another Resource Negotiator(YARN).

Hadoop - History

- Hadoop was started with Doug Cutting and Mike Cafarella in the year 2002 when they both started to work on Apache Nutch project.
- Apache Nutch project was the process of building a search engine system that can index 1 billion pages.
- After a lot of research on Nutch, they concluded that such a system will cost around half a million dollars in hardware, and along with a monthly running cost of $30, 000 approximately, which is very expensive.

Hadoop - History

Hadoop - History

- So, they realized that their project architecture will not be capable enough to the workaround with billions of pages on the web.
- So they were looking for a feasible solution which can reduce the implementation cost as well as the problem of storing and processing of large datasets.

Hadoop - History

- In 2003, they came across a paper that described the architecture of Google's distributed file system, called GFS (Google File System) which was published by Google, for storing the large data sets.
- Now they realize that this paper can solve their problem of storing very large files which were being generated because of web crawling and indexing processes.
- But this paper was just the half solution to their problem.

Hadoop - History

- In 2004, Google published one more paper on the technique MapReduce, which was the solution of processing those large datasets.
- Now this paper was another half solution for Doug Cutting and Mike Cafarella for their Nutch project. These both techniques (GFS & MapReduce) were just on white paper at Google.
- Google didn’t implement these two techniques.

Hadoop - History

- Doug Cutting knew from his work on Apache Lucene (It is a free and open-source information retrieval software library, originally written in Java by Doug Cutting in 1999) that open-source is a great way to spread the technology to more people.
- So, together with Mike Cafarella, he started implementing Google’s techniques (GFS & MapReduce) as open-source in the Apache Nutch project.

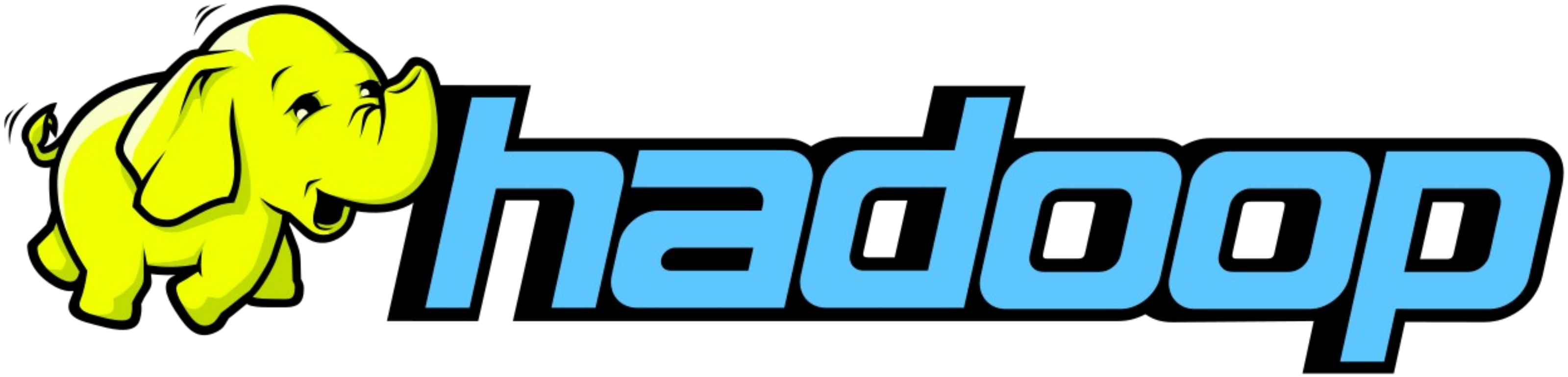
Hadoop - History

- In 2005, Cutting found that Nutch is limited to only 20-to-40 node clusters. He soon realized two problems:
  - (a) Nutch wouldn't achieve its potential until it ran reliably on the larger clusters
  - (b) And that was looking impossible with just two people (Doug Cutting & Mike Cafarella).
- The engineering task in Nutch project was much bigger than he realized. So he started to find a job with a company who is interested in investing in their efforts.
- And he found Yahoo!.Yahoo had a large team of engineers that was eager to work on this there project.

Hadoop - History

- So in 2006, Doug Cutting joined Yahoo along with Nutch project.
- He wanted to provide the world with an open-source, reliable, scalable computing framework, with the help of Yahoo.
- So at Yahoo first, he separates the distributed computing parts from Nutch and formed a new project Hadoop (He gave name Hadoop it was the name of a yellow toy elephant which was owned by the Doug Cutting’s son. and it was easy to pronounce and was the unique word.)
- Now he wanted to make Hadoop in such a way that it can work well on thousands of nodes.
- So with GFS and MapReduce, he started to work on Hadoop.

Hadoop - Logo



Hadoop - History

- In March 2006, Owen O'Malley was the first committer to add to the Hadoop project; Hadoop 0.1.0 was released in April 2006.
- It continues to evolve through contributions that are being made to the project.
- The very first design document for the Hadoop Distributed File System was written by Dhruba Borthakur in 2007.

Hadoop - History

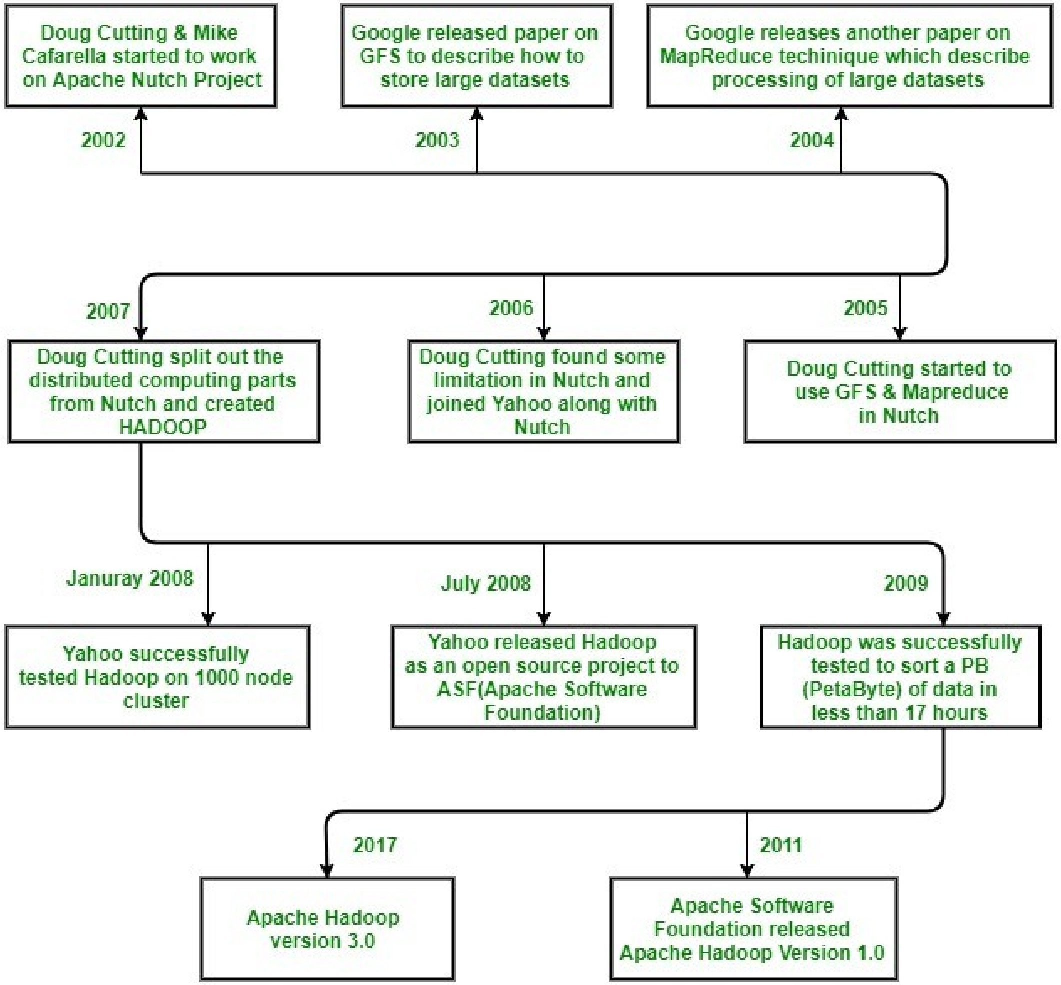
- In 2007, Yahoo successfully tested Hadoop on a 1000 node cluster and start using it.
- In January of 2008, Yahoo released Hadoop as an open source project to ASF(Apache Software Foundation).
- And in July of 2008, Apache Software Foundation successfully tested a 4000 node cluster with Hadoop.

Hadoop - Nodes at Yahoo

Hadoop - History

- In 2009, Hadoop was successfully tested to sort a PB (PetaByte) of data in less than 17 hours for handling billions of searches and indexing millions of web pages.
- And Doug Cutting left the Yahoo and joined Cloudera to fulfill the challenge of spreading Hadoop to other industries.
- In December of 2011, Apache Software Foundation released Apache Hadoop version 1.0.
- And later in Aug 2013, Version 2.0.6 was available.

Hadoop - History



2002

Doug Cutting & Mike Cafarella started to work on Apache Nutch Project

2003

Google released paper on GFS to describe how to store large datasets

2004

Google releases another paper on MapReduce techinique which describe processing of large datasets

2005

Doug Cutting started to use GFS & Mapreduce in Nutch

2006

Doug Cutting found some limitation in Nutch and joined Yahoo along with Nutch

2007

Doug Cutting split out the distributed computing parts from Nutch and created HADOOP

Januray 2008

Yahoo successfully tested Hadoop on 1000 node cluster

July 2008

Yahoo released Hadoop as an open source project to ASF(Apache Software Foundation)

2009

Hadoop was successfully tested to sort a PB (PetaByte) of data in less than 17 hours

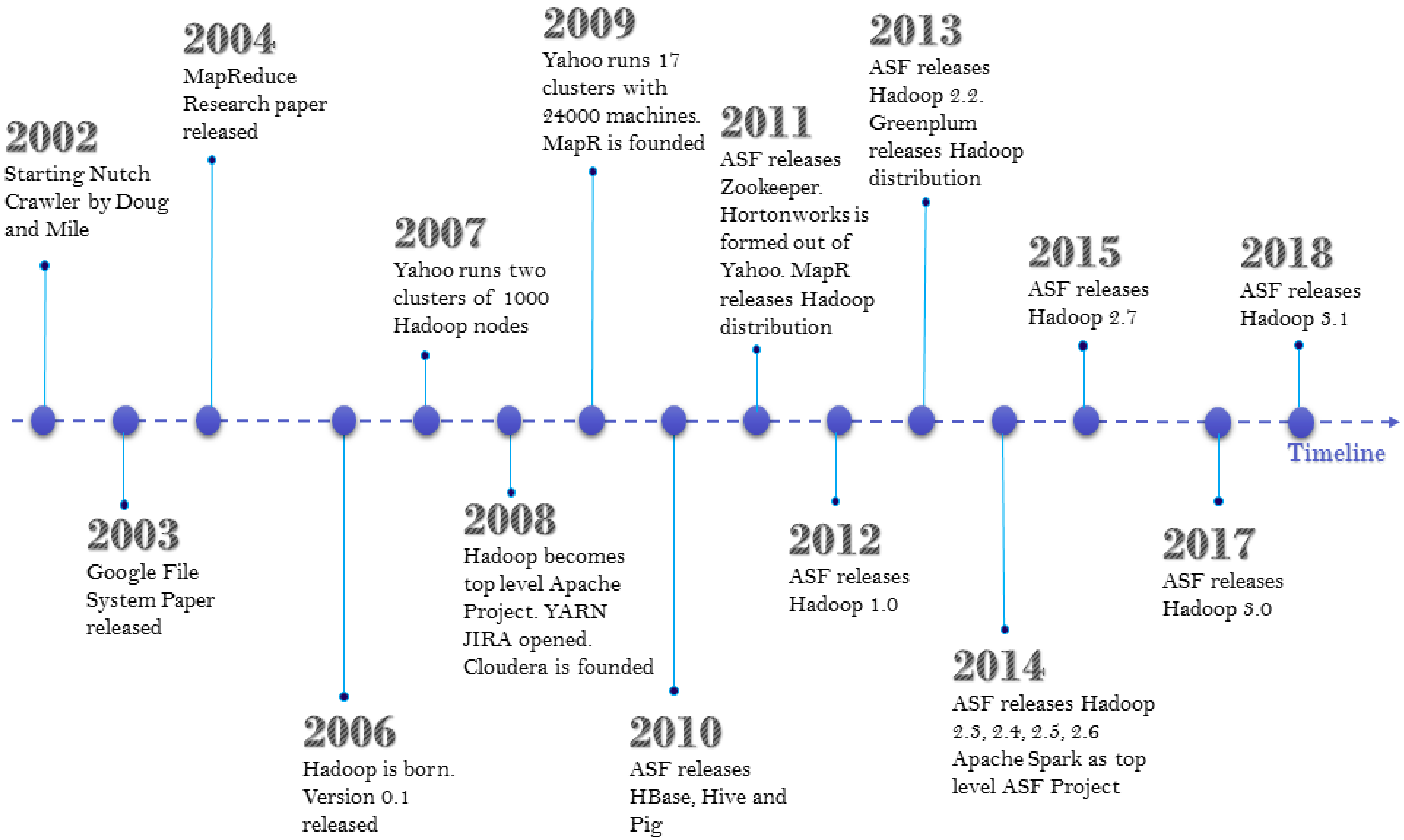
2011

Apache Software Foundation released Apache Hadoop Version 1.0

2017

Apache Hadoop version 3.0

Hadoop - Summary



2002
Starting Nutch Crawler by Doug and Mile

2003
Google File System Paper released

2004
MapReduce Research paper released

2006
Hadoop is born. Version 0.1 released

2007
Yahoo runs two clusters of 1000 Hadoop nodes

2008
Hadoop becomes top level Apache Project. YARN JIRA opened. Cloudera is founded

2009
Yahoo runs 17 clusters with 24000 machines. MapR is founded

2010
ASF releases HBase, Hive and Pig

2011
ASF releases Zookeeper. Hortonworks is formed out of Yahoo. MapR releases Hadoop distribution

2012
ASF releases Hadoop 1.0

2013
ASF releases Hadoop 2.2. Greenplum releases Hadoop distribution

2014
ASF releases Hadoop 2.3, 2.4, 2.5, 2.6 Apache Spark as top level ASF Project

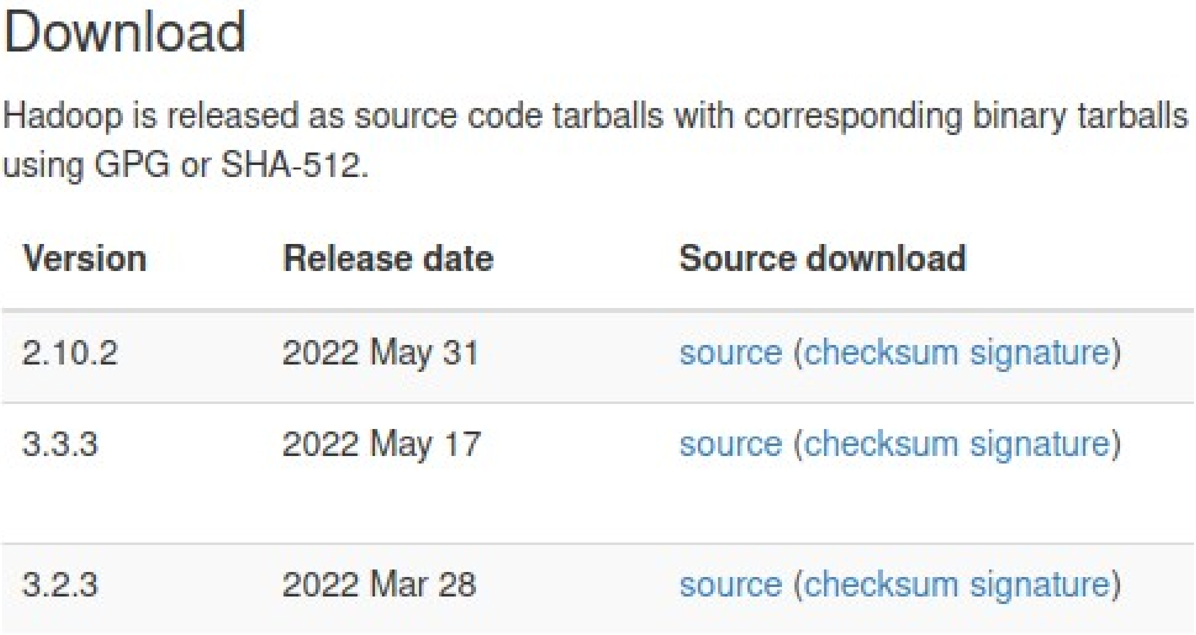
2015
ASF releases Hadoop 2.7

2017
ASF releases Hadoop 3.0

2018
ASF releases Hadoop 3.1

Timeline

Current Versions



Download

Hadoop is released as source code tarballs with corresponding binary tarballs using GPG or SHA-512.

Version	Release date	Source download
2.10.2	2022 May 31	source (checksum signature)
3.3.3	2022 May 17	source (checksum signature)
3.2.3	2022 Mar 28	source (checksum signature)

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com