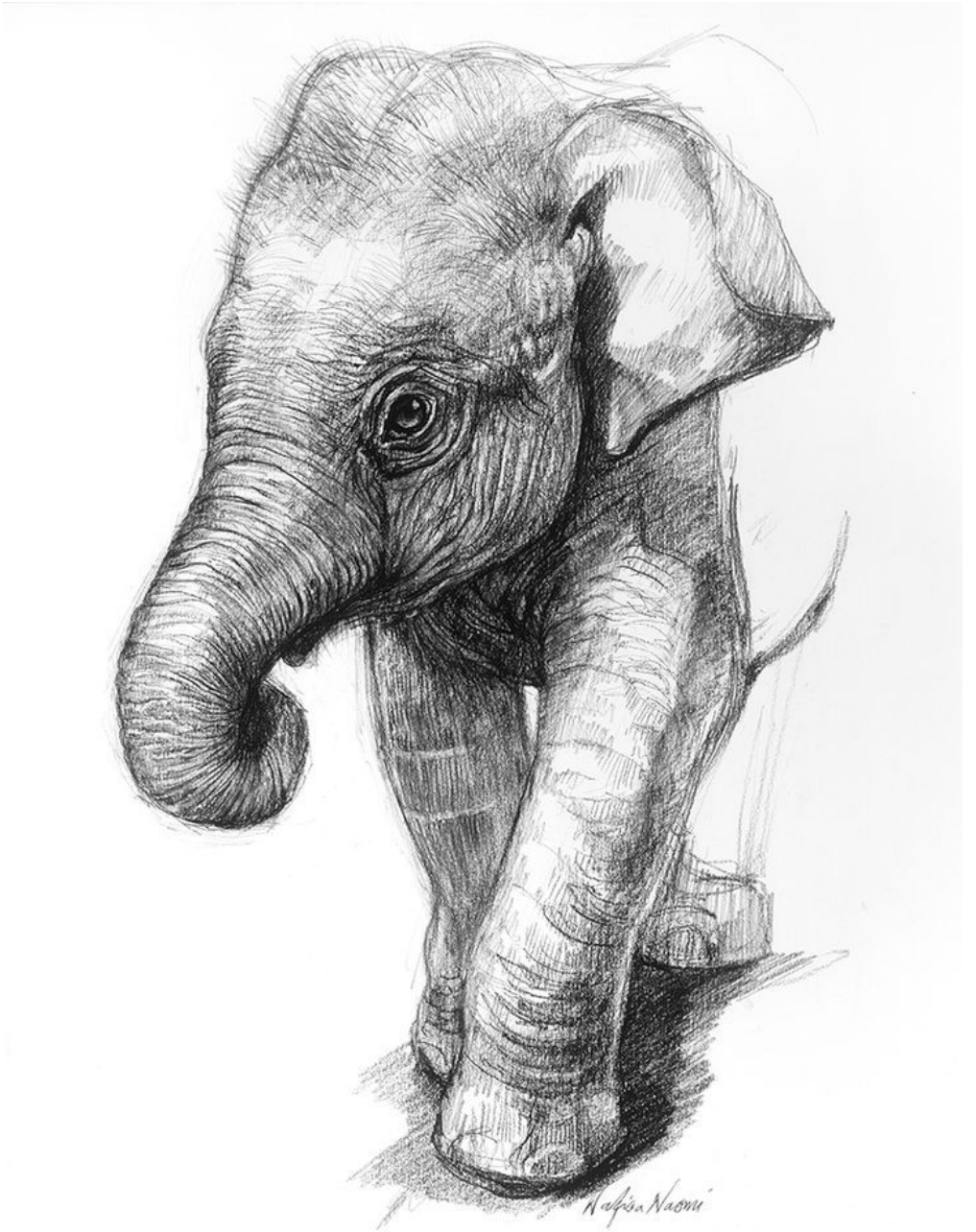
Apache Hadoop HDFS

Presentation

Patric Weber



Agenda

1. Apache Hadoop
2. Usage
3. Motivation & Goal
4. Architecture
5. Hadoop Distributed File System (HDFS)
6. Conclusions
7. Bibliography

Apache Hadoop

- ***"Apache Hadoop is a framework written in Java for scalable, distributed software. It is based on the MapReduce algorithm from Google Inc. as well as suggestions from the Google file system and makes it possible to perform intensive computing processes with large amounts of data in the petabyte range on computer clusters to carry out a test.“***

- *I'm sure you're wondering what the word "Hadoop" means. The first pronounced word of the child by the Lucene inventor, Doug Cutting, was "Hadoop."*

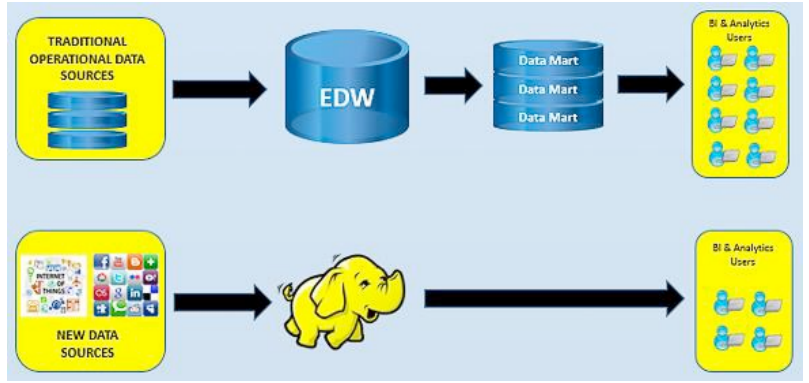
Usage

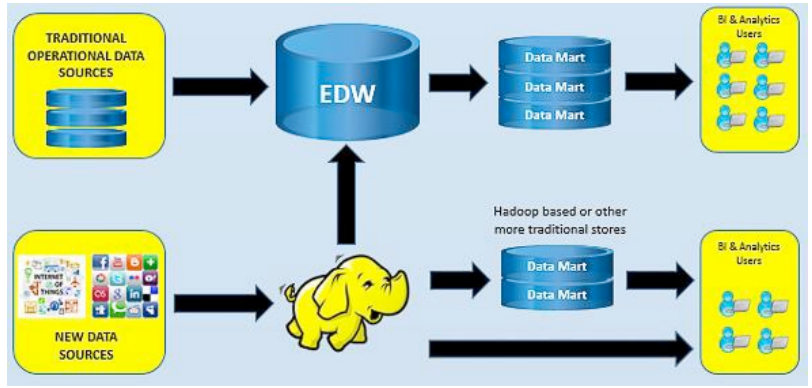
The Three Hadoop Applications in Data Warehousing

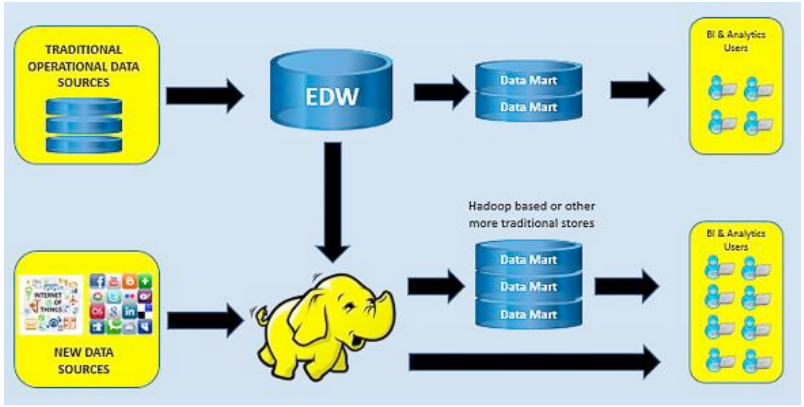
1. **Hadoop as a new data storage device**

2. **Hadoop as a data platform and additional Input for the Enterprise Data Warehouse.**

3. **Hadoop as a data platform and Basis for BI and Analytics.**







Motivation

Apache Hadoop is open-source!

1. **The automatic collection** of data (log files, share prices, social security, etc.)Networks, ...) generates huge amounts of data
2. **Storage** capacities of hard disks are strong in the last decades risen
3. **Average access** times and transfer rates have proved to be comparatively low bettered
4. **The time to read** out a complete hard disk is getting longer and longer. 2000: 40 GB, 32 MB/s, 9.5 ms → 21 minutes 2009: 1000 GB, 125 MB/s, 8.5 ms → 136 minutes
5. **Multiple independent computers** with their own disk memories become clusters connected (distributed storage of data)
6. **Parallel processing** of large amounts of data possible, e.g. Petabytes

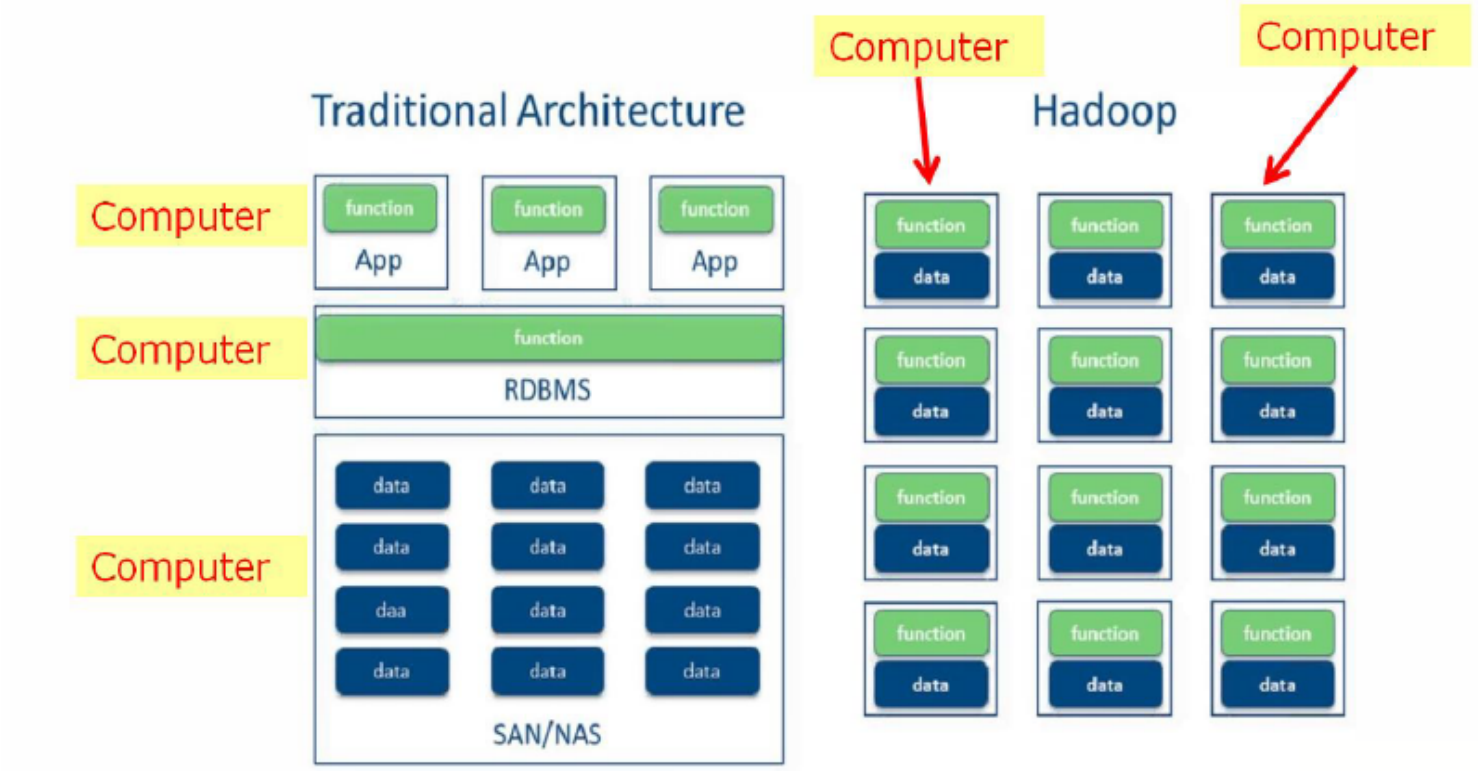
What can Hadoop do?

Master Big Data

- Hadoop can handle intensive computing processes with large amounts of data (Big Data, Petabyte range) on computer clusters. A computer cluster is a "collection" of several computers.
- Therefore, large amounts of data (including unstructured data) can be distributed securely and fault-tolerant on up to several thousand servers inside one clusters.
- Hadoop was originally created to address problems of RDBMS and DWH.

Hadoop Architecture

Master Big Data



Hadoop Distributed File System

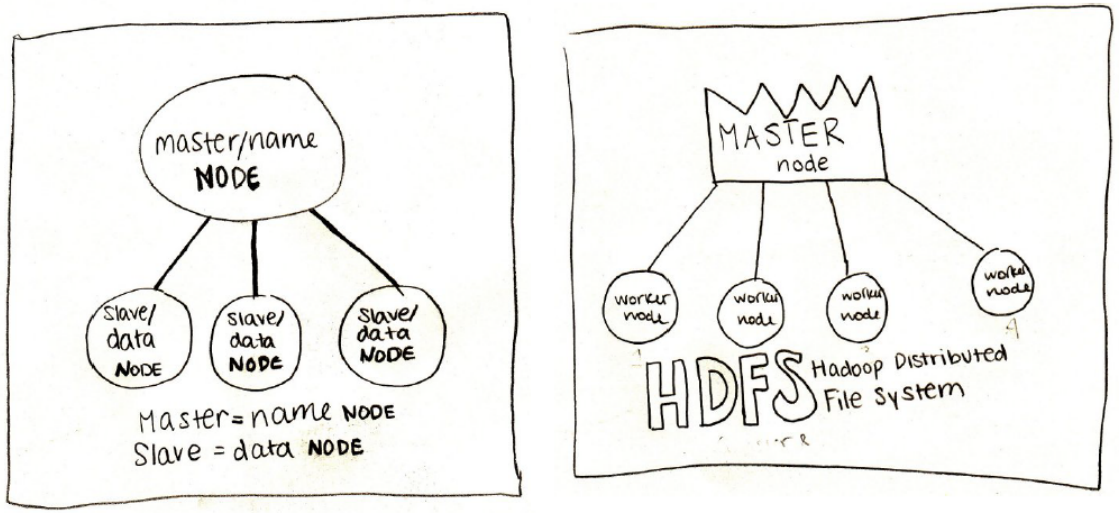
HDFS Design

- HDFS is a filesystem designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware.
- **Very Large files**
- **Streaming data access** (write once ready many times)
- **Commodity Software** (low cost servers)

Hadoop Distributed File System

HDFS Architecture

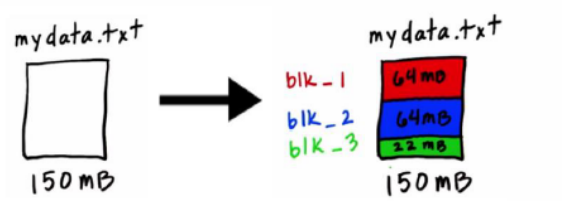
- In an HDFS cluster, there is ONE master node and many worker nodes.
- The **master node** is called the **Name Node (NN)** manages the file systems' operations
- Workers are called **Data Nodes (DN)**. Data nodes actually store the data. They are the workhorses.

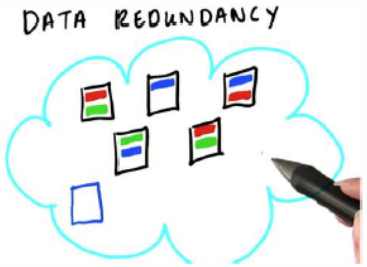


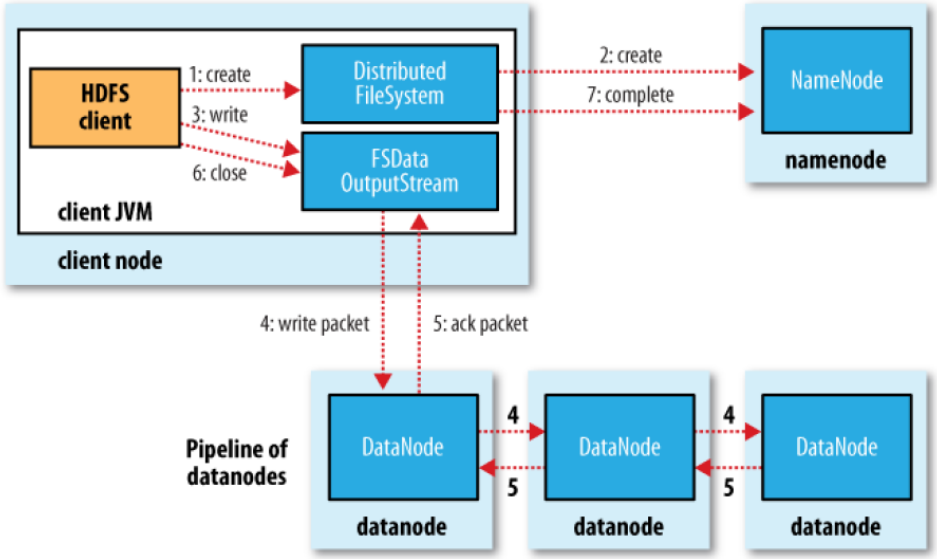
Hadoop Distributed File System

HDFS Architecture

- The Hadoop Filesystem (HDFS) divides files into blocks. A file can, for example, be divided into three blocks (blk_1, blk_2, blk_3). These blocks are then distributed among the different nodes and racks.







Conclusions

Apache's Hadoop open-source project

- Complete new topic, great start with the VM
- **Challenges of Hadoop (HDFS)**
  - Lot's of small Data
  - Super fast data access
  - Multiple writers
  - There's a widely acknowledged talent gap
  - Full data management and governance.

- *Hadoop Illuminated, Kerzner et. al.(2018)*
- *Hadoop; The Definite Guide, White (2012)*

Bibliography

Great Resources

Mark Torr. 3 ways to use Hadoop without throwing out the DWH, October 2014. URL http://www.marktorr.com/3-ways-to-use-hadoop-without-throwing-out-the-dwh/.

Óscar Pereira and Micael Capitão. Figure 2.1: Hadoop ecosystem overview. 1, December 2014a. URL https://www.researchgate.net/figure/Hadoop-ecosystem-overview-1_fig2_270448794.

Óscar Pereira and Micael Capitão. Figure 2.5: Word count program flow executed with MapReduce. 5, December 2014b. URL https://www.researchgate.net/figure/Word-count-program-flow-executed-with-MapReduce-5_fig6_270448794.

Mark Kerzner and Sujee Maniyam. *Hadoop Illuminated*. 2008.

Sarah Sproehnle, Ian Wrigley, and Gundega Dekena. Intro to Hadoop and MapReduce | Udacity, February 2015. URL https://eu.udacity.com/course/intro-to-hadoop-and-mapreduce--ud617.

Bernard Marr. *Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results*. Wiley, Chichester, West Sussex, 2016. ISBN 978-1-119-23138-7.

Mark Van Rijmenam. *Think bigger: developing a successful big data strategy for your business*. AMACOM, American Management Association, New York, 2014. ISBN 978-0-8144-3415-4.

Alex Woodie. Why Hadoop on IBM Power, May 2014. URL https://www.datanami.com/2014/05/12/hadoop-ibm-power/.

Nicole Hemsoth. Cray Launches Hadoop into HPC Airspace, October 2014. URL https://www.hpcwire.com/2014/10/15/cray-launches-hadoop-hpc-airspace/.

Margaret Rouse. What is Hadoop? - Definition from WhatIs.com, 2014. URL https://searchdatamanagement.techtarget.com/definition/Hadoop.

Sachin Bappalige. An introduction to Apache Hadoop for big data, August 2014. URL https://opensource.com/life/14/8/intro-apache-hadoop-big-data.

Frank Kane. What is Hadoop?, March 2018. URL https://www.youtube.com/watch?v=DCaiZq3aBSc&t=1282s.

J Jeffrey Hanson. An introduction to the Hadoop Distributed File System. *IBM - developerWorks*, page 8, January 2011.

Michael Malak. Data Locality: HPC vs. Hadoop vs. Spark, September 2014. URL /content/data-locality-hpc-vs-hadoop-vs-spark.

Tom White. *Hadoop: the definitive guide*. O'Reilly, Beijing, third edition edition, 2012. ISBN 978-1-4493-1152-0.

Dhruba Borthakur. HDFS Architecture Guide, 2018. URL https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

Michael Fertik and David C. Thompson. *The reputation economy: how to optimize your digital footprint in a world where your reputation is your most valuable asset*. Crown Business, New York, first edition edition, 2015. ISBN 978-0-385-34759-4.

Y. Wang, R. Goldstone, W. Yu, and T. Wang. Characterization and Optimization of Memory-Resident MapReduce on HPC Systems. pages 799–808, May 2014. doi: 10.1109/IPDPS.2014.87.