GOETHE UNIVERSITÄT FRANKFURT AM MAIN

www.uni-frankfurt.de

Big Data: A general overview

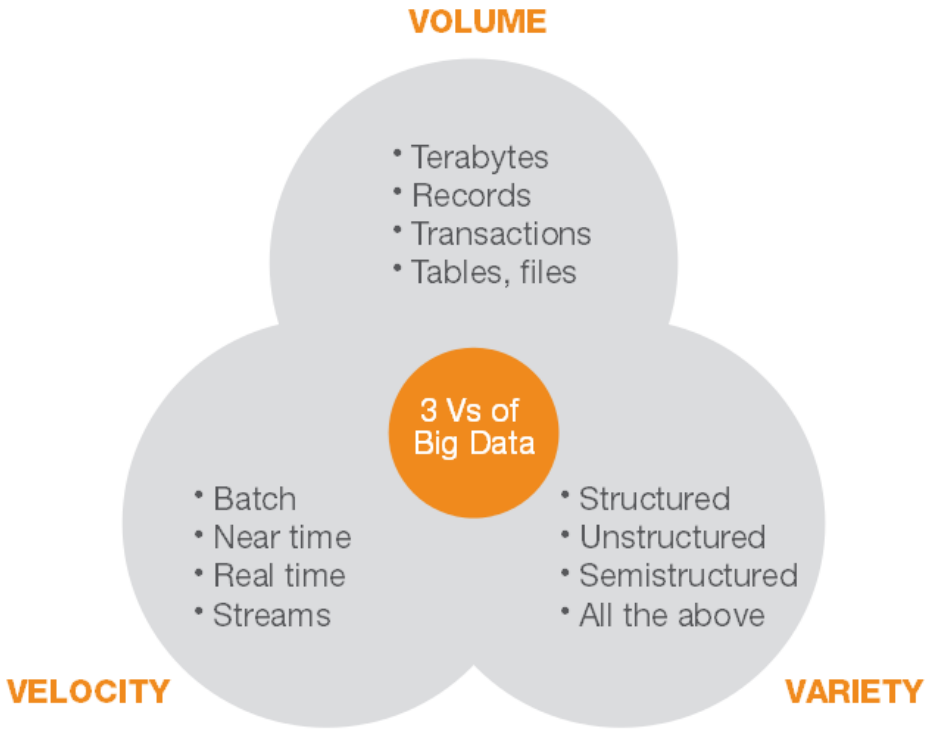
Todor Ivanov

Big Data Laboratory
Chair for Databases and Information Systems

University of Frankfurt

Big Data

- Exponential data growth - **Volume**
  - petabytes/exabytes of user data (text, audio, video, images)
- **Variety** of data sources:
  - Mobile devices
  - Social platforms
  - Sensors (RFID)
  - Web platforms
- **Processing speed - Velocity**
  - How fast are the results available?



Big Data Motivation

- Traditionally,computation has beenprocessor-bound
  - Fine for relatively limited amounts of data
  - Solution: Move computation to Data → Distributed Computing

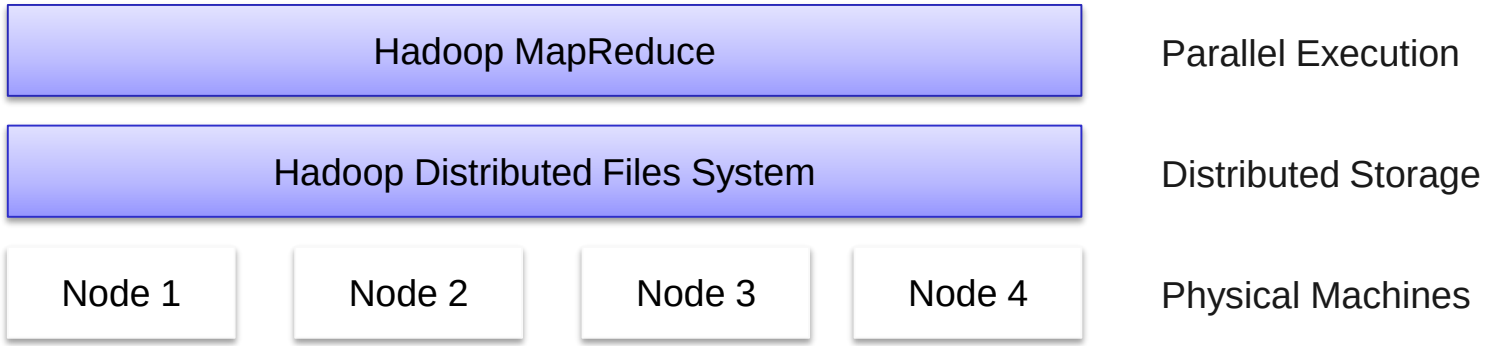
- Architecture effectiveness:
  - Commodity hardware (cost)
  - Multiple commodity servers in a cluster
  - Easy system management

- Business driven - focused on new ways of increasing revenue and reducing costs.
- Accelerate the time-to-value cycle

Hadoop

- Apache Hadoop is **open source framework** (Apache License) for **storing**, **processing** and **analyzing** massive amounts of distributed, unstructured data.

- Originated in Google and implemented by Yahoo & Facebook
- Distributed cluster system
- Platform for massively scalable applications
- Enables parallel data processing
- Built-in replication
- Stores petabytes of unstructured data
- Move computation to data
- Automatically handles node failures
- Written in Java(cross-platform portability)



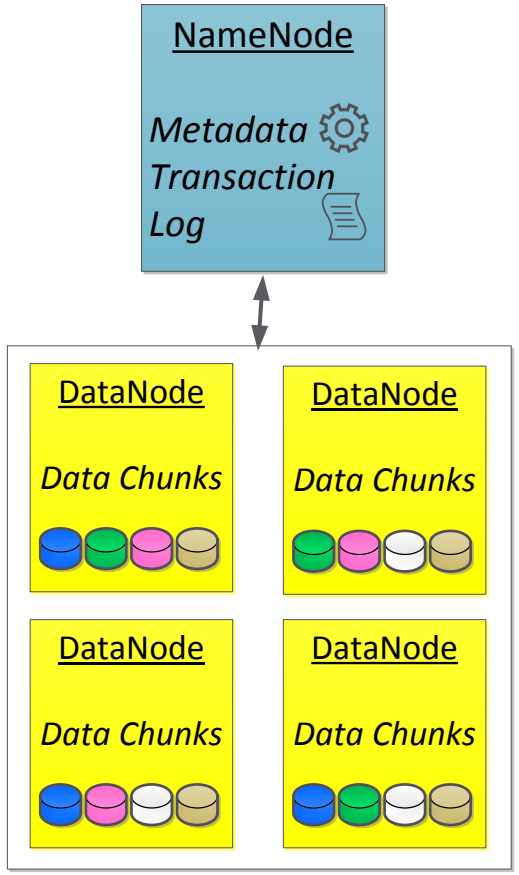
Hadoop Distributions

- Apache Hadoop
- Cloudera CDH
- Hortonworks
- IBM BigInsights
- Greenplum Pivotal HD
- Intel Project Rhino
- MapR
- Microsoft HDInsight

- Steadily growing number of Hadoop distributions!

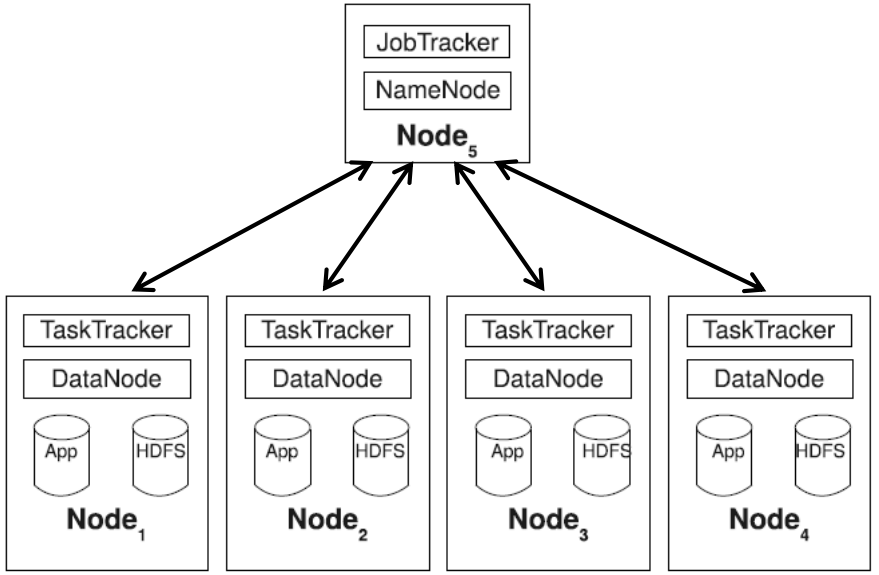
Hadoop Distributed File System

- Distributed filesystem – Master/Slave architecture
- File organization (create, delete, remove ... operations)
- Write-once-read-many access model
- **NameNode** as master
- **Secondary NameNode** as hot backup
- **DataNodes** store the data in blocks
- Default Replication Policy (3 copies of each block)



MapReduce

- **MapReduce** is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes.
- open source impementation of Google‘s MapReduce
- clean API between MapReduce and HDFS
- **JobTracker**
  - splitting into map and reduce tasks
  - scheduling tasks on a cluster node
- **TaskTracker**
  - runs MapReduce tasks periodically



Cloudera Academic Partnership

cloudera®
ACADEMIC PARTNER

- Introduce Apache Hadoop to university students
- Provide students with high-quality **curricula** and **materials**
- Focus on Hadoop Cluster Administration and MapReduce Application Development
- **Certify** students with Hadoop professional credentials to complement their university degrees
- Discount for Cloudera Certifications

Cloudera Academic Training

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

cloudera® ACADEMIC PARTNER

Course Chapters

▪ **Introduction**	**Course Introduction**
▪ The Motivation for Hadoop ▪ Hadoop: Basic Concepts	Introduction to Apache Hadoop and its Ecosystem
▪ Writing a MapReduce Program ▪ Unit Testing MapReduce Programs ▪ Delving Deeper into the Hadoop API ▪ Practical Development Tips and Techniques	Basic Programming with the Hadoop Core API
▪ Common MapReduce Algorithms	Problem Solving with MapReduce
▪ Planning Your Hadoop Cluster ▪ Hadoop Installation	Planning, Installing, and Configuring a Hadoop Cluster
▪ Managing and Scheduling Jobs ▪ Cluster Monitoring and Troubleshooting	Cluster Operations and Maintenance