Fourth Industrial Summer School

Big Data Analytics

Introduction and Fundamentals

Session Objectives

- ✓ Introduction
- ✓ Fundamentals

Big Data Analytics

What is it?

- Processing massive amounts of data that cannot fit in a single computer system
- Loading, analysis, and modeling of big data and making predictions from the learned models

Traditional vs. current approaches

Why Big Data

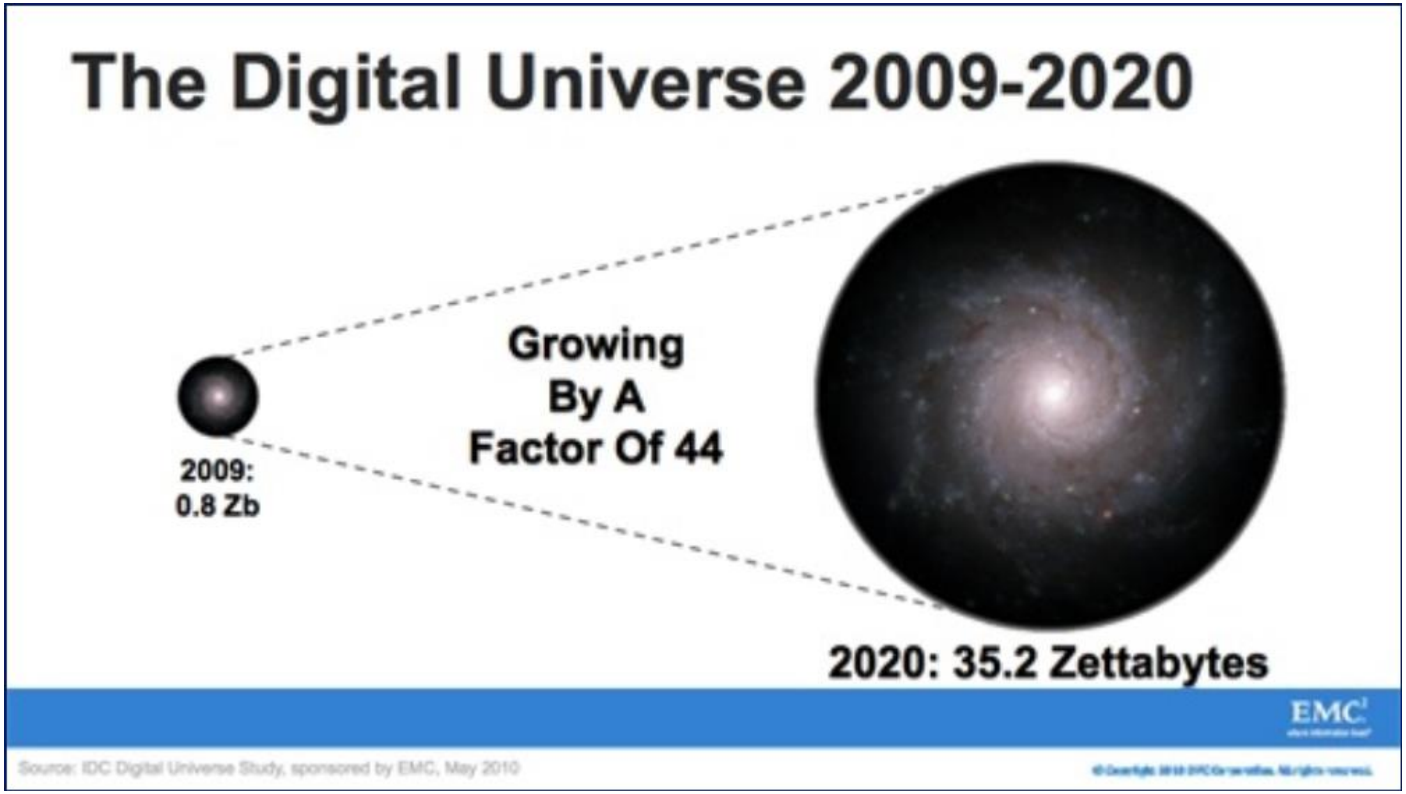
- Data being produced at an exponential rate
- Computing capabilities, commodity clusters
- More data leads to better modeling and predictions, which in turn leads to:
  - Personalized services
  - Recommendation systems
  - Sentiment analysis
  - Location-based adds
  - Smart cities

Interesting insights

- 90% of the information ever generated was generated in the last two years!
- Every Minute:
  - 204 million emails
  - 200,000 photos, 1.8 million likes in Facebook
  - 1.3 million video views, 72 hours of video upload on YouTube

- Source: 25 interesting facts about big data by Bernard Marr, https://www.smartdatacollective.com/big-data-25-facts-everyone-needs-know/)

Interesting insights



The Digital Universe 2009-2020

2009:
0.8 Zb

Growing
By A
Factor Of 44

2020: 35.2 Zettabytes

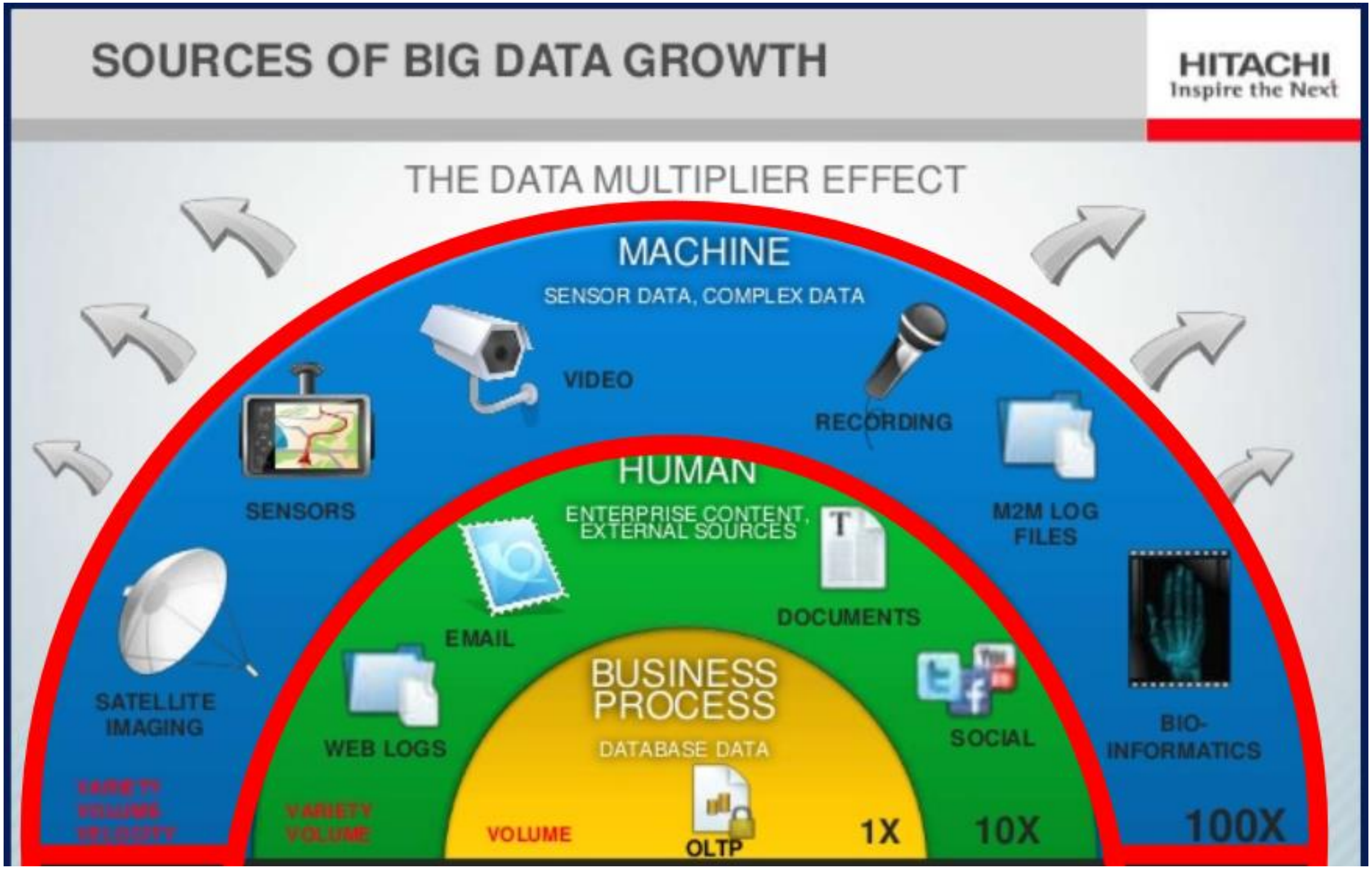
Source: IDC Digital Universe Study, sponsored by EMC, May 2010

- Source: 25 interesting facts about big data by Bernard Marr, https://www.smartdatacollective.com/big-data-25-facts-everyone-needs-know/)

Sources of Big Data

- Machine-generated data:
  - Largest source of data
  - Sensors (machines, smart devices)
- People generated data:
  - Social media, emails, blogs
  - Mainly unstructured and text intensive
  - Facebook produces more data in a day than all the US academic research libraries.
- Organization generated data:
  - Banks, stores, hospitals, governmental institutions
  - Mainly structured data

Sources of Big Data



https://www.slideshare.net/hdscorp/capitalize-on-big-data-through-hitachi-innovation

Big Data–Characteristics

- Data that cannot fit in a single computer
- It is generally identified by Five 'Vs':
  - **Volume**: Challenges related to storage, access, and processing
  - **Velocity:** Real-time processing vs. batch processing
  - **Variety:** Challenges related to integration, storage, and processing
  - **Veracity:** Challenges related to data validity
  - **Valence:** Challenges related to copmlex processing

Big Data–Process

- Depends on defining the right problem
- Big data analytics in the big picture (process):
  - **Acquiring data**: from multiple sources using SQL queries, file parsing, web services.
  - **Exploring:** Understanding data using stats, plots, etc.
  - **Preprocessing:** Cleaning and transformation
  - **Analysis:** Predictive models, clustering, graph analytics, etc.
  - **Reporting**: The results
  - **Actions**: With feedback to close the loop

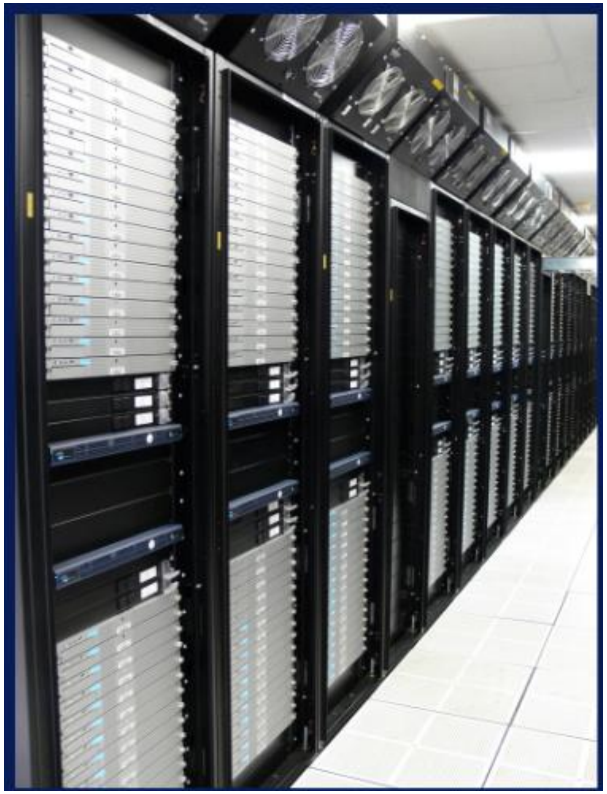
- Value comes from data integration, processing, and modeling.

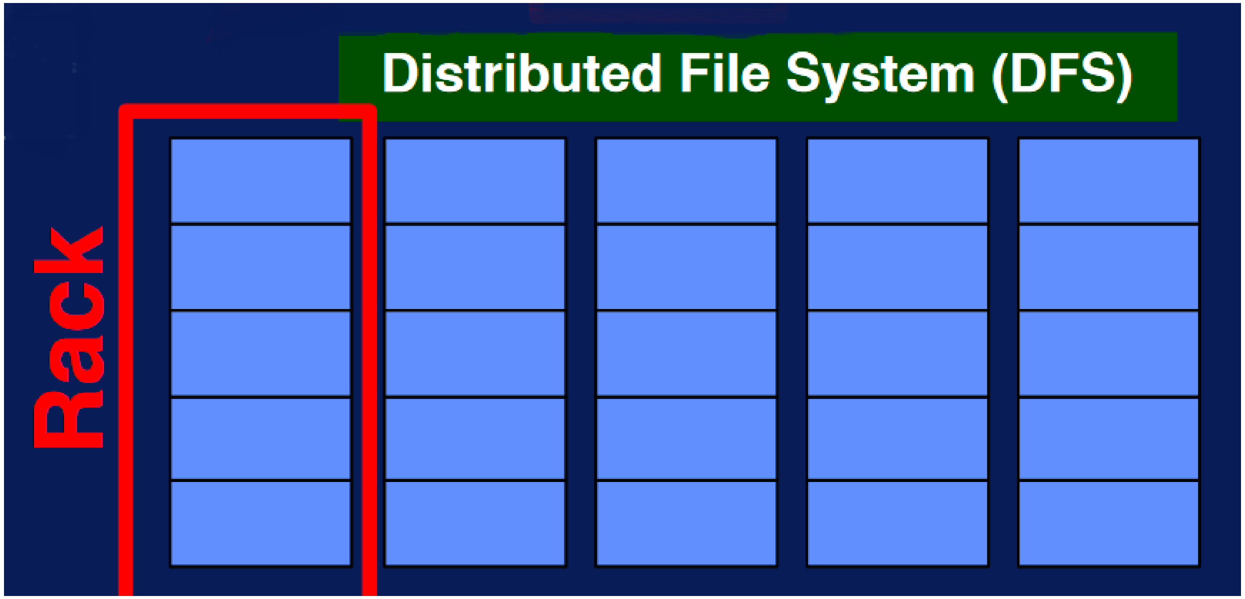
Foundations

Big Data

Distributed File Systems

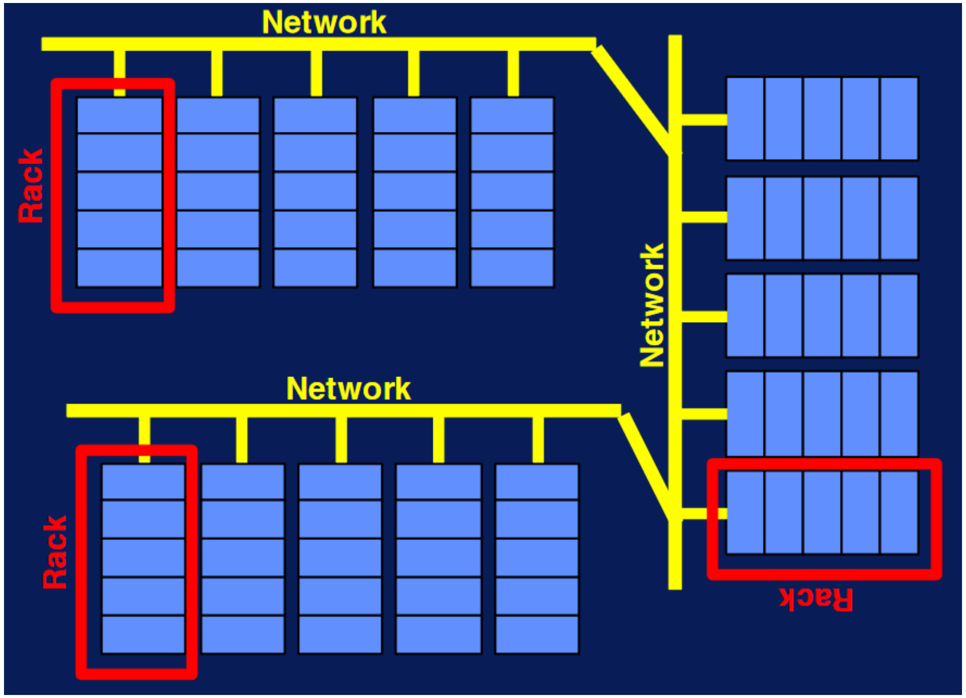
- All the data cannot fit in one computer





- Allows for fault tolerance, scalability and concurrency

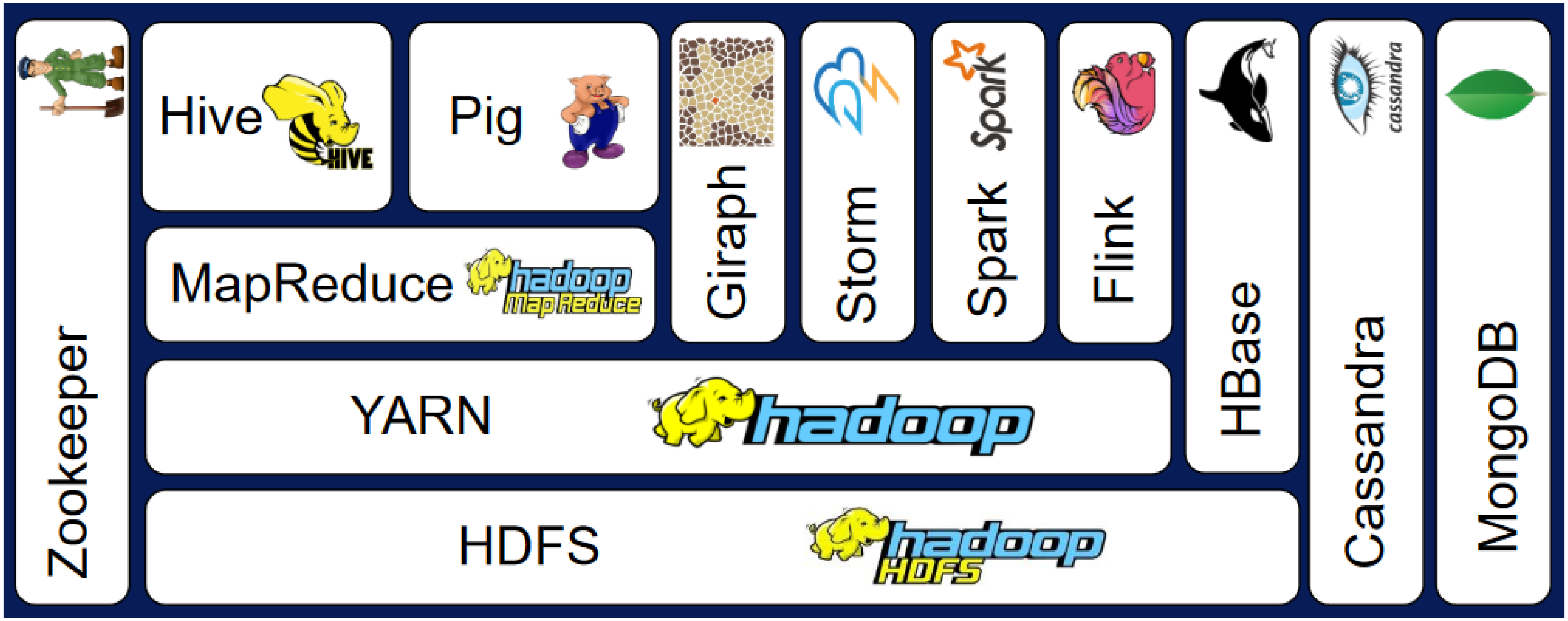
Commodity Clusters vs. Super Computer



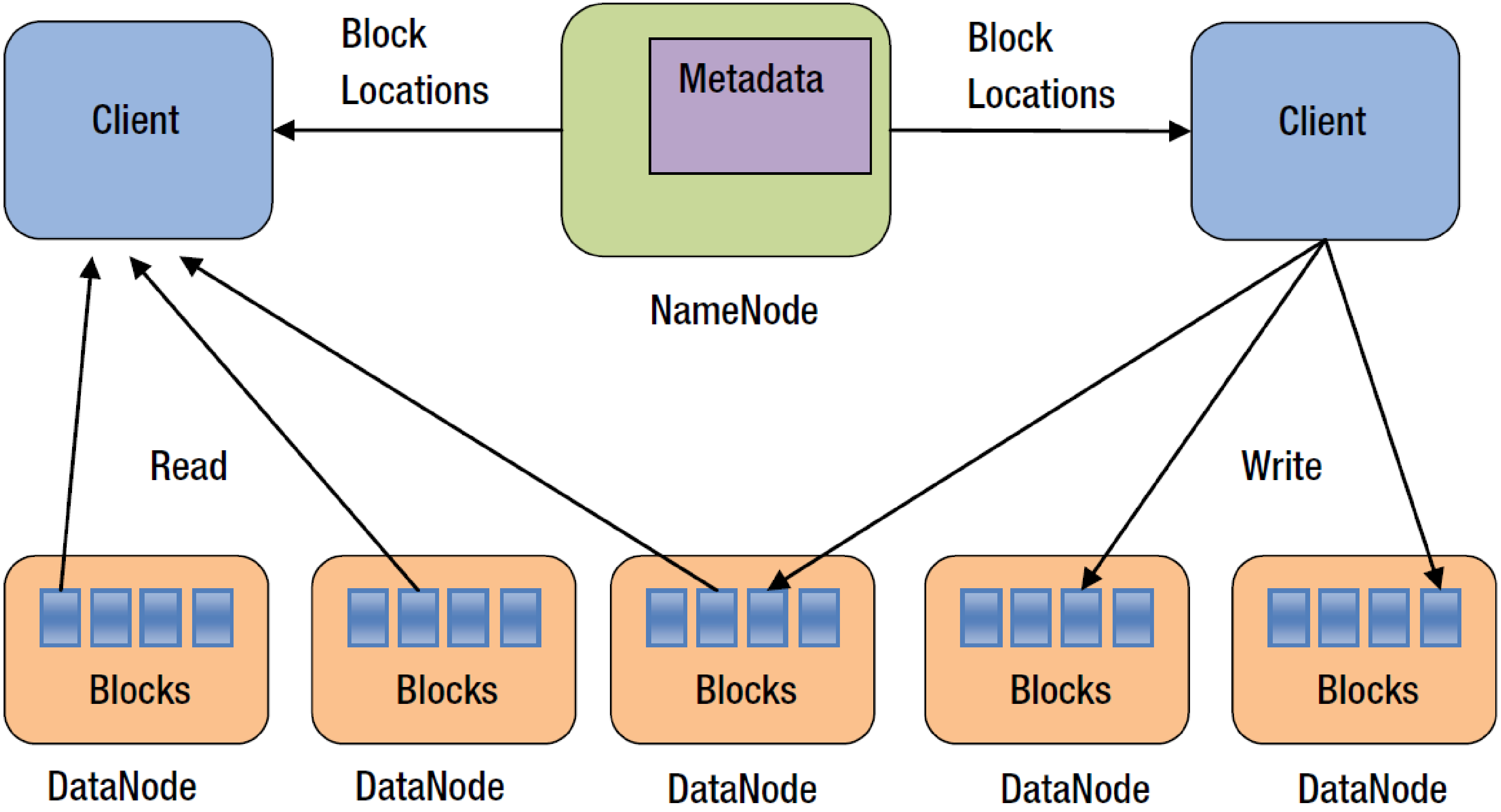
The Hadoop Ecosystem

- Mostly free and open source
- Hadoop is inspired by a system invented at Google in 2004
  - Google file system
  - MapReduce
- Yahoo created Hadoop in 2005
- Layered approach
- Cost benefits along with scalability, fault tolerance, and parallel processing
- Fault tolerance through software is cheaper than implementing it in hardware
- Moving code closer to data and not vice-versa
- No complicated skills to manage parallel computation

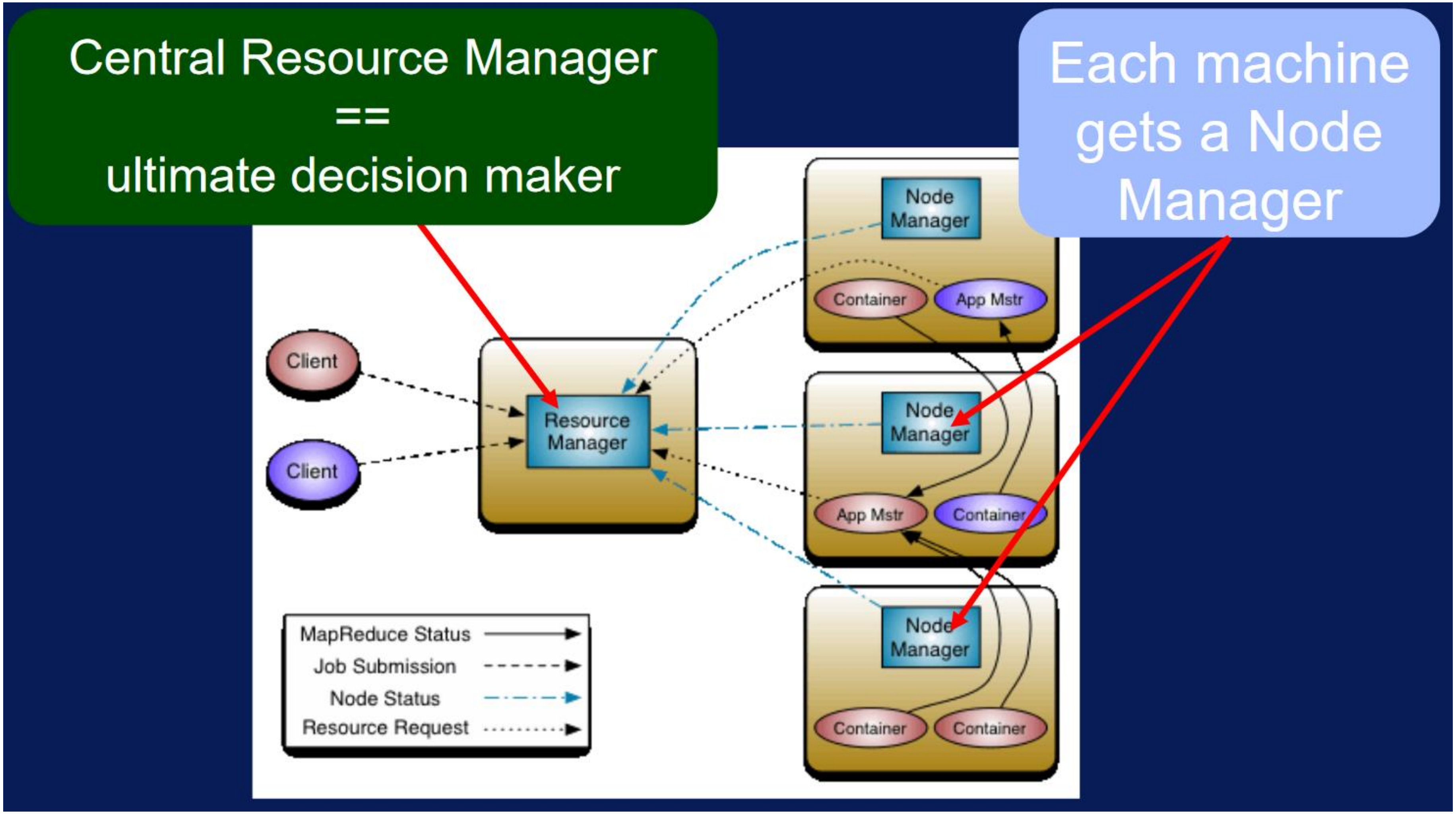
The Hadoop Ecosystem



HDFS Architecture



YARN-Resource Manager and scheduler



Central Resource Manager
==
ultimate decision maker

Each machine gets a Node Manager

MapReduce

- Programming model for Hadoop ecosystem
- WordCount example

Apache Spark

References

- Ilkay Altintas and Amarnath Gupta, Introduction to Big data, University of California San Diego: https://www.coursera.org/learn/big-data-introduction/home/welcome
- Guller, Mohammed. Big data analytics with Spark: A practitioner's guide to using Spark for large scale data analysis. Apress, 2015.
- https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation