Amirkabir University of Technology
(Tehran Polytechnic)

Cloud Computing

Big Data and Analytics in Cloud-Part1

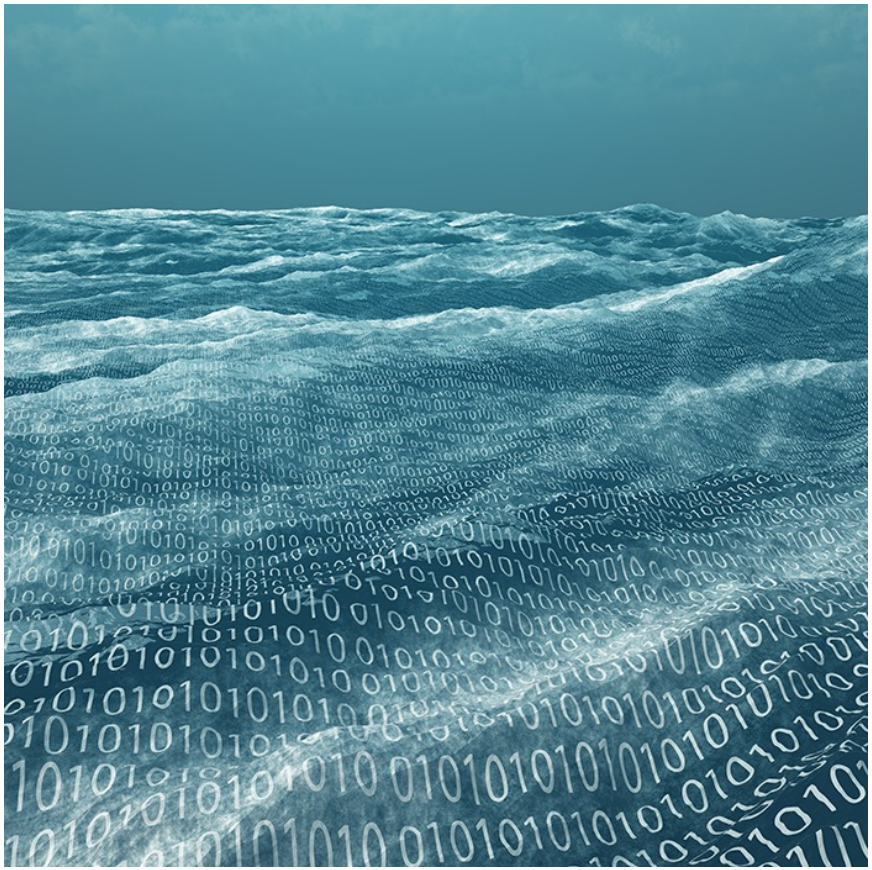
Seyyed Ahmad Javadi

Arian Boukani

Fall 2022

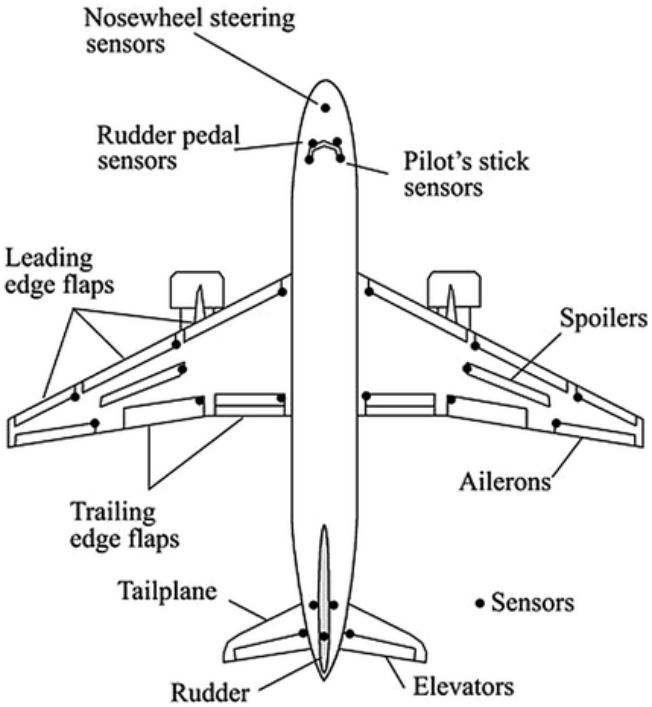
Introduction

- Our world is awash in a rising ocean of data.



Examples of Big Data

- Modern modes of transportation contain thousands of sensors
  - The sensors constantly generate data.
  - Planes, cars, and ships, …



Examples of Big Data (cont.)

- Medical researchers study thousands of genes in millions of patients
  - Attempting to find genes that lead to diseases.

Examples of Big Data (cont.)

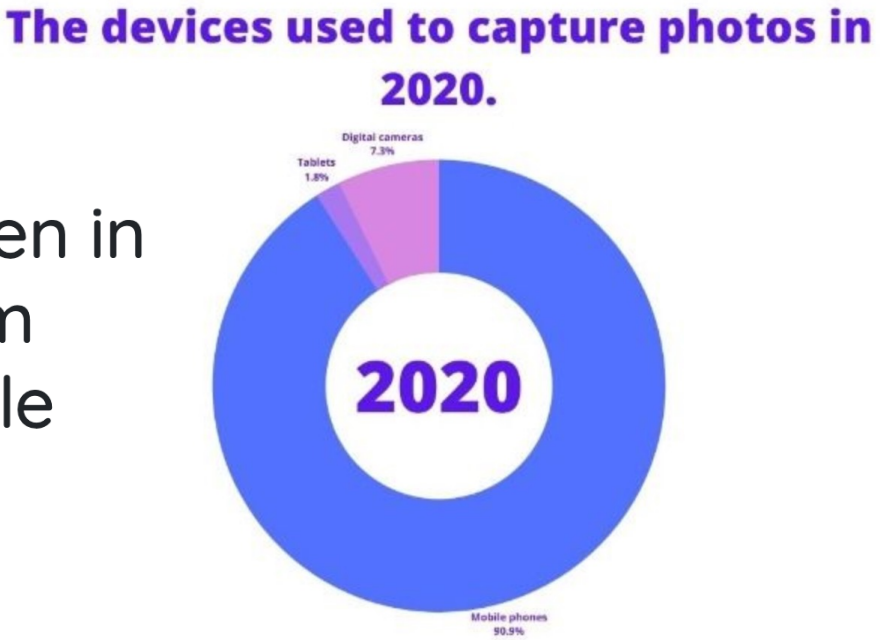
- Online businesses (e.g., Amazon) track customers' online behavior
  - Use the information to suggest books or movies for them to purchase.

Examples of Big Data (cont.)

- Smartphones record audio, video, and images and their owners post them to social media sites.

1.43 trillion photos were taken in 2020 but how many of them were captured on our mobile phones?

Jessica Canning - 30 December 2020



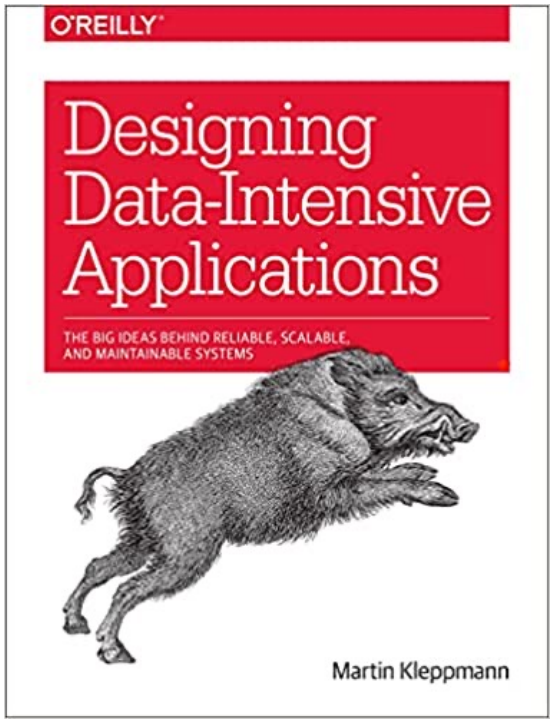
A whopping 90.9% of these photos were taken on mobile phones!

https://www.buymobiles.net/blog/1-43-trillion-photos-were-taken-in-2020-but-how-many-of-them-were-captured-on-our-mobile-phones/

The age of big data has begun, and exploiting big data is changing our world.

Data-Intensive Computing

- Data-intensive computing is about
  - **production,**
  - **manipulation,**
  - **and analysis of large-scale data**



in the range of hundreds of megabytes to ***petabytes & beyond***.

Application Domains

- There are many application domains.
- Exemplar: Computational science
  - Hundreds of GBs of data are produced by telescopes.
  - Bioinformatics applications mine terabytes of data.
  - Earthquake simulators process a massive amount of data.

IT Industry Sectors

- Customer data would easily be in the range of 10-100 terabytes.
  - Processed to generate billing statements
  - Mined to **identify scenarios and patterns** to provide better service.

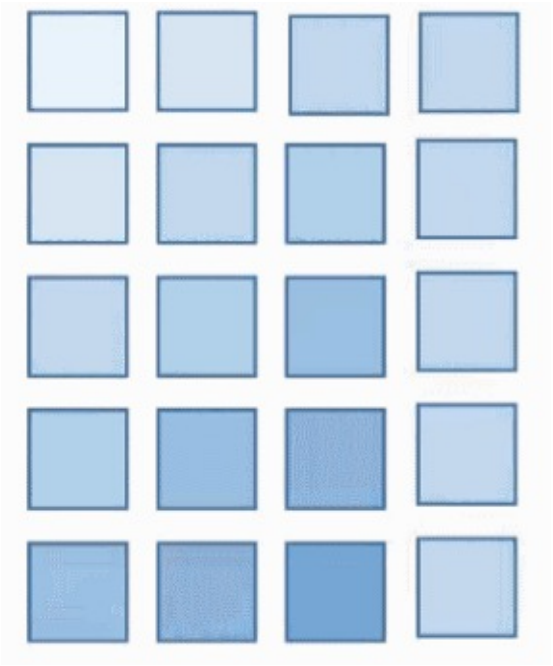
- Google is reported to process about 24 petabytes of information per day and to sort petabytes of data in hours.

Types of Data

- Structured data
- Semi-structured data
- Unstructured data

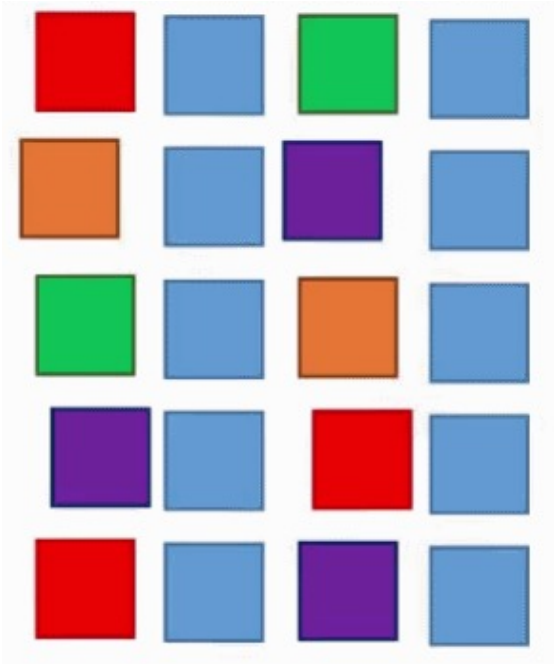
Structured Data

- Data that is organized in a structure
- Examples
  - Fixed fields inside a record (e.g., relational database)
  - Well-formed format (XML or JSON).



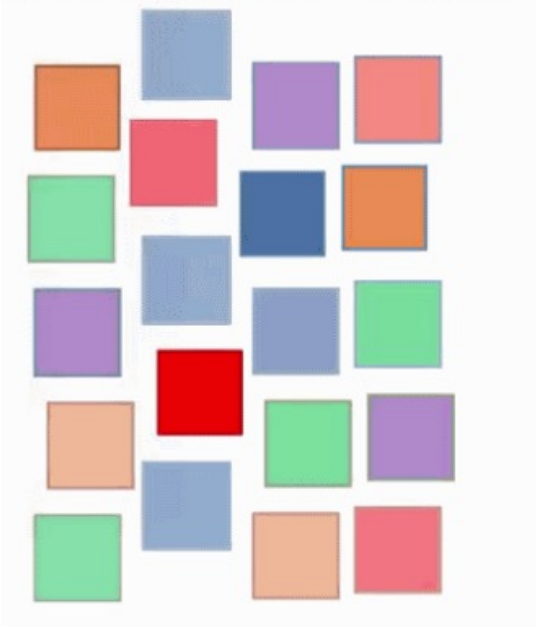
Semi-Structured Data

- Has some structure, but the data isn't expressed in terms of rows and columns.
- Examples:
  - An HTML page



Unstructured Data

- Does not have fields in fixed locations, nor does it follow a standard format such as XML or JSON.
- Examples:
  - Raw text files such as a server log
  - A Microsoft Word document
  - A Portable Document Format (PDF) file.



Sources of Data

- Sources of big data include:
  - Business operational data
  - Scientific data
  - Social networking
  - Web logs
  - Video streaming
  - Sensor data
  - Smartphone data
  - Many more …

Streams

- Some data is produced in streams.
- A data stream is “a sequence of digitally encoded signals used to represent information in transmission”.

Streams-Examples

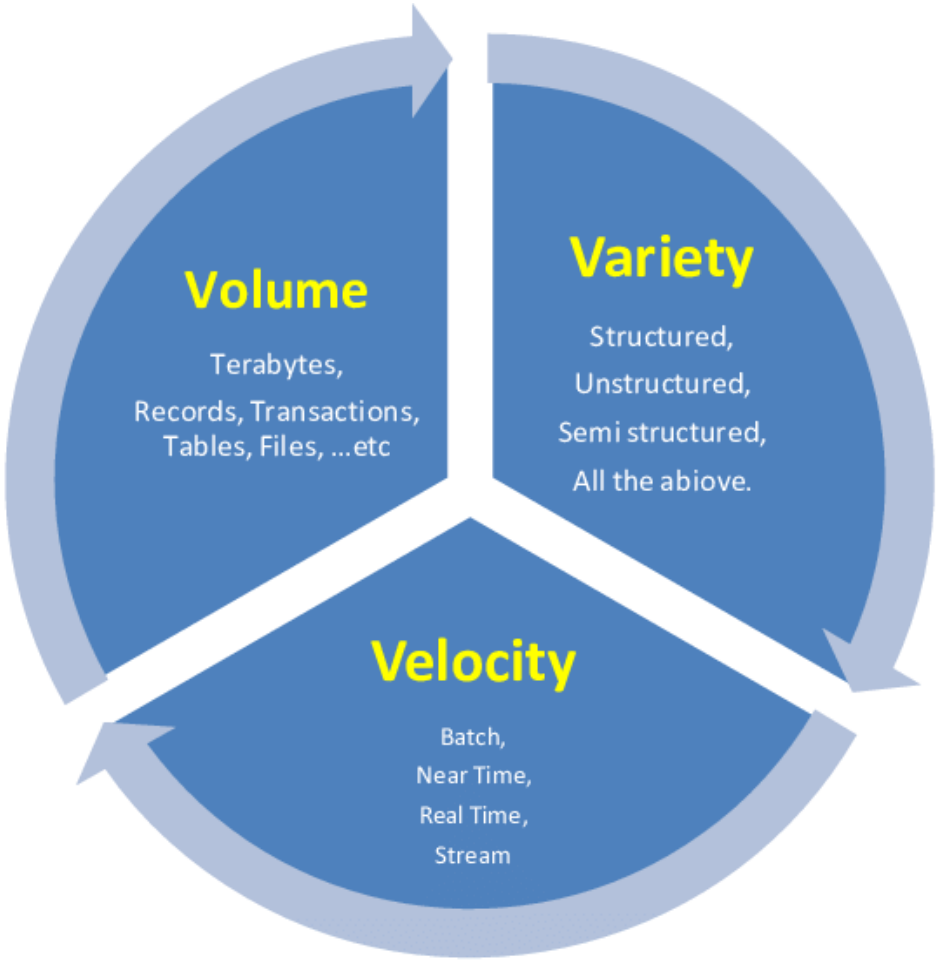
- Click streams
- Packet streams
- Sensor data
- Satellite data
- A video stream produced by an online video camera
- Financial data such as stock-market data.
- ...

Big Data Definition

- **First Definition Butler** (2013)
  - Data that has grown to a size that requires new techniques to store, organize, and analyze the data.

The three “Vs” Definition

- The three “Vs” definition: **volume**, **velocity**, and **variety**.



Three Key Features

- Big data has ***one or more*** of three key features

1. A large volume of data
2. A high velocity with which the data is created
3. A high degree of variety in the data.

Volume

- **Volume** simply means the **amount of data**.
  - How Big?
  - Generally the term “big data” is used to indicate data > 100 GB
  - Often it deals with hundreds of terabytes, and possibly a PB or more.

Three Key Features (cont.)

- **Velocity** means that **the data is being created rapidly**
  - For example, hundreds of messages per second
  - Velocity is typically associated with streams

Three Key Features (cont.)

- **High variety** is usually associated with **unstructured data**

Three Main Preconditions for the Rise of Big Data

- The ability to store large volumes of data in a form that is accessible
  - On hard drives or solid-state drives, not tape drives
- The ability to process big data rapidly at a reasonable cost.
  - Inexpensive computing power (e.g., cloud computing)
- The existence of producers of big data.

Big Data Analytics

- The ability to store and process that data into something that is **understandable by humans in a reasonable amount of time** that has allowed the exploitation of that data.

- Successful big data analytics must be able to process the data into something **smaller than the raw data**
  - Allowing an application to present the results in a way that makes sense to a human.

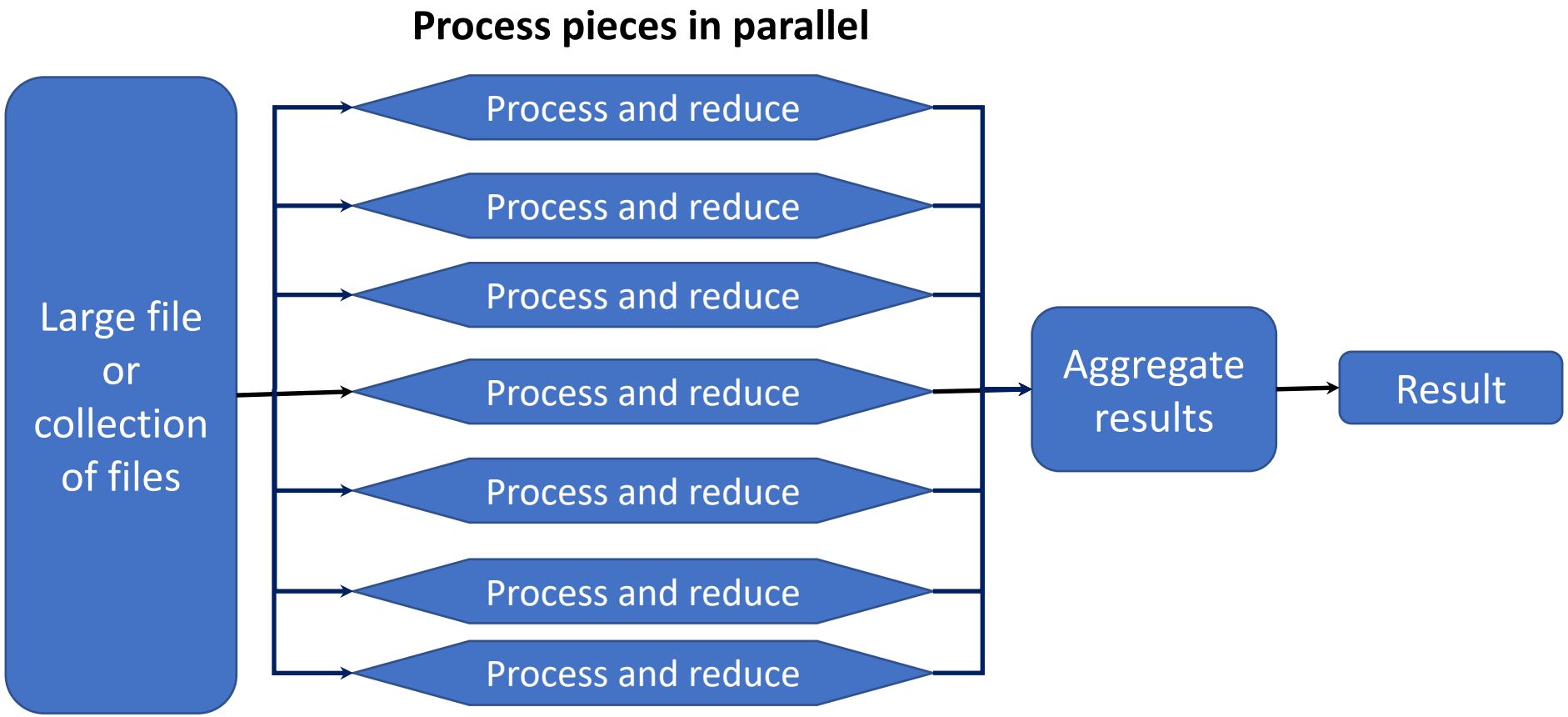
Big Data Analytics (cont.)

- Because big data is so large, **it normally cannot be processed sequentially in a reasonable amount of time.**

- So the **data is broken up into chunks**, which are analyzed by a set of processes running in parallel.

- The results of the parallel analysis are then **joined** together to create the result.

High-level Big Data Analytics



Technologies for Big Data

- **Storage Systems**
  - Distributed file systems and storage clouds
  - NoSQL Databases

- **Programming Platforms**
  - Map-Reduce: Apache Hadoop, Aneka
  - Stream Processing: Heron, Apache Storm, Apache Spark
  - Graph Processing: Pregel, Apache Giraph

Storage Systems

- Traditionally, **database management systems** constituted the de facto storage support for several types of applications.

- The relational model in its original formulation **does not seem to be the preferred solution for supporting data analytics on a large scale.**
  - Due to the explosion of unstructured data (e.g., blogs, Web pages,...),

High-performance distributed file systems and storage clouds

- **Distributed file systems** constitute the primary support for data management.
- They provide an interface whereby to store information in the form of files and later access them for read and write.
- Mostly these file systems constitute the data storage support for large computing clusters, supercomputers, massively parallel architectures, and lately, storage/computing clouds.

High-performance distributed file systems and storage clouds

- Google File System (GFS)
- **Hadoop Distributed File System (HDFS)**
- Amazon Simple Storage Service (S3)
- Lustre
- And many more

Hadoop Distributed File System (HDFS)

- HDFS stores **very large files** (many terabytes and petabytes).
- It could store **tens of millions of files**.
- It can run on **hundreds or thousands of commodity servers**.
- A general assumption about HDFS is that **hardware failure is a norm – not an exception**.
- It is suitable for **big data analytics**
  - **It is optimized to** write **very-big files** once and to read them many times
  - It is not suitable for random reads/writes.

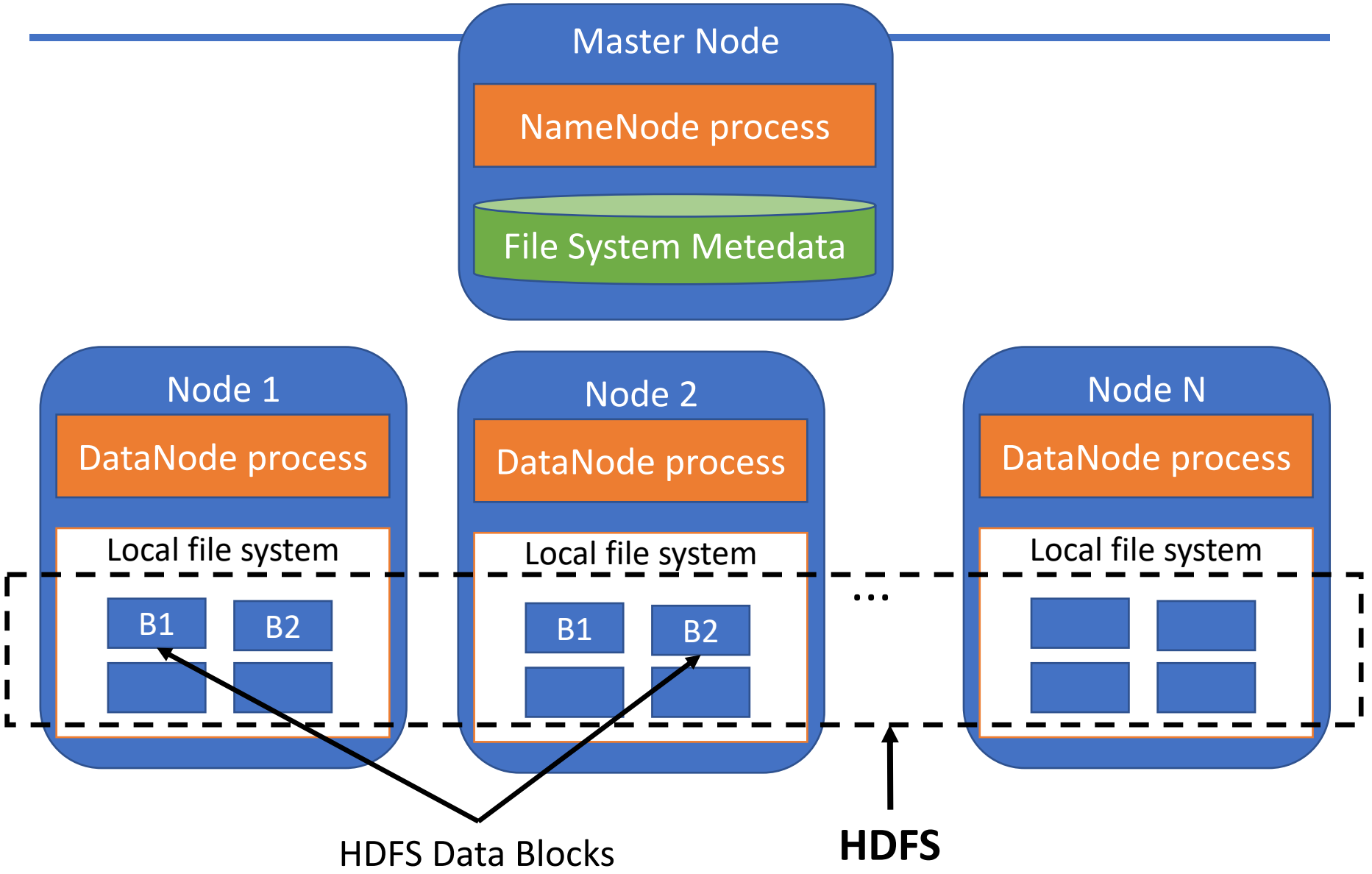
Hadoop Distributed File System (HDFS)

- An HDFS file is a sequence of **blocks stored in a cluster of multiple servers**.
- **Fault tolerance is at block level**.
- HDFS blocks are big ones – 64 MB by default.
- **HDFS is designed for applications that access (streaming) data sets successively**, it is not suitable for small files or for direct reads and writes.
- Hadoop moves the computations to the storage nodes
  - It is the best approach for cases **when the computing programs are relatively small**, and the stored data are big enough.

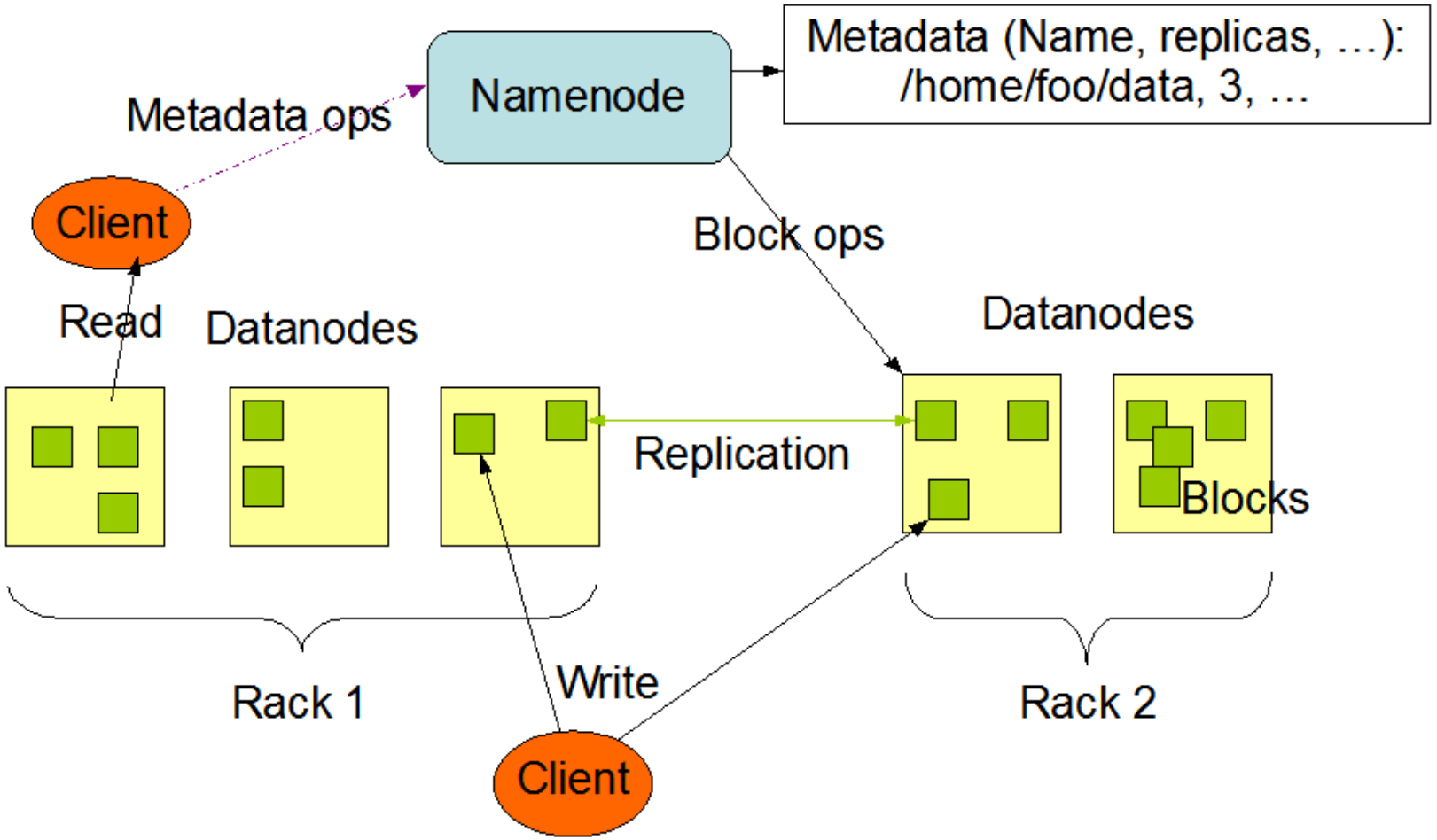
HDFS Structure

- There are namenode and datanode nodes (see next slide)
  - The namenode contains metadata for files, directories, file block locations, and so forth.
  - The datanode stores data blocks.
- The client opens files or directories using the metadata from a namenode, after that, the file datanodes execute the operations.
- The read operations directly access datanodes in a sequential read access mode.
- If a read fails, then the datanode uses a block replica.
- When HDFS has read all the data from a given block, it chooses the next block among all next block replicas.

HDFS Architecture



HDFS Architecture (Cont.)



Source: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html