Channel Islands
CALIFORNIA STATE UNIVERSITY

Big Data & Technologies

Presented by

Pallavi Chavan

Prof. Houman Dallali

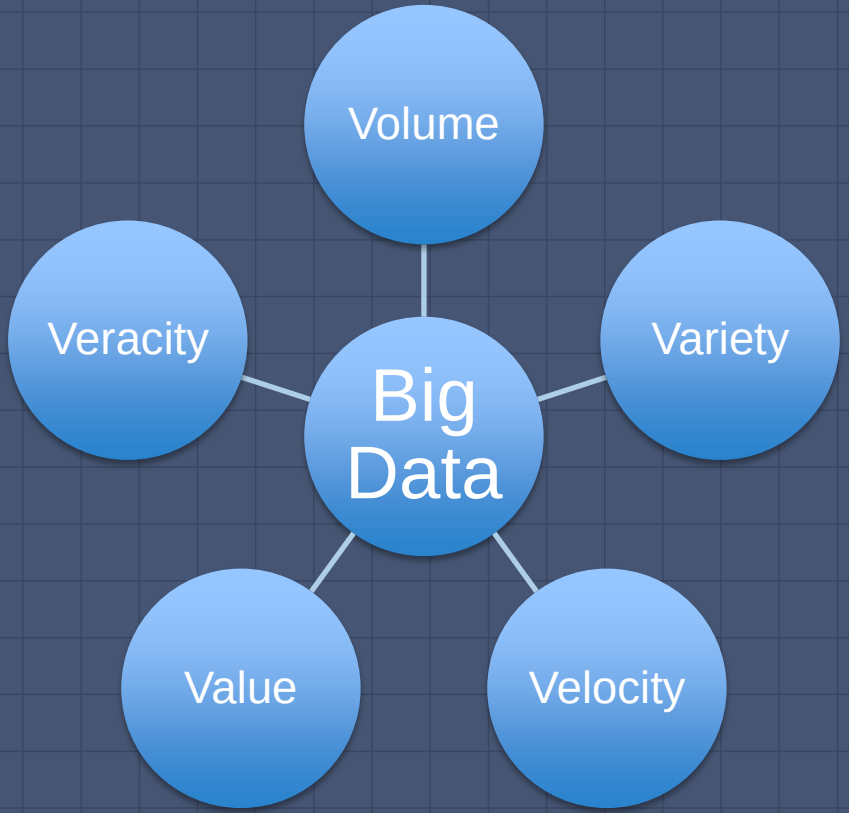
Date: October 9th, 2018

Content

- What is Big Data?
- Big Data Characteristics
- Use cases of Big Data Analytics
- Limitations of traditional approach to handle Big Data
- Solution – Hadoop
- Core components of Hadoop
- Summary
- Conclusion

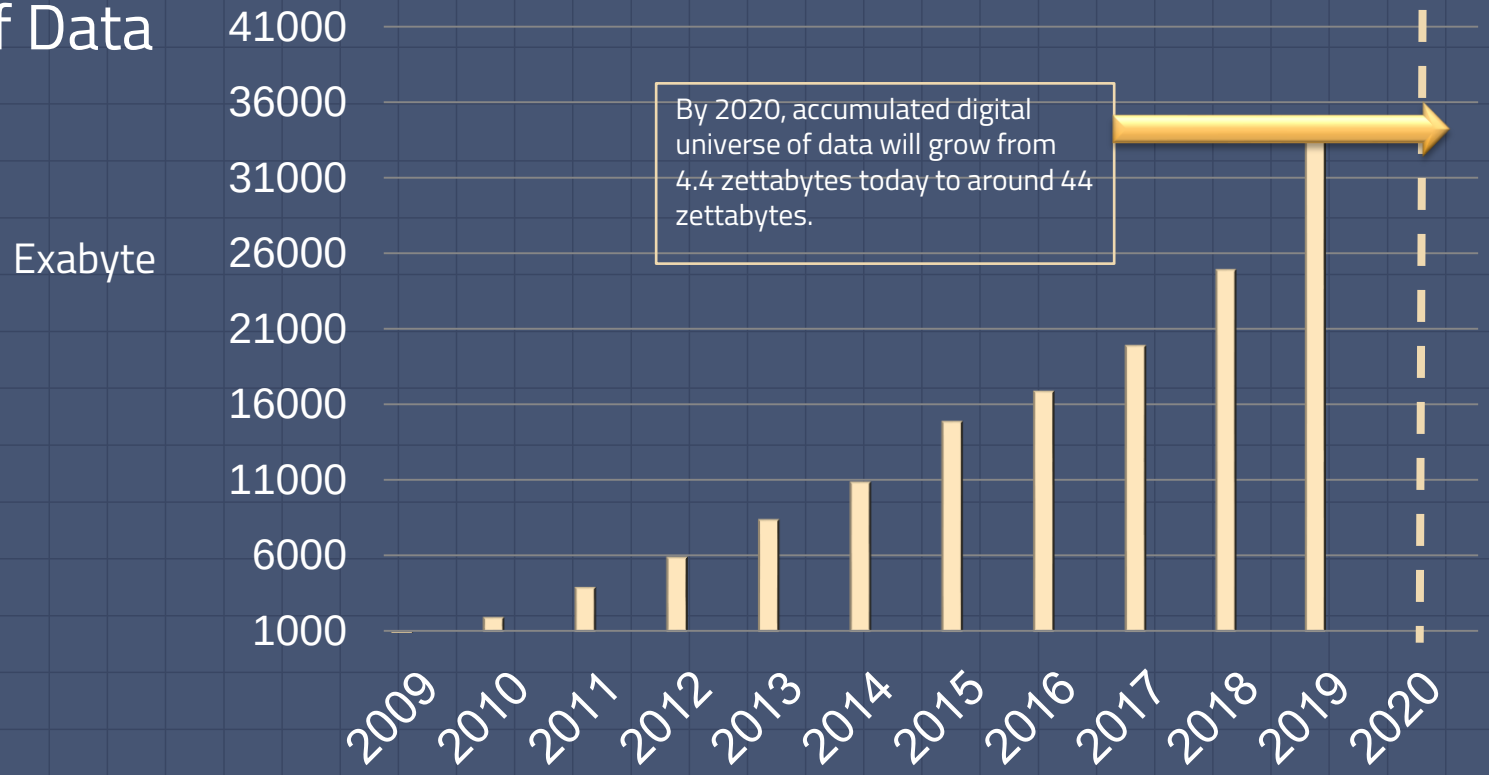
What is Big Data?

Collection of data sets is so large and complex that it becomes difficult to store and process using traditional data management system



Volume

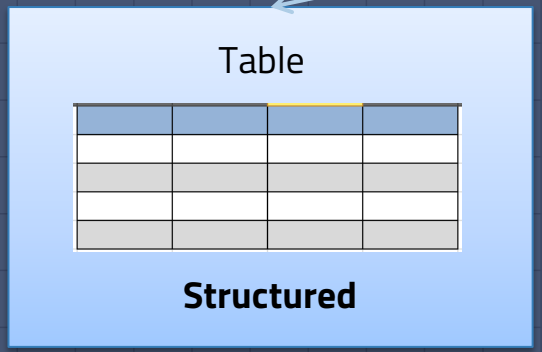
Quantity of Data

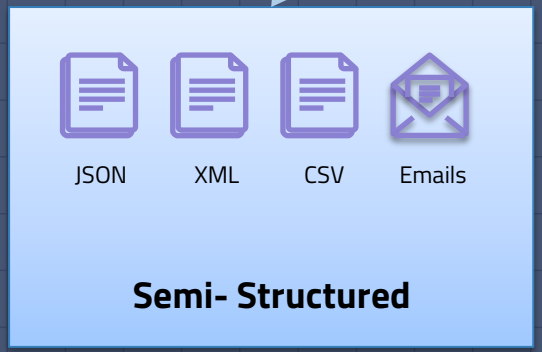


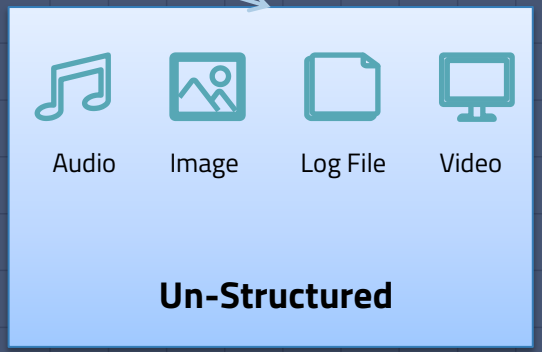
Variety

Type of Data

Different types of data is being generated by different sources



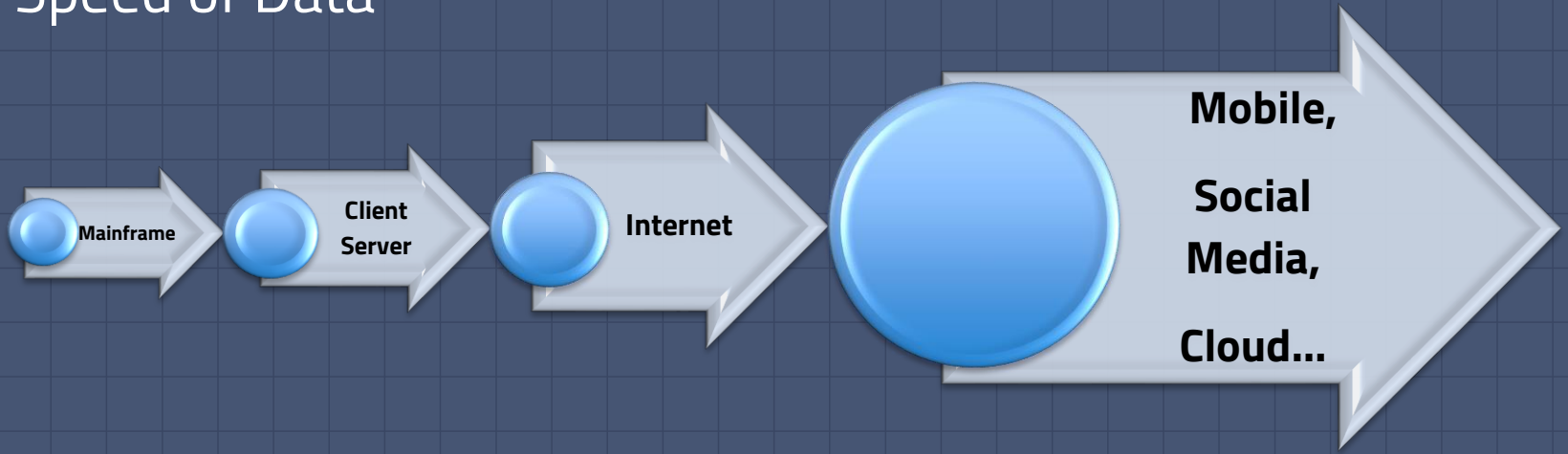




Velocity

Speed of Data

Data is being generated at every 60 seconds



700,000 + status updates

100,000 + tweets

168,000,000 + emails

698,445 Google searches

11,000,000 + instant messages

1820 TB

Value

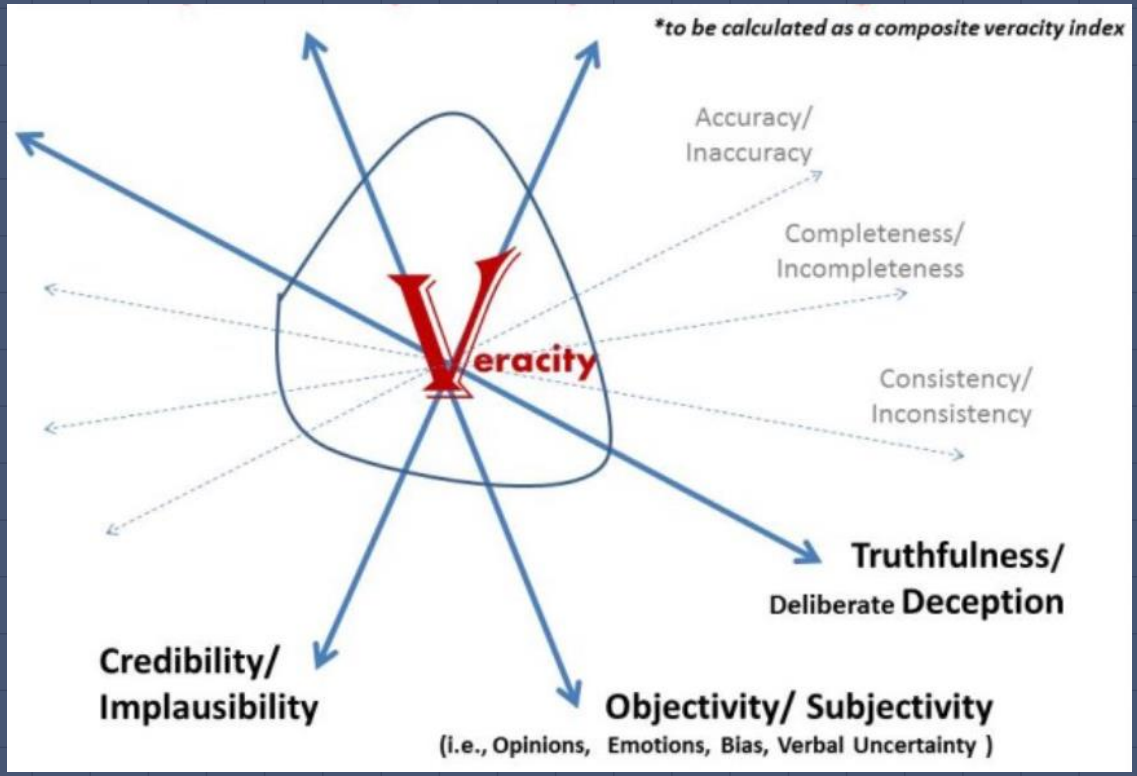
To find out correct meaning out of data

Use of Data

Veracity

Trustworthiness of Data

Quality or meaning of data



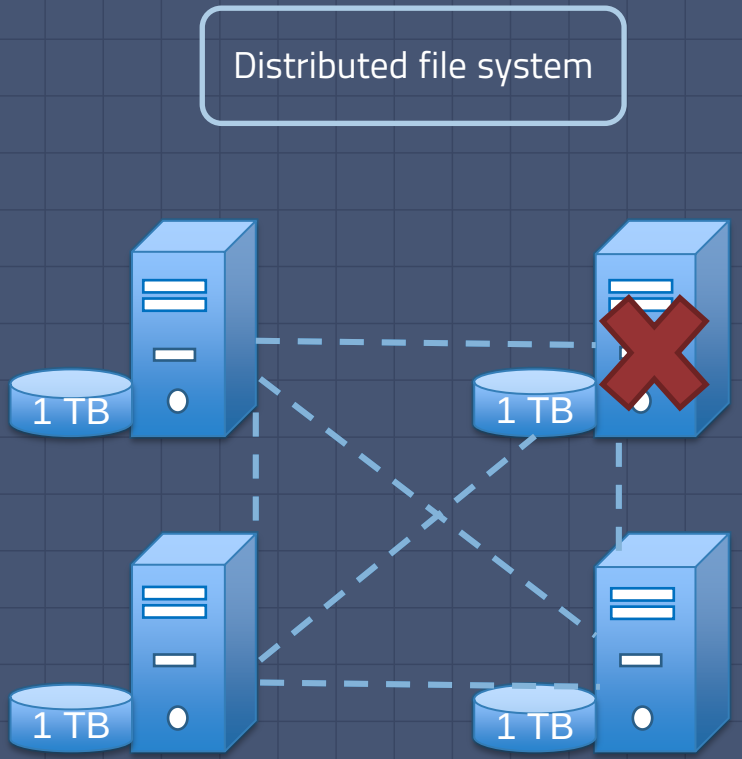
Big Data Analytics Use cases

- Recommendation System
- Smarter HealthCare
- Weather Model
- Predictive Policing
- Homeland Security
- Education System

Limitations of Traditional Processing System

- How to handle petabytes of data in RDBMS?
- How to handle semi structured or unstructured data in RDBMS?
- How can RDBMS handle data coming at high velocity?

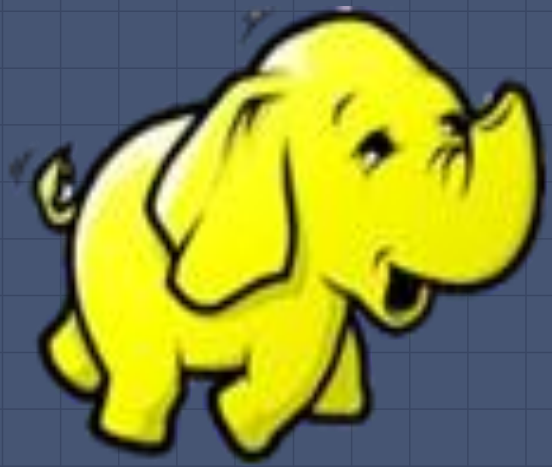
Can we use Distributed File System?



- Hardware Failure
  Many pieces of hardware, higher the chance that one will fail.
- Combine the data after analysis
  For most of the analysis, data read from one disk may need to be combined with data from other disks

HUGE

Helping Hand



hadoop

What is Hadoop?

- Hadoop is a framework for running applications on large cluster built of commodity hardware, using simple programming model

- Based on white papers published by Google –
  i. “The Google File System” published in 2003
  ii. “MapReduce: Simplified data processing on Large Cluster” published in 2004

- Developed by Doug Cutting & Mike Cafarella in 2006

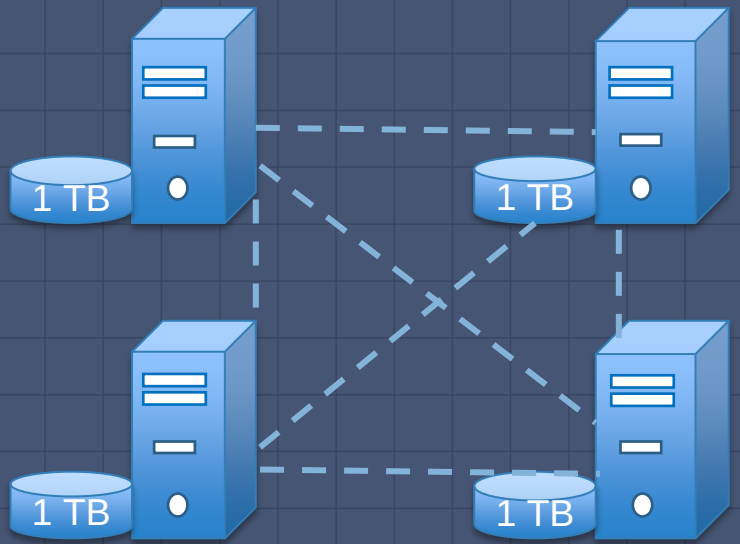
Core Components of Hadoop

To Store Big Data

To Process Big Data

What is HDFS ?

HDFS is a Hadoop Distributed File system that allows us to store large data across large cluster

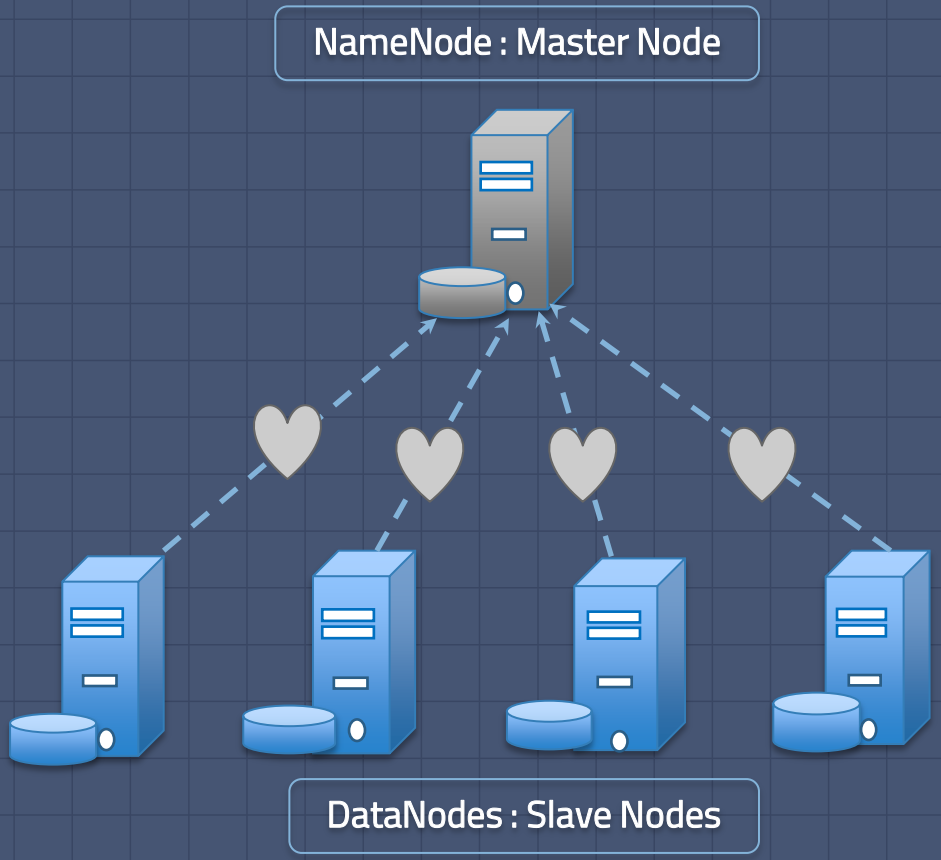


Who manages the data?

Who distributes the data across cluster?

How to access data?

HDFS Architecture



NameNode

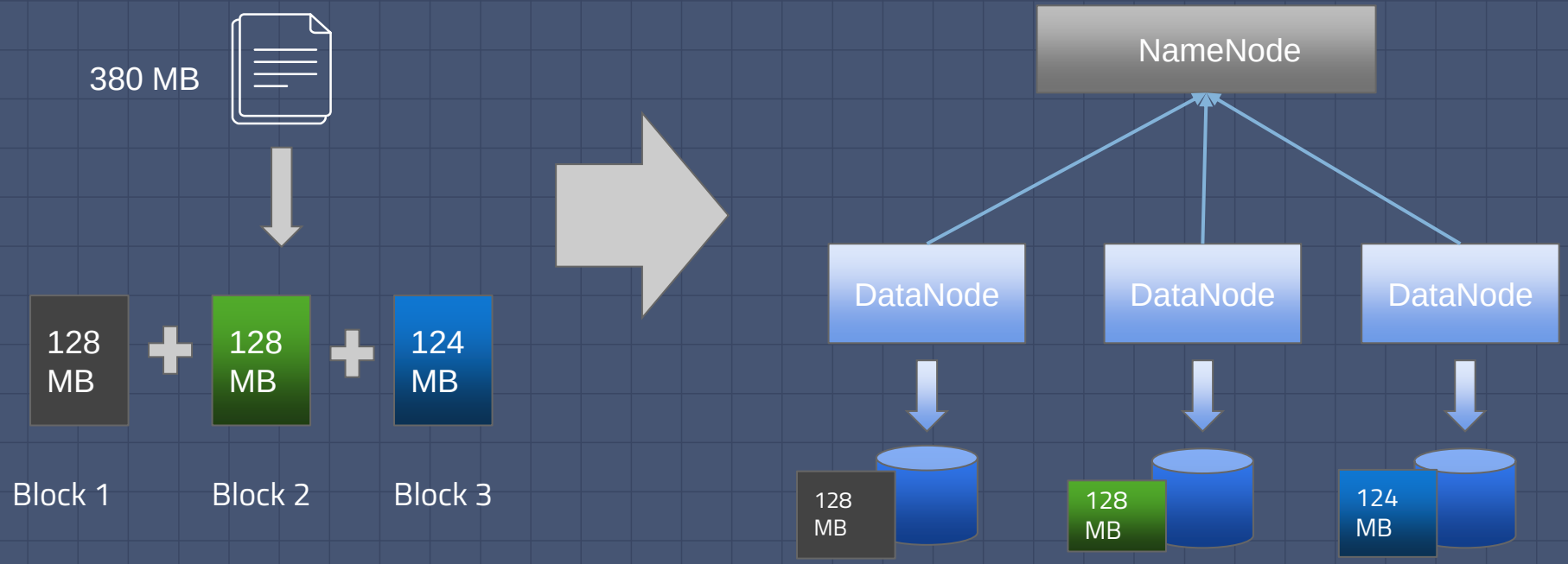
- Master Daemon
- Maintains and manages Datanodes
- Records Metadata
- Receives heartbeat and block report from all datanodes

DataNode

- Slave Daemon
- Stores actual data
- Serves read and write requests from the clients.

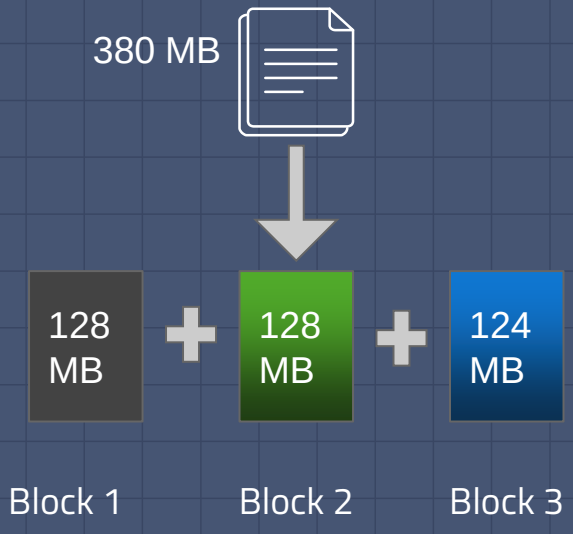
Data in HDFS Data Blocks

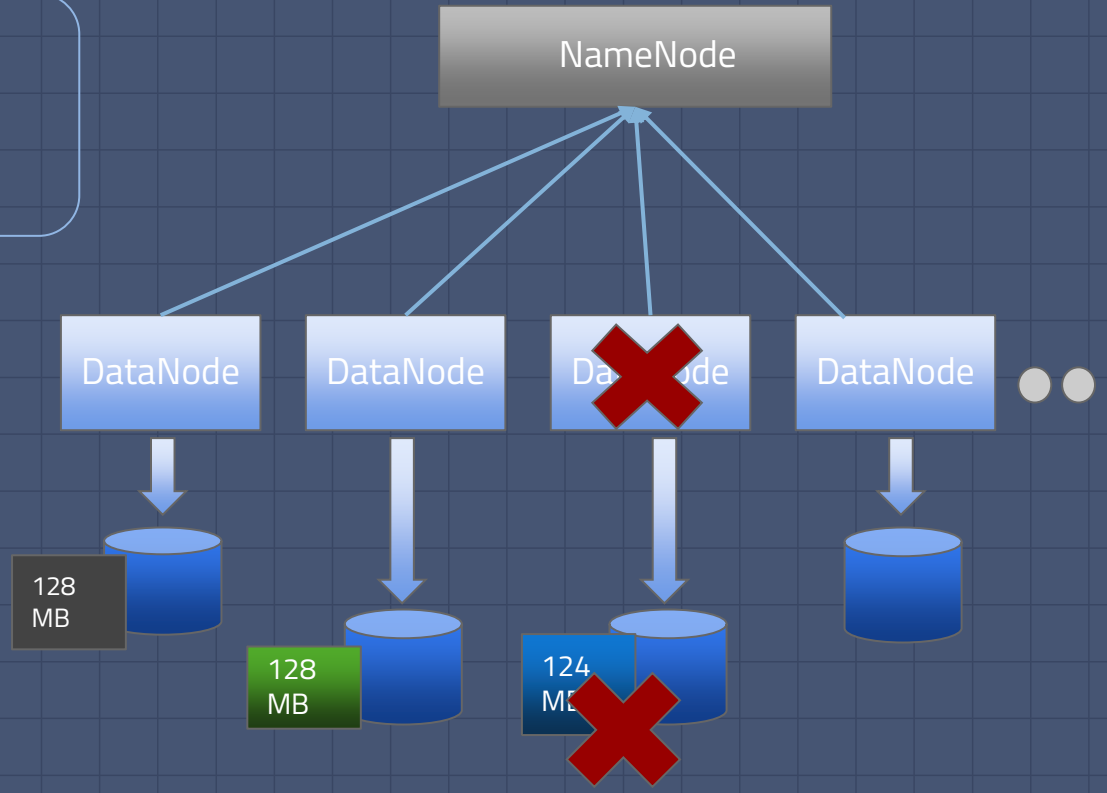
- Each file stored in HDFS is in blocks where each block is 128 MB in Apache Hadoop 2.x
- Files are split into blocks
- Different blocks from the same file will be stored in different machines.



DataNode Failure

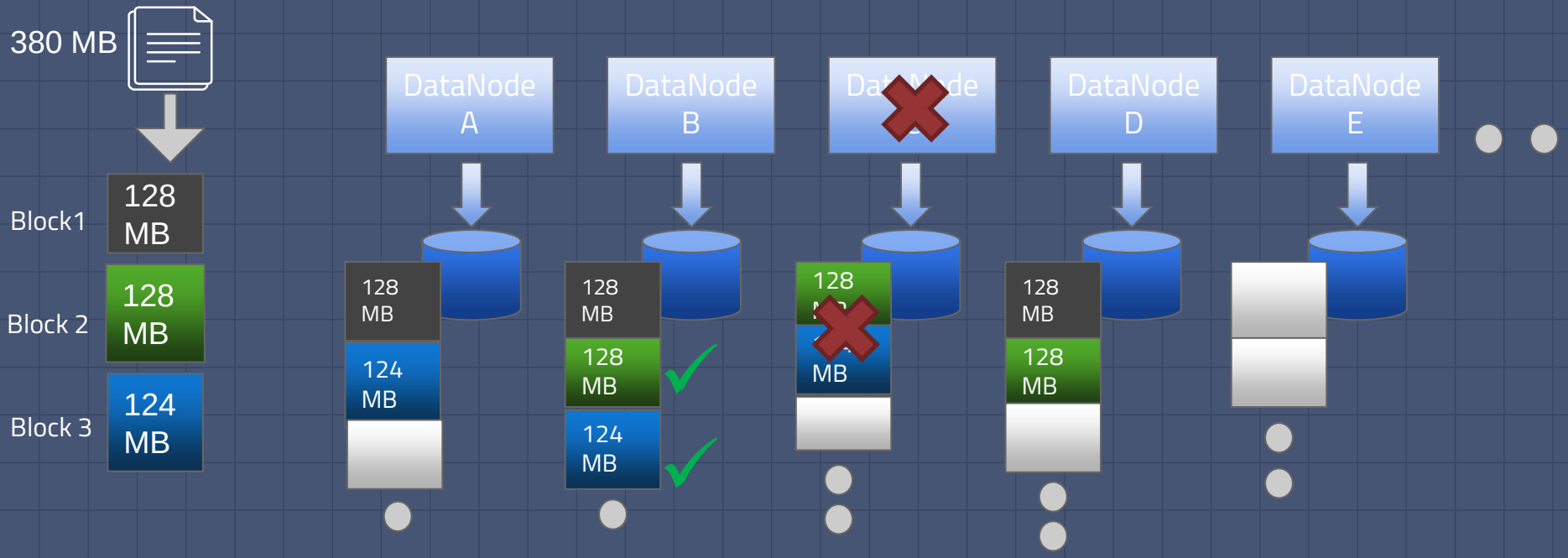
- What if one of the DataNodes crashed?





Replication Factor

- Each data block replicated by default 3 times and distributed across different datanodes



MAP Reduce

- Need of Map Reduce
- “Moving computation to the data”

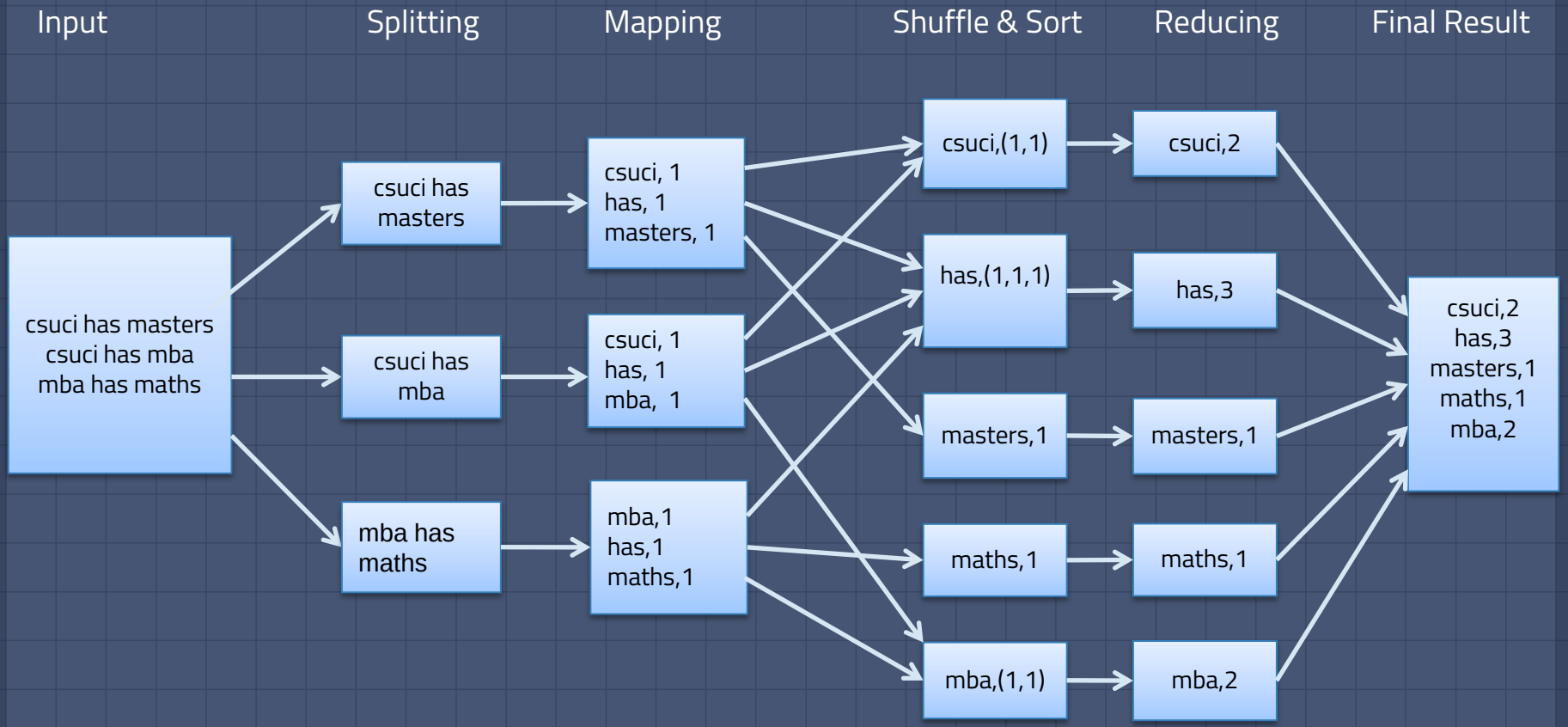
The Mapper

- Reads data as key - value pairs
- Outputs zero or more key – value pairs

The Reducer

- Called once for each unique key
- Gets list of all values associated with a key as input
- The reducer outputs zero or more final key – value pairs
  Usually just one output per input key

Word Count Example



Hadoop Ecosystems

-  Pig : Provides engine for executing data flows with Pig Latin scripting language
-  Hive: Data summarization and ad-hoc query
-  Zookeeper: co-ordination services
-  Flume: Service for collecting, aggregating and moving large amount of data
-  : Workflow scheduler system to manage Apache Hadoop jobs
-  : Provides faster execution with in-memory approach

Who uses Hadoop?

Summary

- Big Data - Enormous amount of data, with different variety is being generated at very high speed
- Hard for traditional processing system to store and process such huge volume of data.
- Solution – Hadoop
- Components of Hadoop - i. HDFS ii. MapReduce
- Tools used in Hadoop framework

Conclusion

- We are staying in the world where everything is connected and generates huge amount of data. Data could aggregate value to society if analyzed well.
- Hadoop is very effective solution to deal with data in petabytes
- Why Hadoop?
  - network bandwidth and seek latency
- Why Map Reduce programming model?
  - Parallel programming
  - large data sets
  - moving computation to data

References

1. O'REILLY Hadoop The Definitive Guide
2. Apache Hadoop!
3. Overview of Big Data and Hadoop for secure healthcare system (https://link.springer.com/article/10.1007/s40747-017-0040-1)
4. https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-014-0019-z
5. https://www.dezyre.com/hadoop-wiki
6. A Literature review on Hadoop Ecosystems (https://link-springer-com.summit.csuci.edu/content/pdf/10.1007%2F978-981-10-8360-0_22.pdf)

Thank you !

Questions?