



Introduction to Big Data

October 2020

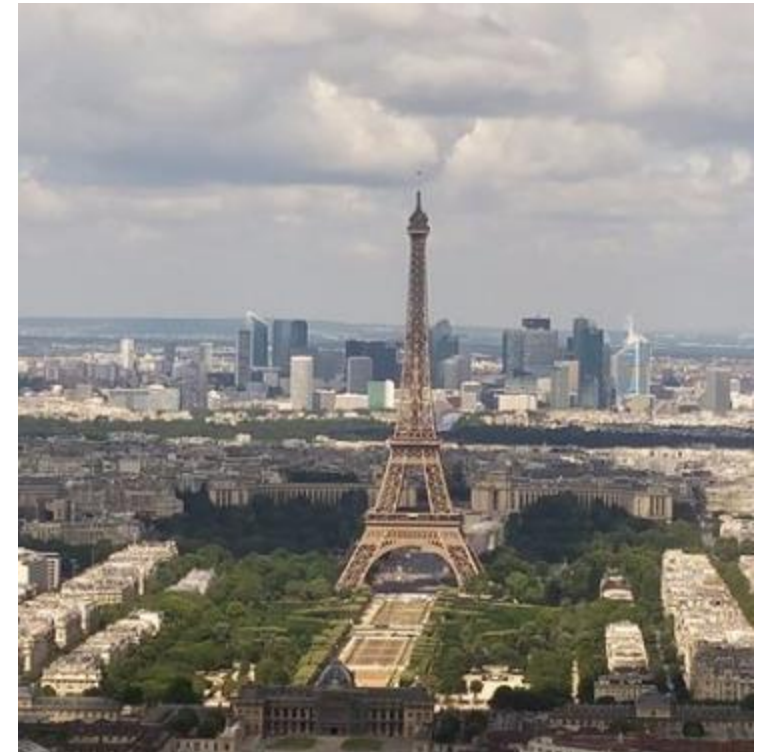
Brolinskyi Sergii



Plan of presentation

- Personal introduction
- Big Data definition and origin
- Big Data characteristics
- Common scenarios
- Real world examples
- Big Data technologies
 - Spark
 - Hadoop
 - Data Lake
- HDFS
- Map Reduce (quick introduction)
- Potential future of Big Data
- Summary

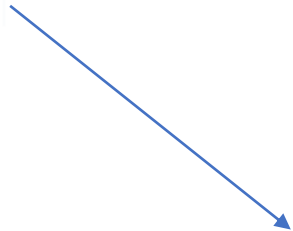
Brolinskyi Sergii



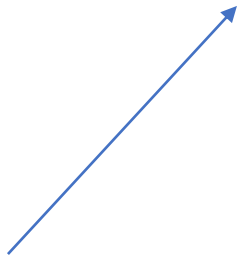
Projects



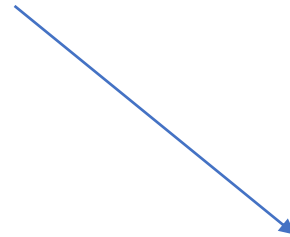
Edge



Power apps




Iris Studio



Movies & TV

Microsoft Store

Home Gaming Entertainment Productivity Deals Microsoft



Yifei Liu
Actress, Director
8/25/1987 (33 years old)

Alberto E. Rodriguez

Job Summary

Preparing Queued Running Done

52 seconds 30 seconds 70,89%

Job Result N/A
Total Duration N/A
Total Compute Time 17,1 minutes

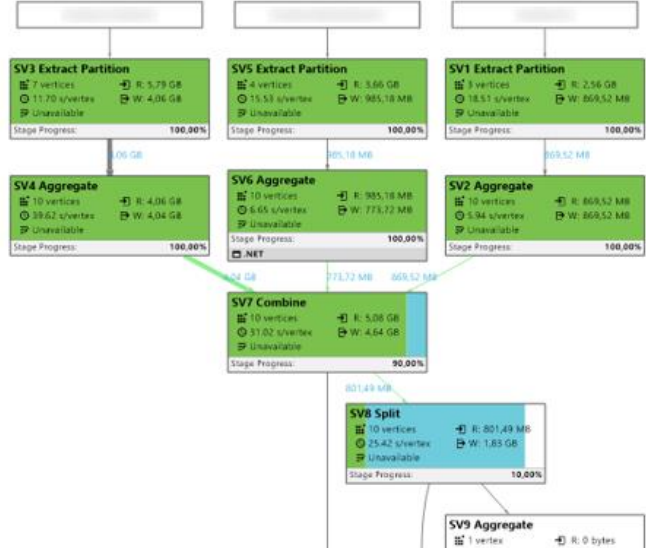

Submit Time 23/01/2020 14:40:22
Start Time 23/01/2020 14:41:23
End Time N/A

Compilation 52 seconds
Queued 30 seconds
Running 2,4 minutes

Account
Author
Runtime soy_yarnpp_release_11ba0f36_201
Priority 1000
Root Process Id 34074462-09d9-453d-af43-c0dbde8f
Application Id application 157973381404? 9577

Job Graph Data State History AU Analysis Diagnostics(4)

Display: Progress Succeeded Retried Failed Running Waiting Search stage





Top deals: Fortnite + Wolverine movies up to 45% off

Fortnite players on Xbox get a \$5 Gift Card with purchase of any featured Wolverine movie

Trailers 360° videos Movies TV

New movies Show all 99+ results



Movies & TV

Data



More data



Even MORE data

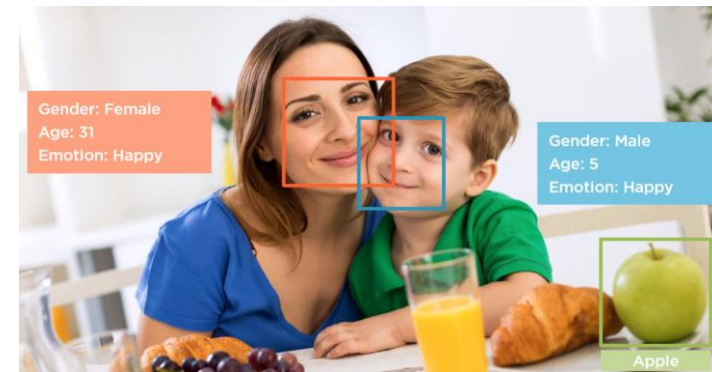


Big Data



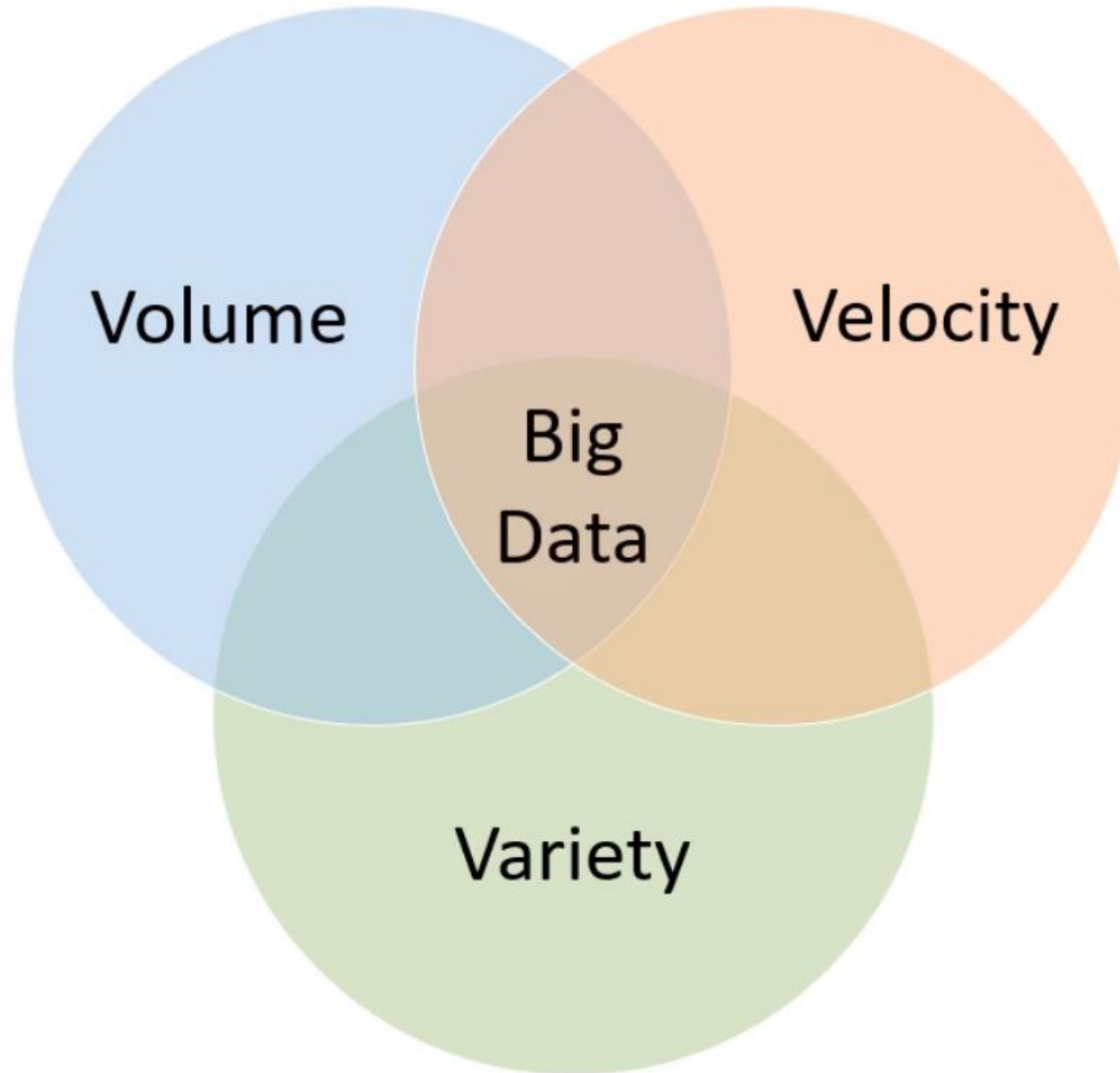
Big Data characteristics

- Volume (10-100 TB of data)
- Variety (structured, semi-structured, unstructured sources)



- Velocity (daily batch is not enough)

Big Data characteristics

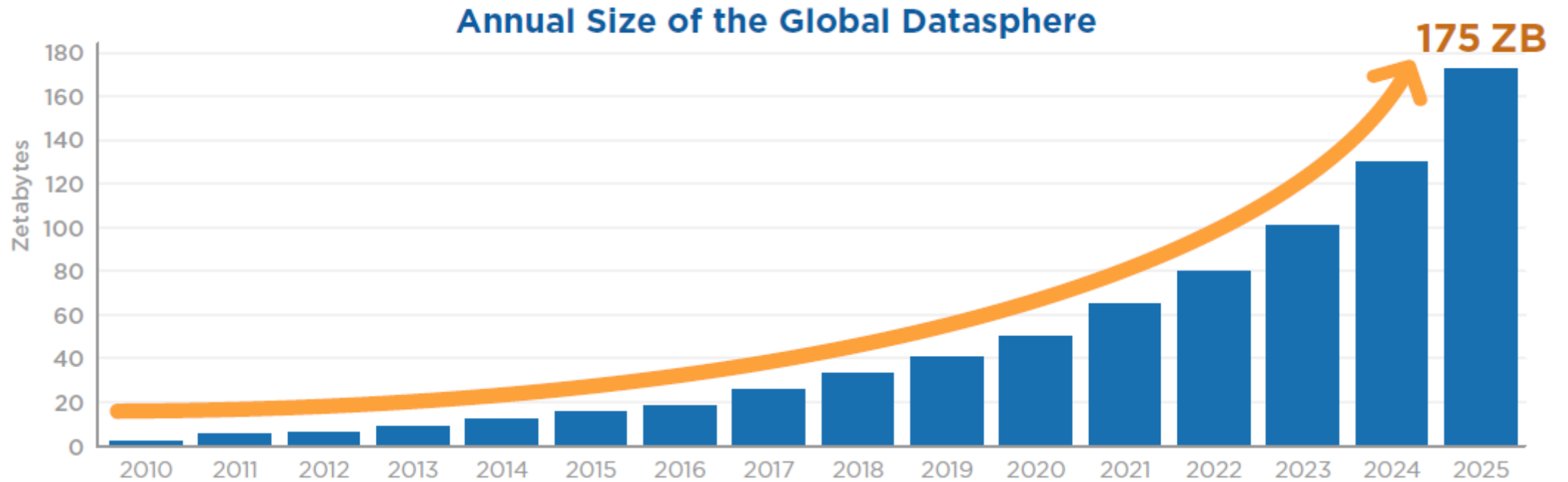


Big Data

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

Data volume is growing exponentially

Figure 1 - Annual Size of the Global Datasphere



Common data producers



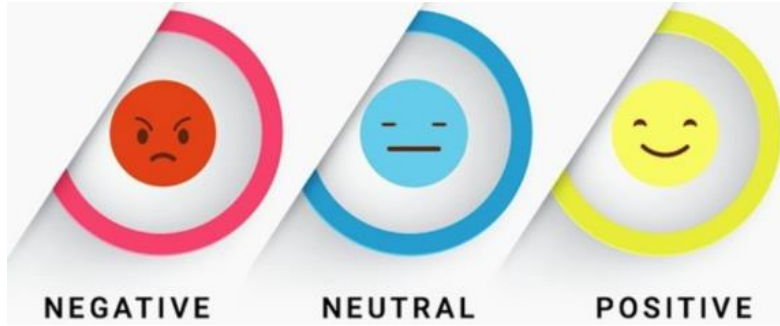
We always had the data but only now we store and analyze it

Example of Big Data in real life



Beer and diapers are often bought together so supermarkets have them close to each other

Example of Big Data in real life



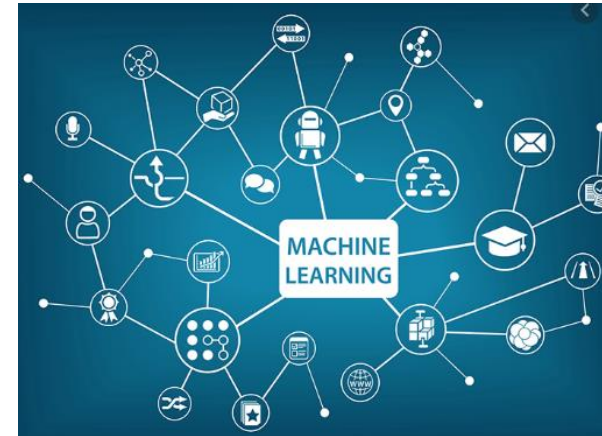
Sentiment analysis



Fraud detection



Forecast



Machine learning

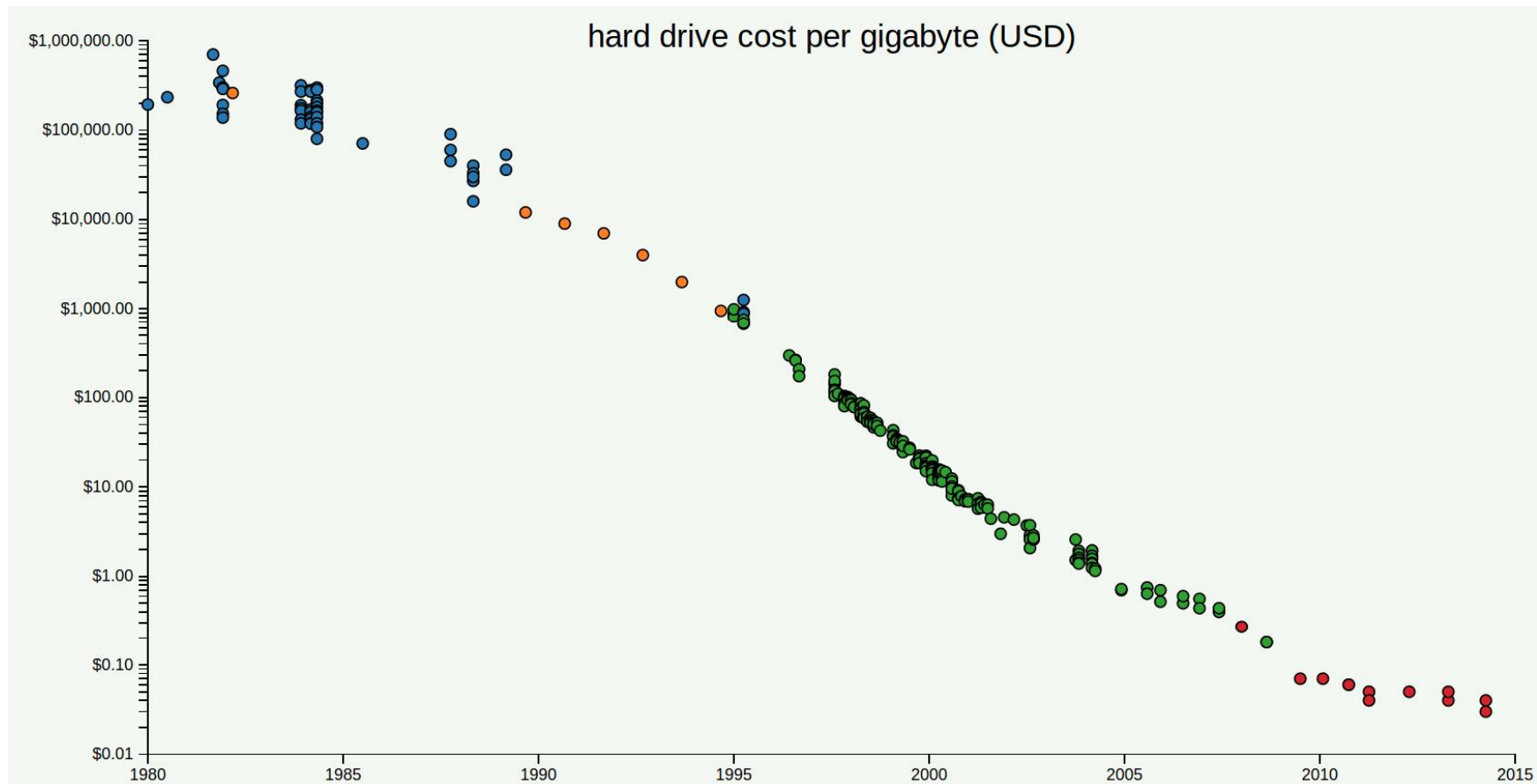
Big Data enablers

Storage getting cheaper

New distributed programming technologies

Web economy

Cloud development



Distributed computing technologies

Spark



Hadoop



Azure Data
Lake



Hadoop



- An Apache project that combines MapReduce engine and a distributed file system (HDFS)
- An Open-Source implementation of Google's MapReduce and Google's distributed file system (GFS)
- Hadoop is typically used in a combination with other technologies and most often Java (those technologies are often referred as a Hadoop stack)

Hadoop stack



Hadoop

- MapReduce
- HDFS

Database

- Hbase
- Cassandra

Query

- HiveQL
- Pig Latin

SQL to Hadoop

- Sqoop

Machine Learning

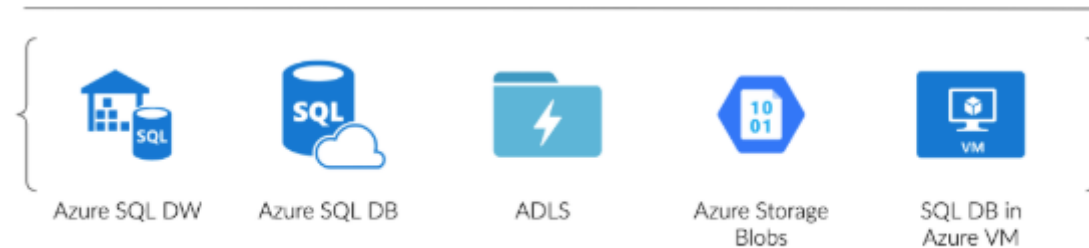
- Mahout

Azure Data Lake



- Data Lake concept - cheap storage of large amounts of unstructured data
- Microsoft owned cloud-based solution that implements HDFS interface
- Data Lake is typically used in a combination with other technologies and most often C# that are part of MS Big Data stack

Azure Data Lake Analytics

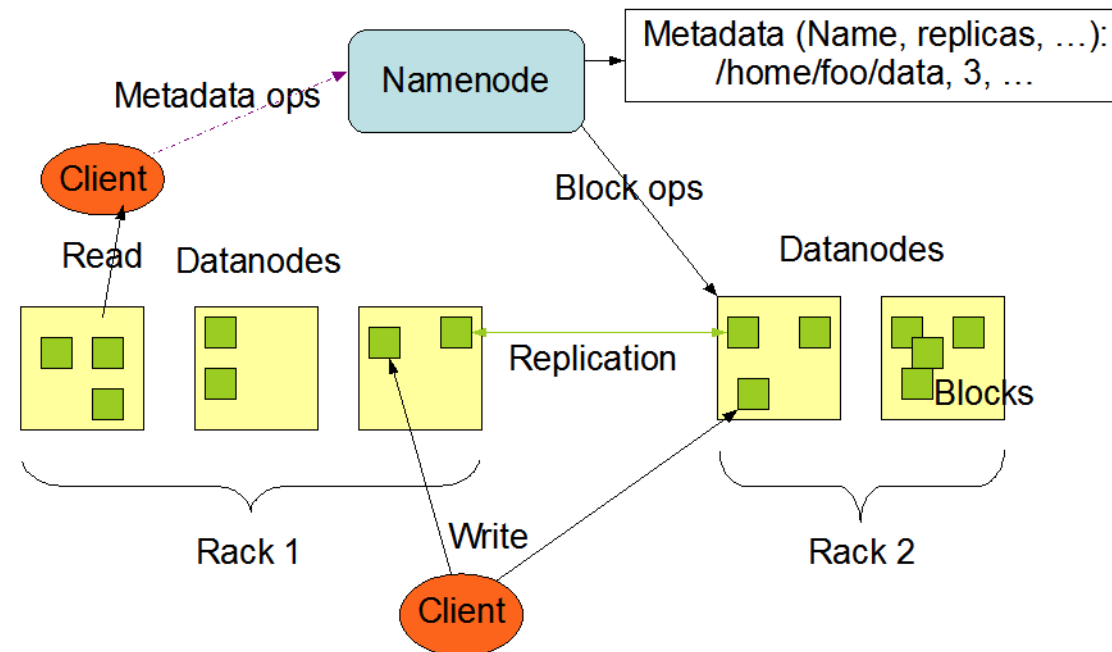


HDFS



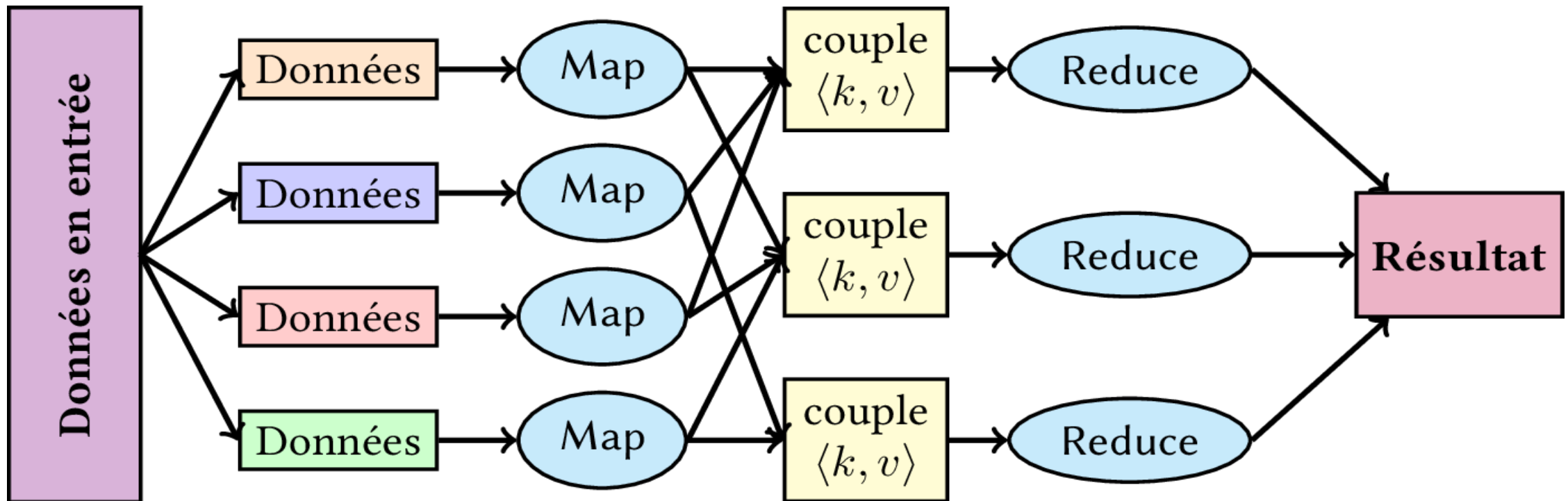
- HDFS – Hadoop Distributed File System
- Files are stored and replicated across different nodes so loss of some of them is not critical and would not cause a data loss
- Files cannot be updated (the file entry can be updated with a new location)

HDFS Architecture



MapReduce (quick intro)

- Basically 2 steps and works with Key-Value format of data
 - Map - split the data and preprocess it
 - Reduce - aggregate the results.

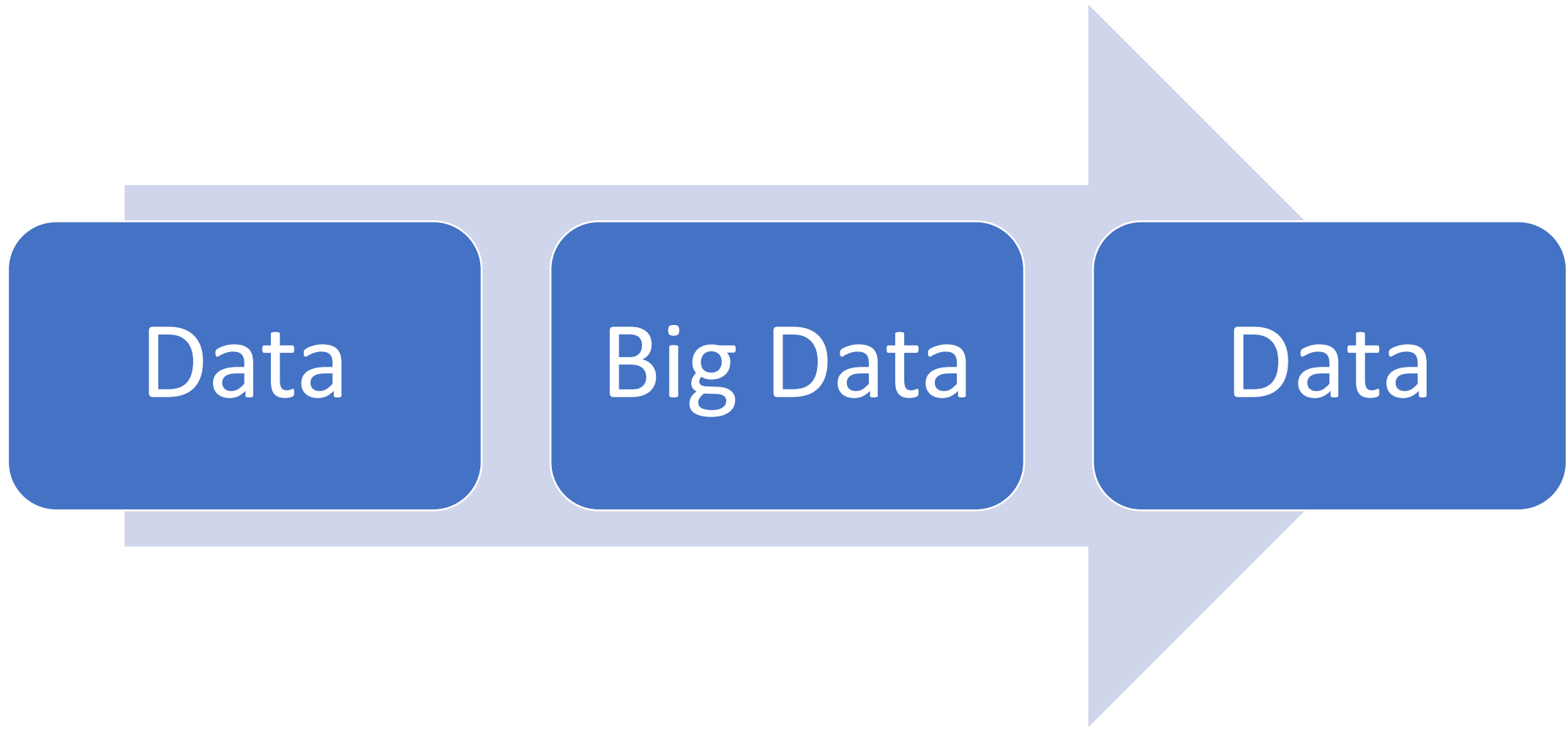


MapReduce (example)

It is the beginning of the year and students want to make a contact list of new profs. Each student is assigned to find a prof by his name in the university and then put them into the group chat.

- Key = professor name and basic information
- Value = professor contact data

Future of Big Data



Summary

Small data when growing becomes big data and can't be processed withing one computer so there is a set of technologies that allow to stock and analyze those data of big volume, velocity and variety. Those technologies form the Big Data stack.

