Decoding Big Data: Unveiling Insights and Applications for the Modern Era

Presented By

Sanjeev Kumar

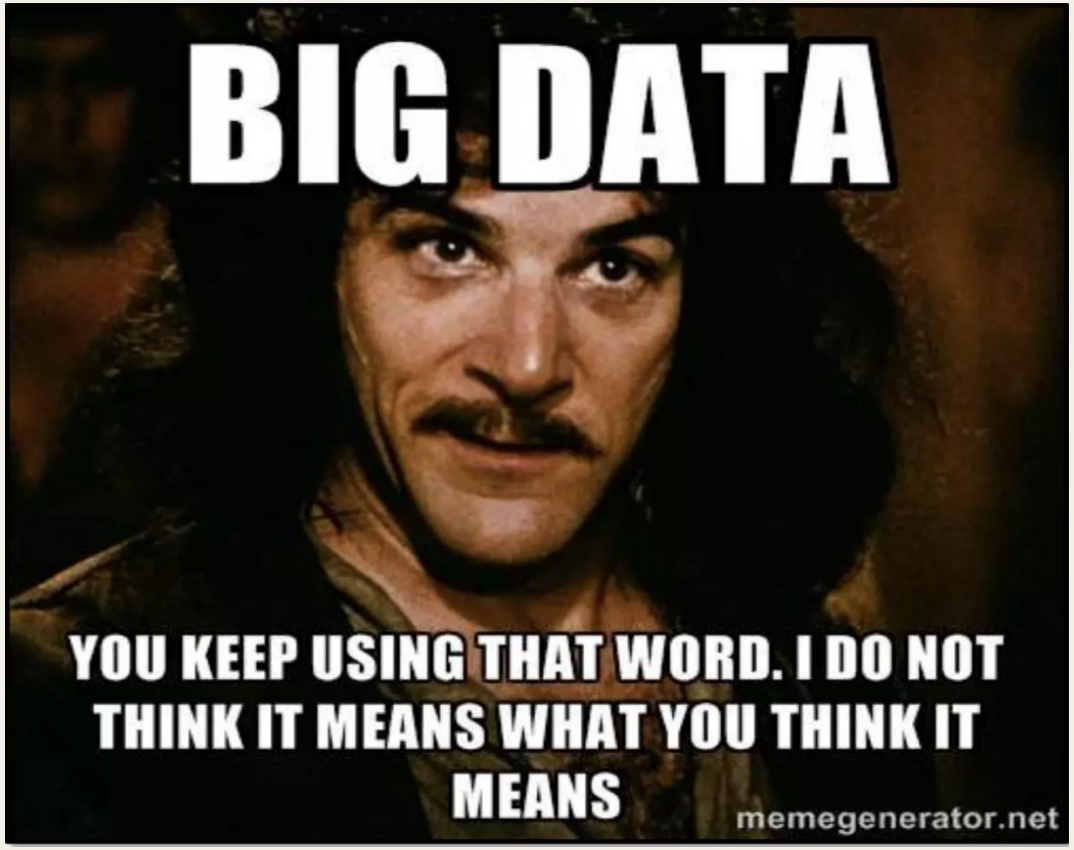
Research Scholar

Department of Computer Science & Engineering

IIT Bombay, Mumbai, India

www.cse.iitb.ac.in/~sanjeev

Defining Big Data



https://knowyourmeme.com/memes/you-keep-using-that-word-i-do-not-think-it-means-what-you-think-it-means

https://www.siliconrepublic.com%2Fjobs%2Fcareer-memes-of-the-week-data-scientist

Defining Big Data

~~Boring~~ Traditional Definition

" In the realm of abundant data, characterized by high volume, velocity, and variety, there arises a need for cost-effective and innovative information processing methods to elevate insights and decision-making capabilities. "

Defining Big Data

Oxford Dictionary Definition

" sets of information that are too large or too complex to handle, analyze or use with standard methods "

Defining Big Data

Oxford Dictionary Definition

" sets of information that are too large or too complex to handle, analyze or use with standard methods "

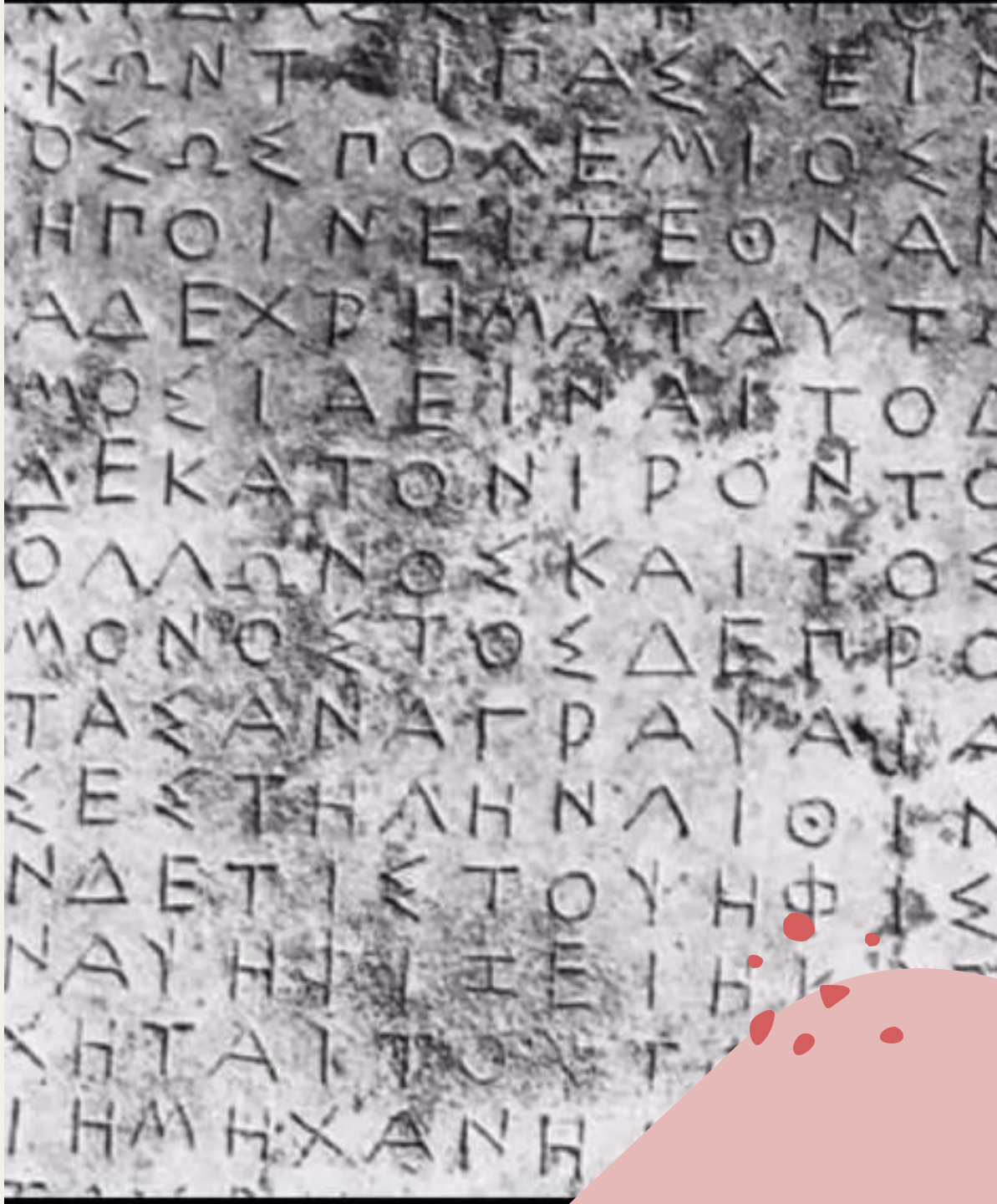
Big Data

- Big data is never about the **size** of the data, it's all about the **value** within the data.

The "Big Data"

1880 US Census

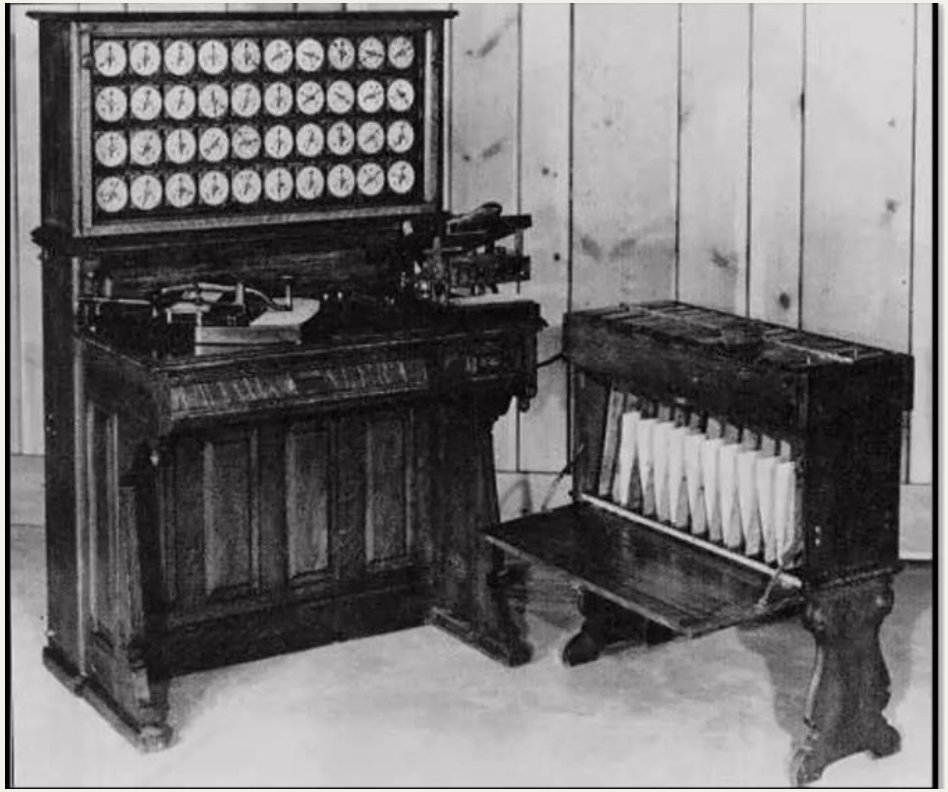
- **50 millions people**
- **Data included: age, gender, no of household**
- **Took 7 year to tabulate**



The "Big Data"

1890 US Census

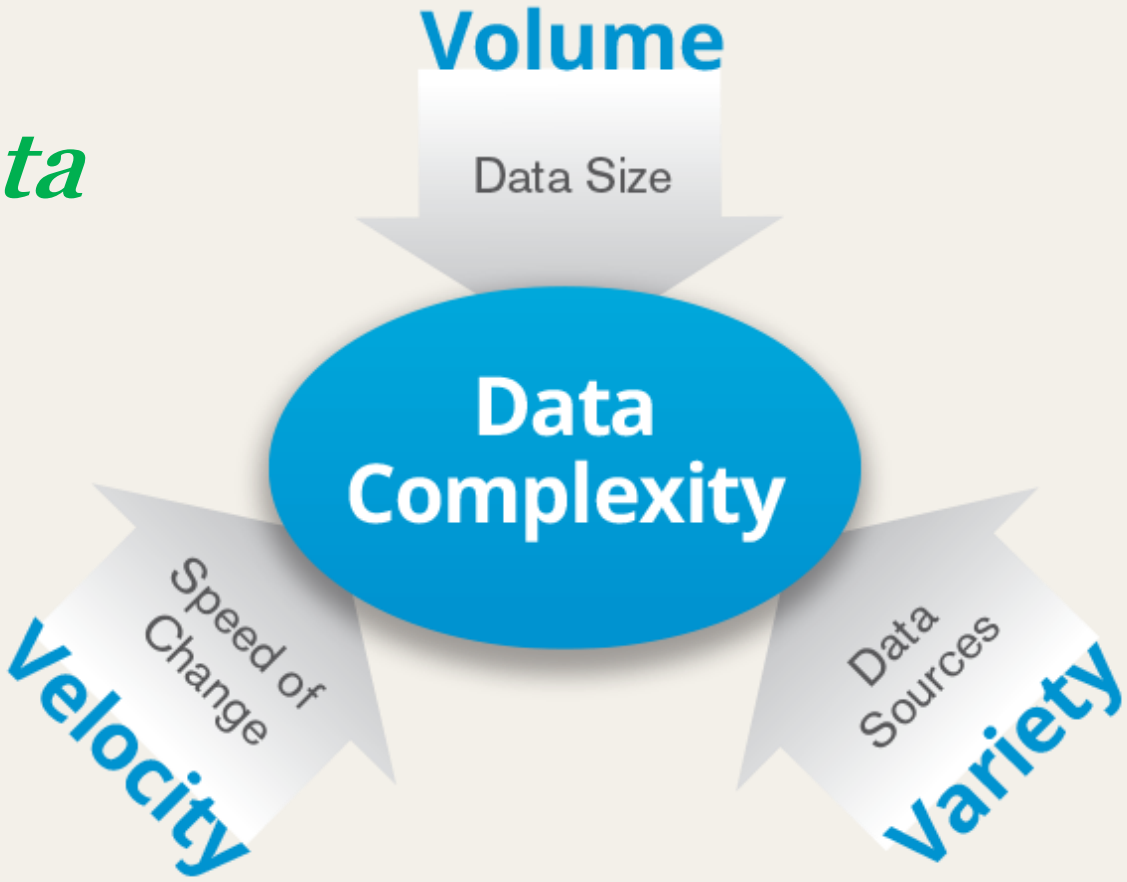
- **65 millions people**
- **Data included: age, gender, no of household, military, citizenships**
- **New technology: Hollerith Tabulating System**
- **Took 6 weeks to tabulate (76x faster)**



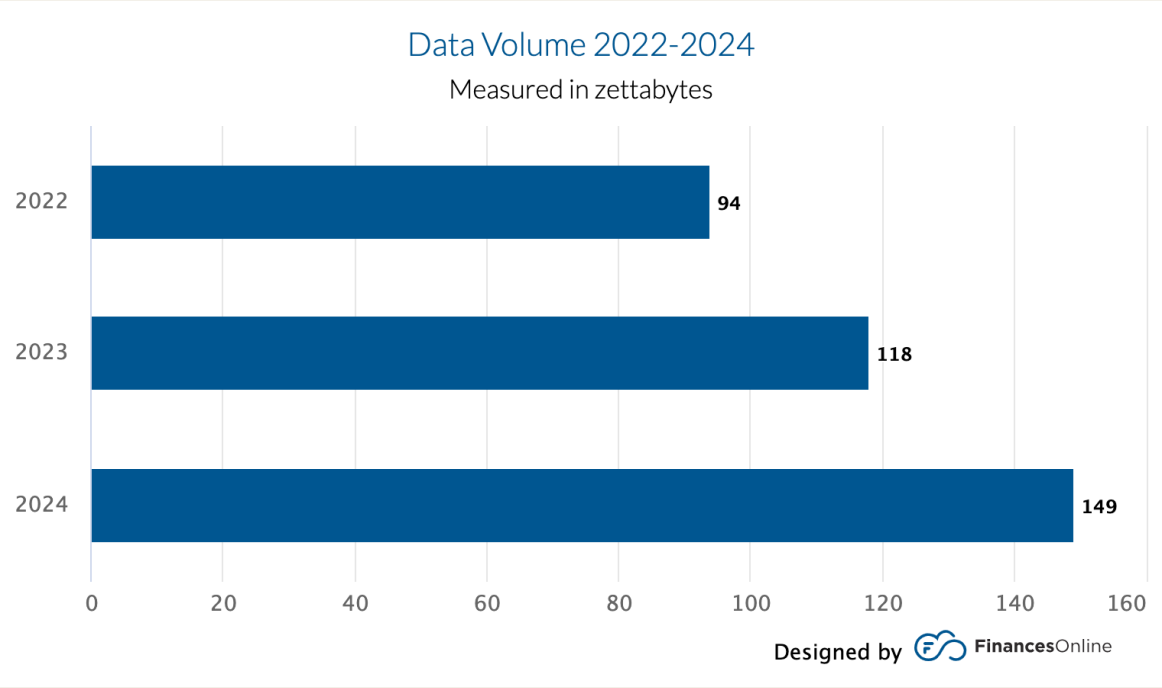
Takeaways:

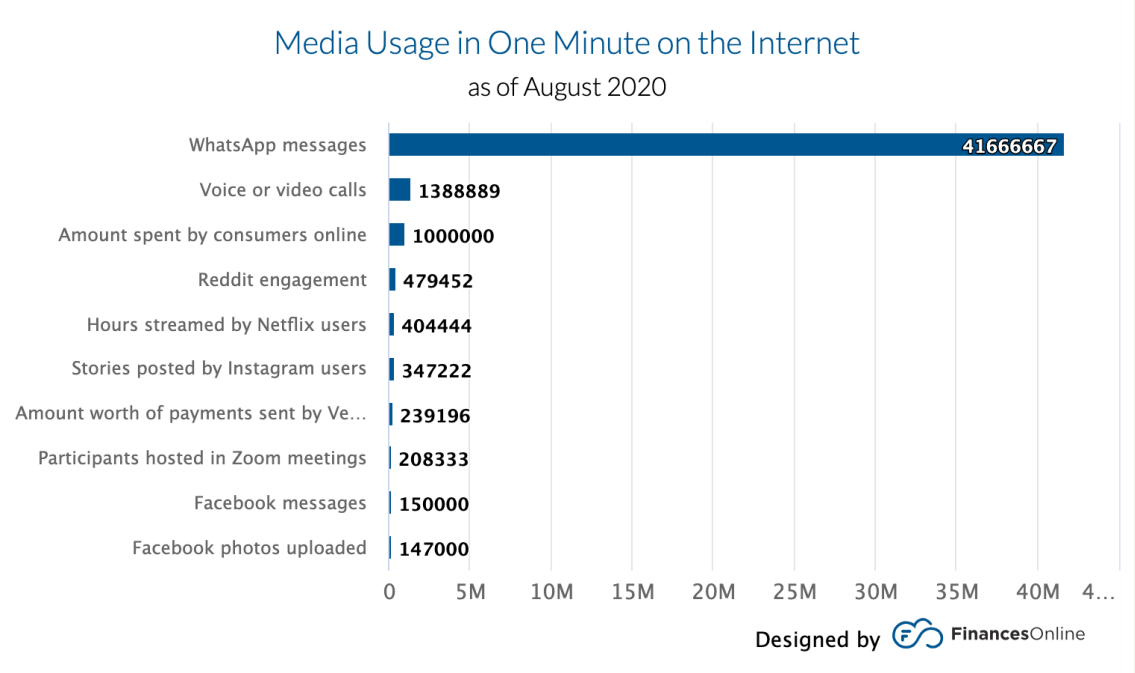
- **Better technology and methodology leads to 76x speedup**

Factors of Big Data



The volume of Big Data





https://financesonline.com/how-much-data-is-created-every-day/

The velocity of Big Data

2 TRILLION
searches on Google by the end of 2021

1.134 TRILLION MB
volume of data created every day

3,026,626
emails sent every second, 67% of which are spam

278,108 PETABYTES
global IP data per month by the end of 2021

230,000
new malware versions created every day

82%
share of video in total global internet traffic at the end of 2021

41,666,667
messages shared by WhatsApp users

1,388,889
video / voice calls made by people worldwide

404,444
hours of video streamed by Netflix users

347,222
stories posted by Instagram users

150,000
messages shared by Facebook users

147,000
photos shared

https://financesonline.com/how-much-data-is-created-every-day/

The Variety of Big Data

Structured

Most traditional data sources
Tabular data Eg: csv, tsv

Semi-structured

XML, JSON

Unstructured

FB logs, Chats, Audio Signals, Videos

The Statistics of Big Data

Big Data is going to be worth $229.4 billion by 2025. (Strategic Tech Investor, 2021)

According to another prediction, there will be 43 billion IoT-connected devices. (McKinsey & Company, 2019)

463 ZB of data will be created every day by 2025. (Raconteur, 2020)

There could be 2 trillion searches on Google by the end of 2023. (Internet Live Stats, 2021)

47 million stories with the Support Small Business Sticker were created on Instagram in the last quarter of **2023**. (Facebook, 2020)

People sent 500 million tweets daily. (TechJury, 2020)

A connected car produced 4 TB of data in one day. (Raconteur, 2020)

The internet population in 2023 will be 66% of the world's total population. (Cisco, 2020)

What to do with these data?

Aggregation & Statistics	Dataware house & OLAP
Indexing, Searching & Querying	Keyword based search, Pattern Matching
Knowledge Discovery	Data Mining, Statistical Modeling

Big Data Analytics

Examining large amount of data.

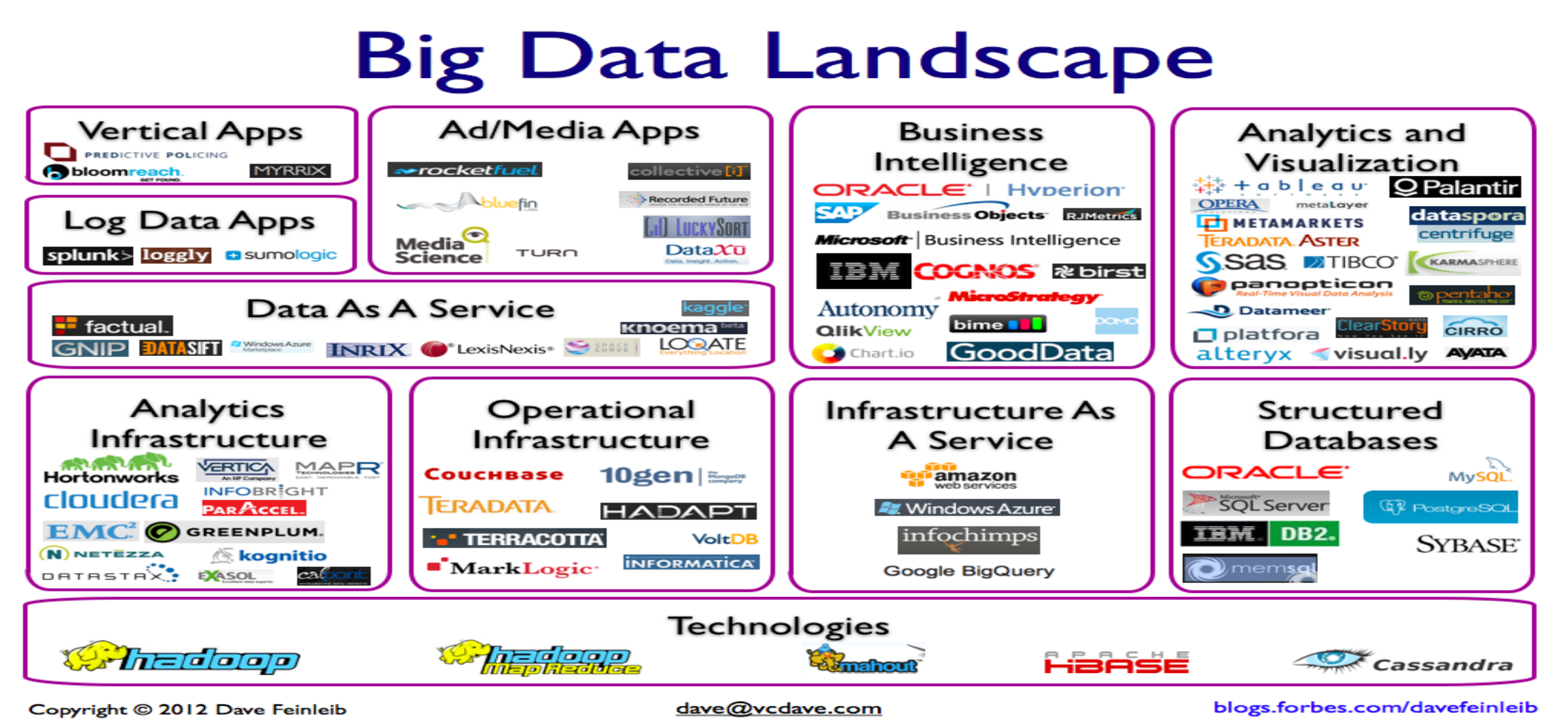
Identifies hidden information, knowledge discovery, correlation.

Better business decision: strategic, operational

Effective marketing, customer satisfaction, increased revenue.

Competitive advantages

Big Data Landscape

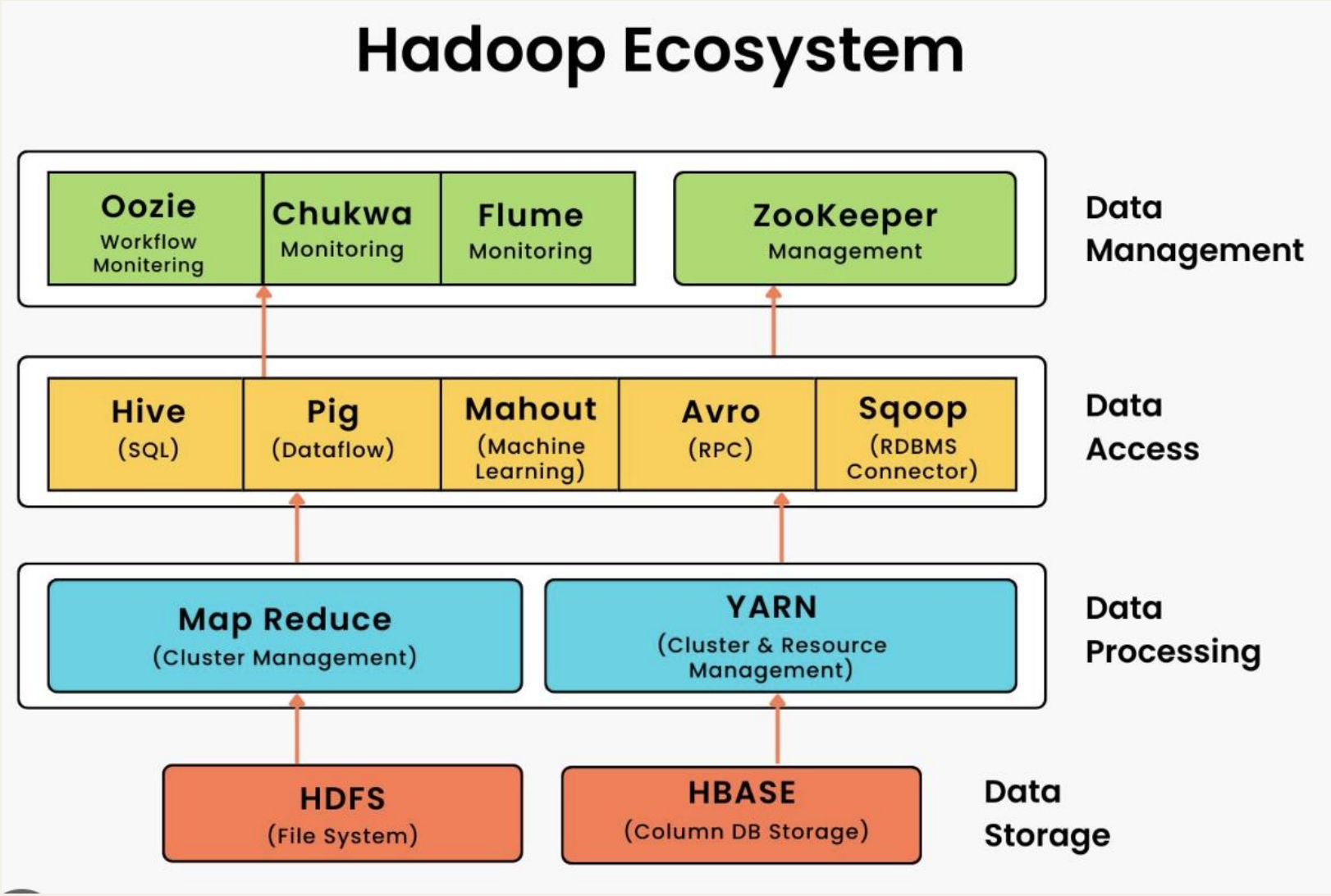


https://www.girikon.com/blog/big-data/

Hadoop & Big Data

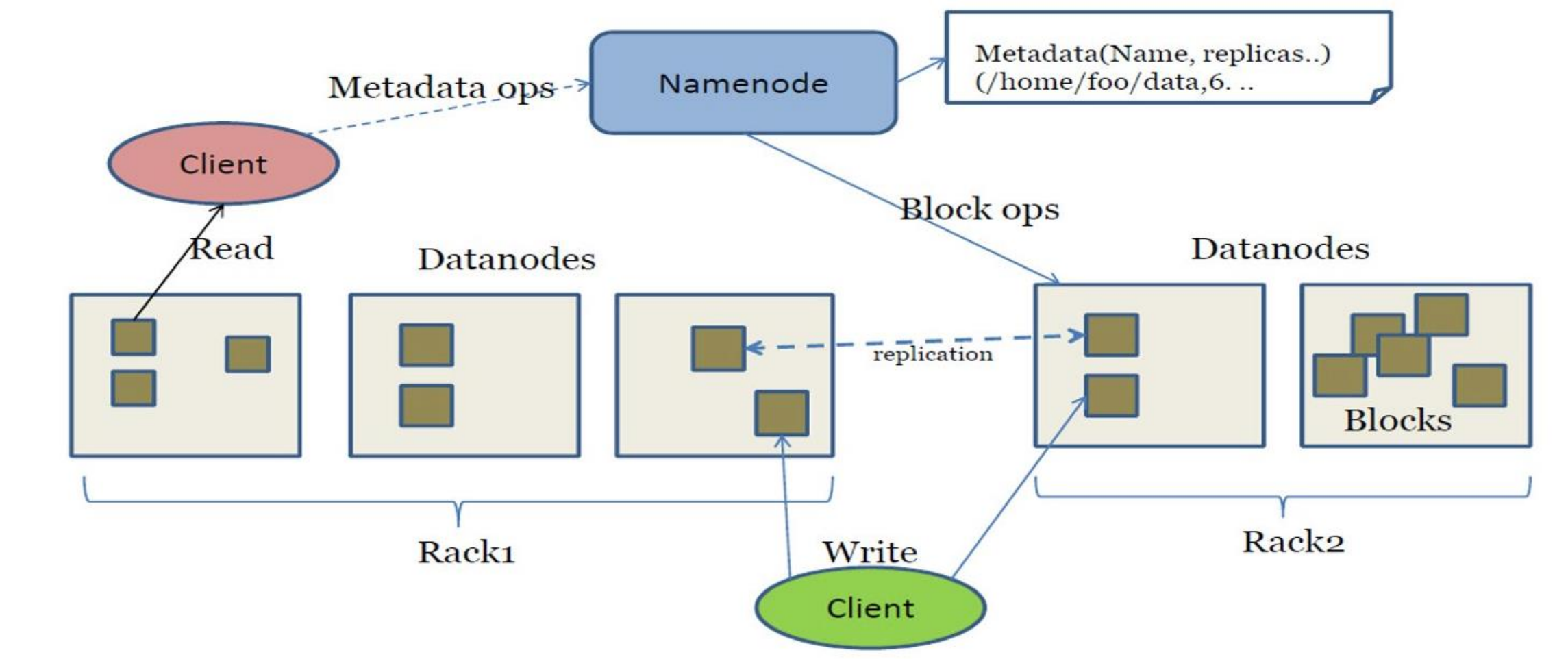
- Won't fit into single computer.
- Distributed data
- Faster computation
- Resource allocator
- Storage

Hadoop



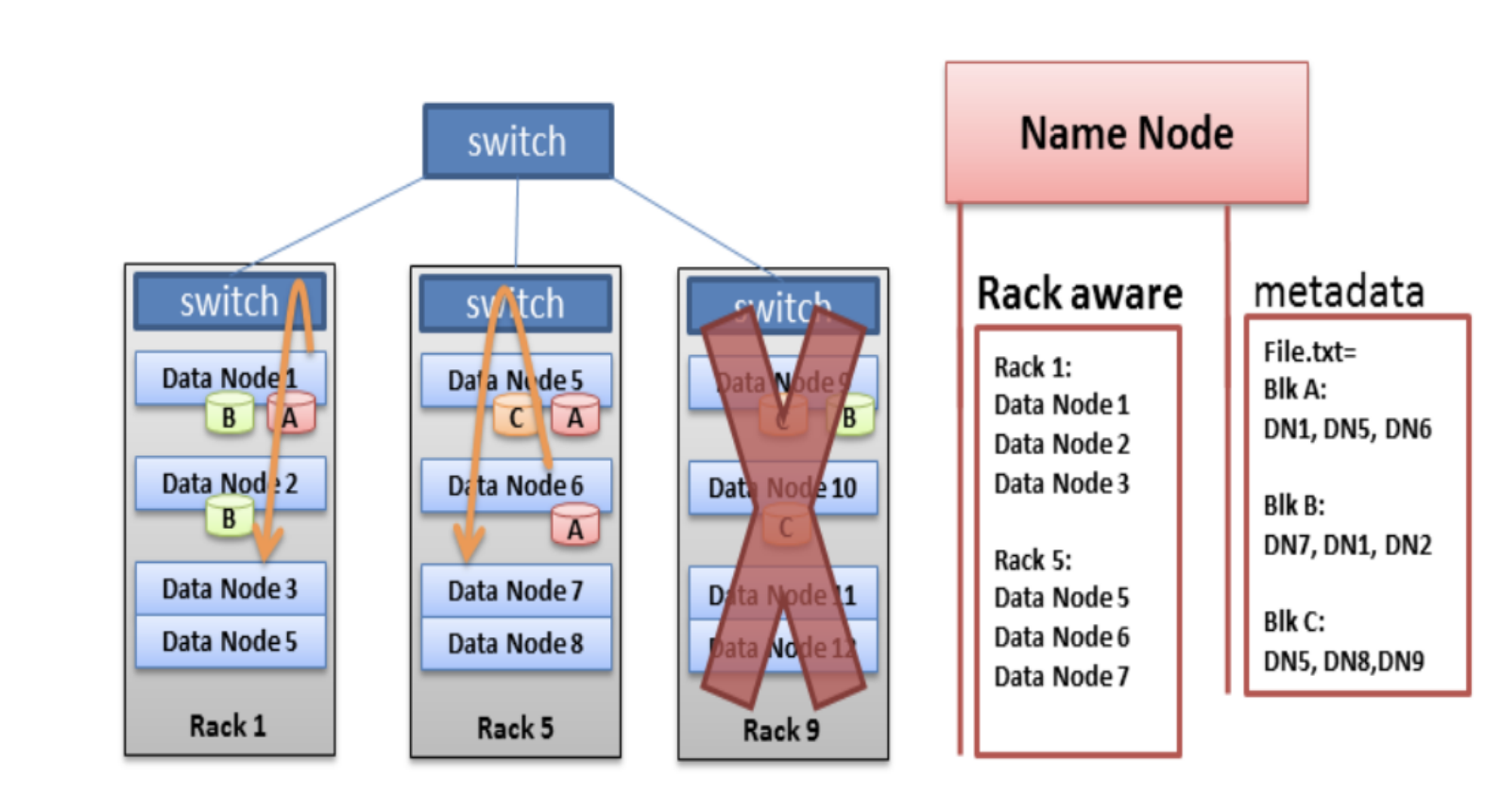
https://www.turing.com/kb/hadoop-ecosystem-and-hadoop-components-for-big-data-problems

Hadoop Architecture



https://searchenginedeveloper.wordpress.com/2013/09/18/hadoop-distributed-file-system/

Hadoop Rack Awareness – Why?



- Never loose all data if entire rack fails.
- Keep bulky flows in-rack when possible.
- Assumption that in-rack is higher bandwidth, lower latency.

Map-Reduce

Map reduce is a programming for processing and generating large datasets.

Map reduce was used to completely regenerate the Google's index of WWW.

Hadoop allows applications to run using the map reduce algorithm.

Users implement interface for 2 functions

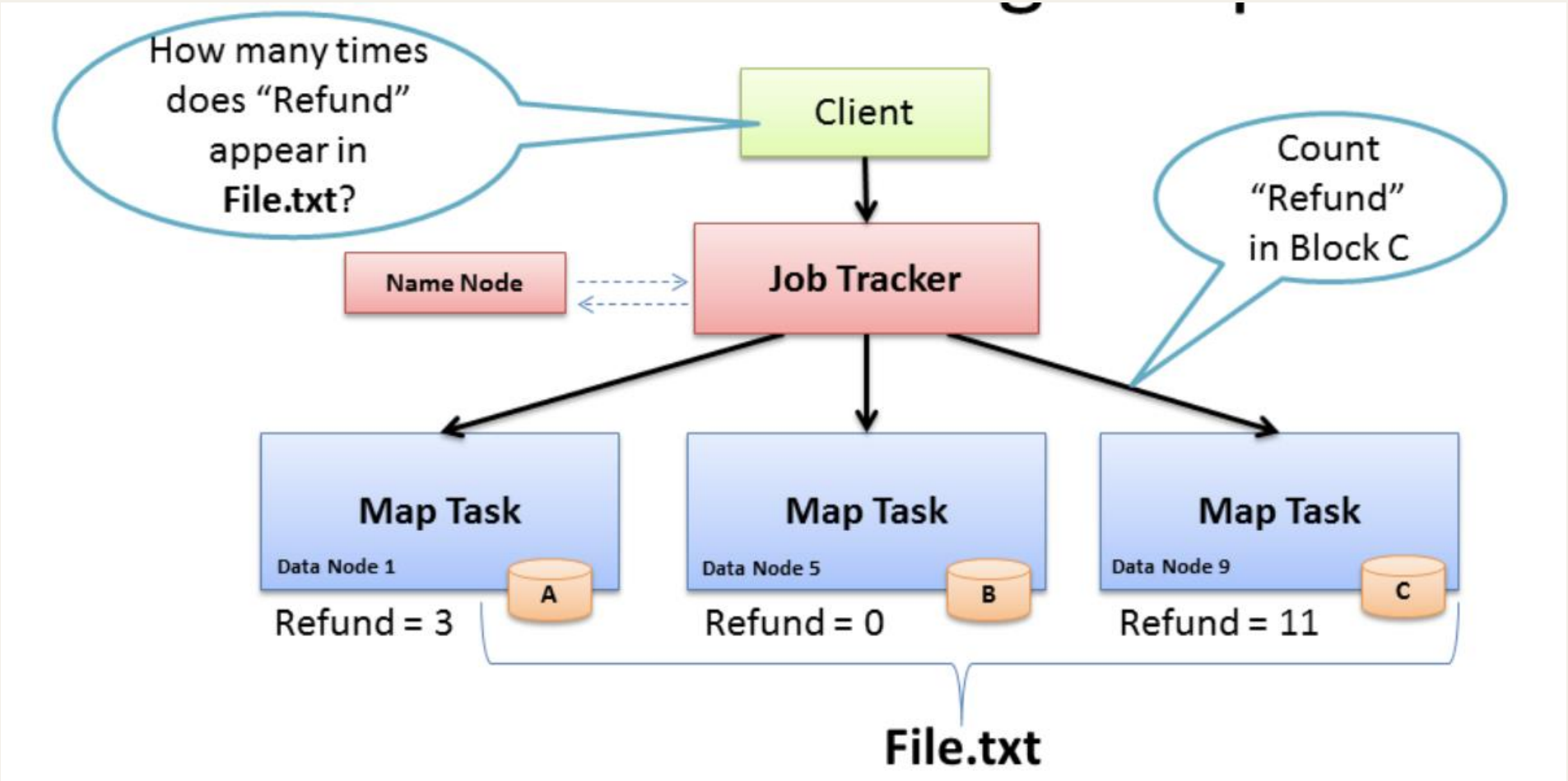
Map

Reduce

Map (in key, in value) --> (out key, intermediate value) list

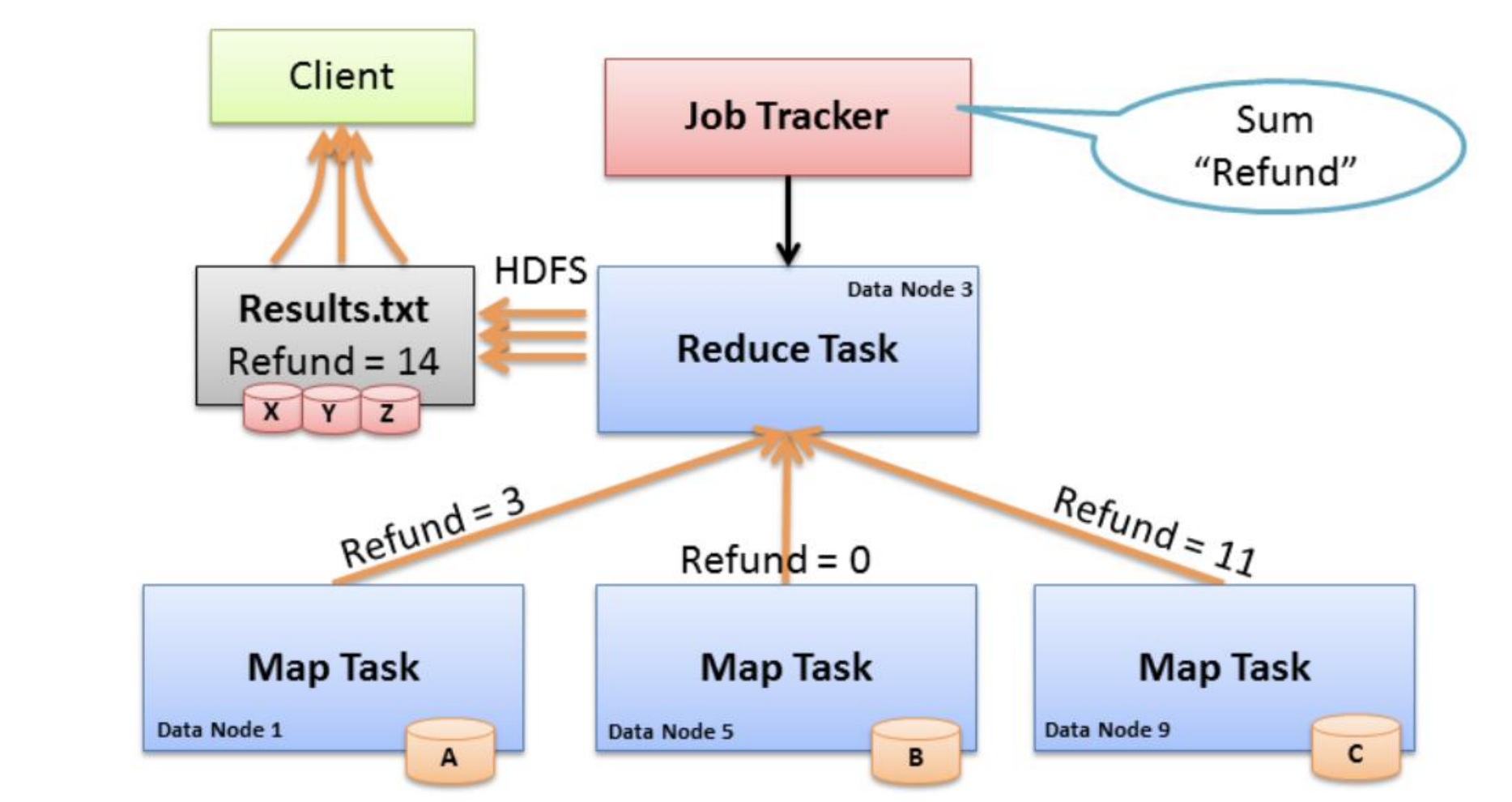
Reduce (out key, intermediate value)--> out_value list

Map



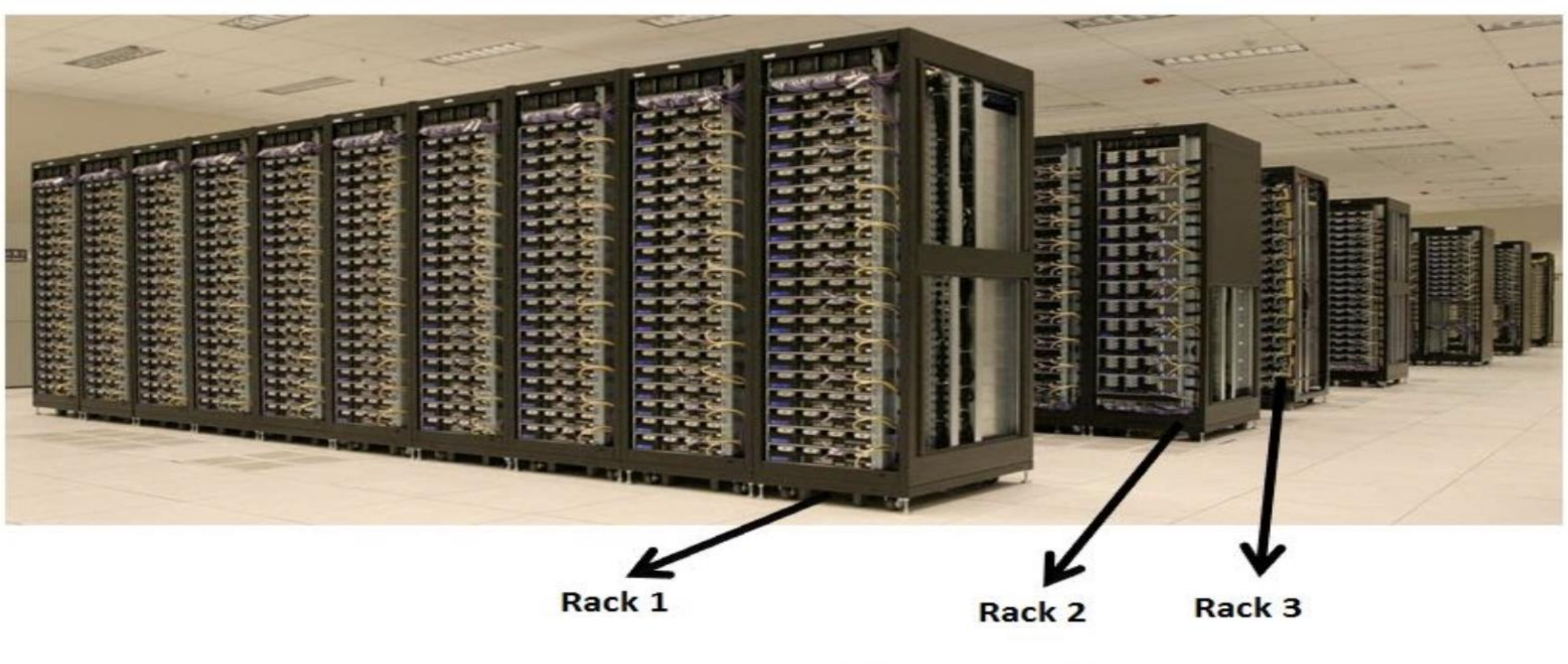
Map: Run this computation on local data

Reduce



Reduce: Run this computation across map results.
Map tasks send output data to Reducer over the network.
Reduce Task data output written to and read from HDFS.

Real time example



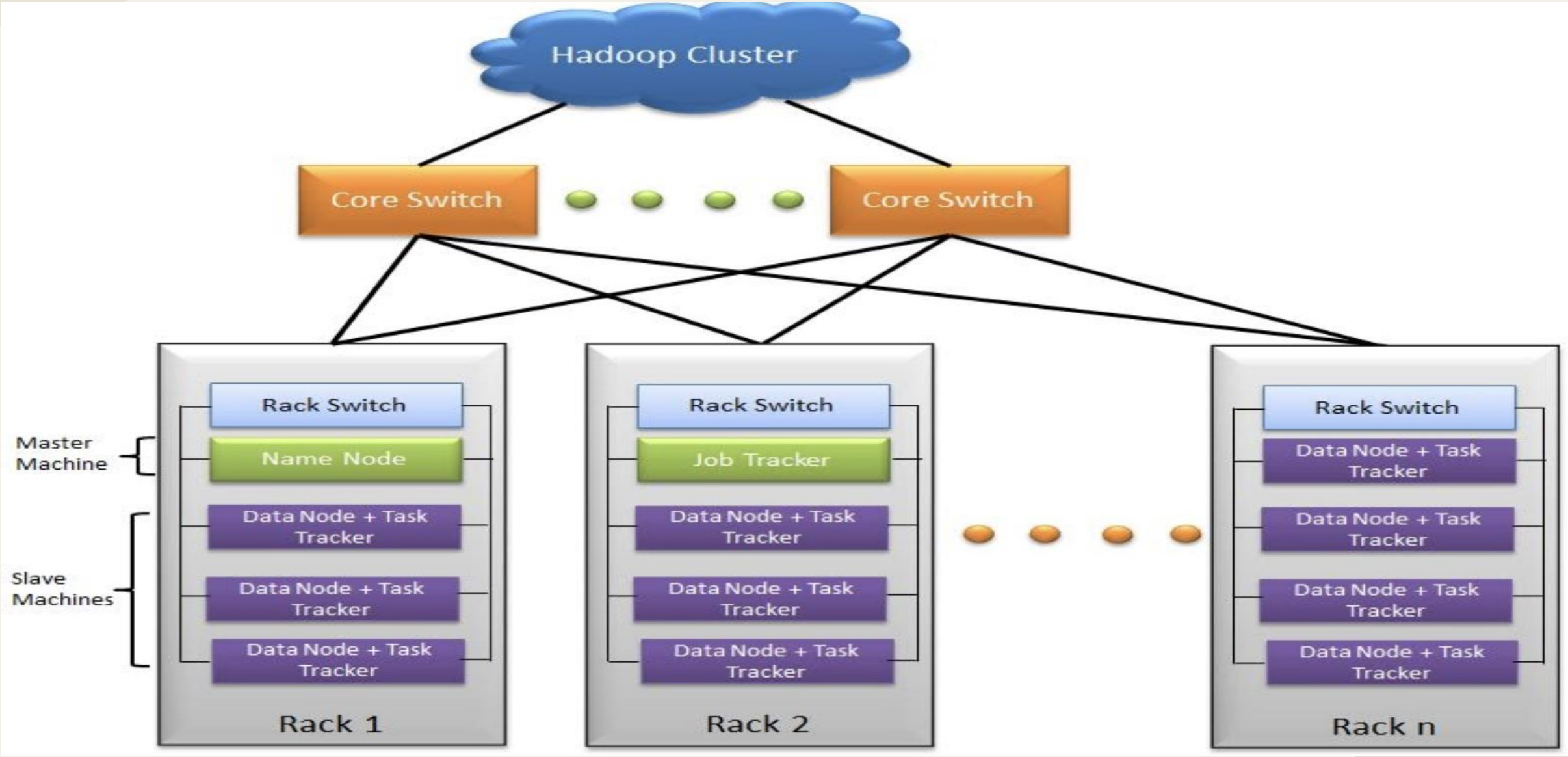
Hadoop Cluster

•Let's try to understand Hadoop Cluster architecture with the help of an example.
What would be typical Hadoop cluster setup for 4500 nodes

•In this case Hadoop Cluster would consists of

- 110 different racks, 40 slave machine
- There are global 8 core switches
- The rack switch has uplinks connected to core switches and hence connecting all other racks with uniform bandwidth, forming the Cluster
- In the cluster, you have few machines to act as Name node and as JobTracker. They are referred as Masters. These masters have different configuration favoring more DRAM and CPU and less local storage.
- The majority of the machines acts as DataNode and Task Trackers and are referred as Slaves. These slave nodes have lots of local disk storage and moderate amounts of CPU and DRAM

Hadoop Cluster



Apache Spark

- Apache Spark is an open-source, distributed computing system designed for big data processing and analytics.

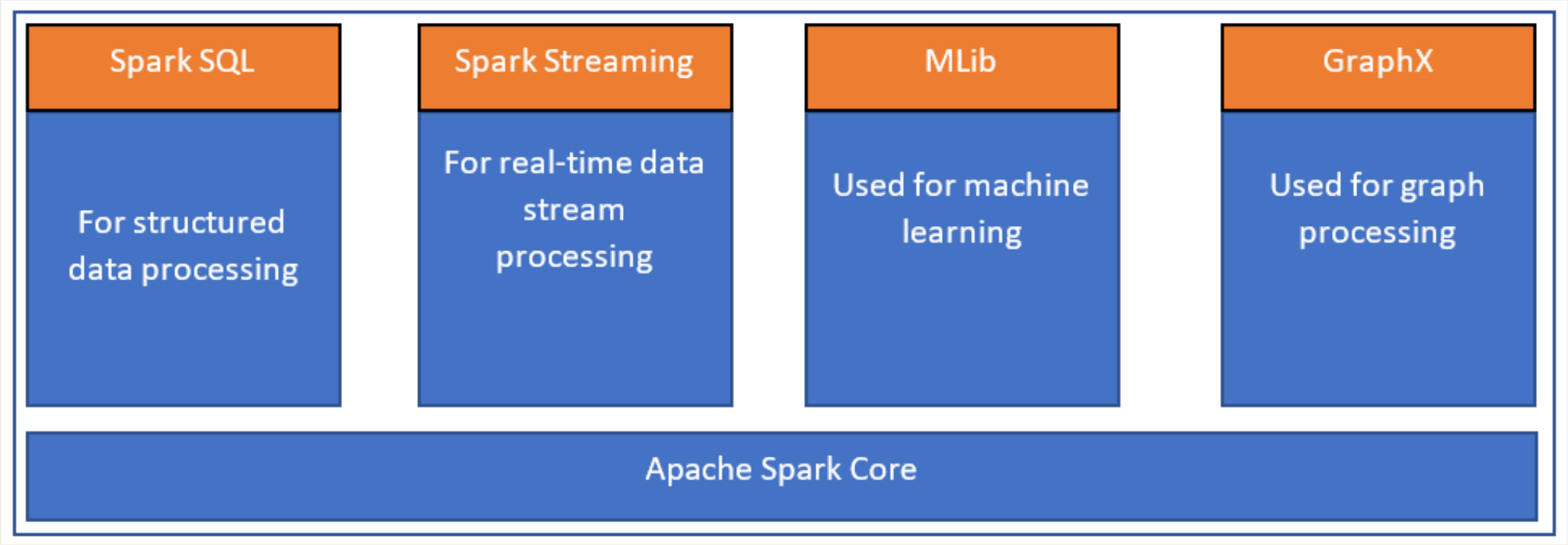
- It provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.

- Spark enables in-memory data processing, making it significantly faster than traditional disk-based systems.

Hadoop vs Spark

hadoop	APACHE Spark
Processing data using MapReduce in Hadoop is slow	Spark processes data 100 times faster than MapReduce as it is done in-memory
Performs batch processing of data	Performs both batch processing and real-time processing of data
Hadoop has more lines of code. Since it is written in Java, it takes more time to execute	Spark has fewer lines of code as it is implemented in Scala

Components of Apache Spark



https://sqlrelease.com/wp-content/uploads/2019/04/Big-Data-processing-using-Apache-Spark-Introduction-Spark-components.png?x66553

Big Data Services

SAAS

PAAS

IAAS

Big Data & Cloud Computing

- **Cloud computing is the use of computing resources (hardware and software) that are delivered as a service (SaaS or IaaS)**
- **In Business View: When it's smarter to rent that to buy...**

  "If you only need milk, would you buy a cow?"

Q & A

"THAT'S your Ark for the Big Data flood? Noah, you will need a lot more storage space!"

Thank You