Large Scale Computing Methods and Systems

Marian Bubak
Department of Computer Science, AGH University of Science and Technology
Kraków, Poland

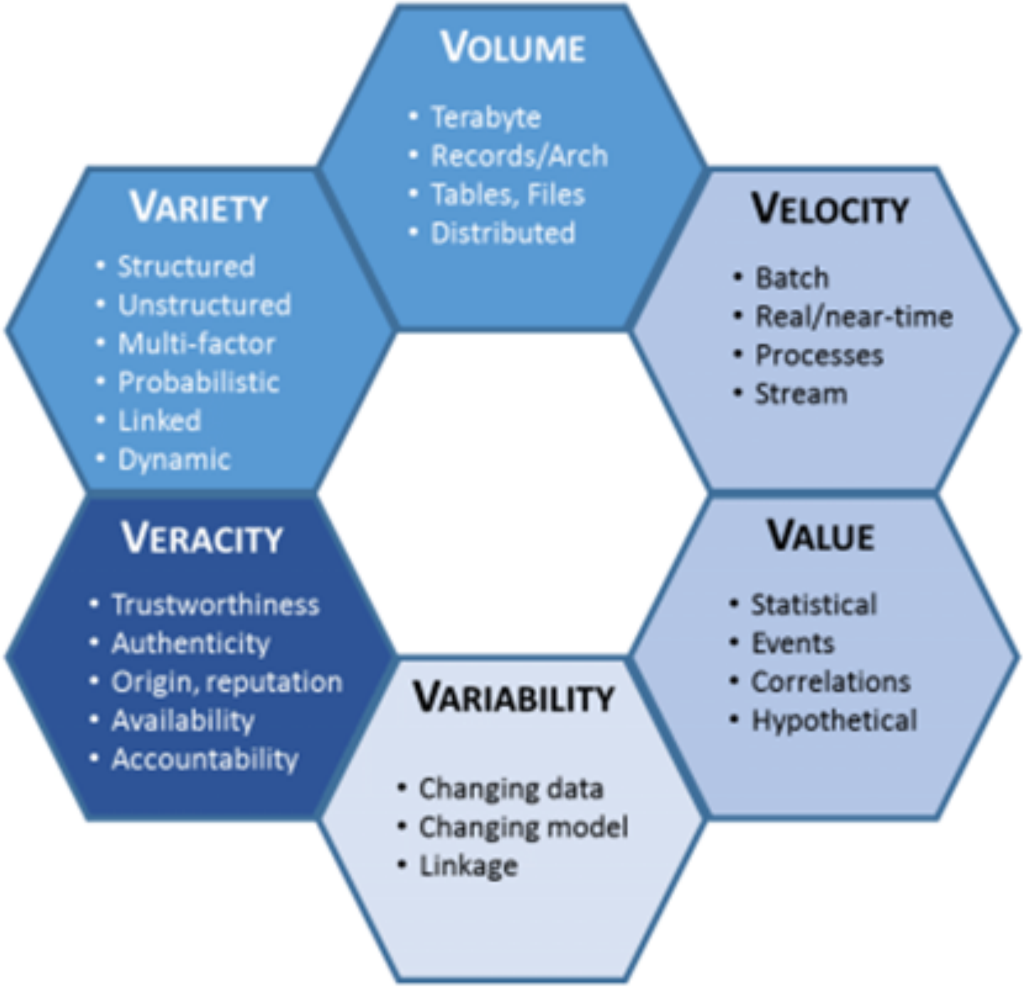
bubak@agh.edu.pl http://dice.cyfronet.pl/

Big Data in Large Scale Computing

Contributors

- Krzysztof Burkat (2019)
- Michał Ćwiertnia (2019)

What is Big Data? 6 Vs model of Big Data



De Mauro, Andrea & Greco, Marco & Grimaldi, Michele. (2014). What is Big Data? A Consensual Definition and a Review of Key Research Topics. 10.13140/2.1.2341.5048.

6 Vs explanation (1/2)

- **Volume**: The ability to ingest, process and store very large datasets. The data can be generated by machine, network, human interactions on system etc. The emergence of highly scalable low-cost data processing technology platforms helps to support such huge volumes. The data is measured in petabytes or even exabyte.
- **Velocity**: Speed of data generation and frequency of delivery. The data flow is massive and continuous which is valuable to researchers as well as business for decision making for strategic competitive advantages. For processing of data with high velocity tools for data processing were introduced. Sampling data helps in sorting issues with volume and velocity.
- **Variety**: It refers to data from different sources and types which may be structured or unstructured. The unstructured data creates problems for storage, data mining and analyzing the data. With the growth of data, even the type of data has been growing fast.

https://community.mis.temple.edu/mis520817/2017/04/07/the-6-vs-of-big-data/

6 Vs explanation (2/2)

- **Variability**: This refers to establishing if the contextualizing structure of the data stream is regular and dependable even in conditions of extreme unpredictability. It defines the need to get meaningful data considering all possible circumstances.
- **Veracity**: It refers to the biases, noises and abnormality in data. This is where we need to be able to identify the relevance of data and ensure data cleansing is done to only store valuable data. Verify that the data is suitable for its intended purpose and usable within the analytic model. The data is to be tested against a set of defined criteria.
- **Value**: Refers to purpose, scenario or business outcome that the analytical solution has to address. Does the data have value, if not is it worth being stored or collected? The analysis needs to be performed to meet the ethical considerations.

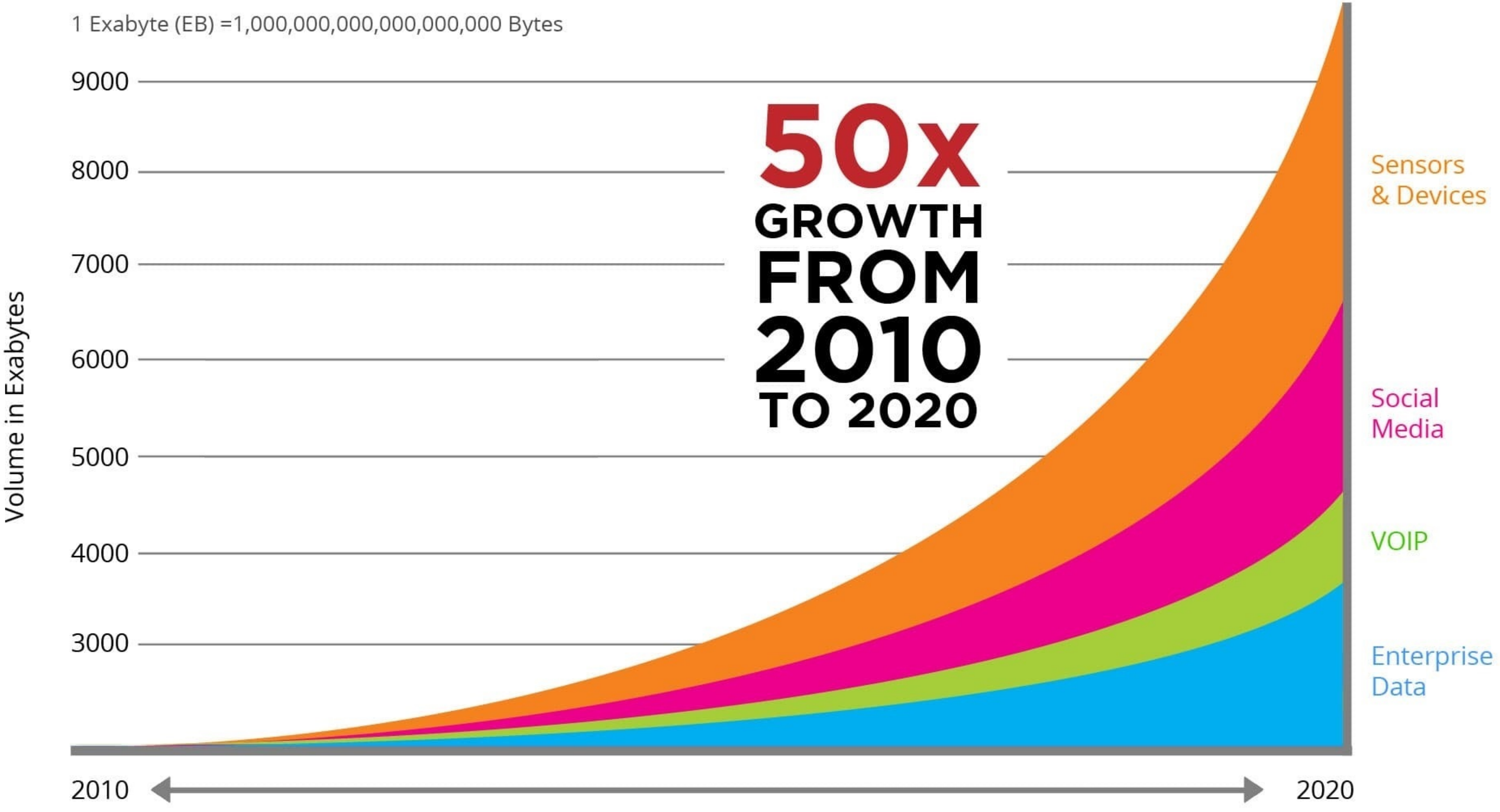
https://community.mis.temple.edu/mis520817/2017/04/07/the-6-vs-of-big-data/

Why Big Data is important?

- The ability to consistently get business value from data is now a trait of successful organizations across every industry, and of every size.
- Data analytics only returns more value when you have access to more data, so organizations across multiple industries have found Big Data to be a rich resource for uncovering profound business insights.

BIG DATA
Mobile
Internet
Networks
Technology
Strategy
Databases
Storage
Intelligence
Discovery
Management
Privacy
Infrastructure
Security
Information

Amount of data



https://corecompete.com/analytics-modernization-v1/big-data-showing-big-growth-min-2/

Introduction to data intensive applications

- **Data-intensive applications** are kind of computing applications which require large volumes of data and devote most of their processing time to I/O and manipulation of data.
- **Characteristics**: explosion of data and massive data processing, central but scalable storage systems, ultra-high speed network for data transfer: 100Gbps and faster networks
- **Examples**: data centres, grid and cloud computing, network storage

Martin Kleppmann: “Designing Data-Intensive Applications”

CAP Theorem

Theorem

It is impossible for a distributed data store to simultaneously provide more than two out of the following three guarantees:

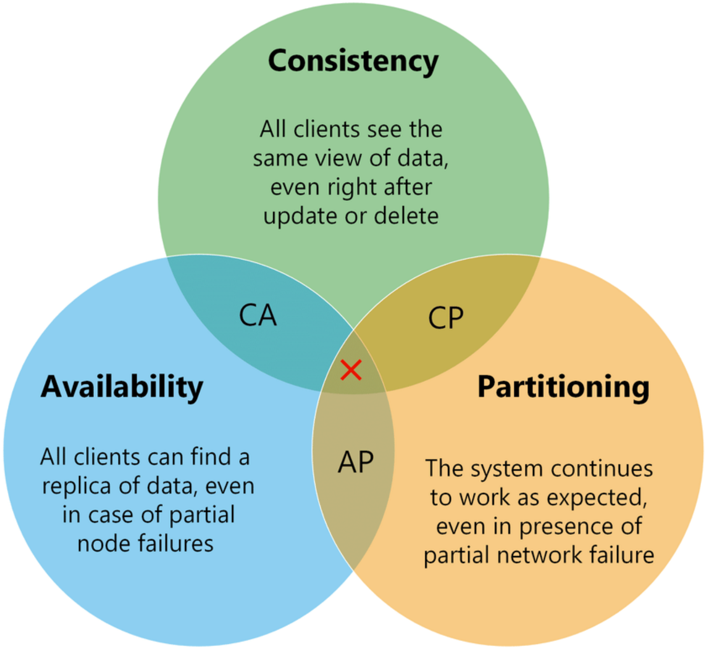
- Consistency – every read receives the most recent write or an error.
- Availability – every request receives a (non-error) response – without the guarantee that it contains the most recent write.
- Partition tolerance – the system continues to operate despite an arbitrary number of messages being dropped (or delayed) by the network between nodes.

Explanation

No distributed system is safe from network failures, thus network partitioning generally has to be tolerated. In the presence of a partition, one is then left with two options: consistency or availability. When choosing consistency over availability, the system will return an error or a time-out if particular information cannot be guaranteed to be up to date due to network partitioning. When choosing availability over consistency, the system will always process the query and try to return the most recent available version of the information, even if it cannot guarantee it is up to date due to network partitioning.

In the absence of network failure – that is, when the distributed system is running normally – both availability and consistency can be satisfied.

CAP is frequently misunderstood as if one has to choose to abandon one of the three guarantees at all times. In fact, the choice is really between consistency and availability only when a network partition or failure happens; at all other times, no trade-off has to be made.


Consistency
All clients see the same view of data, even right after update or delete
CA
CP
×
Availability
All clients can find a replica of data, even in case of partial node failures
AP
Partitioning
The system continues to work as expected, even in presence of partial network failure

[2] https://en.wikipedia.org/wiki/CAP_theorem
[31] https://www.researchgate.net/figure/Visualization-of-CAP-theorem_fig2_282679529

FAIR Data

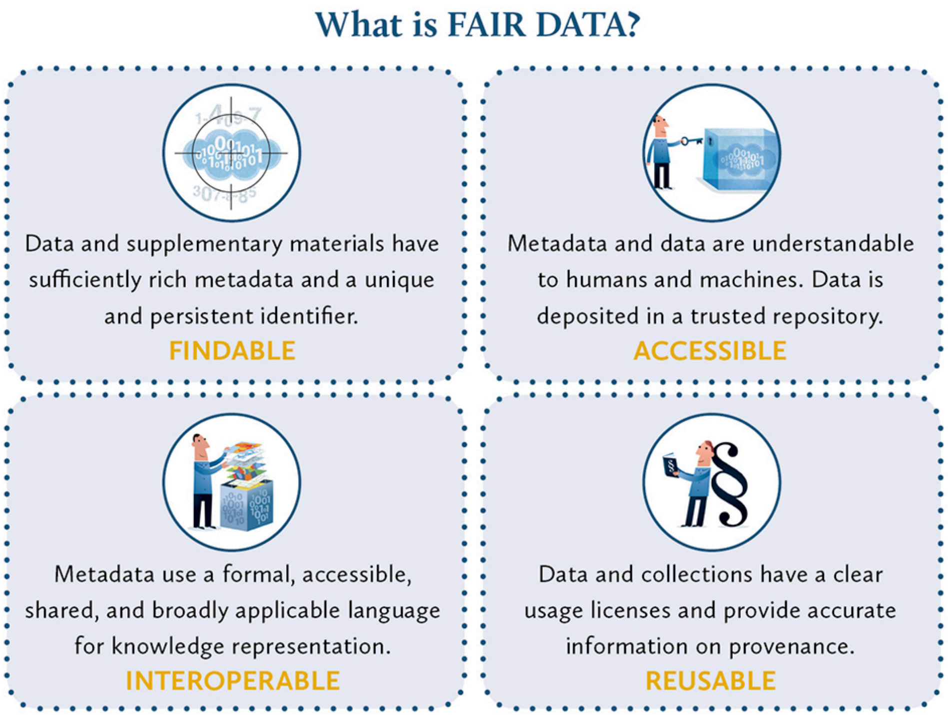
What is FAIR Data?

The attention of researchers is increasingly directed to the phases of the research lifecycle in which data are published, shared, discovered and reused. One of the perceived ways to achieve optimal reuse is to make data FAIR – i.e. Findable, Accessible, Interoperable and Reusable.

The FAIR guiding principles consist of 15 facets which describe a continuum of increasing reusability. Importantly, data should not only be FAIR for humans but also for machines, allowing, for instance, automated search and access to data. Funders like the European Commission have drafted Guidelines on FAIR Data Management for the H2020 programme. Good data management is one way to support the FAIR principles.

FAIR Principles

1. **Findable** – metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.
2. **Accessible** – once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.
3. **Interoperable** – The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.
4. **Reusable** – The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.



What is FAIR DATA?

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.
FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.
ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.
REUSABLE

[32] https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/1.-Plan/FAIR-data
[33] https://www.nature.com/articles/sdata201618
[34] https://www.go-fair.org/fair-principles/

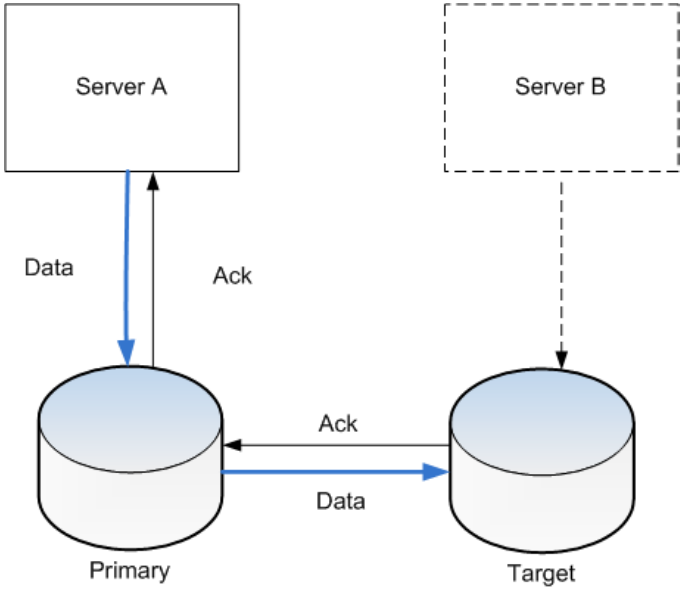
Introduction to data replication

What is data replication?

Data replication is the process of storing data in more than one site or node. This is necessary for improving the availability of data. It also provides consistency and reliability by creating multiple copies of the same data on different sites. Replication also provides minimum access cost, shared bandwidth utilization and delay time by replicating data. The value of replication is to provide transparent, flawless access to resources in the event of a system failure. Replication can be extended across a computer network so that storage devices can be located in physically separated facilities.

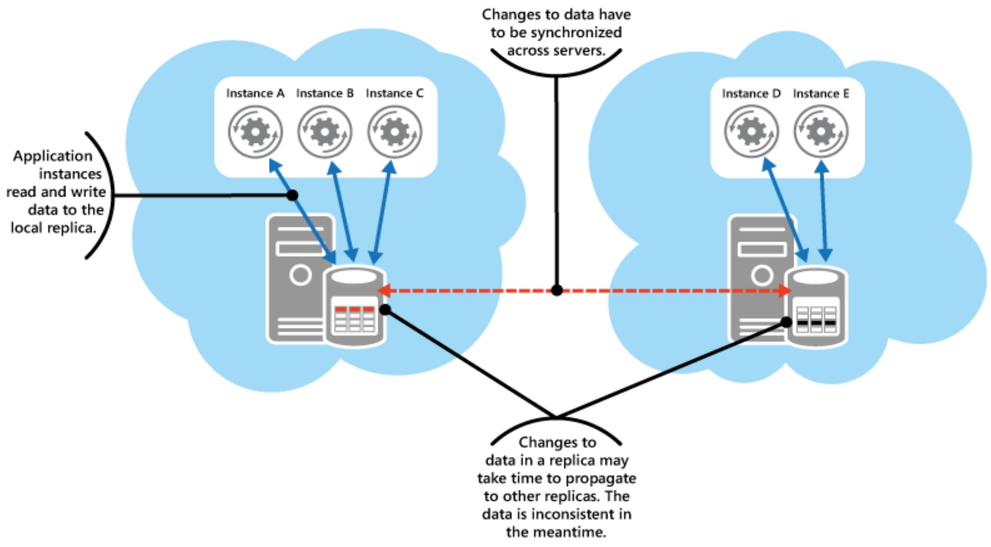
Why data replication is important?

- It allows to achieve high availability – characteristic of a system, which aims to ensure an agreed level of operational performance, usually uptime, for a higher than normal period. Modernization has resulted in an increased reliance on these systems. For example, hospitals and data centers require high availability of their systems to perform routine daily activities. Availability refers to the ability of the user community to obtain a service or good, access the system, whether to submit new work, update or alter existing work, or collect the results of previous work.
- It helps to reduce costs – the study of US data centers quantifies the average cost of an unplanned data center outage at slightly more than US$7,900 per minute (in 2013). Data replication improves reliability, and thanks to that can help
- It improves performance.



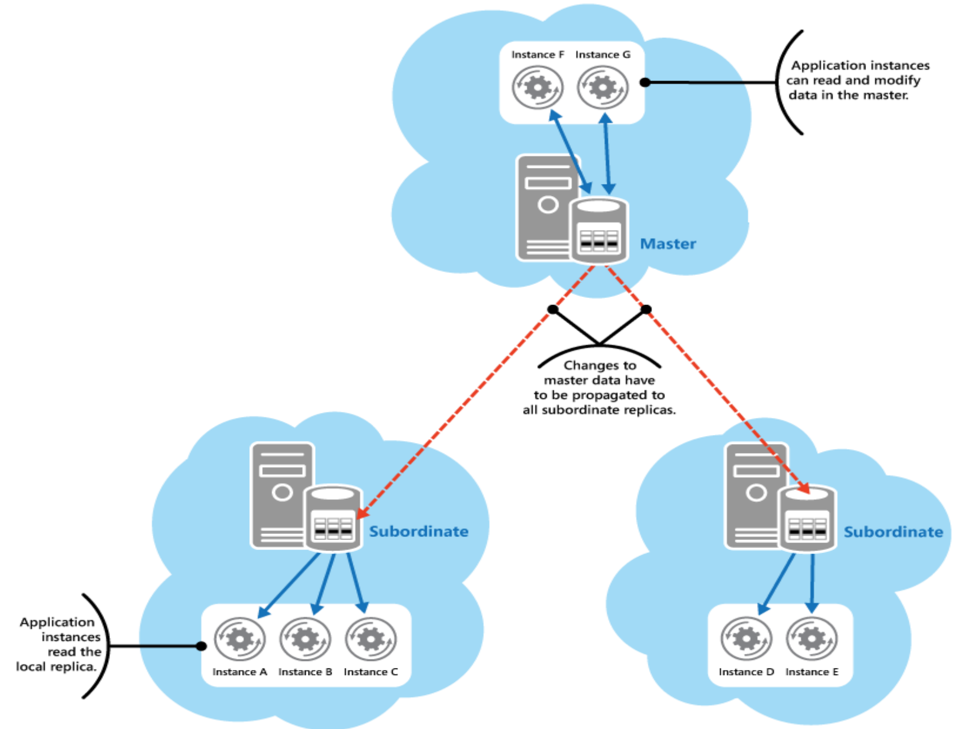
[3] http://ecomputernotes.com/database-system/adv-database/data-replication
[4] https://en.wikipedia.org/wiki/Replication_(computing)
[5] https://en.wikipedia.org/wiki/High_availability
[6] https://www.datacenterdynamics.com/news/one-minute-of-data-center-downtime-costs-us7900-on-average/

Types of data replication


Changes to data have to be synchronized across servers.
Instance A
Instance B
Instance C
Instance D
Instance E
Application instances read and write data to the local replica.
Changes to data in a replica may take time to propagate to other replicas. The data is inconsistent in the meantime.

Master-master replication

In this type of replication data in each replica is dynamic and can be updated. This topology requires a two-way synchronization mechanism to keep the replicas up to date and to resolve any conflicts that might occur. In a cloud application, to ensure that response times are kept to a minimum and to reduce the impact of network latency, synchronization typically happens periodically. The changes made to a replica are batched up and synchronized with other replicas according to a defined schedule. While this approach reduces the overheads associated with synchronization, it can introduce some inconsistency between replicas before they are synchronized.


Instance F
Instance G
Application instances can read and modify data in the master.
Master
Changes to master data have to be propagated to all subordinate replicas.
Subordinate
Subordinate
Application instances read the local replica.
Instance A
Instance B
Instance C
Instance D
Instance E

Master-subordinate replication

In this type of replication data in only one of the replicas is dynamic (the master), and the remaining replicas are read-only. The synchronization requirements for this topology are simpler than that of the Master-Master Replication topology because conflicts are unlikely to occur. However, the same issues of data consistency apply.

[7] https://msdn.microsoft.com/en-us/library/dn589787.aspx

Data synchronization

What is data synchronization?

Data synchronization is the process of establishing consistency among data from a source to a target data storage and vice versa and the continuous harmonization of the data over time. It is fundamental to a wide variety of applications, including file synchronization and mobile device synchronization.

Data synchronization is the process of maintaining the consistency and uniformity of data instances across all consuming applications and storing devices. It ensures that the same copy or version of data is used in all devices - from source to destination.

Strong synchronization vs weak synchronization

Strong synchronization policies enforce a tight replication of „like” patterns of data among the nodes of data region as well as strictly enforcing the replication policies among the nodes of the data region. The strong synchronization mechanism is used when low latency and high consistency are required from the data at the cost of scalability and flexibility.

Weak synchronization policies enable data to be synchronized on an „as needed” basis and sometimes not at all. The weak synchronization mechanism allows for less data consistency but for higher scalability and flexibility.

Challenges

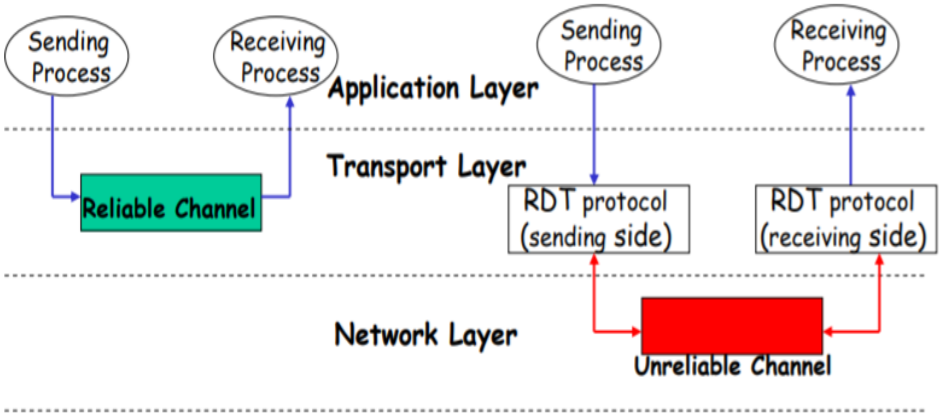
- Data formats complexity
- Real-timeliness
- Security
- Data quality
- Performance
- Maintenance

[8] https://www.techopedia.com/definition/1006/data-synchronization
[9] https://en.wikipedia.org/wiki/Data_synchronization
[10]https://books.google.pl/books?id=6WW73ziknwEC&pg=PA77&lpg=PA77&dq=grid+data+synchronization&source=bl&ots=07kJ37hJH0&sig=_rEUsDL9XOmWHtdW4ZGJTnR_FNk&hl=pl&sa=X&ved=0ahUKEwjjk-3B1ObYAhWEkCwKHc93D2EQ6AEIMDAB#v=onepage&q&f=false
[11] https://www.dataintegration.info/data-synchronization

Introduction to distributed data management

Reliable data transfer

The internet network layer provides only best effort service with no guarantee that packets arrive at their destination. Also, since each packet is routed individually it is possible that packets are received out of order. For connection-oriented service provided by TCP, it is necessary to have a reliable data transfer (RDT) protocol to ensure delivery of all packets and to enable the receiver to deliver the packets in order to its application layer.



GridFTP

GridFTP is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks. The GridFTP protocol is based on FTP, the highly-popular Internet file transfer protocol. Key features of GridFTP Protocol include:

- Performance – GridFTP protocol supports using parallel TCP streams and multi-node transfers to achieve high performance.
- Checkpointing – GridFTP protocol requires that the server send restart markers (checkpoint) to the client.
- Third-party transfers – The FTP protocol on which GridFTP is based separates control and data channels, enabling third-party transfers, that is, the transfer of data between two end hosts, mediated by a third host.
- Security – Provides strong security on both control and data channels. Control channel is encrypted by default. Data channel is authenticated by default with optional integrity protection and encryption.

[12] http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/
[13] https://www.d.umn.edu/~gshute/net/reliable-data-transfer.xhtml
[14] https://pages.cpsc.ucalgary.ca/~mahanti/teaching/W06/CPSC441/lectures/reliable.pdf

Network optimization techniques in context of data intensive applications

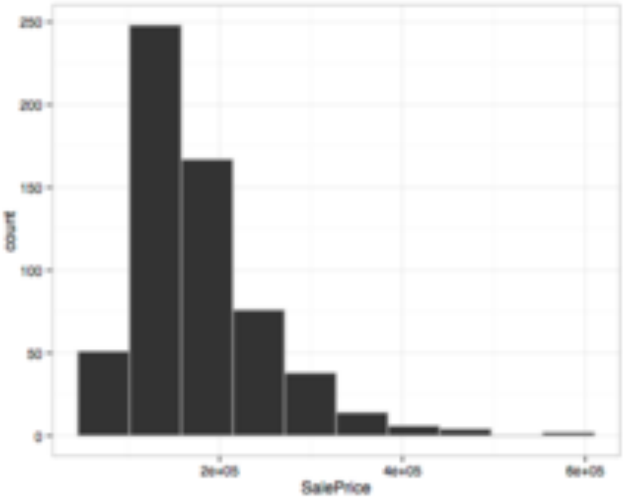
- The problem is, given that applications need to exchange large amounts of data or messages, how infrastructure can be prepared for such situations?
- One simple way is to increase throughput of intermediate infrastructure.
- Other way, by sending and receiving data simultaneously from and to several nodes, parallel applications create concurrent accesses to the resources of the network.
- Parallel file systems are useful when designing such high performance networks.

Martinasso, Maxime, and Jean-François Méhaut. "A contention-aware performance model for HPC-based networks: A case study of the InfiniBand network." European Conference on Parallel Processing. Springer, Berlin, Heidelberg, 2011.

Data visualization introduction

- **Data visualization** involves the creation and study of the visual representation of data.
- Making data visual is a big part of making it understandable and useful. For all the excitement about novel data sources like social computing or the Internet of Things, data analysis will eventually flow into a report or dashboard where someone must make sense of it.

Network visualization



Histogram of housing prices

Michael Friendly (2008). "Milestones in the history of thematic cartography, statistical graphics, and data visualization"

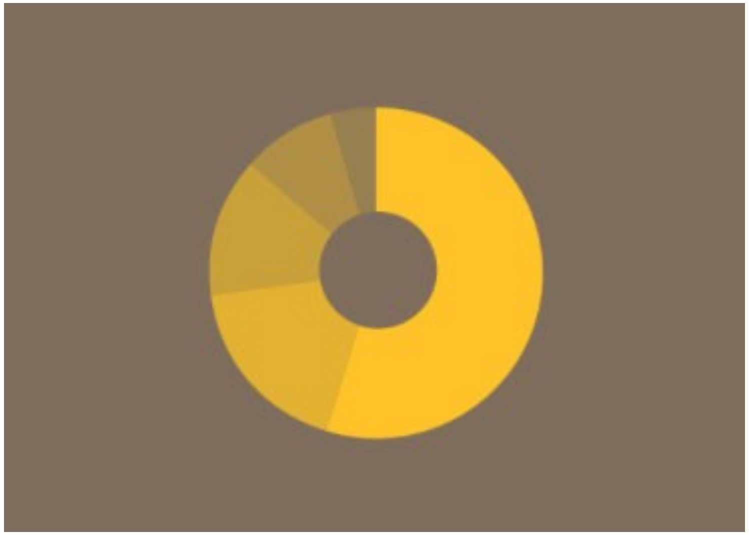
Popular techniques of data visualization

- **Two-dimensional area** - such visual forms are mostly geospatial, which means they represent some certain geographical location on the globe.
- **Multidimensional data visualizations -** this type of data visualization approaches is one of the most widespread, as it combines two or more dimensions to produce easy to grasp images.
- **Hierarchical data visualization -** sometimes it's important to show how one set of data values compares to another one or more data value sets.
- **Network data models -** when we need to describe the way various data sets compare and relate to each other, network data visualization techniques come to our help.
- **Temporal visualizations -** while looking quite like simple linear graphs, temporal visualizations include a start and finish time and some of the items measured might overlap, thus creating a descriptive image shoving the variable adjustment over time.

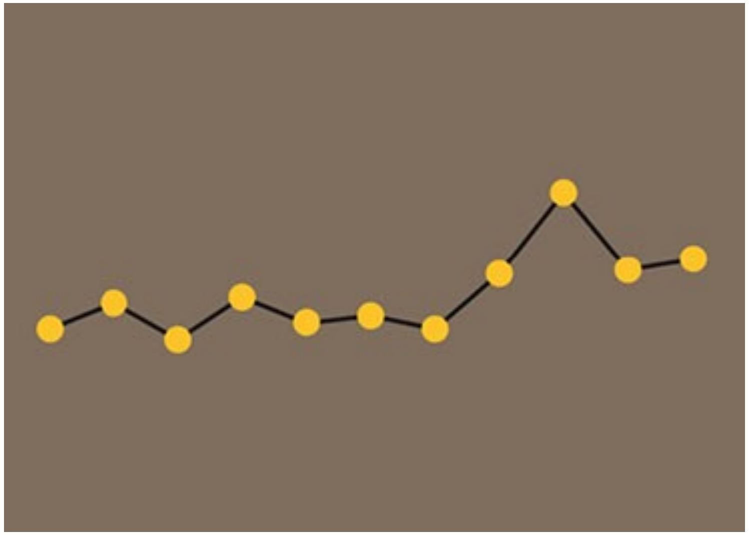
https://towardsdatascience.com/big-data-information-visualization-techniques-f29150dea190

Data visualization techniques examples

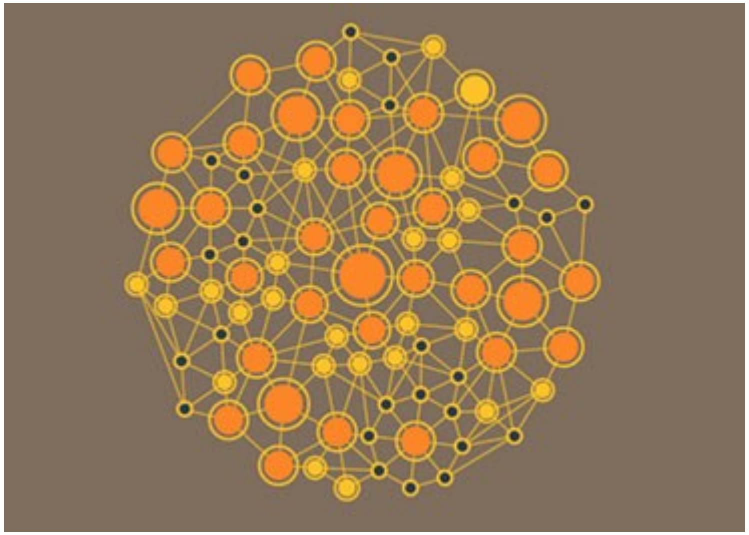
Two-dimensional – choropleth



Multidimensional – pie chart



Temporal – Connected Scatter Plot



Network – node-link diagram

Big Data in cloud computing

- **Cloud computing** provides fundamental support to address the challenges with shared computing resources including computing, storage, networking and analytical software; the application of these resources has fostered impressive Big Data advancements.
- Cloud computing and Big Data combined enable new science discoveries and application developments.

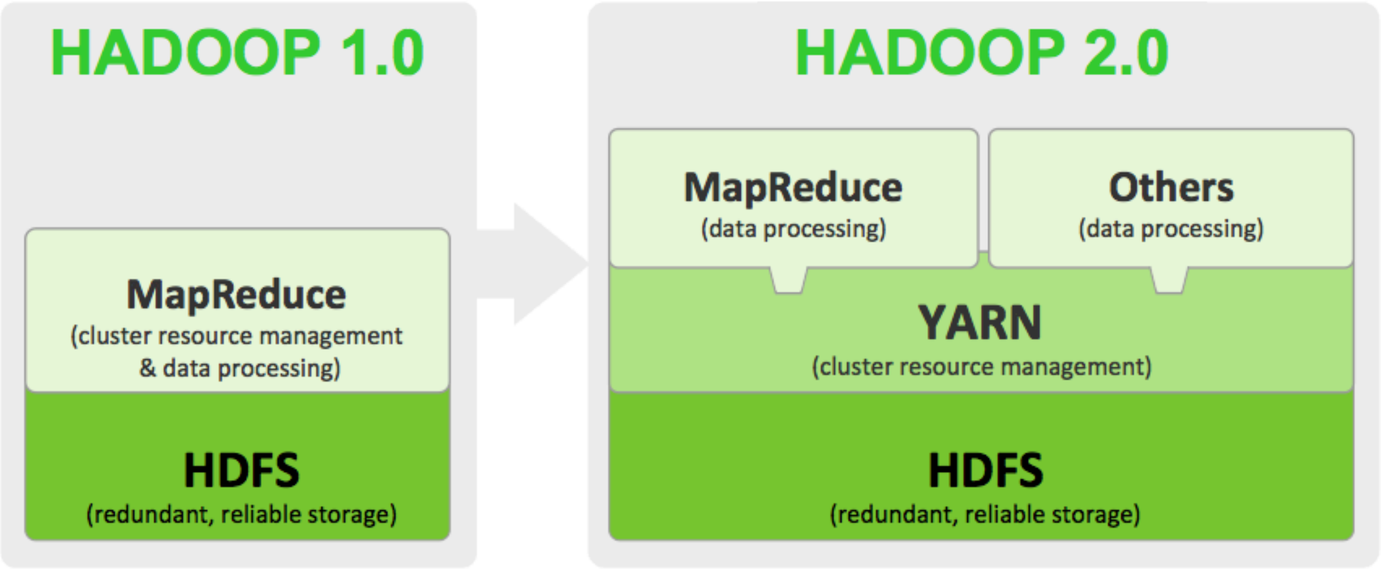
Big Data\cloud computing	**Elasticity**	**Pooled**	**On-demand**	**Self-service**	**Pay-as-you-go**
Volume		x			x
Velocity	x		x		
Variety	x	x		x	
Veracity				x	x
Value	x		x		x
Variability	x		x	x	

Addressing the Big Data challenges with cloud computing

Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu & Fei Hu *"Big Data and cloud computing: innovation opportunities and challenges" International Journal of Digital Earth Vol. 10, Iss. 1, 2017* http://www.tandfonline.com/doi/full/10.1080/17538947.2016.1239771

Big Data technologies – Hadoop

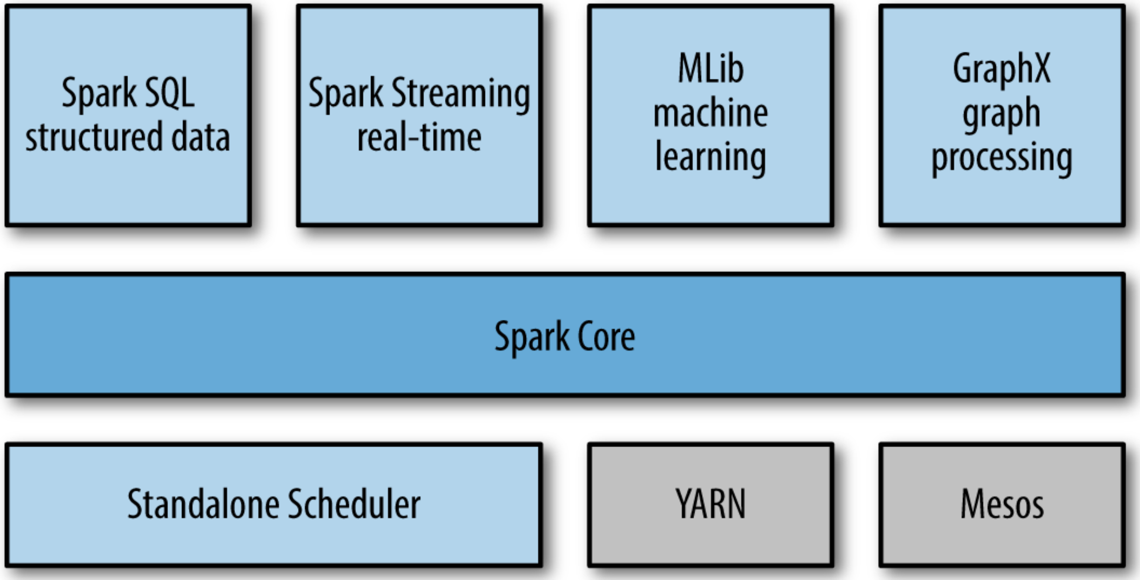
- **The Hadoop Distributed File System** (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications.
- In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size.


HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1-10.

Big Data technologies – Apache Spark

- **Apache Spark** is an open-source engine developed specifically for handling large-scale data processing and analytics. Spark offers the ability to access data in a variety of sources.
- It is designed to accelerate analytics on Hadoop while providing a complete suite of complementary tools.



Zaharia, Matei & Chowdhury, Mosharaf & J. Franklin, Michael & Shenker, Scott & Stoica, Ion. (2010). Spark: Cluster Computing with Working Sets. Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. 10. 10-10.

Dangers of Big Data

- Big Data comes with promise, but it also comes with risk. First is the erosion of privacy. More people know more about each of us than at any point in human history. It's not only easy to find where we live, but where we go, how we live, and what we think.
- This makes individuals and societies more open to manipulation. We can be tricked into giving up our passwords or influenced to vote for candidates. More data offers more ways for advertisers and media companies to shape our desires and value.

- Having all our data somewhere in the cloud leaves it vulnerable to attacks and misuse.
- Data about us could be used to spy on us.

S. Sagiroglu, D. Sinanc "Big data: A review" in 2013 International Conference on Collaboration Technologies and Systems (CTS)

Big data or good data?

- In 2013 Google Flu Trends (GFT) was predicting more than double the proportion of doctor visits for influenza-like illness than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States. This happened despite the fact that GFT was built to predict CDC reports by the use of big data.
- “Big data hubris” is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.
- Google’s algorithm was quite vulnerable to overfitting to seasonal terms unrelated to the flu, like “high school basketball.”
- This shows that not only amount of data matters, it is also important what is the **quality of data and how we use it**.

D. Lazer, R. Kennedy, G. King, A. Vespignani “The Parable of Google Flu: Traps in Big Data Analysis” in Science volume 343

Questions

1) What is the definition of Big Data ?
2) Describe 6 Vs of Big Data.
3) What is CAP Theorem? What are the implications that it poses?
4) What is FAIR Data? How does it help in reusing data?
5) Why is the data replication such an important factor in distributed computing?
6) Describe techniques of data visualization.
7) Explain relation between big data and cloud computing.
8) Describe Apache Hadoop and Apache Spark technologies.
9) What are the dangers of big data ?