“St. Kliment Ohridski” University - Bitola, Macedonia
Faculty of Information and Communication Technologies – Bitola
Department of Software Engineering and Information Systems

1. Properties of NoSQL Databases

Prof. Blagoj Ristevski, PhD

JKU - Johannes Kepler Universität Linz, Austria
Institut für Systemsoftware, November 2024

Overview of Lectures

1. **Properties of NoSQL databases**
2. **Relational vs NoSQL databases**
3. **Types of NoSQL databases**
4. **Column-oriented databases**
5. **Cassandra Data Model**
6. **Casandra Query Language (CQL)**

Introduction

- A huge amount of heterogeneous healthcare, medical, IoT, sensor, biological, multimedia, and administrative data as well as data from social networks are generated daily.
- These data are not only structured but also unstructured or semi-structured
- These data have big data properties so the standard relational databases (DB) are unsuitable for storing and managing them.
- Standard relational DBs are optimized to deal with aggregated structured data stored in tables that are related using relationships with inflexible schemas.
- For instance, graph DBs are appropriate to deal with highly connected data.
- A graph database has a flexible schema and it can store, manage, and update data and relationships.

Non Relational DB

- In contemporary software development, the term non-relational DB regards to data collection management systems that:
  - have no rigid static data structure
  - do not check whether the DB integrity rules are met
  - do not use the query language SQL

Shortcomings of the Relational DBs

- Relative high cost because of
  - updating and creating DBs
  - non-redundancy and strict data structure
  - difficult distribution
  - keeping the data consistency
  - the change of data structure
  - keeping the relationships among tables
  - difficult dealing with:
    - temporal data
    - space data
    - multimedia
    - digital libraries
    - as well with non-structured or semi-structured data

Challenges in Web applications regarding DBs

- Current web applications face the following issues:
  - Concurrent users
    - Millions of users simultaneously
  - huge data amounts
    - TBs and PBs data are created daily
  - data processing
    - Each user’s activity needs some data processing
  - high dynamicity
    - adding new possibilities continuously
    - existing components updating

NON RELATIONAL DBs

- Non-relational DBs DBS that are
  - Non-based only on the relational data model
  - easy distributed
  - Horizontally scalable (their schema is easily changeable)
- Other properties of non-relational DBs:
  - without static schemas
  - easy to replicate
  - They are based on **BASE** principles instead of ACID:
    - **B**asically **A**vailable
    - **S**oft state – their state can change during the time
    - **E**ventual consistency - системот ја постигнува конзистентната состојба со текот на времето
  - they can deal with big data and large data sets

NoSQL DBs

- Non relational DBs are often called NoSQL DBs
  - It means that they deviate from SQL as a symbol of the relational DBs
  - Now they have a tendency to have a query language similar to SQL
  - It is interpreted as “not only SQL” DBs

  Characteristics:

- They are able to deal with complex data structures such as key-values, documents, graphs, etc.
- High level of distribution Intended to handle distributed DBs
- Capable to work with big data
- Working with web applications in real-time

TYPES of NoSQL DBs

- key-value DB
  - data are presented as a collection of key-value pairs, so each possible key appears only once in one collection
    - examples: NoSQLz, Redis, Riak, …
- document-oriented DB
  - They are used to store and manage semi-structured data. XML DBs deal with XML documents. Documents can be XML, YAML, JSON, BSON format.
    - examples: BaseX, CouchDB, CrateDB, Amazon DocumentDB, MongoDB, ...

TYPES of NoSQL DBs

Graph DBs

- Are used for data whose relationships can be presented with graph structure, consisted of elements – **nodes** and **edges.** Nodes are connected with edges representing the relationships.
- Application: in social networks, computer networks, biological networks, network topology data etc.
  - examples: AllegroGraph, Neo4j, InfiniteGraph, Amazon Neptune, …

Column DBs

- Data as tables :

RowId	EmpId	Lastname	Firstname	Salary
001	10	Smith	Joe	60000
002	12	Jones	Mary	80000
003	11	Johnson	Cathy	94000
004	22	Jones	Bob	55000

- They are stored as columns:

10:001,12:002,11:003,22:004;
Smith:001,Jones:002,Johnson:003,Jones:004;
Joe:001,Mary:002,Cathy:003,Bob:004;
60000:001,80000:002,94000:003,55000:004

- examples: SAP HANA, Apache Cassandra, Apache Kudu, C-Store, ...

BIOLOGICAL AND BIOMEDICAL NETWORKS

- There exists a wide variety of biological network data and numerous available public DBs, that organize and provide data to researchers, medical doctors, etc.
- Graph DBs enable the causal relationship between biological and medical entities in the networks. For instance, when the causal relationships between particular diseases are depicted in graph DBs, further deterioration of those diseases can be prevented in time.
- Biological data representing by networks enable visualization and analysis of various phenomena such as drug-target interactions, and diseases' spreading.
- The most commonly used biological networks are PPI networks, genetic interaction networks, gene/microRNA/transcriptional regulatory networks, cell signalling networks, metabolic networks, disease-disease association networks and heterogeneous biological networks.
- To exploit medical and biological big data, network theory provides efficient tools. Biological networks cover experimentally identified physical and functional interactions; computationally and statistically inferred relationships; evolutionary associations based on protein co-evolution, domain families and sequence homology; and functional associations and ontologies representing systematized knowledge.
- To identify specific sub-networks or interactions, querying interfaces are used.
- To study biological and biomedical networks, network properties should be considered.

BIOLOGICAL AND BIOMEDICAL NETWORKS

- Moreover, specific network algorithms like **community detection** and **network motif discovery** algorithms are very appropriate to be applied on graph DBs.
- A community is a subgroup of a network whose internal connections are much denser than external connections. Identifying such communities can be essential for evaluating group behaviour, discovering emergent phenomena, visualization, identifying network sub-clusters and comprehending network structure.
- In the last decade, network medicine as an approach to understanding human diseases from a network theory has had significant results. Also, the study of diseases as non-isolated elements and the understanding of how they relate to each other is crucial to insights into pathogenesis.
- As a part of the medicine network, the disease network succeeds in revealing hidden connections among apparently unconnected biomedical entities such as diseases, physiological processes, signalling pathways, and genes as an intuitive pathway for visual data understanding.
- Some researchers proposed using directed disease networks to facilitate multiple-disease risk prediction, proposing a new framework that combines directed disease networks with system techniques to enhance the prediction risk assessment. The proposed model takes into account multiple-disease risk assessment scores for different diseases based on the patient’s medical history as a recommendation system where input is the medical record of the patient and output is a list of ordered diseases with a fixed length.

BIOLOGICAL AND BIOMEDICAL NETWORKS

- Human disease and health networks can be used to quantify the contagion effects. The estimation and identification of the contagion effects can be important taking into account understanding the spread of human disease and health behaviour as well as various implications for creating effective public health interventions. Disease and health networks also can be used together with traditional statistical analysis and several state-of-the-art statistical methods to quantify some advancement and remaining challenges.
- Graph data to explore medical data focused on the analysis of drugs for different diseases with complex network analysis is used. They used the drug as a node, the relationship as an edge connecting the two nodes, and the co-occurrence frequency of the drug, as the weight of the edge to establish a network graph. The clustering algorithm of the weighted network graph use used also, as the centre of a diffusion method combined with the network topology and the edge weights.
- The purpose is to divide the network graph into communities. Graph DBs are also used to create biological network models that offer a framework for understanding disease by describing the relationships between the mechanisms involved in the regulation of biological processes.
- The mentioned network scoring contains public transcriptomic data for some subset of mechanisms and networks that matched the measured outcomes.

BIOLOGICAL AND BIOMEDICAL NETWORKS

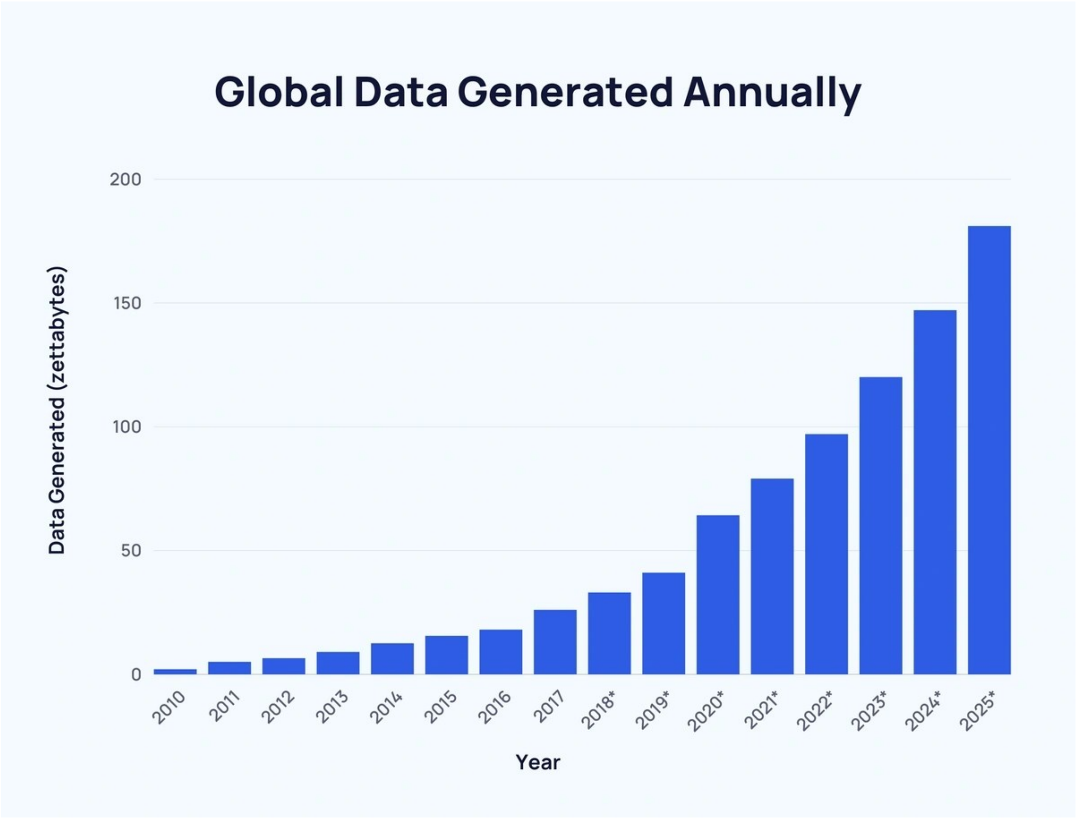
- Many examples show the increased usage of drug-target and disease network to increase the understanding of complex diseases and the concept of a new "multi-target, multi-drug" model, aimed at systemically modulating multiple targets. Identification of the interaction between drugs and target proteins plays an important role in genomic drug discovery.
- The aim is to discover new drugs or novel targets for existing drugs due to the laborious and costly experimental process of drug-target interaction prediction and provide useful information for supporting experimental interaction data.
- In postgenomic drug discovery, the large-scale integration of genomic, proteomic, signalling and metabolomics data can provide information useful to construct complex networks of the cell that would increase the understanding of physiological or pathophysiological states on a molecular basis.
- An emerging paradigm of pharmacology in the postgenomic era that states that drug, target and disease spaces can be correlated to study the effect of drugs on different spaces and their interrelationships can be exploited for designing drugs or cocktails which can effectively target one or more disease states is created.

Big Data in Medicine and Healthcare

- Nowadays, using numerous diverse digital devices results with explosive growth of the different types of a large amount of data, which enable extraction of new information and inherent insights contained in data. Such enormous variety of data collected from various and heterogeneous sources in healthcare and medicine can make valuable comprehension for patients, clinicians, hospitals, pharmacy, insurance companies and other involved parties.
- Additionally, when data from social media, banks and credit cards, IoT data for measuring vital signs are attached to the healthcare and medical data, a holistic view of a patient with environmental factors, which might have influence on patients' health, can be obtained. These linked data can be very sensitive and with different quality attributes. But it is very useful to discover the connections from EHR and coding systems in order to establish common criteria which will be beneficial for big data analysis and visualization for different stakeholders.
- A large volume of healthcare data is generated in hospitals during clinical treatments, labs and administrative procedures.
- This huge volume of data collected from this sector have to be analyzed in order to obtain specific knowledge for all stakeholders in healthcare.
- Usually these data are stored as patients' EHR, medical data records

Amount of Big Data generated

- It is estimated that over 402.74 exabytes 1EB = 10^{18}B are produced daily, as various images, videos, sensor data, ...
- The estimation in 2013 was that about 9 ZB data were generated (1ZB= 10^{21}B, source: Statista)
- The estimation in 2021 was that 79 ZB were generated
- The estimation in 2023 was that 120 ZB were generated



Big Data in Medicine and Healthcare

- Regarding healthcare and medical big data, the first characteristic of these data is complexity that comes with a wide range of activities connected with:
  - patients
  - physicians
  - hospitals
  - clinicians
  - other healthcare providers
  - healthcare insurance companies
  - medical instruments
  - national regulations
  - pharmaceutical companies
  - healthcare research groups
  - healthcare IoT appliances
- These data are also connected with living conditions as environmental data, transportation and communications, social media and advertising data.

Big Data in Medicine and Healthcare

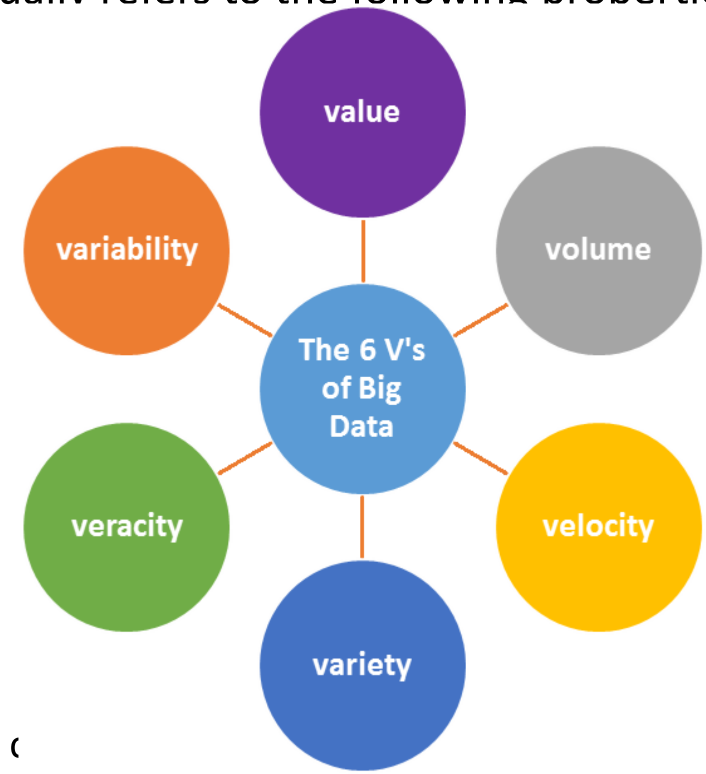
- Healthcare big data can be produced also by many types of data sources:
  - social media and markets
  - instruments, medical devices, mobile devices and services
  - technological and network services
  - EHR
  - physicians' notes
  - medical and pharmaceutical research
- Certain ethical healthcare obligations demand a high quality of services for patients. The aim of medical data storage and created documents is to support the quality management in healthcare institutions.
- Historical diseases data for each patient, healthcare quality system monitoring, specific clinical insights and epidemiological data should also be provided. This could be very useful to gain suitable knowledge that can be used for further training and education of healthcare professionals as well as to assess medical students.
- These data have to be stored in high capacity distributed repositories and used from the other stakeholders for various purposes.
- Suitable data mining techniques are applying on the medical and healthcare data sets. The specific purpose of healthcare data mining procedures should be to support anonymized patient data exchanges among healthcare staff and institutions, to support external demands like law, reimbursement procedures and documentation for planning and control of healthcare services.

Medical and Healthcare Big Data

- Using of numerous diverse digital devices that generate high volume of heterogeneous structured, semi-structured and unstructured data results with explosion and significantly growth of the voluminous complex data. These data enable extraction of new inherent insights in many domains.
- The aim of this enormous variety of data collected from various and heterogeneous sources is to make valuable comprehensions for patients, clinicians, hospitals, pharmacy, so the medical and healthcare data analysis become an extremely challenging task.
- Big data was especially useful for science and technology, business, healthcare and education.
- IoT has appeared, many sensors were connected for enhancing the human life, producing a huge amount of data.
- Big data then was just a part of the wider concept of data science that included many others methods and techniques for big data analysis and visualization usually included in the wider business intelligence and analytics (BIA) concept in order to create a big impact from big data.

Big Data properties

- Nowadays, big data usually refers to the following properties denoted as 6 “V’s”:



- volume - huge amount of
- velocity - data in motion as well as the frequency and speed of creation, processing and analysis of data
- variety - complexity and heterogeneity of multiple data sets
- value - the data coherent analysis, which should be beneficial to the clients, costumers, managers, organizations, corporations and other stakeholders
- variability - data consistency
- veracity - data quality, relevance, uncertainty, reliability and predictive value

Medical and Healthcare Big Data

- Particularly, in medicine and healthcare, in order to obtain more medical related data and to improve disease diagnostics, many wearable sensors, remote monitors, handheld, wearable, smart and capturing devices are used, as well as data generated by using many novel omics technologies.
- Current advances in patients' EHR, their integration with social, behavioral and various omics data have led to novel healthcare models which support personalized and precision medicine.
- Personalized healthcare services provide individuals-tailored diagnoses, drugs and treatments based on the psycho-physiological and spatio-temporal circumstances.
- The main challenge in using healthcare big data effectively is to identify the potential sources of healthcare information and to highlight the value of linking these data together.
- Healthcare data collection is the main data stream in healthcare big data analytics.
- Data sources can be curative, preventive as well as other types of healthcare and medical sources.

Medical and Healthcare Big Data

- Curative data are medical records, labs tests, referral and prescription data.
- Preventive data can be taken from
  - growth cards
  - maternal and child healthcare (MCH) cards
  - school healthcare cards
  - family registration cards.
- Other data sources can deliver comprehensive contents, record filling (patient-retained), layout (self-explanatory), production forms and various environmental data.
- Demographic surveillance should also be taken into account in healthcare data collection because of the need of information such as:
  - causing mortality data
  - sex
  - age
  - age-specific fertility
  - perceived data for mortality and disability
  - expenditure for household healthcare
  - practices
  - service quality and costs covered with healthcare insurance.

Medical and Healthcare Big Data

- System medicine, which combines clinical decision support systems (DSS) and EHR systems, aims toward individualized disease prognosis and treatment of the patients. These prognoses and treatments have to be based on various large amounts of data including phenotype data, omics data and individual preferences of the patients.
- Data used and mutually combined in system medicine can be categorized in three main groups:
  - personal data (behavior data, demographic data)
  - clinical data (examination data, laboratory data, imaging data) and
  - omics data (e.g. genomics, proteomics, metabolomics, transcriptomics, lipidomics, epigenomics, microbiomics, immunomics and exposomics data).
- Data used in the healthcare systems and applications are categorized as:
- **Structured** data have defined data type, structure and format. Such data in healthcare systems are laboratory results, hierarchical terminologies of different diseases, their symptoms, information about diagnosis and prognosis, patients' data, drug and billing information.
- **Semi-structured** data, which are usually generated from various sensor devices for monitoring of patients' conditions and behavior, are organized with minimal structure properties.
- Besides these two categories of data, **unstructured** data have no inherent structure. These data usually include medical prescription, written in natural human languages, biomedical literature, clinical letters and images.

Medical and Healthcare Big Data

- EHR data can be categorized in two main categories:
  - electronic medical records (EMR) and
  - sensor data.
- EMR data are usually consisted of:
  - patients' medical history
  - medical features (e.g. diagnoses, lab tests, medication, procedures, unstructured text data, image data) and
  - socio-demographic information about patients
- Sensor data are collected from various sensors and they originate from huge number of users and devices, hence generating enormously large amounts of real time data streams
- Main characteristics of EMR data are high-dimensionality, missing values, sparsity, irregularity, intrinsic noise and bias
- Problems with missing data values and data sparsity are usually solved by using removal or imputation methods.

Medical and Healthcare Big Data

- Medical imaging is a vigorous source of phenotypic data appropriate for further data analysis, personalized medicine, predictive analytics and artificial intelligence.
- Medical imaging data are generated by imaging techniques: X-ray, mammography, computer tomography (CT), ultrasonography, fluoroscopy, photoacoustic imaging and magnetic resonance imaging (MRI), histology, positron emission tomography (PET), radiography, nuclear medicine, tactile imaging, echocardiography, angiography etc.
- Prescription data contained in the patient EHR generated by physicians, clinical notes, medical research reports are examples of data that contain natural language terms, mathematical symbols and graphs.
- These incompatible data structures and formats along inconsistent data semantics combined with huge data volumes make healthcare big data analytics a challenging and demanding task.
- Large scale omics data can ensure understanding of the molecular basis of particular diseases and disease risks.
- To model and analyze complex interactions that occur between entities in biology, medicine and neuroscience, networks are fundamental and very suitable tools.
- The complex network nodes represent dynamic entities such as genes, microRNAs, proteins, metabolites, whereas the edges represent the links and interactions among nodes.

Big Data Analytics

- Rapidly growing amounts of data are faced with inefficient storage, preprocessing, processing and analysis by traditional relational databases.
- To perform efficient multisite and multivariable searching and querying, **high indexing and efficient data lookup** are required.
- Another stage is **pre-processing of the raw data** that might be inconsistent, inaccurate, erroneous and/or incomplete.
- To make big data analysis more reliable, **integration** of data from heterogeneous sources must be performed to have a conventional structured form.
- To improve the quality of collected data from various sensors/devices and to obtain more reliable analytics results, **identifying and removing** of incomplete, inaccurate and irrelevant data should be made.
- These unreliable data are usually replaced with interpolated values.

Big Data Analytics

- For these reasons, when we are talking about healthcare big data analysis and **visualization**, as most desirable techniques that create rapid information and deep insights, we have to **analyze** who will use results of data analyses and visualization and for which purposes.
- After defining the users/stakeholders, we have to analyze all necessary data sources and data formats and manner how to extract the information and knowledge from these healthcare and medical data, which demands to scrutinize design of the healthcare information system, **data preparation for further data analysis and visualization** and choosing suitable analytics and visualization tools and platforms.
- Big data analytics acquits enormous variety of data from former and current customers to obtain valuable knowledge to improve decision-making, to predict customer habits and behavior and to obtain real-time customer-tailored offers.
- Big data analytics in healthcare and medicine aims to bridge the gap between costs and outcomes in healthcare, which is a result from poor management of insights from research and poor resources management.

Big Data Analytics

- Moreover, data created by omics technologies can significantly improve prediction of diabetes, heart diseases, cancer and other disease.
- The main barriers in computer science related to the integration of omics data into clinic systems are: the development of model of cellular processes that cover non-corresponding omics data types, limitation for data storage and organization of heterogeneous data sets, generated from diverse high-throughput omics technologies, and the lack of suitable multidisciplinary data scientists with wider knowledge in computer science, biology, medicine, bioinformatics and data mining.
- To increase computing features and scalability of different omics data, 3-D memory and scalable methodologies are developed.

Big Data Analytics

- **Data acquisition** is followed by the **data cleansing** step that detects and remove before-mentioned anomalies of data.
- Then, raw data must be **transformed by data normalization and aggregation**. During data transformation **moving, cleaning, splitting, translating, merging, sorting, indexing and validating** of data are performed as sub-steps of this stage in order to make data consistent and easily accessible for further data analysis stages.
- Data integration in a unified standard form is required.
- Sensor data are generated by diverse medical devices at different sampling rate, that make data integration an important step. **Data transmission** is a technique that transfers integrated data into data storage centers, systems or distributed cloud storages for further analysis.
- **Data reduction** is a method that make size reduction of the large data sets in order that suitable data mining techniques and application could be applied. Moreover, some data mining techniques require **discretization** of the continuous attribute intervals, that means discretization techniques should be applied on the data, prior to further **suitable analysis, interpretation and visualization.**

Big Data Analytics

- Since physicians use voluminous clinical notes and unstructured text data, in order to perform optimized full-text search of clinical data, searching and indexing tools are needed.
- Recent advance in healthcare sensor devices elicits the data processing from diverse data sources to be carried out in real-time.
- To discover correlations, patterns and previously unknown knowledge in the large datasets and databases, appropriate data mining techniques should be applied.
- **Data mining algorithms** in medicine and healthcare enable analyze, detection and prediction of specific diseases and hence aid practitioners to make decisions about early disease detection. Such early disease prediction helps to decide and use most suitable treatment considering the symptoms, patient EHR and treatment history.
- Data mining algorithms can be
  - **supervised**
  - **semi-supervised and**
  - **unsupervised learning algorithms**

Big Data Analytics

- Data **visualization** tools in healthcare aids to identify **patterns, trends, outliers, clusters and time series analysis to improve clinical healthcare services and public health policy.**
- Visualization generates outputs like numerous visualization reports (e.g. charts, interactive dashboards), real-time reporting information (e.g. diverse notifications, alerts and key performance indicators) and clinical summaries (e.g. historical reports, statistical analyses, time series comparisons).
- Usually hospitals and clinics have integrated information systems that produce a huge volume data suitable for data analysis, but with restricted possibilities, generating data for decision support for their stakeholders and defining interactions and workflow systems in order to provide healthcare quality systems, patient data privacy as well as optimization of healthcare costs.
- They also have an enormous quantity of reports with statistical data analysis and periodical reports.

Big Data Analytics

- The most important case for data analysis is bringing relevant information for decision making associated with the disease diagnoses, treatments and interventions. Also, there are large amounts of data related to the need of communication help from medical practitioners from different healthcare institutions, to the EHR between two conditions of diagnoses.
- Plans for treatments and storing of data about patient lab tests, plan for creating appropriate healthcare services and medical documents required for the national healthcare bodies.
- The second important reason and motivation of storing huge data sets is administrative healthcare data that supports reimbursement’s procedures, considering the data collected from various healthcare services, interventions and treatments as well as the expenses for these services. The healthcare institutions’ managers have to plan and control the working processes and these data should aid them to increase the transparency and enhance the decision making processes.
- Many implications arise from the need of evidence-based medicine and patient care when healthcare legislation and law are highlighted. There are many law procedures for insufficient document quantities that can have negative implications for healthcare institutions.

Healthcare Information System based on Big Data

- Besides these layers, the frameworks have to address privacy and security issues of the system on the several tiers.
  - Sensitivity tier ensures patient information such as disease name and its status, patient mental status and biometric identifiers.
  - Security tier authenticates patient information such as name, date of birth, doctor name etc. To prevent privacy and security of patient data, the system should adopt a two-level security mechanism.
  - The first security level relates to the authorization of users towards accessing patient data in the hospital by providing temporary user and patient identifiers.
  - To access the patient data at the inter- and intra-hospital levels, an OTP-based security mechanism level should be employed.
- Depending on the purpose of the analysis and data types, analysis of big data is divided into three components: Hadoop MapReduce, stream computing and in-database analytics.

Healthcare Information System based on Big Data

- As a result of the voluminous big data and various data formats, new NoSQL DBMSs are required to integrate and retrieve data sets and to enable data transfer from standard into new operating systems.
- Hadoop MapReduce is an SQL-based programming model, which is able to process large amount of data sets through a Hadoop cluster by provided parallelization, distribution and scheduling services.
- MapReduce allows analysis of structured, semi-structure and unstructured data in a massive parallel processing (MPP) environment Apache Hive is a relational model for querying, searching, analyzing huge data sets stored in Hadoop Distributed File System (HDFS). It uses HiveQL as a query language that transforms typical SQL queries into MapReduce tasks.
- To store data in distributed and scalable DBs, Apache HBase is a suitable system.

Healthcare Information System based on Big Data

- Stream computing supports high-performance big data processing in real time or almost real time.
- Stream computing analysis of healthcare big data can respond to unexpected events that occur, such as customer account misusing and to determine quickly the most appropriate actions.
- Suitable non-Hadoop processing tools for streaming data processing are Spark, Hive, Storm and GraphLab.
- In-database analytics provides high-speed parallel processing, scalability and optimization ensuring secure environment for confidential data.
- Results of in-database analysis are not real time and this analysis in healthcare supports preventive healthcare practice and evident-based medicine.

Big Data Security, Privacy and Governance

- Medical and healthcare data are generated from diverse multiple sources, that means that patients' data security is a big concern, as well their privacy due to the major risk of data leakage since their massive usage of third-party infrastructures and services.
- Cloud storage of these data can be vulnerable by potential malicious outsiders who can access to the cloud platform and for instance, can act man-in the middle attacks.
- Confidentiality of the data can be attained by protecting data from accidental and unauthorized users.
- When big data are stored in databases, encryption methods for data protection can be categorized into table encryption, disk encryption and data encryption.
- Another major legal and ethical issues are related to the **ownership** of data and developed applications that employ patients' data. Particularly, whether patients' data, which are used for development and validation of application and analytics models, can be reused, shared and/or even sold.
- Patients, whose data are crucial for application development and validation, should be pre-awared that the data cannot be reused for other purposes and will not be misused.

Big Data Security, Privacy and Governance

- It is essential for the healthcare organization to secure personal data and to address the risks and legal responsibilities associated with personal data processing, according to the valid national and international laws, policies and regulation for data privacy.
- Storing healthcare big data in a public cloud, which is a cost saving alternative, requires solving security risks and patient privacy control since the data access is controlled by third parties.
- Differently, storing data in a private cloud, which keeps the sensitive data in-house, is a more secure option, but is a more expensive choice. This means that healthcare managers have to make a trade-off between the project budget and security and privacy of sensitive patient data.
- **Data governance** is composed of:
  - **master data management** refers to the governance, policies, standards and tools for data management
  - **life-cycle management** manages business data lifecycle from achieving data, warehousing data, testing and providing diverse application systems.
  - **data security and privacy management** delivers activities related to discovery, monitoring, configuration appraisal, auditing and protection of healthcare big data

Thank you for your attention!

Do you have any questions?