Title:

Data-Driven Health Care Systems Using Hadoop Ecosystem Components

Problem Statement:

The healthcare industry generates massive amounts of data, including patient records, medical imaging, research studies, and clinical trials. The challenge lies in efficiently processing and analyzing this vast volume of data to derive meaningful insights and improve healthcare outcomes. The goal of this case study is to demonstrate how the Hadoop ecosystem components can be leveraged to develop a data-driven healthcare system that can handle large-scale data processing and analytics tasks.

Date: 15/05/2023

Objectives:

1. Implement a scalable and distributed data storage system using HDFS to store healthcare data.

2. Utilize YARN for resource management and job scheduling to ensure efficient processing of data.

3. Apply the MapReduce programming paradigm for distributed data processing and analysis.

4. Explore Spark for in-memory data processing and advanced analytics.

5. Utilize PIG and HIVE for query-based processing and data service functionalities.

6. Integrate HBase as a NoSQL database for real-time reads and writes of healthcare data.

7. Utilize Mahout and Spark MLLib to apply machine learning algorithms for predictive analytics and decision support.

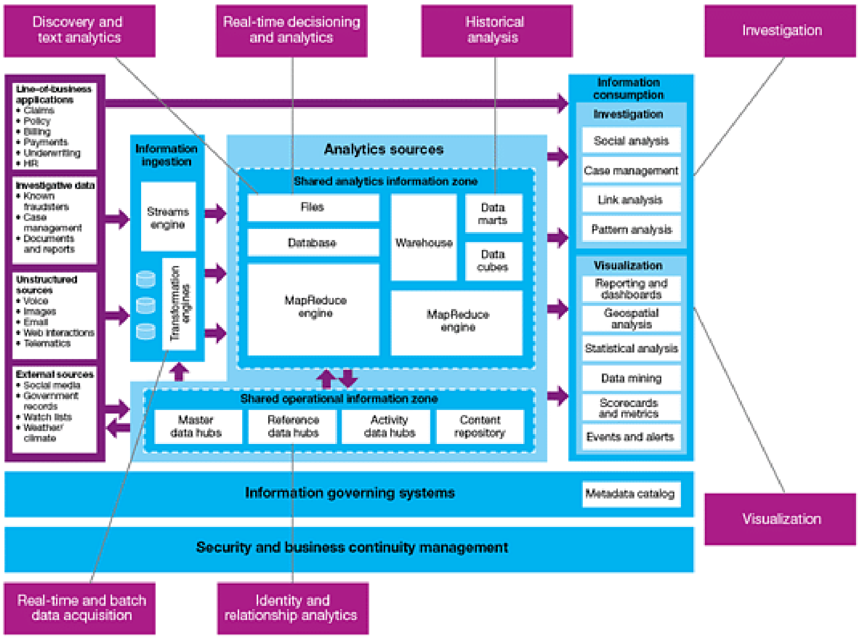
8. Utilize Solar and Lucene for efficient searching and indexing of healthcare data.

Theory:

The Hadoop ecosystem components offer a comprehensive set of tools and technologies for big data processing and analytics. HDFS provides a fault-tolerant distributed file system capable of storing and managing large volumes of data across a cluster of commodity hardware. YARN acts as a resource manager and job scheduler, enabling efficient allocation of resources for data processing tasks. MapReduce is a programming model that allows distributed processing of data in parallel across multiple nodes in a cluster. Spark, on the other hand, provides in-memory data processing capabilities, enabling faster and more efficient analytics. PIG and HIVE offer query-based processing and data service functionalities, simplifying data manipulation and analysis tasks. HBase serves as a scalable NoSQL database capable of handling real-time read and writes operations. Mahout and Spark MLLib provide machine learning algorithm libraries for implementing predictive analytics and decision support systems. Solar and Lucene facilitate efficient searching and indexing of healthcare data.

System Architecture:

At least 10% of Healthcare insurance payments are attributed to fraudulent claims. Worldwide this is estimated to be a multi-billion dollar problem. Fraudulent claims are not a novel problem but the complexity of the insurance frauds seems to be increasing exponentially making it difficult for the healthcare insurance companies to deal with them.



Big Data Analytics helps healthcare insurance companies find different ways to identify and prevent fraud at an early stage. Using Hadoop technology, insurance companies have been successful in developing predictive models to identify fraudsters by making use of real-time and historical data on medical claims, weather data, wages, voice recordings, demographics, cost of attorneys, and call center notes. Hadoop's capability to store large unstructured data sets in NoSQL databases and use MapReduce to analyze this data helps in the analysis and detection of patterns in the field of Fraud Detection.

The upswing for big data in the healthcare industry is due to the falling cost of storage. As early as 5 years ago, the cost of a scalable relational database with a permanent software license was $100,000 per TB along with an additional cost of $20,000 per year for support and maintenance. Now with the advent of Hadoop in Big Data Analytics, it is possible to store, manage and analyze the same amount of data with a yearly subscription of just $1,200. The increasing demand for using Hadoop technology in Healthcare will eliminate the concept of a "one size fits all" kind of medicines and treatments in the healthcare industry. The coming years will see the Healthcare industry provide personalized patient medications at controlled costs.

Methodology:

1. Data Ingestion: Healthcare data from various sources, such as patient records, medical imaging, and clinical trials, are ingested into the HDFS storage system.

2. Data Processing: MapReduce and Spark are utilized to process and analyze healthcare data at scale. This includes data cleansing, transformation, and aggregation tasks.

3. Querying and Analysis: PIG and HIVE are used for query-based processing and data service functionalities. Complex queries are executed to extract specific information and perform analytical tasks.

4. Real-time Operations: HBase is employed to enable real-time reads and writes of healthcare data, supporting instant access to critical patient information.

5. Machine Learning: Mahout and Spark MLLib are leveraged to apply machine learning algorithms on healthcare data for predictive analytics, disease diagnosis, and treatment recommendations.

6. Searching and Indexing: Solar and Lucene are utilized for efficient searching and indexing of healthcare data, enabling quick retrieval of relevant information.

Results:

The implementation of the data-driven health care system using Hadoop ecosystem components has demonstrated several benefits, including:

- Scalability: The system can handle large-scale data processing and analytics tasks, accommodating the ever-increasing volume of healthcare data.

- Performance: In-memory data processing with Spark has improved the speed and efficiency of analytics, enabling real-time insights.

- Flexibility: The integration of various components like HBase, PIG, HIVE, and Mahout provides a diverse range of