

UNIT 3

BIG DATA FOR BUSINESS PERSPECTIVE

The development of big data technologies unlocked a treasure trove of information for businesses. Before that, BI and analytics applications were mostly limited to structured data stored in relational databases and data warehouses -- transactions and financial records, for example.

A lot of potentially valuable data that didn't fit the relational mold was left unused. No more, though.

Before big data platforms and tools were developed, many organizations could use only a small fraction of their data in operational and analytics applications.

The rest often got pushed to the side as so-called *dark data*, which is processed and stored but not put to further use. Effective big data management processes enable businesses to better utilize their data assets.

Being able to do so expands the kinds of data analytics that companies can run and the business value they can get. Big data creates increased opportunities for machine learning, predictive analytics, data mining, streaming analytics, text mining and other data science and advanced analytics disciplines.

Using those disciplines, big data analytics applications help businesses better understand customers, identify operational issues, detect fraudulent transactions and manage supply chains, among other uses.

If done well, the end results include more effective marketing and advertising campaigns, improved business processes, increased revenue, reduced costs and stronger strategic planning -- all of which can lead to better financial results and competitive advantages over business rivals.

In addition, big data contributes to breakthroughs in medical diagnoses and treatments, scientific research and smart city initiatives, law enforcement and other government programs.

Characteristics of Big Data

Big data is a collection of data from many different sources and is often describe by five characteristics: volume, value, variety, velocity, and veracity.

- **Volume:** the size and amounts of big data that companies manage and analyze
- **Value:** the most important “V” from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits
- **Variety:** the diversity and range of different data types, including unstructured data, semi-structured data and raw data
- **Velocity:** the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time
- **Veracity:** the “truth” or accuracy of data and information assets, which often determines executive-level confidence

The additional characteristic of variability can also be considered:

- **Variability:** the changing nature of the data companies seek to capture, manage and analyze – e.g., in sentiment or text analytics, changes in the meaning of key words or phrases

IMPORTANCE OF BIG DATA

Big Data importance doesn't revolve around the amount of data a company has. Its importance lies in the fact that how the company utilizes the gathered data.

Every company uses its collected data in its own way. More effectively the company uses its data, more rapidly it grows.

The companies in the present market need to collect it and analyze it because:

1. Cost Savings

Big Data tools like Apache Hadoop, Spark, etc. bring cost-saving benefits to businesses when they have to store large amounts of data. These tools help organizations in identifying more effective ways of doing business.

2. Time-Saving

Real-time in-memory analytics helps companies to collect data from various sources. Tools like Hadoop help them to analyze data immediately thus helping in making quick decisions based on the learnings.

3. Understand the market conditions

Big Data analysis helps businesses to get a better understanding of market situations.

For example, analysis of customer purchasing behavior helps companies to identify the products sold most and thus produces those products accordingly. This helps companies to get ahead of their competitors.

4. Social Media Listening

Companies can perform sentiment analysis using Big Data tools. These enable them to get feedback about their company, that is, who is saying what about the company.

Companies can use Big data tools to improve their online presence.

5. Boost Customer Acquisition and Retention

Customers are a vital asset on which any business depends on. No single business can achieve its success without building a robust customer base. But even with a solid customer base, the companies can't ignore the competition in the market.

If we don't know what our customers want then it will degrade companies' success. It will result in the loss of clientele which creates an adverse effect on business growth.

Big data analytics helps businesses to identify customer related trends and patterns. Customer behavior analysis leads to a profitable business.

6. Solve Advertisers Problem and Offer Marketing Insights

Big data analytics shapes all business operations. It enables companies to fulfill customer expectations. Big data analytics helps in changing the company's product line. It ensures powerful marketing campaigns.

7. The driver of Innovations and Product Development

Big data makes companies capable to innovate and redevelop their products.

USE CASES OF BIG DATA

1. Product or service development

Companies like P&G and Netflix use big data to analyze target audience groups, social media, and test markets before they plan and launch new products or services. They use big data to anticipate customer demand, classify the key attributes of past and current products or services, and then build predictive models for their new products and services.

2. Create successful email marketing campaigns

Email marketers have a tremendous opportunity for leveraging big data. It can help them deliver high impact, create relevant content to the right audience segments through the right delivery channel at the right time. Create email campaigns that are highly relevant to your subscribers and generate real brand engagement.

3. Improve customer service

Big data enables companies to collect data from social media, web visits, call logs, and other sources and analyze it to improve the customer experience. Marketing and sales teams can deliver personalized offers, reduce customer churn, and handle issues proactively with the help of big data.

4. Drive innovation and creativity.

Big data can provide a company with all the necessary insights to develop new and innovative ideas for products or services or even their hashtag#marketing campaigns. The right data insight on industry trends can improve decisions about financial and planning considerations.

5. Enhance customer retention and acquisition

Big data allows businesses to understand better their perspective customers' needs, interests, behavioral patterns, and usage patterns. These data insights are crucial if a company wants to enhance its offerings, increase conversion rates, generate qualified leads, and retain its existing customers.

6. Better cybersecurity and fraud prevention

Security landscapes and compliance requirements are constantly evolving. Companies can use big data analytics to identify patterns in data that indicate fraud or spot oddities in system behavior. Big data systems can sift through enormous transactional data on servers, databases, apps, files, and devices to identify and mitigate possible fraud.

7. Predictive maintenance

Big data allows hashtag#professionals to detect operational anomalies and potential defects in equipment and processes to address them before failure occurs. The team can deploy more cost-effective maintenance, maximizing uptime for parts and equipment.

DATA IN WAREHOUSE AND DATA IN HADOOP

1. Data Warehouse :

It is a technique for gathering and managing information from different sources to supply significant commercial enterprise insights. A Data warehouse is commonly used to join and analyze commercial enterprise information from heterogeneous sources. It acts as the heart of the BI system which is constructed for data evaluation and reporting.

2. Hadoop :

It is an open-source software program framework for storing information and strolling applications on clusters of commodity hardware. It offers large storage for any sort of data, extensive processing strength, and the potential to deal with actually limitless concurrent duties or jobs.

DIFFERENCE BETWEEN DATA WAREHOUSE AND HADOOP:

S.No .	Data Warehouse	Hadoop
1.	In this, we first analyze the data and then further do the processing.	It can process various types of data such as Structured data, unstructured data, or raw data.
2.	It is convenient for storing a small volume of data.	It deals with a large volume of data.
3.	It uses schema-for-write logic to process the data.	It deals with schema-for-read logic to process the data.
4.	It is very less agile as compared to Hadoop.	It is more agile as compared to Data Warehouse.
5.	It is of fixed configuration.	It can be configured or reconfigured, accordingly.
6.	It has high security for storing different data.	Security is a great concern and It is improving and working on it.
7.	It is mainly used by business professionals.	It mainly deals with Data Engineering and Data Science.

IMPORTANCE OF BIG DATA

Big Data importance doesn't revolve around the amount of data a company has. Its importance lies in the fact that how the company utilizes the gathered data.

Every company uses its collected data in its own way. More effectively the company uses its data, more rapidly it grows.

The companies in the present market need to collect it and analyze it because:

1. Cost Savings

Big Data tools like Apache Hadoop, Spark, etc. bring cost-saving benefits to businesses when they have to store large amounts of data. These tools help organizations in identifying more effective ways of doing business.

2. Time-Saving

Real-time in-memory analytics helps companies to collect data from various sources. Tools like Hadoop help them to analyze data immediately thus helping in making quick decisions based on the learnings.

3. Understand the market conditions

Big Data analysis helps businesses to get a better understanding of market situations.

For example, analysis of customer purchasing behavior helps companies to identify the products sold most and thus produces those products accordingly. This helps companies to get ahead of their competitors.

4. Social Media Listening

Companies can perform sentiment analysis using Big Data tools. These enable them to get feedback about their company, that is, who is saying what about the company.

Companies can use Big data tools to improve their online presence.

5. Boost Customer Acquisition and Retention

Customers are a vital asset on which any business depends on. No single business can achieve its success without building a robust customer base. But even with a solid customer base, the companies can't ignore the competition in the market.

6. Solve Advertisers Problem and Offer Marketing Insights

Big data analytics shapes all business operations. It enables companies to fulfill customer expectations. Big data analytics helps in changing the company's product line. It ensures powerful marketing campaigns.

7. The driver of Innovations and Product Development

Big data makes companies capable to innovate and redevelop their products.

BIG DATA USE CASE

Here are some examples of Big Data applications that affect people every day.

1. Transportation
2. Advertising and Marketing
3. Banking and Financial Services
4. Government

5. Media and Entertainment
6. Meteorology
7. Healthcare
8. Cybersecurity
9. Education
1. **Transportation**

Big Data powers the GPS smartphone applications most of us depend on to get from place to place in the least amount of time. GPS data sources include satellite images and government agencies.

Airplanes generate enormous volumes of data, on the order of 1,000 gigabytes for transatlantic flights. Aviation analytics systems ingest all of this to analyze fuel efficiency, passenger and cargo weights, and weather conditions, with a view toward optimizing safety and energy consumption.

Big Data simplifies and streamlines transportation through:

- Congestion management and traffic control
Thanks to Big Data analytics, Google Maps can now tell you the least traffic-prone route to any destination.
- Route planning
Different itineraries can be compared in terms of user needs, fuel consumption, and other factors to plan for maximize efficiency.
- Traffic safety
Real-time processing and predictive analytics are used to pinpoint accident-prone areas.

2. Advertising and Marketing

Ads have always been targeted towards specific consumer segments. In the past, marketers have employed TV and radio preferences, survey responses, and focus groups to try to ascertain people's likely responses to campaigns. At best, these methods amounted to educated guesswork.

Today, advertisers buy or gather huge quantities of data to identify what consumers actually click on, search for, and "like." Marketing campaigns are also monitored for effectiveness using click-through rates, views, and other precise metrics.

For example, Amazon accumulates massive data stories on the purchases, delivery methods, and payment preferences of its millions of customers. The company then sells ad placements that can be highly targeted to very specific segments and subgroups.

3. Banking and Financial Services

The financial industry puts Big Data and analytics to highly productive use, for:

- Fraud detection
Banks monitor credit cardholders' purchasing patterns and other activity to flag atypical movements and anomalies that may signal fraudulent transactions.

- Risk management
Big Data analytics enable banks to monitor and report on operational processes, KPIs, and employee activities.
- Customer relationship optimization
Financial institutions analyze data from website usage and transactions to better understand how to convert prospects to customers and incentivize greater use of various financial products.
- Personalized marketing
Banks use Big Data to construct rich profiles of individual customer lifestyles, preferences, and goals, which are then utilized for micro-targeted marketing initiatives.

4. Government

Government agencies collect voluminous quantities of data, but many, especially at the local level, don't employ modern data mining and analytics techniques to extract real value from it.

Examples of agencies that do include the IRS and the Social Security Administration, which use data analysis to identify tax fraud and fraudulent disability claims. The FBI and SEC apply Big Data strategies to monitor markets in their quest to detect criminal business activities. For years now, the Federal Housing Authority has been using Big Data analytics to forecast mortgage default and repayment rates.

5. Media and Entertainment

The entertainment industry harnesses Big Data to glean insights from customer reviews, predict audience interests and preferences, optimize programming schedules, and target marketing campaigns.

Two conspicuous examples are Amazon Prime, which uses Big Data analytics to recommend programming for individual users, and Spotify, which does the same to offer personalized music suggestions.

6. Meteorology

Weather satellites and sensors all over the world collect large amounts of data for tracking environmental conditions. Meteorologists use Big Data to:

- Study natural disaster patterns
- Prepare weather forecasts
- Understand the impact of global warming
- Predict the availability of drinking water in various world regions
- Provide early warning of impending crises such as hurricanes and tsunamis

7. Healthcare

Big Data is slowly but surely making a major impact on the huge healthcare industry. Wearable devices and sensors collect patient data which is then fed in real-time to individuals' electronic health records. Providers and practice organizations are now using Big Data for a number of purposes, including these:

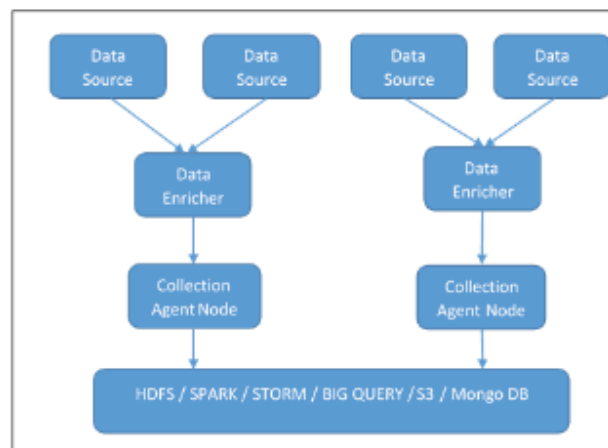
- Prediction of epidemic outbreaks
- Early symptom detection to avoid preventable diseases
- Electronic health records
- Real-time alerting
- Enhancing patient engagement
- Prediction and prevention of serious medical conditions

BIG DATA PATTERN

- Multisource extractor
- Multidestination
- Protocol converter
- Just-in-time (JIT) transformation
- Real-time streaming pattern

Multisource extractor

An approach to ingesting multiple data types from multiple data sources efficiently is termed a *Multisource extractor*. Efficiency represents many factors, such as data velocity, data size, data frequency, and managing various data formats over an unreliable network, mixed network bandwidth, different technologies, and systems:



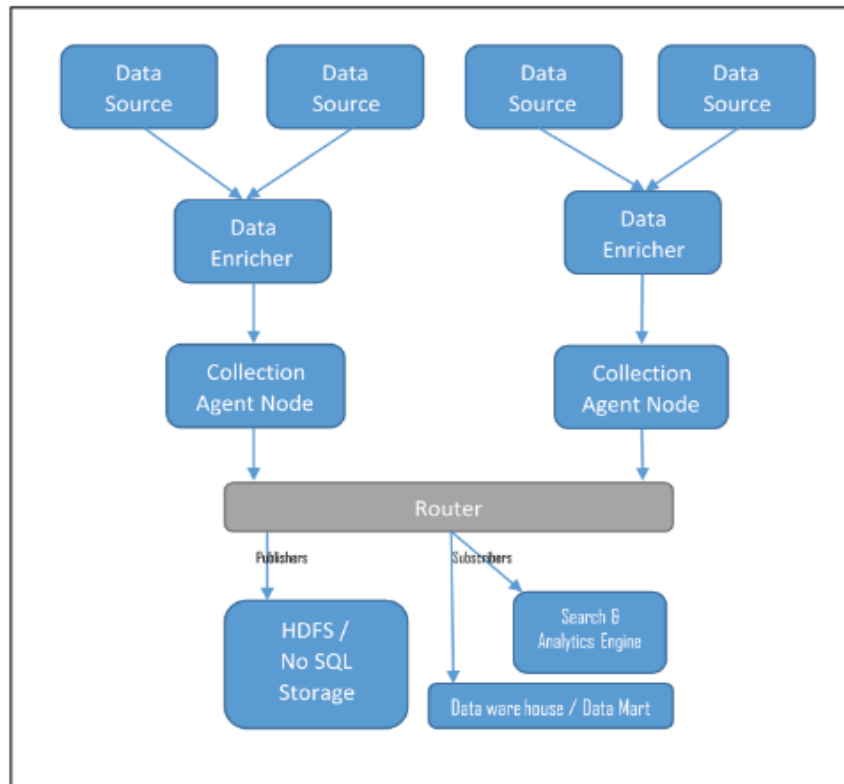
The following are the benefits of the multisource extractor:

- Provides reasonable speed for storing and consuming the data
- Better data prioritization and processing
- Drives improved business decisions
- Decoupled and independent from data production to data consumption
- Data semantics and detection of changed data

- Scalable and fault tolerance system

Multidestination pattern

In multisourcing, we saw the raw data ingestion to HDFS, but in most common cases the enterprise needs to ingest raw data not only to new HDFS systems but also to their existing traditional data storage, such as Informatica or other analytics platforms. In such cases, the additional number of data streams leads to many challenges, such as storage overflow, data errors (also known as data regret), an increase in time to transfer and process data, and so on.

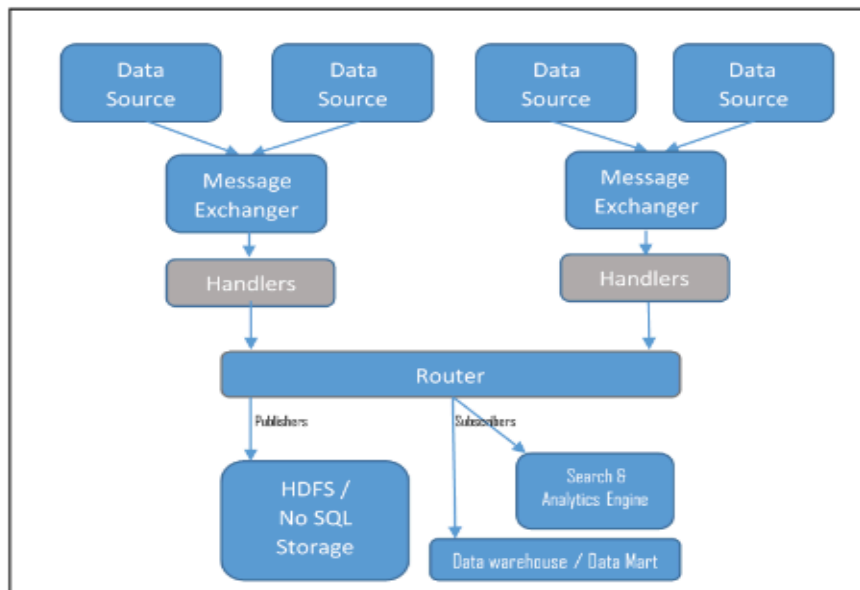


The following are the benefits of the multidestination pattern:

- Highly scalable, flexible, fast, resilient to data failure, and cost-effective
- Organization can start to ingest data into multiple data stores, including its existing RDBMS as well as NoSQL data stores
- Allows you to use simple query language, such as Hive and Pig, along with traditional analytics
- Provides the ability to partition the data for flexible access and decentralized processing
- Possibility of decentralized computation in the data nodes

Protocol converter

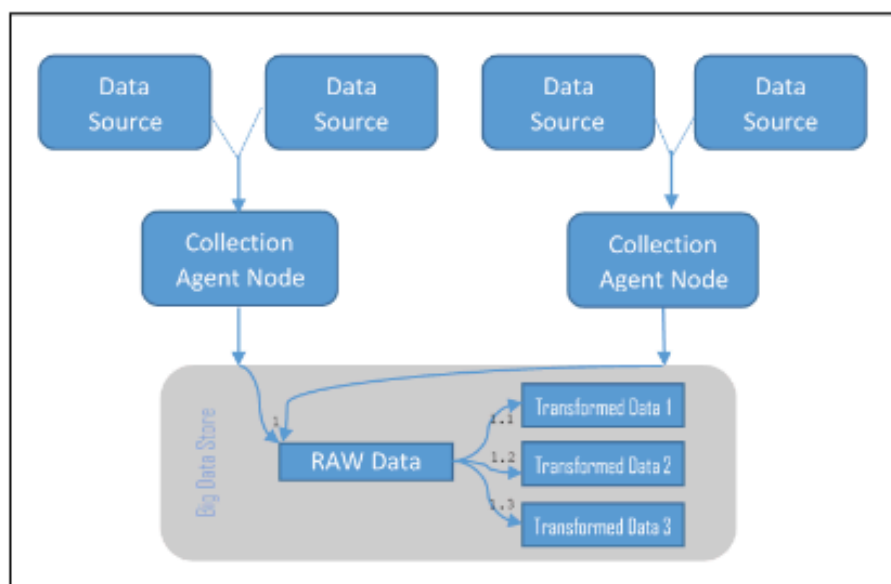
This is a mediatory approach to provide an abstraction for the incoming data of various systems. The protocol converter pattern provides an efficient way to ingest a variety of unstructured data from multiple data sources and different protocols.



In the protocol converter pattern, the ingestion layer holds responsibilities such as identifying the various channels of incoming events, determining incoming data structures, providing mediated service for multiple protocols into suitable sinks, providing one standard way of representing incoming messages, providing handlers to manage various request types, and providing abstraction from the incoming protocol layers.

JUST-IN-TIME (JIT) TRANSFORMATION PATTERN

The JIT transformation pattern is the best fit in situations where raw data needs to be preloaded in the data stores before the transformation and processing can happen. In this kind of business case, this pattern runs independent preprocessing batch jobs that clean, validate, correlate, and transform, and then store the transformed information into the same data store (HDFS/NoSQL); that is, it can coexist with the raw data:

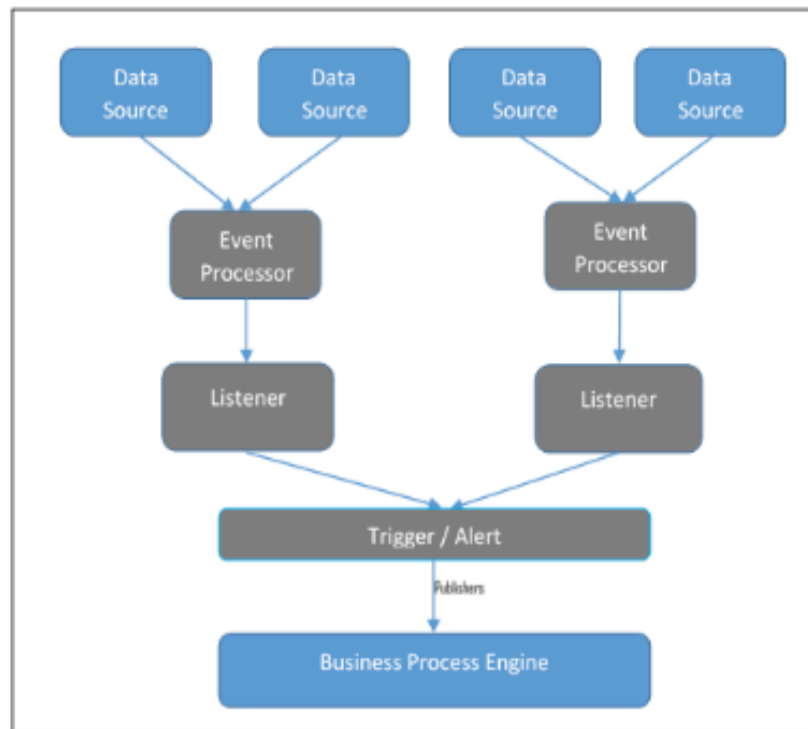


Real-time streaming pattern

Most modern businesses need continuous and real-time processing of unstructured data for their enterprise big data applications.

Real-time streaming implementations need to have the following characteristics:

- Minimize latency by using large in-memory
- Event processors are atomic and independent of each other and so are easily scalable
- Provide API for parsing the real-time information
- Independent deployable script for any node and no centralized master node implementation



Apache Hadoop Applications

1. Finance sectors

Financial organizations use Hadoop for fraud detection and prevention. They use Apache Hadoop for reducing risk, identifying rogue traders, analyzing fraud patterns. Hadoop helps them to precisely target their marketing campaigns on the basis of customer segmentation.

Hadoop helps financial agencies to improve customer satisfaction. Credit card companies also use Apache Hadoop for finding out the exact customer for their product.

2. Security and Law Enforcement

The USA national security agency uses Hadoop in order to prevent terrorist attacks and to detect and prevent cyber-attacks. Big Data tools are used by the Police forces for catching criminals and even predicting criminal activity. Hadoop is used by different public sector fields such as defense, intelligence, research, cybersecurity, etc.

3. Companies use Hadoop for understanding customers requirements

The most important application of Hadoop is understanding Customer's requirements.

Different companies such as finance, telecom use Hadoop for finding out the customer's requirement by examining a big amount of data and discovering useful information from these vast amounts of data. By understanding customers behaviors, organizations can improve their sales.

4. Hadoop Applications in Retail industry

Retailers both online and offline use Hadoop for improving their sales. Many e-commerce companies use Hadoop for keeping track of the products bought together by the customers. On the basis of this, they provide suggestions to the customer to buy the other product when the customer is trying to buy one of the relevant products from that group.

For example, when a customer tries to buy a mobile phone, then it suggests a customer for the mobile back cover, screen guard.

Also, Hadoop helps retailers to customize their stocks based on the predictions that came from different sources such as Google search, social media websites, etc. Based on these predictions retailers can make the best decision which helps them to improve their business and maximize their profits.

5. Real-time analysis of customers data

Hadoop can analyze customer data in real-time. It can track clickstream data as it's for storing and processing high volumes of clickstream data. When a visitor visits a website, then Hadoop can capture information like from where the visitor originated before reaching a particular website, the search used for landing on the website.

Hadoop can also grab data about the other webpages in which the visitor shows interest, time spent by the visitor on each page, etc. This is the analysis of website performance and user engagement.

Enterprises of all types, by implementing Hadoop perform clickstream analysis for optimizing the user-path, predicting the next product to buy, carrying out market basket analysis, etc.

6. Uses of Hadoop in Government sectors

The government uses Hadoop for the country, states, and cities development by analyzing vast amounts of data.

For example, they use Hadoop for managing traffic in the streets, for the development of smart cities, or for improving transportation in the city.

7. Hadoop Uses in Advertisements Targeting Platforms

Advertisements Targeting Platforms use Hadoop for capturing and analyzing clickstream, video, transaction, and social media data. They analyze the data generated by various social media websites such as Facebook, Twitter, Instagram, etc. and then target their interested audience.

GETTING YOUR DATA IN HADOOP:

Data is ubiquitous, but that does not always mean that it's easy to store and access. In fact, many existing pre-Hadoop data architectures tend to be rather strict and therefore difficult to work with and make changes to.

what is a data lake?

With the more traditional database or data warehouse approach, adding data to the database requires data to be transformed into a *pre-determined* schema before it can be loaded into the database. This step is often called "extract, transform, and load" (ETL) and often consumes a lot of time, effort, and expense before the data can be used for downstream applications.

More importantly, decisions about how the data will be used must be made during the ETL step, and later changes are costly. In addition, data are often discarded in the ETL step because they do not fit into the data schema or are deemed un-needed or not valuable for downstream applications.

One of the basic features of Hadoop is a central storage space for all data in the Hadoop Distributed File Systems (HDFS), which make possible inexpensive and redundant storage of large datasets at a much lower cost than traditional systems.

This enables the Hadoop data lake approach, wherein all data are often stored in raw format, and what looks like the ETL step is performed when the data are processed by Hadoop applications.

This approach, also known as schema on read, enables programmers and users to enforce a structure to suit their needs when they access data. The traditional data warehouse approach, also known as schema on write, requires more upfront design and assumptions about how the data will eventually be used.

For data science purposes, the capability to keep all the data in raw format is extremely beneficial since often it is not clear up front which data items may be valuable to a given data science goal.

With respect to big data, the data lake offers three advantages over a more traditional approach:

- All data are **available**. There is no need to make any assumptions about future data use.
- All data are **sharable**. Multiple business units or researchers can use all available data¹, some of which may not have been previously available due to data compartmentalization on disparate systems.
- All **access methods** are available. Any processing engine (MapReduce, Tez, Spark) or application (Hive, Spark-SQL, Pig) can be used to examine the data and process it as needed.

To be clear, data warehouses are valuable business tools, and Hadoop is designed to complement them, not replace them. Nonetheless, the traditional data warehouse technology was developed before the data lake began to fill with such large quantities of data.

The growth of new data from disparate sources including social media, click streams, sensor data, and others is such that we are starting to quickly fill the data lake. Traditional ETL stages may not be able to keep up with the rate at which data are entering the lake. There will be overlap, and each tool will address the need for which it was designed.

