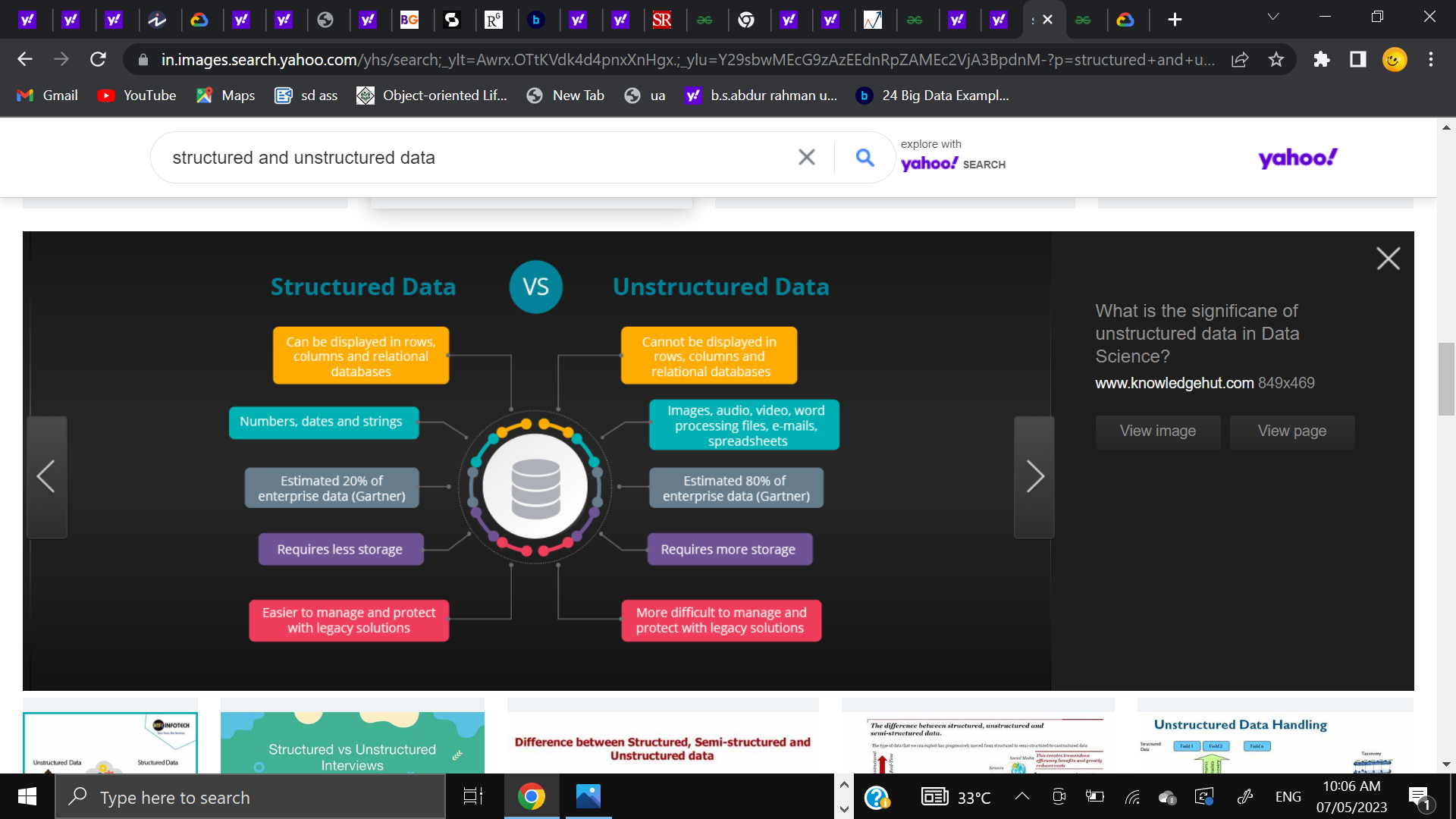
BIG DATA

What is data?

Data is measured, collected and reported, and analyzed, whereupon it is often visualized using graphs, images or other analysis tool .



**What is Big Data?**

"Big data" is the massive amount of data available to organizations that because of

its volume and complexity is not easily managed or analyzed by many business intelligence tools.

Big data overview **:**

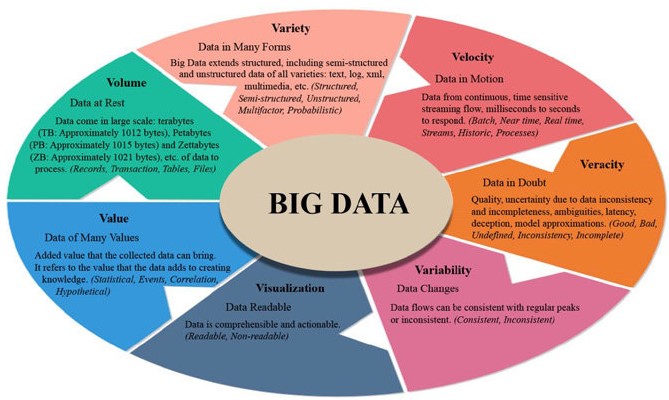
As big data emerged, so did computing models with the ability to store and manage it. centralized or distributed computing systems provide access to big data. Centralized computing means the data is stored on a central computer and processed by computing platforms like [BigQuery](https://cloud.google.com/bigquery/).

Distributed computing means big data is stored and processed on different computers, which communicate over a network. A software framework like [Hadoop](https://cloud.google.com/learn/what-is-hadoop) makes it possible to store the data and run applications to process it.

There are benefits to using centralized computing and analyzing big data where it lives, rather than extracting it for analysis from a distributed system. Insights are accessible to every user in your company—and integrated into daily workflows—when big data is housed in one place and analyzed by one platform.

## Characteristics of big data :

Big data is different from typical data assets because of its volume complexity and need for advanced business intelligence tools to process and analyze it. The attributes that define big data are volume, variety, velocity, and variability. These big data attributes are commonly referred to as the four v’s.

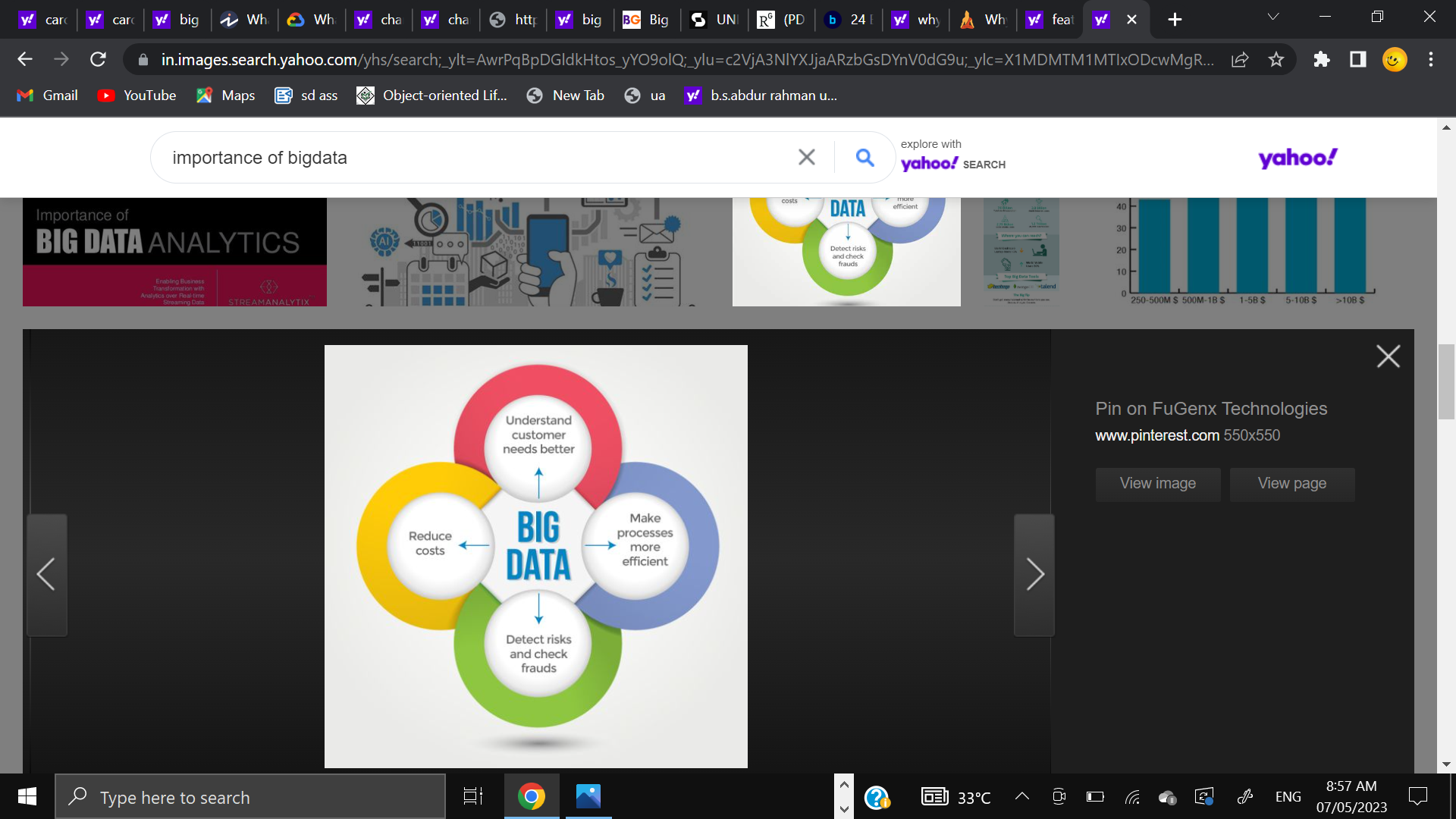


WHY BIG DATA IS IMPORTANT ?

It helps organizations:

* To understand Where, When and Why their customers buy
* Protect the company’s client base with improved loyalty programs
* Seizing cross-selling and upselling opportunities
* Provide targeted promotional information
* Optimize Workforce planning and operations
* Improve inefficiencies in the company’s supply chain
* Predict market trends
* Predict future needs
* Make companies more innovative and competitive
* It helps companies to discover new sources of revenue

IMPORTANCE OF BIG DATA:



## How to store and process Big Data?

The volume and velocity of Big Data can be huge, which makes it almost impossible to store it in traditional data warehouses. Although some and sensitive information can be stored on company premises, for most of the data, companies have to opt for cloud storage or Hadoop.

**Cloud storage** allows businesses to store their data on the internet with the help of a cloud service provider (like Amazon Web Services, Microsoft Azure, or Google Cloud Platform) who takes the responsibility of managing and storing the data. The data can be accessed easily and quickly with an API.

**Top 3 analytics tools in big data :**

1. APACHE Hadoop :

It’s a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel.Today, it is the best big data analytic tool and is popularly used by many tech giants such as Amazon, Microsoft, IBM, etc.

Features :

* Free to use and offers an efficient storage solution for businesses.
* Offers quick access via HDFS (Hadoop Distributed File System).
* Highly flexible and can be easily implemented with MySQL, and JSON.

2. Cassandra :

[**APACHE Cassandra**](https://www.geeksforgeeks.org/introduction-to-apache-cassandra/?ref=gcse) is an open-source NoSQL distributed database that is used to fetch large amounts of data. It’s one of the most popular tools for data analyticsand has been praised by many tech companies due to its high scalability and availability without compromising speed and performance. **of delivering thousands of operations every second**

**Features of APACHE Cassandra:**

* *Data Storage Flexibility:* It supports all forms of data i.e. structured, unstructured, semi-structured, and allows users to change as per their needs.
* *Data Distribution System:* Easy to distribute data with the help of replicating data on multiple data centers.
* *Fast Processing:* Cassandra has been designed to run on efficient commodity hardware and also offers fast storage and data processing.
* *Fault-tolerance:* The moment, if any node fails, it will be replaced without any delay.

3. Qubole :

It’s an open-source big data tool that helps in fetching data in a value of chain using ad-hoc analysis in machine learning. Qubole is a data lake platform that offers end-to-end service with reduced time and effort which are required in moving data pipelines.

**Features of Qubole:**

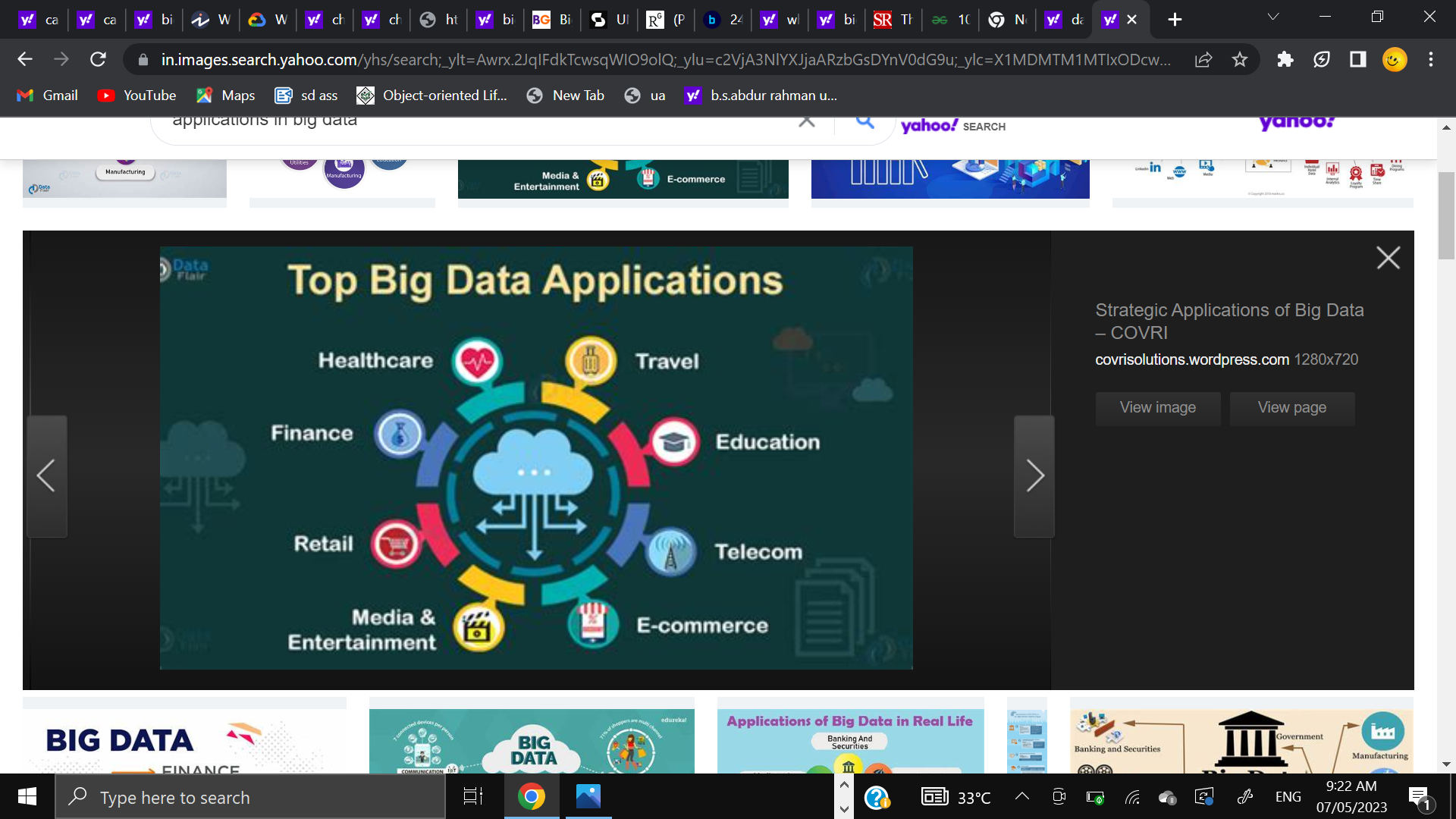
* *Supports ETL process:* It allows companies to **migrate data from multiple sources in one place**.
* *Real-time Insight:* It monitors user’s systems and allows them to view real-time insights
* *Predictive Analysis:* Qubole offers predictive analysis so that companies can take actions accordingly for targeting more acquisitions.

DAY TO DAY EXAMPLES:

**Demand forecasting** has become more accurate with more and more data being collected about customer purchases. This helps companies build forecasting models, that help them forecast future demand, and scale production accordingly. It helps companies, especially those in manufacturing businesses, to reduce the cost of storing unsold inventory in warehouses.

Big data also has extensive use in applications such as product development and fraud detection. Other examples such as:

* Predictive Inventory Ordering
* Discovering Consumer shopping habits
* Mainstream Media Streaming
* Live Road Mapping for Autonomous Vehicles



## Challenges

**1. Data growth**

Managing datasets having terabytes of information can be a big challenge for companies. As datasets grow in size, storing them not only becomes a challenge but also becomes an expensive affair for companies.

To overcome this, companies are now starting to pay attention to data compression and de-duplication. Data **compression**reduces the number of bits that the data needs, resulting in a reduction in space being consumed. Data **de-duplication**is the process of making sure duplicate and unwanted data does not reside in our database.

**2. Data security**

Data security is often prioritized quite low in the Big Data workflow, which can backfire at times. With such a large amount of data being collected, security challenges are bound to come up sooner or later.

Mining of sensitive information, fake data generation, and lack of cryptographic protection (encryption) are some of the challenges businesses face when trying to adopt Big Data techniques.

## Big Data Examples in Marketing

Big data and [marketing](https://builtin.com/marketing) go hand-in-hand, as businesses harness consumer information to forecast market trends, buyer habits and other company behaviors. All of this helps businesses determine what products and services to prioritiz

1.[](https://builtin.com/company/centerfield):

**Location:**Los Angeles, California

Getting more information on customers is a great way to discover their desires and how to meet them. [Centerfield](https://www.centerfield.com/) analyzes customer data to uncover new insights into customer behavior, which influences the marketing and [sales](https://builtin.com/sales) techniques it recommends to clients. The company is able to use this information to discover new customers that fit the same patterns as existing customers.

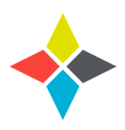
2. [](https://builtin.com/company/disqo) :

**Location:**Glendale, California

With insight help from big data, [DISQO](https://www.disqo.com/) offers products for measuring brand and customer experience. The company specializes in research and marketing lift (sales) efforts, providing API and optimization software for tracking key performance and outcome metrics. Over 125 marketing firms utilize DISQO research tools, while [over 300](https://www.disqo.com/products/ad-measurement/) firms utilize its lift solutions.

## Big Data Examples in Transportation

Navigation apps and databases, whether used by car drivers or airplane pilots, frequently rely on big data analytics to get users safely to their destinations. Insights into routes, travel time and traffic are pulled from several data points and provide a look at travel conditions and vehicle demands in real time.

1. [](https://builtin.com/company/fourkites)[FOURKITES](https://builtin.com/company/fourkites) :

**Location:** Chicago, Illinois

[FourKites](https://www.fourkites.com/)’ platform uses GPS and a host of other location data sources to track packages in real time, whether they’re crossing oceans or traveling by rail. A predictive algorithm then factors in data on traffic, weather and other external factors to calculate the estimated times of arrival for packages, so FourKites clients can give customers advance warning about delays and early deliveries — while also avoiding fees.

1. [](https://builtin.com/company/uber)

**Location:**San Francisco, California

As a rideshare company, [Uber](https://www.uber.com/) monitors its data in order to predict spikes in demand and variations in driver availability. That information allows the company to set the proper pricing of rides and provide incentives to drivers so the necessary number of vehicles are available to keep up with demand. [Data analysis](https://builtin.com/data-science/data-analysis) also forms the basis of Uber’s estimated times of arrival predictions, which goes a long way toward fulfilling customer satisfaction.