1.1 Big Data Overview

Data is created constantly, and at an ever-increasing rate. Mobile phones, social media, imaging technologies to determine a medical diagnosis—all these and more create new data, and that must be stored somewhere for some purpose. Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time. Merely keeping up with this huge influx of data is difficult, but substantially more challenging is analyzing vast amounts of it, especially when it does not conform to traditional notions of data structure, to identify meaningful patterns and extract useful information. These challenges of the data deluge present the opportunity to transform business, government, science, and everyday life.

Several industries have led the way in developing their ability to gather and exploit data:

- Credit card companies monitor every purchase their customers make and can identify fraudulent purchases with a high degree of accuracy using rules derived by processing billions of transactions.
- Mobile phone companies analyze subscribers' calling patterns to determine, for example, whether a caller's frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, the mobile phone company can proactively offer the subscriber an incentive to remain in her contract.
- For companies such as LinkedIn and Facebook, data itself is their primary product. The valuations of these companies are heavily derived from the data they gather and host, which contains more and more intrinsic value as the data grows.

Three attributes stand out as defining Big Data characteristics:

- **Huge volume of data:** Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.
- **Complexity of data types and structures:** Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.
- **Speed of new data creation and growth:** Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

Although the volume of Big Data tends to attract the most attention, generally the variety and velocity of the data provide a more apt definition of Big Data. (Big Data is sometimes described as having 3 Vs: volume, variety, and velocity.) Due to its size or structure, Big Data cannot be efficiently analyzed using only traditional databases or methods. Big Data problems require new tools and technologies to store, manage, and realize the business benefit. These new tools and technologies enable creation, manipulation, and management of large datasets and the storage environments that house them. Another definition of Big Data comes from the McKinsey Global report from 2011:**Big Data is data whose scale,**

distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.

McKinsey & Co.; Big Data: The Next Frontier for Innovation, Competition, and Productivity [1]

McKinsey's definition of Big Data implies that organizations will need new data architectures and analytic sandboxes, new tools, new analytical methods, and an integration of multiple skills into the new role of the data scientist, which will be discussed in Section 1.3. Figure 1.1 highlights several sources of the Big Data deluge.

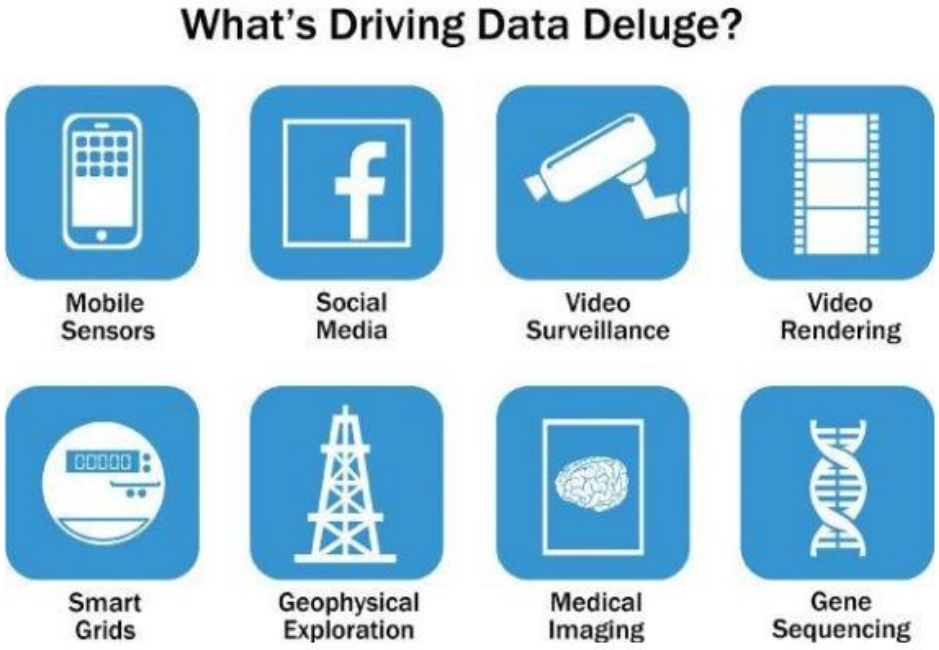


Figure 1.1 What's driving the data deluge

The rate of data creation is accelerating, driven by many of the items in Figure 1.1.

Social media and genetic sequencing are among the fastest-growing sources of Big Data and examples of untraditional sources of data being used for analysis.

For example, in 2012 Facebook users posted 700 status updates per second worldwide, which can be leveraged to deduce latent interests or political views of users and show relevant ads. For instance, an update in which a woman changes her relationship status from "single" to "engaged" would trigger ads on bridal dresses, wedding planning, or name-changing services.

Facebook can also construct social graphs to analyze which users are connected to each other as an interconnected network. In March 2013, Facebook released a new feature called "Graph Search," enabling users and developers to search social graphs for people with similar interests, hobbies, and shared locations.

Another example comes from genomics. Genetic sequencing and human genome mapping provide a detailed understanding of genetic makeup and lineage. The health care industry is looking toward these advances to help predict which illnesses a person is likely to get in his lifetime and take steps to avoid these maladies or reduce their impact through the use

of personalized medicine and treatment. Such tests also highlight typical responses to different medications and pharmaceutical drugs, heightening risk awareness of specific drug treatments.

While data has grown, the cost to perform this work has fallen dramatically. The cost to sequence one human genome has fallen from $100 million in 2001 to $10,000 in 2011, and the cost continues to drop. Now, websites such as 23andme (Figure 1.2) offer genotyping for less than $100. Although genotyping analyzes only a fraction of a genome and does not provide as much granularity as genetic sequencing, it does point to the fact that data and complex analysis is becoming more prevalent and less expensive to deploy.

Figure 1.2 Examples of what can be learned through genotyping, from 23andme.com

As illustrated by the examples of social media and genetic sequencing, individuals and organizations both derive benefits from analysis of ever-larger and more complex datasets that require increasingly powerful analytical capabilities.

1.1.1 Data Structures

1.2.3 Drivers of Big Data

To better understand the market drivers related to Big Data, it is helpful to first understand some past history of data stores and the kinds of repositories and tools to manage these data stores.

As shown in Figure 1.10, in the 1990s the volume of information was often measured in terabytes. Most organizations analyzed structured data in rows and columns and used relational databases and data warehouses to manage large stores of enterprise information. The following decade saw a proliferation of different kinds of data sources—mainly productivity and publishing tools such as content management repositories and networked attached storage systems—to manage this kind of information, and the data began to increase in size and started to be measured at petabyte scales. In the 2010s, the information that organizations try to manage has broadened to include many other kinds of data. In this era, everyone and everything is leaving a digital footprint. Figure 1.10 shows a summary perspective on sources of Big Data generated by new applications and the scale and growth rate of the data. These applications, which generate data volumes that can be measured in exabyte scale, provide opportunities for new analytics and driving new value for organizations. The data now comes from multiple sources, such as these:

- Medical information, such as genomic sequencing and diagnostic imaging
- Photos and video footage uploaded to the World Wide Web
- Video surveillance, such as the thousands of video cameras spread across a city
- Mobile devices, which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones
- Smart devices, which provide sensor-based collection of information from smart electric grids, smart buildings, and many other public and industry infrastructures
- Nontraditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing

Big data can come in multiple forms, including structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings. Contrary to much of the traditional data analysis performed by organizations, most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyze. [2] Distributed computing environments and massively parallel processing (MPP) architectures that enable parallelized data ingest and analysis are the preferred approach to process such complex data.

With this in mind, this section takes a closer look at data structures.

Figure 1.3 shows four types of data structures, with 80–90% of future data growth coming from non-structured data types. [2] Though different, the four are commonly mixed. For example, a classic Relational Database Management System (RDBMS) may store call logs for a software support call center. The RDBMS may store characteristics of the support calls as typical structured data, with attributes such as time stamps, machine type, problem type, and operating system. In addition, the system will likely have unstructured, quasi- or semi-structured data, such as free-form call log information taken from an e-mail ticket of the problem, customer chat history, or transcript of a phone call describing the technical problem and the solution or audio file of the phone call conversation. Many insights could be extracted from the unstructured, quasi- or semi-structured data in the call center data.

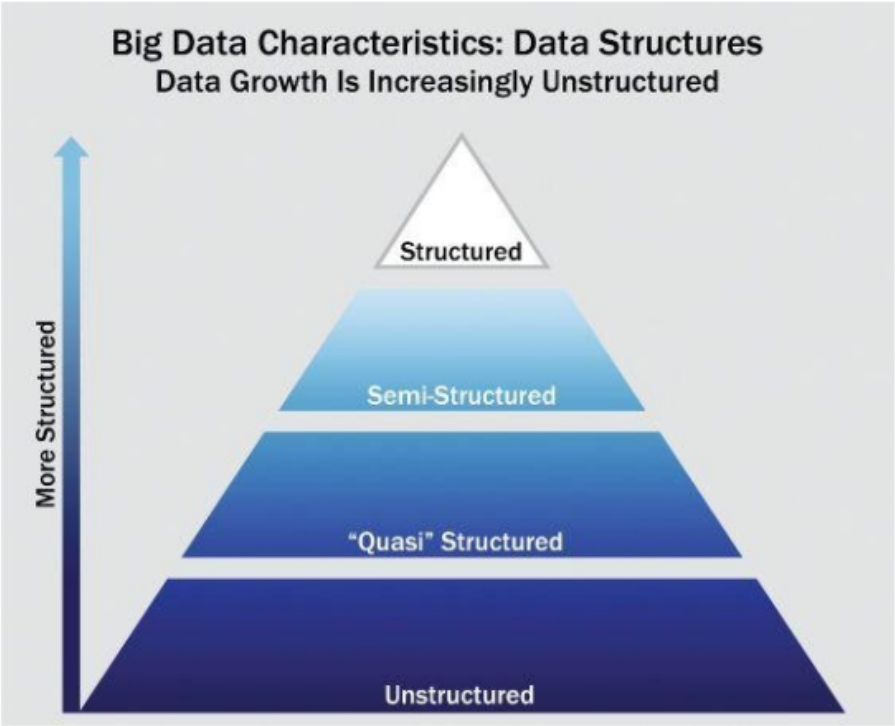


Figure 1.3 Big Data Growth is increasingly unstructured

Although analyzing structured data tends to be the most familiar technique, a different technique is required to meet the challenges to analyze semi-structured data (shown as XML), quasi-structured (shown as a clickstream), and unstructured data.

Here are examples of how each of the four main types of data structures may look.

- **Structured data:** Data containing a defined data type, format, and structure (that is, transaction data, online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spreadsheets). See Figure 1.4.
- **Semi-structured data:** Textual data files with a discernible pattern that enables parsing (such as Extensible Markup Language [XML] data files that are self-describing and defined by an XML schema). See Figure 1.5.
- **Quasi-structured data:** Textual data with erratic data formats that can be formatted with effort, tools, and time (for instance, web clickstream data that may contain inconsistencies in data values and formats). See Figure 1.6.
- **Unstructured data:** Data that has no inherent structure, which may include text documents, PDFs, images, and video. See Figure 1.7.

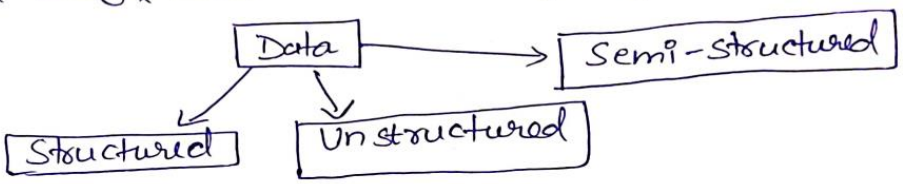
SUMMER FOOD SERVICE PROGRAM 1]

(Data as of August 01, 2011)

Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

Figure 1.4 Example of structured data

* Types of Big Data :- (Forms of Big Data)



(1) Structured Data :- Any data that can be stored, accessed and processed in the form of fixed format known as structured data.

e.g: Data stored in RDBMS. (Employee table in database)

(2) Unstructured Data :- Any data with unknown form or the structure is known as Unstructured data. Its size is huge.

e.g:- Output given by Google Search (It contains, text, images video etc.)
80% is unstructured data.

(3) Semi-Structured data :- It contain both forms of data.

e.g: XML file.

Data Sizes

Term	Size	Example
Kilobyte (KB)	1,000 Bytes	A paragraph of a text document
Megabyte (MB)	1,000 Kilobytes	A small novel
Gigabyte (GB)	1,000 Megabytes	Beethoven's 5th Symphony (classical music video)
Terabyte (TB)	1,000 Gigabytes	All the X-rays in a large hospital
Petabyte (PB)	1,000 Terabytes	Half of the content of all US academic research libraries
Exabyte (EB)	1,000 Petabytes	About one-fifth of the words people has over spoken
Zettabyte (ZB)	1,000 Exabytes	As much information as there are grains of sand on all the world's beaches
Yottabyte (YB)	1,000 Zettabytes	As much information as there are atoms in 7,000 human bodies

History of Big Data

Lots of data got created due to

- Proliferation of Internet
- Social media
- eCommerce

Data Keep on Growing

- Currently (approx.) 328.77 EB of data created a day
- Around 120 ZB of data will be generated in 2024
- 181 ZB of data will be generated in 2025

- Google processes 2.5 EB a day (02/2023)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 300 PB + 4 PB a day (09/2023)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN's Large Hydron Collider (LHC) generates 1 PB of collision data/second – too large to process. Keeping only most 'interesting' ones, CERN Data Centre processes 1 PB/day.

- Videos account for over half of internet data traffic.

What happens in a internet minute

- **$400M sales on Alibaba**
- **439,000** page views on Wikipedia
- **194, 000** apps downloaded
- **31,700** hours of music played on Pandora
- **38,000** photographs uploaded to Instagram
- **4.1** Million searches on Google
- **139,000** hours of video watched on YouTube
- **10** million ads displayed
- **3.3** million shares on Facebook

1 Internet minute

1 Internet minute

- **$400M sales on Alibaba**
- **439,000** page views on Wikipedia
- **194, 000** apps downloaded
- **31,700** hours of music played on Pandora
- **38,000** photographs uploaded to Instagram
- **4.1** Million searches on Google
- **139,000** hours of video watched on YouTube
- **10** million ads displayed
- **3.3** million shares on Facebook

Each of these activities generates

Data

1 Internet minute – for Alibaba

1 Internet minute

- **$400M sales on Alibaba**
- **439,000** page views on Wikipedia
- **194, 000** apps downloaded
- **31,700** hours of music played on Pandora
- **38,000** photographs uploaded to Instagram
- **4.1** Million searches on Google
- **139,000** hours of video watched on Youtube
- **10** million ads displayed
- **3.3** million shares on Facebook

DATA

$400M sales on Alibaba

Product views

Orders

Ratings

Reviews

1 Internet minute - Google

1 Internet minute

- **$400M** sales on Alibaba
- **439,000** page views on Wikipedia
- **194, 000** apps downloaded
- **31,700** hours of music played on Pandora
- **38,000** photographs uploaded to Instagram
- **4.1 Million searches on Google**
- **139,000** hours of video watched on Youtube
- **10** million ads displayed
- **3.3** million shares on Facebook

DATA

4.1 Million searches on Google

Results returned

Results viewed

Results clicked

Data generated in Internet Minute

- Alibaba
- Wikipedia
- Pandora
- Instagram
- Google
- Youtube
- Facebook

And other companies are collecting

Peta Bytes of data every minute

Data generated in a Internet Minute

- This is a 1 TB hard disk drive

Data generated in a Internet Minute

- 1000s of such 1 TB drives are filled up every minute by data collected on the web!!

Collecting truckloads of data - why

Why are web companies collecting truckloads (literally) of data?

Collecting truckloads of data - why

Reason # 1

Because they can afford it

Storage prices have dropped like crazy over the last 2 decades

Collecting truckloads of data - why

Reason # 2

Because they can monetize it

Large scale data can be processed to derive huge amounts of value

Collecting truckloads of data - why

Reason # 2

Because they can monetize it

Large scale data can be processed to derive huge amounts of value

Everything is personalized

Product Recommendations on Amazon,

Newsfeed on Facebook,

Homepage on Netflix

Ads, Offers, Promotions just for you!

How do we go from

Truckloads of data

Monetizable products

Recommendations
Newsfeed
Maps

Giants ....

Companies like

Google,
Apple,
Amazon,
Facebook etc

own Huge Data Centers

Huge Data Centers

covering 100s of acres

with millions of servers

Huge Data Centers

Huge Data Centers
with millions of servers

running
sophisticated
proprietary software

Huge Data Centers

Huge Data Centers
with millions of servers

running
sophisticated
proprietary software

to process
TBs/PBs of data

The Big Data Paradigm

Huge Data Centers

with millions of servers

running sophisticated proprietary software

to process TBs/PBs of data

The Big Data Paradigm

There are only a handful of companies in the world that have all of the above

The Big Data Paradigm

So, should the rest of us even care ?

The Big Data Paradigm

Because of cloud companies like AWS, Microsoft Azure, GCP

Anyone can requisition 100s of servers at a moment's notice

The Big Data Paradigm

Netflix, Pinterest, AirBnB
run their entire business just using cloud services like AWS

The Big Data Paradigm

Open Source Technologies
Hadoop, Spark, HBASE, Hive and many others

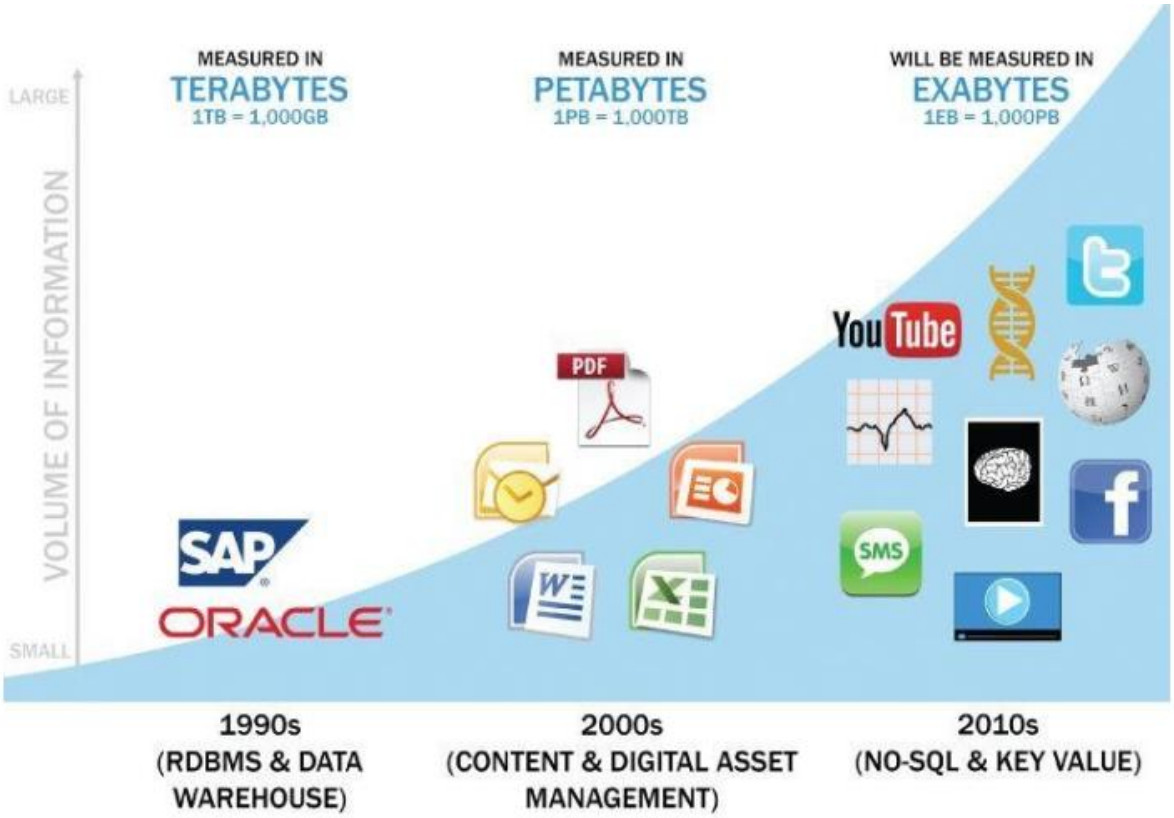


Figure 1.10 Data evolution and rise of Big Data sources

The Big Data trend is generating an enormous amount of information from many new sources. This data deluge requires advanced analytics and new market players to take advantage of these opportunities and new market dynamics, which will be discussed in the following section.

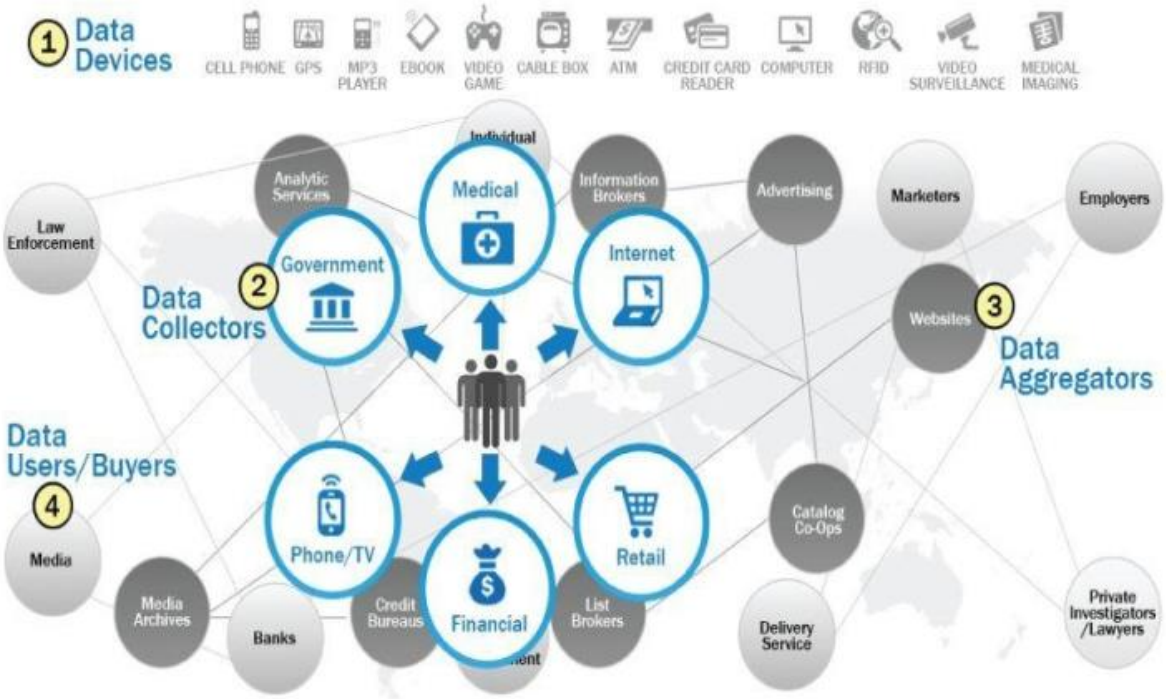


Figure 1.11 Emerging Big Data ecosystems

As illustrated by this emerging Big Data ecosystem, the kinds of data and the related market dynamics vary greatly. These datasets can include sensor data, text, structured datasets, and social media. With this in mind, it is worth recalling that these datasets will not work well within traditional EDWs, which were architected to streamline reporting and dashboards and be centrally managed. Instead, Big Data problems and projects require different approaches to succeed.

Big Data Statistics

- Volume of data
  - 175+ new websites are created every minute of the day. That is over 252 thousand a day.
  - 2.45 billion pieces of content is shared on Facebook each day.
  - X (formerly Twitter) generates 560 GB of data every day
  - YouTube users upload 500 hours of new video content **every minute** of the day. In 2011 it was 48 hrs/min.

- Processing
  - Decoding of the human genome took 13 years from 1990 to 2003 covering 92% of the genome. Now it can be done in 8-9 weeks for full DNA sequencing.
  - Facebook generate 5 Petabytes of data per data. The hive contains 300 petabytes of data.
  - LinkedIn processes and mines Petabytes of user data to power the "People You May Know" feature. This feature processes 100s of terabytes daily.

What is Data?

Representation of facts, concepts, or instructions in a formalized manner.

Characteristics:

- Accuracy
- Completeness
- Reliability
- Relevance
- Timeliness

Big data & Analytics

Wikipedia defines ***"Big Data"*** as a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

What is Big Data?

Big in Big Data refers to:

- Big size is the primary definition.
- Big complexity rather than big volume. It can be small and not all large datasets are big data
- Size matters….but so does accessibility, interoperability and reusability.

Big data is described in V's

- Volume (Internal and External)
  - How much data? TB/PB?
- Velocity: Rate of data creation
  - Growing how fast? GB/s?
  - Sensors; increased transactions and interactions

What is Big Data?

- Variety
  - Different sources, types, formats, schemas
  - Structured and Unstructured data
  - E.g. audio and video files, Facebook & X (Twitter) comments, photos, GPS data, medical files etc.
- Veracity
  - Shows the quality and origin of data
  - Accuracy, trustworthiness
- Value
  - How to turn raw data into useful results
- Variability
  - To what extent, and how fast, is the structure of your data changing?

Characteristics of Big data

Surveillance videos, satellites images, cell phone location, health of power station turbines, furnaces & other industrial machinery, weather and meteorological data, click stream

Archived data:
Patient records, Scanned copies of agreements, records of ex-employees/completed projects, banking transactions older than the compliance regulations

Ex. Trading/stock exchange data, tweets on Twitter, status updates/likes/shares on Facebook

Use of big data

- Recommend cancer medication based on what worked well in similar situations for other patients
- Weather prediction for fishermen, farmers
- Predict equipment malfunctioning in large nuclear power plant, chemical plants, etc.
- Credit card fraud detection

Use of Big Data

Banking and Financial Services
- Fraud Detection to detect the possible fraud or suspicious transactions in Accounts, Credit Cards, Debit Cards, and Insurance etc.

Retail
- Targeting customers with different discounts, coupons, and promotions etc. based on demographic data like gender, age group, location, occupation, dietary habits, buying patterns, and other information which can be useful to differentiate/categorize the customers.

Sentiment Analysis
- Organizations use the data from social media sites like Facebook, Twitter etc. to understand what customers are saying about the company, its products, and services.
- Words like "I like this phone", "This food is too salty", etc.. indicate sentiments
- This type of analysis is also performed to understand which companies, brands, services, or technologies people are talking about.

Use of Big Data …

Customer Service
- IT Services and BPO companies analyze the call records/logs to gain insights into customer complaints and feedback, call center executive response/ability to resolve the ticket, and to improve the overall quality of service.
- Call center data from telecommunications industries can be used to analyze the call records/logs and optimize the price, and calling plan, messaging plan, and data plans

Industrial equipment monitoring & alerting
- A large power plant or chemical factory has thousands of critical equipment that needs to be monitored
- The equipment data needs to be analysed to detect any malfunctioning or danger of accidents

Weather forecasting
- Satellite data from remote sensing satellites need to be analysed at high speed to warn fishermen, farmers and public about potential cyclones, delayed monsoon, etc.

How to Deal with Big Data?

- Analysing Big data requires scale-out solutions not scale-up solutions

- **Move** the analysis to the data.
- Work with scientists to find the most common “20 queries” and make them fast.

Traditional Technology

- Large relational databases
  - On SAN (Storage Area Network)
- Highly parallel processors
- Data may be distributed but processing in one place
- Bring data to process
- Limit on scalability
- High-end hardware ($50,000/TB)

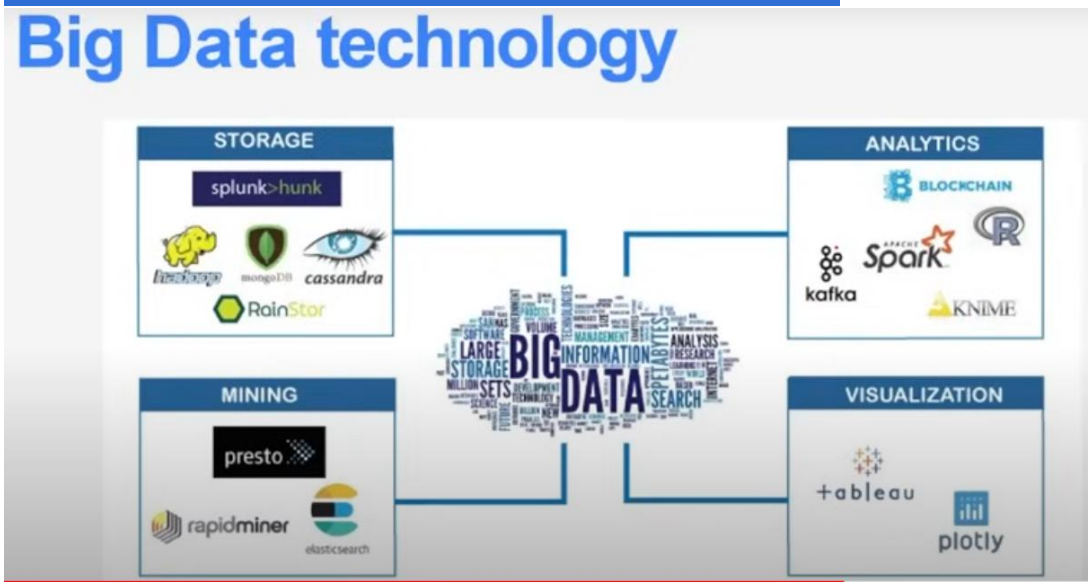
Big Data Technology

- Parallel processing
- Clusters of commodity hardware
- Fault-tolerant processing
- Distributed data and distributed processing
- Data redundancy
- Data locality: Bring process to data
- Commodity hardware ($3,000/TB)
- Data and processing on same machine

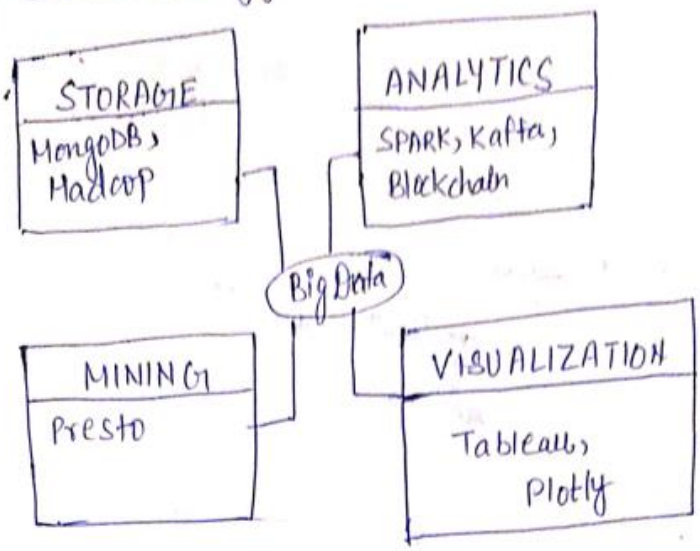
How is the Big Data Technology Different?

- Clusters of commodity class machines
- Distributed data
- Distributed processing

Big Data technology



* Big data technology:-



* Big Data Importance (Why use Big Data?)

1. Cost savings
2. Time-Saving
3. Understand the market conditions
4. Social Media Listening
5. Boost Customer Acquisition & Retention

Big Data Applications

1. Banking and Securities
2. Communications, Media & Entertainment
3. Healthcare Providers
4. Education
5. Government
6. Insurance etc.

Big Data importance

Big Data importance

1. Cost Savings
2. Time-Saving
3. Understand the market conditions
4. Social Media Listening
5. Boost Customer Acquisition and Retention
6. Solve Advertisers Problem
7. The driver of Innovations and Product Development

* Big Data Analytics :-

It is a process of examining large datasets containing a variety of datatypes to uncover hidden patterns, unknown correlation making trends, customer prefences & other useful information.

Challenges :-

1. Insufficient understanding and acceptance of big data.
2. Confusion while big data tools selection.
3. Paying loads of money
4. Data integration
5. Data security
6. Data Analysis

Need of Big Data Analytics :-

1. Optimize business operations by analyzing customer behaviour.

   e.g: amazon

2. Next Generation Products.

   e.g: Netflix , Spotify

What is Big Data Analytics?

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Need for Big Data Analytics

Need for Big Data Analytics

1. Optimize business operations by analyzing customer behaviour

amazon

Need for Big Data Analytics

2. Next Generation Products

NETFLIX

Spotify

Types of Big Data Analytics

Types of Big Data Analytics

1. Descriptive Analysis
2. Predictive Analysis
3. Prescriptive Analysis
4. Diagnostic Analysis

Types of Big Data Analytics

1. Descriptive Analysis

What is happening now based on incoming data.

Types of Big Data Analytics

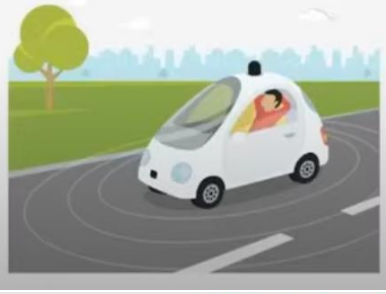
2. Predictive Analysis

What might happen in future.

Types of Big Data Analytics

3. Prescriptive Analysis

What action should be taken.



Google's self-driving car is perfect example of Presciptive Analysis.

Types of Big Data Analytics

4. Diagnostic Analysis

What did it happen

Types of Big Data Analytics :—

1. Descriptive Analysis (WHAT Happened?)
2. Predictive Analysis (What will happen?)
3. Prescriptive Analysis (Recommended / How can we make it happen)?
4. Diagnostic Analysis (WHY DOI it happen?)

(1) **Descriptive Analysis** :- It provides the insights into what has occured in the past and with the trends to dig into more detail. This help in creating reports.

e.g: Performance of 'XYZ' company in year 2022.

(2) **Predictive Analysis** :-

" what will happen in the future?"

=> This type of analytics ensures that the path for the future course of action is predicted. This makes use of historical and present data to predict future events.

e.g: Performance of 'XYZ' company in year 2024.

(3) **Prescriptive Analysis** :- It is the next step in predictive analysis. It explores several possible actions & suggests

actions depending on the results of descriptive & predictive analytics of a given dataset.

"What actions to be taken to achieve predicted results?".

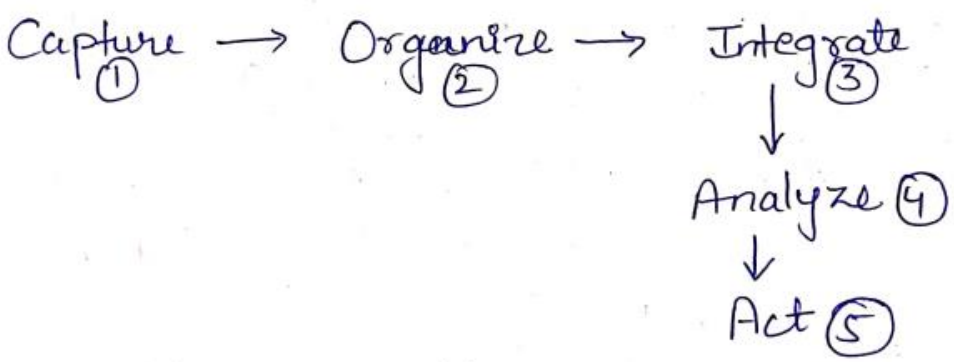
e.g: Self Driving car (Predictive + Prescriptive)

(4) Diagnostic Analytics:- It gives a detailed and in-depth insight into the root cause of a problem.

"Why did it happen?"

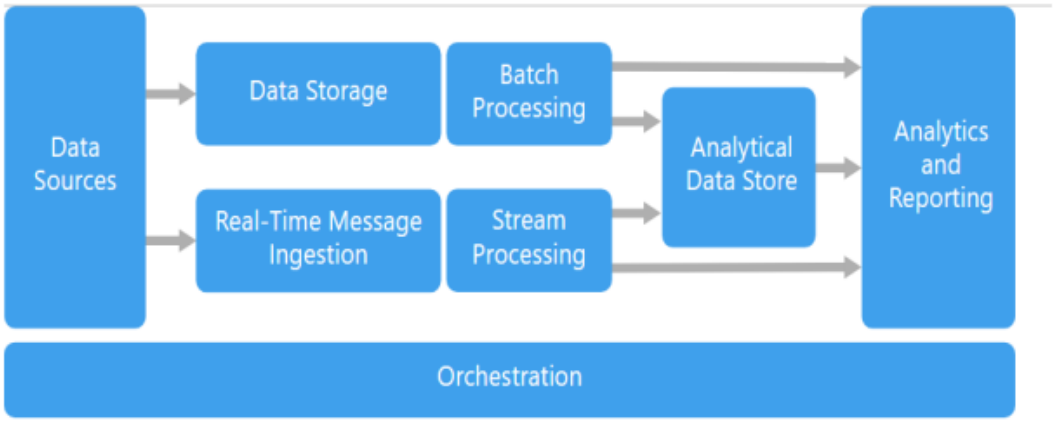
→ It uses techniques such as data discovery, data mining & correlation.

* Big Data Management Cycle:-



Big data architecture

A big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems.



Most big data architectures include some or all of the following components:

1. **Data sources**: All big data solutions start with one or more data sources. Examples include:
    - Application data stores, such as relational databases.
    - Static files produced by applications, such as web server log files.
    - Real-time data sources, such as IoT devices.

2. **Data storage**: Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats. This kind of store is often called a *data lake*. Options for implementing this storage include Azure Data Lake Store or blob containers in Azure Storage.

3. **Batch processing**: Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Usually, these jobs involve reading source files, processing them, and writing the output to new files. Options include running U-SQL jobs in Azure Data Lake Analytics, using Hive, Pig, or custom Map/Reduce jobs in an HDInsight Hadoop cluster, or using Java, Scala, or Python programs in an HDInsight Spark cluster.

4. **Real-time message ingestion**: If the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing. This might be a simple data store, where incoming messages are dropped into a folder for processing. However, many solutions need a message ingestion store to act as a buffer for messages, and to support scale-out processing, reliable delivery, and other message queuing semantics. Options include Azure Event Hubs, Azure IoT Hubs, and Kafka.

5. **Stream processing**: After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis. The processed stream data is then written to an output sink. Azure Stream Analytics provides a managed stream processing service based on perpetually running SQL queries that operate on unbounded streams. You can also use open-source Apache streaming technologies like Storm and Spark Streaming in an HDInsight cluster.

6. **Analytical data store**: Many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools. The analytical data store used to serve these queries can be a Kimball-style relational data warehouse, as seen in most traditional business intelligence (BI) solutions. Alternatively, the data could be presented through a low-latency NoSQL technology such as HBase, or an interactive Hive database that provides a metadata abstraction over data files in the distributed data store. Azure Synapse Analytics provides a managed service for large-scale, cloud-based data warehousing. HDInsight supports Interactive Hive, HBase, and Spark SQL, which can also be used to serve data for analysis.

7. **Analysis and reporting**: The goal of most big data solutions is to provide insights into the data through analysis and reporting. To empower users to analyze the data, the architecture may include a data modeling layer, such as a multidimensional OLAP cube or tabular data model in Azure Analysis Services. It might also support self-service BI, using the modeling and visualization technologies in Microsoft Power BI or Microsoft Excel. Analysis and reporting can also take the form of interactive data exploration by data scientists or data analysts. For these scenarios, many Azure services support analytical notebooks, such as Jupyter, enabling these users to leverage their existing skills with Python or R. For large-scale data exploration, you can use Microsoft R Server, either standalone or with Spark.

8. **Orchestration**: Most big data solutions consist of repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report or dashboard. To automate these workflows, you can use an orchestration technology such Azure Data Factory or Apache Oozie and Sqoop.