Big Data

Unit-I:

Overview of Big Data Analytics: Introduction to Data Analytics and Big Data, Evolution of big data, Challenges with Traditional Large Scale Systems, characteristics (V‟s , types (structured, semi-structured and unstructured) and sources of Big Data, Distributed, Parallel Computing and Cloud Computing for big data. Analytics Toolkit: Components of the analytics toolkit, Analytical Sandbox: Internal, External and Hybrid.

What's Big Data?

- **Big data** is the term for 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.

- The challenges include **capture, curation, storage, search, sharing, transfer, analysis, and visualization.**
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "**spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions**."

Other definitions

- Big Data is also **data** but with a **huge size**.
- Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.
- In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

from Wikipedia:

Big data is the term for – 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.

Facts and Figures

- **Walmart** handles 1 million customer transactions/hour.
- **Facebook** handles 40 billion photos from its user base!
- **Facebook** inserts 500 terabytes of new data every day.
- **Facebook** stores, accesses, and analyzes 30+ Petabytes of user generated data.
- **A flight generates** 240 terabytes of flight data in 6-8 hours of flight.
- **More than 5 billion people** are calling, texting, tweeting and browsing on mobile phones worldwide.
- **Decoding the human genome** originally took 10 years to process; now it can be achieved in one week.[8]
- **The largest AT&T database** boasts titles including the largest volume of data in one unique database (312 terabytes) and the second largest number of rows in a unique database (1.9 trillion), which comprises AT&T's extensive calling records.

An Insight

Byte: One grain of rice

KB(3): One cup of rice:

MB (6): 8 bags of rice:

GB (9): 3 Semi trucks of rice:

TB (12): 2 container ships of rice

PB (15): Blankets ½ of Jaipur

Exabyte (18): Blankets West coast Or 1/4th of India

Zettabyte (21): Fills Pacific Ocean

Yottabyte(24): An earth-sized rice bowl

Brontobyte (27): Astronomical size

Desktop

Internet

Big Data

Future

What's making so much data?

- Sources: People, machine, organization: Ubiquitous computing
- More people carrying data-generating devices (Mobile phones with facebook, GPS, Cameras, etc.)

- **Data on the Internet:**
- Internet live stats

What is big data analytics?

- **Big data analytics is the process of finding patterns, trends, and relationships in massive datasets. These complex analytics require specific tools and technologies, computational power, and data storage that support the scale.**

How does big data analytics work?

Big data analytics follows five steps to analyze any large datasets:

1.Data collection
2.Data storage.
3.Data processing
4.Data cleansing
5.Data analysis

Benefits of Big Data Analytics:

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data—which means a completely different approach to tackling problems.

Evolution of big data

- The History of Big Data Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.

- Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyze big data sets) was developed that same year. NoSQL also began to gain popularity during this time.

- The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store.
- In the years since then, the volume of big data has skyrocketed.

- With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.

- While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.

Where is the problem?

- Traditional RDBMS queries isn't sufficient to get useful information out of the huge volume of data
- To search it with traditional tools to find out if a particular topic was trending would take so long that the result would be meaningless by the time it was computed.
- Big Data come up with a solution to store this data in novel ways in order to make it more accessible, and also to come up with methods of performing analysis on it.

Challenges

- Capturing
- Storing
- Searching
- Sharing
- Analysing
- Visualization

Challenges with Traditional Large Scale Systems

The following list of challenges has been dominating in the case Traditional systems in real time scenarios:

- **Uncertainty of Data Management Landscape**
- **The Big Data Talent Gap**
- **The talent gap that exists in the industry Getting data into the big data platform**
- **Need for synchronization across data sources**
- **Getting important insights through the use of Big data analytics**

Uncertainty of Data Management Landscape:

- Because big data is continuously expanding, there are new companies and technologies that are being developed everyday.
- A big challenge for companies is to find out which technology works bests for them without the introduction of new risks and problems.

The Big Data Talent Gap:

- While Big Data is a growing field, there are very few experts available in this field.
- This is because Big data is a complex field and people who understand the complexity and intricate nature of this field are far few and between.

The talent gap that exists in the industry Getting data into the big data platform:

- Data is increasing every single day. This means that companies have to tackle limitless amount of data on a regular basis.
- The scale and variety of data that is available today can overwhelm any data practitioner and that is why it is important to make data accessibility simple and convenient for brand mangers and owners.

Need for synchronization across data sources:

- As data sets become more diverse, there is a need to incorporate them into an analytical platform.
- If this is ignored, it can create gaps and lead to wrong insights and messages.

Getting important insights through the use of Big data analytics:

- It is important that companies gain proper insights from big data analytics and it is important that the correct department has access to this information.
- A major challenge in the big data analytics is bridging this gap in an effective fashion.

Characteristics of Big Data

Back in 2001, Gartner analyst Doug Laney listed the 3 ‘V’s of Big Data – Variety, Velocity, and Volume. Let’s discuss the characteristics of big data. These characteristics, isolated, are enough to know what big data is

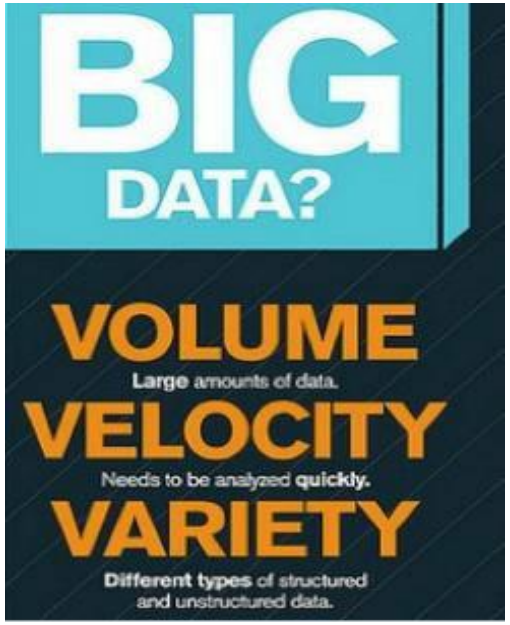
Variety: Variety of Big Data refers to structured, unstructured, and semi-structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

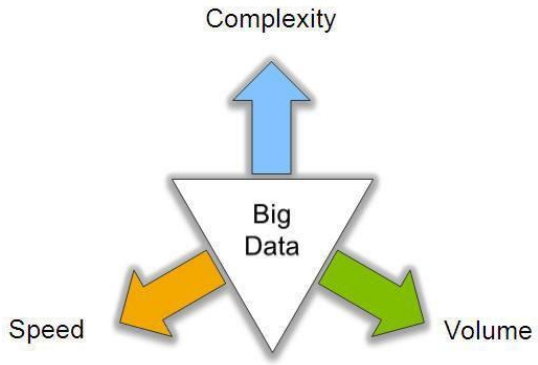
Velocity: Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

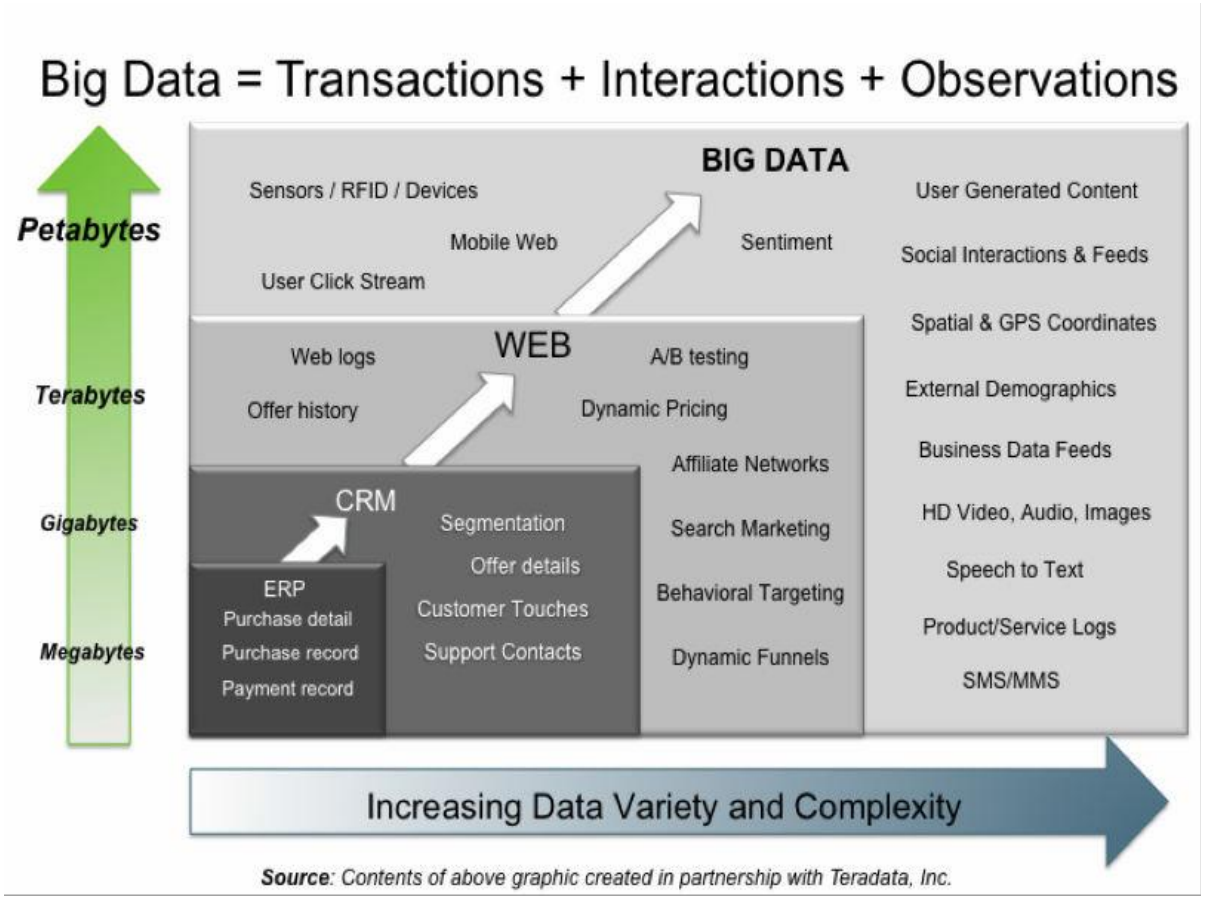
Volume: Volume is one of the characteristics of big data. We already know that Big Data indicates huge ‘volumes’ of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data is stored in data warehouses.

IBM considers Big Data (3V's): Characteristics.

- The 3V's: Volume, Velocity and Variety.



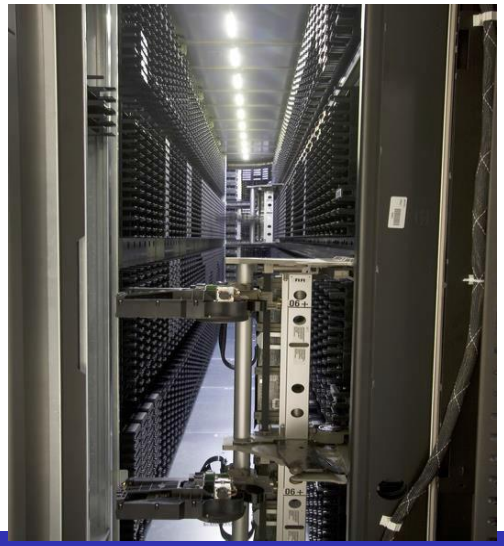




Volume (Scale)

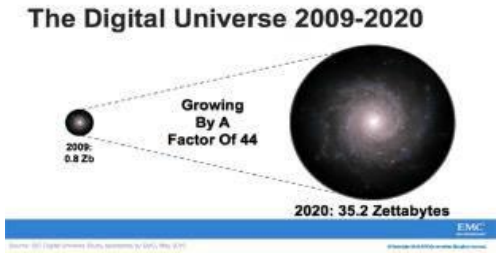
Volume: Enterprises are awash with ever-growing data of all types, easily amassing terabytes even Petabytes of information.

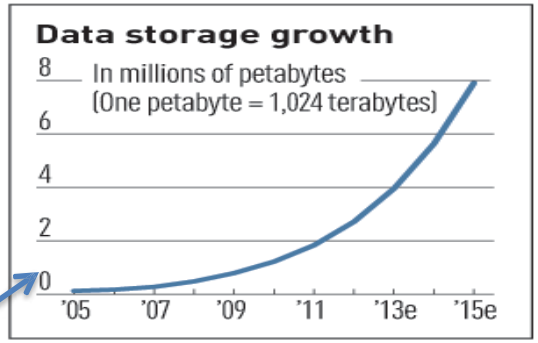
- Turn 12 terabytes of Tweets created each day into improved product sentiment analysis
- Convert 350 billion annual meter readings to better predict power consumption

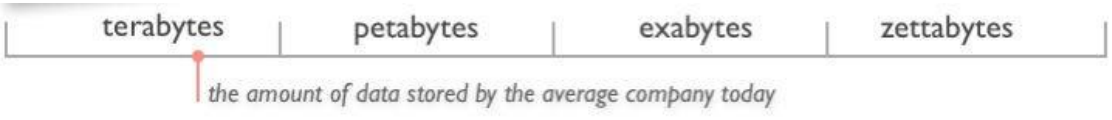


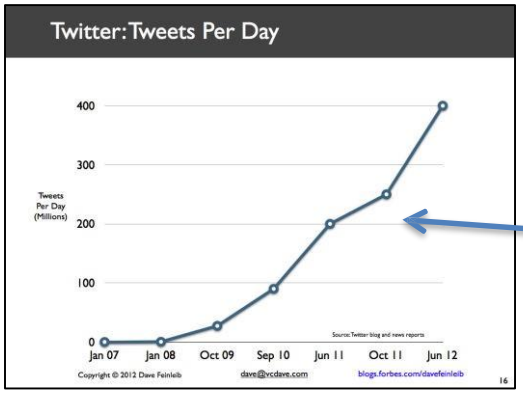
Volume (Scale)

- **Data Volume**
  44x increase from 2009 2020 From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially









Exponential increase in collected/generated data

Example 1: CERN's Large Hydron Collider(LHC)

CERN's Large Hydron Collider (LHC) generates 15 PB a year

Velocity (Speed)

- **Velocity:** Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.
  - Scrutinize 5 million trade events created each day to identify potential fraud
  - Analyze 500 million daily call detail records in real-time to predict customer churn faster

Examples: Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**

E-Promotions: Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you

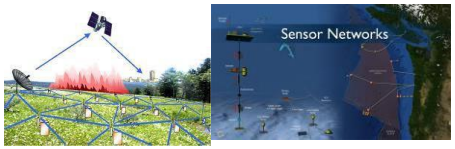
Healthcare monitoring: sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

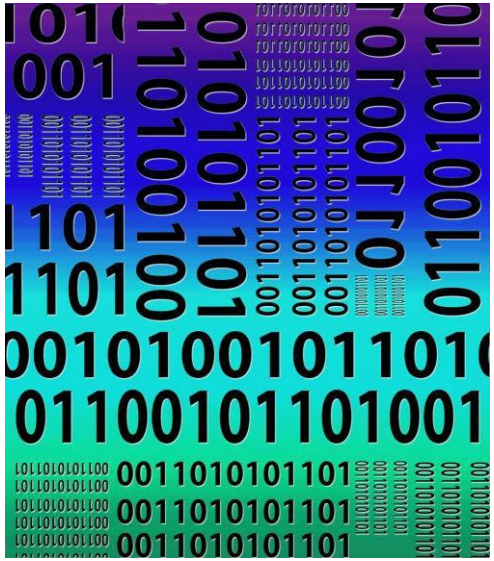


Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data But, by the ability to manage, analyze, summarize, visualize, and
- discover knowledge from the collected data in a timely manner and in a scalable fashion

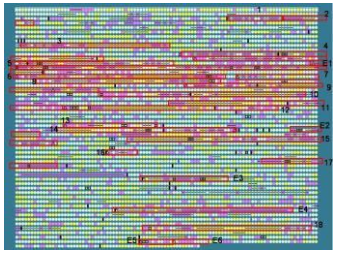
Variety (Complexity)

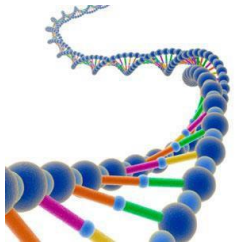
- **Variety:** Big data is any type of data –
  - Structured Data (example: tabular data)
  - Unstructured –text, sensor data, audio, video
  - Semi Structured : web data, log files

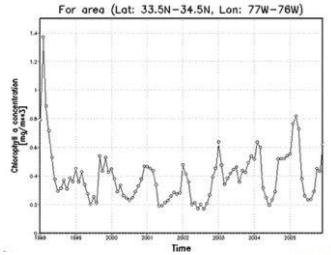


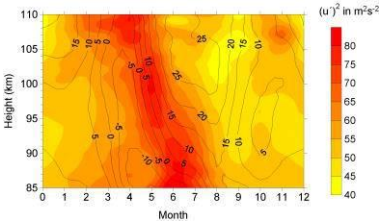
Examples: Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
  Text Data (Web)
- Semi-structured Data (XML) Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
  You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)





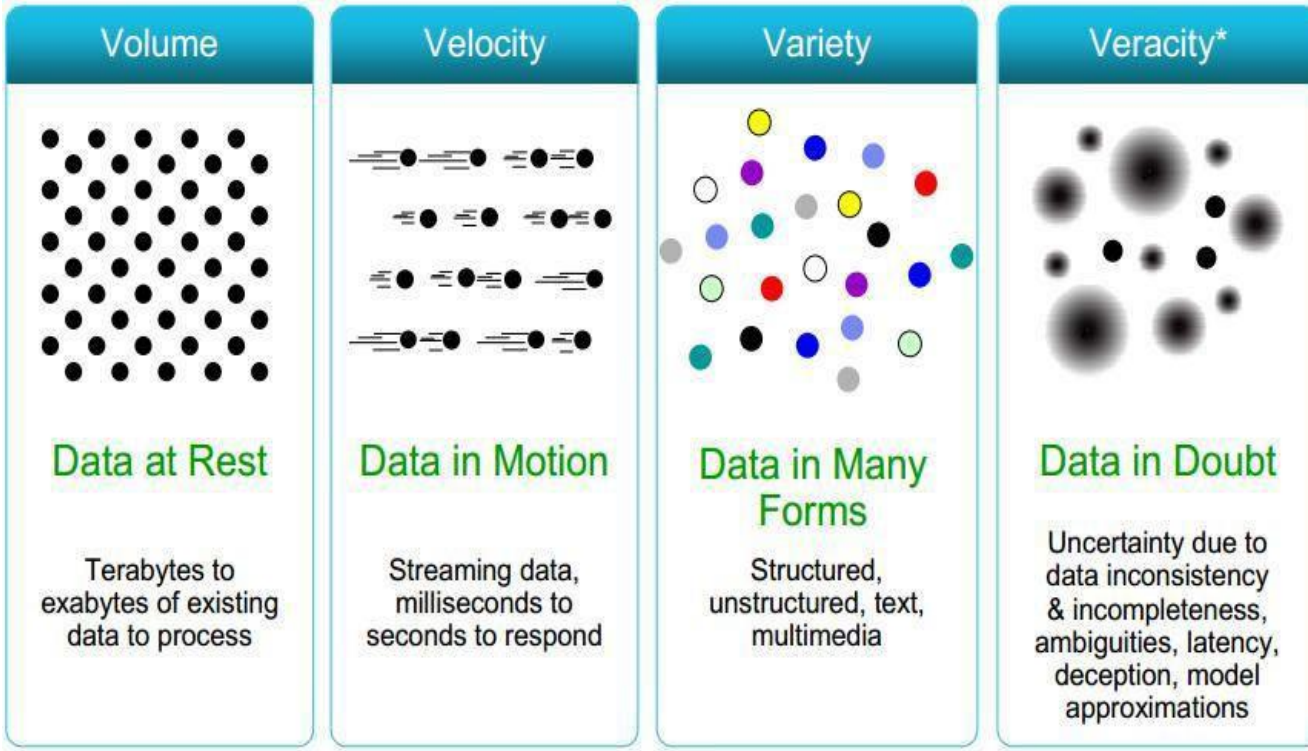




To extract knowledge → all these types of data need to linked together

The 3 Big V's (+1)

- **Big 3V's**
  - Volume
  - Velocity
  - Variety
- **Plus 1**
  - Value



The 3 Big V's (+1) (+ N more)

- **Plus many more**

  Veracity Validity

  Variability

  Viscosity & Volatility

  Viability,

  Venue,

  Vocabulary, Vagueness,

  …

Value

- Integrating Data
  - Reducing data complexity Increase
  - Unify your data systems
  - All 3 above will lead to increased data collaboration
  - -> add value to your big data

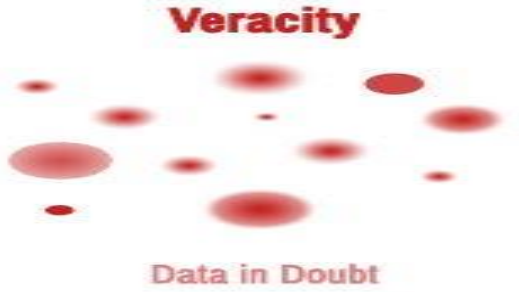
Veracity

- **Veracity** refers to the biases ,noise and abnormality in data, trustworthiness of
- data. 1 in 3 business leaders don't trust the information

  they use to make decisions.

  How can you act upon information if you don't trust it?

  Establishing trust in big data presents a huge challenge as the variety and numbe grows.



Valence

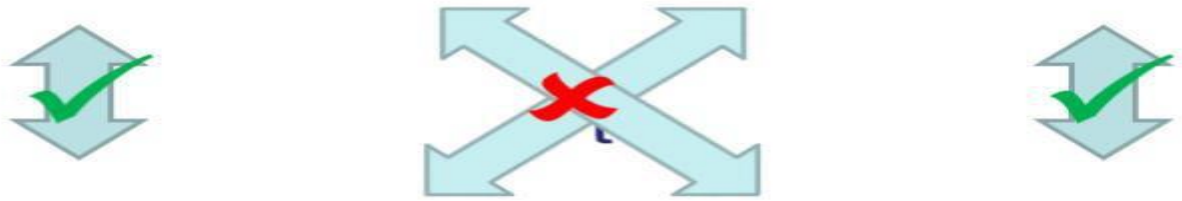
- **Valence** refers to the connectedness of big data.
- Such as in the form of graph networks

Validity

Accuracy and correctness of the data relative to a particular use

- Example: **Gauging storm intensity**

  satellite imagery vs social media posts

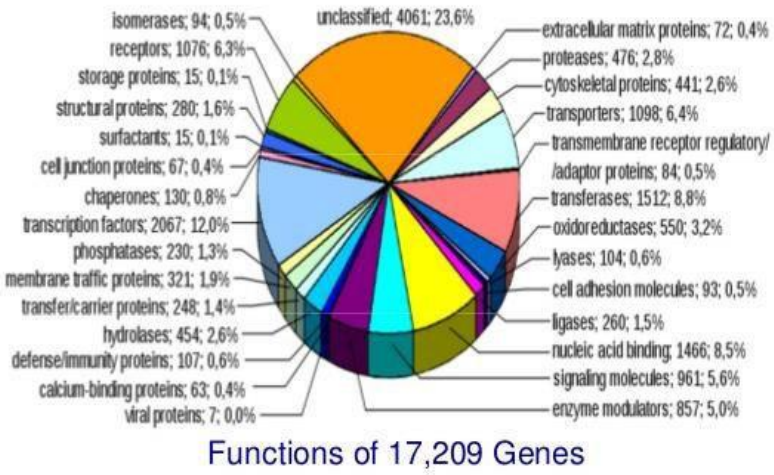


- prediction quality vs human impact

Variability

How the meaning of the data changes over time

-
- Language evolution Data availability Sampling
- processes
- Changes in characteristics of the data source



Functions of 17,209 Genes

Viscosity & Volatility

- Both related to velocity
- **Viscosity:** *data velocity relative to timescale of event being studied*
- **Volatility:** *rate of data loss and stable lifetime of data*

  Scientific data often has practically unlimited lifespan, but social / business data may evaporate in finite time

More V's

- **Viability**
  - Which data has meaningful relations to questions of interest?
- **Venue**
  - Where does the data live and how do you get it?
- **Vocabulary**
  - Metadata describing structure, content, & provenance
  - Schemas, semantics, ontologies, taxonomies, vocabularies
- **Vagueness**
  - Confusion about what "Big Data" means

Dealing with Volume

- Distill big data down to small information Parallel and
- automated analysis Automation requires
- standardization Standardize by reducing Variety:
- Format Standards Structure

Types Of Big Data https://www.geeksforgeeks.org/types-of-big-data/

1.Big Data could be found in three forms:

Structured

Unstructured

Semi-structured

What is Structured Data?

•Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

•Developed techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.

•Foreseeing issues of today :

– when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zetta bytes.

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured Data

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge,

- un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
- A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.
- Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.
- Example of Unstructured data - The output returned by 'Google Search'

Semi-structured Data

Semi-structured data can contain both the forms of data. Semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.

Example of semi-structured data is – a data represented in an XML file.

Personal data stored in an XML file.

```
<rec>
<name>Prashant Rao</name>
<sex>Male</sex>
<age>35</age>
```

Sources of Big Data:

These data come from many sources like

- **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

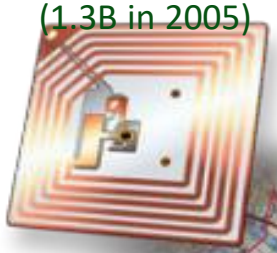
Source of Data Generation

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs **of log data every day**

30 billion RFID tags today (1.3B in 2005)



4.6 billion camera phones world wide

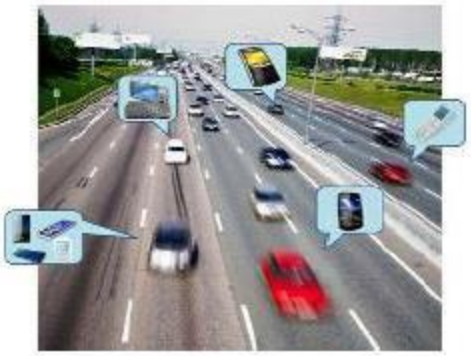
100s of millions of GPS enabled devices sold annually

76 million smart meters in 2009… 200M by 2014

2+ billion people on the Web by end 2011

An Example of Big Data at Work

Crowdsourcing



Computing

Sensing

Real time traffic info

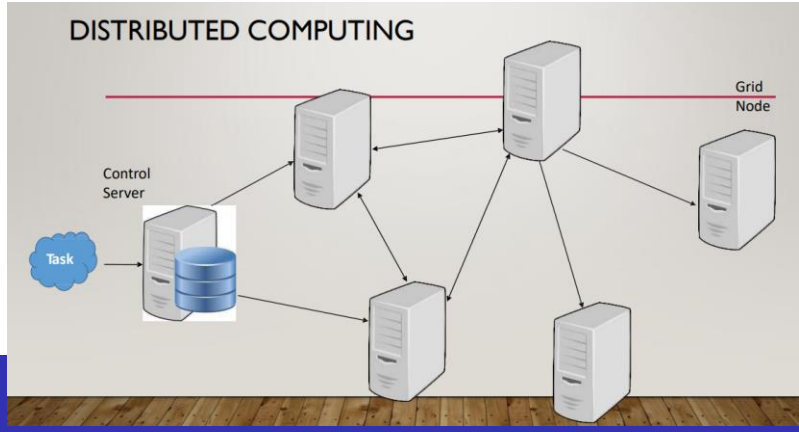
Distributed, Parallel Computing and Cloud Computing for big data.

Big Data can’t be handled by traditional data storage and processing systems. For handling such type of data, Distributed and Parallel Technologies are more suitable.

Distributed Computing

Multiple computing resources are connected in a network and computing tasks are distributed across these resources.

- Increases the Speed for storing data.
- Increases the Efficiency more suitable to process huge amount of data in a limited time.



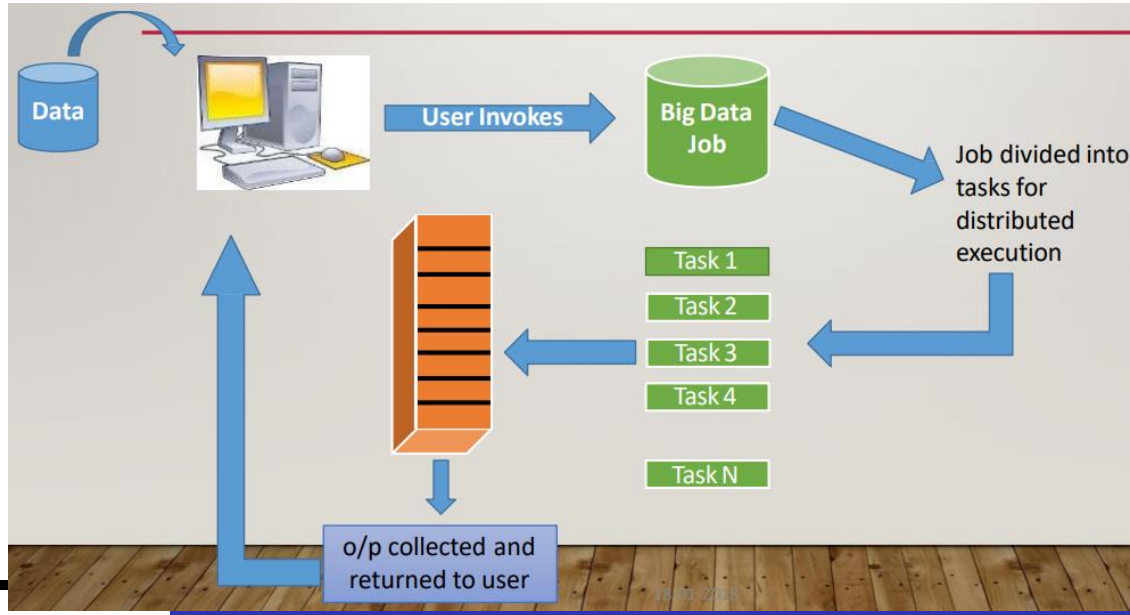
MERITS OF THE SYSTEM

Scalability –

The system with added scalability, can accommodate the growing amounts of data more efficiently and flexibly.

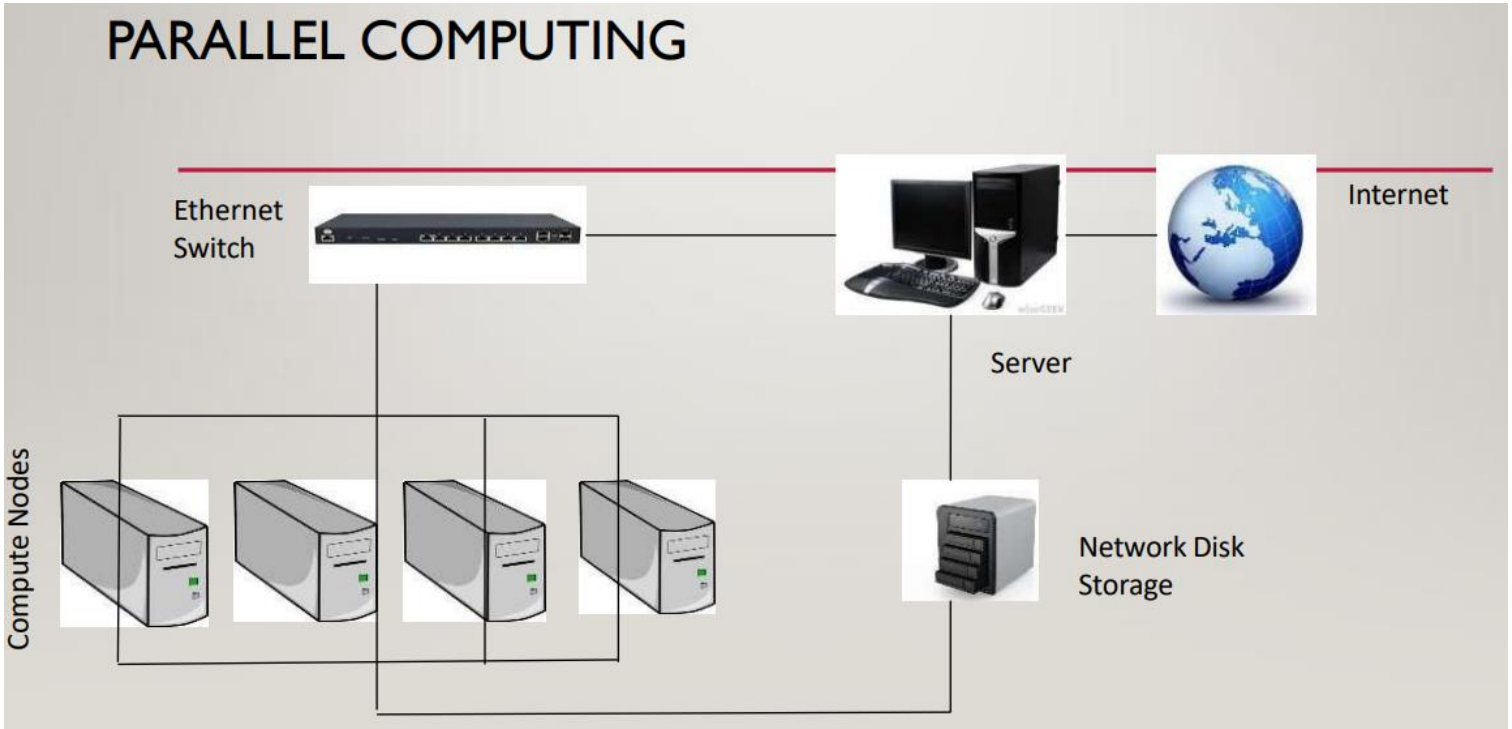
Virtualization and Load Balancing Features –

The sharing of workload across various systems. Virtualization – creates a virtual environment h/w platform, storage device and OS



Parallel Computing

- Divide complex computations into subtasks, handled individually by processing units, running in parallel.
- Improves the processing capability of a computer system by adding additional computational resources to it.



PARALLEL COMPUTING TECHNIQUES

1) Cluster or Grid Computing

Primarily used in Hadoop. Based on a connection of multiple servers in a network (clusters) servers share the workload among them.

2) Massively Parallel Processing (MPP)

Used in data warehouses.
Single machine working as a grid is used in the MPP platform.
Capable of handling the storage, memory and computing activities.
Software written specifically for MPP platform is used for optimization.
MPP platforms, EMC Greenplum, ParAccel , suited for high-value use cases.

3) High Performance Computing (HPC)

Offer high performance and scalability by using IMC.
Suitable for processing floating point data at high speeds.
Used in research and business organization where the result is more valuable than the cost or where strategic importance of project is of high priority.

DIFFERENCE B/W DISTRIBUTED AND PARALLEL SYSTEMS

Distributed System

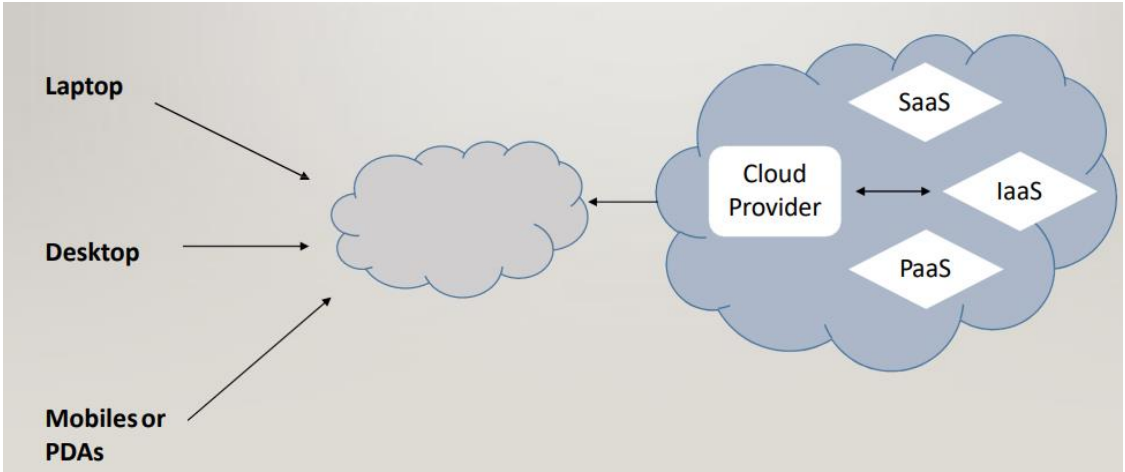
- Independent autonomous system connected in a n/w for accomplishing specific task.
- Coordination is possible b/w connected computers that have their own memory and CPU.
- Loose coupling of computers connected in a n/w, providing access to data and remotely located resources.

Parallel System

- Computer systems with several processing units attached to it.
- Common shared memory can be directly accessed by every processing unit in a n/w.
- Tight coupling of processing resources that are used for solving a single, complex problem.

CLOUD COMPUTING

- Cloud Computing is the delivery of computing services—servers, storage, databases, networking, software, analytics and more—over the Internet (“the cloud”).
- Companies offering these computing services are called cloud providers and typically charge for cloud computing services based on usage, similar to how you are billed for water or electricity at home



FEATURES OF CLOUD COMPUTING

Scalability

- Addition of new resources to an existing infrastructure.
- Increase in the amount of data , requires organization to improve hardware components.
- The new hardware may not provide complete support to the software, that used to run properly on the earlier set of hardware. - Solution to this problem is using cloud services that employ the distributed computing technique to provide scalability.

Elasticity

- Hiring certain resources, as and when required, and paying for those resources.
- No extra payment is required for acquiring specific cloud services.
- A cloud does not require customers to declare their resource requirements in advance.

Resource Pooling

- Multiple organizations, which use similar kinds of resources to carry out computing practices, have no need to individually hire all the resources.

Self Service

- Cloud computing involves a simple user interface that helps customers to directly access the cloud services they want.

Low Cost

- Cloud offers customized solutions, especially to organizations that cannot afford too much initial investment.
- Cloud provides pay-us-you-use option, in which organizations need to sign for those resources only that are essential.

Fault Tolerance

- Offering uninterrupted services to customers

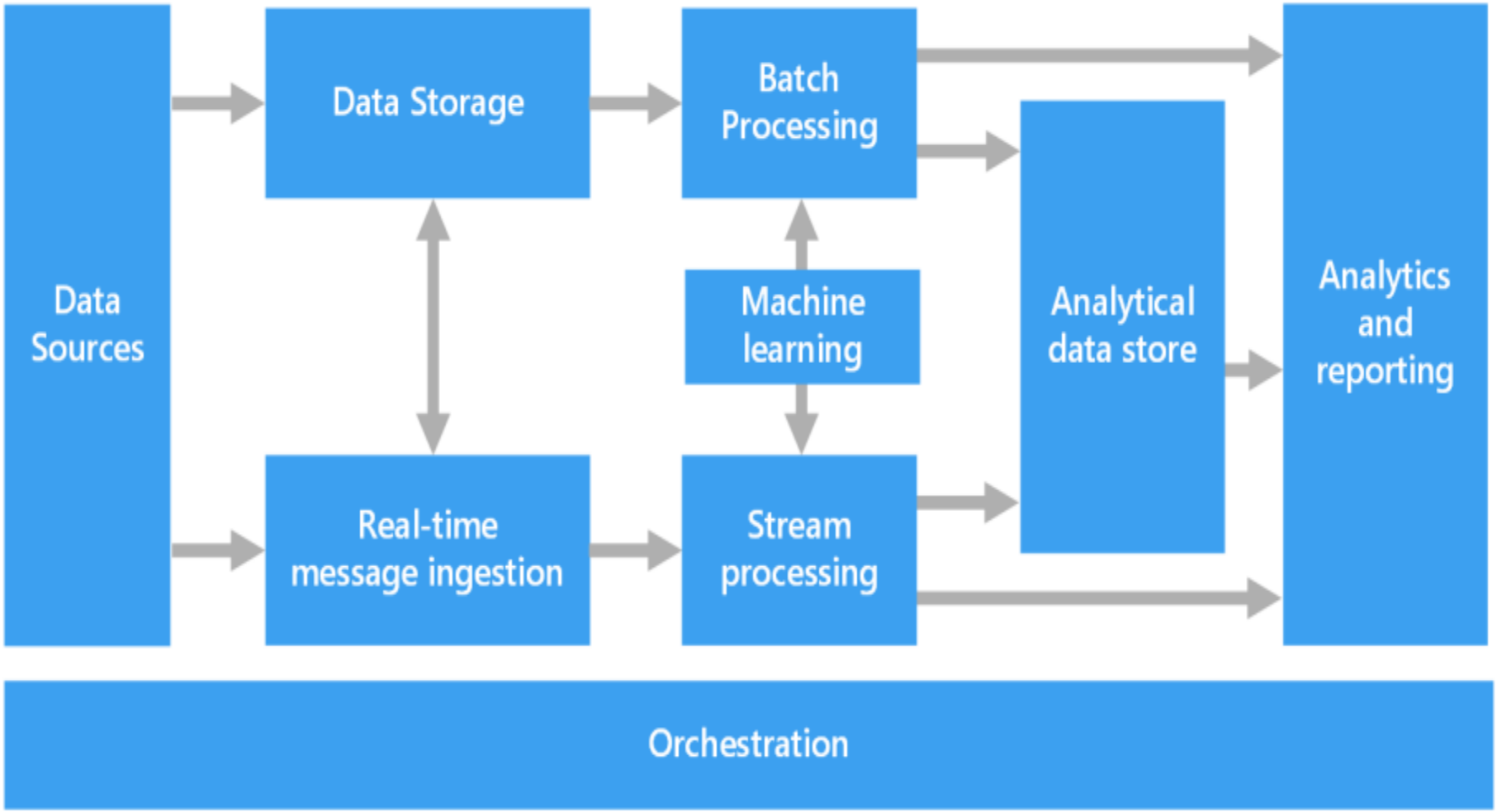
CLOUD SERVICES FOR BIG DATA

In big data Iaas, Paas and Saas clouds are used in following manner.

IAAS:- Huge storage and computational power requirement for big data are fulfilled by limitless storage space and computing ability obtained by iaas cloud.

PAAS:- Offerings of various vendors have stared adding various popular big data platforms that include mapreduce, Hadoop. These offerings save organisations from a lot of hassels which occur in managing individual hardware components and software applications.

SAAS:- Various organisation require identifying and analysing the voice of customers particularly on social media. Social media data and platform are provided by SAAS vendors.In addition, private cloud facilitates access to enterprise data which enable these analyses



Components of a Big Data Architecture

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

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.

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.

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. This portion of a streaming architecture is often referred to as stream buffering. Options include Azure Event Hubs, Azure IoT Hub, and Kafka.

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.

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.

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.

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.

Components of the Analytics Toolkit

Analytics Toolkit consists of Hadoop and Spark, which can be installed both locally on the user's machine as well as on the cloud; and it has R and Python, both of which can be installed on the user's machine as well as on a cloud platform. Your Analytics Toolkit will consist of:

Software/platform	Used for data mining	Used for machine learning
Hadoop	X	
Spark	X	X
Redis	X	
MongoDB	X	
Open Source R	X	X
Python (Anaconda)	X	X
Vowpal Wabbit		X
LIBSVM, LIBLINEAR		X
H2O		X

TOOLS:

Where processing is hosted? – Distributed Servers / Cloud (e.g. Amazon EC2)

Where data is stored? – Distributed Storage (e.g. Amazon S3)

What is the programming model? – Distributed Processing (e.g. MapReduce)

How data is stored & indexed? – High-performance schema-free databases (e.g. MongoDB)

What operations are performed on data? – Analytic / Semantic Processing

Big data tools for HPC and supercomputing – MPI

Big data tools on clouds – MapReduce model

- – Iterative MapReduce model
- – DAG model – Graph model
- – Collective model

Other BDA tools

- – SaS – R – Hadoop

Thus, the BDA tools are used throughout the BDA applications development.

Analytical Tools

Different types of analytical tools available:

- GUI Based: Excel, SPSS, SAS ,Rstudio
- Visualization: Tableau, MicroStrategy
- Coding Based: SAS, R

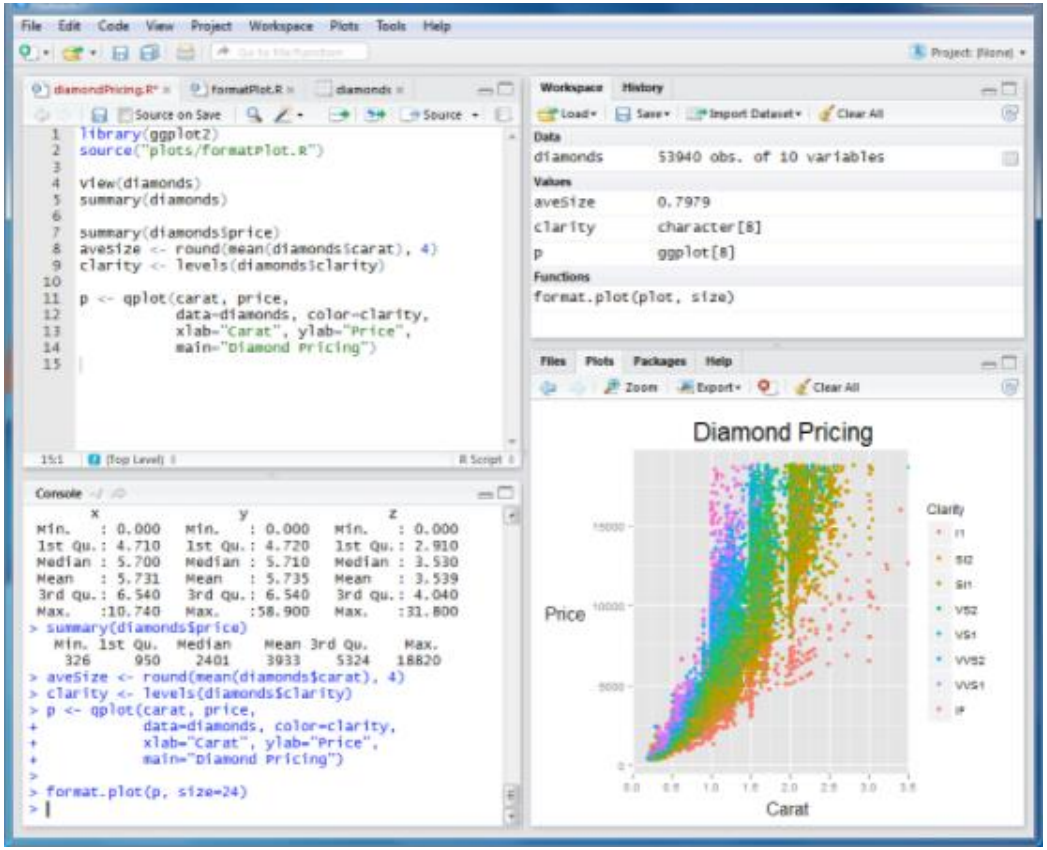


Fig 3: Interface of R studio

Analytic Sandbox

- **An analytic sandbox provides a set of resources with which in-depth analysis can be done to answer critical business questions.**
- **An analytic sandbox is ideal for data exploration, development of analytical processes, proof of concepts, and prototyping.**

- **A sandbox is going to be leveraged by a fairly small set of users. There will be data created within the sandbox that is segregated from the production database.**

- **Data in a sandbox will have a limited shelf life. The idea isn't to build up a bunch of permanent data**

Analytic Sandbox Benefits.

Benefits from the view of an **analytic professional**:

- **Independence**. Analytic professionals will be able to work independently on the database system without needing to continually go back and ask for permissions for specific projects.

- **Flexibility**. Analytic professionals will have the flexibility to use whatever business intelligence, statistical analysis, or visualization tools that they need to use.
- **Efficiency**. Analytic professionals will be able to leverage the existing enterprise data warehouse or data mart, without having to move or migrate data. Analytic professionals can reduce focus on the administration of systems and babysitting of production processes by shifting those maintenance tasks to IT.
- **Speed**. Massive speed improvement will be realized with the move to parallel processing. This also enables rapid iteration and the ability to "fail fast" and take more risks to innovate.

Benefits from the view of IT:

Centralization. IT will be able to centrally manage a sandbox environment just as every other database environment on the system is managed.

Streamlining. A sandbox will greatly simplify the promotion of analytic processes into production since there will be a consistent platform for both development and deployment.

Simplicity. There will be no more processes built during development that have to be totally rewritten to run in the production environment.

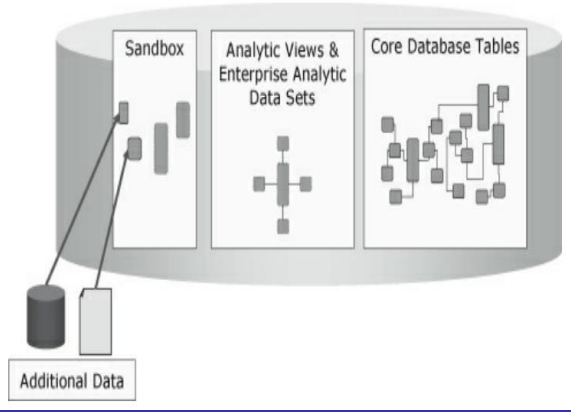
Control. IT will be able to control the sandbox environment, balancing sandbox needs and the needs of other users. The production environment is safe from an experiment gone wrong in the sandbox.

Costs: Big cost savings can be realized by consolidating many analytic data marts into one central system.

An Internal Sandbox

For an internal sandbox, a portion of an enterprise data warehouse or data mart is carved out to serve as the analytic sandbox. In this case, the sandbox is physically located on the production system. However, the sandbox database itself is not a part of the production database. The sandbox is a separate database container within the system

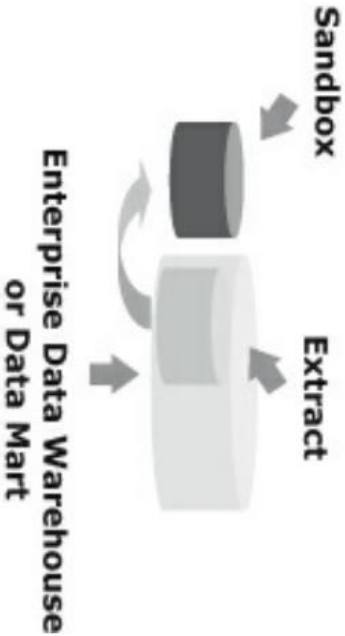
- One strength of an internal sandbox is that it will leverage existing hardware resources and infrastructure already in place.
- The biggest strength of an internal sandbox is the ability to directly join production data with sandbox data.
- An internal sandbox is very cost-effective since no new hardware is needed
- One such weakness is that there will be an additional load on the existing enterprise data warehouse or data mart



External Sandbox

An external sandbox, a physically separate analytic sandbox is created for testing and development of analytic processes. It's relatively rare to have an environment that's purely external. It is important to understand what the external sandbox is, however, as it is a component of a hybrid sandbox environment

- The biggest strength of an external sandbox is its simplicity. The sandbox is a stand-alone environment, dedicated to advanced analytics development.
- Most organizations have a test and/or development environment, independent of their production system, for application and business intelligence work.
- Another strength of an external sandbox is reduced workload management. When only analytic professionals are using the system, it isn't necessary to worry much about tuning and balancing.



- A major weakness of an external sandbox is the additional cost of the stand-alone system that serves as the sandbox platform.
- Another weakness is that there will be some data movement. It will be necessary to move data from the production system into the sandbox before developing a new analysis

Hybrid Sandbox:

A hybrid sandbox environment is the combination of an internal sandbox and an external sandbox. It allows analytic professionals the flexibility to use the power of the production system when needed, but also the flexibility of the external system for deep exploration or tasks that aren't as friendly to the database.

- The strengths of a hybrid sandbox environment are similar to the strengths of the internal and external options, plus having ultimate flexibility in the approach taken for an analysis.
- Another advantage is if an analytic process has been built and it has to be run in a "pseudoproduction" mode temporarily while the full production system process is being deployed.
- The weaknesses of a hybrid environment are similar to the weaknesses of the other two options, but with a few additions.