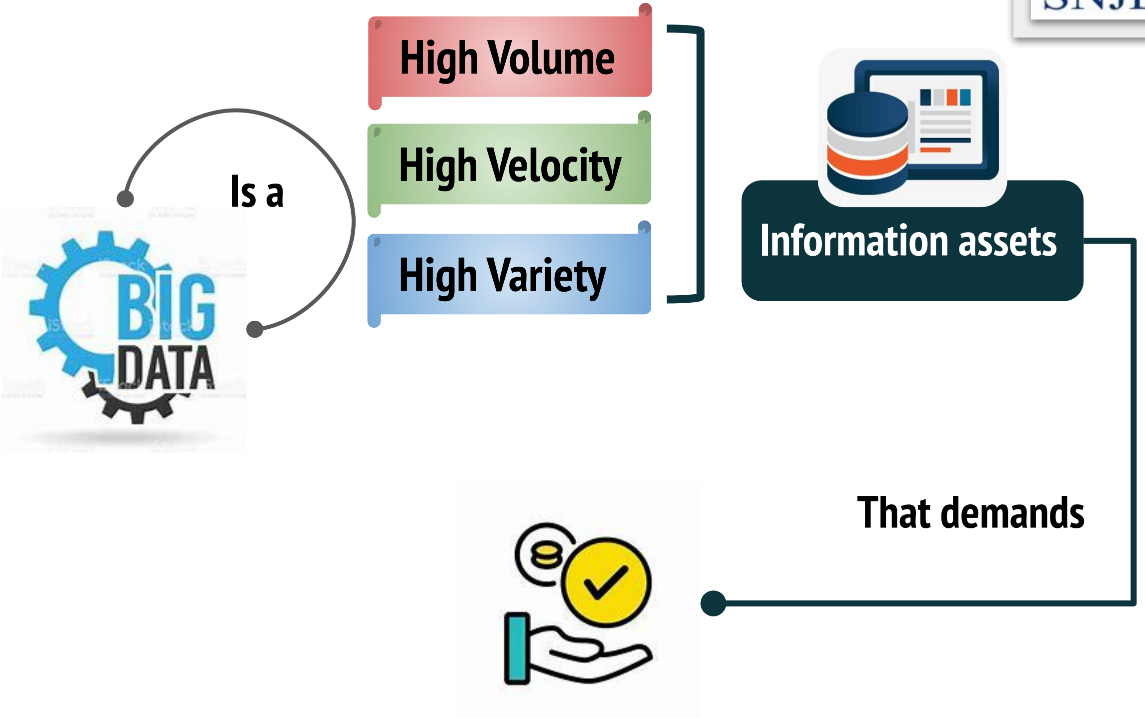
Outline

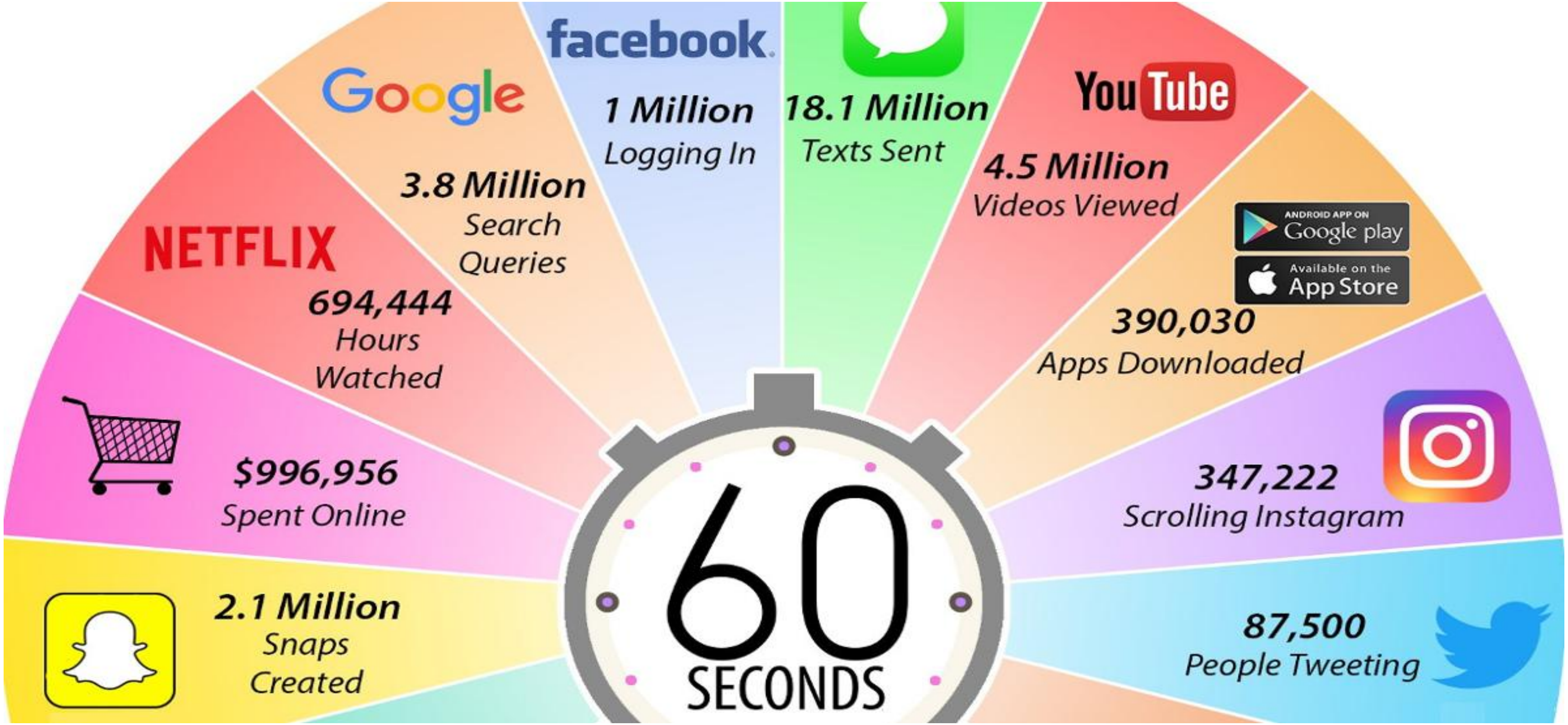
- Introduction to big data
- Sources of Big Data
- Data Analytic Lifecycle

Introduction to Big Data

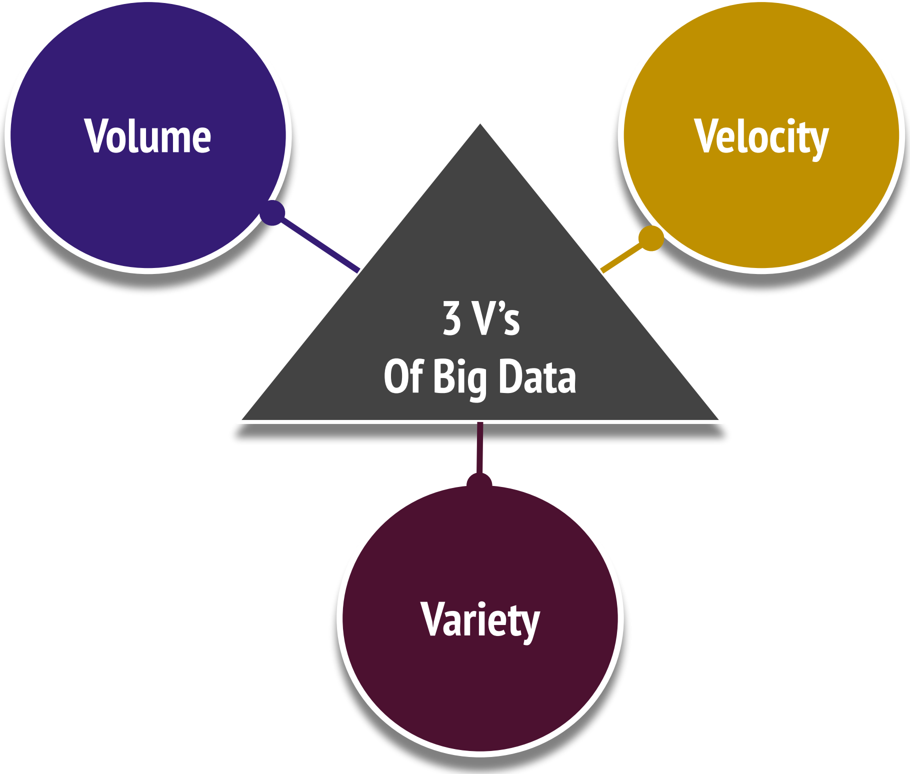


Cost effective innovation forms of information cost effective innovative forms of information processing for enhanced insight and decision making

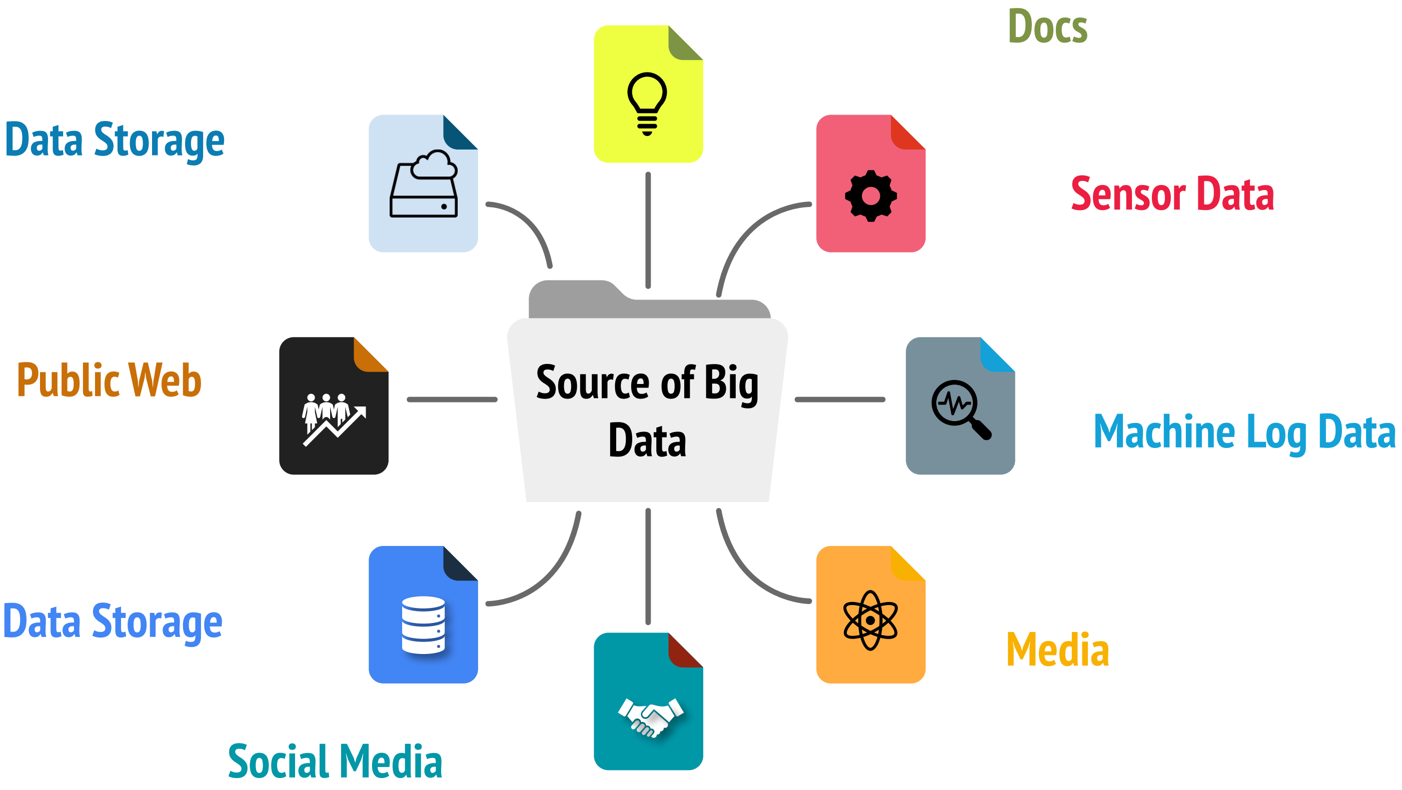
Data Generated per Minute On The Internet

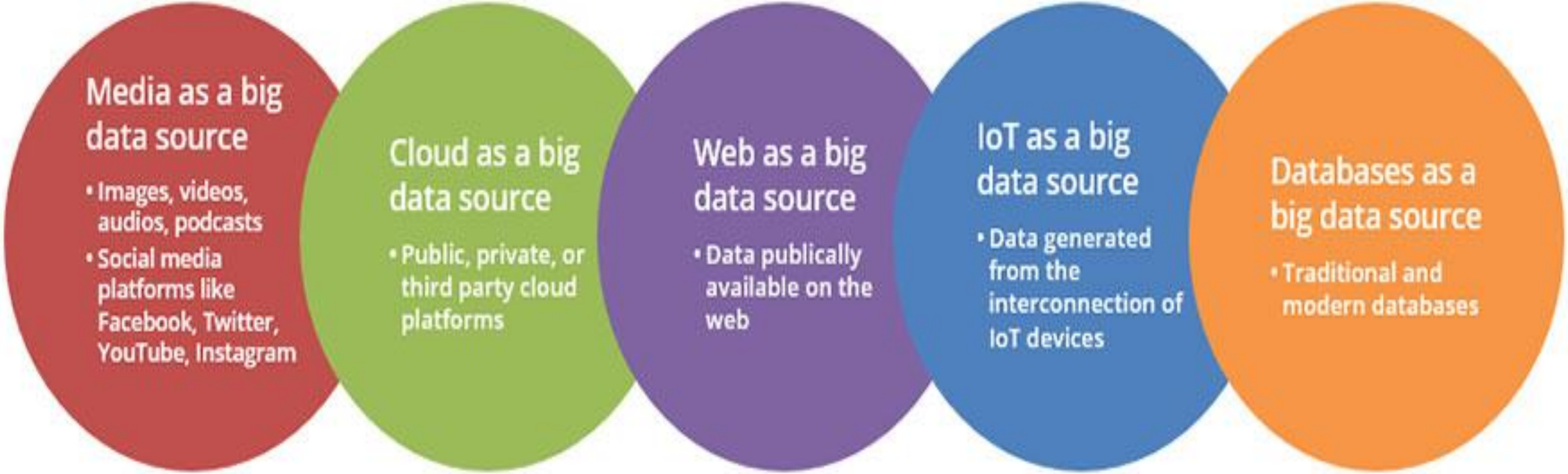


Introduction to big data



Sources of Big Data





Media as a big data source

- Images, videos, audios, podcasts
- Social media platforms like Facebook, Twitter, YouTube, Instagram

Cloud as a big data source

- Public, private, or third party cloud platforms

Web as a big data source

- Data publically available on the web

IoT as a big data source

- Data generated from the interconnection of IoT devices

Databases as a big data source

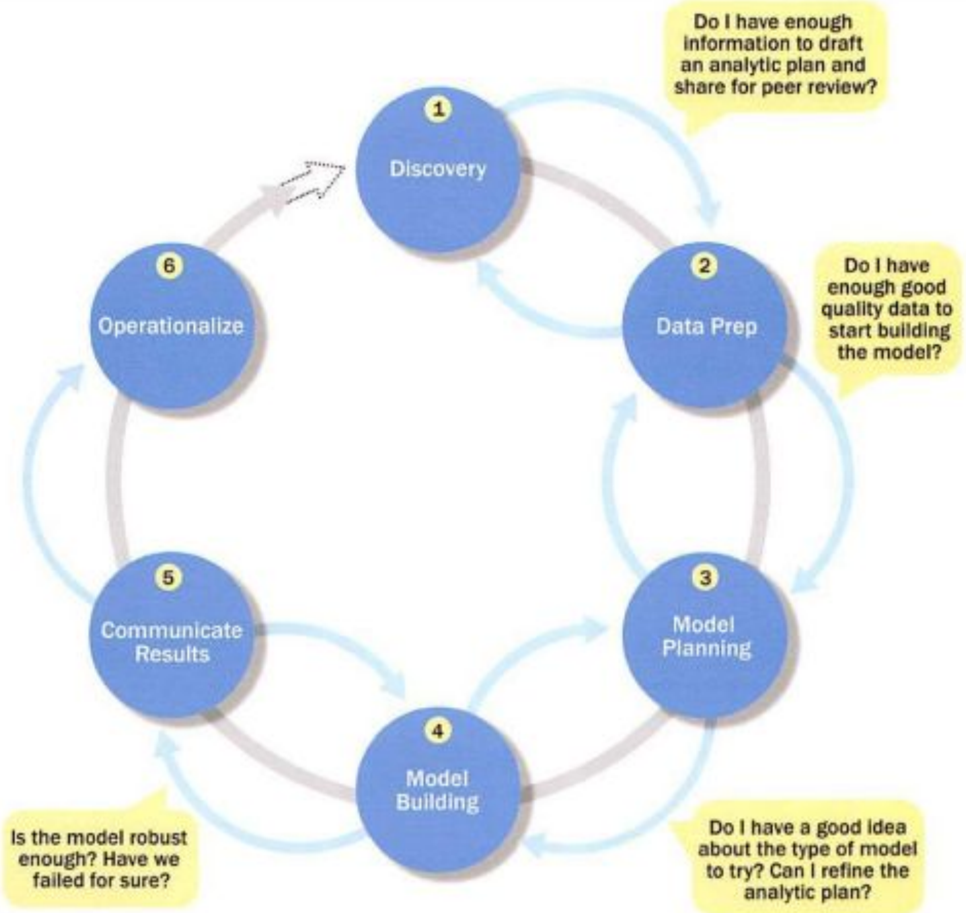
- Traditional and modern databases

Life Event Situations

- Big Data Analytics in different areas: retail,IT infrastructure, and social media.

- Big Data presents many opportunities to improve sales and marketing analytics.

Data Science Life Cycle


1
Discovery
2
Data Prep
3
Model Planning
4
Model Building
5
Communicate Results
6
Operationalize
Do I have enough information to draft an analytic plan and share for peer review?
Do I have enough good quality data to start building the model?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
Is the model robust enough? Have we failed for sure?

Data Science Life Cycle

Phase 1 : Discovery

- In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn.
- The team assesses the resources available to support the project in terms of people, technology, time, and data.
- Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data

Data Science Life Cycle

Phase 1 : Discovery

- Learning the business domain.
- Resources
- Framing the problem.
- Identifying key stakeholders.
- Interviewing the Analytics Sponsor.
- Developing initial hypotheses.
- Identifying **potential data sources.**

Data Science Life Cycle

Phase 2 : Data preparation

- Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.
- The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox.
- The ELT and ETL are sometimes abbreviated as ETLT.
- Data should be transformed in the ETLT process so the team can work with it and analyze it.
- In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data

Data Science Life Cycle

Phase 2 : Data preparation

- Explore
- Pre-process
- Condition data.
- Do I have enough good quality data to start building the model?
- ELTL : ETL software, or Extract-Transform-Load, is used to manage all aspects of data preparation
- 50 % of time.

Data Science Life Cycle

Phase 2 : Data preparation

- Preparing the Analytic Sandbox.
- Performing ETLT
- Learning about data
- Data Conditioning
- Survey and Visualize
- Tools- hadoop,alpine miner,openrefine
- Data Wrangler

Data Science Life Cycle

Phase 3 : Model planning

- Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.
- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

Data Science Life Cycle

Phase 3 : Model planning

- Data Exploration and variable selection
- Model selection
- Do I have a good idea about the type of model to try? Can I refine the analytic plan?
- Common tools

  * R –5000 packages for data analysis and graphical.

  * SQL Analysis services

  * SAS/ACCESS

Phase 4 : Model building

- In Phase 4, the team develops datasets for testing, training, and production purposes.
- In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
- The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and work flows (for example, fast hardware and parallel processing, if applicable).

Data Science Life Cycle

Phase 4 : Model building

- Develop Analytical model and train it.
- Model build on training data, fit on train data and evaluated with test.
- Is the model robust? Have we failed for sure?
- Is the model robust? Have we failed for sure?
- **Common tools**
- **Commercial tools**
- – SAS Enterprise Miner, SPSS Modeller, MatLab, Statistica
- **Free and Open Source Tools**
- – R, Octave,Weka,Python,SQL

Data Science Life Cycle

Phase 5 : Communicate results

- In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1.
- The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

Data Science Life Cycle

Phase 5 : Communicate results

- **Result as success or failure.**
- Identify key findings, quantify the business value.
- Develop narrative to summarize and convey findings to stakeholders.

Data Science Life Cycle

Phase 6 : Operationalize

- In Phase 6, the team delivers final reports, briefings, code, and technical documents.
- In addition, the team may run a pilot project to implement the models in a production environment.

Data Science Life Cycle

Phase 6 : Operationalize

- **Team delivers final reports, briefings, code and technical documents.**
- May run a pilot project to implement the models in a production environment.
- Free or open source tools - Octave, WEKA, MADlib.

Data Science Life Cycle

Class B

- Introduction to Big Data and Sources of Big Data 53 Tejal Khairnar
- Phase 1 Akansha
- Phase 2: Mayuri
- Phase 3: Govinda
- Phase 4::Gayatri
- Phase 5: Timple
- Phase 6: Amol

Case Study

- **Global Innovation Social Network and Analysis (GINA).**
- EMC's **Global Innovation Network and Analytics** (GINA) team is a group of senior technologists located in centers of excellence (COEs) around the world.