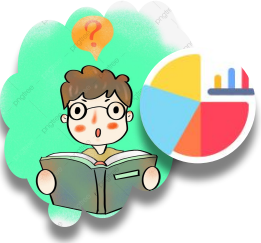
Outline

- Introduction to Data Visualization
- Challenges to Big data visualization
- Types of data visualization
- Data Visualization Techniques
- Tools used in Data Visualization
- Hadoop ecosystem, Map Reduce, Pig, Hive,



Introduction

Data Visualization

is

Graphical

Representation

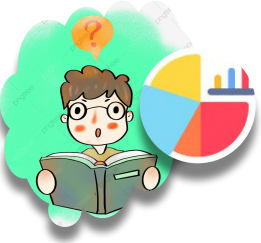
of

Information

Data

Data Visualization

- Data visualization is a graphical representation of any data or information.
- Visual elements such as charts, graphs, and maps are the few data visualization tools that provide the viewers with an easy and accessible way of understanding the represented information.
- Data visualization enables you or decision-makers of any enterprise or industry to look into analytical reports and understand concepts that might otherwise be difficult to grasp.



Introduction

Data Visualization Tools

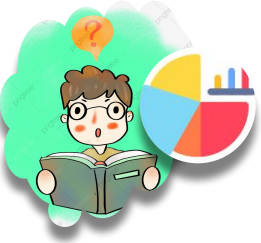
Provide

Accessible way

To

See & Understand

trends, outliers, and patterns in data.



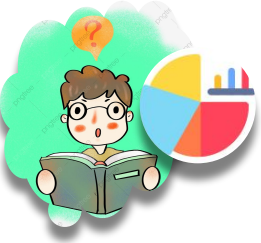
Introduction

Visual Elements

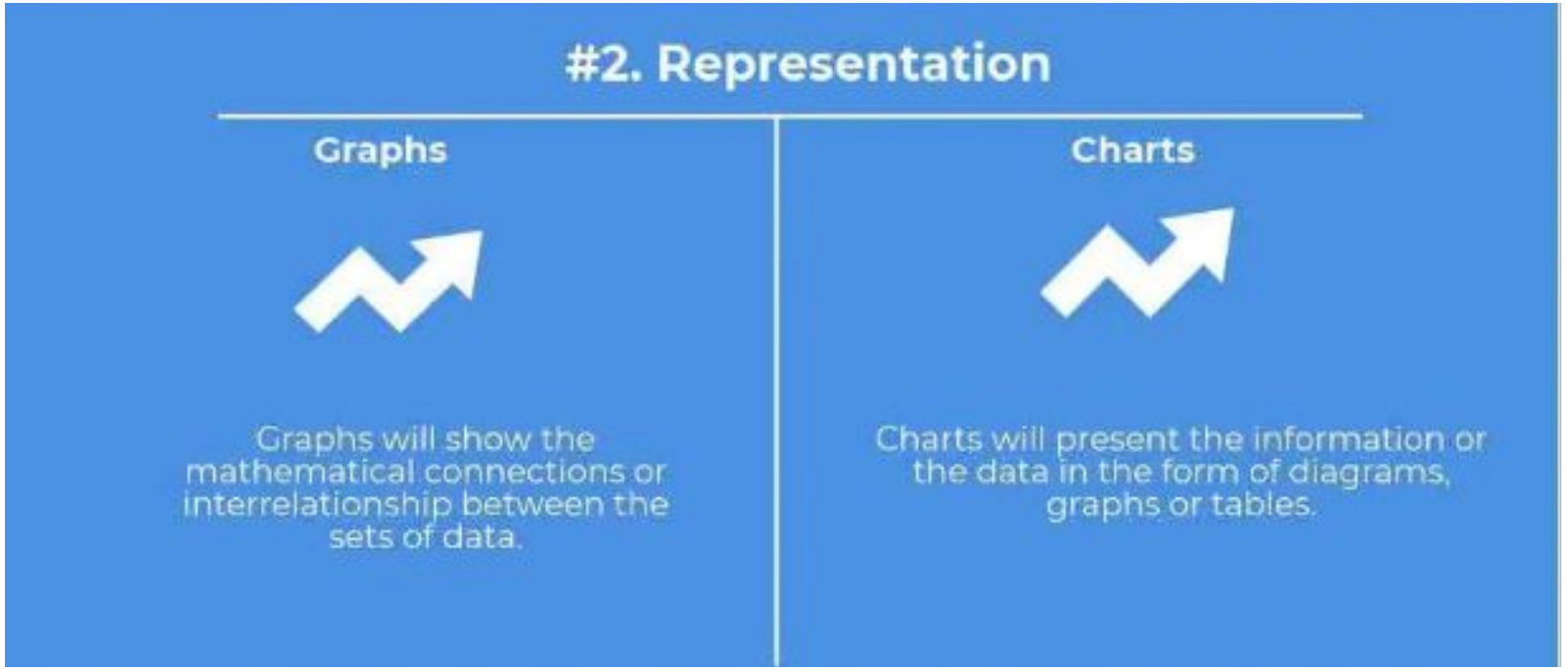
Charts

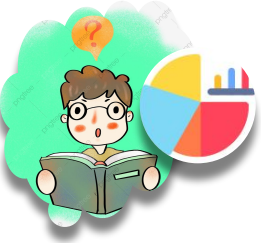
Graphs

Maps



Introduction

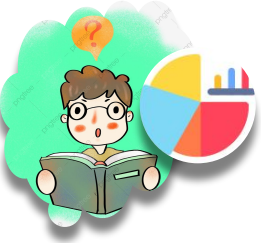




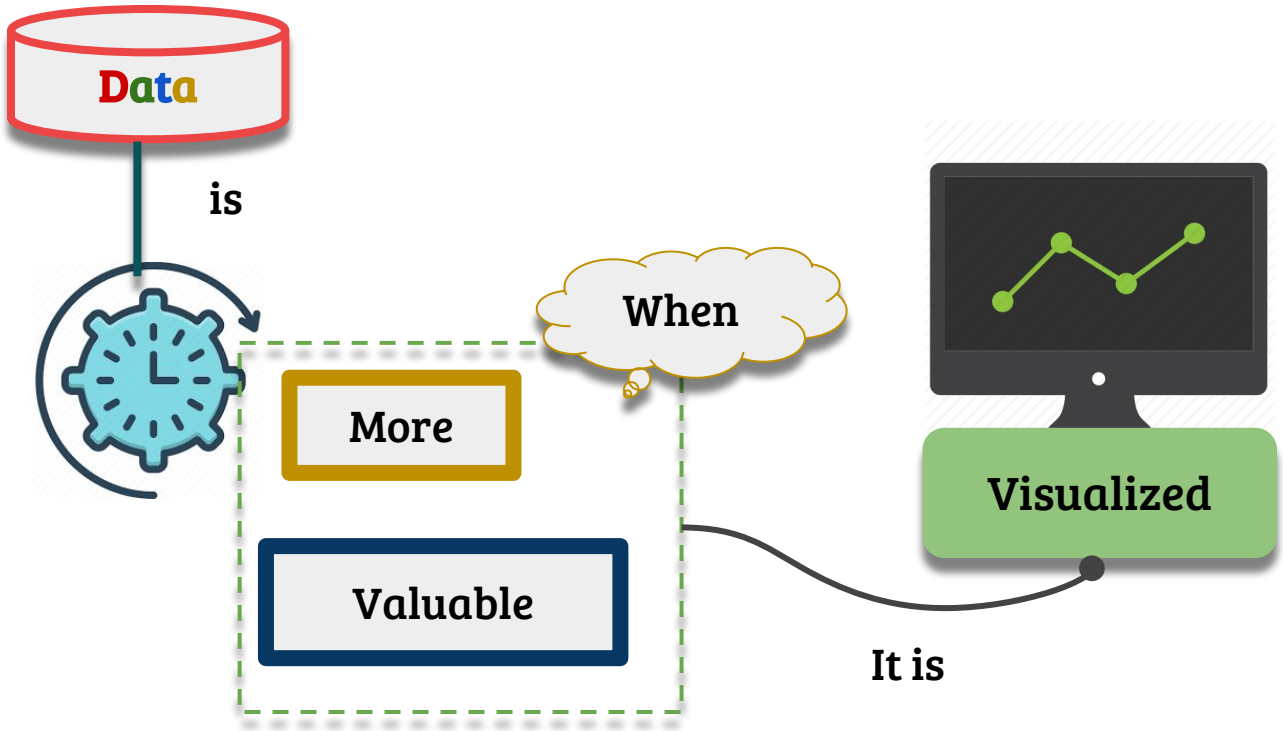
Introduction

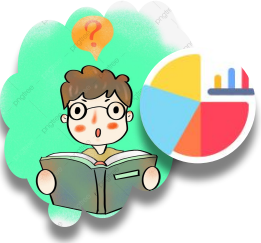
#3. Are they same?

Graphs	Charts
	
All kind of graphs are charts; hence graphs are one kind of chart.	All kind of charts are not graphs.

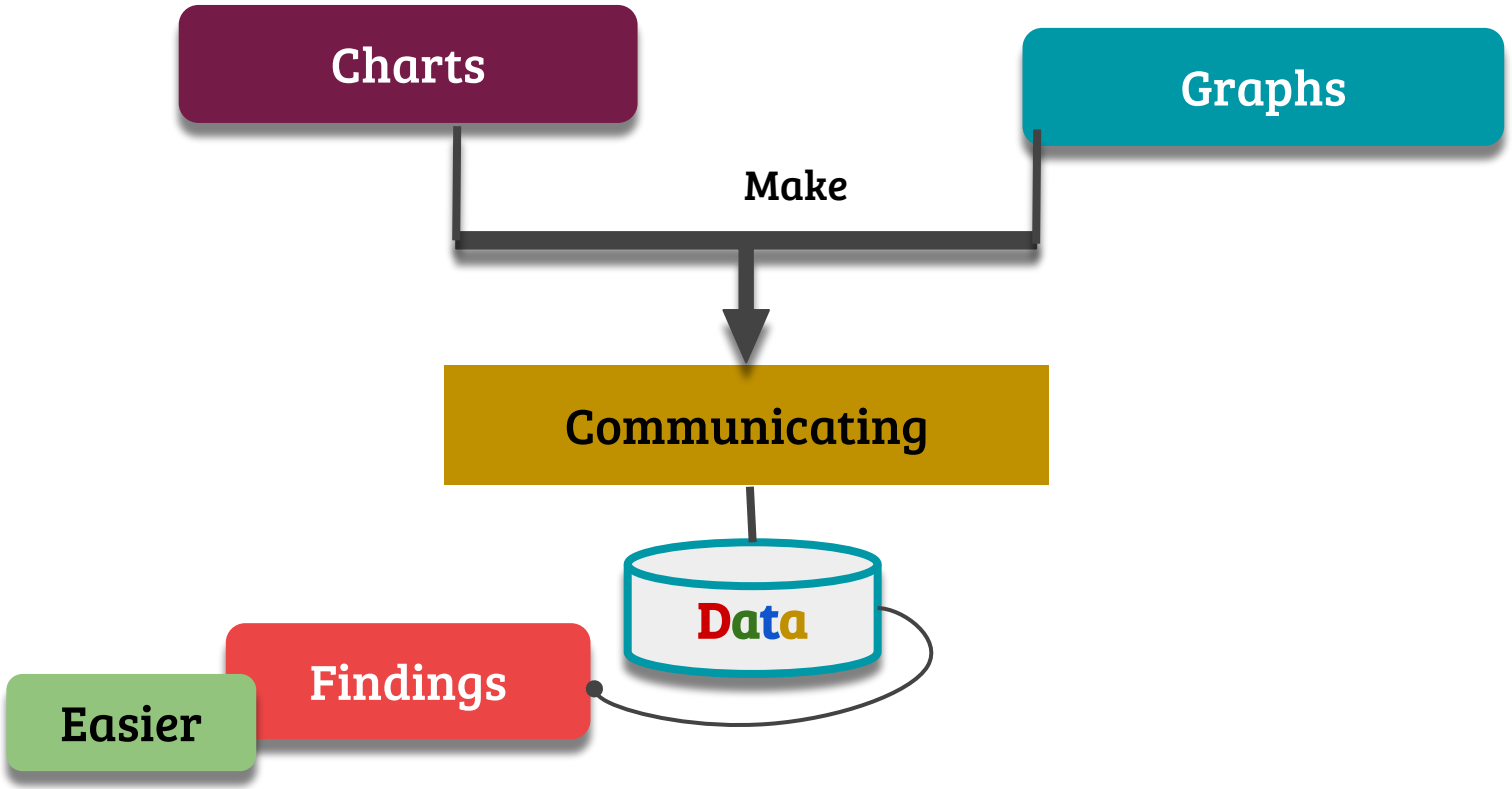


Its Need



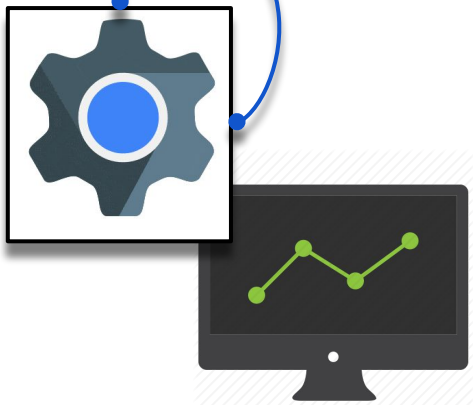


Its Need



Its Need

Without

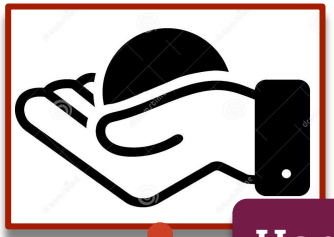


Visual representation

of

insights

It can be



Hard

For the

Audience to Grasp

the

True Meaning

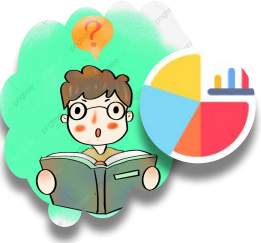
of

Findings

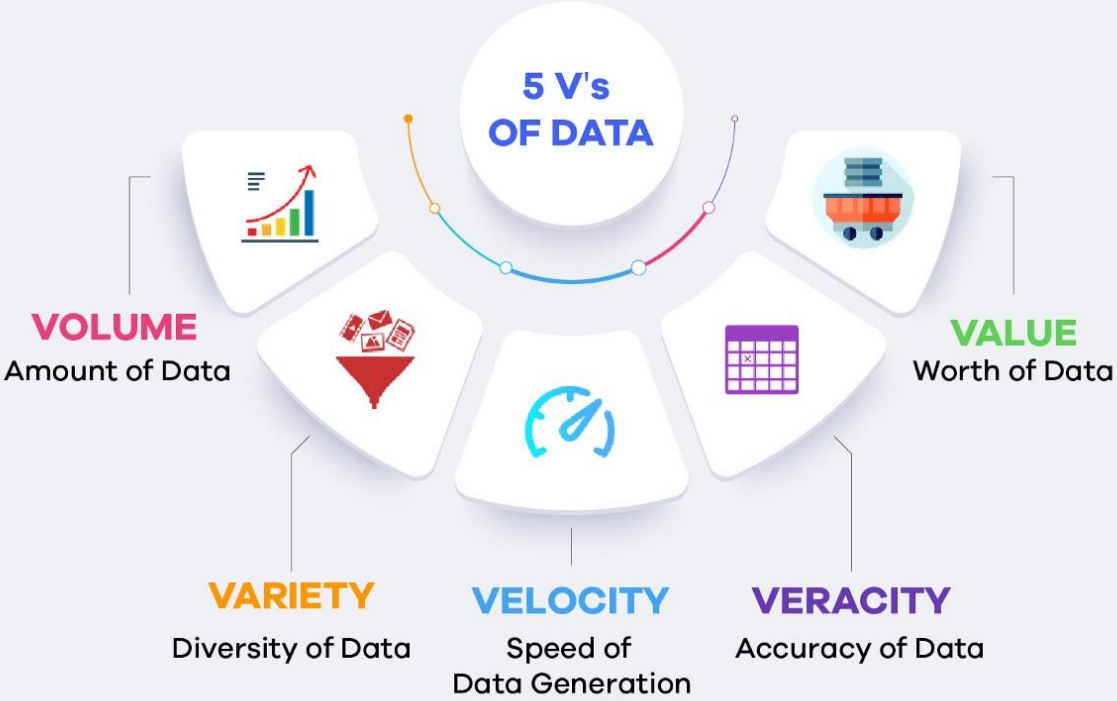
#SNJBCOE #ComputerEngineering

Outline

- Introduction to Data Visualization
- Challenges to Big data visualization
- Types of data visualization
- Data Visualization Techniques
- Tools used in Data Visualization
- Hadoop ecosystem, Map Reduce, Pig, Hive,



Big data visualization Challenge

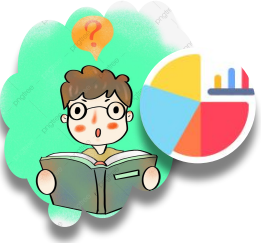


Big data visualization Challenge

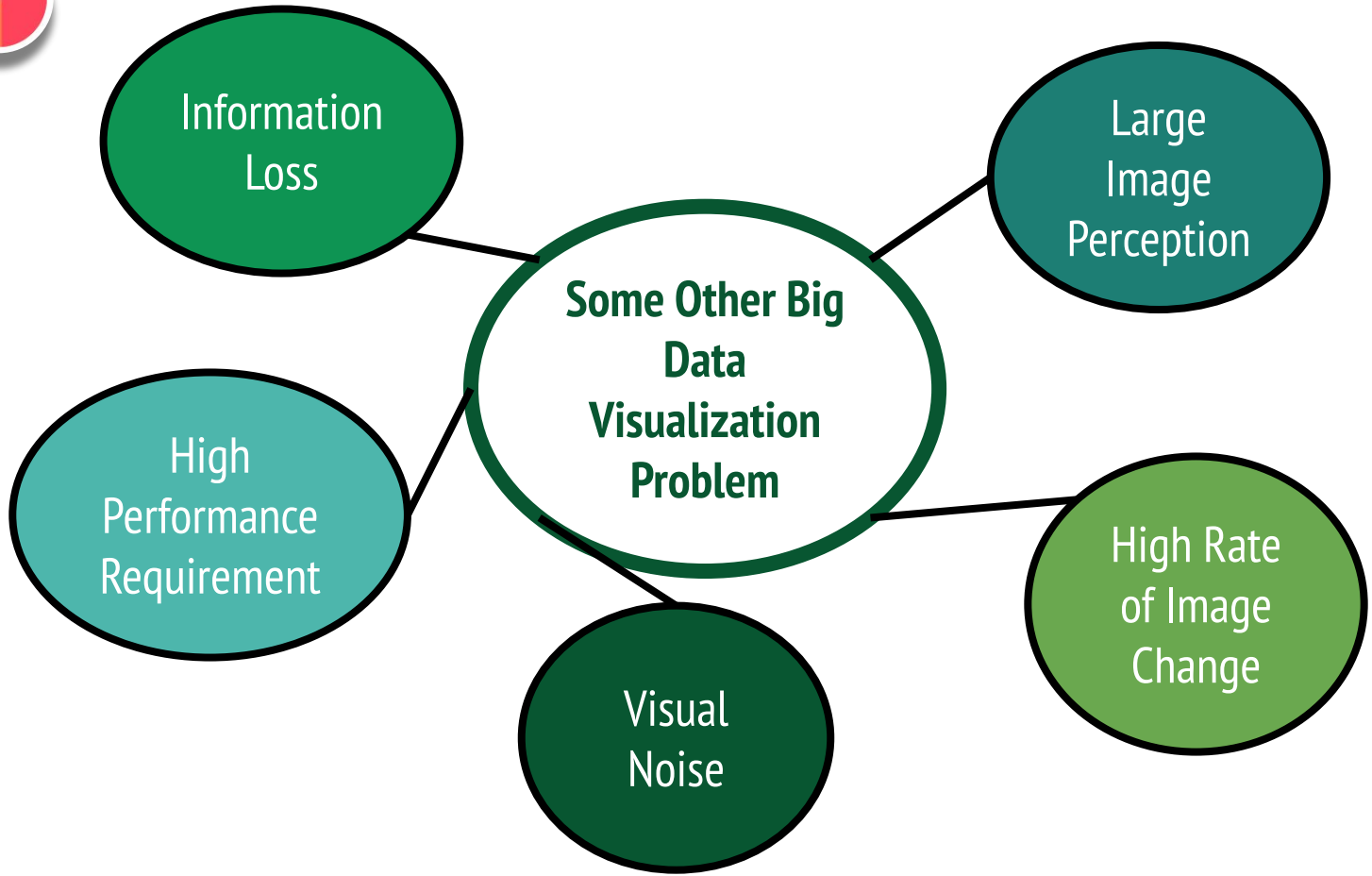
- Visualization of big data with diversity and heterogeneity (structured, semi-structured, and unstructured) is a big problem.
- Speed is the desired factor for the big data analysis.
- Designing a new visualization tool with efficient indexing is not easy in big data.
- Cloud computing and advanced graphical user interface can be merged with the big data for the better management of big data scalability

Big data visualization Challenge

- Visualization systems must contend with unstructured data forms such as graphs, tables, text, trees, and other metadata.
- Big data often has unstructured formats.
- Due to bandwidth limitations and power requirements, visualization should move closer to the data to extract meaningful information efficiently.
- Visualization software should be run in an in smooth manner. Because of the big data size, the need for massive parallelization is a challenge in visualization.



Big data visualization Challenge



Big data visualization Challenge

Visual noise

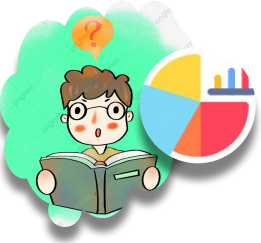
Most of the objects in dataset are too relative to each other. Users cannot divide them as separate objects on the screen.

Information loss

Reduction of visible data sets can be used, but leads to information loss

Large image perception

Data visualization methods are not only limited by aspect ratio and resolution of device, but also by physical perception limits.



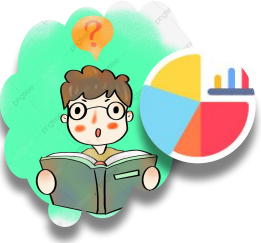
Big data visualization Challenge

High rate of image change:

Users observe data and cannot react to the number of data change or its intensity on display.

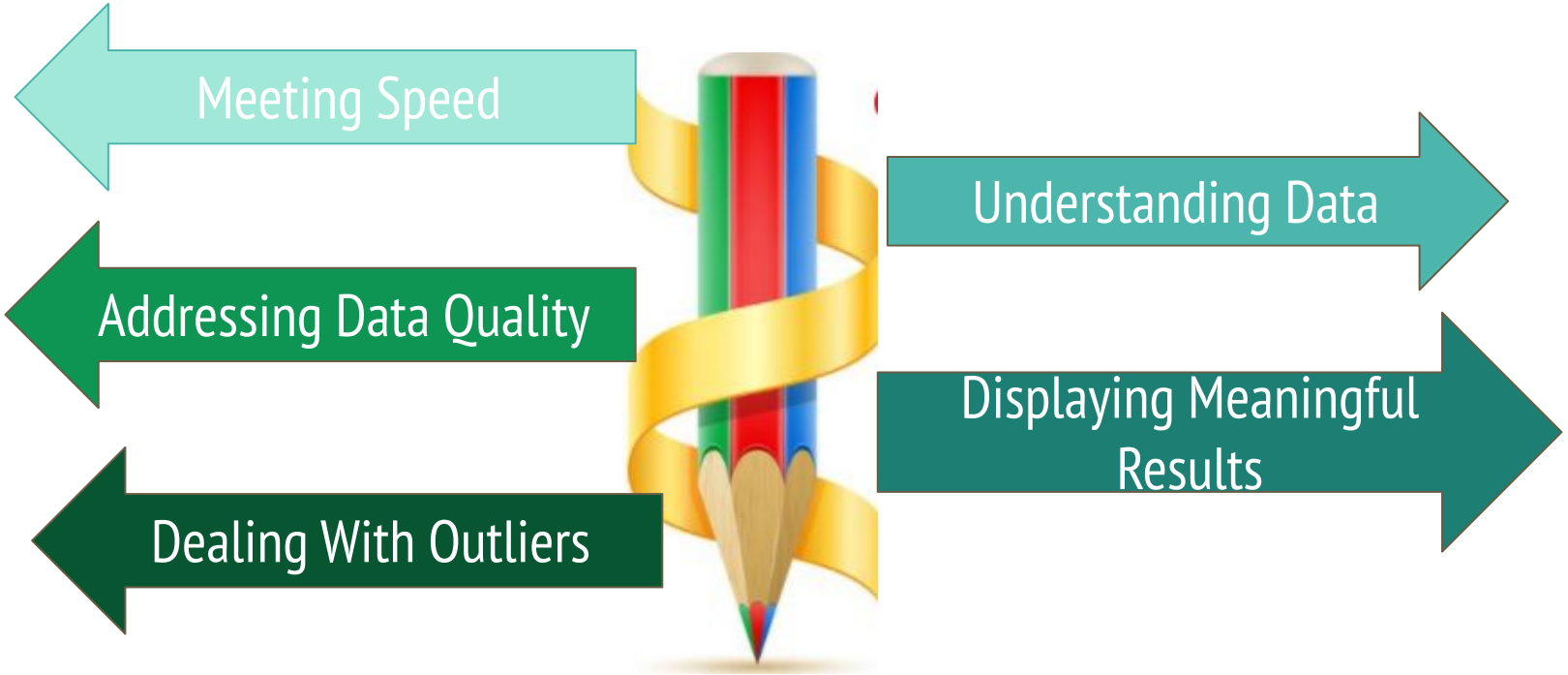
High performance requirements

It can be hardly noticed in static visualization because of lower visualization speed requirements , high performance requirement.



Big data visualization Challenge

Solution



#SNJBCOE #ComputerEngineering

Big data visualization Challenge

Meeting the need for speed

One possible solution is hardware. Increased memory and powerful parallel processing can be used. Another method is putting data in-memory but using a grid computing approach, where many machines are used.

Understanding the data

One solution is to have the proper domain expertise in place.

Big data visualization Challenge

Addressing data quality

It is necessary to ensure the data is clean through the process of data governance or information management.

Displaying meaningful results

One way is to cluster data into a higher-level view where smaller groups of data are visible and the data can be effectively visualized

Dealing with outliers

Possible solutions are to remove the outliers from the data or create a separate chart for the outliers.

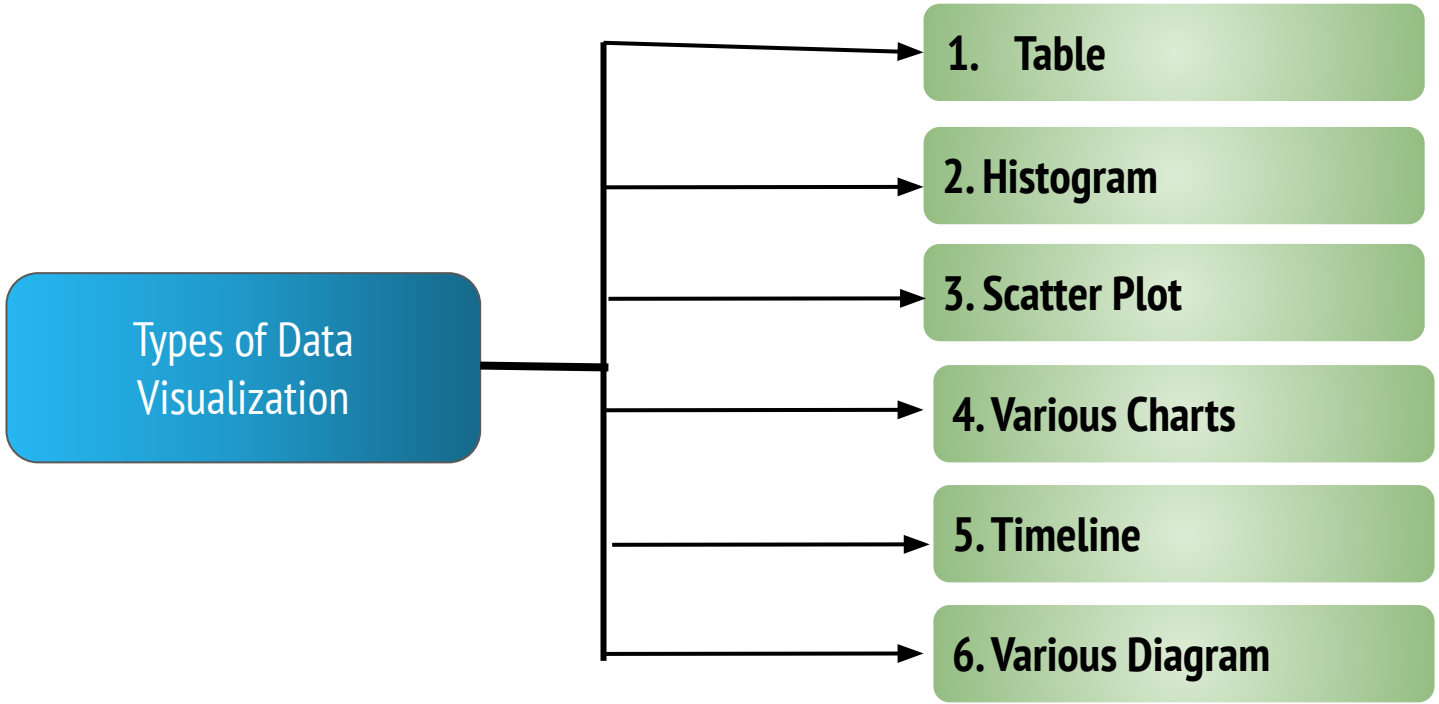
Outline

- Introduction to Data Visualization
- Challenges to Big data visualization
- Types of data visualization
- Data Visualization Techniques
- Tools used in Data Visualization
- Hadoop ecosystem, Map Reduce, Pig, Hive,

Types of data visualization

- Many conventional data visualization methods are often used.
- They are:
  - Table, histogram, scatter plot, line chart, bar chart, pie chart, area chart, flow chart, bubble chart,
  - multiple data series or combination of charts, timeline,
  - Venn diagram, data flow diagram, and entity relationship diagram, etc.
  - The additional methods are: parallel coordinates, treemap, cone tree, and semantic network, etc

Types of data visualization



Types of data visualization

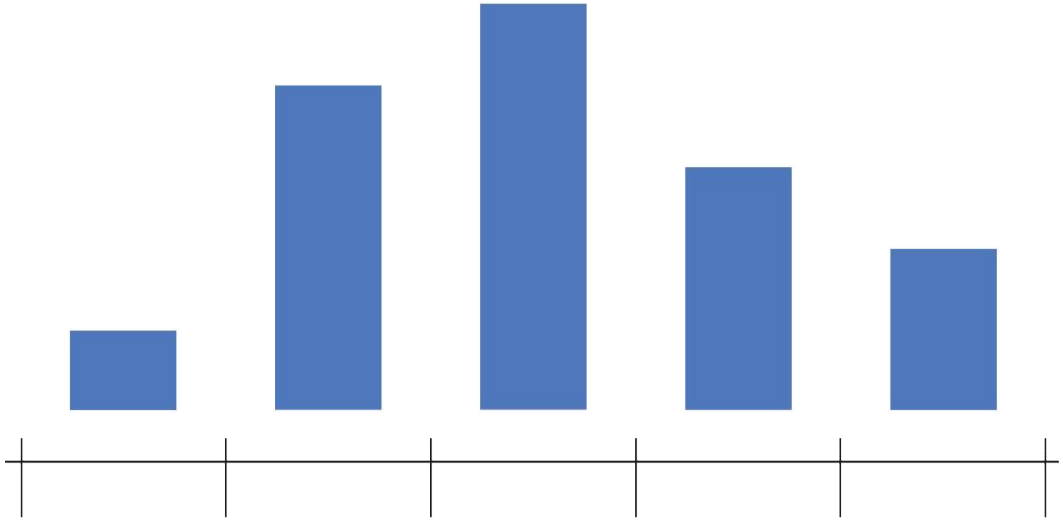
Table

		619 ENTRY RATING						
C EXIT RATING		1	2	3	4	5	6	7
24	1	21	2	1	0	0	0	0
114	2	30	77	7	0	0	0	0
275	3	47	54	150	14	7	0	3
334	4	32	58	54	151	18	7	14
550	5	40	75	74	44	267	25	25
459	6	26	53	72	31	45	205	27
140	7	4	13	18	9	15	20	61
1896		200	332	376	249	352	257	130

Types of data visualization

Histogram

- The data is grouped into ranges(eg.10-29) & then plotted as connected bars



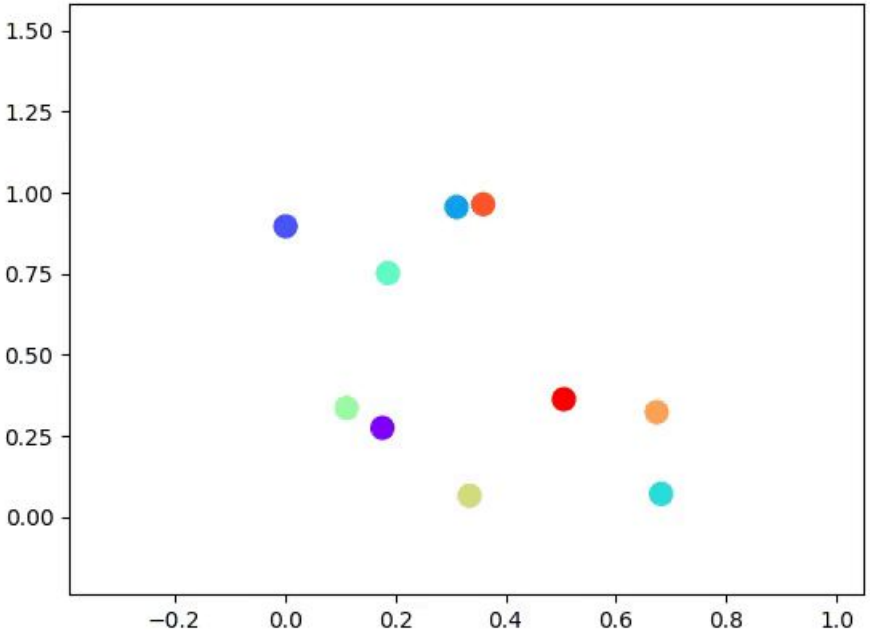
Types of data visualization

- An approximate representation of the distribution of numerical data. Divide the entire range of values into a series of intervals and then count how many values fall into each interval this is called binning
- For example, determining frequency of annual stock market percentage returns within particular ranges (bins) such as 0-10%, 11-20%, etc. The height of the bar represents the number of observations (years) with a return % in the range represented by the respective bin.

Conventional data visualization methods

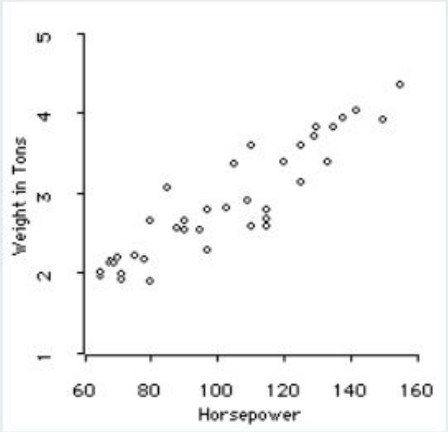
Scatter plot

- It display collection of all points for the set of data
- When you have multiple data points and need to examine the correlation between X and Y variables.
- Consequently, variables should depend on each other or influence each other in some way.
- For example, supply is usually related to demand.

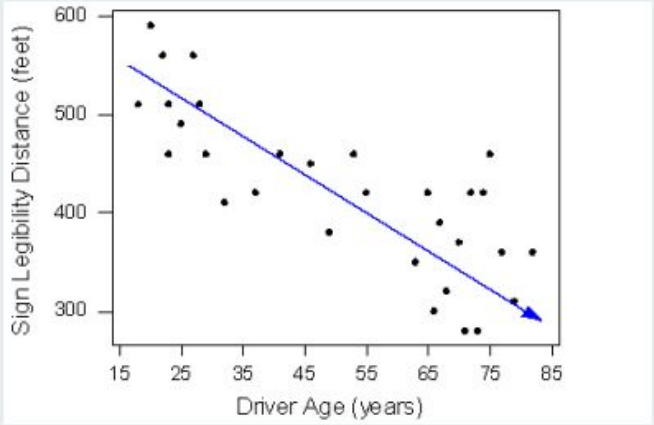


Conventional data visualization methods

Scatter plot



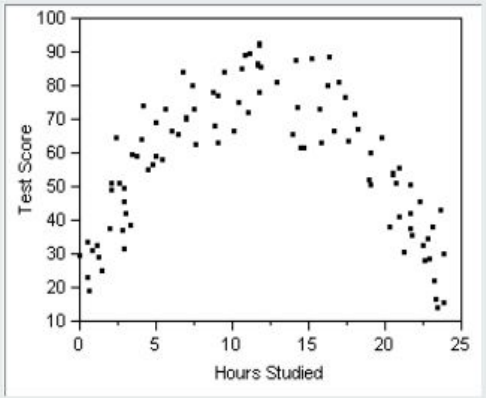
Linear Association



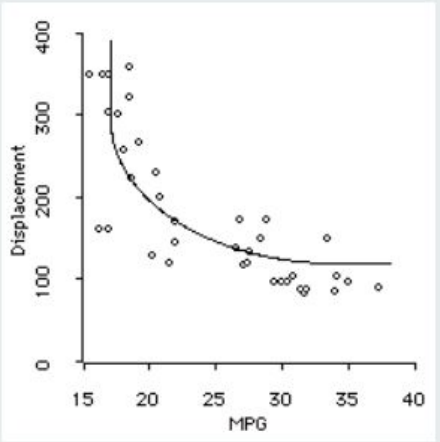
Linear Association

Conventional data visualization methods

Scatter plot



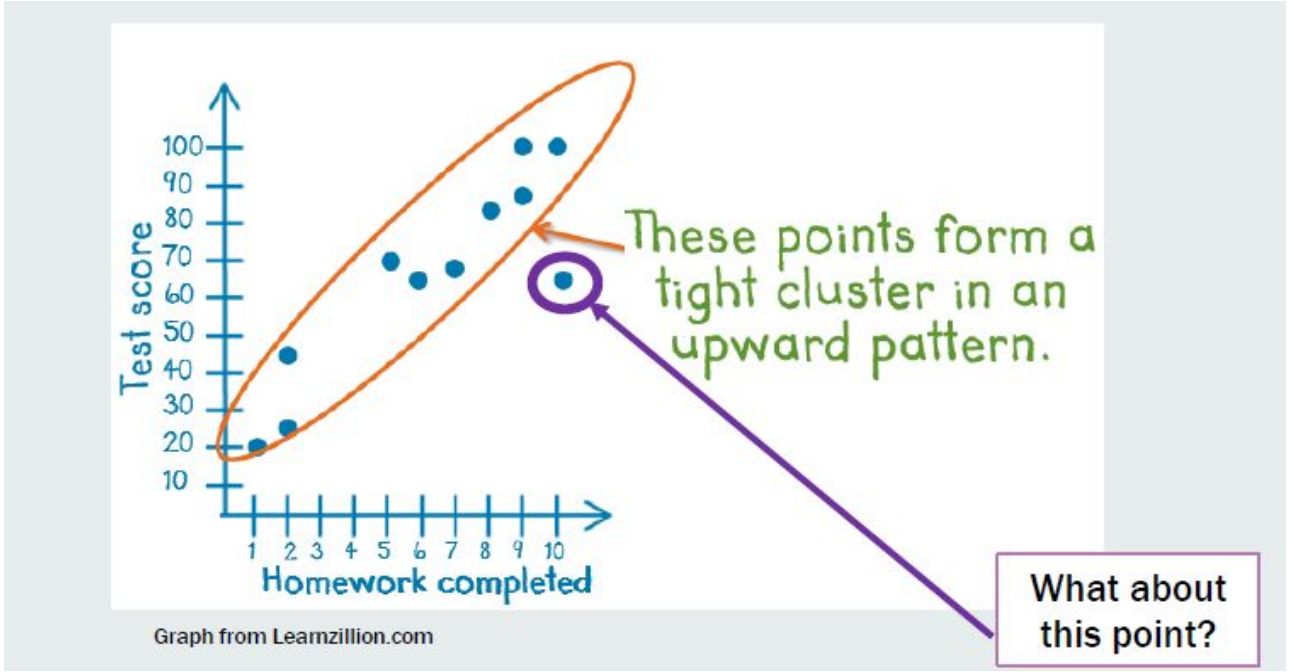
Non- Linear Association



Non- Linear Association

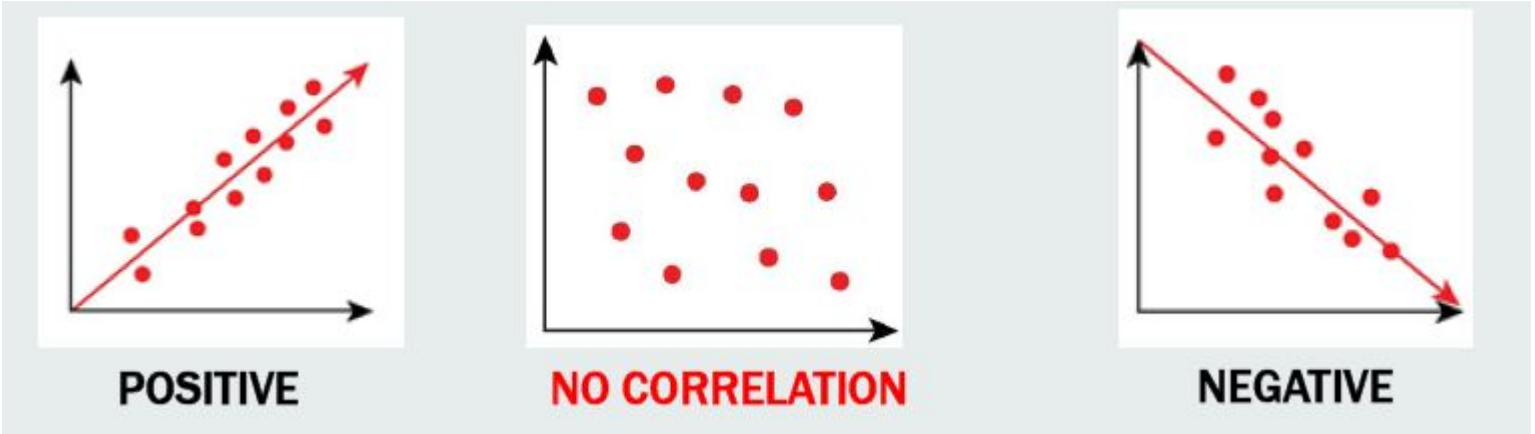
Conventional data visualization methods

Scatter plot



Conventional data visualization methods

Scatter plot

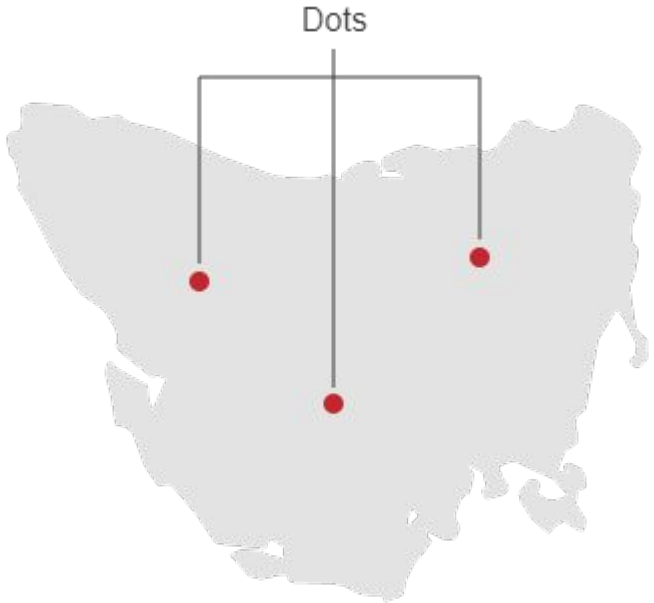


Conventional data visualization methods

Different Types of Chart

Dot distortion map

- Dot symbol to represent a feature on the map

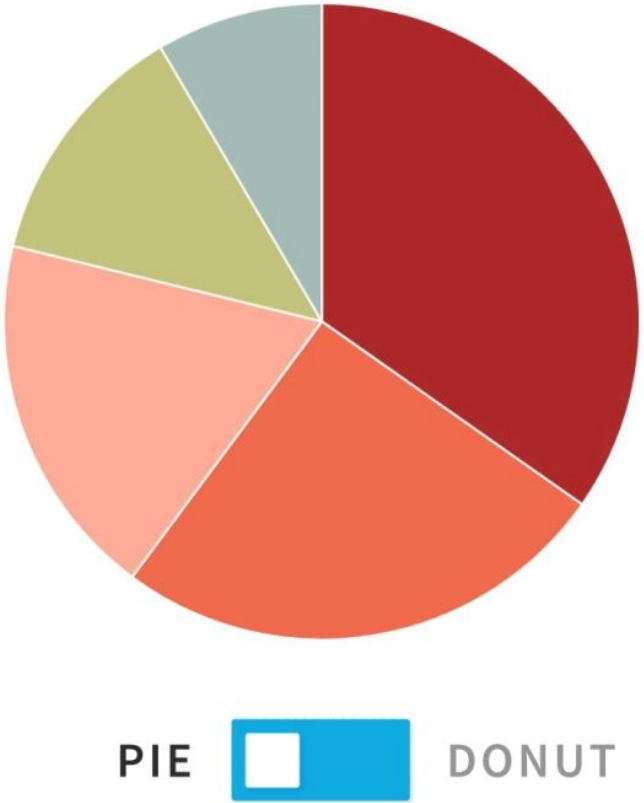


Conventional data visualization tools

Different Types of Chart

Pie Chart

- The circle is divided into sector to represented numeric proportion



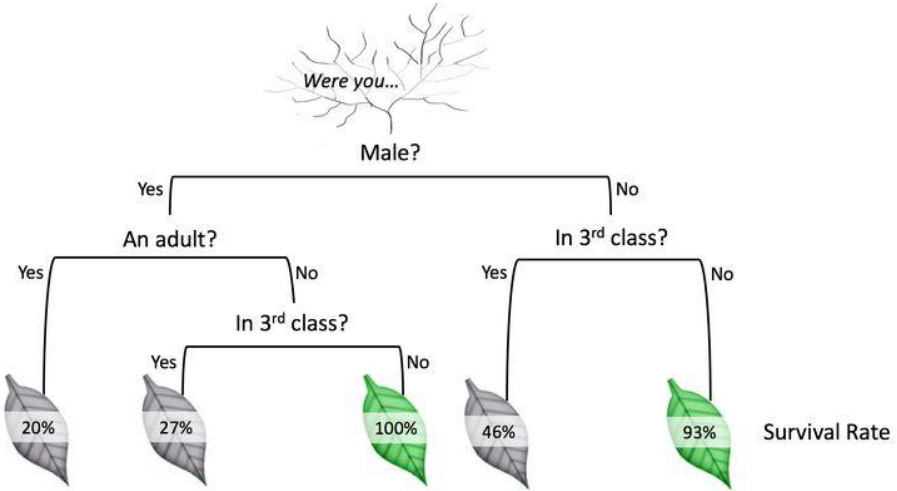
Conventional data visualization methods

Different Types of Chart

3) hierarchical

Tree Diagram

- It represent data or the hierarchy in the graph form



Conventional data visualization methods

Different Types of Chart

Node Link Diagram

- Node - Visualize as a Dot
- Link - Line Segment to display data connection

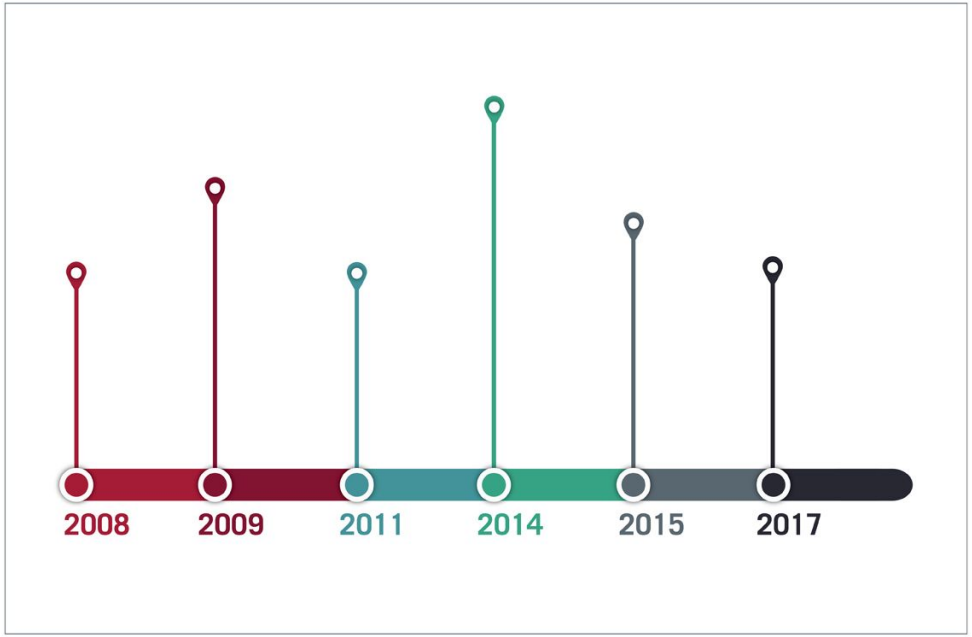
Conventional data visualization methods

Different Types of Chart- Click Here

Conventional data visualization methods

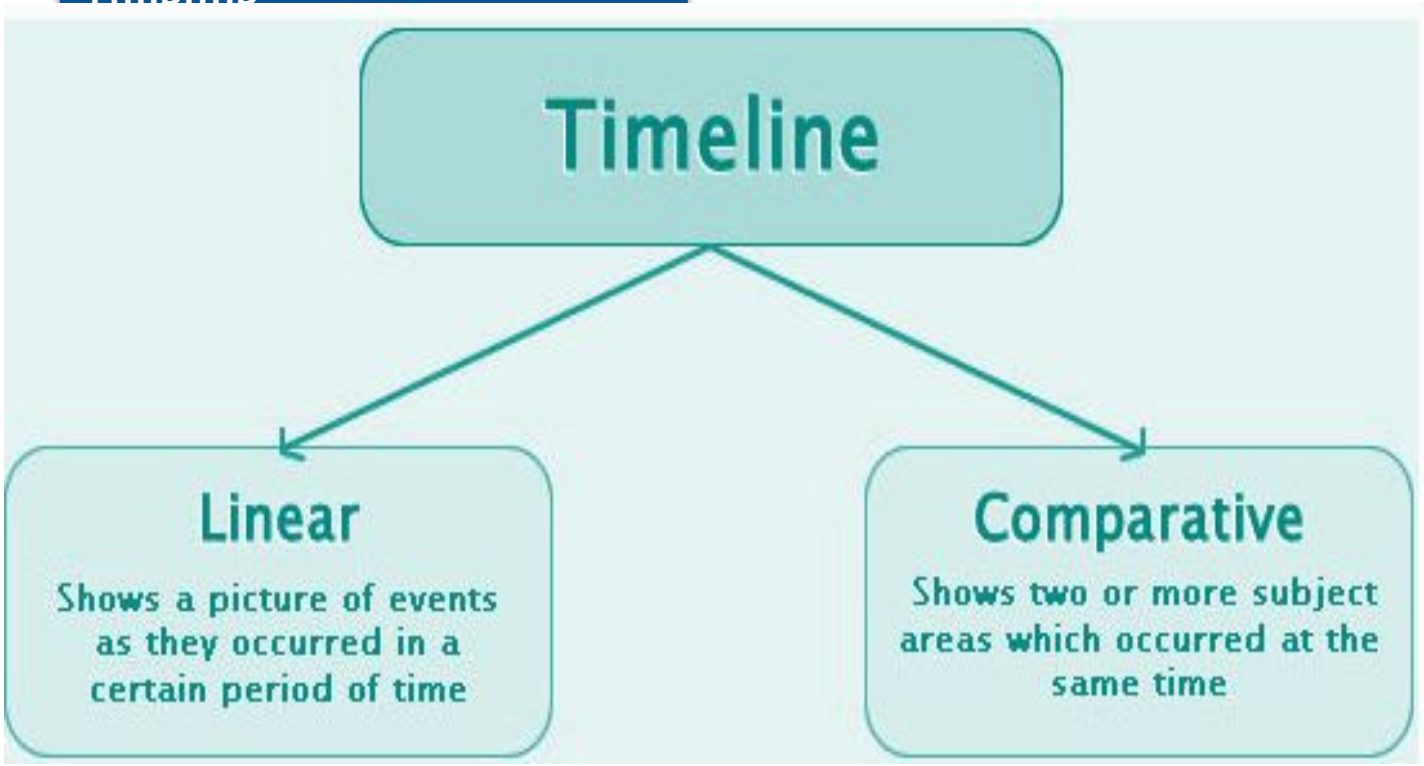
Timeline

- The most effective way to visualize a sequence of events in chronological order. T
- They are typically linear, with key events outlined along the axis.
- Timelines are used to communicate time-related information and display historical data.



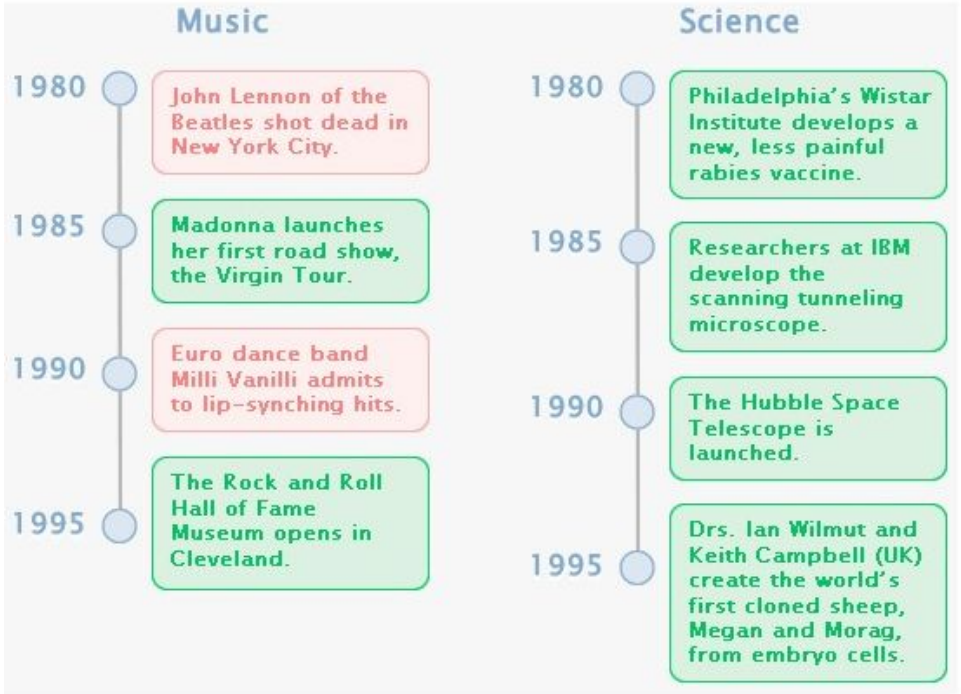
Conventional data visualization methods

Timeline



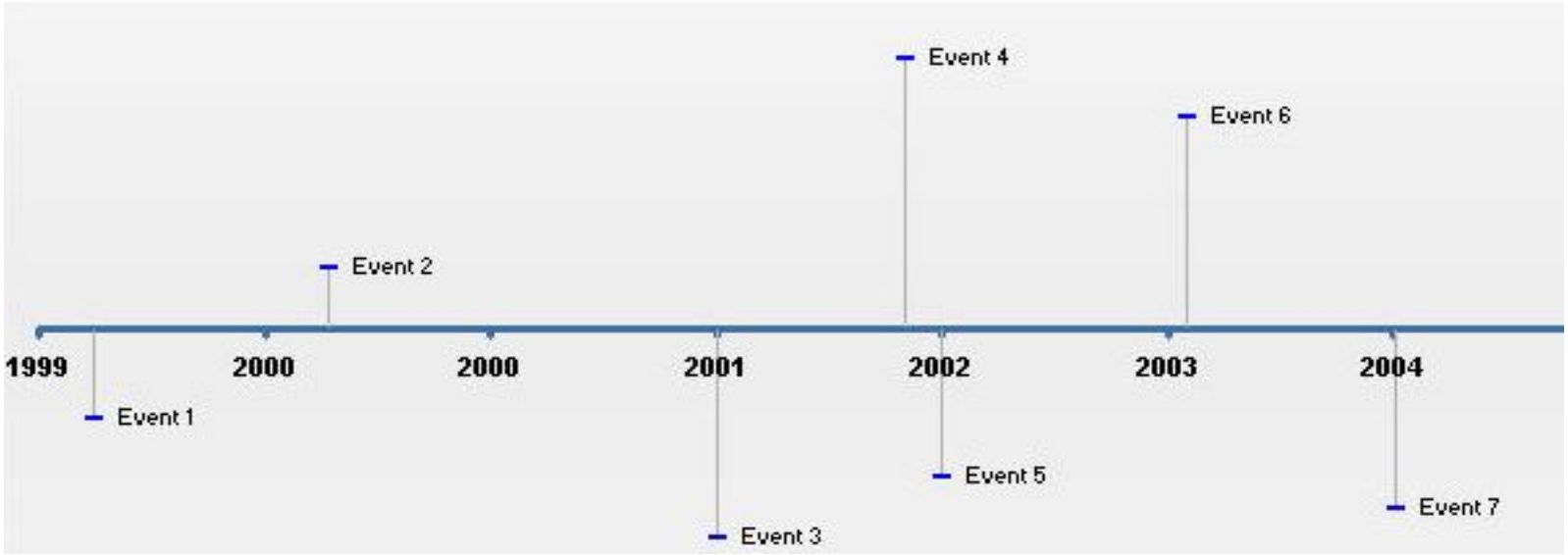
Conventional data visualization methods

Timeline



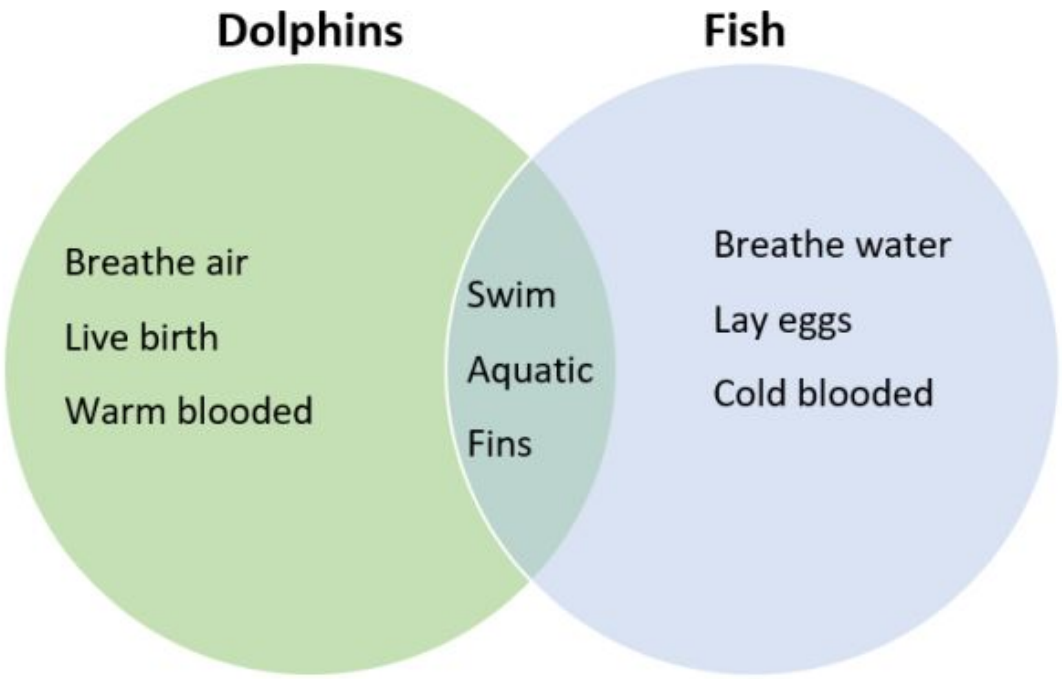
Conventional data visualization methods

Timeline



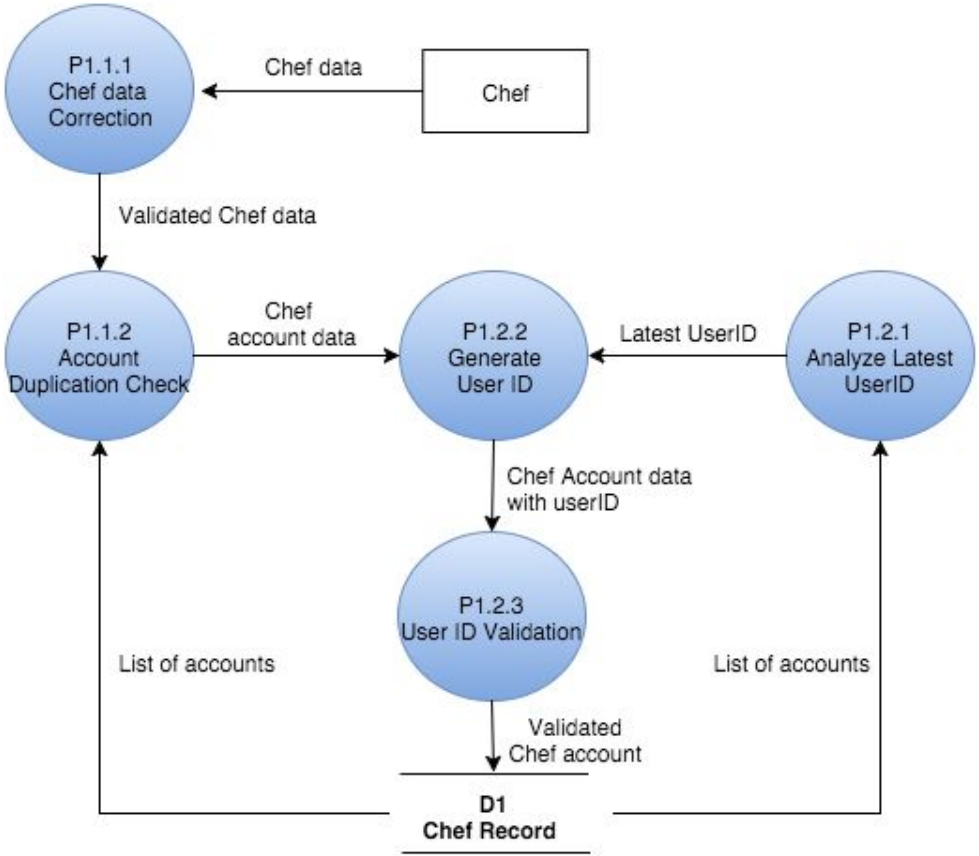
Conventional data visualization methods

Venn Diagram



Conventional data visualization methods

- Data Flow Diagram

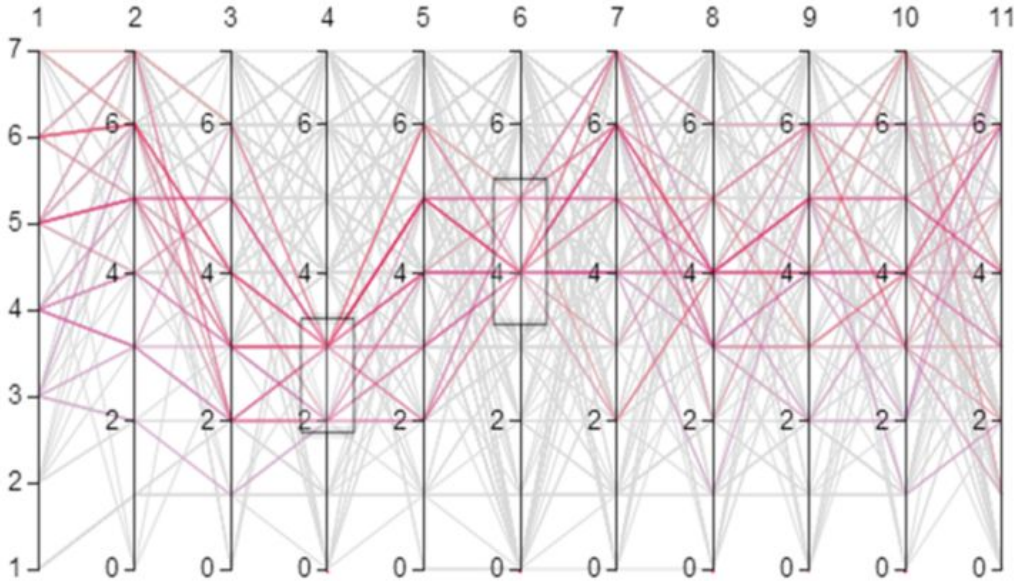


#SNJBCOE #ComputerEngineering

Conventional data visualization methods

Parallel Coordinates

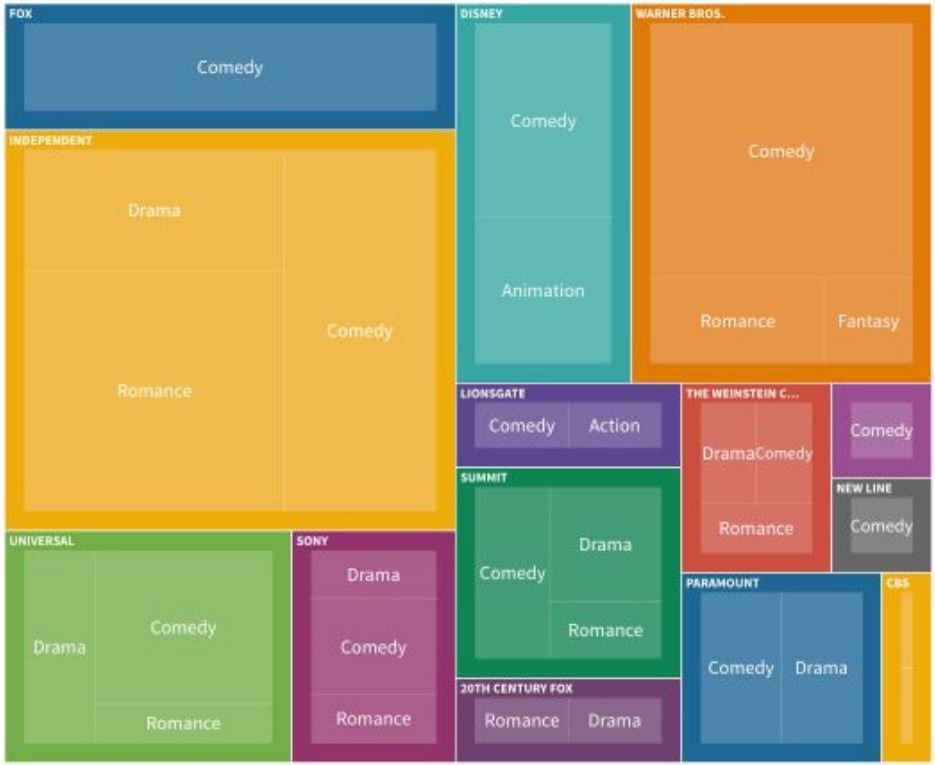
- to plot individual data elements across many dimensions.
- Parallel coordinate is very useful when to display multidimensional data



Conventional data visualization methods

Treemap

- an effective method for visualizing hierarchies.
- The size of each sub-rectangle represents one measure, while color is often used to represent another measure of data.
- streaming music and video tracks in a social network community.



Conventional data visualization methods

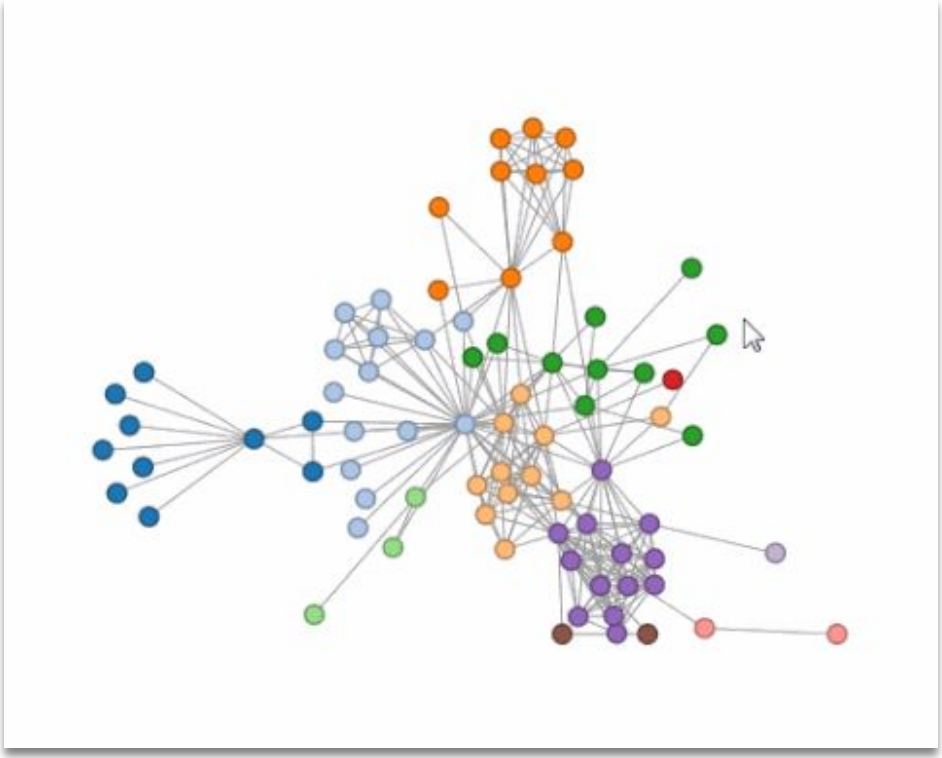
Cone tree

- It is method for displaying hierarchical data such as organizational body in three dimensions.
- The branches grow in the form of cone.

Conventional data visualization methods

Semantic Network

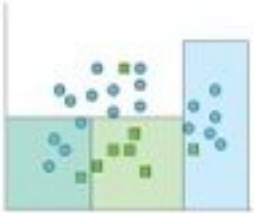
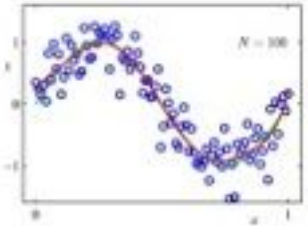
- a graphical representation of logical relationship between different concepts.
- It generates directed graph, the combination of nodes or vertices, edges or arcs, and label over each edge



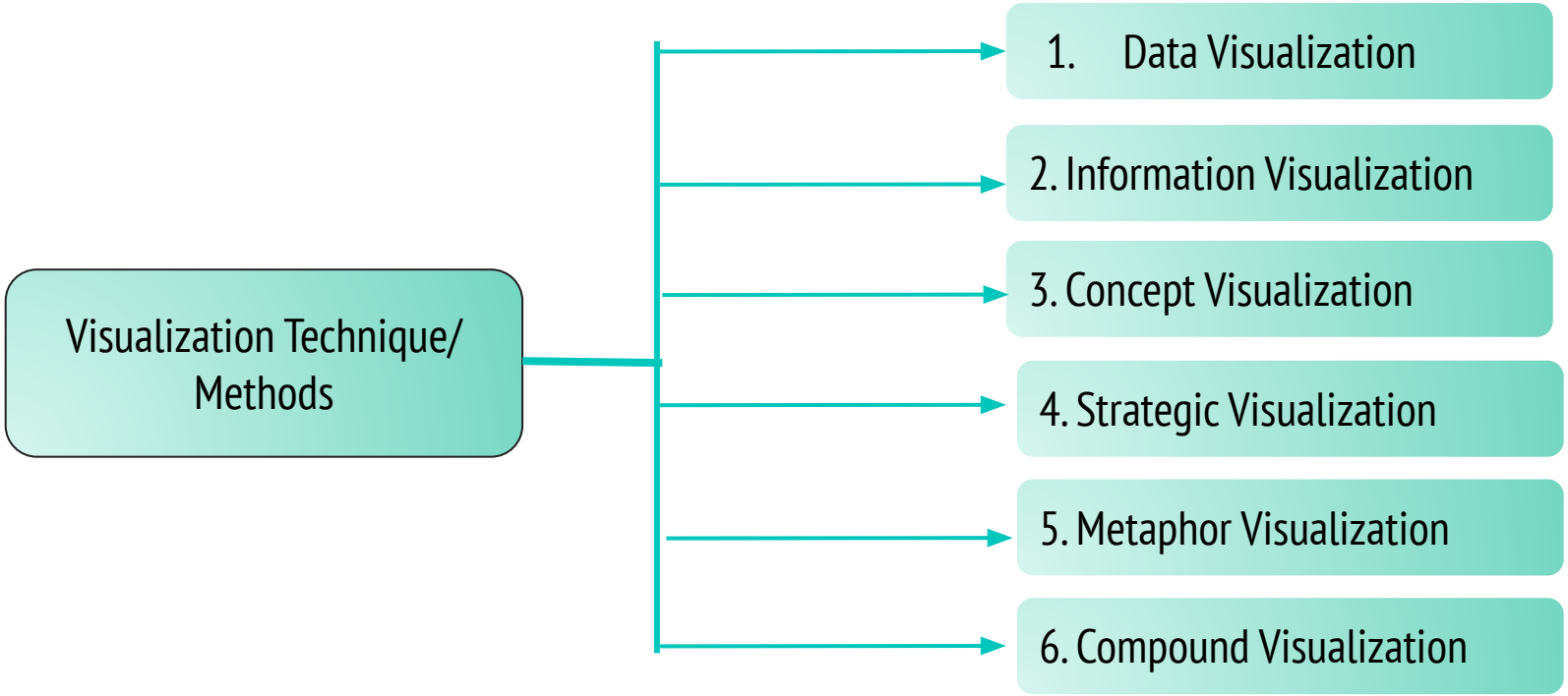
Outline

- Introduction to Data Visualization
- Challenges to Big data visualization
- Types of data visualization
- Data Visualization Techniques
- Tools used in Data Visualization
- Hadoop ecosystem, Map Reduce, Pig, Hive,

Data visualization techniques

Predictive methods	Descriptive methods
Classification  Learns a method for predicting the instance class from pre-labeled (classified) instances	**Clustering**  Finds "natural" grouping of instances given un-labeled data
Regression  An attempt to predict a continuous attribute	**Association Rules**  Method for discovering interesting relations between variables in large DBs

Data visualization techniques



https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.4640&rep=rep1&type=pdf

Big data visualization Challenge

Information visualization

Visually represents quantitative data with or without axes in schematic or diagrammatic forms e.g. Table, Line chart, Pie chart, Histogram, and Scatter plot etc

Understanding the data

An interactive interface of data to increase cognition or perception ability. Transform data into a changeable image, through which users can interact during manipulation, e.g. Data map, Tree map, Clustering, Semantic network, Timeline, and Venn/ Euler diagram etc.

Big data visualization Challenge

Concept visualizations

These methods used to elaborate ideas, plan, concepts, and analyze it easily, e.g. Mindmap, Layer chart, Concentric circle, Decision tree, Pert chart etc.

Strategic visualization

A systematic approach in which an organization visually represent it strategies of development, formulation, communication, implementation, and some time its analysis, e.g. Organizational chart, Strategy map, Failure tree, and Portfolio diagram etc

Big data visualization Challenge

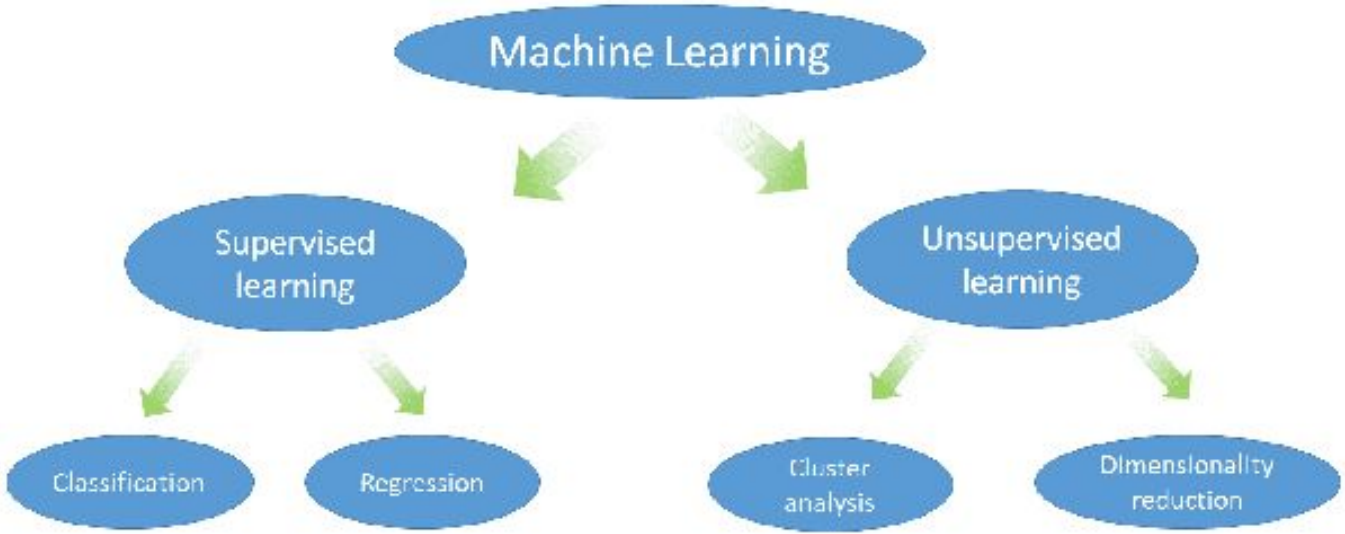
Metaphor visualization

It organizes and structure information graphically. They convey insight of information through key characteristics of metaphor that is employed, e.g. Metro map, Story template, Funnel, and Tree etc.

Compound visualization

The complementary use of different graphic representation formats in one single schema or frame, e.g. Cartoon, Rich picture, Knowledge map, and Learning map etc

Data visualization techniques



Outline

- Introduction to Data Visualization
- Challenges to Big data visualization
- Types of data visualization
- Data Visualization Techniques
- Tools used in Data Visualization
- Hadoop ecosystem, Map Reduce, Pig, Hive,

data visualization Tools

Infogram

Sisense

Adaptive Insights

Dundas BI

FineReport

Google Charts

WhataGraph

Power BI

Grafana

Tableau

What is Tableau

Tableau

is

Software Company

produce

Interactive Data Visualization Products

Focus on

Business Products

Founded

in

What is Tableau

Tableau

Allow

all

Customer

to

Spend more time

on

Data analysis

&

less on data wrangling

#SNJBCOE #ComputerEngineering

What is Google Charts

Google chart tools are powerful, simple to use, and free.

We can use interactive charts and data tools.

Content Beyond Syllabus- Dashboard

Decision - making process

Data Accuracy

gain information quickly and accurately

Dashboard

A good Presentation

of

Information

in

Visual form

that enables

Decision making easily

Dashboard

Large sum of available data

Can be

Challenge

For

Organization

to

Sort & Process

Dashboard

Large amount of data

Difficult to understand

Dashboard come into picture

Dashboard

Dashboard Using Google Sheet | Charts

Dashboard

Demo | Hands on

ALGO

Outline

- Introduction to Data Visualization
- Challenges to Big data visualization
- Types of data visualization
- Data Visualization Techniques
- Tools used in Data Visualization
- Hadoop ecosystem, Map Reduce, Pig, Hive

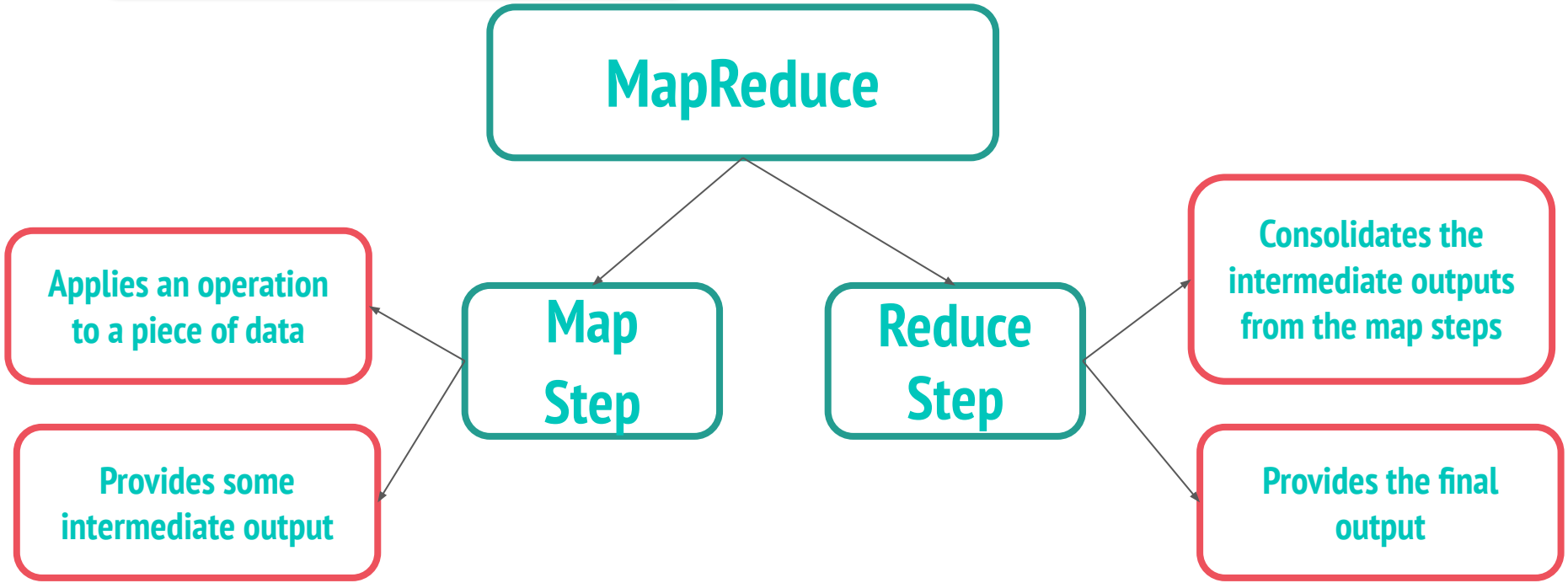
Hadoop ecosystem,Map Reduce, Pig, Hive

Map Reduce

- The MapReduce paradigm offers the means
  - **to break a large task into smaller tasks,**
  - **run tasks in parallel, and**
  - **consolidate the outputs of the individual tasks into the final output.**
- **Apache Hadoop includes a software implementation of MapReduce.**

Hadoop ecosystem,Map Reduce, Pig, Hive

Map Reduce



Hadoop ecosystem,Map Reduce, Pig, Hive

Map Reduce

Map Step

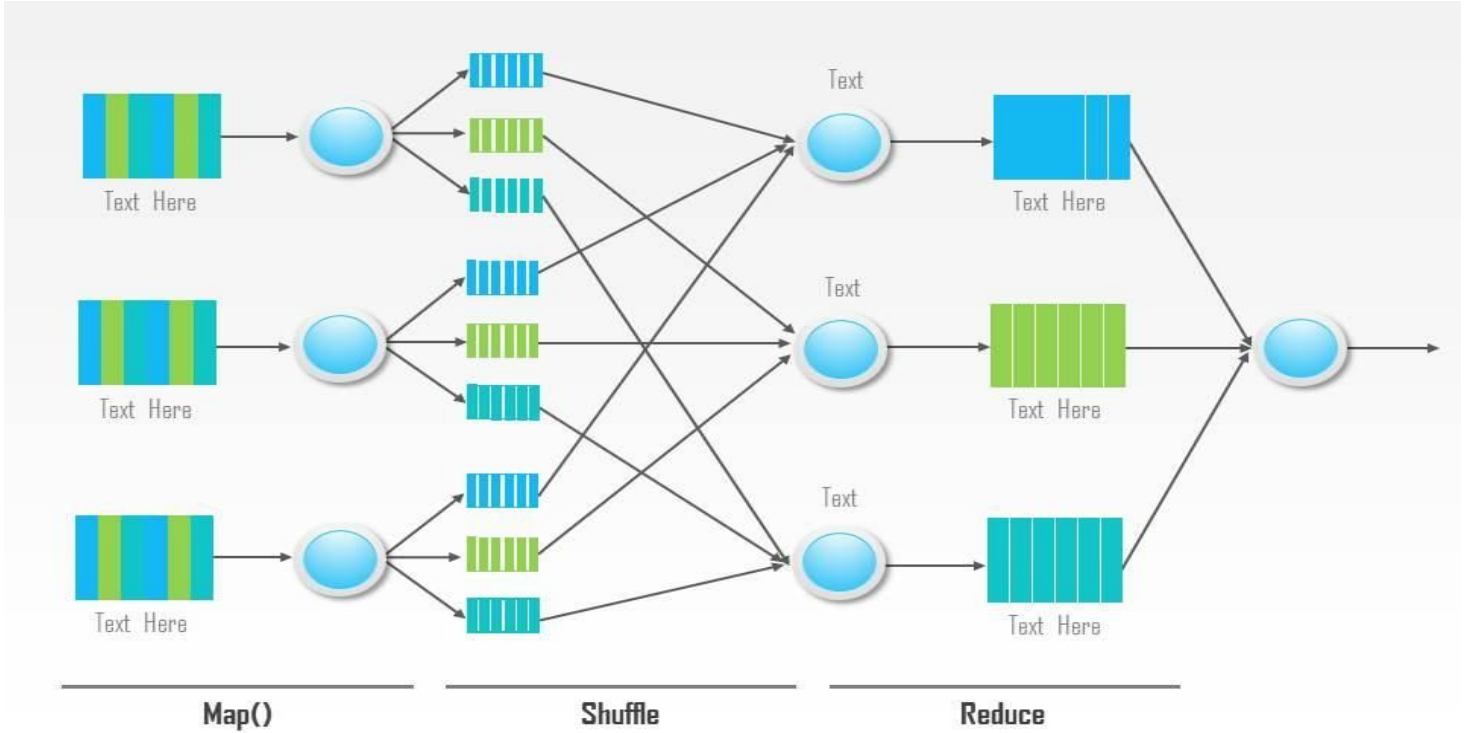
Reduce Step

Uses <Key, Value> Pair

As
I/P & O/P

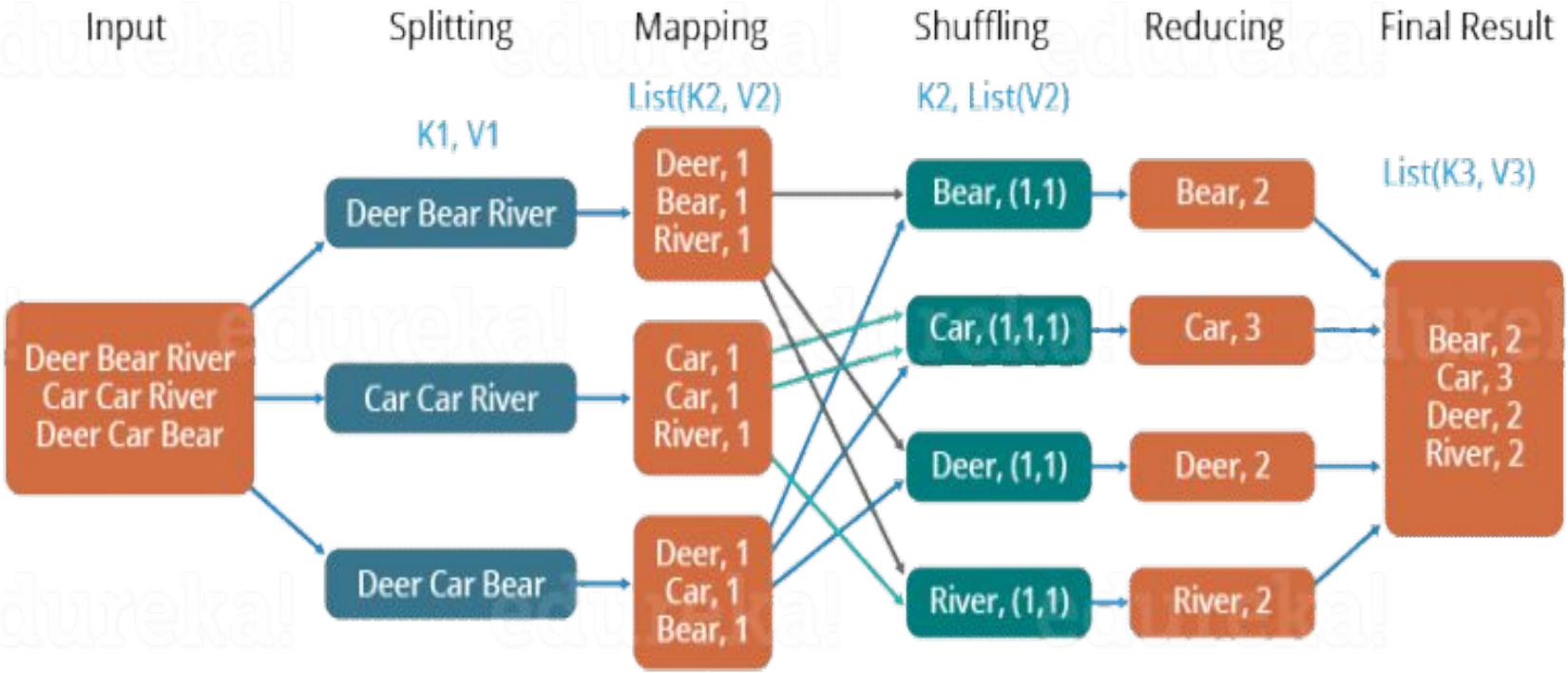
Hadoop ecosystem,Map Reduce, Pig, Hive

Map Reduce



Hadoop ecosystem,Map Reduce, Pig, Hive

Map Reduce



Hadoop ecosystem,Map Reduce, Pig, Hive

Map Reduce

- Hadoop is an o**pen source framework**, from the Apache foundation,
- Capable of processing large amounts of **heterogeneous data sets** in a distributed fashion across clusters of **commodity computers and hardware using a simplified programming model.**
- Hadoop provides a **reliable shared storage and analysis system**.

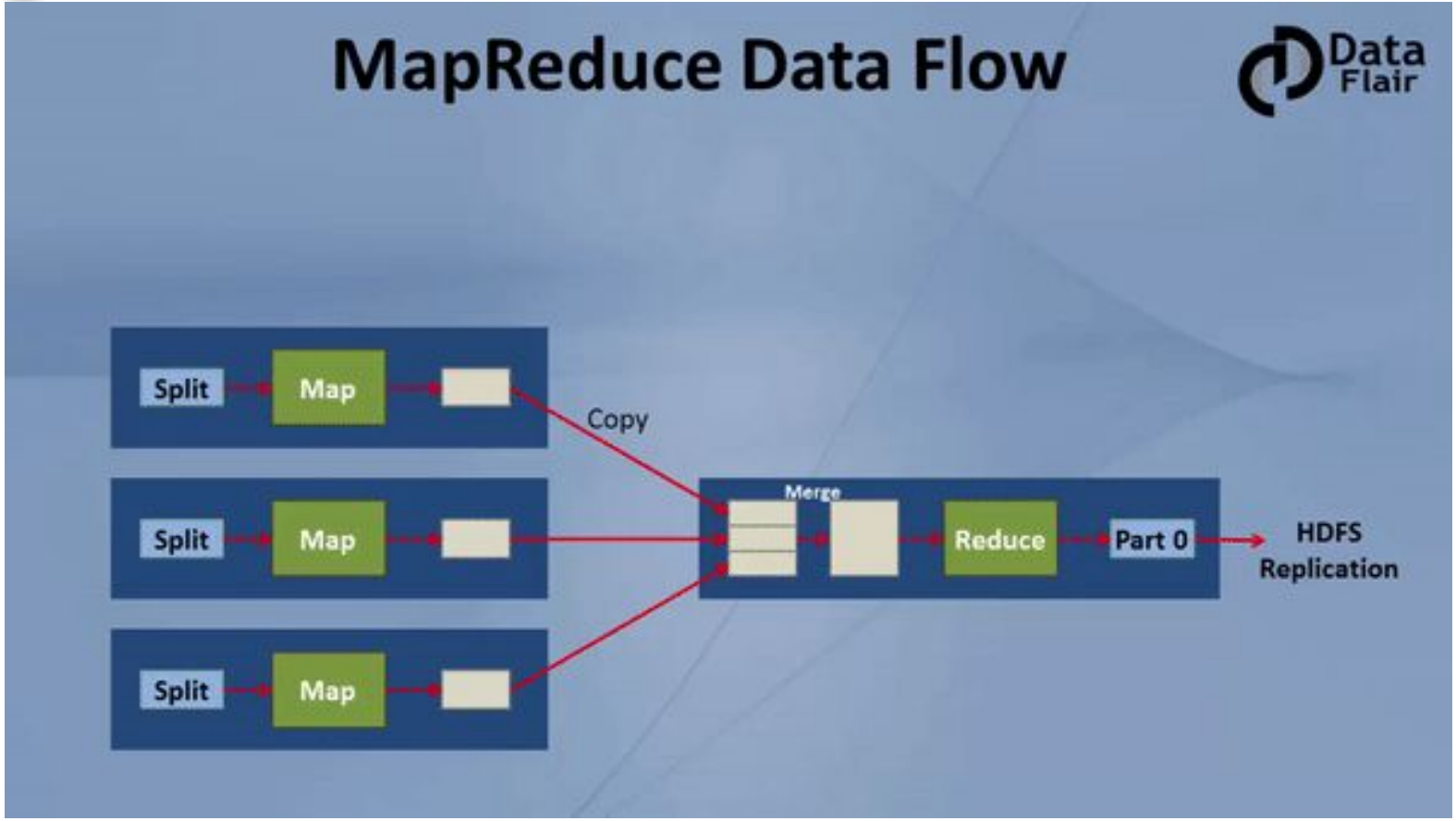
Hadoop ecosystem,Map Reduce, Pig, Hive

Hadoop Distributed File System HDFS

- HDFS Architecture provides **a complete overview of HDFS Namenode and data nodes and their functionality**.
- **Namenode** will **store metadata** and **data nodes** will **store actual data**.
- The **client will interact with the Namenode** in the cluster to perform the task.
- **Data nodes** will keep sending a **heartbeat to Namenode** to indicate that it's alive.

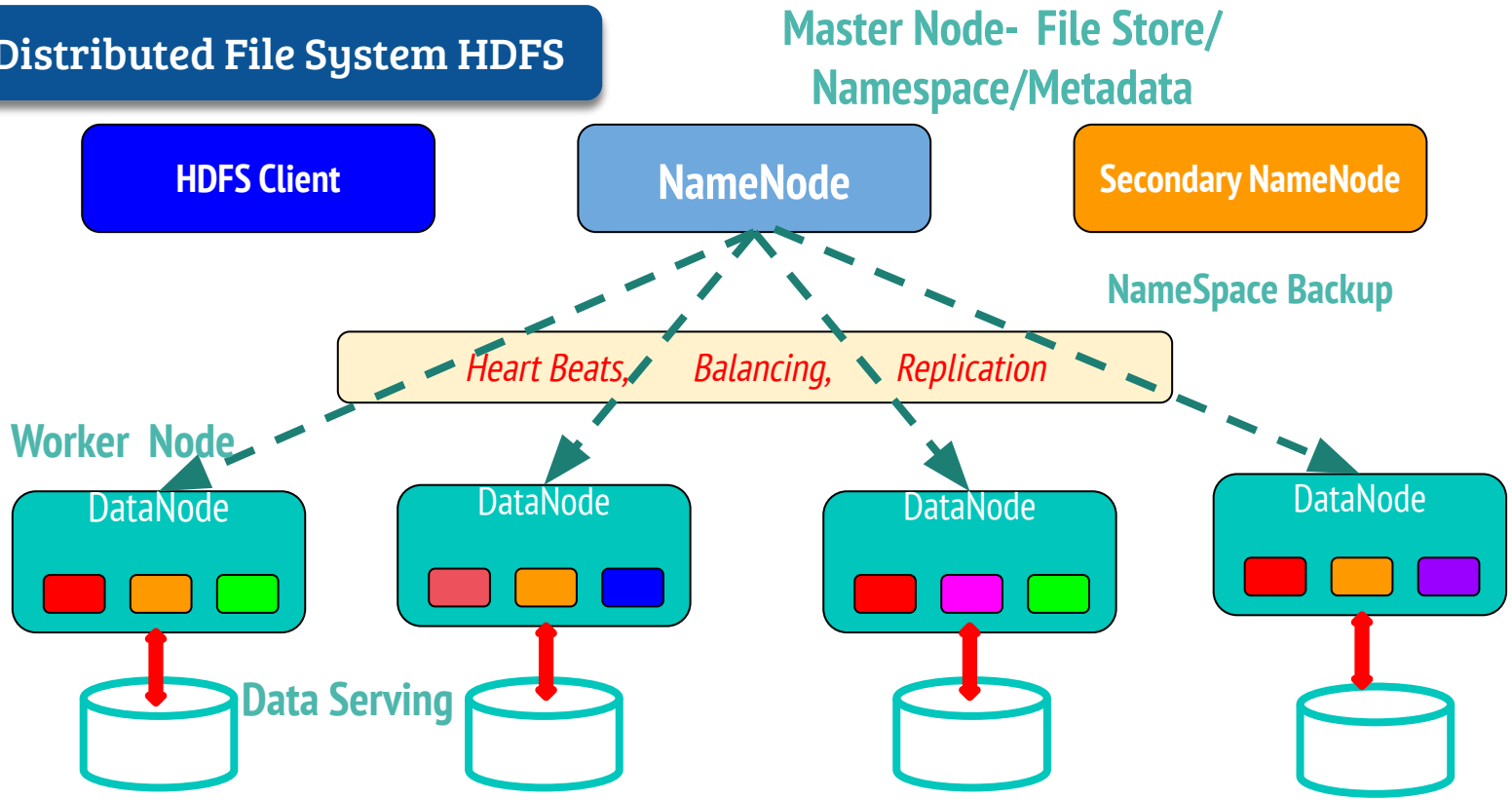
https://www.cloudduggu.com/hadoop/hdfs/

Hadoop ecosystem,Map Reduce, Pig, Hive



Hadoop ecosystem,Map Reduce, Pig, Hive

Hadoop Distributed File System HDFS



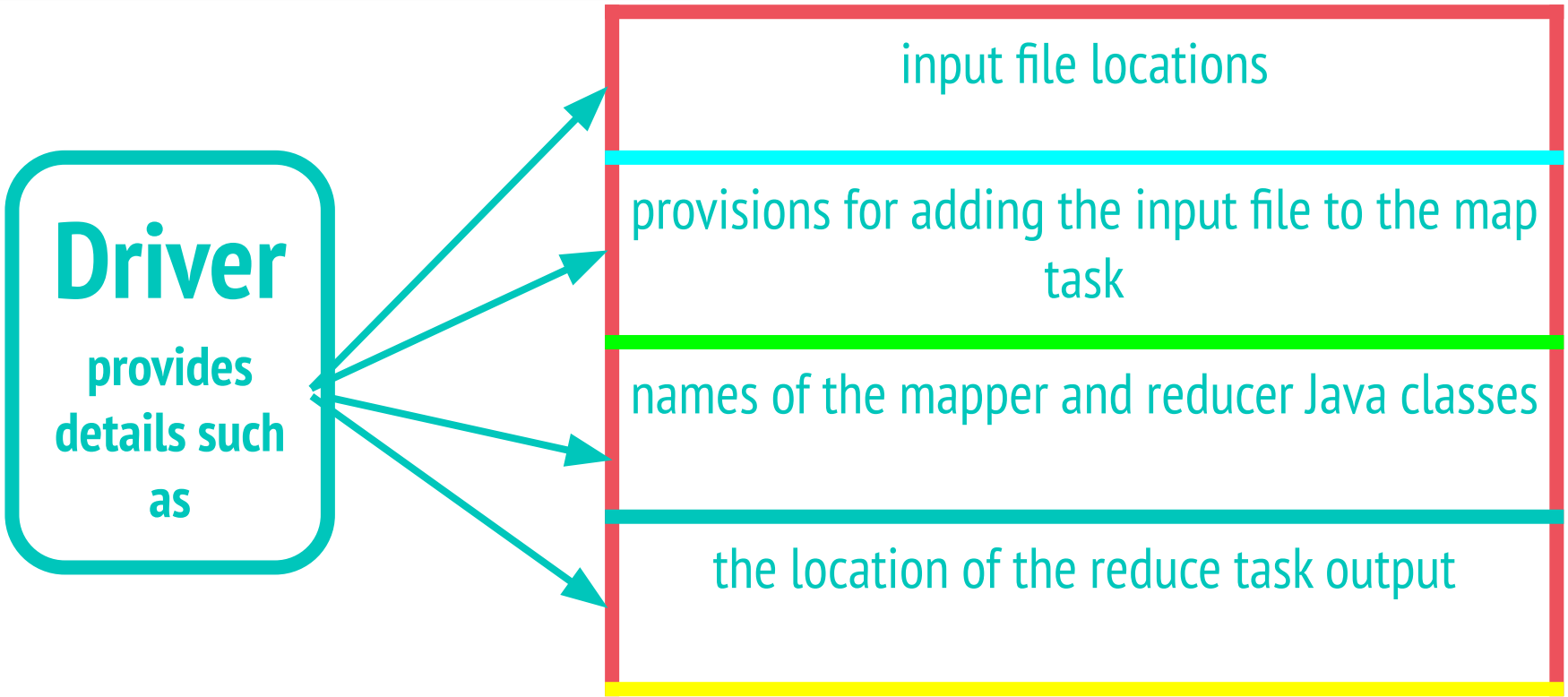
Hadoop Distributed File System HDFS

How a MapReduce job is run in Hadoop?

A typical MapReduce program in Java consists of three classes

- Driver
- Mapper
- Reducer

Hadoop Distributed File System HDFS

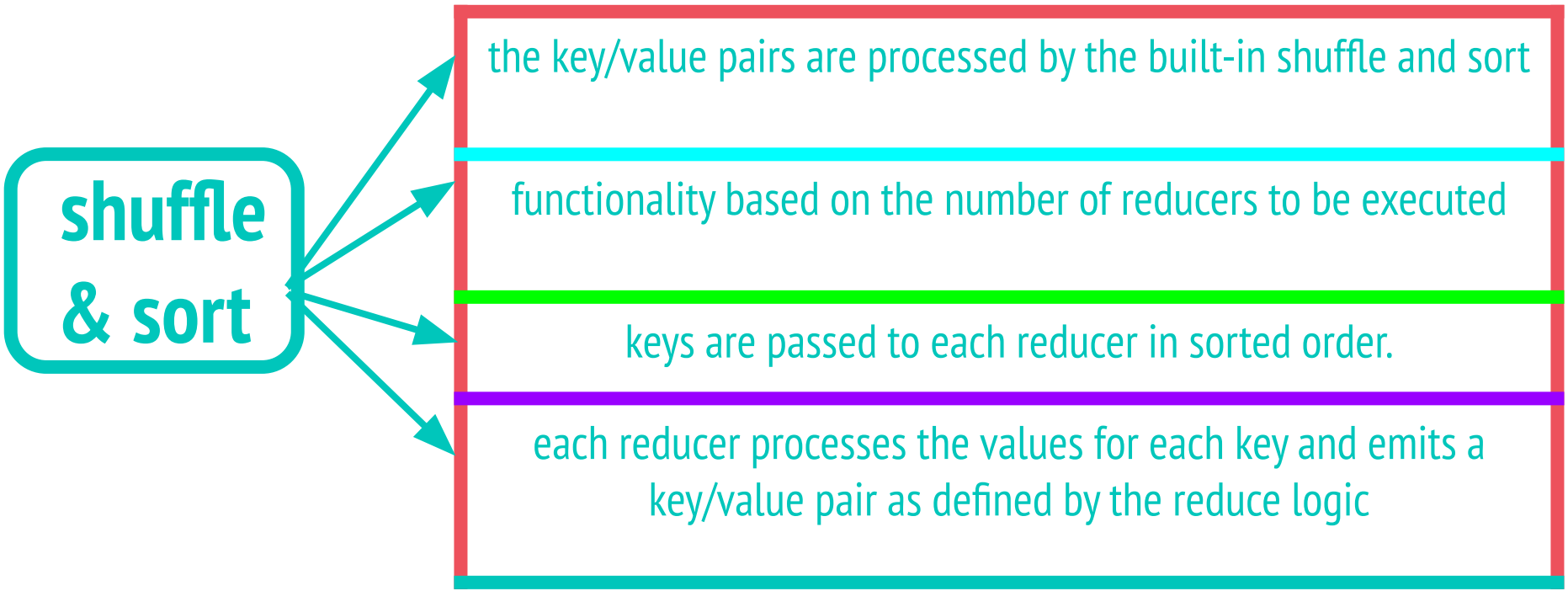


Hadoop Distributed File System HDFS

mapper

- provides the logic to be processed on each data block corresponding to the specified input files in the driver code
- map task is instantiated on a worker node where a data block resides.
- The key/value pairs are stored temporarily in the worker node's memory

Hadoop Distributed File System HDFS



Hadoop Distributed File System HDFS

Several Hadoop features provide additional functionality to a MapReduce job.

Combiner

Apply between map task & shuffle & sort

minimizes the amount of intermediate map output

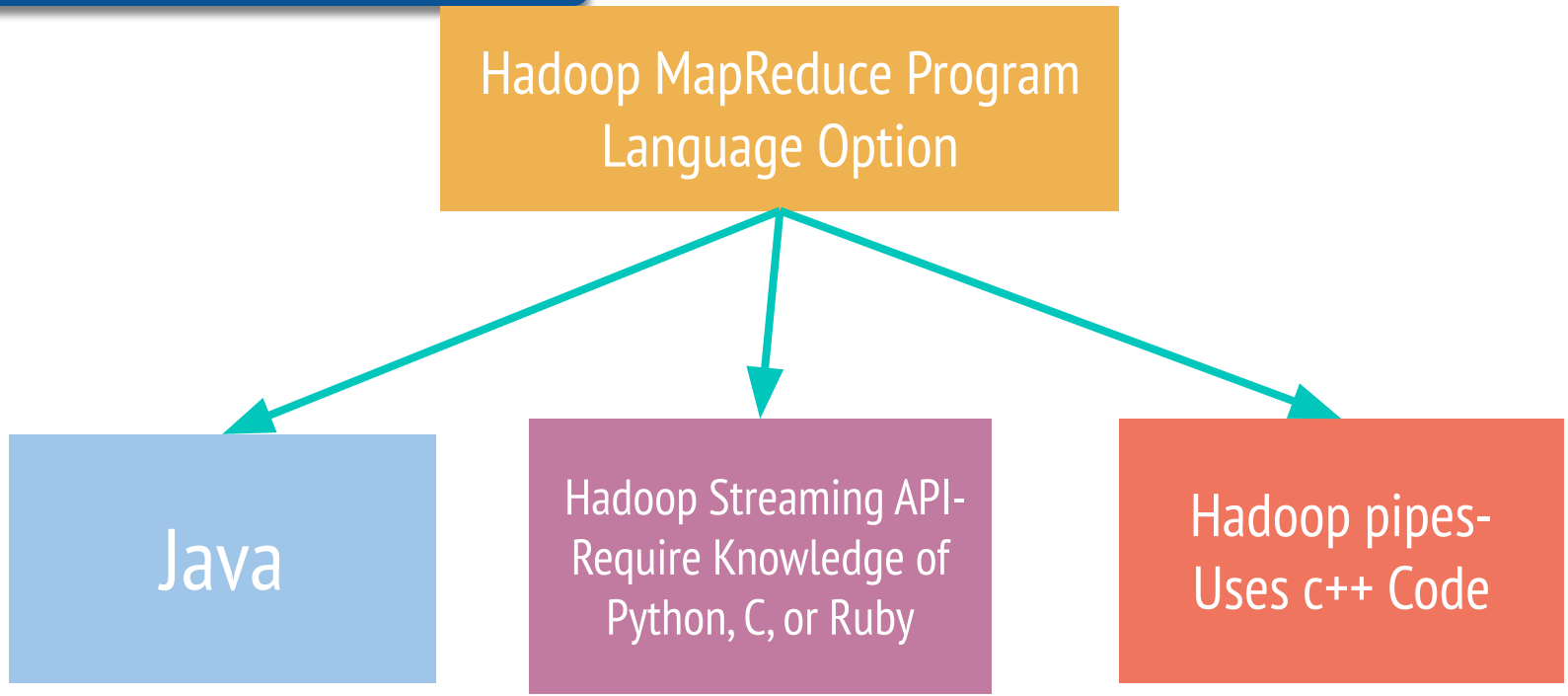
partitioner

separate the output into separate files for for subsequent analysis.

ensure that the workload is evenly distributed across the reducers

Hadoop ecosystem,Map Reduce, Pig, Hive

Hadoop Distributed File System HDFS



Hadoop ecosystem,Map Reduce, Pig, Hive

The Hadoop Ecosystem

Hadoop User Experience (HUE)

HUE

Data Exchange

Sqoop

Zoo Keeper

Coordination

Pig

Scripting

Hive

SQL

HIVE

Mahout

ML

mahout

Oozie

Workflow

OOZIE

APACHE HBASE

Hbase

Columnar data store

Flume

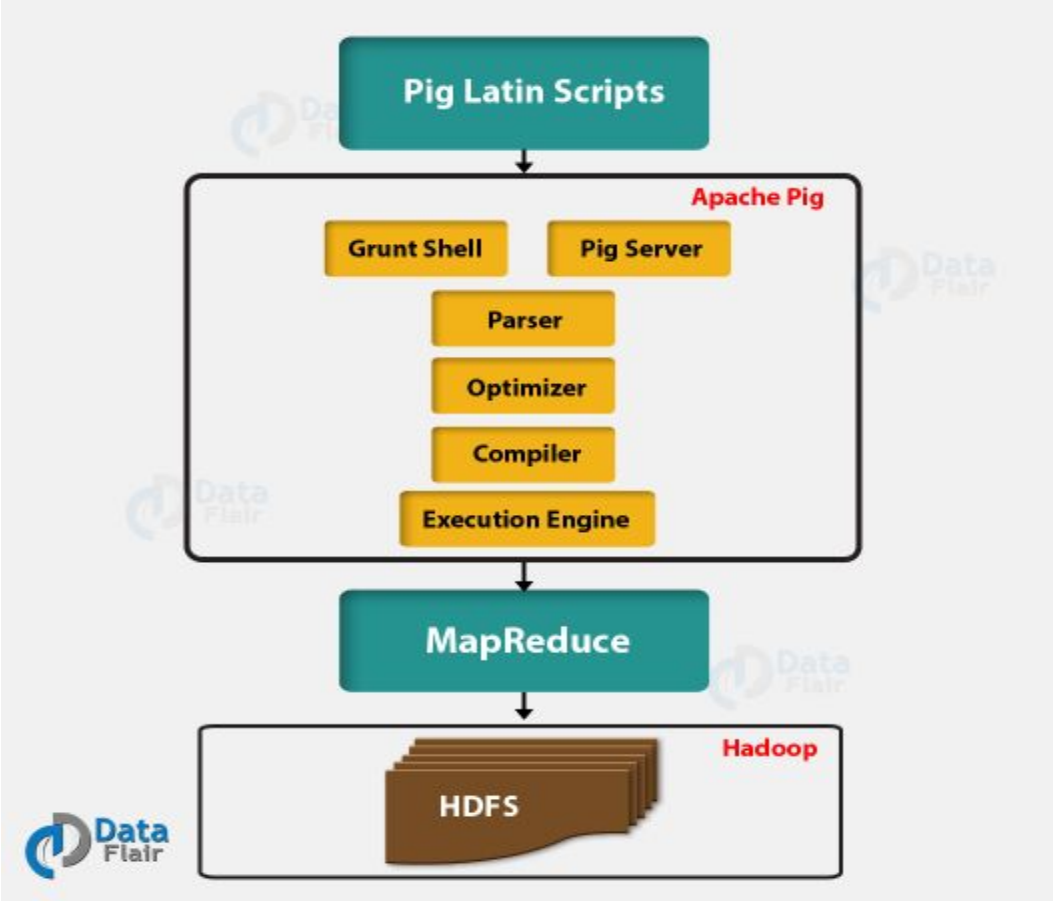
Log Control

YARN/Map Reduce V2

Hadoop Distributed File System

Hadoop ecosystem,Map Reduce, Pig, Hive

The Hadoop Ecosystem - Pig



The Hadoop Ecosystem - Pig

- Apache Pig consists of a
  - data flow language,
  - Pig Latin, and
  - environment to execute the Pig code.
- The main benefit of using Pig is to utilize the power of MapReduce in a distributed system,

  while simplifying the tasks of developing and executing a MapReduce job.

The Hadoop Ecosystem - Pig

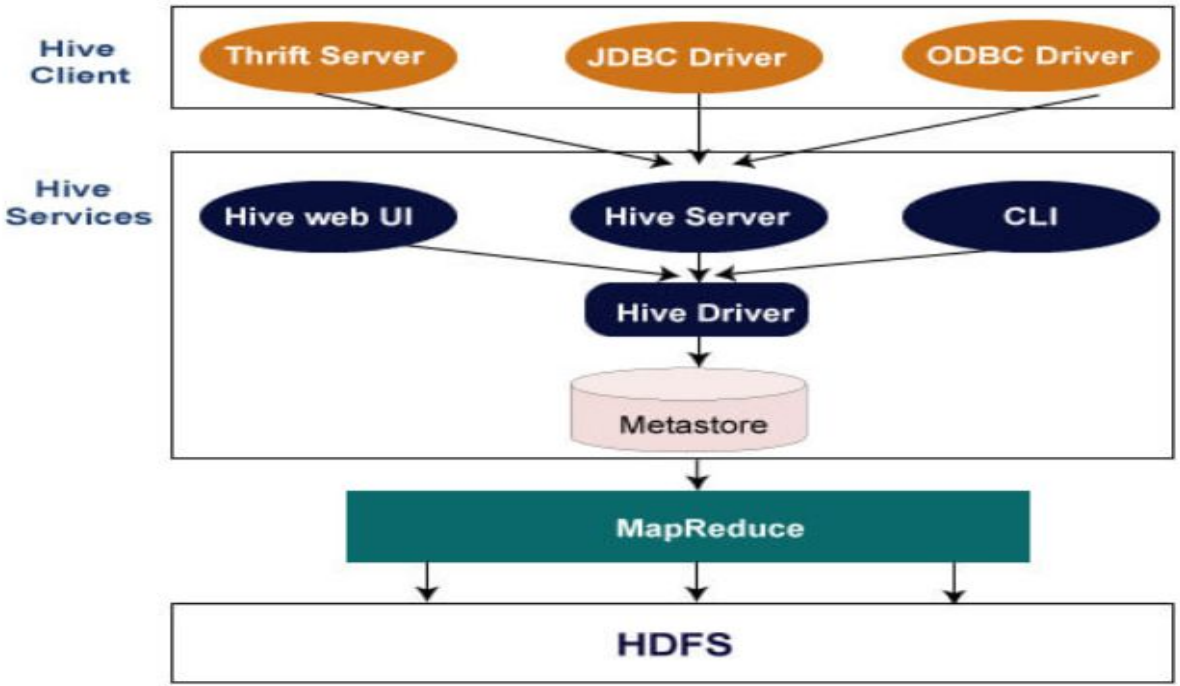
- Pig include entering the Pig execution environment by typing pig at the command prompt and then entering a sequence of Pig instruction lines at the grunt prompt.
- Example :

```
$ pig grunt> records = LOAD '/user/customer.txt' AS (cust_id:INT, first_name:CHARARRAY,
last_name:CHARARRAY, email_address:CHARARRAY);
grunt> filtered_records = FILTER records BY email_address matches '.*@isp.com';
grunt> STORE filtered_records INTO '/user/isp_customers';
grunt> quit
```

The Hadoop Ecosystem - Pig Builtin Functions

- Eval
- Load/Store
- Math
- String
- DateTime

The Hadoop Ecosystem - HIVE



Hadoop ecosystem,Map Reduce, Pig, Hive

The Hadoop Ecosystem - HIVE

- Apache Hive enables users **to process data without explicitly writing MapReduce code**.
- Hive language, H**iveQL (Hive Query Language), resembles Structured Query Language (SQL)**
- A Hive table **structure consists of rows and columns.**
- The rows typically correspond to some record, transaction, or particular entity (for example, customer) detail.
- The values of the corresponding columns represent the various attributes or characteristics for each row.
- Additionally, a user may consider using Hive if the user has experience with SQL and the data is already in HDFS.
- **Hive is not intended for real-time querying**

The Hadoop Ecosystem - HIVE

- When to use Hive?
  - Data easily fits into a table structure.
  - Data is already in HDFS. (Note: Non-HDFS files can be loaded into a Hive table.)
  - Developers are comfortable with SQL programming and queries.
  - There is a desire to partition datasets based on time. (For example, daily updates are added to the Hive table.)
  - Batch processing is acceptable.

The Hadoop Ecosystem - HIVE

- To start hive simply type hive on command prompt.
  $ hive
- hive>
  From this environment, a user can define new tables, query them, or summarize their contents.
- **hive> create table customer (cust_id bigint, first_name string, last_name string, email_address string) row format delimited fields terminated by '\t';**
  ('\t')-delimited HDFS file

The Hadoop Ecosystem - HIVE

- To load the customer table with the contents of HDFS file, customer.txt

  hive> load data inpath '/user/customer.txt' into table customer;

- HiveQL query is executed to count the number of records in the newly created table, customer.

  hive> select count(*) from customer;

The Hadoop Ecosystem - HIVE use cases

- Exploratory or ad-hoc analysis of HDFS data: **Data can be queried, transformed, and exported to analytical tools, such as R.**
- Extracts or data feeds to r**eporting systems, dashboards, or data repositories such as HBase:** Hive queries can be scheduled to provide such periodic feeds.
- Combining external structured data to data already residing in HDFS: **Hadoop is excellent for processing unstructured data, but often there is structured data residing in an RDBMS**, such as Oracle or SQL Server, t**hat needs to be joined with the data residing in HDFS.**
- The data from an **RDBMS can be periodically added to Hive tables for querying with existing data in HDFS.**

Outline

- Introduction to Data Visualization
- Challenges to Big data visualization
- Types of data visualization
- Data Visualization Techniques
- Tools used in Data Visualization
- Hadoop ecosystem, Map Reduce, Pig, Hive

Thank You