UNIT V

5 Big Data Vizualization

Syllabus

Introduction to Data visualization, *Challenges to Big data visualization, Conventional data visualization tools, Techniques for visual data representations, Types of data visualization, Visualizing Big Data, Tools used in data visualization, Propriety Data Visualization tools, Open - source data visualization tools,* ***Case Study :*** *Analysis of a business problem of Zomato using visualization, Analytical techniques used in Big data visualization, Data Visualization using Tableau Introduction to : Candela, D3.js, Google Chart API.*

Contents

5.1 Introduction to Data Visualization

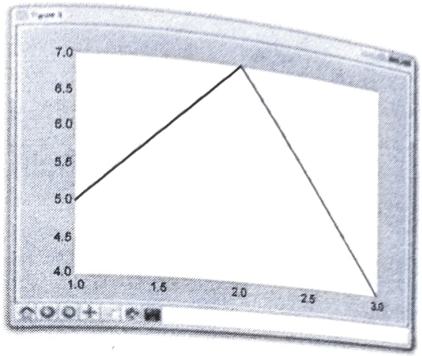
SPPU : May-18, Dec.-18

- Data visualization is the presentation of quantitative information in a graphical form. In other words, data visualizations turn large and small datasets into visuals that are easier for the human brain to understand and process.
- Good data visualizations are created when communication, data science and design collide. Data visualizations done right offer key insights into complicated datasets in ways that are meaningful and intuitive.
- Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers and new insights about the information represented in the data.
- In order to craft a good data visualization, you need to start with clean data that is well sourced and complete. Once your data is ready to visualize, you need to pick the right chart. This can be tricky, but there are many resources available to help you choose the right type of chart for your data.
- A graph is simply a visual representation of numeric data. MatplotLib supports a large number of graph and chart types.
- MatplotLib is popular Python package used to build plots. MatplotLib can also be used to make 3D plots plots and animations.
- Line plots can be created in Python with MatplotLib's pyplot library. To build a line plot, first import MatplotLib. It is a standard convention to import MatplotLib's pyplot library as plt.
- To define a plot, you need some values, the matplotLib. pyplot module and an idea of what you want to display.

```
import MatplotLib.pyplot as plt
plt.plot([1, 2,3], [5,7,4])
plt.show( )
```

- The plt.plot will "draw" this plot in the background, but we need to bring it to the screen when we're ready, after graphing everything we intend to.
- plt.show() : With that, the graph should pop up. If not, sometimes it can pop under or you may have gotten an error.

- Your graph should look like :



- This window is matplotlib window, which allows us to see our graph, as well as interact with it and navigate it.
- **Three principal drivers of this technology :**
  1. **Visual :** Data are represented in a graphic/visual format.
  2. **Insight :** Data visualization, helps manager to understand data immediately and provides advice and suggestions on the possible actions to take.
  3. **Sharing :** Advice and suggestions on the possible actions can be easily shared across the company which will lead to a consequent.
- To fully document graph, user usually have to resort to labels, annotations and legends. Each of these elements has a different purpose, as follows :
  1. **Label :** Make it easy for the viewer to know the name or kind of data illustrated.
  2. **Annotation :** Help extend the viewer's knowledge of the data, rather than simply identify it.
  3. **Legend :** Provides cues to make identification of the data group easier.
- Benefits of data visualization :
  1. Constructing ways in absorbing information. Data visualization enables users to receive vast amounts of information regarding operational and business conditions.
  2. Visualize relationships and patterns in businesses.
  3. More collaboration and sharing of information
  4. More self-service functions for the end users.

- Big data visualization is important because :
  1. It provides clear knowledge about patterns of data.
  2. Detects hidden structures in data.
  3. Identify areas that need to be improved.
  4. It helps us to understand which products to place where.
  5. Clarify factors which influence human behaviour.

5.1.1 Challenges to Big Data Visualization

- Big data analytics plays a key role through reducing the data size and complexity in big data applications. Visualization is an important approach to helping big data get a complete view of data and discover data values.
- Scalability and dynamics are two major challenges in visual analytics.
- Volume : The methods are developed to work with an immense number of datasets and enable to derive meaning from large volumes of data.
- Variety : The methods are developed to combine as many data sources as needed.
- Velocity : With the methods, businesses can replace batch processing with real - time stream processing.
- Value : The methods not only enable users to create attractive info graphics and heat maps, but also create business value by gaining insights from big data.
- Visualization of big data with diversity and heterogeneity (structured, semi - structured and unstructured) is a big problem. Speed is the desired factor for big data analysis.
- There are also following problems for big data visualization :
  1. **Visual noise** : Most of the objects in the dataset are too relative to each other. Users cannot divide them as separate objects on the screen.
  2. **Information loss** : Reduction of visible data sets can be used, but leads to information loss.
  3. **Large image perception :** Data visualization methods are not only limited by aspect ratio and resolution of device, but also by physical perception limits.
  4. **High rate of image change** : Users observe data and cannot react to the number of data changes or its intensity on display.
  5. **High performance requirements** : It can be hardly noticed in static visualization because of lower visualization speed requirements - high performance requirements.

- Following problems are encountered during visualizing big data :

a) Scalability and dynamics are two major challenges in visual analytics.

b) Visualization of big data with diversity and heterogeneity (structured, semi-structured and unstructured) is a big problem. Speed is the desired factor for big data analysis. Designing a new visualization tool with efficient indexing is not easy in big data.

c) Cloud computing and advanced graphical user interface can be merged with the big data for the better management of big data scalability.

d) Visualization systems must contend with unstructured data forms such as graphs, tables, text, trees and other metadata. Big data often has unstructured formats.

e) 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 situ manner.

f) Visual noise : It is messy to represent the whole array of data being studied on the screen. This problem comes when most of the objects share too much of relativeness, and that's the only reason why viewers cannot view them as separate objects.

g) High - performance requirements : The graphical analysis does not stop at just static picture representation, so the above issues turn out to be more critical in unique perception.

h) Large image perception : This problem occurs due to the human perceptions which different for different entities. In spite of the higher level of graphical data visualizations, it has its own limitations when compared with the table representation.

i) High rate of image change : This issue turns into the hugest in checking assignments, when a man who analyses the information just can't respond to the quantity of information changes or its power on display.

Review Questions

1. *What are the major challenges in visualizing the big data and how to overcome these challenges.* **SPPU : May-18 (End Sem), Marks 8**
2. *Explain various challenges in big data visualization and explain the mechanism to overcome the challenges.* **SPPU : Dec.-18 (End Sem), Marks 8**
3. *What is the need of data visualization ? Also explain advantages of data visualization.* **SPPU : Dec.-18 (End Sem), Marks 8**

5.2 Types of Data Visualization

SPPU : Dec.-19

- Various types of data visualization are as follows :
  1. Multidimensional : 2D Area
  2. Temporal
  3. Hierarchical
  4. Network

Sr. No.	Types	Descriptions
1.	Multidimensional : 2D Area	1. **Cartogram :** It distorts map space to express information such as travel time or population of the alternate variable. It mainly consists of two main types : Area based and distance - based cartograms. 2. **Choropleth :** It is used to represent the statistical measurement such as population density rate or website visitors count per city. 3. **Dot distortion map :** It uses a dot symbol to represent a feature on the map, depending on the visual scatter for displaying spatial patterns.
2.	Temporal	1. **Pie chart :** The circle is divided into sectors to represent numeric proportions. The length of the arc and angle length of the sector is proportional to the particular quantity it represents. 2. **Histogram :** In a histogram, the data are grouped into ranges (e.g. 10 - 19, 20 - 29) and then plotted as connected bars. Each bar represents a range of data. The width of each bar is proportional to the width of each category and the height is proportional to the frequency or percentage of that category. 3. **Scatter plot :** It displays collection of all the points for the set of data limited only for two values.
3.	Hierarchical	1. **Dendrogram :** It is nothing but a tree diagram used to represent clusters generated by hierarchical clustering. 2. **Ring chart :** It is a multi-level pie chart which is represented by the nested circles. 3. **Tree diagram :** It represents the data or the hierarchy in the graph form, which can be visualized from left to right or top to bottom.
4.	Network	1. **Alluvial diagram :** It is a flow diagram which visualizes over time changes in network structure. 2. **Node link diagram :** In this representation, nodes are visualized as dots whereas links are represented as line segments to display the data connection. 3. **Matrix :** It shows relation between two to four groups of information and gives information regarding the same.

Review Question

1. *Describe different types of data visualization.* SPPU : Dec.-19 (End Sem), Marks 8

5.3 Data Visualization Techniques

SPPU : May-18, Dec.-18, 19

- Whenever collection of data is started and the range of data increases rapidly, an efficient and convenient technique for representing data is needed.
- Higher authorities do not have enough time to go through whole reports regarding the progress of their firm or organization, so it is required for presenting the data in such a manner that enables readers to interpret the important data with minimum effort and time.
- Data visualization techniques are helping user, to avoid overloading the working memory.
- Techniques for data presentation are broadly classified in two ways :
  1. Non graphical techniques : Tabular form, case form
  2. Graphical techniques : Pie chart, bar chart, line graphs, geometrical diagrams.

5.3.1 Line Graph

- It is also called stick graphs. It gives relationships between variables.
- Line graphs are usually used to show time series data - that is how one or more variables vary over a continuous period of time. They can also be used to compare two different variables over time.
- Typical examples of the types of data that can be presented using line graphs are monthly rainfall and annual unemployment rates.
- Line graphs are particularly useful for identifying patterns and trends in the data such as seasonal effects, large changes and turning points. Fig. 5.3.1 show line graph. (Refer Fig. 5.3.1 on next page).
- As well as time series data, line graphs can also be appropriate for displaying data that are measured over other continuous variables such as distance.
- For example, a line graph could be used to show how pollution levels vary with increasing distance from a source, or how the level of a chemical varies with depth of soil.
- In a line graph the x-axis represents the continuous variable (for example year or distance from the initial measurement) whilst the y-axis has a scale and indicates the measurement.

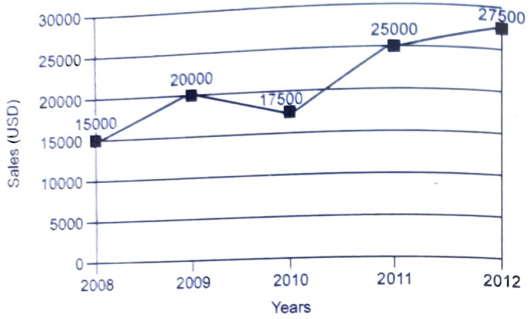


Fig. 5.3.1 Line graph

- Several data series can be plotted on the same line chart and this is particularly useful for analysing and comparing the trends in different datasets.
- Line graph is often used to visualize the rate of change of a quantity. It is more useful when the given data has peaks and valleys. Line graphs are very simple to draw and quite convenient to interpret.

5.3.2 Pie Chart

- A type of graph in which a circle is divided into sectors that each represent a proportion of the whole. Each sector shows the relative size of each value.
- A pie chart displays data, information and statistics in an easy to read "pie slice" format with varying slice sizes telling how much of one data element exists.
- Pie chart is also known as circle graph. The bigger the slice, the more of that particular data was gathered. The main use of a pie chart is to show comparisons. Fig. 5.3.2 shows pie chart.
- Various applications of pie charts can be found in business, school and at home. For business pie charts can be used to show the success or failure of certain products or services.
- At school, pie chart applications include showing how much time is allotted to each subject. At home pie charts can be useful to see expenditure of monthly income in different needs.

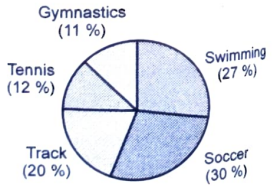


Fig. 5.3.2 Pie chart

- Reading of pie chart is as easy as figuring out which slice of an actual pie is the biggest.

- Pie charts can be drawn using the function pie() in the pyplot module. The below python code example draws a pie chart using the pie() function. By default the pie() fucntion of pyplot arranges the pies or wedges in a pie chart in counter clockwise direction.

```
# import the pyplot library
import matplotlib.pyplot as plotter

# The slice names of a student distribution pie chart
pieLabels = 'Rakshita', 'Ritesh', 'Rupali', 'Rutu', 'Rushi', 'Radhika'

# marks data
marksShare = [59.69, 16, 9.94, 7.79, 5.68, 0.54]
figureObject, axesObject = plotter.subplots()

# Draw the pie chart
axesObject.pie(marksShare, labels=pieLabels, autopct='%1.2f', startangle=90)

# Aspect ratio - equal means pie is a circle
axesObject.axis('equal')
plotter.show()
```

- The essential part of a pie chart is the values. You could create a basic pie chart using just the values as input.
- Limitations of pie chart :
  a) It is difficult to tell the difference between estimates of similar size.
  b) Error bars or confidence limits cannot be shown on pie graph.
  c) Legends and labels on pie graphs are hard to align and read.
  d) The human visual system is more efficient at perceiving and discriminating between lines and line lengths rather than two-dimensional areas and angles.
  e) Pie graphs simply don't work when comparing data.

5.3.3 Venn Diagram

- Venn diagram is a diagram that visually displays all the possible logical relationships between collections of sets. Each set is typically represented with a circle.
- Venn diagram shows the similarities and differences of two or more data sets by using overlapping circles. The overlapping areas show the similarities and the non-overlapping areas show the differences.
- Venn diagrams may also be called primary diagrams, set diagrams, or logic diagrams.

- Venn diagrams can be useful tools for analysis or support the decision-making process. Although Venn diagrams can have unlimited circles (each circle representing a data set), they usually have just two or three overlapping circles.
- By the size of the circle, you can show the importance of an organization or projects. The bigger a circle is, the more important is a project.
- Overlapping circles represent interacting organizations. There is also the possibility of a subset. This means that a small circle is placed within a larger circle.
- The small circle stands for a component in a big organization or project which is symbolized by a big circle.
- Example : There are a total of 55 books, 23 available in hard copy, 20 available on Kindle, and 12 books available in both formats. Fig. 5.3.3 shows venn diagram of this data.

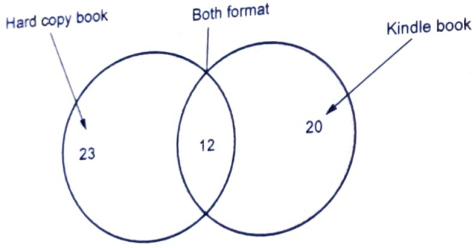


Fig. 5.3.3 Venn diagram

- Venn diagrams can become much more complex with more data sets and are often shaded to help better visualize the relationships between data sets.

Advantages of Venn Diagram

1. Easy way to show similarities and differences amongst systems.
2. Works without much technical equipment.
3. A tool which is easy to understand and to use
4. Clearly orientated towards output
5. To solve complex mathematical problem.

Disadvantages of Venn Diagram

1. Venn diagram is often a snapshot of a group interaction and negotiations.
2. Growing complexity if more than four circles are drawn.
3. If the Venn diagrams are done by groups, the views of weaker actors are likely to be submerged.

5.3.4 Scatter Diagram

- **Scatter diagram** is also called scatter plot, X-Y graph. The scatter plot is the model of data visualization depicting two sets of unconnected dots as parameter values.
- Scatter plots which use horizontal and vertical axes to plot data points and display how much one variable is affected by another. The position of each dot on the horizontal and vertical axis indicates values for an individual data point.
- Fig. 5.3.4 shows scatter plots of two variables.

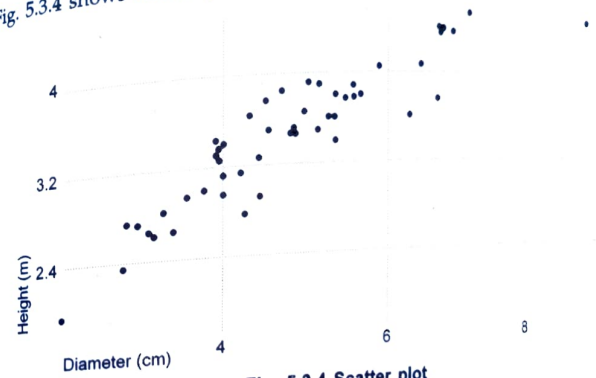


Fig. 5.3.4 Scatter plot

- The example scatter plot above shows the diameters and heights for a sample of fictional trees. Each dot represents a single tree; each point's horizontal position indicates that tree's diameter (in centimeters) and the vertical position indicates that tree's height (in meters).
- From the plot, we can see a generally tight positive correlation between a tree's diameter and its height. We can also observe an outlier point, a tree that has a much larger diameter than the others.
- While working with statistical data it is often observed that there are connections between sets of data.
- A scatter diagram is a tool for analyzing relationships between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis.
- The pattern of their intersecting points can graphically show relationship patterns. Commonly a scatter diagram is used to prove or disprove cause-and-effect relationships.
- Scatter plot's primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of

individual data points, but also patterns when the data are taken as a whole. Identification of correlational relationships are common with scatter plots.

- A scatter plot can also be useful for identifying other patterns in data. We can divide data points into groups based on how closely sets of points cluster together. Scatter plots can also show if there are any unexpected gaps in the data and if there are any outlier points. This can be useful if we want to segment the data into different parts, like in the development of user personas.

Merits :

a) Scatter diagrams are easy to draw.

b) It can be easily understood and interpreted.

c) Shows both positive and negative types of graphical correlation.

Demerits :

a) You cannot use scatter diagrams to show the relation of more than two variables.

b) Interpretation can be subjective.

Review Questions

1. *Explain any two visual data representation techniques with sample data set.* **SPPU : May-18 (End Sem), Marks 8**
2. *Explain different techniques of data visualization in detail.* **SPPU : Dec.-18 (End Sem), Marks 9**
3. *Explain pie chart and scatter plot.* **SPPU : Dec.-19 (End Sem), Marks 8**

5.4 Visualizing Big Data

- Big data visualization is the process of displaying data in charts, graphs, maps and other visual forms.
- There are various analytical techniques used in big data processing in order to extract, collect, store, process and analyze the huge amount of data coming very fast with the different variety.

1. Machine Learning :

- A machine learning algorithm then takes these examples and produces a program that does the job. The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers. If we do it right, the program works for new cases as well as the ones we trained it on.

- Main goal of machine learning is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example." Another goal is to develop computational models of the human learning process and perform computer simulations.
- The goal of machine learning is to build computer systems that can adapt and learn from their experience.
- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instructions. It should carry out to transform the input to output.
- For example, The addition of four numbers is carried out by giving four number as input to the algorithm and output is the sum of all four numbers. For the same task, there may be various algorithms. It is interesting to find the most efficient one, requiring the least number of instructions or memory or both.
- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology is being used by global enterprises to gain a competitive edge.
- Machine learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.
- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behaviour of a system for any set of input values, after an initial training phase.
- In contrast to supervised learning, unsupervised or self-organized learning does not require an external teacher. During the training session, the neural network receives a number of different input patterns, discovers significant features in these patterns and learns how to classify input data into appropriate categories.
- Unsupervised learning algorithms aim to learn rapidly and can be used in real - time. Unsupervised learning is frequently employed for data clustering, feature extraction etc.
- Reinforcement learning : This is an advanced machine learning technique. This is based on probability theory where mapping can be done based on input received and changes based on the environment around it.
- Deep learning : This is also advanced machine learning technique which has multiple processing layers so as to produce non-linear response based on input data. There are so many small processors called as **neuron working** parallel in data processing.

- Predictive analytics : This technique refers to prediction based on past experience and it uses both data mining and machine learning.
- Association rule learning : This is used to identify interesting relations between different attributes from large datasets.

5.5 Tools used in Data Visualization

SPPU : Dec.-19

- Traditional data visualization tools are often inadequate to handle big data. Methods for interactive visualization of big datawere presented.
- First, design space of scalable visual summaries that use data reduction approaches was described to visualize a variety of data types.
- Methods were then developed for interactive querying among binned plots through a combination of multivariate data tiles and parallel query processing.
- Lot of big data visualization tools run on the Hadoop platform. The common modules in Hadoop are : Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop Map Reduce.
- They analyze big data efficiently, but lack adequate visualization. Some software with the functions of visualization and interaction for visualizing data has been developed.

5.5.1 Pentaho

- Pentaho tightly couple's data integration with full business analytics to solve data integration challenges while providing business analytics in a single, seamless platform.
- Pentaho's Java-based data integration engine integrates with the MapRHadoop cache for automatic deployment as a MapReduce task across every data node in a Hadoop cluster, making use of the massively parallel processing and high availability of Hadoop.
- Pentaho's open-source heritage drives our continued innovation in a modern, integrated, embeddable platform built for the future of analytics, including diverse and big data requirements.
- Within a single platform it provides visual tools to extract and prepare our data plus the visualizations and analytics that will change the way we run our business.
- Pentaho's modern, simplified and interactive approach empowers business users to access, discover and blend all types and sizes of data. With a spectrum of increasingly advanced analytics, from basic reports to predictive modeling, users can analyze and visualize data across multiple dimensions, all while minimizing dependence on IT.

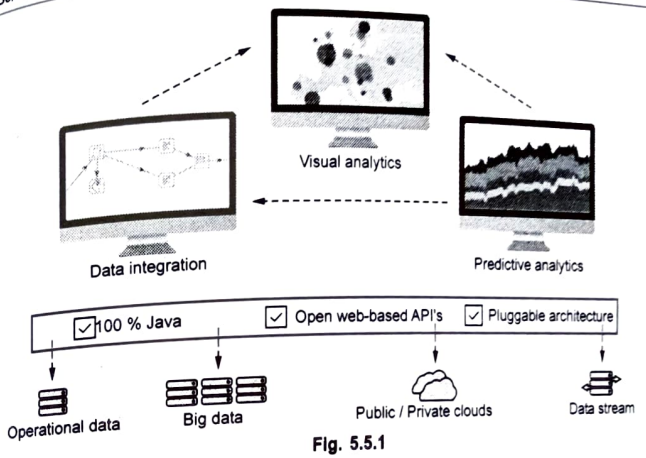


Fig. 5.5.1

- The business analytics platform is a web application that allows users to publish and manage reports within an enterprise business intelligence system.
- The business analytics platform offers many capabilities, including the management and execution of Pentaho reports. By combining Pentaho reporting and Pentaho's business analytics platform, information technologists may utilize Pentaho reporting in their environment without writing any code.
- In addition to the publishing and execution of reports, the open source business analytics platform allows for scheduling, background execution, security and much more.

Advantages

1. Pentaho is an intuitive platform, where IT as well as business people can access and visualize data easily.
2. Easy access to data from diverse sources ranging from Excel to Hadoop.
3. Reporting is fast due to in-memory caching techniques.
4. Detailed visualisation and easy to understand infographics, with drilling and filters available. Seamless integration with third party applications, such as Google Maps.
5. The devices supported covers almost every platform : Android, iPhone, iPad, Mac, Web-based, Windows.

Disadvantages

1. All the products in Pentaho suite are inconsistent in the manner in which they work.
2. The metadata layer is cumbersome to use and understand. The documentation also is of little help at times.
3. There is no system of perpetual licensing. The usage rights have to be bought every year, at the same price.
4. Advanced analytics and corresponding data visualisation need more improvement, when compared with the same in Tableau.

5.5.2 Datameer

- Datameer's flipside provides simple, highly accessible, visual data profiling that lets users easily spot outliers in data, quickly and early in the analytics process. Datameer runs natively on Hadoop.
- Datameer, an end-to-end big data analytics platform, is built on Apache Hadoop to perform integration, analysis and visualization of massive volumes of both structured and unstructured data. It can be rapidly integrated with any data sources such as new and existing data sources to deliver an easy-to-use, cost-effective and sophisticated solution for big data analytics.
- It simplifies data extraction, data transformation, data loading and real-time data retrieval. It helps gain actionable insights from complex organizational data through data preparation and analytics.
- Fig. 5.5.2 shows all Datameer functionality occurs across three major components.

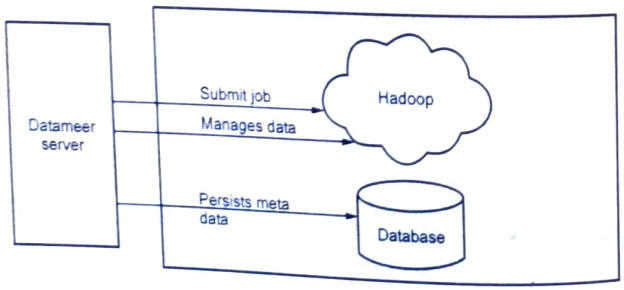


Fig. 5.5.2 Datameer functionality

- The Datameer server : Server is also called conductor. This server orchestrates all work and manages the configuration of all jobs performed on the Hadoop cluster.

It also hosts the web app that lets users interact via the software's web UI. All processing done during the design of a workbook in real time on the Datameer server. Datameer provides real-time feedback during the design phase using intelligent previews generated by our smart sampling technology.

- Database for metadata storage : Datameer uses a database to store all metadata.
- Hadoop cluster : The Hadoop cluster provides persistent storage for all data, pre-views and other job artifacts, as well as a big data processing framework for executing long-running operations.
- Fundamental to the design of Datameer software is the fact that all resource - intensive processes are submitted to Hadoop clusters. This approach allows Datameer to scale up and scale out easily by distributing work across the entire Hadoop cluster.

5.5.3 JasperReport

- JasperReports is a powerful open source reporting package, but generating reports with data from multiple sources is hard and often impossible without the enterprise version.
- Fig. 5.5.3 shows JasperReport.

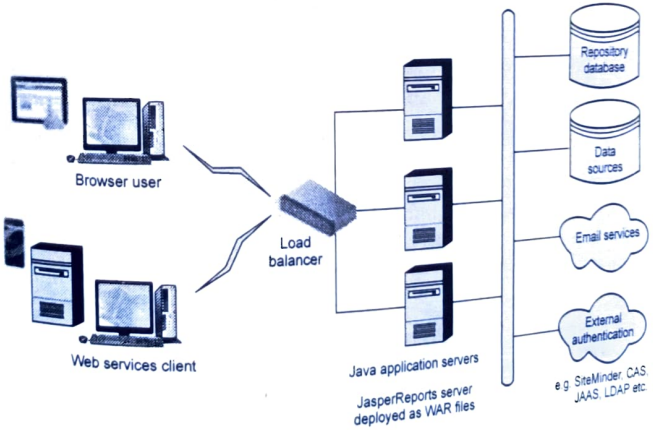


Fig. 5.5.3 JasperReport

- JasperReports is an open source java reporting engine. JasperReports is a Java class library and it is meant for those Java developers who need to add reporting capabilities to their applications.

- The main purpose of JasperReports is to create page oriented, ready to print documents in a simple and flexible manner.
- JasperReports Server is a stand-alone and embeddable reporting server.
- It provides reporting and analytics that can be embedded into a web or mobile application as well as operate as a central information hub for the enterprise by delivering mission critical information on a real-time or scheduled basis to the browser, mobile device, or email inbox in a variety of file formats.
- JasperReports Server is optimized to share, secure and centrally manage your Jaspersoft reports and analytic views.
- Data sources are structured data containers. While generating the report, JasperReports engine obtains data from the datasources. Data can be obtained from the databases, XML files, arrays of objects and collection of objects.
- JasperReports has a feature <style> which helps to control text properties in a report template. This element is a collection of style settings declared at the report level.
- Properties like foreground color, background color, whether the font is bold, italic, or normal, the font size, a border for the font and many other attributes are controlled by <style> element.

5.5.4 Dygraphs

- Dygraphs is an open-source JavaScript library that produces interactive, zoomable charts of time series. It is designed to display dense data sets and enable users to explore and interpret them.
- It can handle large data sets with millions of plot points. It works in all browsers and zooms down for mobile devices. The dygraphs package is an R interface to the dygraphs JavaScript charting library.
- This library can be used to develop interactive charts on the X and Y axis and to display powerful diagrams. Dygraphs.js can use five types of input : CSV data, URL, array, function, DataTable.
- Some of the features of dygraphs :
  1. Plots time series without using an external server or flash
  2. Works in Internet Explorer (using excanvas)
  3. Lightweight (69 kB) and responsive
  4. Displays values on mouseover, making interaction easily discoverable
  5. Supports error bands around data series
  6. Interactive zoom

7. Displays annotations on the chart
8. Adjustable averaging period
9. Can intelligently chart fractions
10. Customizable click-through actions
11. Compatible with the Google Visualization API.

- The dygraphs package is available on CRAN now and can be installed with :

```
install.packages("dygraphs")
```

- Dygraphs work primarily with time series. If you have a DSS dataset with a "date" column, user will need to convert dataframe to a time series or XTS object.
- For example, the following will create a time-series of revenue by order_ts

```
library(xts)
df <- dkuReadDataset("orders")
timeseries <- xts(df$revenue, order.by=as.Date(df$order_ts))
# You can then plot timeseries
dkuDisplayDygraph(dygraph(timeseries) %>% dyRangeSelector())
```

- It allows users to explore and interpret dense data sets. All the charts are inter-active : It can be used mouse over to highlight individual values, or click and drag to zoom. It is possible to change the number and hit enter to adjust the averaging period. Dygraphs handles huge data sets.

5.5.5 1-D, 2-D and 3-D Data

- Every data set has a general structure. It is always characterised by a group of variables and the records the database contains. The first group consists of one - dimensional, two-dimensional, three - dimensional and high - dimensional data sets.
- The variable in one - dimensional data is usually time. An example is the log of interrupts in a processor.
- Two-dimensional data can often be found in statistics like the number of financial transactions in a certain period of time.
- Three - dimensional data can be positioned in three - dimensional space or points on a surface whereas time varies. High - dimensional data contains all those sets of data that have more than three considered variables. Examples are locations in space that vary with time.
- **Two-dimensional data** can be visualized in different ways. A very common visualization form is the **scatter plot.** In a scatterplot the frame for the data presentation is a Cartesian coordinate system, in which the axes correspond to the two dimensions.

- Another important visualization technique for two - dimensional data is the line graph. The difference to scatter plots is that this time the relation between the dimension on the horizontal axis and the one on the vertical axis is definite.
- **Three - dimensional data** : The two - dimensional techniques can easily be extended to three dimensions. The third - dimension is achieved in scatter plots and bar charts by adding a further axis, orthogonal to the other two.
- A scatter plot, more commonly called a graph of y versus x, shows the relationship of 2 variables and with the addition of colour can represent a 3rd variable. A scatterplot matrix of n variables is obtained by projection of the data onto n*(n–1) scatter plots, i.e., all possible combinations of scatter plots are drawn as illustrated in Fig. 5.5.4 which is an example for pressure, temperature and velocity data.

Pressure	PvT	PvV
TvP	Temperature	TvV
VvP	VvT	Velocity

Fig. 5.5.4

Review Questions

1. *Explain different data vizualization tools.* **SPPU : Dec.-19 (End Sem), Marks 8**
2. *List the conventional data visualization tools. Explain any two.* **SPPU : Dec.-19 (End Sem), Marks 8**

5.6 Case Study : Analysis of a Business Problem of Zomato using Visualization

- Founded in 2008 Zomato is a major food delivery aggregator with a markdown cap of 1 Trillion INR. It started as Foodiebay, a restaurant recommendation product, at its peak, it has 35000 menus and ₹ 60 Lakh monthly revenue. Foodiebay.com reroutes to zomato.com now.
- Zomato offered customer : A solution to have access to all the restaurants through a database, the type of food menu, location of the eateries and most importantly their feedback and the reviews.
- Zomato Kitchens under the banner of Zomato Infrastructure Services provides cloud kitchens to the best and reliable restaurants only. It provides kitchen equipment, tech stack, POS, and delivery, and tracking systems. Zomato earns a share of restaurants profit, thus making sure it's a win-win situation.

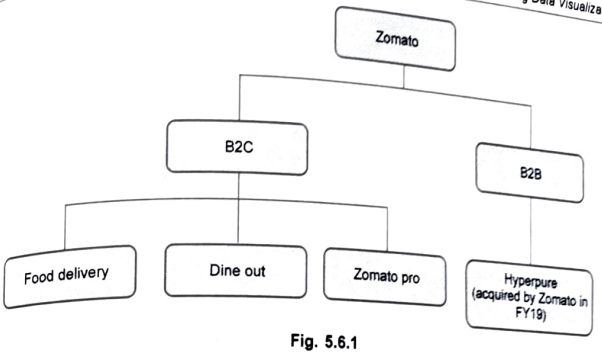


Fig. 5.6.1

- Zomato is dynamic on Instagram, Facebook and Twitter. Beginning at July 2019, it has 154 K followers on Instagram, 1,899,405 supporters and 1.42 Million lovers on Twitter.
- The dataset of restaurant was carried out by the researchers based on Zomato registered restaurant through Zomato API and it is publicly available on "www.kaggle.com". The dataset has multiple different variety of columns which are used to analyze and identify which city has highest number of good restaurants based on ratings, votes and analyzing pattern of expensive restaurant with quality of food.

5.7 Analytical Techniques used in Big Data Visualization

- There are various analytical techniques used in big data processing in order to extract, collect, store, process and analyze the huge amount of data coming very fast with the different variety.

Machine learning :

- A machine learning algorithm then takes these examples and produces a program that does the job. The program produced by the learning algorithm may look very different from a typical hand - written program. It may contain millions of numbers. If we do it right, the program works for new cases as well as the ones we trained it on.
- Main goal of machine learning is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example". Another goal is to develop computational models of human learning process and perform computer simulations.

- The goal of machine learning is to build computer systems that can adapt and learn from their experience.
- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instructions. It should carry out to transform the input to output.
- For example, for addition of four numbers is carried out by giving four numbers as input to the algorithm and output is sum of all four numbers. For the same task, there may be various algorithms. It is interested to find the most efficient one, requiring the least number of instructions or memory or both.
- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine learning provide business insight and intelligence. Decision makers are provided with greater insights into organizations. This adaptive technology is being used by global enterprise to gain a competitive edge.
- Machine learning algorithm discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.
- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behavior of a system for any set of input values, after an initial training phase.
- In contrast to supervised learning, unsupervised or self - organized learning does not require an external teacher. During the training session, the neural network receives a number of different input patterns, discovers significant features in these patterns and learns how to classify input data into appropriate categories.
- Unsupervised learning algorithms aim to learn rapidly and can be used in real - time. Unsupervised learning is frequently employed for data clustering, feature extraction, etc.
- Reinforcement learning : This is advanced machine learning technique. This is based on probability theory where mapping can be done based on input received and changes based on environment around it.
- Deep learning : This is also advanced machine learning technique which has multiple processing layer so as to produce non-linear response based on input data. There are so many small processors called as neuron working parallel in data processing.
- Predictive analytics : This technique refers to prediction based on past experience and it uses both data mining and machine learning.
- Association rule learning : This is used to identify interesting relations between different attributes from large datasets.

5.8 Data Visualization using Tableau

SPPU : May-18, Dec.-19

- Tableau is one of the fastest evolving Business Intelligence (BI) and data visualization tool. Tableau server is a business intelligence application that provides browser - based analytics anyone can use. It's a rapid-fire alternative to the slow pace of traditional business intelligence software.
- A business intelligence and data visualization tool allowing users to make sense of their data through interactive charts, graphs, and diagrams.
- Why uses Tableau ?
  1. Traditional BI tools require complex installations
  2. Rapid results to useful information
  3. Easy to use for all skill levels
  4. Excellent migration path for excel users
  5. It can use many different sources of data.
- Tableau uses a visual query language. The tableau data engine is a breakthrough in-memory analytics database designed to overcome the limitations of existing databases and data silos.
- Capable of being run on ordinary computers, it leverages the complete memory hierarchy from disk to L1 cache. It shifts the curve between big data and fast analysis.
- Tableau allows the users to directly connect to databases, cubes and data warehouses etc. After analysing the data, the results can be shared live with just a few clicks. The dashboard can be published to share it live on web and mobile devices.
- Tableau is relatively new in the business intelligence market but its market share is growing on a daily basis. It is being nearly all industries, from transportation to healthcare.
- Tableau software does not support expanded analytics such as box plots, network graphs, tree - maps, heat-maps, 3D-scatter plots, profile charts or data relationships tool which allow users to mine data for relationships like another data visualization software does.
- Tableau connects and extracts the data stored in various places. It can pull data from any platform imaginable. A simple database such as an excel, pdf, to a complex database like Oracle, a database in the cloud such as Amazon webs services, Microsoft Azure SQL database, Google Cloud SQL and various other data sources can be extracted by Tableau.

- Tableau saves time when updating daily and weekly reports that currently reside in spreadsheets. That's because Tableau separates the data layer from the presentation layer and makes updating a spreadsheet data source a trivial append to the bottom of your source spreadsheet.
- Tableau is not an ETL engine for cleaning-up bad data, although it can be very helpful in identifying missing or erroneous data in your existing data sources. Visualizing data via time series, bar charts, scatter plots or in maps highlights errors and outliers more effectively than grids of data in a spreadsheet.

Review Question

1. *Explain data visualization with Tableau.*

SPPU : May-18 (End Sem), Dec.-19 (End Sem), Marks 8

5.9 Introduction to Candela

SPPU : May-18

- As an open-source suite of web visualization components that make use of the Python language, Candela emphasizes scalable, rich visualizations created with a normalized API for use in real-world data science situations.
- Candela is an open-source suite of inter-operable web visualization components. It provides library for JavaScript, package for Python and R.
- The tool works on Kitware's resonant platform and offers a range of elements for data visualization. It allows user to make super-rich visualizations that are scalable and available within a normalized API.
- Candela is a JavaScript library that provides reusable visualization components for the web. Let's consider following points :

1. Reusable : Candela provides a general API not tied to any particular framework or library, so the components user create with it can be ported from application to application very easily.
2. Visualization : That general API does not provide many constraints; the main one is that user must implement a function called render() which will carry out visualization semantics, whatever they may be.
3. Components : Use object-oriented concepts to implement the notion of a visualization component.
4. For the web : Candela uses modern JavaScript, taking advantage of features and modern tooling to produce a library that can be used to do visualization almost anywhere on the web.

5.9.1 D3.js

- D3.js is also known as D3, short for Data-Driven Documents. It is an open-source JavaScript library, which is certainly the most excellent in terms of developing interactive data visualization in the web browser.
- D3.js also lets influential data visualization mechanism and carries a data-driven tactic to DOM (Document Object Model) handling. It functions on a random series of connections, better known as selections.
- D3 also uses web technologies like HTML, CSS, SVG and JavaScript.
- A JavaScript library for creating data visualizations with an emphasis on web standards.
- Using HTML, SVG and CSS, bring documents to life with a data-driven approach to DOM manipulation, all with the full capabilities of modern browsers and no constraints of proprietary frameworks.
- Key features :
  a) Bind arbitrary data to DOM
  b) Create interactive SVG bar charts
  c) Generate HTML tables from data sets
  d) Variety of components and plugins to enhance capabilities
  e) Built-in reusable components for ease of coding.
- For example, user can use D3 to generate an HTML table from an array of numbers. Or, use the same data to create an interactive SVG bar chart with smooth transitions and interaction.

5.9.2 Google Chart API

- The Google Chart API is an interactive web service that creates graphical charts from user - supplied data.
- Google servers create a PNG image of a chart from data and formatting parameters specified by a user's HTTP request. The service supports a wide variety of chart information and formatting.
- The Google Chart Tools enable adding live charts to any web page. They provide advantages such as a rich gallery of visualizations provided as image charts and interactive charts and they can read live data from a variety of data sources.
- Users embed the data and formatting parameters in an HTTP request and Google returns a PNG image of the chart. Many types of chart are supported and by making the request into an image tag the chart can be included in a web page.

Review Question

1. *Explain : i) Google chat API ii) Cloudera.* **SPPU : May-18 (End Sem), Marks 8**

5.10 Multiple Choice Questions

Q.1 3D scatter plots are used to plot data points on three axes in the attempt to show the relationship between _____ variables.

a two b three
c four d six

Q.2 _____ projection techniques help users find interesting projections of multidimensional data sets.

a Geometric b Pixel oriented
c Circle segments d None

Q.3 List categorization of visualization methods.

a Pixel-oriented visualization techniques
b Geometric projection visualization techniques
c Icon-based visualization techniques
d All of these

Q.4 Line graph is also called _____ graph.

a X-Y b stick
c column d row

Q.5 Treemaps display hierarchical data using _____.

a rectangles b square
c triangle d circle

Q.6 What does D3.js mean ?

a It is a JavaScript framework to display D3 models.
b It is a JavaScript library for changing native objects to D3 objects.
c It is Node.js to parse a server's data to objects with D3 features.
d It is a JavaScript library for creating and manipulating documents based on data.

Q.7 What are the types of filters are used in Tableau ?

a Custom filters b Context filters
c Normal filters d all of the above

Q.8 Which of the following is not a challenge in big data visualization ?

a Velocity b Volume
c Version d Variety

Answer Keys for Multiple Choice Questions :

Q.1	b	Q.2	a	Q.3	d	Q.4	b	Q.5	a
Q.6	d	Q.7	d	Q.8	c				

❑❑❑