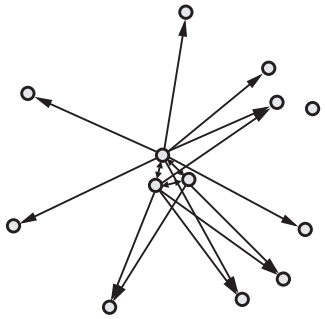


Fig. 5.4.2 (a) : Core - periphery structure

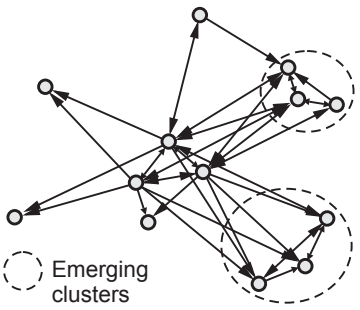


Fig. 5.4.2 (b) Cluster structure

1. Affiliation network :

- **Affiliation networks** contain information about the relationships between two sets of nodes : A set of subjects and a set of affiliations. An affiliation network can be formally represented as a bipartite graph, also known as a two-mode network.
- Affiliation networks are **two mode networks** that allow one to study the dual perspectives of the actors and the events. They look at collections or subsets of actors or subsets rather than ties between pairs of actors. Connections among members of one of the modes as based on linkages established through the second mode.
- An affiliation network is a network in which actors are joined together by common membership of groups or clubs of some kind.
- A distinctive feature of affiliation networks is **duality** i.e. events can be described as collections of individuals affiliated with them and actors can be described as collections of events with which they are affiliated.
- Based on two-mode matrix data, affiliation networks consist of sets of relations connecting actors and events, rather than direct ties between pairs of actors as in one-mode data. Familiar affiliation networks include persons belonging to associations, social movement activists participating in protest events, firms creating strategic alliances, and nations signing treaties.
- The representation of two-mode data should facilitate the visualization of three kinds of patterning :
  a. The actor-event structure
  b. The actor-actor structure
  c. The event-event structure

- Many ways to represent affiliation networks :
  1. Affiliation network matrix
  2. Bipartite graph
  3. Hypergraph
  4. Simplicial complex

Benefits of affiliations network

1. Affiliations of actors with events provide a direct linkage between actors through memberships in events, or between events through common memberships.
2. Affiliations provide conditions that facilitate the formation of pairwise ties between actors.
3. Affiliations enable us to model the relationships between actors and events as a whole system.

2. Bipartite graph :

- Nodes are partitions into two subsets and all lines are between pairs of nodes belonging to different subsets. Fig 5.4.3 shows bipartite network. As there are g actors and h events, there are g + h nodes.

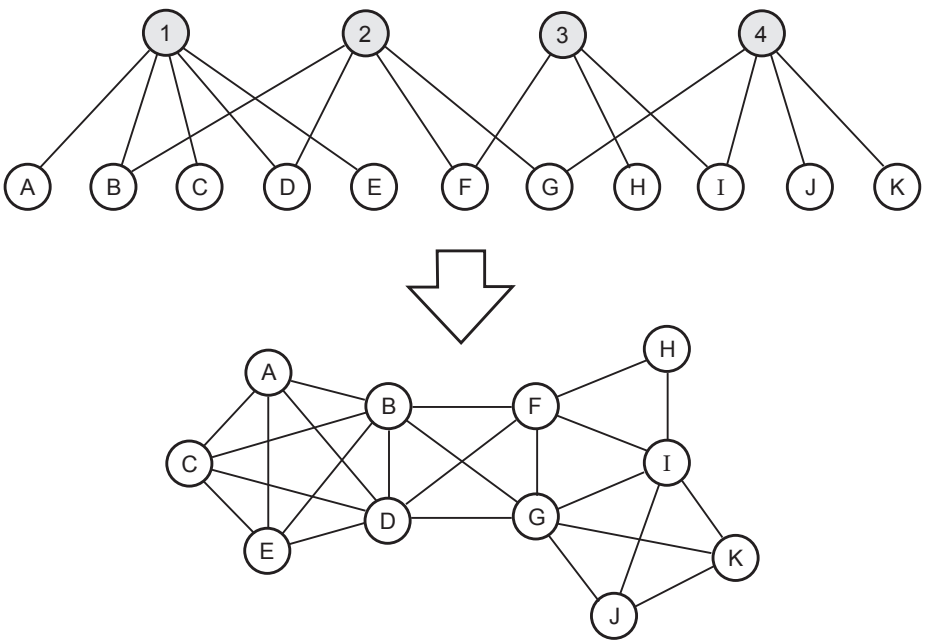


Fig. 5.4.3 Bipartite graph

- The lines on the graph represent the relation "is affiliated with" from the perspective of the actor and the relation "has as a member" from the perspective of the event.
- No two actors are adjacent and no two events are adjacent. If pairs of actors are reachable, it is only via paths containing one or more events. Similarly, if pairs of events are reachable, it is only via paths containing one or more actors.

Advantages

1. They highlight the connectivity in the network, as well as the indirect chains of connection.
2. Data is not lost and we always know which individuals attended which events.

Disadvantage

1. They can be unwieldy when used to depict larger affiliation networks.

5.4.5 Application of Social Network Analysis

- Social Network Analysis (SNA) is an important and valuable tool for knowledge extraction from massive and un-structured data. Social network provides a powerful abstraction of the structure and dynamics of diverse kinds of inter-personal connection and interaction.
- Facebook is a social networking service and website that connects people with other people, and share data between people. A user can create a personal profile, add other users as friends, exchange data, create and join common interest communities.
- Twitter is a social net-working and microblogging service. The users of Twitter can exchange text-based posts called tweets. A tweet is a maximum 140 characters long but can be augmented by pictures or audio recording. The main concept of Twitter was to build a social network formed by friends and followers. Friends are people who you follow, followers are those who follow you.
- The role of social networks in labor markets deserves attention for at least two reasons : First, because of the central role networks play in disseminating information about job openings they place a critical role in determining whether labor markets function efficiently; and second, because network structure ends up having implications for things like human capital investment as well as inequality.
- Social Network Analysis (SNA) primarily focuses on applying analytic techniques to the relationships between individuals and groups, and investigating how those relationships can be used to infer additional information about the individuals and groups.

- SNA is used in a variety of domains. For example, business consultants use SNA to identify the effective relationships between workers that enable work to get done; these relationships often differ from connections seen in an organizational chart.

5.5 Introduction to Business Analysis

- Business Analysis is a discipline and practice of defining business needs and recommending solutions to business problems. Business analysis deals with the current state of each company, desired future state, stakeholders' needs, processes, software and more.
- A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms.
- The adoption of a business intelligence system tends to promote a scientific and rational approach to the management of enterprises and complex organizations.
- Obstacle to business intelligence in an organization are as follows :

1. Lack of BI strategy : Organizations should proactively define the problems they trying to solve. Only then they will be able to identify the right Business Intelligence solution that will suit their requirements.

2. Business intelligence :

- When You Don't Know How to Code. Now a days, executives find it difficult to access the right data at right time. And even if they do find what they're looking for, data formats are typically so complex and unstructured it's hard to find out meaningful and relevant data.
- Now unless they are using Excel extensively, they probably would not get much satisfaction (or value) from their BI system.
- A good practice would be to replace Excel Sheets with intuitive dashboards to make data more engaging, meaningful and eventually very powerful.

3. Lack of training and execution :

- Many a times, companies might have well-articulated requirements, a sound BI strategy, and a good tool solution, but lack technical skills like designing, building, maintaining, and supporting BI applications.
- This results in BI applications to run slowly, break frequently, deliver uncertain results and eventually leading to rising cost of using the BI solution. The causes of lack of execution often are multiple and varied, as are its remedies.

4. Lack of BI impact (Low utilization) :

- Management might always wonder why there is no change in business results attributable to BI and might feel that business value of BI investments not captured. This indicates that the organization is not utilizing the BI solution at par with global standards and best practices.

5. Business intelligence with unstructured data :

- Most of the times data is unstructured for BI to analyze. This lead to a problem when users need to perform simple BI processes.
- Businesses may invest in big data analytics but cannot complete the tasks in time. They may result to people spending hours on cleaning and structuring the data first and then using the BI solution.

6. Installation and deployment :

- A painful BI solution installation and deployment would be difficult to maintain. Even an unplanned and rushed deployment would be unsuccessful so often.
- Doing this may leave users void with time to understand the system and develop the skills using the solution effectively.

5.6 Model Evaluation and Selection

SPPU : Dec.-18

- A binary classification rule is a method that assigns a class to an object, on the basis of its description.
- The performance of a binary classifier can be assessed by tabulating its predictions on a test set with known labels in a contingency table or confusion matrix, with actual classes in rows and predicted classes in columns.
- Measures of performance need to satisfy several criteria :
  1. They must coherently capture the aspect of performance of interest ;
  2. They must be intuitive enough to become widely used, so that the same measures are consistently reported by researchers, enabling community - wide conclusions to be drawn ;
  3. They must be computationally tractable, to match the rapid growth in the scale of modern data collection.
  4. They must be simple to report as a single number for each method - dataset combination.
- Performance metrics for binary classification are designed to capture trade-offs between four fundamental population quantities : True positives, false positives, true negatives and false negatives.

- The evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, named confusion matrix.
- Confusion matrix is also called a contingency table.
  1. **False positives :** Examples predicted as positive, which are from the negative class.
  2. **False negatives :** Examples predicted as negative, whose true class is positive.
  3. **True positives :** Examples correctly predicted as pertraining to the positive class.
  4. **True negatives :** Examples correctly predicted as belonging to the negative class.

$$\text{Accuracy rate} = \frac{|\text{True negatives}| + |\text{True positives}|}{|\text{False negatives}|+|\text{True positive}|+|\text{True negatives}|+|\text{True positives}|}$$

- The complement of accuracy rate is the error rate, which evaluates a classifier by its percentage of incorrect predictions.

$$\text{Error rate} = \frac{|\text{False negatives}| + |\text{False positives}|}{|\text{False negatives}|+|\text{False positive}|+|\text{True negatives}|+|\text{True positives}|}$$

$$\text{Error rate} = 1 - (\text{Accuracy rate})$$

- The recall and specifity measures evaluate the effectiveness of a classifier for each class in the binary problem. The recall is also known as sensitivity or true positive rate. Recall is the proportion of examples belonging to the positive class which were correctly predicted as positive.
- The **specificity** is a statistical measure of how well a binary classification test correctly identifies the negatives cases.

$$\text{Recall (R)} = \frac{|\text{True positive}|}{|\text{True positive}|+|\text{False negative}|}$$

$$\text{Specificity} = \frac{|\text{True negatives}|}{|\text{False positives}|+|\text{True negative}|}$$

- True Positive Rate (TPR) is also called sensivity, hit rate, and recall.

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}}$$

- A statistical measure of how well a binary classification test correctly identifies a conditon. Probability of correctly labeling members of the target class.

- No single measure tells the whole story. A classifier with 90 % accuracy can be useless if 90 percent of the population does not have cancer and the 10 % that do are misclassified by the classier. Use of multiple measures recommended.
- Binary classification accuracy metrics quantify the two types of correct predictions and two types of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Each metric measures a different aspect of the predictive model.

5.6.1 Issues Regarding Classification and Prediction

Preparing the data for classification and prediction

Following pre-processing steps may be applied to the data to help improve the accuracy, efficiency, and scalability of the classification or prediction process.

a) **Data cleaning :** Preprocess data in order to reduce noise and handle missing values.

b) **Relevance analysis (Feature selection) :** Remove the irrelevant or redundant attributes.

c) **Data transformation :** Generalize and/or normalize data

Feature wise comparison between classification and prediction :

1. **Accuracy :** Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
2. **Speed :** This refers to the computational cost in generating and using the classifier or predictor.
3. **Robustness :** It refers to the ability of classifier or predictor to make correct predictions from given noisy data.
4. **Scalability :** Scalability refers to the ability to construct the classifier or predictor efficiently ; given large amount of data.
5. **Interpretability :** It refers to what extent the classifier or predictor understands.

5.6.2 Holdout Method

- The data is split into two different datasets labelled as a training and a testing dataset. This can be a 60/40 or 70/30 or 80/20 split. This technique is called the hold-out validation technique.
- Suppose we have a database with house prices as the dependent variable and two independent variables showing the square footage of the house and the number of rooms.

- Now, imagine this dataset has 30 rows. The whole idea is that you build a model that can predict house prices accurately.
- To 'train' your model, or see how well it performs, we randomly subset 20 of those rows and fit the model.
- The second step is to predict the values of those 10 rows that we excluded and measure how well our predictions were.
- As a rule of thumb, experts suggest to randomly sample 80 % of the data into the training set and 20 % into the test set.

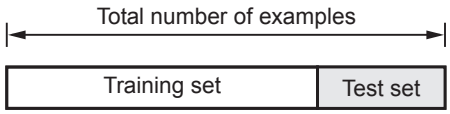


Fig. 5.6.1

- Training set : Used to train the classifier.
- The holdout method has two, basic drawbacks :
  1. It requires extra dataset
  2. It is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split.

5.6.3 Random Subsampling

- Random subsampling performs K data splits of the entire sample. For each data split, a fixed number of observations is chosen without replacement from the sample and kept aside as the test data.
- The prediction model is fitted to the training data from scratch for each of the K splits and an estimate of the prediction error is obtained from each test set.
- Let the estimated PE in the i^{th} test set be denoted by E_i. The true error estimate is obtained as the average of the separate estimates E_i.

$$\frac{1}{K}\sum_{i=1}^{K} E_i$$

1. Cross-Validation

- Cross-validation is a technique for evaluating estimating performance by training several machine learning models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, i.e., failing to generalize a pattern.
- In general, machine learning involves deriving models from data, with the aim of achieving some kind of desired behaviour, e.g., prediction or classification.

- Fig. 5.6.2 shows cross-validation.

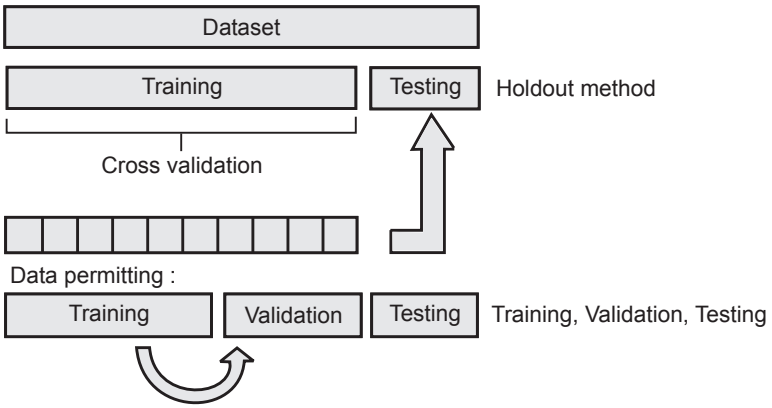


Fig. 5.6.2 Cross validation

- But this generic task is broken down into a number of special cases. When training is done, the data that was removed can be used to test the performance of the learned model on "new" data. This is the basic idea for a whole class of model evaluation methods called **cross validation**.
- Types of cross validation methods are holdout, K-fold and Leave-one-out.
- The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximate fits a function using the training set only.
- The K-fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times.
- Each time, one of the k subsets is used as the test set and the other k - 1 subsets are put together to form a training set. Then the average error across all k trials is computed.
- Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set.
- That means that N separate times, the function approximate is trained on all the data except for one point and a prediction is made for that point.
- Cross-validation ensures non-overlapping test sets.

K-fold cross-validation :

- In this technique, k – 1 folds are used for training and the remaining one is used for testing as shown in Fig. 5.6.3.

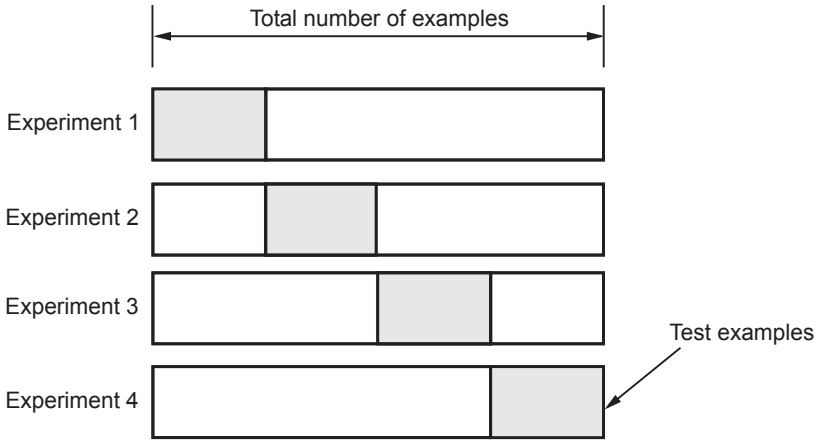


Fig. 5.6.3 K-fold cross validation

- The advantage is that entire data is used for training and testing. The error rate of the model is average of the error rate of each iteration.
- This technique can also be called a form the repeated hold-out method. The error rate could be improved by using stratification technique.

Review Question

1. *Explain any three of classification performance measures.*

SPPU : Dec.-18 (End Sem), Marks 6

5.7 Clustering and Time-series Analysis using Scikit-learn

- Time series data is widely used to analyse different trends and seasonalities of products over time by various industries. Sktime is a unified python framework/library providing API for machine learning with time series data and sklearn compatible tools to analyse, visualize, tune and validate multiple time series learning models such as time series forecasting, time series regression and classification.
- Time series are a stream of data that are created by making measures of something such as sales, temperature, stocks, etc. in fixed frequency. They have to be indexed in time order and usually used in weather forecasting, econometrics, earthquake prediction, signal processing, etc.
- Clustering of unlabeled data can be performed with the module sklearn.cluster.

5.7.1 Scikit-learn

- Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

- It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.
- Scikit-learn is a library, i.e. a collection of classes and functions that users import into Python programs. Using scikit-learn therefore requires basic Python programming knowledge
- The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes :
  1. NumPy : Base n-dimensional array package
  2. SciPy : Fundamental library for scientific computing
  3. Matplotlib : Comprehensive 2D/3D plotting
  4. IPython : Enhanced interactive console
  5. Sympy : Symbolic mathematics
  6. Pandas : Data structures and analysis
- Scikit-learn is the package for machine learning and data science experimentation favoured by most data scientists. It contains a wide range of well-established learning algorithms, error functions, and testing procedures.

5.7.2 Understanding Classes in Scikit-learn

- Scikit-learn takes a highly object-oriented approach to machine learning models. Every major Scikit-learn class inherits from ***sklearn.base.BaseEstimator***.
- Scikit-learn features some base classes on which all the algorithms are built. Apart from *BaseEstimator*, the class from which all other classes inherit, there are four class types covering all the basic machine learning functionalities :
  1. Classifying
  2. Regressing
  3. Grouping by clusters
  4. Transforming data
- Scikit-learn takes a highly object-oriented approach to machine learning models. Every major Scikit-learn class inherits from ***sklearn.base.BaseEstimator***.
- All objects within scikit-learn share a uniform common basic API consisting of three complementary interfaces : an **estimator** interface for building and fitting models, a **predictor** interface for making predictions and a **transformer** interface for converting data.

1. Estimators

- The estimator interface is at the core of the library. It defines instantiation mechanisms of objects and exposes a fit method for learning a model from training data.

- All supervised and unsupervised learning algorithms (e.g., for classification, regression or clustering) are offered as objects implementing this interface. Machine learning tasks like feature extraction, feature selection or dimensionality reduction are also provided as estimators.

2. Predictors

- The predictor interface extends the notion of an estimator by adding a predict method that takes an array X test and produces predictions for X test, based on the learned parameters of the estimator.
- In the case of supervised learning estimators, this method typically returns the predicted labels or values computed by the model.

3. Transformers

- Since it is common to modify or filter data before feeding it to a learning algorithm, some estimators in the library implement a transformer interface which defines a transform method.
- It takes as input some new data X test and yields as output a transformed version of X test.
- Preprocessing, feature selection, feature extraction and dimensionality reduction algorithms are all provided as transformers within the library.

5.8 Confusion Matrix

SPPU : May-19

- A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The numbers displayed give the frequency of each data point.
- The confusion matrix for binary classification shown below :

	Predicted class	
True class	**Positive**	**Negative**
Positive	True negative	False negative
Negative	False positive	True negative

- A confusion matrix contains information about actual and predicted classification done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. Confusion matrix is also called a contingency table.
  1. **False positives :** Examples predicted as positive, which are from the negative class.

2. **False negatives :** Examples predicted as negative, whose true class is positive.
3. **True positives :** Examples correctly predicted as pertraining to the positive class.
4. **True negatives :** Examples correctly predicted as belonging to the negative class.

- Binary classification accuracy metrics quantify the two types of correct predictions and two types of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Each metric measures a different aspect of the predictive model.
- Accuracy (ACC) measures the fraction of correct predictions. Precision measures the fraction of actual positives among those examples that are predicted as positive. Recall measures how many actual positives were predicted as positive. F1-measure is the harmonic mean of precision and recall many actual positives were predicted as positive. F1-measure is the harmonic mean of precision and recall.

5.8.1 ROC Curve

- Receiver Operating Characteristics (ROC) graphs have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates noisy channel. Recent years have seen an increase in the use of ROC graphs in the machine learning community.
- An ROC plot plots true positive rate on the Y-axis against false positive rate on the X-axis ; a single contingency table corresponds to a single point in an ROC plot.
- The performance of a ranker can be assessed by drawing piecewise linear curve in an ROC plot, known as an ROC curve. The curve starts in (0, 0), finishes in (1, 1) and is monotonically non-decreasing in both axes.
- A useful technique for organizing classifiers and visualizing their performance. Especially useful for domains with skewed class distribution and unequal classification error costs.
- It allows to create ROC curve and a complete sensitivity / specificity report. The ROC curve is a fundamental tool for diagnostic test evaulation.
- In a ROC curve the true positive rate (sensitivity) is plotted in function of the false positive rate (100 -specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity / specificity pair corresoponding to a particular decision threshold. The area under the ROC curve is a measure of how well a parameter two segments.

- Each point on an ROC curve connecting two segments corresponds to the true and false positive rates achieved on the same test set by the classifier obtained from the ranker by splitting the ranking between those two segments.
- An ROC curve is convex if the slopes are monotonically non-increasing when moving along the curve from (0, 0) to (1, 1). A concavity in an ROC curve, i.e. in an ROC curve, i.e. two or more adjacent segment with increasing slopes, indicates a locally worse than random ranking. In this case, we would get better ranking performance by joining the segments involved in concavity, thus creating a coarser classifier.

Review Question

1. *Explain the term confusion matrix.* **SPPU : May-19 (End Sem), Marks 4**

5.9 Elbow Plot

- The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k.
- If k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.
- It involves running the algorithm multiple times over a loop, with an increasing number of cluster choice and then plotting a clustering score as a function of the number of clusters.
- The Elbow and Silhouette methods are the two state-of-the-art methods used to identify the correct cluster number in the dataset.
- The Elbow method is the oldest method to distinguish the potential optimal cluster number for the analyzed dataset, whose basic idea is to specify K = 2 as the initial optimal cluster number K, and then keeps increasing K by step 1 to the maximal specified for the estimated potential optimal cluster number, and finally distinguish the potential optimal cluster number K corresponding to the plateau.
- The optimal cluster number K is distinguished by the fact that before reaching K, the cost rapidly decreases to the called cost peak value, and after exceeding K, it continues to increase with the called cost peak value almost unchanged, as shown in Fig. 5.9.1 (a) with an explicit elbow point.

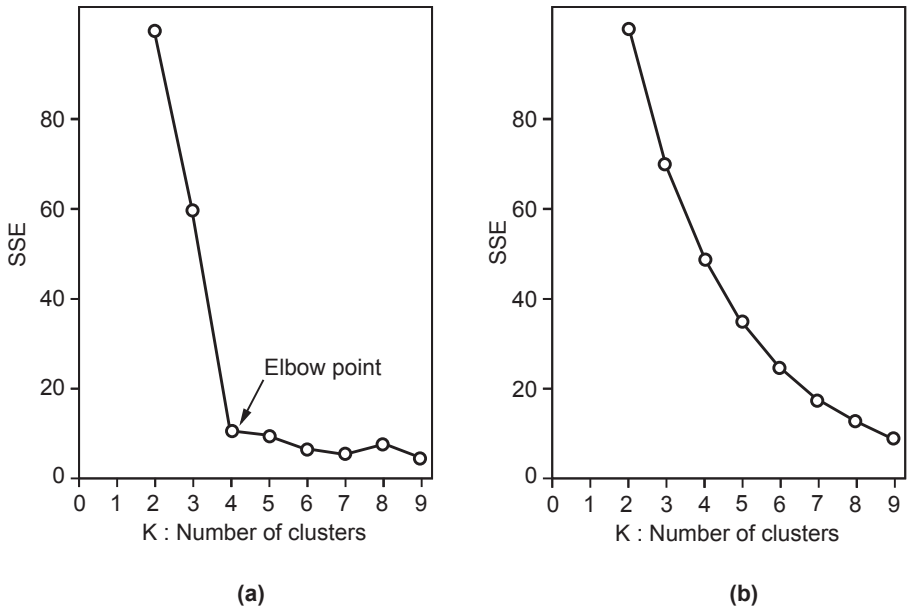


Fig. 5.9.1 Elbow point

- Meanwhile, the optimal cluster number corresponding to the elbow point depends on the manmade selection. There is, however, a problem with the Elbow method in that the elbow point cannot be unambiguously distinguished by the experienced analysts when the plotted curve is fairly smooth, as shown in Fig. 5.9.1 (b) with an ambiguous elbow point.
- To select the best K, we need to plot the mean in-cluster distance for each K. As K increases from 1, before reaching the optimal K, the decrease speed is relatively fast because the number of centers are too low from the very beginning and each new center will incur a large decrease in the mean distance.
- But after the optimal K, the decrease is slow since the correct cluster structure is already discovered and any newly added center will appear in a certain cluster already formed. That will not decrease the mean in-cluster distance too much. The entire curve looks like an L shape and the best K lies in the turning point or the elbow of the L shape.

5.10 Multiple Choice Questions

Q.1 A _____ is a a flowchart-like tree structure, where each internal node denotes a test on an attribute.

a desicion tree b binary tree

c cluster d none of these

Q.2 A node without further branches is called as _____.

a internal node
b root node
c lead node
d binary node

Q.3 _____ occurs when the gap between the training error and test error is too large.

a Underfitting
b Overfitting
c Overloaded
d Purning

Q.4 What is the approach of basic algorithm for decision tree induction ?

a Greedy
b Top down
c Procedural
d Step by Procedural

Q.5 What are two steps of tree pruning work ?

a Pessimistic pruning and Optimistic pruning
b Postpruning and Prepruning
c Cost complxity pruning and time complexity pruning
d None of the options

Q.6 How will you counter over-fitting in decision tree ?

a By pruning the longer rules
b By creating new rules
c Both by pruning the longer rules and by creating new rules.
d None of the above

Q.7 Which of the following is not a forecasting technique ?

a Judgemental
b Time series
c Time horizon
d Associative

Q.8 Which of the following is not true for forecasting ?

a Forcasts are rarely perfect.
b The underlying casual system will remain same in the future.
c Forecast for group of items is accurate than individual item.
d Short range forcasts are less accurate than long range forecasts.

Q.9 ETL stand for _____.

a Extract Transform and Load

b Exact Transfer and Language

c Extract Transmission and Language

d None

Q.10 ARIMA is a _____ model that uses time series data to either better understand the data set or to predict future trends.

a statistical analysis b analytical

c descriptive d all of these

Answer Keys for Multiple Choice Questions :

Q.1	a	**Q.2**	c	**Q.3**	b	**Q.4**	a	**Q.5**	b
Q.6	a	**Q.7**	c	**Q.8**	d	**Q.9**	a	**Q.10**	a

❑❑❑

UNIT VI

6 Data Visualization and Hadoop

Syllabus

Introduction to Data Visualization, Challenges to Big data visualization, Types of data visualization, Data Visualization Techniques, Visualizing Big Data, Tools used in Data Visualization, Hadoop ecosystem, Map Reduce, Pig, Hive, Analytical techniques used in Big data visualization. ***Data Visualization using Python :*** *Line plot, Scatter plot, Histogram, Density plot, Box- plot.*

Contents

6.1 Introduction to Data Visualization . . . . . **Dec.-18, May-19** · · · · · · · · · · Marks 9

6.2 Types of Data Visualization . . . . . **Dec.-18 , 19** · · · · · · · · · · · · · Marks 9

6.3 Data Visualization Techniques . . . . . **Dec.-18, 19** · · · · · · · · · · · · · · Marks 8

6.4 Visualizing Big Data . . . . . **Dec.-18, 19, May-19** · · · · · · · Marks 8

6.5 Tools used in Data Visualization . . . . . **May-19, Dec.-19** · · · · · · · · · Marks 8

6.6 Hadoop Ecosystem . . . . . **Dec.-18 , 19, May-19** · · · · · · Marks 9

6.7 Multiple Choice Questions

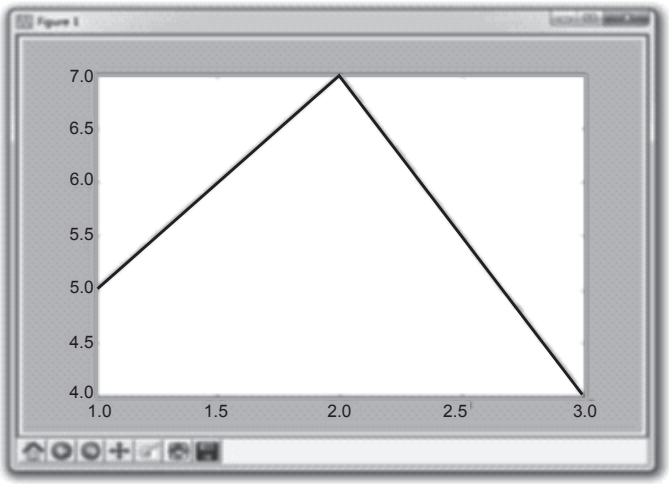
6.1 Introduction to Data Visualization

SPPU : Dec.-18, May-19

- Data visualization is the presentation of quantitative information in a graphical form. In other words, data visualizations turtan 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, we need to start with clean data that is well sourced and complete. Once data is ready to visualize, we need to pick the right chart. This can be tricky, but there are many resources available to help us to choose the right type of chart for 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, we 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 we may have gotten an error. 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 he may take.
  3. **Sharing :** Advice and suggestions on the possible actions can be easily shared across the company which will lead to a consequent.
- For fully document graph, we 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.

6.1.1 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 challenges in big data visualization ?* **SPPU : Dec.-18 (End Sem), Marks 8**
2. *What is data visualization ? Describe any four data visualization techniques* **SPPU : May-19 (End Sem), Marks 8**
3. *Why is it difficult to visualize big data ? Also explain analytical techniques used in big data visualization* **SPPU : May-19 (End Sem), Marks 9**

6.2 Types of Data Visualization

SPPU : Dec.-18, 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 Questions

1. *What is data visualization ? Explain any four data visualization techniques*

SPPU : Dec.-18 (End Sem), Marks 9

2. *Explain data visualization with respect to 1-D, 2-D and 3-D data.*

SPPU : Dec.-19 (End Sem), Marks 9

6.3 Data Visualization Techniques

SPPU : 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 you 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.

6.3.1 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. 6.3.1 show line graph.
- 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.

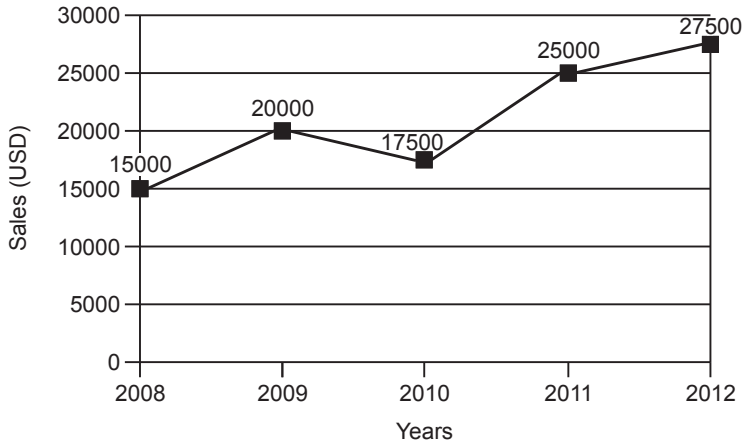


Fig. 6.3.1 Line graph

- 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.
- 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.

6.3.2 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. 6.3.2 shows pie chart.

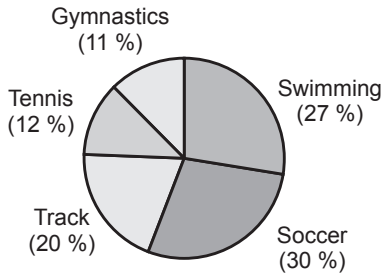


Fig. 6.3.2 Pie chart

- 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.
- 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.

6.3.3 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, we 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. 6.3.3 shows venn diagram of this data.

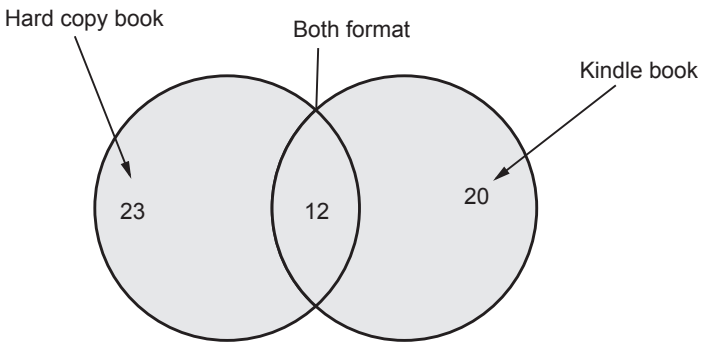


Fig. 6.3.3 : Venn diagram

- 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.

6.3.4 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. 6.3.4 shows scatter plots of two variables.

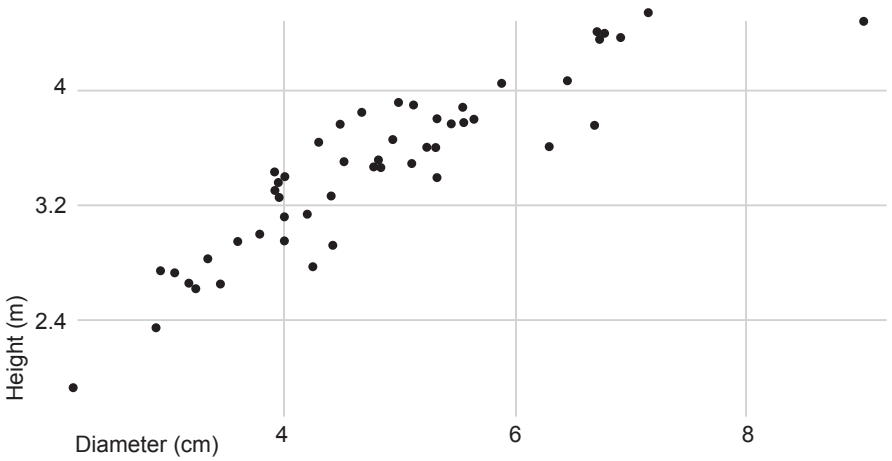


Fig. 6.3.4 : Scatter plot

- 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 how data visualization is done or visually represented, if data is 1-D, if data 2-D and data is 3-Dimentional ?* **SPPU : Dec.-18 (End Sem), Marks 6**
2. *Explain analytical techniques used in big data visualization.* **SPPU : Dec.-18 (End Sem), Marks 3**
3. *Why it is difficult to visualize big data ?* **SPPU : Dec.-19 (End Sem), Marks 8**

6.4 Visualizing Big Data

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

- 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.
- 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

Review Questions

1. *Explain big data visualization tools in short (any four tools).*

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

2. *Explain various tools to visualize big data. (Any four)*

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

6.5 Tools used in Data Visualization

SPPU : May-19, 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.

6.5.1 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.
- 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.

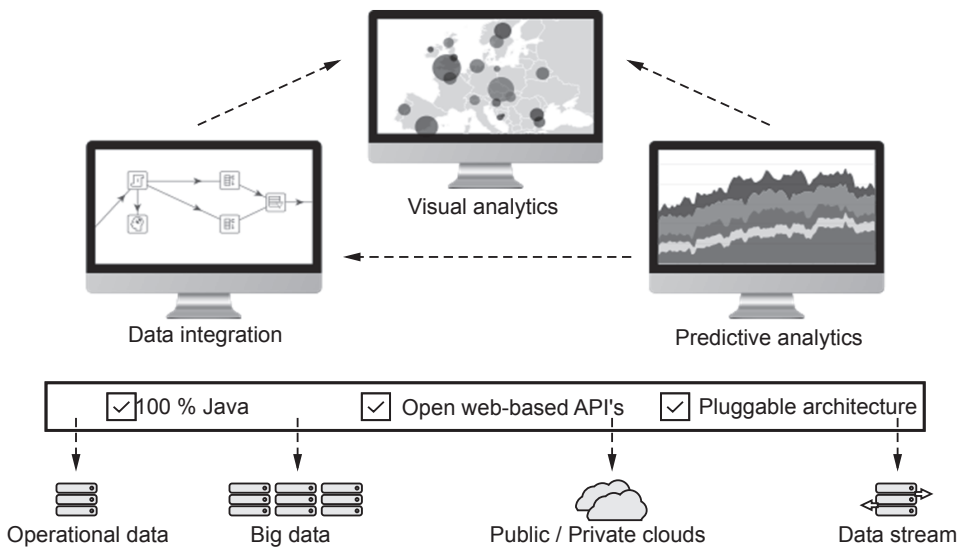


Fig. 6.5.1

- 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.

6.5.2 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. 6.5.2 shows all Datameer functionality occurs across three major components.

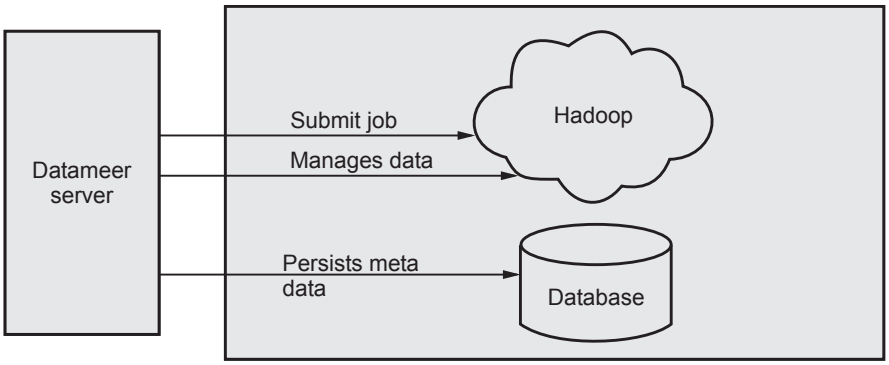


Fig. 6.5.2 Datameer functionality

- 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.

6.5.3 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. 6.5.3 shows JasperReport.

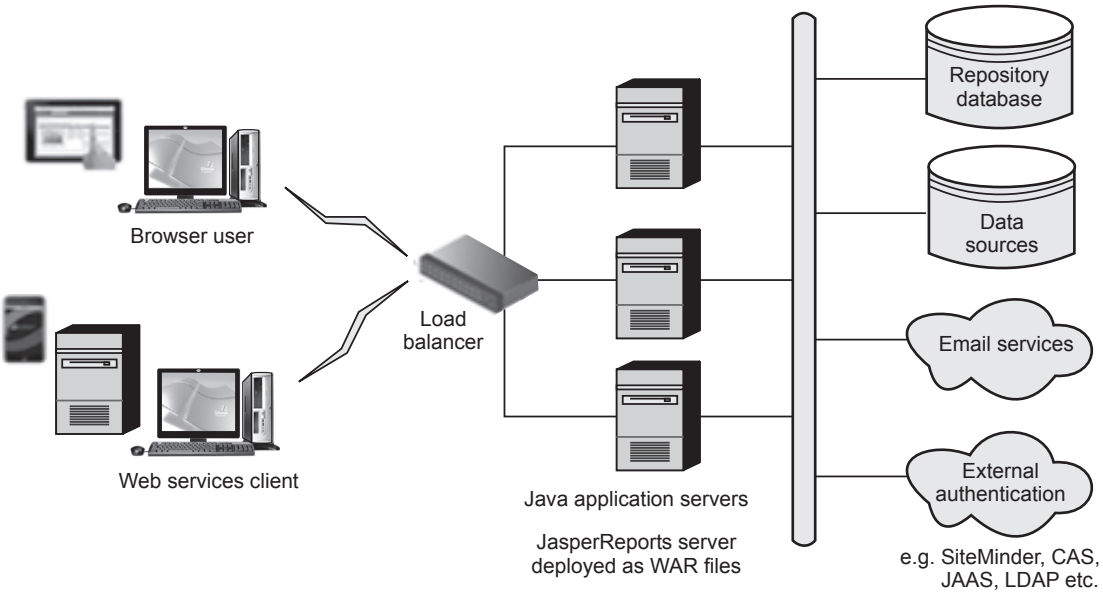


Fig. 6.5.3 JasperReport

- 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 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.

6.5.4 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, you'll need to convert your 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.

6.5.5 Tableau

- Tableau is one of the fastest evolving Business Intelligence (BI) and data visualization tools. 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 use 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 used Tableau.
- 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 tools 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 web 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 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.

6.5.6 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. 6.5.4 which is an example for pressure, temperature and velocity data.

Pressure	PvT	PvV
TvP	Temperature	TvV
VvP	VvT	Velocity

Fig. 6.5.4

Review Question

1. *Explain data visualization tool - Tableau.* **SPPU : Dec.-19, May-19 (End Sem), Marks 8**

6.6 Hadoop Ecosystem

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

- Hadoop ecosystem is neither a programming language nor a service, it is a platform or framework which solves big data problems.
- The Hadoop ecosystem refers to the various components of the Apache Hadoop software library, as well as to the accessories and tools provided by the Apache software foundation for these types of software projects and to the ways that they work together.
- Hadoop is a Java-based framework that is extremely popular for handling and analysing large sets of data. The idea of a Hadoop ecosystem involves the use of different parts of the core Hadoop set such as MapReduce, a framework for handling vast amounts of data and the Hadoop Distributed File System (HDFS), a sophisticated file-handling system. There is also YARN, a Hadoop resource manager.
- In addition to these core elements of Hadoop, Apache has also delivered other kinds of accessories or complementary tools for developers.
- Some of the most well-known tools of the Hadoop ecosystem include HDFS, Hive, Pig, YARN, MapReduce, Spark, HBase, Oozie, Sqoop, Zookeeper, etc.

- Fig. 6.6.1 shows Apache Hadoop ecosystem.

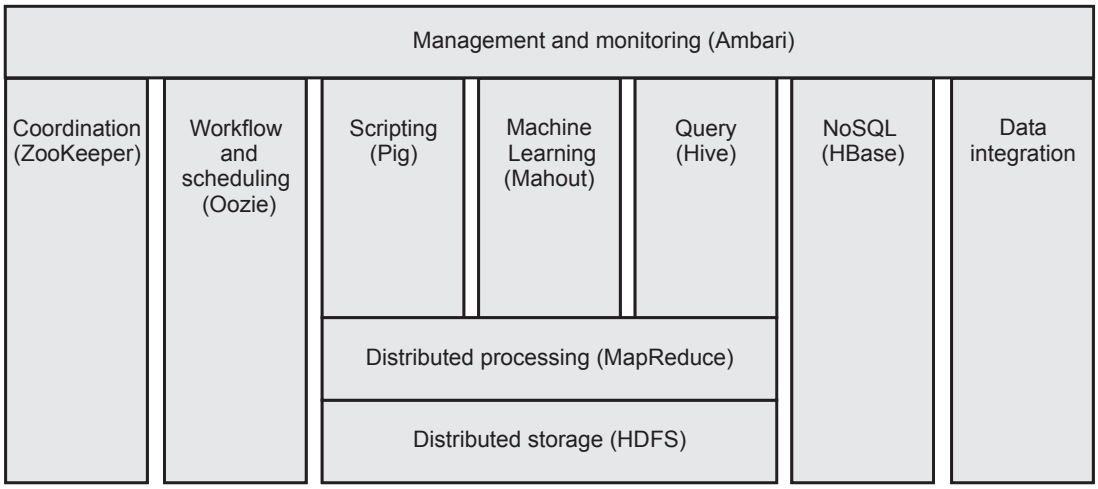


Fig. 6.6.1 : Apache Hadoop ecosystem

- Hadoop Distributed File System (HDFS), is one of the largest Apache projects and primary storage system of Hadoop. It employs a NameNode and DataNode architecture. It is a distributed file system able to store large files running over the cluster of commodity hardware.
- YARN stands for Yet Another Resource Negotiator. It is one of the core components in open source Apache Hadoop suitable for resource management. It is responsible for managing workloads, monitoring and security controls implementation.
- Hive is an ETL and data warehousing tool used to query or analyze large datasets stored within the Hadoop ecosystem. Hive has three main functions : Data summarization, query and analysis of unstructured and semi-structured data in Hadoop.
- Map-Reduce : It is the core component of processing in a Hadoop ecosystem as it provides the logic of processing. In other words, Map-Reduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside the Hadoop environment.
- Apache Pig is a high-level scripting language used to execute queries for larger datasets that are used within Hadoop.
- Apache Spark is a fast, in-memory data processing engine suitable for use in a wide range of circumstances. Spark can be deployed in several ways, it features Java, Python, Scala, and R programming languages and supports SQL, streaming

data, machine learning and graph processing, which can be used together in an application.

- Apache HBase is a Hadoop ecosystem component which is a distributed database that was designed to store structured data in tables that could have billions of rows and millions of columns. HBase is a scalable, distributed, and NoSQL database that is built on top of HDFS. HBase provide real-time access to read or write data in HDFS.

6.6.1 Hadoop Architecture

- Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It provides a software framework for distributed processing of large datasets in real-time applications.
- Hadoop manages to process and store vast amounts of data by using interconnected affordable commodity hardware. Hundreds or even thousands of low-cost dedicated servers working together to store and process data within a single ecosystem.
- Hadoop provides the basic platform for big data processing. The Hadoop architecture has mainly two parts : Hadoop Distributed File System (HDFS) and the MapReduce engine.
- Fig. 6.6.2 shows HDFS archtiecture

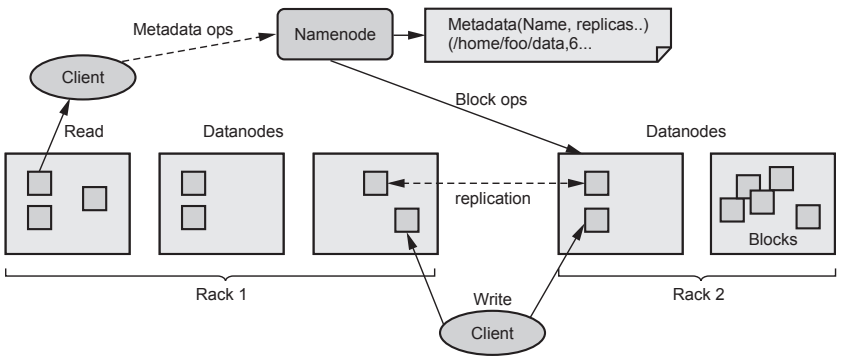


Fig. 6.6.2 Hadoop architecture

- Hadoop distributed file system is a block-structured file system where each file is divided into blocks of a pre-determined size. These blocks are stored across a cluster of one or several machines.

- Apache Hadoop HDFS architecture follows a master/slave architecture, where a cluster comprises of a single NameNode (Master node) and all the other nodes are DataNodes (Slave nodes).
- DataNodes process and store data blocks, while NameNodes manage the many DataNodes, maintain data block metadata and control client access.

1. NameNode and DataNode

- Namenode holds the meta data for the HDFS like Namespace information, block information etc. When in use, all this information is stored in main memory. But this information also stored in disk for persistence storage.
- Namenode manages the file system namespace. It keeps the directory tree of all files in the file system and metadata about files and directories.
- DataNode is a slave node in HDFS that stores the actual data as instructed by the NameNode. In brief, NameNode controls and manages a single or multiple data nodes.
- DataNode serves to read or write requests. It also creates, deletes and replicates blocks on the instructions from the NameNode.
- Fig. 6.6.3 shows Namenode. It shows how NameNode stores information on disk.

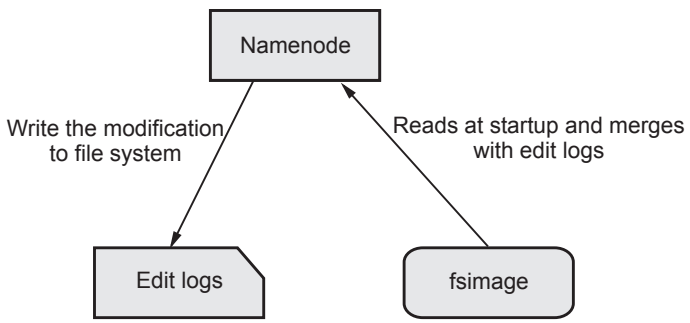


Fig. 6.6.3 Name node

- Two different files are :
  1. **fsimage :** It's the snapshot of the file system when name node started.
  2. **Edit logs :** It's the sequence of changes made to the file system after name node started.
- Only in the restart of namenode, edit logs are applied to fsimage to get the latest snapshot of the file system.

- But namenode restart are rare in production clusters which means edit logs can grow very large for the clusters where namenode runs for a long period of time.
- The following issues we will encounter in this situation :
  1. Editlog become very large, which will be challenging to manage it.
  2. Namenode restart takes long time because lot of changes to be merged.
  3. In the case of crash, we will lost huge amount of metadata since fsimage is very old.
- So to overcome this issues we need a mechanism which will help us reduce the edit log size which is manageable and have up to date fsimage, so that load on namenode reduces.
- Secondary Namenode helps to overcome the above issues by taking over responsibility of merging editlogs with fsimage from the namecode.
- Fig. 6.6.4 shows secondary Namenode.

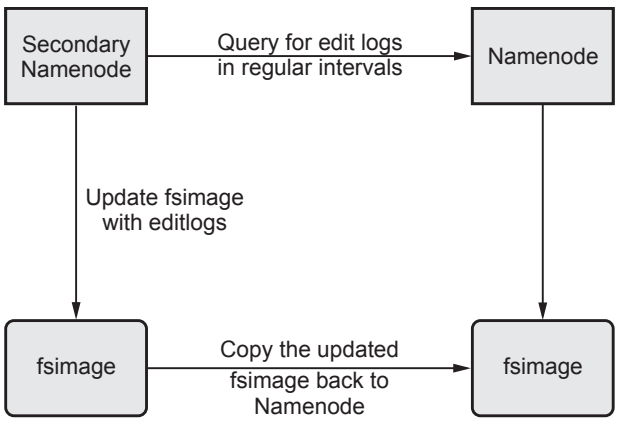


Fig. 6.6.4 Secondary Namenode

- Working of secondary Namenode :
  1. It gets the edit logs from the Namenode in regular intervals and applies of fsimage.
  2. Once it has new fsimage, it copies back to Namenode.
  3. Namenode will use this fsimage for the next restart, which will reduce the startup time.
- Secondary Namenode's whole purpose is to have a checkpoint in HDFS. Its just a helper node for Namecode. That's why it also known as checkpoint node inside the community.

Hadoop Distributed File System :

- Hadoop Distributed File System (HDFS) is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds of nodes.
- A block is the minimum amount of data that it can read or write. HDFS blocks are 128 MB by default and this is configurable. When a file is saved in HDFS, the file is broken into smaller chunks or "blocks".
- HDFS is a fault-tolerant and resilient system, meaning it prevents a failure in a node from affecting the overall system's health and allows for recovery from failure too. In order to achieve this, data stored in HDFS is automatically replicated across different nodes.
- HDFS supports a traditional hierarchical file organization. A user or an application can create directories and store files inside these directories. The file system namespace hierarchy is similar to most other existing file systems; one can create and remove files, move a file from one directory to another, or rename a file.
- Hadoop Distributed File System is a block-structured file system where each file is divided into blocks of a pre-determined size. These blocks are stored across a cluster of one or several machines.
- Apache Hadoop HDFS Architecture follows a Master/Slave Architecture, where a cluster comprises of a single NameNode (MasterNode) and all the other nodes are DataNodes (Slave nodes).
- HDFS can be deployed on a broad spectrum of machines that support Java. Though one can run several DataNodes on a single machine, but in the practical world, these DataNodes are spread across various machines

6.6.2 MapReduce

- MapReduce is a programming model and software framework first developed by Google. Intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.
- Characteristics of MapReduce :
  1. Very large scale data : peta, exa bytes
  2. Write once and read many data. It allows for parallelism without mutexes
  3. Map and reduce are the main operations : simple code
  4. All the maps should be completed before reduced operation starts

5. Map and reduce operations are typically performed by the same physical processor
6. Number of map tasks and reduced tasks are configurable.

- MapReduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside the Hadoop environment.
- In a MapReduce program, Map() and Reduce() are two functions.
  1. The Map function performs actions like filtering, grouping and sorting.
  2. While the reduce function aggregates and summarizes the result produced by the map function.
  3. The result generated by the Map function is a key value pair (K, V) which acts as the input for Reduce function.
- MapReduce works by breaking the processing into two phases :
  1. Map phase 2. Reduce phase
- Each phase has key-value pairs as input and output. In addition the programmer also specifies two functions: map function and reduce function.
- Map function takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
- Reduce function takes the output from the Map as an input and combines those data tuples based on the key and accordingly modifies the value of the key.
- Every Map/Reduce program must specify a Mapper and typically a Reducer. The Mapper has a map method that transforms input (key, value) pairs into any number of intermediate (key', value') pairs. The Reducer has a reduce method that transforms intermediate (key', value'*)aggregates into any number of output (key'', value'') pairs.
- Fig. 6.6.5 shows MapReduce logical data flow.

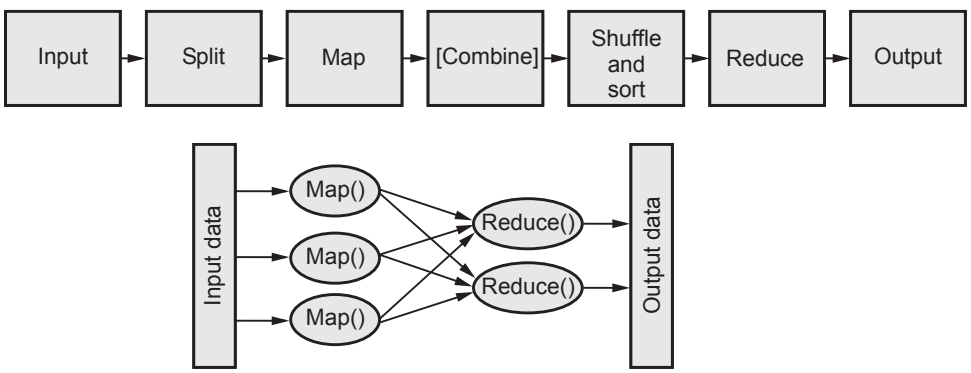


Fig. 6.6.5 Map-Reduce logical data flow

1. **Input :** This is the input data / file to be processed.
2. **Split :** Hadoop splits the incoming data into smaller pieces called "splits".
3. **Map :** In this step, MapReduce processes each split according to the logic defined in map() function. Each mapper works on each split at a time. Each mapper is treated as a task and multiple tasks are executed across different TaskTrackers and coordinated by the JobTracker.
4. **Combine :** This is an optional step and is used to improve the performance by reducing the amount of data transferred across the network. Combiner is the same as the reduce step and is used for aggregating the output of the map() function before it is passed to the subsequent steps.
5. **Shuffle and Sort :** In this step, outputs from all the mappers are shuffled, sorted to put them in order and grouped before sending them to the next step.
6. **Reduce :** This step is used to aggregate the outputs of mappers using the reduce() function. Output of reducer is sent to the next and final step. Each reducer is treated as a task and multiple tasks are executed across different TaskTrackers and coordinated by the JobTracker.
7. **Output :** Finally the output of reduce step is written to a file in HDFS.

- Map tasks write their output to the local disk, not to HDFS. Map output is intermediate output : It's processed by reducing tasks to produce the final output and once the job is complete, the map output can be thrown away. So, storing it in HDFS with replication would be overkill. If the node running the map task fails before the map output has been consumed by the reduced task then Hadoop will automatically rerun the map task on another node to re-create the map output.
- The map tasks partition their output, each creating one partition for each reduce task. There can be many keys and their associated values in each partition, but the records for any given key are all in a single partition. The partitioning can be controlled by a user-defined partitioning function, but normally the default partitioner, which buckets keys using a hash function, works very well.
- The partitions are sorted and transferred across the network to the node where the respective reduce task is running, where they are merged and passed to the user-defined reduce function.
- Consider an **ecommerce system** that receives a million requests every day to process payments. There may be several exceptions thrown during these requests such as "payment declined by a payment gateway," "out of inventory," and "invalid address."

- A developer wants to analyze last four days' logs to understand which exception is thrown how many times.

1. Map

- Let's assume that the Hadoop framework runs just four mappers. Mapper 1, Mapper 2, Mapper 3 and Mapper 4.
- The value input to the mapper is one record of the log file. The key could be a text string such as "file name + line number." The mapper, then, processes each record of the log file to produce key value pairs. Here, we will just use a filler for the value as '1.' The output from the mappers look like this :

```
Mapper 1 -> <Exception A, 1>, <Exception B, 1>, <Exception A, 1>, <Exception C, 1>,
<Exception A, 1>
Mapper 2 -> <Exception B, 1>, <Exception B, 1>, <Exception A, 1>, <Exception A, 1>
Mapper 3 -> <Exception A, 1>, <Exception C, 1>, <Exception A, 1>, <Exception B, 1>,
<Exception A, 1>
Mapper 4 -> <Exception B, 1>, <Exception C, 1>, <Exception C, 1>, <Exception A, 1>
```

- Assuming that there is a combiner running on each mapper - Combiner 1 … Combiner 4 - that calculates the count of each exception (which is the same function as the reducer), the input to Combiner 1 will be :

```
<Exception A, 1>, <Exception B, 1>, <Exception A, 1>, <Exception C, 1>,
<Exception A, 1>
```

2. Combine : The output of Combiner 1 will be :

```
<Exception A, 3>, <Exception B, 1>, <Exception C, 1>
```

- The output from the other combiners will be :

```
Combiner 2: <Exception A, 2> <Exception B, 2>
Combiner 3: <Exception A, 3> <Exception B, 1> <Exception C, 1>
Combiner 4: <Exception A, 1> <Exception B, 1> <Exception C, 2>
```

3. Partition : After this, the partitioner allocates the data from the combiners to the reducers. The data is also sorted for the reducer.

- The input to the reducers will be as below :

```
Reducer 1: <Exception A> {3,2,3,1}
Reducer 2: <Exception B> {1,2,1,1}
Reducer 3: <Exception C> {1,1,2}
```

- If there were no combiners involved, the input to the reducers will be as below :

```
Reducer 1: <Exception A> {1,1,1,1,1,1,1,1,1}
Reducer 2: <Exception B> {1,1,1,1,1}
Reducer 3: <Exception C> {1,1,1,1}
```

- Now, each reducer just calculates the total count of the exceptions as :

```
Reducer 1: <Exception A, 9>
Reducer 2: <Exception B, 5>
Reducer 3: <Exception C, 4>
```

- The data shows that exception A is thrown more often than others and requires more attention. When there are more than a few weeks or months of data to be processed together, the potential of the MapReduce program can be truly exploited.
- With HDFS, we are able to distribute the data so that data is stored on hundreds of nodes instead of a single large machine. Mapreduce provides the framework for highly parallel processing of data across clusters of commodity hardware.
  Fig. 6.6.6 shows MapReduce data processing.

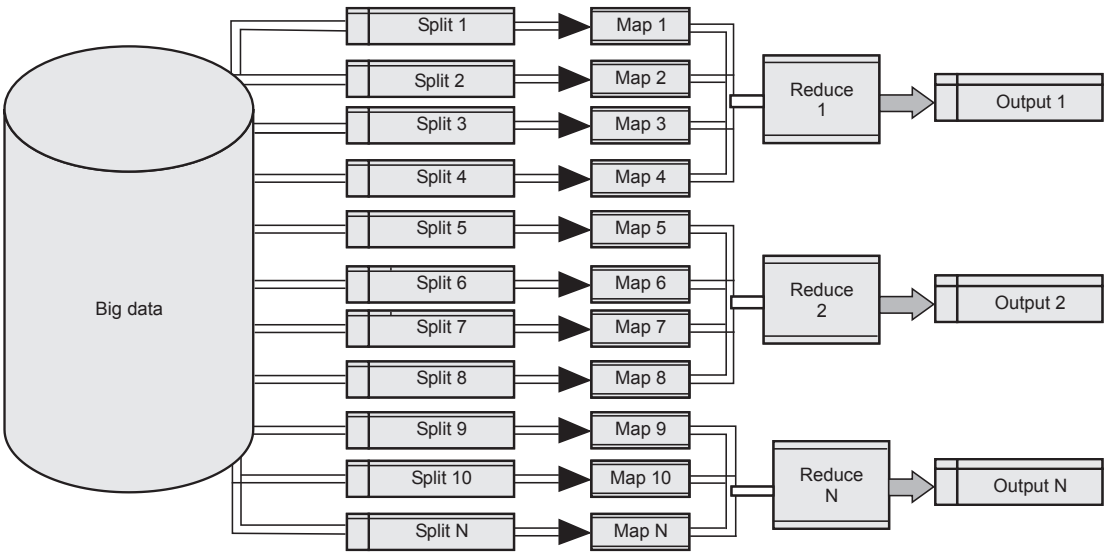


Fig. 6.6.6 Map-Reduce data processing

- It removes the complicated programming part from the programmers and moves into the framework. Programmers can write simple programs to make use of the parallel processing.
- The framework splits the data into smaller chunks that are processed in parallel on cluster of machines by programs called **mappers.**
- The output from the mappers is then consolidated by reducers into desired result. The share nothing architecture of mappers and reducers make them highly parallel.
- Data locality is achieved by mapreduce by working closely with HDFS. When you specify the file system as HDFS for mapreduce, it automatically schedules the mappers on the same node as where the block of data exists.

- Mapreduce can get the blocks from HDFS and process them. The final output from Mapreduce also can be stored in HDFS file system. However, the intermediate files between mappers and reducers are not stored in HDFS and are stored on the local file system of the mappers.

Function of job tracker :

- There is a single job tracker that runs on the master node. It is the driver for the mapreduce jobs. Its functions are :
  1. Accepts jobs from client and divides into tasks
  2. Schedules tasks on worker nodes called **task trackers**
  3. Keeps heartbeat info from task trackers on worker nodes
  4. Reschedules the task on alternate worker if a worker fails.

Function of task tracker :

- Task tracker runs on each worker node and there are as many task trackers as the worker nodes. If HDFS is also used, then data nodes of HDFS also become worker nodes for task tracker. The functions of a task tracker are :
  a. Takes assignments from job tracker
  b. Executes the tasks locally
  c. Each worker node has specific number of mapper and reducer tasks it can take at one time
  d. The tasks assigned are run in parallel
  e. Normally they can take more map jobs than reduce tasks
  f. Task tracker does a task attempt before executing task
  g. Task tracker may do multiple attempts before declaring a task as failed.
  h. Task tracker maintains a connection with the task attempt called **umbilical** protocol
  i. Task tracker sends a regular heartbeat signal to job tracker indicating its status including available map and reduce tasks
  j. Task tracker runs each task attempt in a separate JVM. So even if the task has bad code due to which it fails, it will not cause task tracker to abort.

- Hadoop configs are contained under /etc/hadoop/conf in CDH

Sr. No.	Name of File	Description
1.	hadoop-env.sh	• Used for environment-specific settings • It update the JAVA path to configure user JAVA_HOME. • It also specify JVM options for various Hadoop components
2.	core-site.xml	• System-level Hadoop configuration items, such as the HDFS URL, • It configure the Hadoop temporary directory and script locations for rack-aware Hadoop clusters
3.	hdfs-site.xml	• Used for HDFS settings such as File replication count, the block size, permissions
4.	mapred-site.xml	• Hadoop distributed file settings i.e. no. of reduce tasks, memory sizes
5.	Masters	• List of hosts that are Hadoop masters, i.e. secondary name nodes
6.	Slaves	• List of set of hosts that are going to act as slaves

- The default settings for above configuration are available at
  http://hadoop.apache.org/common/docs/r1.0.0/core-default.html

6.6.3 Pig

- Pig is an open-source high level data flow system. A high-level platform for creating MapReduce programs used in Hadoop. It translates into efficient sequences of one or more MapReduce jobs.
- Pig offers a high-level language to write data analysis programs which we call as Pig Latin. The salient property of pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.
- Pig makes use of both, the Hadoop Distributed File System as well as the MapReduce.

Features of Pig Hadoop :

1. In-built operators : Apache Pig provides a very good set of operators for performing several data operations like sort, join, filter, etc.
2. Ease of programming.
3. Automatic optimization : The tasks in Apache Pig are automatically optimized.
4. Handles all kinds of data : Apache Pig can analyze both structured and unstructured data and store the results in HDFS.

- Fig. 6.6.7 shows Pig architecture.

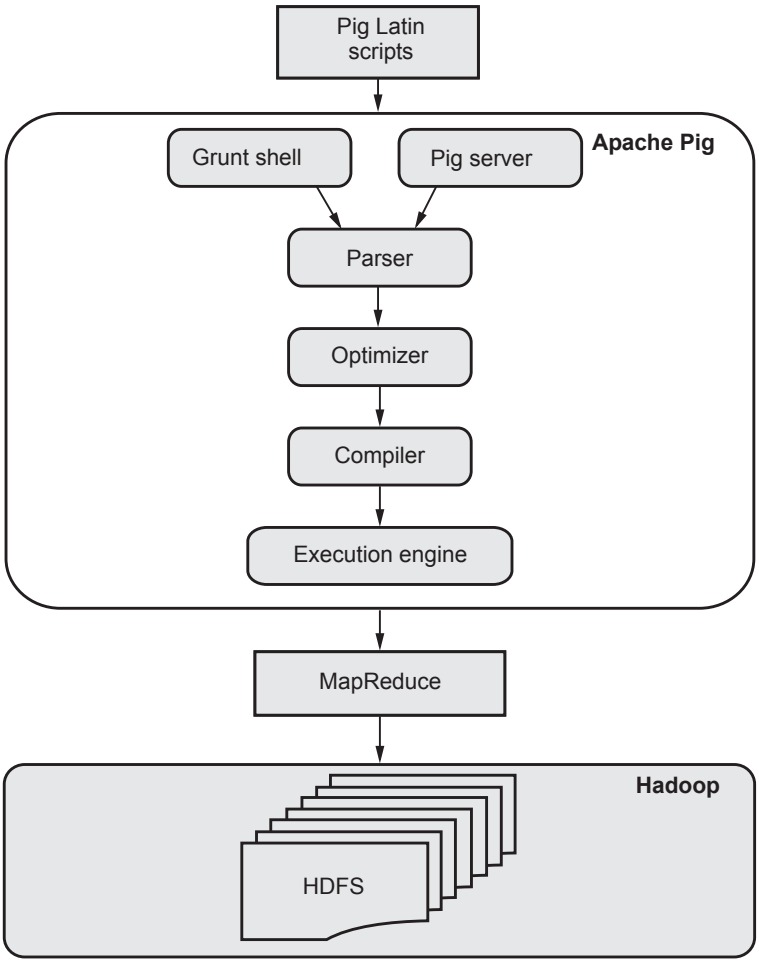


Fig. 6.6.7 Pig architecture

- Pig has two execution modes :

1. **Local mode :** To run pig in local mode, we need access to a single machine; all files are installed and run using local host and file system. Specify local mode using the -x flag (pig-x local).

2. **Mapreduce mode :** To run pig in mapreduce mode, we need access to a Hadoop cluster and HDFS installation. Mapreduce mode is the default mode; but don't need to, specify it using the -x flag

- Pig Hadoop framework has four main components :

1. **Parser :** When a Pig Latin script is sent to Hadoop Pig, it is first handled by the parser. The parser is responsible for checking the syntax of the script, along with

other miscellaneous checks. Parser gives an output in the form of a Directed Acyclic Graph (DAG) that contains Pig Latin statements, together with other logical operators represented as nodes.

2. **Optimizer :** After the output from the parser is retrieved, a logical plan for DAG is passed to a logical optimizer. The optimizer is responsible for carrying out the logical optimizations.
3. **Compiler :** The role of the compiler comes in when the output from the optimizer is received. The compiler compiles the logical plan sent by the optimize the logical plan is then converted into a series of MapReduce tasks or jobs.
4. **Execution Engine :** After the logical plan is converted to MapReduce jobs, these jobs are sent to Hadoop in a properly sorted order and these jobs are executed on Hadoop for yielding the desired result.

- Pig can run on two types of environments : The local environment in a single JVM or the distributed environment on a Hadoop cluster.
- Pig has variety of scalar data types and standard data processing options. Pig supports Map data; a map being a set of key-value pairs.
- Most pig operators take a relation as an input and give a relation as the output. It allows normal arithmetic operations and relational operations too.
- Pig's language layer currently consists of a textual language called **Pig Latin.** Pig Latin is a data flow language. This means it allows users to describe how data from one or more inputs should be read, processed and then stored to one or more outputs in parallel.
- These data flows can be simple linear flows, or complex workflows that include points where multiple inputs are joined and where data is split into multiple streams to be processed by different operators. To be mathematically precise, a Pig Latin script describes a directed acyclic graph (DAG), where the edges are data flows and the nodes are operators that process the data.
- The first step in a Pig program is to LOAD the data, which we want to manipulate from HDFS. Then run the data through a set of transformations. Finally, DUMP the data to the screen or STORE the results in a file somewhere.

Advantages of Pig :

1. Fast execution that works with MapReduce, Spark and Tez.
2. Its ability to process almost any amount of data, regardless of size.
3. A strong documentation process that helps new users learn Pig Latin.

4. Local and remote interoperability that lets professionals work from anywhere with a reliable connection.

Pig disadvantages :

1. Slow start-up and clean-up of MapReduce jobs
2. Not suitable for interactive OLAP analytics
3. Complex applications may require many user defined function.

6.6.4 Hive

- Apache Hive is an open source data warehouse software for reading, writing and managing large data set files that are stored directly in either the Apache Hadoop Distributed File System (HDFS) or other data storage systems such as Apache HBase.
- Data analysts often use Hive to analyze data, query large amounts of unstructured data and generate data summaries.
- Features of Hive :
  1. It stores schema in a database and processes data into HDFS.
  2. It is designed for OLAP.
  3. It provides SQL type language for querying called HiveQL or HQL.
  4. It is familiar, fast, scalable and extensible.
- Hive supports variety of storage formats : TEXTFILE for plaintext, SEQUENCEFILE for binary key-value pairs, RCFILE stores columns of a table in a record columnar format
- Hive table structure consists of rows and columns. The rows typically correspond to some record, transaction, or particular entity detail.
- The values of the corresponding columns represent the various attributes or characteristics for each row.
- Hadoop and its ecosystem are used to apply some structure to unstructured data. Therefore, if a table structure is an appropriate way to view the restructured data, Hive may be a good tool to use.
- Following are some Hive use cases :
  1. Exploratory or ad-hoc analysis of HDFS data : Data can be queried, transformed and exported to analytical tools.
  2. Extracts or data feeds to reporting systems, dashboards, or data repositories such as HBase.

3. Combining external structured data to data already residing in HDFS.

Advantages :

1. Simple querying for anyone already familiar with SQL.
2. Its ability to connect with a variety of relational databases, including Postgres and MySQL.
3. Simplifies working with large amounts of data.

Disadvantages :

1. Updating data is complicated
2. No real time access to data
3. High latency.

- **Program Example :** Write a code in JAVA for a simple Word Count application that counts the number of occurrences of each word in a given input set using the Hadoop Map-Reduce framework on local-standalone set-up.

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {
  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    public void map(Object key, Text value, Context context )
       throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
      }
    }
  }
  public static class IntSumReducer
       extends Reducer<Text,IntWritable,Text,IntWritable> {
```

```
  private IntWritable result = new IntWritable();
  public void reduce(Text key, Iterable<IntWritable> values,
                     Context context
                     ) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values) {
      sum += val.get();
    }
    result.set(sum);
    context.write(key, result);
  }
 }
 public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  Job job = Job.getInstance(conf, "word count");
  job.setJarByClass(WordCount.class);
  job.setMapperClass(TokenizerMapper.class);
  job.setCombinerClass(IntSumReducer.class);
  job.setReducerClass(IntSumReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);
  FileInputFormat.addInputPath(job, new Path(args[0]));
  FileOutputFormat.setOutputPath(job, new Path(args[1]));
  System.exit(job.waitForCompletion(true) ? 0 : 1);
 }
}
```

6.6.5 Difference between Pig and Hive

Sr. No.	Pig	Hive
1.	Pig used for data transformations and processing.	Hive used for warehousing and querying data.
2.	Pig works on structured, semi-structured and unstructured data.	Hive works only on structured data.
3.	Pig does not support web interface.	Hive support web interface.
4.	Pig is a scripting platform that runs on Hadoop clusters, designed to process and analyze large datasets. Pig uses a language called Pig Latin, which is similar to SQL.	Hive is a data warehouse system used to query and analyze large datasets stored in HDFS. Hive uses a query language called HiveQL, which is similar to SQL.
5.	Pig support Avro file format.	Hive does not support Avro file format.
6.	Creating schema is not required to store data in Pig.	Hive supports schema.

7.	Pig loads data quickly.	Hive takes time to load but executes quickly.
8.	Pig works on the client-side of the cluster.	Hive works on the server-side of the cluster.
9.	Used for programming.	Used for reporting.

6.6.6 HBase

- HBase is an open source, non-relational, distributed database modeled after Google's BigTable. HBase is an open source and sorted map data built on Hadoop. It is column oriented and horizontally scalable.
- It is a part of the Hadoop ecosystem that provides random real-time read/write access to data in the Hadoop file system. It runs on top of Hadoop and HDFS, providing Big Table-like capabilities for Hadoop.
- HBase supports massively parallelized processing via MapReduce for using HBase as both source and sink.
- HBase supports an easy-to-use Java API for programmatic access. It also supports Thrift and REST for non-Java front-ends.
- HBase is a column oriented distributed database in Hadoop environment. It can store massive amounts of data from terabytes to petabytes. HBase is scalable, distributed big data storage on top of the Hadoop eco system.
- The HBase physical architecture consists of servers in a Master-Slave relationship. Typically, the HBase cluster has one Master node, called HMaster and multiple Region Servers called HRegionServer. Fig. 6.6.8 shows Hbase architecture.

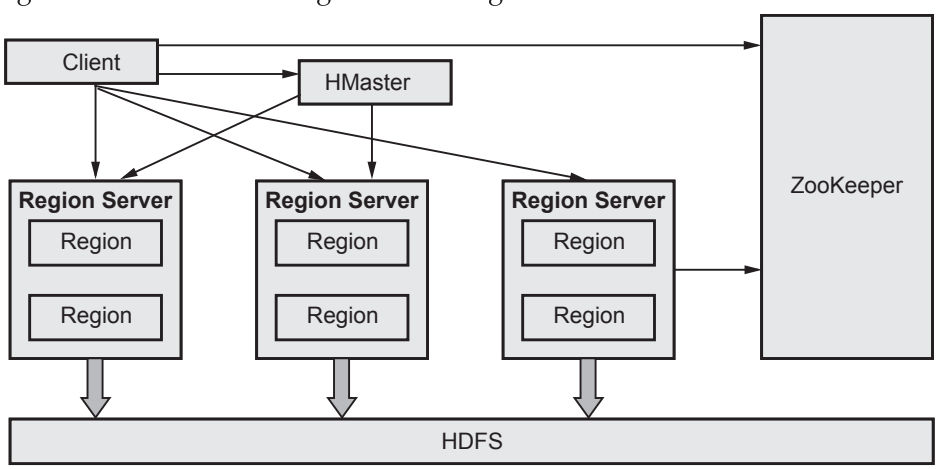


Fig. 6.6.8 Hbase architecture

- Zookeeper is a centralized monitoring server which maintains configuration information and provides distributed synchronization. If the client wants to communicate with regions servers, client has to approach Zookeeper.
- HMaster in the master server of Hbase and it coordinates the HBase cluster. HMaster is responsible for the administrative operations of the cluster.
- HRegions servers : It will perform the following functions in communication with HMaster and Zookeeper.
  1. Hosting and managing regions.
  2. Splitting regions automatically.
  3. Handling read and writes requests.
  4. Communicating with clients directly
- HRegions : For each column family, HRegions maintain a store. Main components of HRegions are Memstore and Hfile.
- Data model in HBase is designed to accommodate semi-structured data that could vary in field size, data type and columns.
- HBase is a column-oriented, non-relational database. This means that data is stored in individual columns and indexed by a unique row key. This architecture allows for rapid retrieval of individual rows and columns and efficient scans over individual columns within a table.
- Both data and requests are distributed across all servers in an HBase cluster, allowing user to query results on petabytes of data within milliseconds. HBase is most effectively used to store non-relational data, accessed via the HBase API.

6.6.7 Difference between HDFS and HBase

Sr. No.	HDFS	HBase
1.	HDFS is a distributed file system suitable for storing large files.	HBase is a database built on top of the HDFS.
2.	HDFS does not support fast individual record lookups.	HBase provides fast lookups for larger tables.
3.	It provides high latency batch processing; no concept of batch processing.	It provides low latency access to single rows from billions of records (Random access).
4.	It provides only sequential access of data.	HBase internally uses Hash tables and provides random access and it stores the data in indexed HDFS files for faster lookups.

5.	HDFS are suited for high latency operations.	HBase is suited for low latency operations.
6.	In HDFS, data are primarily accessed through Map Reduce jobs.	HBase provides access to single rows from billions of records.
7.	HDFS doesn't have the concept of random read and write operations.	HBase data is accessed through shell commands, client API in Java, REST, Avro or Thrift.

6.6.8 Mahout

- Mahout is an open source machine learning library from Apache written in java. It also supports a number of clustering algorithms like k-means, mean-shift and canopy.
- The primitive features of Apache Mahout include :
  1. The algorithms of Mahout are written on top of Hadoop, so it works well in distributed environment.
  2. Mahout uses the Apache Hadoop library to scale effectively in the cloud.
  3. Mahout offers the coder a ready-to-use framework for doing data mining tasks on large volumes of data.
  4. Mahout lets applications to analyze large sets of data effectively and in quick time
  5. Includes several MapReduce enabled clustering implementations such as k-means, fuzzy k-means, Canopy etc.
  6. Supports Distributed Naive Bayes and Complementary Naïve Bayes classification implementations.
  7. Comes with distributed fitness function capabilities for evolutionary programming.
  8. Includes matrix and vector libraries.
- Mahout is an open source machine learning library built on top of Hadoop to provide distributed analytics capabilities. Mahout incorporates a wide range of data mining techniques including collaborative filtering, classification and clustering algorithms.

Review Questions

1. *Explain MapReduce paradigm with example.* **SPPU : Dec.-18 (End Sem), Marks 6**
2. *Explain Hadoop distributed file system.* **SPPU : Dec.-18 (End Sem), Marks 5**
3. *Explain the Hadoop Ecosystem in detail with Pig, Hive, HBase and Mahout.* **SPPU : Dec.-18 (End Sem), Marks 8**
4. *Explain working of Apache Hadoop with HDFS and MapReduce.* **SPPU : Dec.-19 (End Sem), Marks 9**
5. *Explain following terms : i) Pig ii) Hive iii) HBase iv) Mahout* **SPPU : Dec.-19 (End Sem), Marks 8**
6. *What is Map-Reduce ? Explain working of Map-Reduce with example.* **SPPU : May-19 (End Sem), Marks 9**
7. *Explain HDFS with respect to NameNode, DataNodes, Secondary NameNode with example.* **SPPU : May-19 (End Sem), Marks 8**

6.7 Multiple Choice Questions

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

a two three b three

c four d six

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

a Geometric b Pixel oriencted

c Circle segments d None

Q.3 List categorization of visualization methods.

a Pixel-oriented visulization techniques.

b Geometric visualization techniques

c Icon-based visulization projection technique

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 circule

Q.6 Mahout is an open-source machine learning library from Apache written in _____.

a C b C++

c Python d Java

Q.7 HBase is a _____ , non-relational database.

a row-oriented b column-oriented

c horizontal d vertical

Q.8 Pig support _____ file format.

a mp3 b jpeg

c doc d Avro

Q.9 Pig works on the _____ of the cluster

a server-side b master node

c client-side d none

Q.10 Hive works on the _____ of the cluster.

a server-side b master node

c client-side d none

Q.11 MapReduce is a programming model and software framework first developed by _____.

a Microsoft b Amazon

c TCS d Google

Q.12 HDFS blocks are _____ MB by default and this is configurable.

a 32 b 64

c 128 d 256

Q.13 Apache Hadoop HDFS architecture follows a _____ architecture.

a client/server b master/slave

c peer to peer d all of these

Q.14 Hive is _____ and data warehousing tool used to query or analyze large datasets stored within the Hadoop ecosystem.

a STL b HDFS

c ETL d data mining

Q.15 Hadoop ecosystem include _____.

a Hive b Pig

c YARN d all of these

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	b	**Q.8**	d
Q.9	c	**Q.10**	a
Q.11	d	**Q.12**	c
Q.13	b	**Q.14**	c
Q.15	d		

❑❑❑

SOLVED MODEL QUESTION PAPER (In Sem)
Data Science and Big Data Analytics
T.E. (Computer) Semester - VI (As Per 2019 Pattern)

Time : 1 Hour] **[Maximum Marks : 30**

N. B. :

i) Attempt Q.1 or Q.2, Q.3 or Q.4.

ii) Neat diagrams must be drawn wherever necessary.

iii) Figures to the right side indicate full marks.

iv) Assume suitable data, if necessary.

Q.1 **a)** *What is data reduction ? Explain its strategies.* **(Refer section 1.10)** **[3]**

b) *What is big data ? Explain 3V's of big data.* **(Refer section 1.3)** **[4]**

c) *Explain data integration and transformation.* **(Refer section 1.9)** **[8]**

OR

Q.2 **a)** *What is data discretization ?* **(Refer section 1.11)** **[3]**

b) *Explain data science life cycle.* **(Refer section 1.5)** **[5]**

c) *What is data wrangling ? Explain process of data wrangling.* **(Refer section 1.7)** **[7]**

Q.3 **a)** *What is Bayes theorem ?* **(Refer section 2.4)** **[4]**

b) *Explain difference between null hypothesis and alternative hypothesis.* **(Refer section 2.5.3)** **[4]**

c) *What is Chi-square test ? List its characteristics ? Explain Chi -sqaure test for independence of attributes.* **(Refer section 2.7)** **[7]**

OR

Q.4 **a)** *What is need of statistics in data science and big data analytics ?* **(Refer section 2.1)** **[3]**

b) *Explain Wilcoxon Rank - sum test* **(Refer section 2.8.1)** **[5]**

c) *Explain various measures of central tendency.* **(Refer section 2.2)** **[7]**

SOLVED MODEL QUESTION PAPER (End Sem)

Data Science and Big Data Analytics

T.E. (Computer) Semester - VI (As Per 2019 Pattern)

Time : $2\frac{1}{2}$ Hours] **[Maximum Marks : 70**

N. B. :

i) Attempt Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.

ii) Neat diagrams must be drawn wherever necessary.

iii) Figures to the right side indicate full marks.

iv) Assume suitable data, if necessary.

Q.1 a) *What is big data ? Explain big data ecosystem.* **(Refer section 3.1)** **[8]**

b) *Explain data analytics life cycle.* **(Refer section 3.3)** **[10]**

OR

Q.2 a) *What is analytics sandbox ? Explain.* **(Refer section 3.2.4)** **[5]**

b) *Explain data analytics architecture with suitable diagram.* **(Refer section 3.1.4)** **[6]**

c) *What is data repository ? Explain advantages and disadvantages of data repository. Which are the factor responsible for data volume in big data.* **(Refer sections 3.2.1, 3.2.3 and 3.2.5)** **[7]**

Q.3 a) *What is regression ? Explain logistics regression. What is the difference between linear and logistics regression ?* **(Refer section 4.8)** **[8]**

b) *What is decision tree ? Explain how decision tree is constructed using ID3 algorithm.* **(Refer section 4.10)** **[9]**

OR

Q.4 a) *Generate frequent itemsets and generate association rules based on it using apriori algorithm. Minimum support is 50 % and minimum confidence is 70 %.* **(Refer example 4.6.1)** **[8]**

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

b) *What is data pre-processing ? How to remove duplicates ? How it handles missing data value ?* **(Refer section 4.3)** **[9]**

Q.5 **a)** *What is clustering ? Explain hierarchical clustering.* **(Refer section 5.1.6)** **[6]**

b) *What is confusion matrix ? Explain ROC curve.* **(Refer section 5.8)** **[6]**

c) *What is time series analysis ? Explain assumptions of ARIMA model.* **(Refer section 5.2)** **[6]**

OR

Q.6 **a)** *What is social network analysis ? How to develop social network analysis ?* **(Refer section 5.4)** **[6]**

b) *Explain random sampling.* **(Refer section 5.6.3)** **[6]**

c) *Explain various text pre-processing techniques.* **(Refer sections 5.3.2)** **[6]**

Q.7 **a)** *Explain the following data visualization techniques :* **(Refer section 6.3)** **[9]**

a. Venn diagram b. Line graph c. Pie chart

b) *Explain Hadoop ecosystem in details.* **(Refer section 6.6)** **[8]**

OR

Q.8 **a)** *What is data visualization ? Explain challenges to big data visualization.* **(Refer section 6.1)** **[8]**

b) *Explain the following data visualization tools :* **(Refer section 6.5)** **[9]**

a. Pentaho b. Datameter c. Tableau

❑❑❑

TEXT BOOKS FOR T.E. (COMP) SEM VI

Compulsory Subjects

1. Web Technology *(A. A. Puntambekar)*
2. Data Science and Big Data Analytics *(I. A. Dhotre, Dr. Kalpana V. Metre)*
3. Artificial Intelligence *(Anamitra Deshmukh-Nimbalkar, Dr. Vaishali P. Vikhe)*

Elective Subjects

4. Information Security *(I. A. Dhotre, Dr. Swati Nikam)*
5. Augmented and Virtual Reality *(Dr. Ninad More, Sunita Patil)*
6. Cloud Computing *(I. A. Dhotre)*
7. Software Modeling and Architecture *(A. A. Puntambekar)*