UNIT I

1 Introduction : Data Science and Big Data

Syllabus

Introduction to Data science and Big Data, *Defining Data science and Big Data, Big Data examples,* ***Data Explosion :*** *Data Volume, Data Variety, Data Velocity and Veracity. Big data infrastructure and challenges* ***Big Data Processing Architectures :*** *Data Warehouse, Re-Engineering the Data Warehouse, shared everything and shared nothing architecture, Big data learning approaches.* ***Data Science*** *- The Big Picture : Relation between AI, Statistical Learning, Machine Learning, Data Mining and Big Data Analytics.*

Contents

1.1 Introduction to Data Science

SPPU : April-20

- Data is a collection of facts and figures which relay something specific, but which are not organized in any way. It can be numbers, words, measurements, observations or even just descriptions of things. We can say, data is raw material in the production of information.
- Types of data are record data, data matrix, document data, transaction data, graph data and ordered data.
- Data science is an interdisciplinary field that seeks to extract knowledge or insights from various forms of data. At its core, data science aims to discover and extract actionable knowledge from data that can be used to make sound business decisions and predictions.
- Data science uses advanced analytical theory and various methods such as time series analysis for predicting the future. From historical data, instead of knowing how many products sold in the previous quarter, data science helps in forecasting future product sales and revenue more accurately.
- Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information and make business decisions. Data science uses complex machine learning algorithms to build predictive models.
- Data science enables businesses to process huge amounts of structured and unstructured big data to detect patterns.

1.1.1 Applications of Data Science

- Asking a personal assistant like Alexa or Siri for a recommendation demands data science. So does operating a self-driving car, using a search engine that provides useful results or talking to a chatbot for customer service. These are all real-life applications for data science.
- Following are some main reasons for using data science technology :
  - With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
  - Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, which handle the huge amount of data, are using data science algorithms for better customer experience.
  - Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.
- Data science can help in different predictions such as various surveys, elections, flight ticket confirmation, etc.

1. Healthcare : Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.
2. Gaming : Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.
3. Image recognition : Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.
4. Logistics : Data science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.
5. Predict future market trends : Collecting and analyzing data on a larger scale can enable you to identify emerging trends in your market. Tracking purchase data, celebrities and influencers and search engine queries can reveal what products people are interested in.
6. Recommendation systems : Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase or browse on their platforms.
7. Streamline manufacturing : Another way you can use data science in business is to identify inefficiencies in manufacturing processes. Manufacturing machines gather data from production processes at high volumes. In cases where the volume of data collected is too high for a human to manually analyze it, an algorithm can be written to clean, sort and interpret it quickly and accurately to gather insights.

1.1.2 Relationship between Data Science and Information Science

- Data science, as the interdisciplinary field, employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science and computer science. Data science and information science are twin disciplines by nature. The mission, task and nature of data science are consistent with those of information science.
- Data science is heavy on computer science and mathematics. Information science is used in areas such as knowledge management, data management and interaction design.
- Information science is the science and practice dealing with the effective collection, storage, retrieval and use of information. It is concerned with recordable information and knowledge and the technologies and related services that facilitate their management and use.

1.1.3 Business Intelligence versus Data Science

Business Intelligent (BI)	Data Science
BI tends to provide reports, dashboards, and queries on business questions for the current period or in the past.	Data science tends to use disaggregated data in a more forward-looking, exploratory way, focusing on analyzing the present and enabling informed decisions about the future.
BI systems make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understand how much of a given product was sold in a prior quarter or year.	Data science tends to be more exploratory in nature and may use scenario optimization to deal with more open-ended questions.
BI helps monitor the current state of business data to understand the historical performance of a business.	Data science, as used in business, is basically data-driven, where many interdisciplinary sciences are applied together to extract meaning.
BI is designed to handle static and highly structured data.	Data science can handle high-speed, high-volume and complex, multi-structured data from a wide variety of data sources.

1.1.4 Data Science Life Cycle

- A data science life cycle is an iterative set of data science steps you take to deliver a project or analysis. Fig 1.1.1 shows data science life cycle.

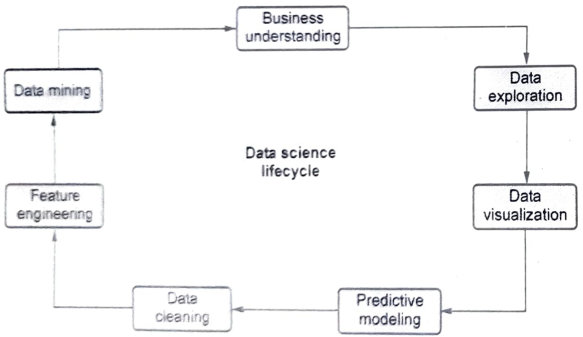

Fig. 1.1.1 Data science life cycle

a) Business understanding : Understand the basic problem you are trying to solve.

b) Data exploration : Understand the pattern and bias in your data.

c) Data visualization : Create and study of the visual representation of data.

d) Predictive modeling : It is the stage where the machine learning finally comes into your data.

e) Data cleaning : Detecting and correcting corrupt or inaccurate records.

f) Feature engineering : It is the process of cutting down the features.

g) Data mining : Gathering your data from different source.

Review Question

1. *Explain data science and its various applications.* SPPU : April-20 (In Sem), Marks 4

1.2 Defining Big Data

SPPU : April-18

- Big data can be defined as very large volumes of data available at various sources, in varying degrees of complexity, generated at different speed i.e., velocities and varying degrees of ambiguity, which cannot be processed using traditional technologies, processing methods, algorithms or any commercial off-the-shelf solutions.
- 'Big data' is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short, such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.
- The processing of big data begins with the raw data that isn't aggregated or organized and is most often impossible to store in the memory of a single computer.
- Big data processing is a set of techniques or programming models to access large-scale data to extract useful information for supporting and providing decisions. Hadoop is the open-source implementation of MapReduce and is widely used for big data processing.

1.2.1 Difference between Data Science and Big Data

Data science	Big data
It is a field of scientific analysis of data in order to solve analytically complex problems and the significant and necessary activity of cleansing, preparing of data.	Big data is storing and processing large volume of structured and unstructured data that can not be possible with traditional applications.
It is used in Biotech, energy, gaming and insurance.	Used in retail, education, healthcare and social media.
Goals : Data classification, anomaly detection, prediction, scoring and ranking.	Goals : To provide better customer service, identifying new revenue opportunities, effective marketing etc.

1.2.2 Benefits of Big Data Processing

Benefits of big data processing :

1. Improved customer service.
2. Business can utilize outside intelligence while taking decisions.
3. Reducing maintenance costs.
4. Re-develop your products : Big data can also help you understand how others perceive your products so that you can adapt them or your marketing, if need be.
5. Early identification of risk to the product / services, if any
6. Better operational efficiency.

1.2.3 Big Data Challenges

- Collecting, storing and processing big data comes with its own set of challenges :
  1. Big data is growing exponentially and existing data management solutions have to be constantly updated to cope with the three Vs.
  2. Organizations do not have enough skilled data professionals who can understand and work with big data and big data tools.

Review Question

1. *Justify your answer with example "Data science and big data are same or different".*

SPPU : April-18 (In Sem), Marks 5

1.3 Data Explosion

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

- The essence of computer applications is to store things in the real world into computer systems in the form of data, i.e., it is a process of producing data. Some data are the records related to culture and society and others are the descriptions of phenomena of the universe and life. The large scale of data is rapidly generated and stored in computer systems, which is called data explosion.
- Data is generated automatically by mobile devices and computers, think facebook, search queries, directions and GPS locations and image capture.
- Sensors also generate volumes of data, including medical data and commerce location-based sensors. Experts expect 55 billion IP - enabled sensors by 2021. Even storage of all this data is expensive. Analysis gets more important and more expensive every year.
- Fig. 1.3.1 shows the big data explosion by the current data boom and how critical it is for us to be able to extract meaning from all of this data.

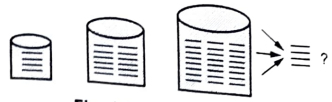

Fig. 1.3.1 Data explosion

- The phenomena of exponential multiplication of data that gets stored is termed as "Data Explosion". Continuous inflow of real-time data from various processes, machinery and manual inputs keeps flooding the storage servers every second.
- Sending emails, making phone calls, collecting information for campaigns; each day we create a massive amount of data just by going about our normal business and this data explosion does not seem to be slowing down. In fact, 90 % of the data that currently exists was created in just the last two years.
- Reason for this data explosion is **Innovation**.
  1. Business model transformation : Innovation changed the way in which we do business, provide services. The data world is governed by three fundamental trends are business model transformation, globalization and personalization of services.
     - Organizations have traditionally treated data as a legal or compliance requirement, supporting limited management reporting requirements. Consequently, organizations have treated data as a cost to be minimized.
     - The businesses are required to produce more data related to product and provide services to cater each sector and channel of customer.
  2. Globalization : Globalization is an emerging trend in business where organizations start operating on an international scale. From manufacturing to customer service, globalization has changed the commerce of the world. Variety and different formats of data are generated due to globalization.
  3. Personalization of services : To enhance customer service, the form of one-to-one marketing in the form of personalization of service is opted by the customer. Customers expect communication through various channels increases the speed of data generation.
  4. New sources of data : The shift to online advertising supported by the likes of Google, Yahoo and others is a key driver in the data boom. Social media, mobile devices, sensor networks and new media are on the fingertips of customers or users. The data generated through this is used by corporations for decision support systems like business intelligence and analytics. The growth of technology helped to emerge new business models over the last decade or more. Integration of all the data across the enterprise is used to create business decision support platform.

1.3.1 V's of Big Data

- We differentiate big data characteristics from traditional data by one or more of the five V's : *Volume, velocity, variety, veracity and value.*

1. **Volume :** Volumes of data are larger than that conventional relational database infrastructure can cope with. It consisting of terabytes or petabytes of data.
   - Fig. 1.3.2 shows big data volume.

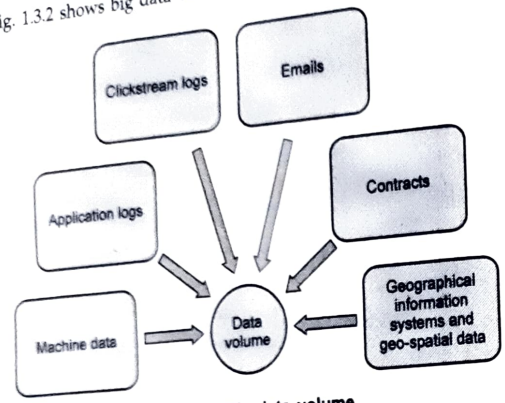

Fig. 1.3.2 Big data volume

2. **Velocity :** The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. It is being created in or near real-time.
3. **Variety :** It refers to heterogeneous sources and the nature of data, both structured and unstructured.
   - Fig. 1.3.3 (a) and Fig. 1.3.3 (b) shows big data velocity and data variety.

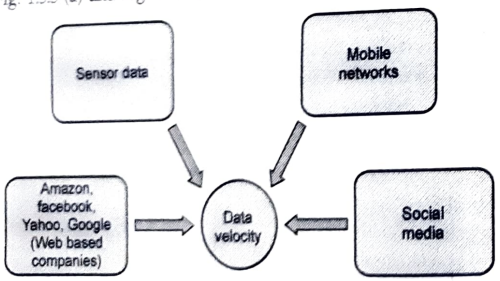

Fig. 1.3.3 (a) Data velocity

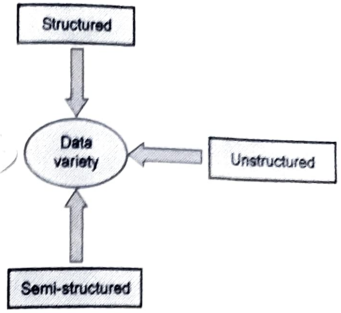

Fig. 1.3.3 (b) Data variety

4. **Value :** It represents the business value to be derived from big data.
   - The ultimate objective of any big data project should be to generate some sort of value for the company doing all the analysis. Otherwise, you're just performing some technological task for technology's sake.
   - For real-time spatial big data, decisions can be enhanced through visualization of dynamic change in such spatial phenomena as climate, traffic, social-media-based attitudes and massive inventory locations.
   - Exploration of data trends can include spatial proximities and relationships. Once spatial big data are structured, formal spatial analytics can be applied, such as spatial autocorrelation, overlays, buffering, spatial cluster techniques and location quotients.
5. **Veracity :** Big data must be fed with relevant and true data. We will not be able to perform useful analytics if much of the incoming data comes from false sources or has errors. Veracity refers to the level of trustiness or messiness of data and if higher the trustiness of the data, then lower the messiness and vice versa. It relates to the assurance of the data's quality, integrity, credibility and accuracy. We must evaluate the data for accuracy before using it for business insights because it is obtained from multiple sources.

1.3.2 Compare Cloud Computing and Big Data

Cloud computing	Big data
It provides resources on demand.	It provides a way to handle huge volumes of data and generate insights.
It refers to internet services from SaaS, PaaS to IaaS.	It refers to data, which can be structured, semi-structured or unstructured.

Cloud is used to store data and information on remote servers.	It is used to describe a huge volume of data and information.
Cloud computing is economical as it has low maintenance costs centralized platform no upfront cost and disaster safe implementation.	Big data is a highly scalable, robust ecosystem and cost-effective.
Vendors and solution providers of cloud computing are Google, Amazon web service, Dell, Microsoft, Apple and IBM.	Vendors and solution providers of big data are Cloudera, Hortonworks, Apache and MapR.
The main focus of cloud computing is to provide computer resources and services with the help of network connection.	Main focus of big data is about solving problems when a huge amount of data generating and processing.

Review Questions

1. *State one example of big data and explain how all V's are applied for big data example.*

SPPU : Dec.-18 (End Sem), Marks 4

2. *Explain 5 V's for defining big data along with the factors responsible for data explosion.*

SPPU : April-19 (In Sem), Marks 5

3. *Explain big data along with 5 V's.*

SPPU : April-18 (In Sem), Marks 5, Dec.-19 (End Sem), Marks 6

1.4 Big Data Examples

- Machine data consists of information generated from industrial equipment real-time data from sensors that track parts and monitor machinery and even web logs that track user behavior online.
- At arcplan client CERN, the largest particle physics research center in the world the Large Hadron Collider (LHC) generates 40 terabytes of data every second during experiments.
- Regarding transactional data, large retailers and even B2B companies can generate multitudes of data on a regular basis considering that their transactions consist of one or many items, product IDs, prices, payment information, manufacturer and distributor data and much more.

Factors responsible for data volume in big data are as follows :

1. Machine data : Machine data contains a definitive record of all activity and behavior of your customers, users, transactions, applications, servers, networks factory machinery and so on. It's configuration data, data from APIs and message queues, change events, the output of diagnostic commands and call detail records, sensor data from remote equipment and more.

2. Application log : Most homegrown and packaged applications write local logfiles, logging services built into application servers like WebLogic, WebSphere and JBoss. These files are critical for day-to-day debugging of production applications by developers and application support. When developers put timing information into their log events, they can also be used to monitor and report on application performance.
3. Business process logs : Complex events processing and business process management system logs are treasure troves of business and IT relevant data. These logs will generally include definitive records of customer activity across multiple channels such as the web, IVR / contact center or retail.
4. Clickstream data : User activity on the Internet is captured in clickstream data. This provides insight into a user's website and web page activity. This information is valuable for usability analysis, marketing and general research.
5. Third party data : The sensitive data that's not in databases is on file systems. In some industries such as healthcare, the biggest data leakage risk is consumer records on shared file systems. Different OS, third-party tools and storage technologies provide different options for auditing read access to sensitive data at the file system level. This audit data is a vital data source for monitoring and investigating access to sensitive data.
6. Electronic mails : Every company have large collection of emails generated by customers, employees and executives on daily basis. These email communication are an important asset to an organization, which are audited case-by-case basis and entire life cycle management of emails is done.

- Some of the examples of big data are :

1. **Social media :** Social media is one of the biggest contributors to the flood of data we have today. Facebook generates around 500 + terabytes of data everyday in the form of content generated by the users like status messages, photos and video uploads, messages, comments etc.
2. **Stock exchange :** Data generated by stock exchanges is also in terabytes per day. Most of this data is the trade data of users and companies.
3. **Aviation industry :** A single jet engine can generate around 10 terabytes of data during a 30 minute flight.
4. **Survey data :** Online or offline surveys conducted on various topics which typically has hundreds and thousands of responses and need to be processed for analysis and visualization by creating a cluster of population and their associated responses.
5. **Compliance data :** Many organizations like healthcare, hospitals, life sciences, finance etc, has to file compliance reports.

1.5 Data Processing Infrastructure Challenges

SPPU : May-18

- Data processing infrastructure challenges are storage, transportation, processing and throughput.

1. Storage : The increase in the volume of data, increases the need for storing the data and processing of data. Big data technology has changed the way we gather and store data, including data storage device, data storage architecture and data access techniques. It requires more sophisticated storage medium with higher I/O speed to meet the challenges of big data issues. Direct-Attached Storage (DAS), Network-Attached Storage (NAS) and Storage Area Network (SAN) are the enterprise storage architecture that are commonly are in use.
2. Transportation : Data is transfer from one place to other place and process the data then load into memory for manipulation. The data is transported between computer and storage layers. Increase in bandwidth is not solution to this problem.
3. Processing : Data processing needs to combine the logic and mathematical computation in one cycle. This processing can be accomplished by CPU or processor, memory and software. With each generation CPU processing speed is increased, have improved processing capabilities. Memory is required for compute and processing. Memory has become cheaper and faster with evolution of processor capability. Software are used to write the programs for transforming and processing of data.
4. Speed and throughout : This is the major challenges for data processing. Various architecture layers like hardware, software's networking and storage are responsible for storage and are added. Each layer has its own limitation, causes limitation in the overall throughput in the data processing.

Review Question

1. List and explain data processing infrastructure challenges in big data.

SPPU : May-18 (End Sem), Marks 6

1.6 Big Data Processing Architectures

SPPU : April-18,19,20, May-18

1.6.1 Data Warehouse

- A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process. A data warehouse stores historical data for purposes of decision support.

- A database is an application-oriented collection of data that is organized, structured, coherent, with minimum and controlled redundancy, which may be accessed by several users in due time.
- Data warehousing provides architectures and tools for business executives to systematically organize, understand and use their data to make strategic decisions.
- A data warehouse is a subject-oriented collection of data that is integrated, time-variant, non-volatile, which may be used to support the decision-making process.
- Data warehouses are databases that store and maintain analytical data separately from transaction-oriented databases for the purpose of decision support.
- The main objective of a data warehouse is to support the decision-making system, focusing on the subjects of the organization. The objective of a database is to support the operational system and information is organized on applications and processes.

Multitier architecture of data warehouse :

- Data warehouse architecture is a data storage framework's design of an organization. A data warehouse architecture takes information from raw sets of data and stores it in a structured and easily digestible format.
- Data warehouse system is constructed in three ways. These approaches are classified the number of tiers in the architecture.
  a) Single-tier architecture.
  b) Two-tier architecture.
  c) Three-tier architecture (Multi-tier architecture).
- *Single tier* warehouse architecture focuses on creating a compact data set and minimizing the amount of data stored. While it is useful for removing redundancies. It is not effective for organizations with large data needs and multiple streams.
- *Two-tier* warehouse structures separate the resources physically available from the warehouse itself. This is most commonly used in small organizations where a server is used as a data mart. While it is more effective at storing and sorting data. Two-tier is not scalable and it supports a minimal number of end-users.

Three tier (Multi-tier) architecture :

- Three tier architecture creates a more structured flow for data from raw sets to actionable insights. It is the most widely used architecture for data warehouse systems.

- Fig. 1.6.1 shows three tier architecture. Three tier architecture sometimes called multi-tier architecture.

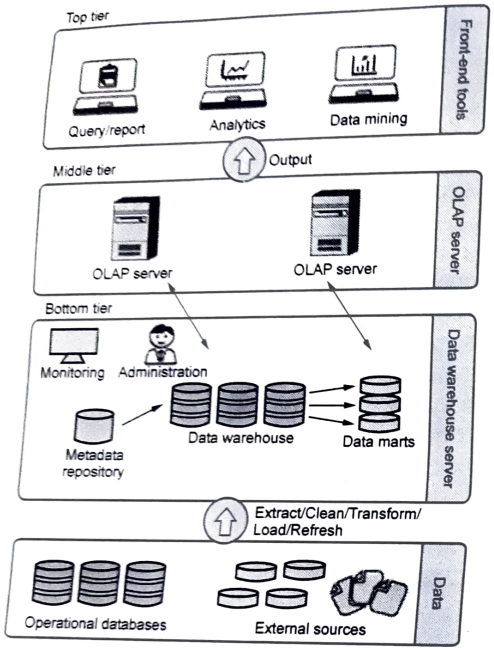

Fig. 1.6.1 Three tier architecture

- The bottom tier is the database of the warehouse, where the cleansed and transformed data is loaded. The bottom tier is a warehouse database server.
- The middle tier is the application layer giving an abstracted view of the database. It arranges the data to make it more suitable for analysis. This is done with an OLAP server, implemented using the ROLAP or MOLAP model.
- OLAPS can interact with both relational databases and multidimensional databases, which lets them collect data better based on broader parameters.
- The top tier is the front-end of an organization's overall business intelligence suite. The top-tier is where the user accesses and interacts with data via queries, data visualizations and data analytics tools.

- The top tier represents the front-end client layer. The client level which includes the tools and Application Programming Interface (API) used for high-level data analysis, inquiring and reporting. User can use reporting tools, query, analysis or data mining tools.

1.6.2 Shared Everything Architecture

- This architectural model consists of nodes that share all resources within the system. Each node has access to the same computing resources and shared storage. Shared-everything architecture refers to system architecture where all resources are shared including storage, memory and the processor.
- Fig. 1.6.2 shows shared everything architecture.

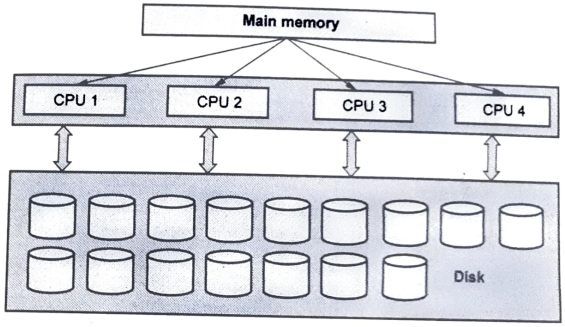

Fig. 1.6.2 Shared everything architecture

- The main idea behind such a system is maximizing resource utilization. The disadvantage is that shared resources also lead to reduced performance due to contention.
- Scalability is the main problem. Oracle RAC uses this architecture.
- Symmetric multiprocessing (SMP) and Distributed Shared Memory (DSM) are the types of shared everything architecture.
- In the SMP architecture, all the CPUs share a single pool of memory for read-write access concurrently and uniformly without latency. Sometimes this is referred to as Uniform Memory Access (UMA) architecture.

- The DSM architecture addresses the scalability problem by providing multiple pools of memory for processors to use. In the DSM architecture, the latency to access memory depends on the relative distances of the processors and their dedicated memory pools. This architecture is also referred to as Nonuniform Memory Access (NUMA) architecture.

1.6.3 Shared Nothing Architecture

- Shared nothing architecture is a distributed computing architecture that consists of multiple separated nodes that don't share resources. The nodes are independent and self-sufficient as they have their own disk space and memory.
- Each node has its own private memory (M), processor (CPUs) and storage devices independent of any other node in the configuration. This means that every node stores its own lock table and buffer pool.
- Fig. 1.6.3 shows shared nothing architecture.

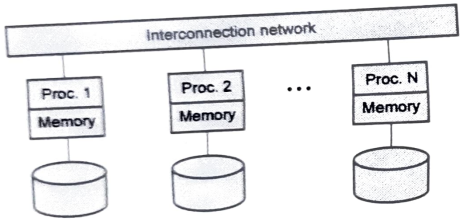

Fig. 1.6.3 Shared nothing architecture

- The key feature of shared-nothing architecture is that the operating system, not the application server, owns responsibility for controlling and sharing hardware resources. Each node is under the control of its own copy of the operating system and thus can be viewed as a local site.
- Shared nothing is also known as Massively Parallel Processing (MPP) solutions typically employed by large data warehouse systems.
- Data is horizontally partitioned across nodes, such that each node has a subset of the rows from each table that was distributed and all the replicated tables.
- Shared nothing can be made to scale to hundreds or even thousands of machines. Because of this, it is generally regarded as the best-scaling architecture. Shared-nothing architecture scales better and is well suited for a cloud data warehouse considering very low-cost commodity PCs and networking hardware.

1.6.4 Re - engineering the Data Warehouse

- Re-engineering the data warehouse means building a next generation data warehouse. Fig 1.6.4 shows re-engineering the data warehouse.

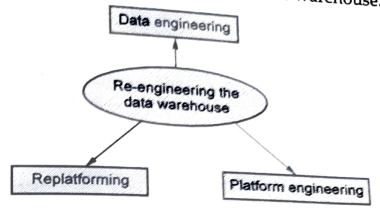

Fig. 1.6.4 Re-engineering the data warehouse

- Various methods of re-engineerings are replatforming, platform engineering and data engineering.
  1. **Replatforming** : Replatforming means, new infrastructure and hardware. Depending on organizations requirement, new technologies such as warehouse appliances, tiered storage, private cloud can be deployed.
- **Advantages** : Scalability, reliability, security, lower maintenance, code optimization.
- **Disadvantages** : It is time consuming and leads to disruption of business activities.
  2. **Data engineering** : It is re-engineering of data structures for creating better performance. There is a change in initial data model of data warehouse to make new data model. It includes partitioning the tables in vertical or horizontal partition, colocation of related tables in same storage region, distribution of data, adding new data types and adding new database functions for performance boost.
  3. **Platform engineering** : It is related to modifying some parts of the infrastructure, which helps to gain better scalability and improved performance.

     Platform engineering is popular concept in automotive industry as product parts are crafted to offer improved quality, service and cost.

Review Questions

1. *Explain the role of shared everything and shared nothing architecture in big data.* **SPPU : April-18 (In Sem), Marks 5**
2. *What is data warehouse ? Explain design and architecture of data warehouse.* **SPPU : May-18 (End Sem), Marks 6**
3. *Explain shared-everything and shared-nothing architectures in detail with respect to big data.* **SPPU : April-19 (In Sem), Marks 5**

4. *List and explain choices for re-engineering the data warehouse.* **SPPU : April-19 (In Sem), Marks 5**
5. *Discuss the processing complexities associated with the big data.* **SPPU : April-19 (In Sem), Marks 5**
6. *What are the pitfalls of data warehouse ? Why companies are shifting to big data using hadoop.* **SPPU : April-20 (In Sem), Marks 4**
7. *Draw and explain big data processing architecture with technologies used at each of the stage of big data processing.* **SPPU : April-20 (In Sem), Marks 6**

1.7 Big Data Learning Approaches

SPPU : April-18, 20

- Data is a boon for machine learning systems. The more data a system receives, the more it learns to function better for businesses. Hence, using machine learning for big data analytics happens to be a logical step for companies to maximize the potential of big data adoption.
- Big data refers to extremely large sets of structured and unstructured data that cannot be handled with traditional methods. Big data analytics can make sense of the data by uncovering trends and patterns. Machine learning can accelerate this process with the help of decision-making algorithms. It can categorize the incoming data, recognize patterns and translate the data into insights helpful for business operations.
- Machine learning algorithms are useful for collecting, analyzing and integrating data for large organizations. They can be implemented in all elements of big data operation, including data labeling and segmentation, data analytics and scenario simulation.
- Machine Learning (ML) is considered as a very fundamental and vital component of data analytics. In fact ML is predicted to be the main drivers of the big data revolution for obvious reasons for its ability to learn from data and provide with data driven insights, decisions and predictions.
- 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.
- Supervised and unsupervised learning are the different types of machine learning methods.

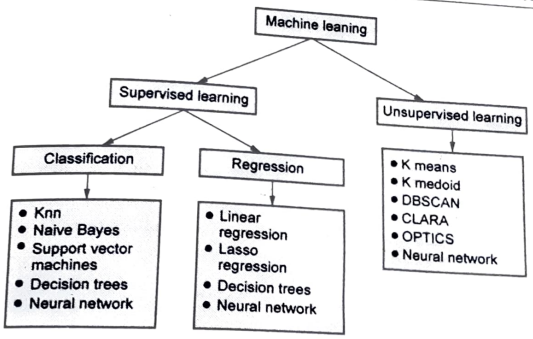

Fig. 1.7.1

- **Supervised learning** is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behavior of a system for any set of input values, after an initial training phase. Supervised learning is also called classification.
- **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. Unsupervised learning is also called clustering.

Review Questions

1. *Enlist the impact of learning approaches in big data ? Explain different kinds of learning approaches.* **SPPU : April-18 (In Sem), Marks 5**
2. *Explain different learning approaches in big data. Explain with example.* **SPPU : April-20 (In Sem), Marks 6**

1.8 Data Science : The Big Picture

SPPU : Dec.-19

- The field of data science is fundamentally about organizing and using data to provide insights for human decision-making. Developing the capability to glean insights from data has become crucial for start-ups and fortune 500 companies alike.
- Many organizations have been collecting unprecedented amounts of data from both physical sensors and the online activity of millions of people but a pile of unorganized data will not yield insights on its own.

- This is where data scientists come into play by cleaning up and organizing data to make it suitable for analysis. They also build the statistical models necessary for analyzing data to reveal notable patterns or trends.
- Several key types of data analysis :
  a) Descriptive analytics aims to gain insight into either current or historical data trends.
  b) Predictive analytics looks to gain insight into future unknowns by using the best available data to make predictions.
  c) Prescriptive analytics can recommend what humans should do given the available data insights.

1.8.1 Relation between AI and Machine Learning

- The emergence of modern AI based on machine learning has significantly boosted predictive and prescriptive analytics.
- AI refers to a set of tools that can automate machine actions in ways that mimic intelligent behaviors. A small sample of such applications might include :
  a) Visually identifying cats and dogs in social media photos.
  b) Translating between different languages in the text found on websites.
  c) Detecting possible signs of cancer in patients' X-ray images.

Machine learning in modern AI systems :

- Most modern AI is based on machine learning : A category of computer algorithms that can automatically learn from data. Instead of relying on humans to program each step, ML models train on large datasets to identify notable patterns within the data and make their own predictions based on that information.
- They can then apply the lessons learned from their training datasets to analyzing completely new and unfamiliar datasets in the real world.
- The importance of the training data means that ML performance depends greatly upon having access to large and diverse datasets of high quality.
- For example, a machine learning model that trains to recognize dogs by only looking at 100 images of Siberian Huskies is unlikely to perform well when suddenly tasked with identifying tens of thousands of images from a diverse array of dog breeds.
- ML models can follow several different approaches :
  a) Supervised learning relies heavily upon hand-labeled training datasets and is the most common type of machine learning.

  b) Unsupervised learning sifts through unlabelled data to try and find unusual patterns that might escape the human eye.
  c) Reinforcement learning uses trial and error to learn from mistakes and get closer to achieving a specific goal.
- Here is just one example of how data science can intersect with AI based on machine learning. Let us assume that an Internet search engine company wants to provide and monetize the most relevant online searches in response to the query : "Allergy medicine for kids."
- Data scientists help collect and organize large datasets containing millions of user search results related to allergy medication for kids. Then, they work with software developers and engineers to build machine learning models that learn from these datasets.
- Through training, machine learning models can identify user preferences for various search results, such as information about what allergy medications come in the form of syrups and chewable tablets. This helps to continuously update the search engine, so that it delivers more relevant results and ranks them higher.
- The search and click trends identified by the machine learning models also provide information about people's medical needs and shopping habits, such as certain allergy medicine brands being more popular among families in a specific geographic area at a certain time of year.
- Data scientists analyze these trends to find business insights that they can share with corporate leaders and online advertisers.

1.8.2 Data Mining and Big Data Analytics

- Data mining refers to extracting or mining knowledge from large amounts of data. It is a process of discovering interesting patterns or knowledge from a large amount of data stored either in databases, data warehouses or other information repositories.
- It is the computational process of discovering patterns in huge data sets involving methods at the intersection of AI, machine learning, statistics and database systems.
- To make predictions, predictive mining tasks perform inference on the current data. Predictive analysis provides answers of the future queries that move across using historical data as the chief principle for decisions.
- It involves the supervised learning functions used for the prediction of the target value. The methods fall under this mining category are the classification, time-series analysis and regression.

- Descriptive analytics is the conventional form of business intelligence and data analysis, seeks to provide a depiction or "summary view" of facts and figures in an understandable format, to either inform or prepare data for further analysis.
- Descriptive analytics helps organizations to understand what happened in the past. It helps to understand the relationship between product and customers.
- Big data analytics is the often complex process of examining big data to uncover information such as hidden patterns, correlations, market trends and customer preferences that can help organizations make informed business decisions.

Review Question

1. *Explain machine learning approaches in big data.* **SPPU : Dec.-19 (End Sem), Marks 6**

1.9 Multiple Choice Questions

Q.1 Three characteristics of big data are _____ _____ _____.

a volume, velocity, variety
b value, variable, variance
c volume, vanish, various
d velocity, volume, vault

Q.2 Data is collection of data objects and their _____.

a information b attributes
c characteristics d none

Q.3 In big data, _____ refer to heterogeneous sources and the nature of data, both structured and unstructured.

a volume b variety
c velocity d all of these

Q.4 Various types of data analytics are _____ .

a descriptive model b predictive model
c prescriptive model d all of these

Q.5 Machine learning is inherently a _____ field.

a interdisciplinary b multidisciplinary
c single d none

Q.6 Symmetric multiprocessing and distributed shared memory are the types of _____ architecture.

a data mining b machine learning
c shared nothing d shared everything

Q.7 Data _____ is characterized by the amount of data that is generated continuously.

a variety b velocity
c volume d all of above

Q.8 _____ architecture is a distributed computing architecture where multiple systems are networked to form a scalable system.

a Shared-nothing b Shared-everything
c Machine learning d Big data

Q.9 The key feature of shared-nothing architecture is that the _____ not the application server owns responsibility for controlling and sharing hardware resources.

a CPU b storage
c operating system d none

Q.10 In the SMP architecture, all the processors share a single pool of _____ for read-write access concurrently and uniformly without latency.

a CPU b memory
c cache d storage

Q.11 DSM architecture is also referred to as _____ architecture.

a UMA b COMA
c NUMA d all of these

Answer Keys for Multiple Choice Questions :

Q.1	a	Q.2	b	Q.3	b	Q.4	d
Q.5	b	Q.6	d	Q.7	c	Q.8	a
Q.9	c	Q.10	b	Q.11	c		

❑❑❑