ISYS5050 Notes

GENERAL INFORMATION

Topic	Type	Marks	Due Date
Assignment	Individual	15%	Week 8
Group Project	Group	35%	Week 13
Final Exam	Individual	50%	Exam Week

EXAM INFORMATION

Topic	Description
Differences between	Provide brief description between given two things such as Semantic graphs and property graphs.
Explanations	Give brief explanations for terminologies such as DW, ETL, Queries, explanation, etc.
Case study	Give solution to the problem of the organization.
Group Project	1 answer will be there where your approach, key findings, etc. was taken and can explain the project.

WEEK 1

Introduction

Data is usually referred to as raw facts. When these data points are connected, we get information. Information is an aggregation of the analysis done on the data. By linking different pieces of information gives us the ability to make informed decisions and thus we create knowledge.

We have different types of data such as unstructured data and structured data. Unstructured data means that it is datasets aren't stored in a structured database format. Structured data is a standardized format for providing information about a page and classifying the page content. Some examples of these are given below:

Unstructured data

- Social Media
- Clickstream data
- Machine-Generated Data (e.g. IoT, logs)
- Internal Documents
- Notes (e.g. Patient Charts)
- Images
- Video
- Sound
- Streaming data
- Geolocation data

Structured data

- Data stored in relational data bases, data warehouses
- E.g. Sales data, customer data, financial transactions, hotel reservation systems
- E.g. Phone numbers, credit card info, addresses, product details

With the advent of social media, the amount of unstructured data collection has increased. Organisations need a way to analyse and manage this unstructured data and convert this into structured data. Following which they can extract information from them, analyse them, and thus tap into the benefits of all the data. The process of data to information is given below:

Knowledge is about how pieces of information can be derived from the collected data and how can we use this information to accomplish the goals set by the organisation. When we uncover relationships that are not explicitly stated as information, we gain deeper insights and turn information into knowledge. Wisdom, is basically the knowledge applied in action. Knowledge management is concerned with facilitating the creation, sharing, combining, transferring, and application of knowledge and the system which allow these are called knowledge management systems. The knowledge management cycle is as follows:

The importance of Knowledge Management System is given below:

- To gain or enhance competitive advantage.
- To improve performance.
- To boost innovation.
- To gain and share insights.
- To make better decisions.
- To connect people looking for knowledge to those who have it.
- To work smarter and more efficiently.

The goals of Knowledge Management Systems is as follows:

- Make knowledge visible.
- Encourage knowledge sharing and transfer.
- Build the knowledge infrastructure.
- Develop a knowledge focused culture.
- Promote knowledge focused and sharing communities.
- Improve the knowledge capture processes.
- Increase access to organisational knowledge.
- Maintain knowledge as an organisational asset.

WEEK 2

Assigned readings: Knowledge Management and Knowledge Management Systems

Knowledge is a broad and abstract notion that has defined epistemological debate in western philosophy since the classical Greek era. In the past few years, however, there has been a growing interest in treating knowledge as a significant organizational resource. To be credible, KMS research and development should preserve and build upon the significant literature that exists in different but related fields. A knowledge-based perspective of the firm has emerged in the strategic management literature.

The knowledge-based perspective postulates that the services rendered by tangible resources. depend on how they are combined and applied, which is in turn a function of the firm's knowledge. This knowledge is embedded in and carried through multiple entities including. organization culture and identity, routines, policies, systems, and documents, as well as individual employees. Because knowledge-based resources are usually difficult to imitate and socially complex, the knowledge- based view of the firm posits that these knowledge

assets may produce long-term sustainable competitive advantage. There are different types of knowledge perspectives and their implications, a few of them are as follows:

Table 1. Knowledge Perspectives and Their Implications

Perspectives		Implications for Knowledge Management (KM)	Implications for Knowledge Management Systems (KMS)
Knowledge vis-à-vis data and information	Data is facts, raw numbers. Information is processed/ interpreted data. Knowledge is personalized information.	KM focuses on exposing individuals to potentially useful information and facilitating assimilation of information	KMS will not appear radically different from existing IS, but will be extended toward helping in user assimilation of information
State of mind	Knowledge is the state of knowing and understanding.	KM involves enhancing individual's learning and understanding through provision of information	Role of IT is to provide access to sources of knowledge rather than knowledge itself
Object	Knowledge is an object to be stored and manipulated.	Key KM issue is building and managing knowledge stocks	Role of IT involves gathering, storing, and transferring knowledge
Process	Knowledge is a process of applying expertise.	KM focus is on knowledge flows and the process of creation, sharing, and distributing knowledge	Role of IT is to provide link among sources of knowledge to create wider breadth and depth of knowledge flows
Access to information	Knowledge is a condition of access to information.	KM focus is organized access to and retrieval of content	Role of IT is to provide effective search and retrieval mechanisms for locating relevant information
Capability	Knowledge is the potential to influence action.	KM is about building core competencies and understanding strategic know-how	Role of IT is to enhance intellectual capital by supporting development of individual and organizational competencies

Knowledge taxonomies and examples:

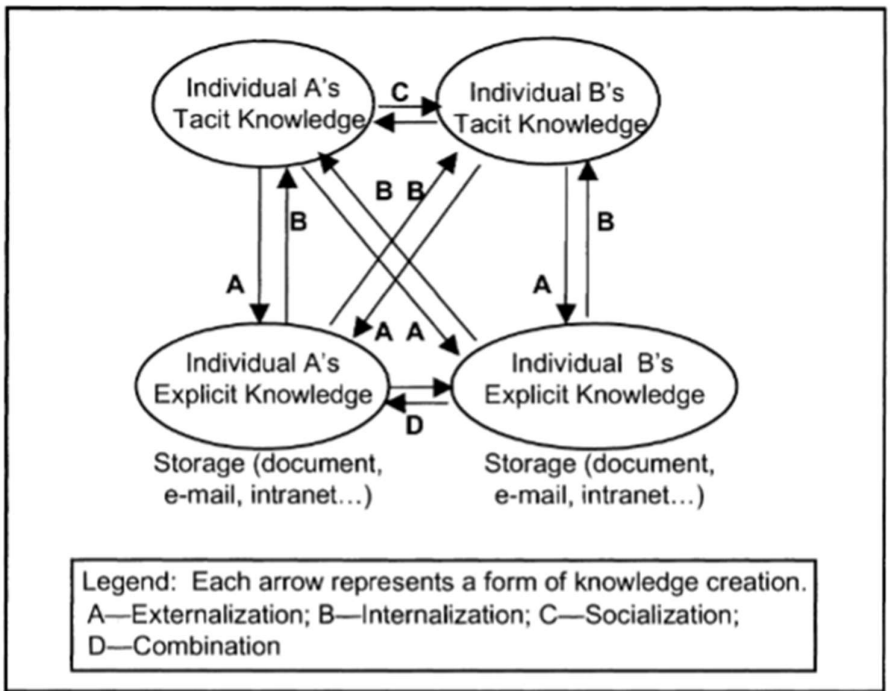
Table 2. Knowledge Taxonomies and Examples		
Knowledge Types	**Definitions**	**Examples**
Tacit	Knowledge is rooted in actions, experience, and involvement in specific context	Best means of dealing with specific customer
Cognitive tacit:	Mental models	Individual's belief on cause-effect relationships
Technical tacit:	Know-how applicable to specific work	Surgery skills
Explicit	Articulated, generalized knowledge	Knowledge of major customers in a region
Individual	Created by and inherent in the individual	Insights gained from completed project
Social	Created by and inherent in collective actions of a group	Norms for inter-group communication
Declarative	Know-about	What drug is appropriate for an illness
Procedural	Know-how	How to administer a particular drug
Causal	Know-why	Understanding why the drug works
Conditional	Know-when	Understanding when to prescribe the drug
Relational	Know-with	Understanding how the drug interacts with other drugs
Pragmatic	Useful knowledge for an organization	Best practices, business frameworks, project experiences, engineering drawings, market reports

Knowledge management systems (KMS) refer to a class of information systems applied to managing organizational knowledge. That is, they are IT-based systems developed to support and enhance the organizational processes of knowledge creation, storage/retrieval, transfer, and application. There are 3 common application of organisational knowledge management:

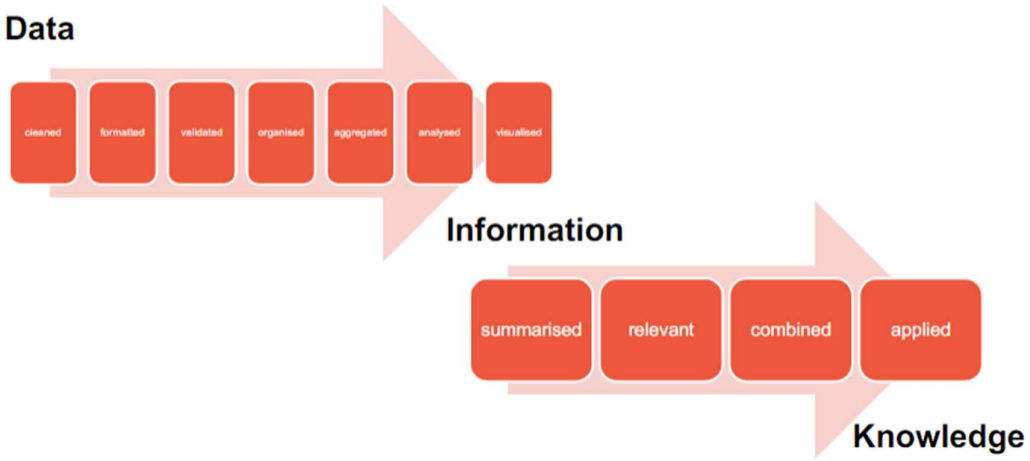
- The coding and sharing of best practices.
- The creation of corporate knowledge directories (mapping of internal expertise).
- The creation of knowledge networks.

Organizational knowledge creation involves developing new content or replacing existing content within the organization's tacit and explicit knowledge. Four modes of knowledge creation have been identified: socialization, externalization, internalization, and combination. The four knowledge creation modes are not pure, but highly interdependent and intertwined. That is, each mode relies on, contributes to, and benefits from other modes. The knowledge creation modes are shown in the following diagram.

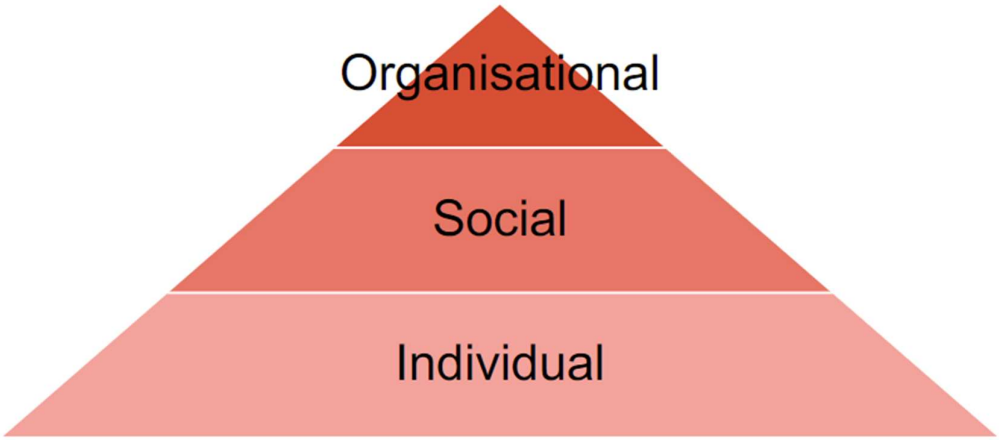
An important aspect of the knowledge-based theory of the firm is that the source of competitive advantage resides in the application of the knowledge rather than in the knowledge itself.



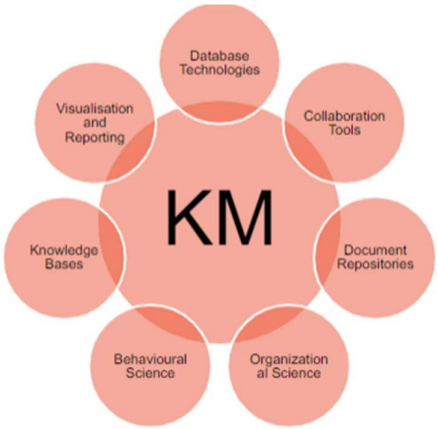
Introduction



Data	Information	Knowledge
Sales data	Average sales in the last 5 years	• Are sales going up or down? • How marketing costs correlate with sales?
Census records	Demographics (e.g. aga or income distribution among the population)	• What is the growth rate of the population? • What will be the average population in 3, 5, 10 year time?
Whether data	Average temperature and rainfall in a city for a period of time	• What is the overall trend of temperature in the country? • Is whether patterns getting more extreme?
Netflix Viewing Activity	Sci-fi and action movies were popular last year	• Which movies would the user watch next? • Which genres are trending?



The knowledge held by an individual, the knowledge held by a group or a team and finally the knowledge hold by an entire organisation comprising of different groups and teams.



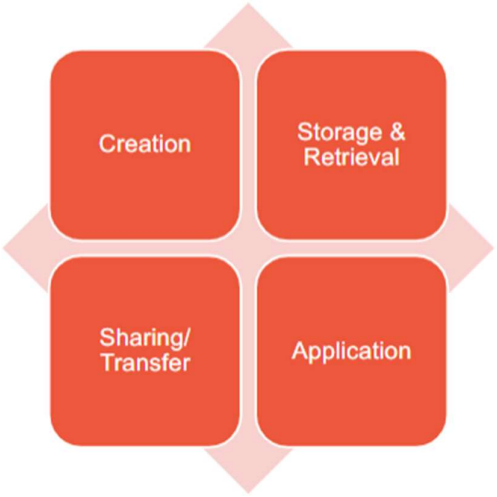
Knowledge Management is multi-disciplinary in nature as seen from the above diagram. It involves database technologies for storage, collaboration tools for communication, document repositories for documentation storage, organisational science, behavioural science, knowledge bases, and much more. The goal of Organisational Knowledge Management is to maintain knowledge as an organisational asset.

Types of Organisational Knowledge

There exists 2 different forms of organisational knowledge. These are Explicit and Tacit. Explicit knowledge is predefined and documented. It is usually structured and stored in some databases. Our main aim for knowledge management systems is to make sure that the knowledge we have in the organisation is explicit in nature. This is because explicit knowledge can be transferred from one person to another quite easily and more efficiently. The difference between explicit and tacit knowledge is given below:

	Explicit Knowledge	Tacit Knowledge
	"Academic knowledge" or "know-what" that is described in formal language, print, or electronic media, often based on established work processes, use people-to-documents approach.	"Practical, action-oriented knowledge" or "know-how" based on practice, acquired by personal experience, seldom expressed openly, often resembles intuition.
Learn	On the job, trial and error, self-directed in areas of greatest expertise; meet work goals and objectives set by organization.	Supervisor or team leader facilitates and reinforces openness and trust to increase sharing of knowledge and business judgment.
Teach	Trainer designed using syllabus, uses formats selected by organization, based on goals and needs of the organization, may be outsourced.	One-on-one, mentor, internships, coach, on-the-job training, apprenticeships, competency based, brainstorm, people to people.
Type of thinking	Logical, based on facts, use proven methods, primarily convergent thinking.	Creative, flexible, unchartered, leads to divergent thinking, develop insights.
Share knowledge	Extract knowledge from person, code, store, and reuse as needed for customers, e-mail, electronic discussions, and forums.	Altruistic sharing, networking, face-to-face contact, videoconferencing, chatting, storytelling, personalize knowledge.

Knowledge Processes



There are 4 stages to knowledge process which are as follows:

- Knowledge creation
  - There are 4 modes of knowledge creation, these are:
    - Tacit to Tacit (Socialisations)
    - Tacit to Explicit (Externalisation)
    - Explicit to Explicit (Combinations)
    - Explicit to Tacit (Internalisation)
- Knowledge storage and retrieval
  - Organisations can learn from past experiences and forget knowledge which might be bad.
  - The process of storage, organisation, and retrieval of organisational knowledge is known as organisational memory.
  - It could be embedded in organisational procedures, routines, policies, manuals, databases, repositories, archives, cultures, and individuals.
- Knowledge sharing/transfer

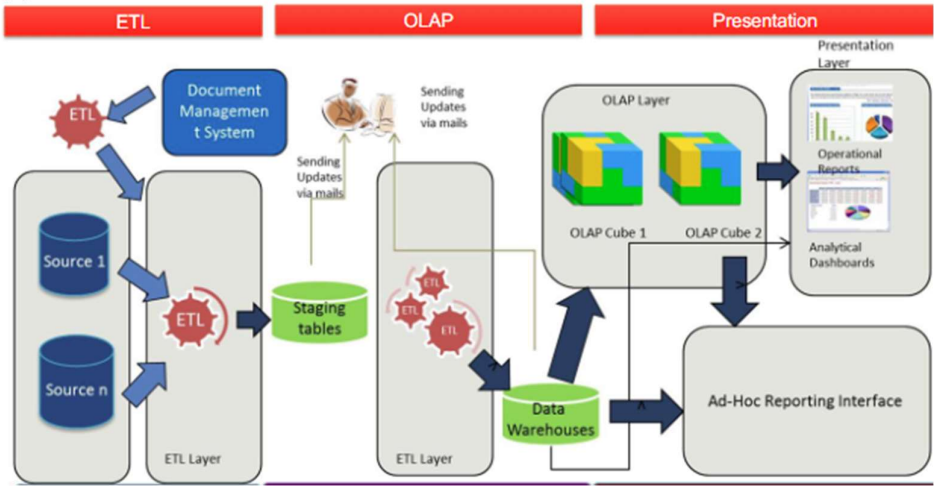
    - Knowledge transfer can take place through any modes of communication but none are as efficient as formal modes of knowledge sharing.
    - There are multiple channels or modes of knowledge sharing, these are:
        - Informal (unscheduled meets, coffee break chats, etc.)
        - Formal (training sessions, guided tours, etc.)
        - Personal (mentorships)
        - Impersonal (Training videos, seminars, etc.)
    - Knowledge transfer depends on the ability of the person who is transferring the knowledge and the capacity of the person who is receiving the knowledge.
- Knowledge Application
    - Competitive advantage of an organisation is in the application of the knowledge rather than in the knowledge itself.
    - There are several methods of application such as:
        - Directives: Rules, standards, procedures, and instructions.
        - Routines: Interaction protocols and process specifications.
        - Specialised task teams: Teams of individuals with the required knowledge and speciality.

Business Intelligence

Business Intelligence is the process of applying the knowledge management systems into our organisations and thus drive business decisions. Business intelligence is a set of techniques and tools for the transformation of raw data into meaningful and useful information to support business decision making. There are various benefits of Business intelligence, a few of them are:

- Improving decision-making process.
- Helps to know the business.
- Reduce the risk of bottlenecks.
- Helps identify the waste in the system.
- Enables real time analysis with quick navigation.
- Makes it easy to access and share information.

An organisation has an abundance of data and this data needs to be made sense of as the data might be messy, need formatting and cleaning up, etc. The main idea behind business intelligence will be to analyse data and present them in a meaningful way as it is easier to communicate and present facts via visuals. Business intelligence can be described in the following phases:



The phases are:

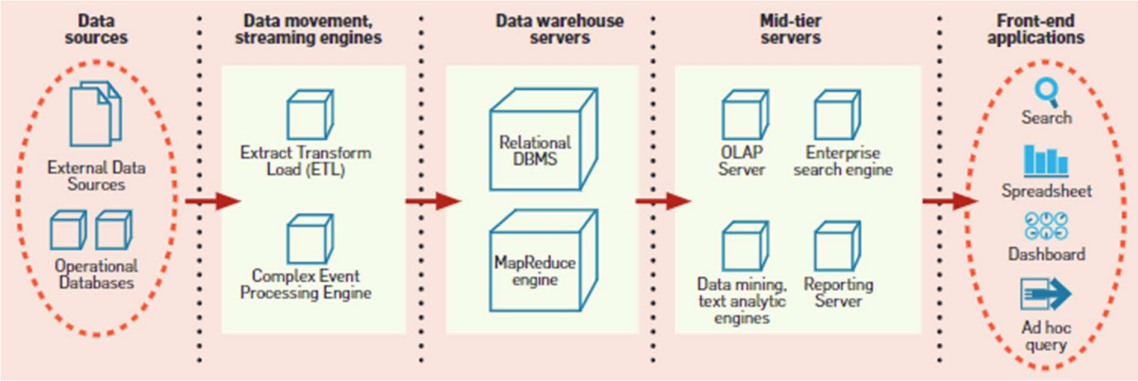
- ETL
  - It stands for Extract (from various sources), Transform (into desired structure) and Load (into a data warehouse).
  - The main idea of this phase is to collect relevant data from different sources and clean them to transform it into a consistent format which can then be used for visualisations.
  - ETL allows migrating data between different databases and applications. A few such ETL tools are:
    - Relational Databases (E.g., Oracle, MySQL, SQL Server, etc.)
    - Cloud Based and SaaS applications (E.g., SAP, Salesforce, etc.)
    - Files (E.g., XML, Excel, CSV, etc.)
- OLAP
  - It stands for Online Analytical Processing step.
  - Here the cleaned data is explored via different methods to create the right settings of visualizations that are accurate and are not overloaded with lots of irrelevant information.
  - It is the process of performing operations on multi-dimensional data cubes for the purpose of analysing and visualizing.
  - Data cubes are representation of data with dimensions and measures. A data cube can be multi-dimensional. The individual facts are called measures.
  - A few operations of OLAP are:
    - Roll up (consolidation): Summarizes data along the dimension.
    - Drill down: Allows one to navigate deeper into the dimensions of the data.
    - Slicing: Enables one to take one level of information for display.
    - Dicing: Enables to select data from multiple dimensions to analyse.
    - Pivot: Gain a new view of data by rotating the data axes of the cube.
- Presentation
  - The analysed data is transformed into interactive visualizations, dashboards, etc.
  - Data visualization is the process of presenting information graphically in which relationships, patterns, similarities, and differences are encoded through shapes, colours, positions, and size.
  - It is necessary to visualize data to communicate the trends, patterns, etc.
  - It is necessary to transform data into information and visualization does it.
  - Visualisations are done to show evidence to the audience.
  - There are quite a few methods to visualize data, a few are:
    - Comparison:
      Comparing values or quantities over time or each category.
    - Proportion
      Displaying individual parts of a whole.
    - Distribution
      Showing possible values of the data and how often they occur.
    - Relationships, Locations and Trends

WEEK 3

Assigned reading: An overview of Business Intelligence Technology

Business Intelligence (Bi) Software is a collection of decision support technologies for the enterprise aimed at enabling knowledge workers such as executives, managers, and analysts to make better and faster decisions. The data over which BI tasks are performed often comes

from different sources typically from multiple operational databases across departments within the organization, as well as external vendors. Different sources contain data of varying quality, use inconsistent representations, codes, and formats, which must be reconciled. Thus the problems of integrating, cleansing, and standardizing data in preparation for BI tasks can be rather challenging. Efficient data loading is imperative for BI. A typical business intelligence architecture is given below:

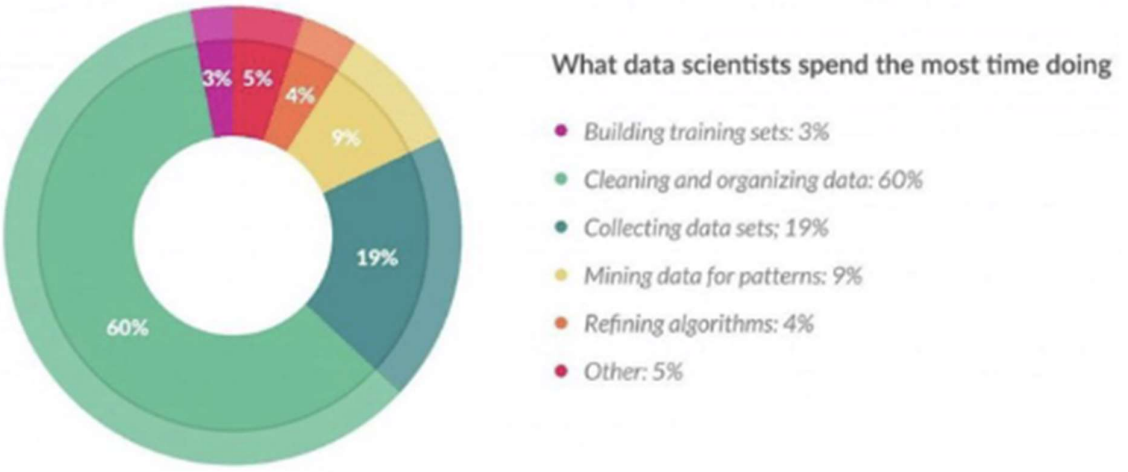


The data over which BI tasks are performed is typically loaded into a repository called the data warehouse that is managed by one or more data warehouse servers. A popular choice of engines for storing and querying warehouse data is relational database management systems (RDBMS). Data warehouse servers are complemented by a set of mid-tier servers that provide specialized functionality for different BI scenarios. Online analytic processing (OLAP) servers efficiently expose the multidimensional view of data to applications or users and enable the common BI operations such as filtering, aggregation, drill-down and pivoting.

Extract-Trans-Form-Load (ETL) refers to a collection of tools that play a crucial role in helping discover and correct data quality issues and efficiently load large volumes of data into the warehouse.

ETL

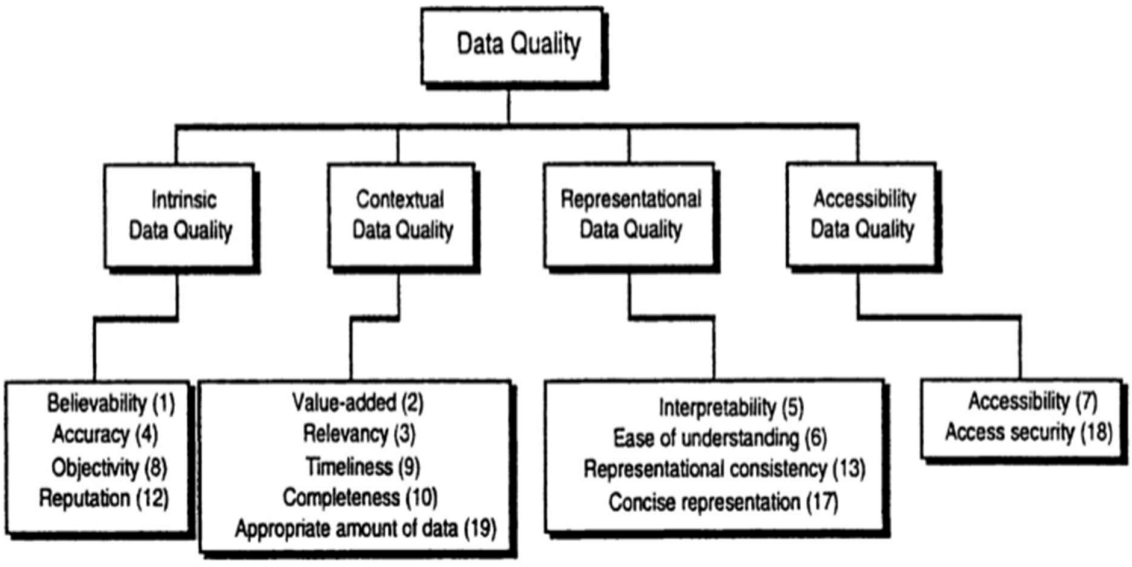
ETL stands for Extract-Transform-Load. Data preparation accounts for about 80% of the work of data scientists. It is the most time-consuming part of the process.



Data comes in all shapes and sizes in different types of file types such as CSV, excel, JSON, PDF, Word, Images, videos, etc. Data comes from anywhere and can be messy. ETL is the process of extracting data from various sources, transforming them into the desired structure and then load the data into a staging environment for future use. ETL is the data quality control step. The process of ETL is iterative and as follows:

The data quality is made up of 4 parts. These are:

- Accuracy: To degree to which data is error free, correct, flawless, and reliable. It relates to credibility, completeness, objectivity, traceability, and reputation of the data.
- Relevancy: The degree to which data is applicable, helpful, relevant, interesting, and usable. It relates to the value added by the data, its timeliness, ease of operation and flexibility.
- Representation: The degree to which data is interpretable and easy to understand and represented concisely and consistently. It relates to the interpretability, ease of understanding, representational consistency, and concise representation.
- Accessibility: The degree to which data is available or easily and quickly retrievable. It relates to the cost effectiveness and access security.



Data quality attributes

Attribute	What it means	Example of good practice	Example of bad practice	Metrics
Consistency	No matter where you look in the database, you won't find any contradictions in your data.	Your payment system shows that Jane Brown has made 5 purchases this month, and CRM system contains the same information.	Your payment system shows that Jane Brown has made 5 purchases this month, while CRM system shows she has made only 4.	The number of inconsistencies.
Accuracy	The information your data contains corresponds to reality.	Your customer's name is Jane Brown. And this is exactly how it's reflected in your CRM.	In your CRM, the customer's name is spelled Jane Brawn, though her actual name is Jane Brown.	The ratio of data to errors.
Completeness	All available elements of the data have found their way to the database.	You know that Jane Brown is born on 11/04/1975.	You have no idea how old Jane Brown is, as the date of birth cell is empty.	The number of missing values.
Auditability	Data is accessible and it's possible to trace introduced changes.	You can track down the changes made in Jane's data record. For example, on 12/5/2018, her phone number was changed.	It's impossible to trace down the changes in Jane's record.	% of cells where the metadata about introduced changes is not accessible.

Data quality attributes

Attribute	What it means	Example of good practice	Example of bad practice	Metrics
Orderliness	The data entered has the required format and structure.	The entry for December 11, 2018 is in the format 12/11/2018.	The entry for December 11, 2018 is in the format 12/11/18, 12/11/2018 and even 11/12/18 (in your European stores).	The ratio of data of inappropriate format.
Uniqueness	A data record with specific details appears only once in the database.	You have only one record for Jane Brown, born on 11/04/1975, who lives in Seattle.	You have multiple duplicate records for Jane Brown.	The number of duplicates revealed.
Timeliness	Data represents reality within a reasonable period of time or in accordance with corporate standards.	On 02/15/2018, the customer informed you that her name is misspelled in the emails you send her. The customer's name was corrected the next day.	On 02/15/2018, the customer informed you that her name is misspelled in the emails you send her. Her name was corrected only in a month.	Number of records with delayed changes.

Data Warehouses

The transformed data from multiple sources through ETL is stored in a common repository for the purpose of supporting data analytics and decision making. This common repository is the data warehouse. Data warehouses are optimized for read and query access rather than daily operations and transactions. Data warehouse is defined in many ways but is basically the tools and techniques for managing data to support business intelligence systems. It is a database that is maintained separately from the organization's operational database. Data warehousing is the process of constructing and using 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.

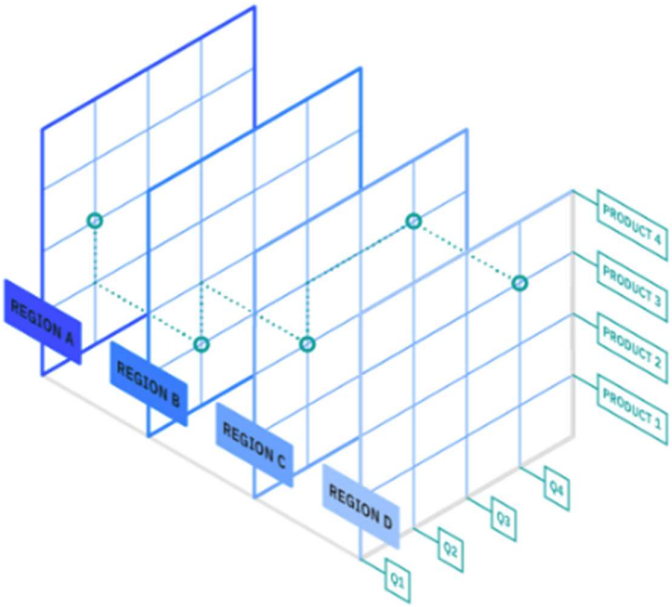
The following are the attributes of a data warehouse:

- It is subject oriented.
    - It is organized around major subjects such as customer, product, sales, etc.
    - It focuses on the modelling and analysis of data for decision makers not on a daily operations or transaction processing.
    - Provide a simple and concise view around a particular subject by excluding data that are not useful in the decision support process.
- It is integrated.
    - It means that the data comes from multiple sources such as relational databases, flat files, etc.
    - Data cleaning and data integration techniques are applied to ensure consistency in naming convention, encoding structures, attribute measures, etc.
    - When data is moved to the warehouse, it is converted into the similar format as that of the data in the warehouse.
- It is time variant.
    - The time horizon for the data warehouse is significantly longer than that of operational systems.
    - Operational databases provide with current data whereas data warehouse data provides information from a historical perspective for the purpose of data analytics and decision support.
    - Every key structure in the data warehouse contains an element of time explicitly or implicitly, but the key in operational data may or may not contain the time element.
- It is non-volatile.
    - It means that a physical separate store of transformed data exists apart from the operational data.
    - Operational data doesn't update data in the warehouse environment as it does not require transaction processing, recovery, and concurrency control mechanisms.
    - It requires only and optimized for 2 operations:
        - Initial loading of the data
        - Access of the data

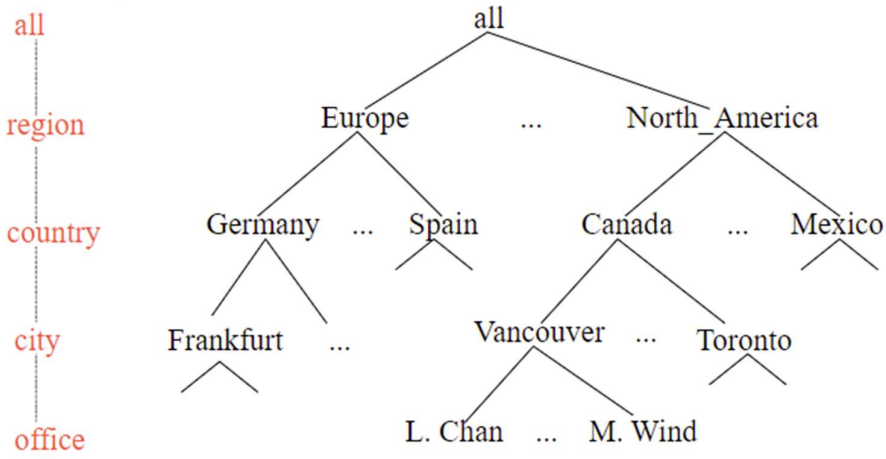
A data warehouse is kept separate as it provides high performance for both systems, DBMS (tuned for transactional processing) and warehouse (tuned for analytical processing). Data warehouse has different functions and a different data all together.

A data warehouse is based on a multi-dimensional data model which views data in the form of a data cube. A data cube can be viewed from multiple dimensions and each fact table

contains measures and keys to each of the related dimensional table. An example of a data cube is given below:

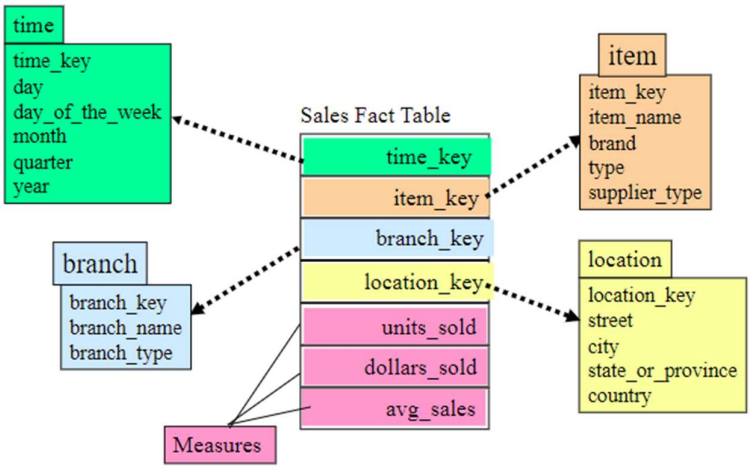


A dimension shows an aspect of the data and is usually categorical in nature. It is single level and has a hierarchy of dimensions such as various levels of one aspect. A measure is an actual quantifiable data in the cube which is also known as facts. This is usually numerical in nature. A data cube is formed by the combination of 2 or more dimensions with one or more measures. An example of hierarchical dimension is given below:

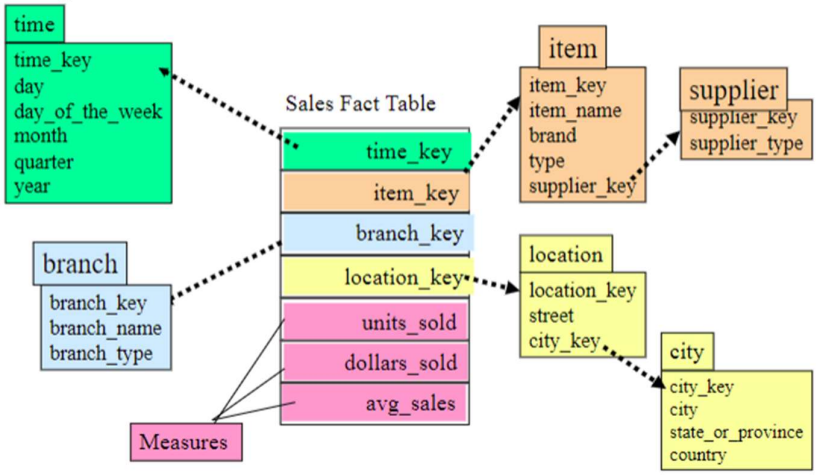


In data cubes we form a data in the data warehouse The data warehouse stores the data in the data cubes using different types of schemas and presentational formats. A few examples of schemas are:

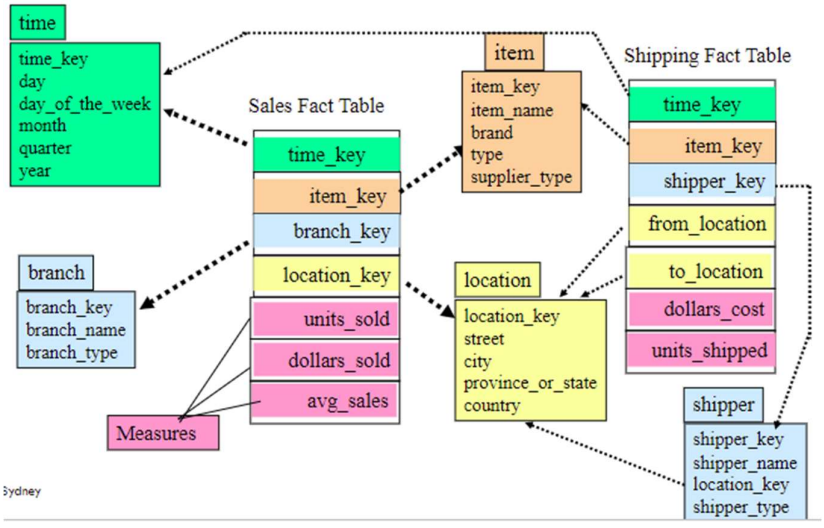
- Star schema
  A fact table in the middle connected to a set of dimension tables just like the example given below:



- Snowflake schema
  It tries to optimize on normalized tables to reduce redundancy. An example of such a schema is given below:



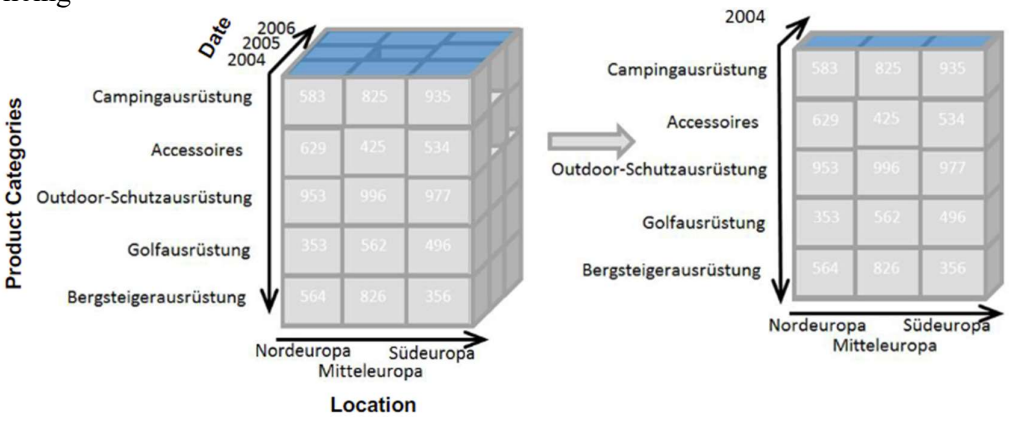
- Fact constellation
  It has multiple fact tables and multiple measures that are related to different sets of dimensions. This normally produces a complex structure. An example of this is given below:



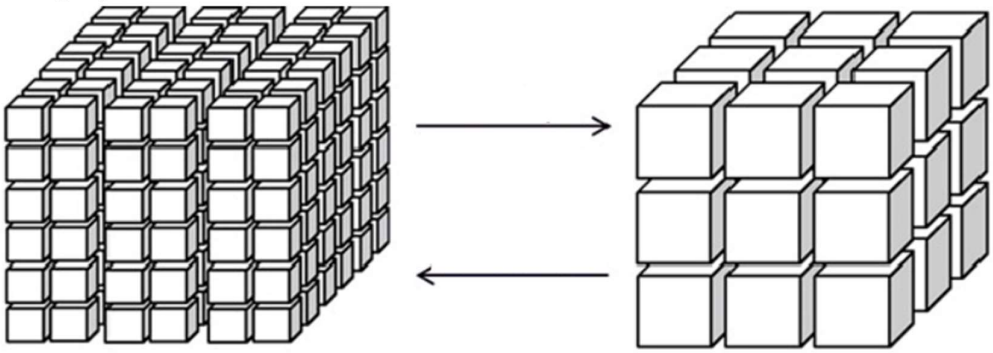
Online Analytical Processing

The process of performing operations on multi-dimensional data cubes stored in a data warehouse for the purpose of analyzing and visualizing data is called as Online Analytical Processing. There are 4 basic operations to OLAP as mentioned earlier.

- Slicing



- Drilling Up/down

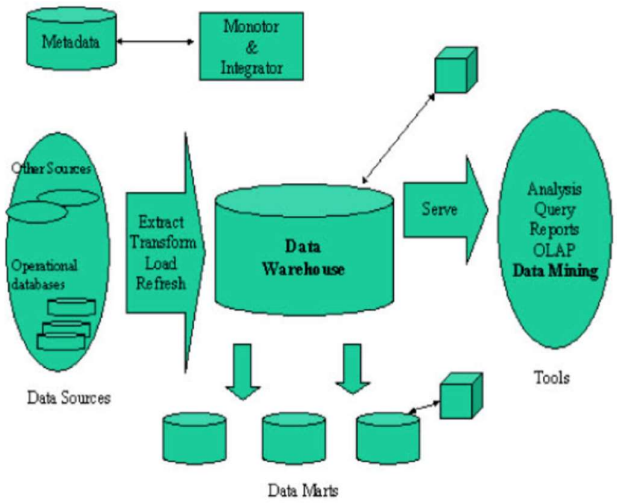


WEEK 4

Assigned Readings: An overview of data warehousing.

In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting and data analysis and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources.

The data needed to provide reports, analytic applications and ad hoc queries all exist within the set of production applications that supports the organization. Table are structured to optimize data entry and validation performance, making them hard to use for retrieval and analysis. IT systems can be divided into transactional (OLTP) and analytical (OLAP) wherein OLTP systems provides source data to warehouses and OLAP systems helps to analyze it. An example of data warehouse architecture is given below:



A star schema is usually used to create a conceptual model of the data. Star schema contains a large central table. A snowflake schema is refinement of star schema where some dimensional hierarchy is achieved by further splitting the data into a set of small dimensions. A few concept hierarchies are defined as follow:

- Data cube
- Design of Data Warehouse
- Three Data warehouse models
- Metadata repository

Data warehouse is a subject oriented, integrated, time variant, and non-volatile collection of data in support of a manager's decision-making process.

Assigned Readings: Data warehousing and online analytical processing.

Data warehouses generalize and consolidate data in multidimensional space. The construction of data warehouses involves data cleaning, data integration, and data transformation, and can be viewed as an important preprocessing step for data mining. Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse systems are valuable tools in today's competitive, fast-evolving world. A few key features of Data warehousing are:

- Subject Oriented
- Integrated
- Time variant
- Non-volatile

The utilization of a data warehouse often necessitates a collection of decision support technologies. This allows "knowledge-workers" (e.g., managers, analysts, and executives) to use the warehouse to obtain an overview of the data quickly and conveniently, and to make sound decisions based on information in the warehouse.

The major task of online operational database systems is to perform online transaction and query processing. These systems are called online transaction processing (OLTP) systems. The major distinguishing bit between OLTP & OALP are as follow:

- Users and system orientation
- Data contents

- Database design
- View
- Access patterns

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds

Introduction

Data warehouses is a collection of transformed data from multiple sources (transformed using ETL tools) stored in a common repository for the purpose of supporting data analytics and decision-making. It can also be defined as a collection of tools and technologies to manage data to support Business Intelligence or as a database that is maintained separately from the organization's operational database. The process of constructing and using data warehouses is called Data Warehousing.

Data cubes, dimensions, and measures

Data warehouse looks at the data in the form of a data cube. A data cube has 2 or more dimensions and one or more measures. A dimension shows an aspect of the data, usually categorical and is a single level entity. Dimensions can be hierarchical in nature such as Product categories and their subcategories, a location (country > state > city > suburb), date (year > quarter > month > week > day), etc. Measures are actual quantifiable data in the cube which is usually numerical in nature. Examples of measures are sales, revenue, cost, salary, etc.

OLAP

OLAP stands for Online Analytical Processing. OLAP refers to the ability to perform operations on multi-dimensional data cubes stored in a data warehouse for the purpose of analyzing and visualizing data. A few operations will be discussed below:

- Drill up (Summarization operation)

  Navigating from lower levels of hierarchical data to higher levels of dimensional data. Example would be to total revenue per state from the total revenue of cities.

- Drill down (Summarization operation)

  Navigating from high levels of hierarchical data to lower levels of dimensional data. Example would be to total revenue of a product subcategory from the total revenue of a category.
- Roll up. (Summarization operation)

  Summarize the data measures according to a formula or a calculation (average, max, min, etc.). An example would be to analyze the average sales of a particular year.
- Slice (Extraction Operation)

  Creating a new cube by keeping a single value for a dimension of the source cube. An example would be to compare product sales for a particular year across different categories.
- Dice (Extraction Operation)

  Creating a new cube by keeping multiple values for one or more dimension of the source cube. An example would be to compare the product sales across different product categories across different years.
- Nesting (Extraction operation)

  Placing 2 or more dimensions together in a row or column. Basically, a cube within another cube. An example would be to compare product sales in different categories across various regions.
- Pivot (Extraction operation)

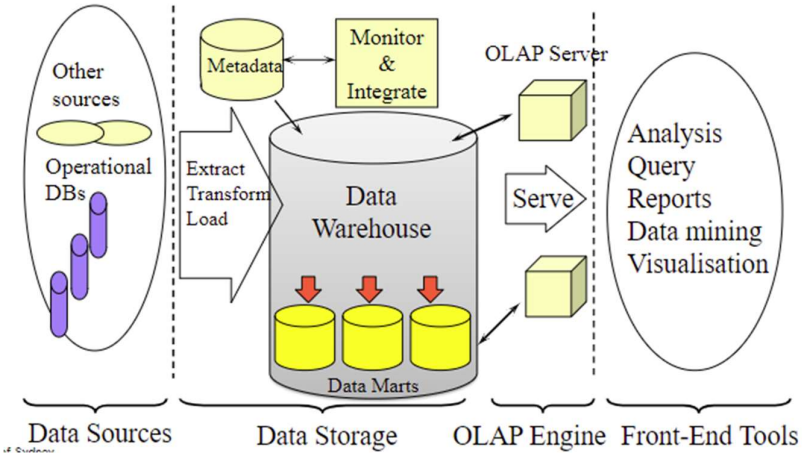
  Rotating the cube to see data from another aspect. For example, comparing product sales in various regions across time instead of different product categories.

OLTP

OLTP stands for Online Transactional Processing. It enables the real time execution of large numbers of database transactions by large number of users for everyday operations. The major difference between OLAP and OLTP is summarized below:

	OLTP (on-line transaction processing)	OLAP (on-line analytical processing)
Definition	Major task of traditional relational DBMS	Major task of data warehouse system
Function	Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.	Data analysis and decision support, data mining
Users	Clerk, IT professional	Knowledge worker, data scientists, business analysts
DB Design	Application-oriented, ER	Subject-oriented, star-schema, etc.
Data	Current, up-to-date, detailed, relational, isolated	Historical, summarised, multidimensional, integrated, consolidated
Usage	Repetitive	Ad-hoc
Access	Lots of read/write	Lots of scans and complex read query execution
Unit of Work	Short, simple transactions (CRUD)	Complex read-only queries
# Records Accessed	Tens	Millions
# Users	Thousands	Hundreds
Response Time	Milliseconds	Seconds
DB Size	100MB-GB	100GB-TB
Metric	Transaction throughput	Query throughput
Optimisation	Concurrency, access, large number of transactions, data recovery	Complex data analysis and fast read query execution
Normalisation	Highly normalised for fast inserts and updates	Highly denormalised for fast read (select) queries
Availability	Critical, 24/7/365	Less critical
Backup	Frequent, concurrent backups	Less-frequent or one-time backup of historical data

Multi-Tiered architecture of Date Warehouse



The data is collected from the data sources. It is extracted, Transformed, and loaded into the Data Storage. This stored data is then processed using OLAP engine to answer the Front-end tools. There are different types of Data Warehouse Models which are:

- Enterprise Warehouse

  This type of data warehouse collects all the information about the subjects spanning the entire organization. This enables enterprise level decision-making. The cost of Enterprise Warehouse is high.
- Data Mart

  This is a subset of the corporate wide data that is of value to a specific group of users. Its scope is confined to specific, selected groups and provides easier access to data required by specific teams within the organization. Data Marts is usually preferred as it provides quicker access to data, provides faster insights, and hence facilitates faster decision making. I require access to few datasets and is less resource intensive making it simpler and faster to implement. Data Marts also enables very specific data analytics projects that are short lived (transient analysis).
- Virtual Warehouse

  In this type of data warehouse, the actual data is not stored. Only the metadata is, and the data is extracted from the source as and when required.

The Data warehouse also has the Data warehouse Metadata Repository. Metadata is the data defining warehouse objects. It stores the structure of the data warehouse and has the schemas, view, dimensions, hierarchies, etc. Operational metadata includes the data lineage and history of the migrated and transformation path. It also includes the currency of data, monitoring information and the algorithms, configurations used for summarization.

OLAP Server Architecture

- MOLAP

  MOLAP is called Multidimensional OLAP. In this, data is stored in a multi-dimensional cube. It uses a sparse array based (matrix) multidimensional storage engine. It has pre-computed queries to fit in to the cube and pre-summarized stored data. It has fast data retrieval, optimal for slicing, dicing, drilling up and down. It can perform complex calculations and aggregations and has fast performance for large datasets.

- ROLAP

  ROLAP is called Relational OLAP. In this, a relational DBMS is used to store, query, and manage warehouse data while enabling slice and dice operations. Relies mainly on SQL to query the relational Databases. Complex computations can be difficult in this. The ROLAP includes optimization of RDBMS backend, implementation of aggregation logic, and additional tools and services. Queries are made on demand and no precomputation of information is present. Performance can be slow due to large and inefficient joins between large tables.
- HOLAP

  HOLAP is called Hybrid OLAP. This is part MOLAP and part ROLAP. It is flexible in nature and uses the low-level source data in relational data (Part ROLAP) and uses high level aggregations in sparse matrix (Part MOLAP).

Data warehouse usage and applications

Data warehousing gives us 3 distinct use cases. Data warehouse allows us to process the information by supporting querying, basic statistical analysis, and reporting using visuals, tables, etc. It also allows us to analytically process the data through the multidimensional analysis of data. This is done through basic OLAP operations, such as slice-dice, drilling, pivoting, etc. Finally, data warehousing also allows us data mining as we can discover knowledge from hidden patterns. Data warehouse supports associations, constructing analytical models, performing classification and prediction, etc.

OLAM

OLAM stands for Online Analytical Mining. High quality data is present in data warehouses in an integrated, consistent, and cleaned format. The available information processing structure of data warehouses is advanced and OALP based exploratory data analysis is possible through mining using the ETL tools. OLAM is becoming more mainstream with the online and on demand selection of data mining functions such as integration and swapping of multiple mining functions, algorithms, and tasks.

Data Analytics for Business Intelligence

There are different types of data analytics for Business Intelligence, and they are:

- Descriptive (What happened?)

  Collects and analyses historical data and it focuses on what has already happened. This is not used to draw inferences or make predictions from its findings. It uses aggregation, data mining, data discovery, and exploration to discover patterns, and trends in the data. It is easy to gain insights to support decision making through descriptive analysis.
- Diagnostic (Why did it happen?)

  Analyses the historical data to understand why something happened. It is focused on determining the causes of trends and correlations. It uses drill down, data discovery, data mining, and correlation analysis. It compares coexisting trends, uncovers correlation between variables, and determining causal relationships wherever possible.
- Predictive (What could happen?)

  Analyzing the past data patterns and trends by looking at historical data and focused on understanding what could happen in the future. It uses probabilities, data mining, statistical modelling, and machine learning to forecast possible future outcomes and their likelihoods. It is based on probabilities and is hence never completely accurate.
- Prescriptive (What should happen?)

The most advanced stage in Business Analysis process is Prescriptive analysis. In this analysis is descriptive and predictive analytics results and recommends the best possible course of action. It uses statistics, and machine learning algorithms to anticipate what, when, and why something may happen. It requires a large amount of data to produce useful insights.

Correlation and causation are different. Correlation means that the directional movement of 2 or more variables is related. 2 variables can be positive, negative or have no correlation. Causation means that the change in one variable causes a change in another variable. Correlation doesn't imply causation, but causation always implies correlation. Correlation can offer insights but finding the cause requires controlled experiments.

WEEK 5

Assigned Reading: Lakehouse

Lakehouse is a new architectural pattern which is based on open direct-access data formats such as Apache Parquet, has first class support for Machine Learning and Data Science, and offers state of the art performance. Lakehouses can help address several major challenges with data warehouses, including data staleness, reliability, total cost of ownership, data lock-in, and limited use-case support.

A decade ago, the first-generation systems started to face several challenges. First, they typically coupled compute and storage into an on-premises appliance. This forced enterprises to provide and pay for the peak of user load and data under management, which became very costly as datasets grew. Second, not only were datasets growing rapidly, but more and more datasets where completely unstructured which data warehouses could not store and query at all. To solve these problems, the second-generation data analytics platforms started offloading all the raw data into data lakes: low-cost storage systems with a file API that hold data in generic and usually open file formats, such as Apache Parquet and ORC. This approach started with the Apache Hadoop movement, using the Hadoop File System (HDFS) for cheap storage. From 2015 onwards, cloud data lakes, such as S3, ADLS and GCS, started replacing HDFS. They have superior durability, geo-replication, and most importantly, extremely low cost with the possibility of automatic, even cheaper, archival storage, etc.

In today's architectures, data is first ETLed into lakes, and then again ELTed into warehouses, creating complexity, delays, and new failure modes. Moreover, enterprise use cases now include advanced analytics such as machine learning, for which neither data lakes nor warehouses are ideal. Specifically, today's data architectures commonly suffer from four problems:

- Reliability
- Data staleness
- Limited support for advanced analytics
- Total cost of ownership

We define a Lakehouse as a data management system based on low-cost and directly accessible storage that also provides traditional analytical DBMS management and performance features such as ACID transactions, data versioning, auditing, indexing, caching, and query optimization. Lakehouses thus combine the key benefits of data lakes and data warehouses: low-cost storage in an open format accessible by a variety of systems from the former, and powerful management and optimization features from the latter. The architecture of the Lakehouse is defined below:

- The system stores the data in a low-cost object store such as Amazon S3 using a standard file format such as Apache Parquet but implements a transactional metadata layer on top of the object store that defines which objects are part of a table version.
- Implement metadata layers for Data Management.
- Implement SQL performance optimization techniques such as caching, auxiliary data, and data layouts.

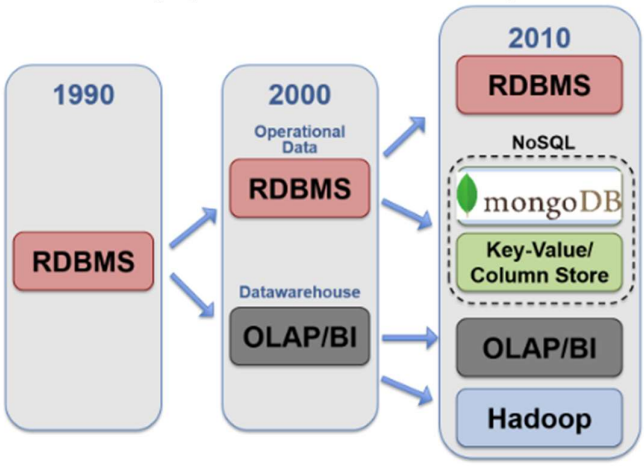
We have argued that a unified data platform architecture that implements data warehousing functionality over open data lake file formats can provide competitive performance with today's data warehouse systems and help address many of the challenges facing data warehouse users. Although constraining a data warehouses' storage layer to open, directly accessible files in a standard format appears like a significant limitation at first, optimizations such as caching for hot data and data layout optimization for cold data can allow Lakehouse systems to achieve competitive performance. We believe that the industry is likely to converge towards Lakehouse. designs given the vast amounts of data already in data lakes and the opportunity to greatly simplify enterprise data architectures.

Big Data

There are mainly 2 types of data which are structured and unstructured data. Unstructured data usually means that the type of data doesn't have a specific schema allocated to it whereas Structured data usually has a predefined schema. Examples of unstructured data would be social media, notes, images, videos, etc. Examples of structured data would include data stored in databases, data warehouses such as sales data, financial transactions, credit card information, etc. In today's era the amount of unstructured data being accumulated is increasing.

The major data sources for big data are machine log data, social media, sensor data, public web sources, multimedia files, internal documents, archived data, etc. There are 3 Vs of Big Data which are Volume, Variety, and Velocity. Each V describes a characteristic of Big Data. The rising volumes of data and drastic reduction in storage costs, falling cost of data management tools, rising number of data scientists and added competitive advantage is driving Big Data in today's world.

New DB technologies are coming up as seen in the below image:



NoSQL

NoSQL stands for Not Only Structured Query Language. It is a type of Database Management System which is more versatile than traditional database systems. It can store, retrieve, and query unstructured documents. It has a flexible schema wherein data is stored without a predefined schema and big data sets can be analyzed in parallel by assigning them to different servers, results of which are then collected and aggregated and can be further used in conjunction with relational database systems.

NoSQL is more efficient than Relational database and sharing and replication is simpler. The NoSQL is more tolerant to failures and is focused on horizontal scalability. The relational model takes data and separates it into tables whereas in NoSQL, it isn't necessary. NoSQL has a dynamic schema, high data velocity and allows storage of data that is structured, semi-structured, and unstructured. There are different types of NoSQL databases such as the key value stores, document stores, column oriented, graph networks, object oriented, etc.

Big Data and Business Intelligence

The Web becomes the data repository as there are many sources of external data coming through the web. New challenges in extraction, integration, and analysis come from the heterogenous data sources. Business Intelligence applications moving to the web as a service provided from the cloud combines historical and real-time data. There are limits to data warehousing for unstructured data, especially if it is big data. These limits are:

- Data warehouses usually handle primarily structured data. There has been massive growth in less structured, diverse data.
- Data warehouses are not the most flexible as they are schema driven, hence they face problems with the unstructured data.
- Data warehouses are not designed to easily support the massive amounts of diverse data.
- It is expensive to design and implement an efficient ETL pipeline of Big Data.

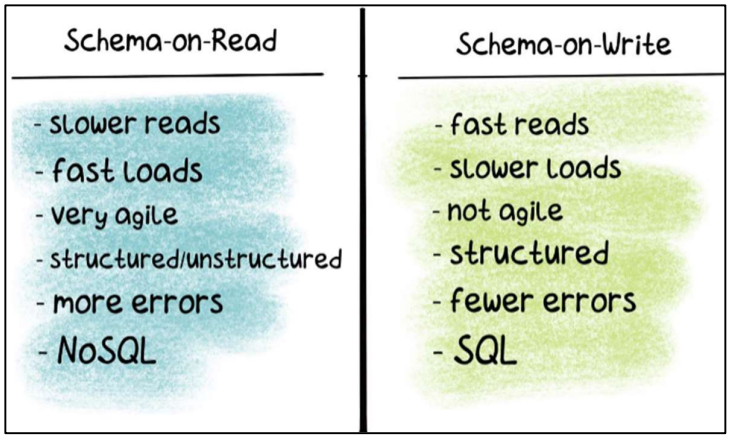
Data Lakes

Data lakes are alternative approaches to big data storage, exploration, and analytics. A data lake is a repository of data stored in its natural raw format that creates a large, catalogued, centralized, cheap storage system. Blob storage systems on the cloud are scalable, inexpensive, write only file systems which are used to access these stored data. Data lake is a method of organizing large volumes of highly diverse data from diverse sources. The main principle behind data lakes is to ingest first and prep later, meaning that the main aim of data lakes is to make data available faster for multiple use cases.

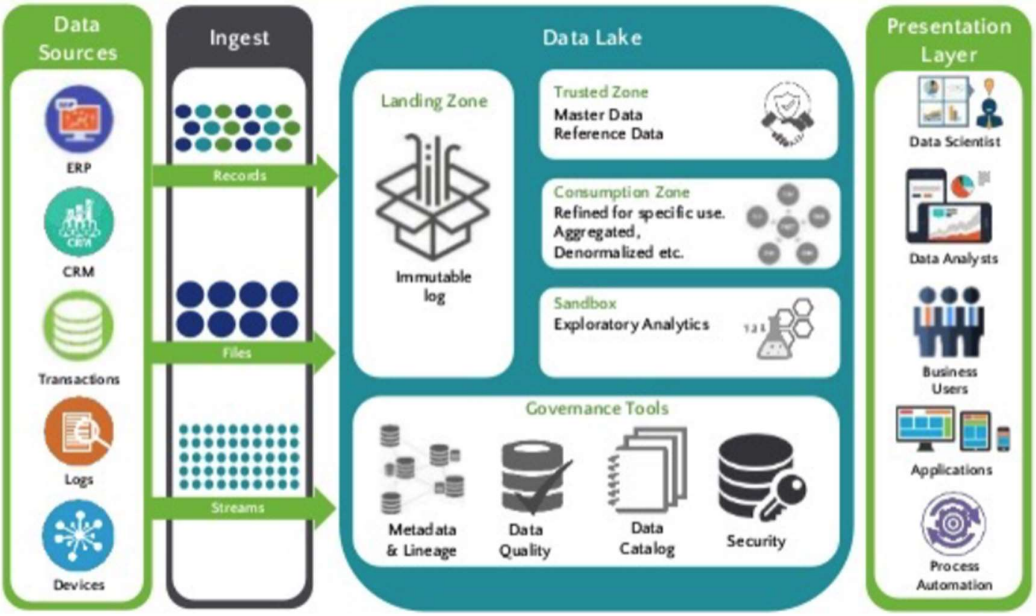
The organizations want to realize the value from all the data being collected and have more data incoming through all the data coming online and hence are driven to adopt big data practices. The differences between Data lakes and traditional Data warehouses are given below:

Data Lake	Traditional On-Premise Data Warehouse
• Data is stored in native format • Store forever • Flexible access to raw data • Schema-on-Read • Seperate storage & compute	• Data requires transformation • Expensive to store large volumes • Transformed before loading (ETL) • Schema-on-Write • Tightly coupled storage & compute

The main difference between schema on read and schema on write is given below:



The basic architecture of ETL pipeline in a data warehouse is Source data → ETL infrastructure → Data warehouse. The basic architecture of ETL pipeline in a Data lake is Source data → ingest → Data lake → Presentation layer.



There are limits to the OLAP approach for data warehouses which are covered by data lakes. The main advantage being that organizations must manage multiple data types, from many sources, and add new data sources and formats regularly. All this can be done using data lakes. Data lakes allow an organization to load and store all their data, structured and unstructured, in one centralized repository. The main benefits of data lakes are the speed to insight, it is a single source of truth, agility, future proofing, and cost.

Trends in Business Intelligence

There are multiple new trends in Business Intelligence. A few are mentioned below:

- Embedded Analytics
  Business Intelligence applications are not integrated into the workflow systems in organizations. Embedded Analytics involves Business Intelligence involves Business Intelligence and dashboards, reports, and visualizations that are directly integrated into the workflow applications. It presents analytics users with information in context within the applications. One can experience complete analytics experience from inside a workflow or other applications.

- Cloud Computing
  Cloud computing is a system that Business Intelligence tools can take advantage of. Cloud computing alters the way computing, storage, and networking resources are allocated. Through virtualization cloud architecture gains a service centered approach. Applications do not need physical resources, but virtual resources are dynamically allocated based on demand. There are different types of cloud models such as On-premises, Infrastructure As A Service, Platform As A Service, and Software As A Service.

- Social Media Analytics
  It is another source of unstructured data which is valuable for a lot of organizations. It is useful for trend analysis, marketing optimization and even sentiment analysis. Through these services, Social Media Analytics has become a key part of marketing analysis that Organizations conduct.

- Web Analytics
  It is the measurement, collection, analysis, and reporting of internet data for understanding and optimizing web usage is called Web analytics.

- Unstructured data and text analytics
  Text analytics delivers clarity, speed, and breadth. It extracts and classifies unstructured data into multiple languages. It discovers patterns in events and opinions and categorizes them and models customer behavior based on qualitative analysis. This is done using Natural Language Processing (NLP) such as Sentiment Analysis.

WEEK 6

Data Visualization

Data visualization is the process of presenting information graphically in which relationships, patterns, and differences are encoded through shapes, colors, positions, and size. It is an art of presenting data in a way which is useful for the users. There are multiple methods of visualizing data such as:

- Comparison
  Comparing values or quantities over time or various categories. We use this to illustrate the similarities and differences among categories. A few visuals which are a good fit for comparisons are vertical bar chart, column bar chart, and horizontal bar chart.
- Proportion
  Displaying individual parts of a whole. We use this to show summaries of data, similarities, anomalies, and a percentage related to the whole data. A few visuals that are a good fit for proportions are pie charts, stacked bar chart, stacked area chart, tree maps, etc.

- Distribution
  Showing possible values (or intervals) of the data and how often they occur. We use this to reveal outliers, shape of distributions, frequencies, range of values, minimum, maximum, median values, etc. A few graphs which would be a good fit for Distribution are histograms, density plots, and box plots.
- Relationship
  Showing how one or more variables are related to other variables. This is used to show outliers and correlations (positive and negative). A few graphs that are a good fit for relationship diagrams are scatterplots, scatterplot matrix, bubble chart, etc.
- Location
  Visualizing data in relation to geography. It is used to demonstrate similarities and differences by location, density, or counts. A few charts which are a good fit for Location type questions are Choropleth (filled maps), point maps, symbol maps, etc.
- Trends
  Showing how one or more variables change over time. It requires time dimensions, and it illustrates change over time, cycles, or comparisons over time. A few charts which are a good fit for Trends are Line chart, sparkline chart, and area graph.
- Word frequency and sentiment
  Visualizing textual data. It provides insights of the frequency or count of words and phrases. It is useful for sentiment analysis and a word cloud is the best chart type for this question.

Dashboards

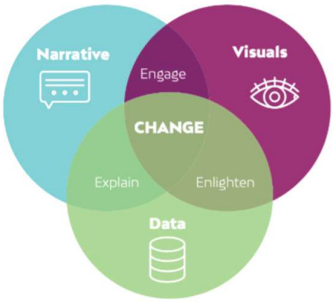
Dashboards are a collection of charts which provide insights immediately. It has several advantages such as:

- Fast and effective decision making.
- Immediate access to key performance metrics and indicators.
- Monitor processes and identify bottlenecks.
- Access to relevant information when making decisions.
- Avoids information overload.

To create a successful dashboard, one needs to experiment and iterate the process until a good dashboard is formed. Dashboards need to be simplistic and not too overdesigned. Simple, less is more. Overloading the dashboard is not the way to deliver insights. Creativity is always appreciated in Dashboards. Having interactive dashboards is an effective way of delivering more information through a smaller number of graphs.

Data Story telling

It is the combination of Data, Visualization, and Narration. Data story telling is the process of using data to tell a story, communicate insights, and influence the desired change. There are 3 main components of Data story telling:



Using data storytelling, we make an impact on the audience. It is much easier to remember than statistics. It is more persuasive and motivating and it boosts engagement with the audience more easily. Following are the steps to tell a story with data:

- Find and define the story in the data.
- Define the perspective.
- Create a hierarchy.
- Organize.
- Plot.
- Let the data tell the story.
- Choose the right visuals.
- Review, test, and edit.

WEEK 7

Evolution of the Web

Before the internet, knowledge was all stored-on papers and documents and books. It was all connected through citations and if one needed to access it, one must travel physically to the library or the storage area to find that resource. The internet created a web of all these resources, and this was done using hyperlinks. Clicking on the hyperlink allowed one to travel online to another resource location all together. Through the internet, all information is now available to us at our fingertips. This phase was called the Web of Documents stage.

This transformed into Application Silos. Application silos were basically data which was stored in different applications. This gave rise to a problem wherein one must go to each application to update the data stored in it. This gave rise to the next generation of the web, the Web of Data.

The Web of Data abstracts away the data from the application layer. This way the data is all centralized. This is also called the Semantic Web. Semantic Web is an extension to the current Web which standardizes the way Web documents are published. This web has a fatal error. This version of the web was designed for humans and not for machines. Semantic web is about the meanings, key concepts, or content of a webpage.

Knowledge Graph

A knowledge graph, also known as a semantic network, represents a network of real-world entities (objects, events, situations, etc.) and illustrates the relationship between them. This information is usually stored in a graph database and visualized as a graph structure. A knowledge graph is a large semantic network of interconnected data objects representing formal and structured definitions of knowledge in a domain.

The purpose of having semantic knowledge graphs for knowledge management is to share a common understanding among people in an organization or among partner organizations. It is useful to express facts and statements in an unambiguous language, can make domain knowledge explicit, infer new knowledge from existing knowledge, reuse and link other knowledge graphs, and much more.

WEEK 8

Assigned Readings: The Semantic Web

Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully. Computers can adeptly parse Web pages for layout and routine processing but in general, computers have no reliable way to process the semantics.

The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The essential property of the World Wide Web is its universality. The power of a hypertext link is that "anything can link to anything".

For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. Traditional knowledge-representation systems typically have been centralized, requiring everyone to share the same definition of common concepts. But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable. Moreover, these systems usually carefully limit the questions that can be asked so that the computer can answer reliably— or answer at all. Semantic Web researchers, in contrast, accept that paradoxes and unanswerable questions are a price that must be paid to achieve versatility.

Adding logic to the Web—the means to use rules to make inferences, choose courses of action and answer questions—is the task before the Semantic Web community now. A mixture of mathematical and engineering decisions complicates this task. The logic must be powerful enough to describe complex properties of objects but not so powerful that agents can be tricked by being asked to consider a paradox.

Two important technologies for developing the Semantic Web are already in place: eXtensible Markup Language (XML) and the Resource Description Framework (RDF). XML lets everyone create their own tags that annotate Web pages or sections of text on a page. In short, XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean. Meaning is expressed by RDF, which encodes it in sets of triples, each triple being rather like the subject, verb, and object of an elementary sentence. These triples can be written using XML tags. In RDF, a document makes assertions that things have properties with certain values. This structure turns out to be a natural way to describe most of the data processed by machines. The triples of RDF form webs of information about related things.

Ideally, the program must have a way to discover such common meanings for whatever databases it encounters. A solution to this problem is provided by the third basic component of the Semantic Web, collections of information called ontologies. For Artificial-intelligence and Web researchers an ontology is a document or file that formally defines the relations among terms. The most typical kind of ontology for the Web has a taxonomy and a set of inference rules. The taxonomy defines classes of objects and relations among them. Classes, subclasses, and relations among entities are a very powerful tool for Web use. We can express many relations among entities by assigning properties to classes and allowing subclasses to inherit such properties. Inference rules in ontologies supply further power.

With ontology pages on the Web, solutions to terminology (and other) problems begin to emerge. The meaning of terms or XML codes used on a Web page can be defined by pointers from the page to an ontology. Ontologies can enhance the functioning of the Web in many ways. They can be used in a simple fashion to improve the accuracy of Web searches—the search program can look for only those pages that refer to a precise concept instead of all the ones using ambiguous keywords. More advanced applications will use ontologies to relate the information on a page to the associated knowledge structures and inference rules.

The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine-readable Web content and automated services (including other agents) become available. The Semantic Web promotes this synergy: even agents that were not expressly designed to work together can transfer data among themselves when the data comes with semantics. Another vital feature will be digital signatures, which are encrypted blocks of data that computers and agents can use to verify that the attached information has been provided by a specific trusted source. Many automated Web-based services already exist without semantics, but other programs such as agents have no way to locate one that will perform a specific function. This process, called service discovery, can happen only when there is common language to describe a service in a way that lets other agents "understand" both the function offered and how to take advantage of it.

In the next step, the Semantic Web will break out of the virtual realm and extend into our physical world. For instance, what today is called home automation requires careful configuration for appliances to work together. Semantic descriptions of device capabilities and functionality will let us achieve such automation with minimal human intervention. The semantic web is not "merely" the tool for conducting individual tasks that we have discussed so far. In addition, if properly designed, the Semantic Web can assist the evolution of human knowledge.

Introduction

There are 3 generations of Knowledge Management Tools and World Wide Web. The first generation was the first stage of WWW evolution which was the static web. It had centralized repositories based on Web 1.0 as its Knowledge Management System. In the second generation we got the read-write web, the social web. The knowledge management system in this generation was collaborative, social Knowledge Management based on Web 2.0. Lastly, the third generation is the transition stage from web of documents to web of data which is standardized and machine understandable in nature. The knowledge management in this scenario is Semantic Knowledge Management based on Web 3.0 and has Knowledge Graphs in it. The first two generations were designed for humans and not for machines.

Knowledge Repositories

Knowledge repositories are electronic repositories that systematically capture, categorize, and store a wide range of documents. Designed to be easy to retrieve and use. It Can help organizations connect people with useful, contextualized information and each other. It has a lot of limitations such as:

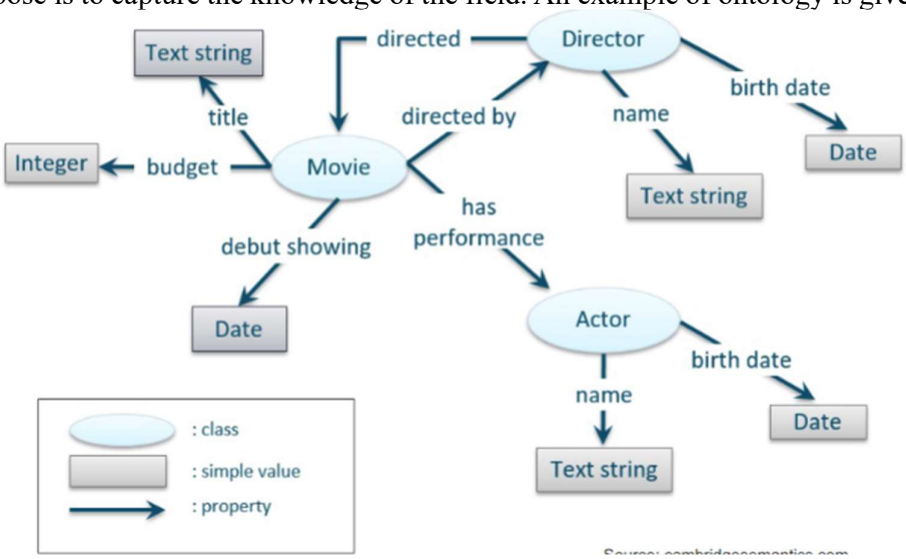
- The keywords do not capture the meaning.
- Difficulties in accessing the most relevant documents.
- Difficulties in understanding the semantic relationship among the relevant documents.
- Over time the amount of information collected might lead to information overload.
- The document collections may not be up-to-date, consistent, or correct.
- No standard representation.

Semantic Web

It is an extension of the current Web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation. It standardizes the way Web documents are published. It is about the definition of the information than presentation of it.

Ontology

It is an explicit (a concrete, formal and machine readable) specification of a conceptualization (an abstract model describing a particular field of knowledge or domain). It is a formal specification of concepts in a domain and the relationships between the concepts. It includes the concepts (classes), properties (attributes of concepts and relations between concepts), constraints and inferencing rules, and sometimes individual instances. Ontology defines a common vocabulary and intended meaning; it enables a shared understanding. Its main purpose is to capture the knowledge of the field. An example of ontology is given below:



The following are the reasons for developing ontologies:

- To share common understanding of the structure of information among people or software agents.
- To enable reuse of domain knowledge.
- To reuse ontologies built by others.
- To make domain assumptions explicit.
- Analyzing domain knowledge.
- To infer new knowledge from existing knowledge.

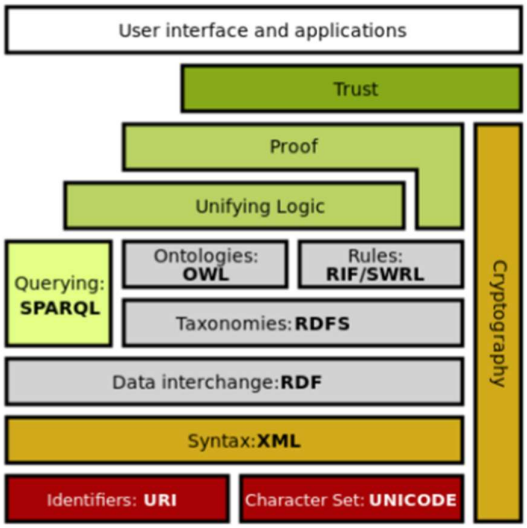
Deriving new, implicit data from known, explicit data based on a set of inferencing rules is Inferencing. Rules here are a set of axioms based on which new data is asserted. The following is the Ontology development process:

- Determine the domain and scope of the ontology.
- Consider reusing existing ontologies.
- Define the class and the class hierarchies.
- Define the attribute of classes.
- Define the relations between classes.

There are multiple languages used for creating ontologies such as:

- RDF (Resource Description Framework): The data modelling language.
- RDFS (RDF Schema): The schema language of RDF.
- OWL (Web ontology Language): An expressive Ontology language.
- SHACL (Shapes Constraint Language): Data validation and constraints.
- SPARQL: The query language for Semantic Web and ontologies.

The semantic web stack is shown below:



Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. There are 3 types of Linked Data:

- Private: it can be used inside organizations and shared between business partners to provide easier integration and to facilitate interoperability.
- Public: It can also be open. Datasets in various domains which are published on the web based on linked data principles and are publicly available for machine and human consumption.
- Mixed: Linking a private linked data to a public one.

Linked open data is a community effort which aims to publish linked data on the web using semantic web technologies. There are 5 principles of Linked Open Data:

- It is available on the web.
- It is available on the web as a machine-readable structured data.
- All the above and is in a non-proprietary format.
- All the above and uses open standard to identify things.
- All the above plus data is linked to other people's data to provide content.

Given below is the advantages of 5-star Linked Data:

Data Consumer	Data Publisher
Discover related data while consuming data.	Make your data discoverable.
Directly learn about the data schema.	Benefit from other data publishers linking into your data.
Link to it from any other place.	Have fine granular control over the data items and optimize their access.
Combine the data with other data.	Increase the value of your data.
Reuse existing tools and libraries.	

WEEK 9

Assigned Readings: What is a Knowledge Graph

Knowledge graphs provide an opportunity to expand our understanding of how knowledge can be managed on the Web and how that knowledge can be distinguished from more conventional Web-based data publication schemes such as Linked Data. Knowledge graph meaning is expressed as structure, the statements are unambiguous, and use a limited set of relation types. It includes explicit provenance and may include uncertainty assessments.

Knowledge graphs are a critical component of the Semantic Web and serve as information hubs for general use as well as for domain-specific applications. Most knowledge graphs seek to aggregate knowledge from third party sources, whether from external databases, from data aggregated though crawling the Web, or through the application of entity and relationship extraction methods. Knowledge graphs are not simply aggregations of RDF or linked data, but critically provide time-invariant information about entities of general interest. Their structures tend to be focused on a limited set of relations adhering to a coherent knowledge model, setting them apart from the linked data cloud in general, which usually has relied on the open framework of the Semantic Web to accommodate a completely free-form use of vocabularies and ontologies.

Semantic Web

Semantic Web is an extension to the current Web which standardizes the way Web documents are published. It standardizes the web by using standard vocabularies (ontologies), format (RDF), Representation language (RDFS, OWL, SHACL), and query language (SPARQL). The Semantic Web abstracts away the documents and applications layer. It creates a web of data. An RDF graph is a labelled, directed graph made of triples. This is as shown below:

Ontology

A formal specification of concepts in a domain and the relationships between the concepts. It includes the concepts (classes), properties (attributes of concepts and relations between concepts), constraints and inferencing rules, and sometimes individual instances. Ontology defines a common vocabulary and intended meaning; it enables a shared understanding. Its main purpose is to capture the knowledge of the field.

Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. There are 3 types of Linked Data:

- Private: it can be used inside organizations and shared between business partners to provide easier integration and to facilitate interoperability.
- Public: It can also be open. Datasets in various domains which are published on the web based on linked data principles and are publicly available for machine and human consumption.
- Mixed: Linking a private linked data to a public one.

Linked open data is a community effort which aims to publish linked data on the web using semantic web technologies. There are 5 principles of Linked Open Data:

- It is available on the web.
- It is available on the web as a machine-readable structured data.
- All the above and is in a non-proprietary format.
- All the above and uses open standard to identify things.
- All the above plus data is linked to other people's data to provide content.

Given below is the advantages of 5-star Linked Data:

Data Consumer	Data Publisher
Discover related data while consuming data.	Make your data discoverable.
Directly learn about the data schema.	Benefit from other data publishers linking into your data.
Link to it from any other place.	Have fine granular control over the data items and optimize their access.
Combine the data with other data.	Increase the value of your data.
Reuse existing tools and libraries.	

Knowledge Graph

A knowledge graph is a large semantic network of interconnected data objects representing a formal and structured definition of knowledge in a domain. A few qualities of knowledge graphs are given below:

- It is a labelled, directed graph.
- The meaning is expressed as a structure.
- Statements are unambiguous.
- All identities are identifiable using global identifiers.
- Good Knowledge Graphs include explicit provenance.

A few knowledge graph technologies are:

- Ontology based data access.
- REDF graphs are based on Semantic web technologies and standards.
- Property graphs are directed graph models represented by a set of labelled nodes and edges.

A few characteristic features of Knowledge Graphs are:

- Reuse: KGs can reuse existing, well-established ontologies and other semantic KGs.
- Interoperability: Semantic KGs can be shared among applications and on the web.
- Flexibility: Semantic KGs can operate in an open environment in which classes can be defined dynamically.
- Adaptability: Semantic KGs can adapt to changing requirements by creating flexible data and schema layers.
- Consistency and Quality: Checking across models can be done automatically.
- Reasoning: KGs can benefit from automated reasoning.

Knowledge graphs are required when:

- Many concepts with complex relationships among them are present.
- Combining internal and external data.
- Complex logical conditions and rules.

- Need a flexible data model.
- Unifying various formats of data.

WEEK 10

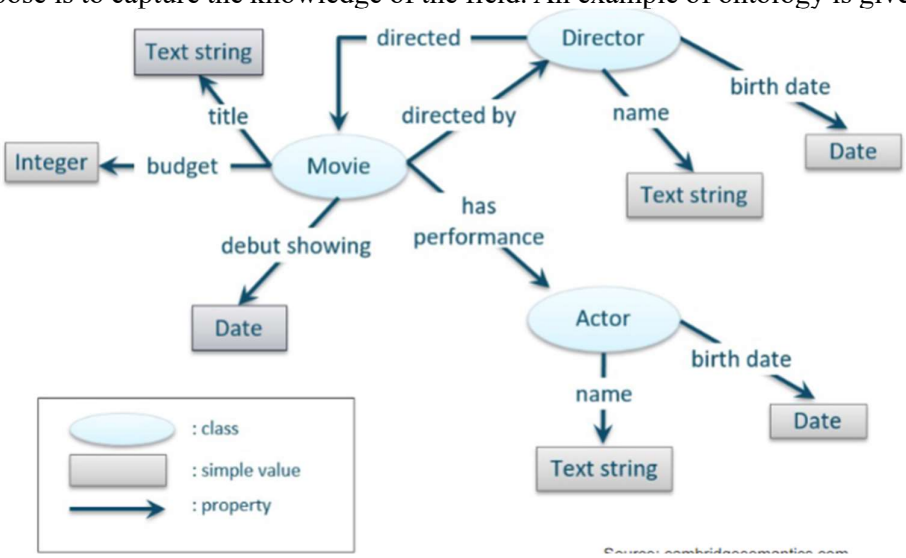
Assigned Readings: Why Ontology will be a big word in your company's future.

Your business has its own language. This language is not only critical to being able to communicate with others in your organization as well as with your customers, but it also influences how the programmers and data scientists identify those things that their data systems track, accept and analyze. While there are several different ways you can describe this language, one of the more useful is called ontology, which quite literally means the study of the names of things.

The concepts, relationships/properties, and individual data, and are collectively referred to as an ontology. In effect, the ontology describes the things, the sets of descriptions and the relationships, while the data is then given as a particular data cell. The goal of semantics is to make your business language machine-readable. Semantics also makes building such web applications easier and more cost-effective. Comprehensive use of ontologies means that you can make full cycle pipelines from data acquisition to user interfaces to data analysis to dashboards.

Ontology

It is an explicit (a concrete, formal and machine readable) specification of a conceptualization (an abstract model describing a particular field of knowledge or domain). It is a formal specification of concepts in a domain and the relationships between the concepts. It includes the concepts (classes), properties (attributes of concepts and relations between concepts), constraints and inferencing rules, and sometimes individual instances. Ontology defines a common vocabulary and intended meaning; it enables a shared understanding. Its main purpose is to capture the knowledge of the field. An example of ontology is given below:



The following are the reasons for developing ontologies:

- To share common understanding of the structure of information among people or software agents.
- To enable reuse of domain knowledge.
- To reuse ontologies built by others.
- To make domain assumptions explicit.

- Analyzing domain knowledge.
- To infer new knowledge from existing knowledge.

Deriving new, implicit data from known, explicit data based on a set of inferencing rules is Inferencing. Rules here are a set of axioms based on which new data is asserted. The following is the Ontology development process:

- Determine the domain and scope of the ontology.

  Domain ontologies define concepts specific to a particular domain of knowledge. Upper ontologies aim at defining generic concepts regardless of the domain of knowledge. They can be used to integrate or align multiple domain ontologies. Upper ontologies can be used to integrate or align multiple domain ontologies.
- Consider reusing existing ontologies.
- Define the class and the class hierarchies.
- Define the attribute of classes.
- Define the relations between classes.
- Add constraints, restrictions, and validation rules.

Some semantic technologies are standard vocabularies (ontologies), format (RDF), Representation language (RDFS, OWL, SHACL), and query language (SPARQL). The ontology languages are:

- RDF (Resource Description Framework): The data modelling language.
- RDFS (RDF Schema): The schema language of RDF.
- RDFS-Plus: An addition to RDFS with basic constructs from OWL.
- OWL (Web ontology Language): An expressive Ontology language.
- SHACL (Shapes Constraint Language): Data validation and constraints.
- SPARQL: The query language for Semantic Web and ontologies.

Level of Expressivity refers to the variety and quantity of ideas that can be represented and communicated using a language or a model. The following is a graph representing the level of expressivity of ontology languages:

An RDF graph is a labelled, directed graph made of triples. This is as shown below:

A resource is a thing that can have things said about it. Literals, in contrast, are used for values such as strings, numbers, and dates. A URI provides a global identification for a resource across the Web. URIs (Uniform Resource Identifier) can be abbreviated using QNames (Qualified Names).

RDF

A framework for representing a basic statement about any subject in the form of triples. There are different syntaxes:

- RFD/XML
  The original serialization format of RDF represents RDF triples in XML.
- N-Triple
  The simplest form of RDF serialization represents raw RDF triples using fully qualified, unabbreviated URIs.
- Turtle
  A compact form of RDF serialization representing RDF triples using abbreviated QNames.

RDFS

A schema language for RDF with basic level of expressivity to define classes, subclasses, and properties. The Following are RDFS constructs:

rdf:type	To specify that a resource is instance of a class.
rdf:Class	To define a Class.
rdf:Property	To define a Property.
rdfs:domain	To specify the class(es) for which the property is defined.
rdfs:range	To specify the possible values (class or datatype) for a property.
xsd datatypes	A set of standard literal datatypes that can be used as the range of properties.
rdfs:subClassOf	To create a class hierarchy to represent generalisation-specification and inheritance
rdfs:subPropertyOf	To create a property hierarchy to represent generalisation-specification and inheritance

The following are the limits of RDFS Expressivity:

- Cardinality restrictions.
- Disjoint classes.
- Defining classes as Boolean combination of other classes.
- Property characteristics.
- Lack of ability to define inferencing rules.

OWL

It is an ontology language for expressing detailed constraints on classes and properties. A few basic OWL constructs are given below:

owl:inverseOf	To specify that a Property is inverse of another Property. :hasChild owl:inverseOf :hasParent . :employs owl:inverseOf :employedBy . :marriedTo owl:inverseOf :marriedTo . (Symmetric Property)
owl:SymmetricProperty	To define a Property that is inverse of itself. :marriedTo a owl:SymmetricProperty . owl:inverseOf a owl:SymmetricProperty .
owl:TransitiveProperty	To define a transitive Property (e.g., Given that Sydney locatedIn NSW and NSW locatedIn Australia then Sydney locatedIn Australia) :locatedIn a owl:TransitiveProperty .

owl:FunctionalProperty	To define a Property that can have only one (unique) value for any given instance. :hasBiologicalMother a owl:FunctionalProperty . :studentID a owl:FunctionalProperty .
owl:InverseFunctionalProperty	If a Property is defined to be inverse-functional, then the object of a property statement uniquely determines the subject. :studentID a owl:InverseFunctionalProperty .
owl:sameAs	To specify that two resources are the same. data1:student1 :studentID “s123” . data2:student2 :studentID “s123” . data1:student1 owl:sameAs data2:student2 .
owl:DatatypeProperty	To define a Property that can take literal values, i.e. a datatype as its range (also referred to as Attributes). :dateOfBirth a owl:DatatypeProperty ; rdfs:domain :Person ; rdfs:range xsd:date .
owl:ObjectPropery	To define a Property that can take instances of a class as values, i.e. a Class as its range (also referred to as Relations). :marriedTo a owl:ObjectProperty ; rdfs:domain :Person ; rdfs:range :Person .

Some advanced OWL constructs are used for defining restrictions for properties and classes as well as combining classes and restrictions.

- A full set theory language
- Specifying cardinality restrictions
- Disjoint sets

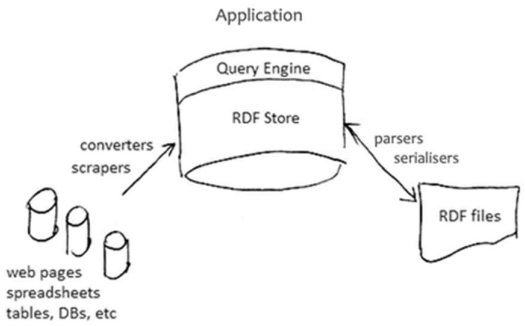
Some of the best practices for creating Ontologies:

- No space or special characters in namespaces and URIs.
- Single nouns for class names.
- Verbs for relations.
- UpperCamelCase for classes and instance names.
- LowerCamelCase for attributes and relations.
- Separate ontology concepts from instances using different namespaces.
- Build molecular ontologies that split your domain into logical parts.

Semantic Application Architecture

Semantic Application Architecture

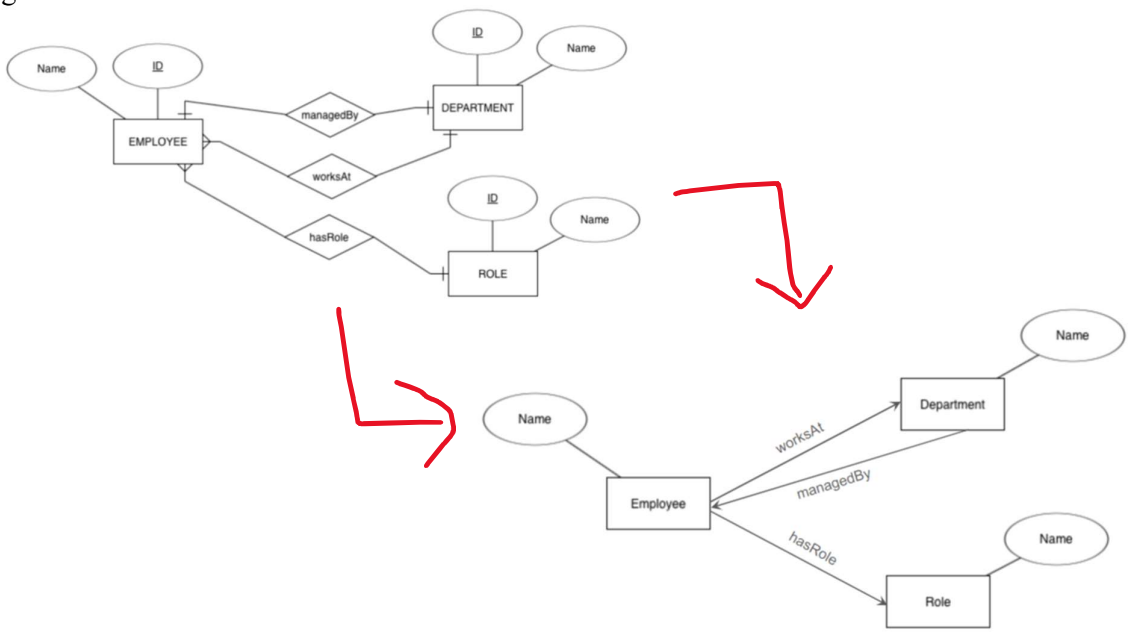
- RDF files in TTL (turtle) or RDF/XML format.
- RDF Store (triple store)
  - Virtuoso, GraphDB, Amazon Neptune, Apache TDB, Stardog
  - Reasoning and inference rules
  - SPARQL Query Engine
- Application Code
  - Libraries: Apache Jena, RDF4J
  - User interface, data entry, charts, analytics



WEEK 11

From Relational to RDF

Using the Entity Relationship Diagram, we create the ontology. An example of this is given below:



SPARQL

SPARQL stands for SPARQL Protocol and RDF Query Language. It has 2 types of components called the protocol layer and the query language. SPARQL query is like SQL query. A few types of SPARQL queries are given below:

- Select: Returns the raw results.
- Construct: Returns the results as a new RDF graph.
- Ask: Returns a Boolean result based on the query.
- Describe: Returns a valid RDF graph describing a resource.
- Update: Delete and Insert.

WEEK 12

Knowledge Graphs Database Architecture

There is no all-in-one solution as the database type will change on the basis of the use case. There are relational databases, analytical (OLAP), Key-Value, Column-Family, Graph, Document, etc. One needs a Knowledge Graph when:

- Many concepts with complex relationships among them are present.
- Combining internal and external data.
- Complex logical conditions and rules.
- Need a flexible data model.
- Unifying various formats of data.

Knowledge graphs can be implemented using various Knowledge Graph technologies such as the RDF graphs which are based on Semantic Web Technologies and Standards or the Property Graphs which are directed graph models represented by a set of labelled nodes and edges. Property Graphs are also known as Labelled Property Graphs (LPG). Both nodes and edges have an ID, a type, and can have a set of properties in the form of Key-Value Pairs. The difference between RDF graphs and Property Graphs are given below:

	RDF Graph	Property Graphs
Nodes	Resources (classes, instances), Attributes (literal values), nodes have no internal structure	Unique ID + key-value pairs
Edges	Relations, relations have no internal structure	Unique ID + key-value pairs
Identifier	URIs for Resources and Relations	Unique IDs for Nodes and Edges
Expressivity	Complex descriptions via links to other nodes. No properties on edges (resolved in RDF*)	Properties (key-value pairs) on edges, it uniquely identify instances of edges
Formal Semantics	Yes. Standard, schema-driven data modelling using ontologies enables reuse.	No formal model representation.
Standardisation	All technologies are W3C approved	Different competing companies
Query Language	SPARQL (query + protocol specifications)	No standard: Cypher, PGOL, GQL, …
Serialisation format	XML, Turtle, N-Triple, N3, JSON-LD, …	No serialisation
Schema Language	RDFS, OWL, SHACL	None

	RDF Graph	Property Graphs
Designed for	Linked Data, Semantic Web, formal representation of knowledge, decentralised data control	Graph representation of data
Reasoning	Yes. Via inference rules.	No.
Strengths	Modelling domain knowledge with complex relations among concepts, global identifiers, data federation, interoperability via standard languages and technologies, data validation and constraints.	Easy to understand and get started with. Graph traversal.
Common User Cases	Knowledge Graphs, data integration, metadata management (sharing common terms and vocabulary), model driven applications, knowledge sharing	Graph analytics, path finding, complex networks (computer, citation, social, …)

Multi Graph Design

We can have logical separation in the ontological designs. Multi Graph designs give a lot more flexibility based on multiple factors. Data is naturally fragmented in multiple graphs and separated by domains, and access control (organizational unit or geographical).

Social Linked Data: SOLID

Solid changes the current model where users must hand over personal data to digital giants in exchange for perceived value. SOLID hands back this control over personal data to us. Solid is an open-source framework to decentralize the way Web applications work today resulting in improved data ownership and privacy. Solid is based on RDF and Semantic Web technologies providing powerful data independence and data management mechanisms.

Usually, data is segmented and partitioned through (non-interoperable) mobile apps (provider). Solid makes it possible to keep the data in control of the user providing a streamlined experience across different apps. User decides where to store personal data: in the Personal Online Datastore (POD).

Reasoning

The process of adding new knowledge (inferences) based on existing knowledge using a set of inference rules. It is based on the transitive rule such as if A → B and B → C then A → C. There are 2 types of reasoning, and these are:

- Deductive Reasoning: Extracting new facts from given existing facts and a set of rules about the world and/or domain of interest.
- Inductive Reasoning: Based on statistics, graph analytics, machine learning, etc.

There are 2 types of reasoning mechanisms:

- Forward Chaining
  It is data driven. Inference rules are applied to available data to extract additional data until the goal is achieved. New facts are added to the database (materialization). This mechanism has faster query responses but slower updates and is mostly used in planning, design, and monitoring tasks.
- Backward Chaining
  It is goal driven. It starts from the goals and moves backwards to see if any data supports achieving this goal. New facts are inferred at query time, and this makes query responses slower and makes updates faster. It is mostly used in classification and diagnosis tasks.

An example for Forward/Backward Chaining

Variables/Assumptions: true or false

A = Have $10,000
B = Younger than 30
C = Education at college level
D = Annual income of at least $40,000
E = Invest in securities
F = Invest in growth stocks
G = Invest in IBM stock (the potential goal)

Example Query/Scenario:

An investor has $10,000 (i.e., that A is true) and that she is 25 years old (i.e., that B is true). She would like advice on investing in IBM stock (yes or no for the goal G)

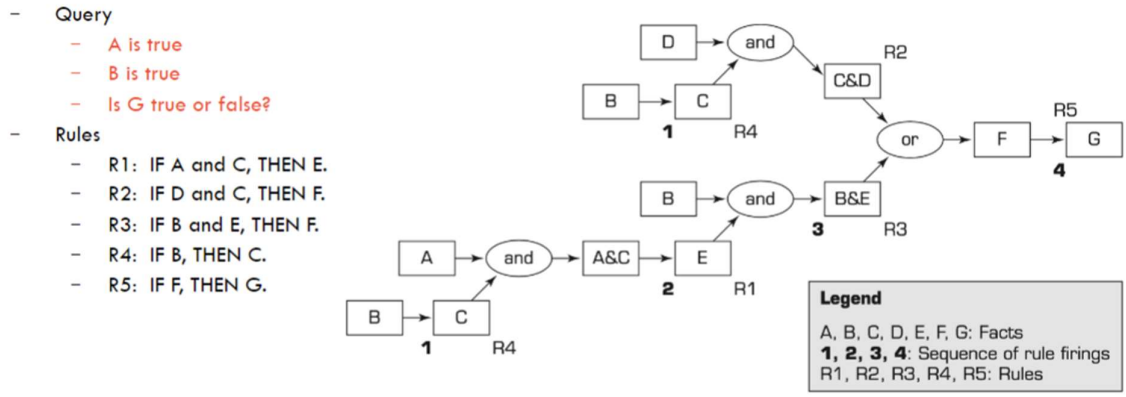
Inference Rules:

R1: IF a person has $10,000 to invest and she has a college degree,
THEN she should invest in securities.
R2: IF a person's annual income is at least $40,000 and she has a college degree,
THEN she should invest in growth stocks.
R3: IF a person is younger than 30 and she is investing in securities,
THEN she should invest in growth stocks.
R4: IF a person is younger than 30,
THEN she has a college degree.
R5: IF a person wants to invest in a growth stock,
THEN the stock should be IBM.

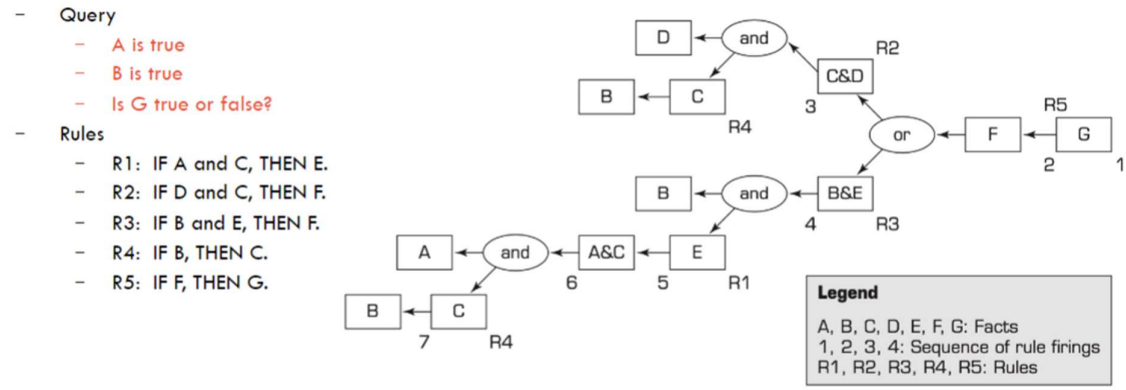
These rules can be written as follows:

R1: IF A and C, THEN E.
R2: IF D and C, THEN F.
R3: IF B and E, THEN F.
R4: IF B, THEN C.
R5: IF F, THEN G.

Forward Chaining



Backward Chaining



GUEST LECTURE

There were 3 stages of Web. The first stage was WWW evolution, the read web. It was a period of simple static websites and was a centralized repository for documents. The second stage was the read-write web, the social web. This was where collaborative, social Knowledge management took place. Finally, the third stage was where read-write-executable web, dynamic applications, interactive services, all came into existence. It transited from web of documents to web of data.

The first stage of web was based on publishing model and eCommerce. People were publishing documents (crating and disseminating information). It involved a consistent look and feel and structure for everyone. It had clear authors and ownerships. It had a wide distribution of key information which was controlled and authoritative. There was consistent messaging and was the only source of information. It had its own set of disadvantages such as being process heavy and having a limited number of content authors. It was slow and time consuming to produce content and it was not suitable for dynamic environments and small teams. The emergence of Web 2.0 started with the creation and spread of social media.

The second stage of web was based on collaboration. A continual process of sharing and refining information by a community. The main base for Web 2.0 was social media which paid greater attention to informal learning, collaboration and inter-personal learning. Social media are basically tools of social software. There are different types of social software such as blogging, social networking, instant messaging, etc. but the main purpose is collaboration.

Social media saw a rapid growth in participation and is still increasing globally. It is more diversified and has many communities. There are a lot of opportunities for knowledge acquisition and collaboration through social and other interactions.

Communities of Practice are a collection of people who engage on an ongoing basis in a common endeavor or shared interest or passion. Essentially bottom-up, typically evolve naturally but can also be created. goal of gaining knowledge related to their field. It is through the process of sharing information and experiences with the group that the members learn from each other and have an opportunity to develop themselves personally and professionally.

Social media technology provides a conduit and means for people to share their knowledge, insight, and experience. It's no longer about storing knowledge but about knowledge that is needed at the right time. Web 2.0 spawned the crowdsourcing phenomenon. Crowdsourcing refers to the idea that the World Wide Web can facilitate the aggregation and/or selection of useful inputs and solutions from a potentially large number of people connected to the internet.

The web 3.0 is the next paradigm shift of the internet taking the best of web 2.0, including rich internet applications and social media, and bringing them to mobile devices, netbooks, and digital signage. Information is searched for filtered, personalized, and delivered to end users based on preferences, biofeedback, and location.