Chapter 10:
Emerging Database Technologies & Applications

Contents

1 Distributed Databases & Client-Server Architectures

2 Spatial and Temporal Database

3 Multimedia Databases

4 Geographic Information Systems

5 XML

6 Data Warehousing

7 Outsourcing database services

8 Big Data

Contents

1	**Distributed Databases & Client-Server Architectures**
2	Spatial and Temporal Database
3	Multimedia Databases
4	Geographic Information Systems
5	XML
6	Data Warehousing
7	Outsourcing database services
8	Big Data

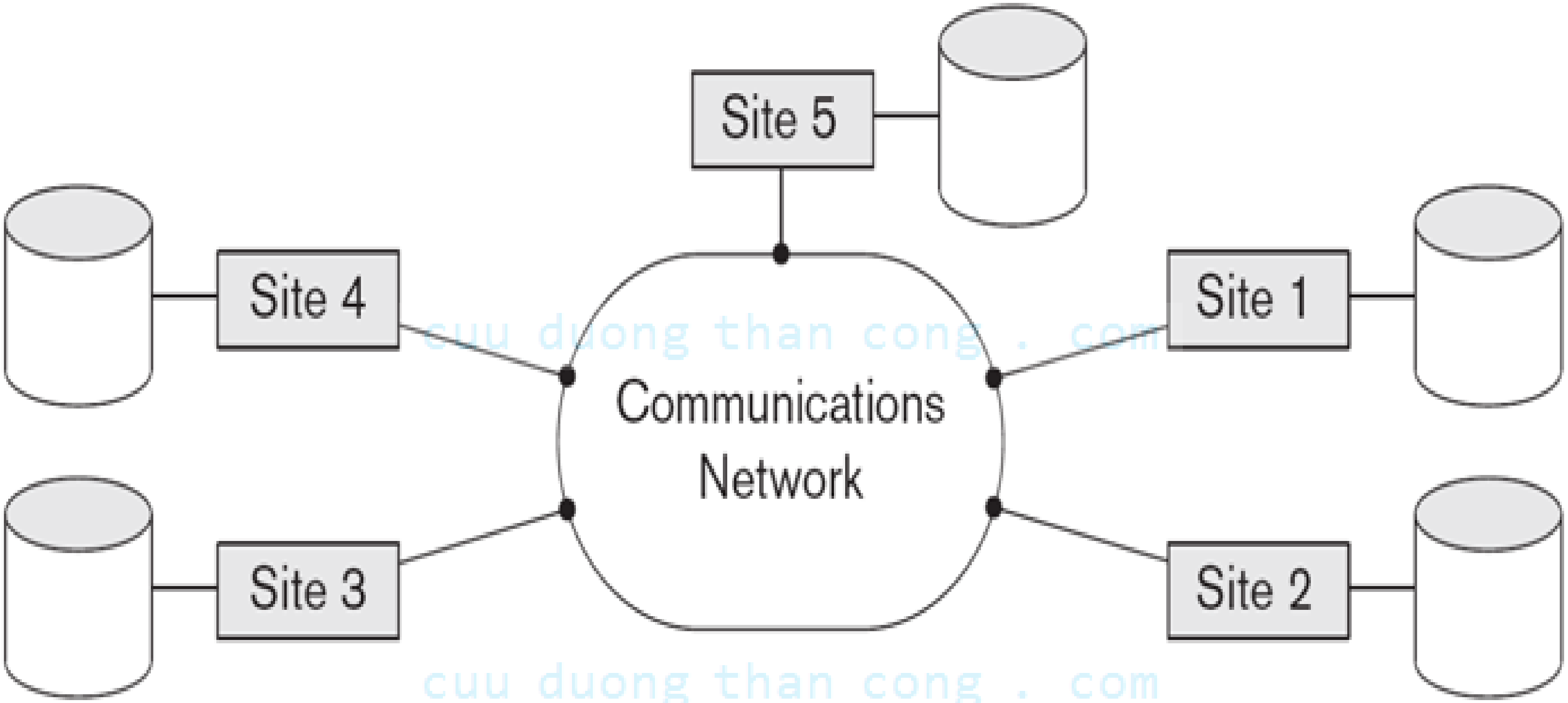
Distributed Databases & Client-Server Architectures

- Distributed Database Concepts
- Data Fragmentation, Replication and Allocation
- 3-Tier Client-Server Architecture

Distributed Database Concepts

- A transaction can be executed by multiple networked computers in a unified manner.
- A **distributed database (DDB)** processes a unit of execution (a transaction) in a distributed manner.
- DDB is a collection of multiple logically related database distributed over a computer network, and a distributed database management system as a software system that manages a distributed database while making the distribution ***transparent*** to the user.

Distributed Database System



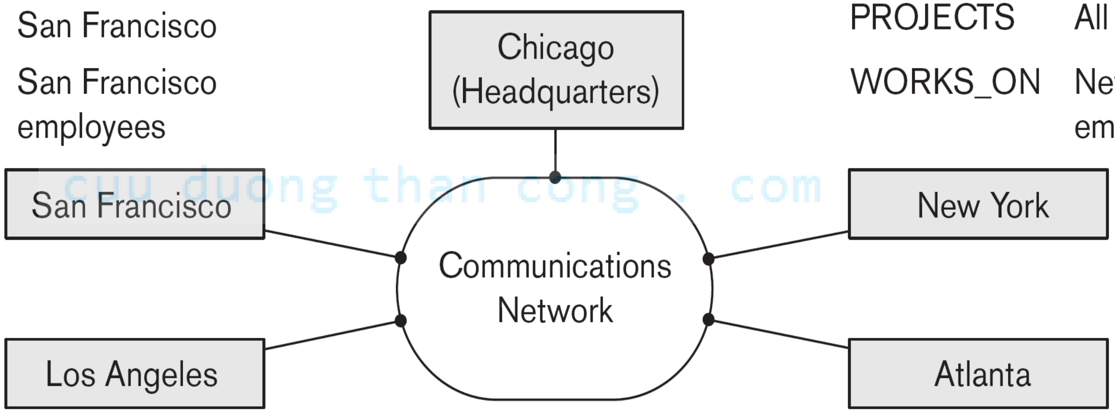
Distributed Database System

Data distribution and replication among distributed databases.

EMPLOYEES	San Francisco and Los Angeles
PROJECTS	San Francisco
WORKS_ON	San Francisco employees

EMPLOYEES	All
PROJECTS	All
WORKS_ON	All

EMPLOYEES	New York
PROJECTS	All
WORKS_ON	New York employees



EMPLOYEES	Los Angeles
PROJECTS	Los Angeles and San Francisco
WORKS_ON	Los Angeles employees

EMPLOYEES	Atlanta
PROJECTS	Atlanta
WORKS_ON	Atlanta employees

Distributed Database System

- Types of Transparency:
  - **Data organization transparency (Distribution and Network transparency)**
    - Users do not have to worry about operational details of the network.
    - *Location transparency* refers to freedom of issuing command from any location without affecting its working.
    - *Naming transparency* allows access to any names object (files, relations, etc.) from any location.

Distributed Database System

- Types of Transparency:
  - **Replication transparency**:
    - It allows to store copies of a data at multiple sites.
    - It minimizes access time to the required data.
  - **Fragmentation transparency**:
    - Allows to fragment a relation horizontally (create a subset of tuples of a relation) or vertically (create a subset of columns of a relation).

Distributed Database System

- Types of Transparency:
  - **Design transparency:**
    - Refer to freedom from knowing how the distributed database is designed
  - **Execution transparency:**
    - Refer to freedom from knowing where a transaction executes

Distributed Database System

- Advantages of Distributed Database System
  - **Improved ease and flexibility of application development**
    - Developing and maintaining applications at geographically distributed sites of an organization is facilitated owing to transparency of data distribution and control.

Distributed Database System

- Advantages of Distributed Database System
  - **Increased reliability and availability**:
    - Reliability refers to system live time, that is, system is running efficiently most of the time. Availability is the probability that the system is continuously available (usable or accessible) during a time interval.
    - A distributed database system has multiple nodes (computers) and if one fails then others are available to do the job.

Distributed Database System

- Advantages of Distributed Database System
  - **Improved performance**:
    - A distributed DBMS fragments the database to keep data closer to where it is needed most.
    - This reduces data management (access and modification) time significantly.
  - **Easier expansion (scalability)**:
    - Allows new nodes (computers) to be added anytime without chaining the entire configuration.

Data Fragmentation, Replication and Allocation

- **Data Fragmentation**
  - Split a relation into logically related and correct parts. A relation can be fragmented in two ways:
    - **Horizontal Fragmentation:** It is a horizontal subset of a relation which contain those of tuples which satisfy selection conditions
    - **Vertical Fragmentation:** It is a subset of a relation which is created by a subset of columns.

Data Fragmentation, Replication and Allocation

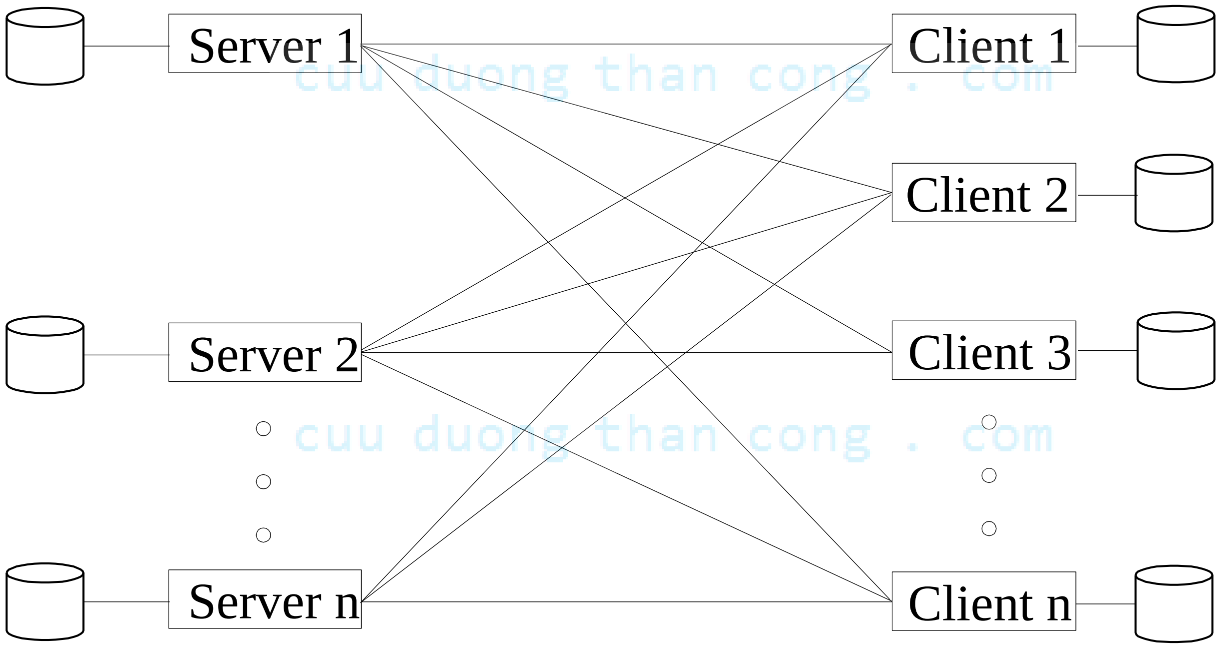
- **Fragmentation schema**
  - A definition of a set of fragments (horizontal or vertical or horizontal and vertical) that includes all attributes and tuples in the database that satisfies the condition that the whole database can be reconstructed from the fragments by applying some sequence of UNION (or OUTER JOIN) and UNION operations.
- **Allocation schema**
  - It describes the distribution of fragments to sites of distributed databases. It can be fully or partially replicated or can be partitioned.

Data Fragmentation, Replication and Allocation

- **Data Replication**
  - Database is replicated to all sites.
  - In full replication the entire database is replicated and in partial replication some selected part is replicated to some of the sites.
  - Data replication is achieved through a replication schema.
- **Data Distribution (Data Allocation)**
  - This is relevant only in the case of partial replication or partition.
  - The selected portion of the database is distributed to the database sites.

Client-Server Database Architecture

- It consists of clients running client software, a set of servers which provide all database functionalities and a reliable communication infrastructure.



Client-Server Database Architecture

- Clients reach server for desired service, but server does reach clients.
- The server software is responsible for local data management at a site, much like centralized DBMS software.
- The client software is responsible for most of the distribution function.
- The communication software manages communication among clients and servers.

Client-Server Database Architecture

- The processing of a SQL queries goes as follows:
  - Client parses a user query and decomposes it into a number of independent sub-queries. Each subquery is sent to appropriate site for execution.
  - Each server processes its query and sends the result to the client.
  - The client combines the results of subqueries and produces the final result.

Contents

1 Distributed Databases & Client-Server Architectures

2 Spatial and Temporal Database

3 Multimedia Databases

4 Geographic Information Systems

5 XML

6 Data Warehousing

7 Outsourcing database services

8 Big Data

Temporal Database Concepts

- Time Representation
- Calendars
- Time Dimensions

Temporal Database Concepts

- Time Representation
  - Time is considered ordered sequence of points in some granularity
  - Use the term chronon instead of point to describe minimum granularity
  - A calendar organizes time into different time units for convenience.
  - Accommodates various calendars
    - Gregorian (western), Chinese, Islamic, Hindu, etc.

Temporal Database Concepts

- Time Representation
  - Point events
    - Single time point event
      - E.g., bank deposit
    - Series of point events can form a time series data
  - Duration events
    - Associated with specific time period
    - Time period is represented by start time and end time

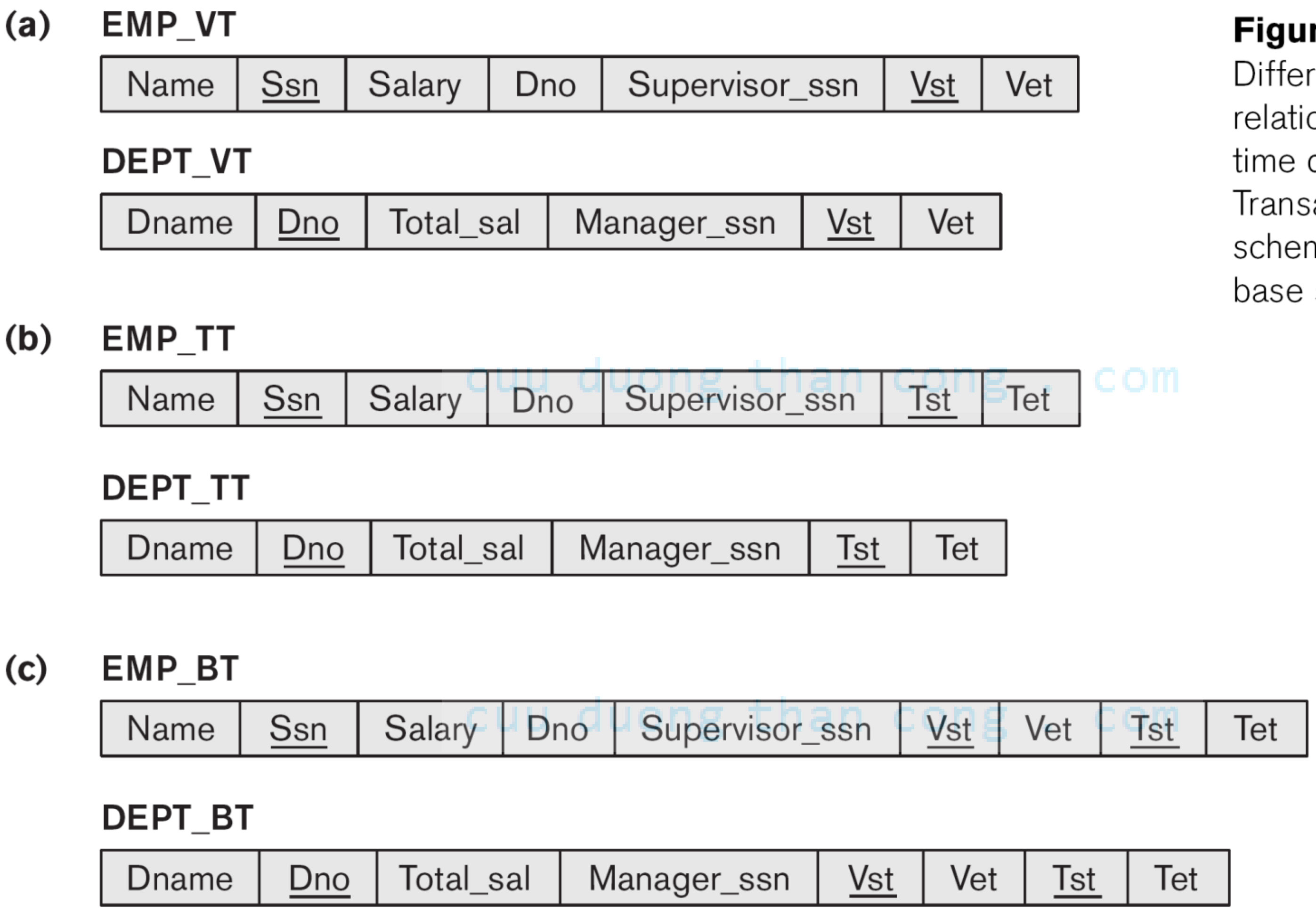
Temporal Database Concepts

- Time Representation
  - Transaction time
    - The time when the information from a certain transaction becomes valid
  - Bitemporal database
    - Databases dealing with two time dimensions

Temporal Database Concepts

- Incorporating Time in Relational Databases Using Tuple Versioning
  - Add to every tuple
    - Valid start time
    - Valid end time

Temporal Database Concepts



(a) EMP_VT

Name	Ssn	Salary	Dno	Supervisor_ssn	Vst	Vet

DEPT_VT

Dname	Dno	Total_sal	Manager_ssn	Vst	Vet

(b) EMP_TT

Name	Ssn	Salary	Dno	Supervisor_ssn	Tst	Tet

DEPT_TT

Dname	Dno	Total_sal	Manager_ssn	Tst	Tet

(c) EMP_BT

Name	Ssn	Salary	Dno	Supervisor_ssn	Vst	Vet	Tst	Tet

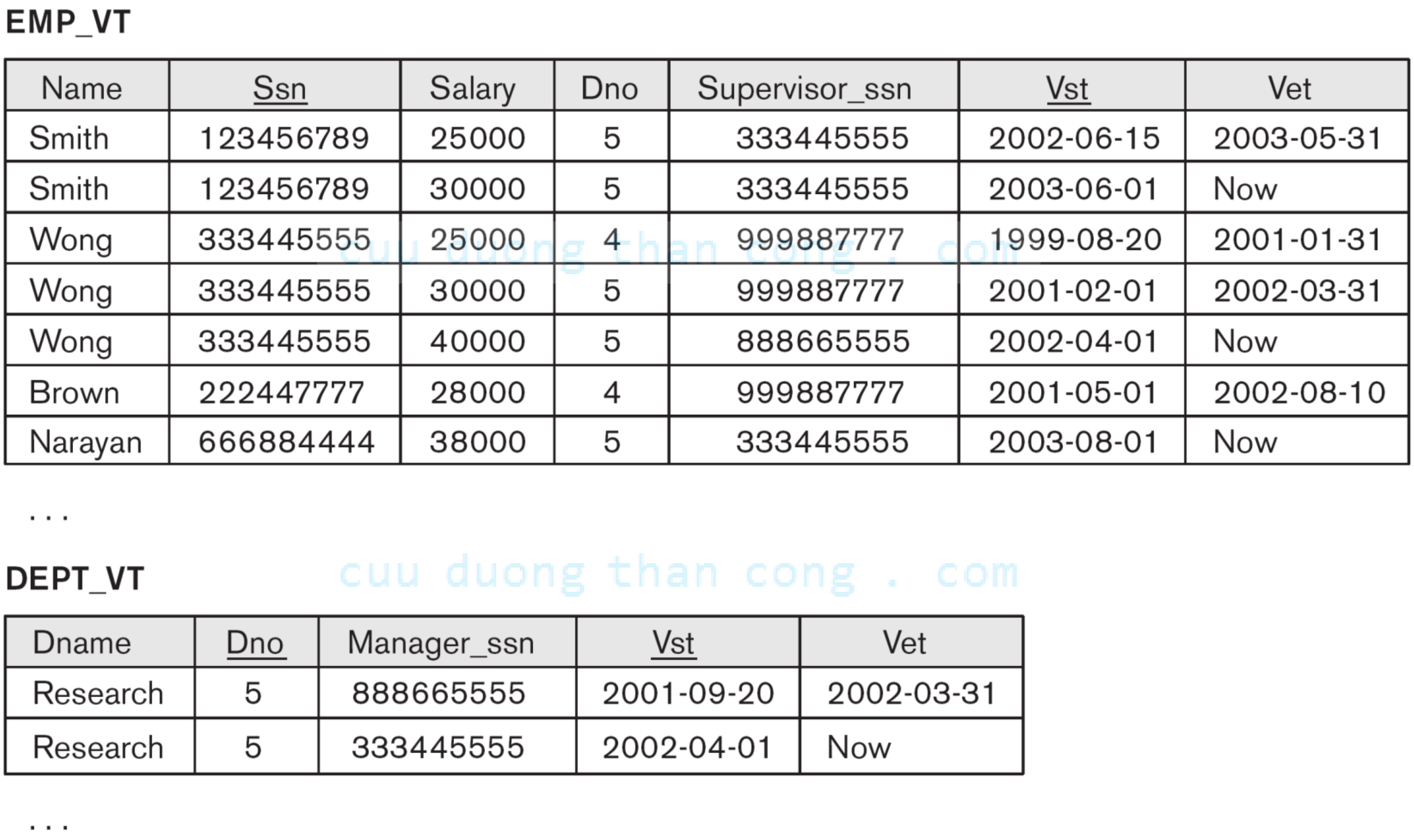
DEPT_BT

Dname	Dno	Total_sal	Manager_ssn	Vst	Vet	Tst	Tet

Figure 26.7
Different types of temporal relational databases. (a) Valid time database schema. (b) Transaction time database schema. (c) Bitemporal database schema.

Temporal Database Concepts

Figure 26.8
Some tuple versions in the valid time relations EMP_VT and DEPT_VT.



EMP_VT

Name	Ssn	Salary	Dno	Supervisor_ssn	Vst	Vet
Smith	123456789	25000	5	333445555	2002-06-15	2003-05-31
Smith	123456789	30000	5	333445555	2003-06-01	Now
Wong	333445555	25000	4	999887777	1999-08-20	2001-01-31
Wong	333445555	30000	5	999887777	2001-02-01	2002-03-31
Wong	333445555	40000	5	888665555	2002-04-01	Now
Brown	222447777	28000	4	999887777	2001-05-01	2002-08-10
Narayan	666884444	38000	5	333445555	2003-08-01	Now

. . .

DEPT_VT

Dname	Dno	Manager_ssn	Vst	Vet
Research	5	888665555	2001-09-20	2002-03-31
Research	5	333445555	2002-04-01	Now

. . .

Temporal Database Concepts

- Incorporating Time in Object-Oriented Databases Using Attribute Versioning
  - A single complex object stores all temporal changes of the object
  - Time varying attribute
    - An attribute that changes over time
    - E.g., salary
  - Non-Time varying attribute
    - An attribute that does not changes over time
    - E.g., date of birth

Temporal Database Concepts

```
class TEMPORAL_SALARY
{ attribute Date Valid_start_time;
attribute Date Valid_end_time;
attribute float Salary; };

class TEMPORAL_DEPT
{ attribute Date Valid_start_time;
attribute Date Valid_end_time;
attribute DEPARTMENT_VT Dept; };

class TEMPORAL_SUPERVISOR
{ attribute Date Valid_start_time;
attribute Date Valid_end_time;
attribute EMPLOYEE_VT Supervisor; };
```

Common operations used in queries

[T.Vst, T.Vet] INCLUDES [T1, T2]

⇔ T1 ≥ T.Vst AND T2 ≤ T.Vet

[T.Vst, T.Vet] INCLUDED_IN [T1, T2]

⇔ T1 ≤ T.Vst AND T2 ≥ T.Vet

[T.Vst, T.Vet] OVERLAPS [T1, T2]

⇔(T1 ≤ T.Vet AND T2 ≥ T.Vst)

[T.Vst, T.Vet] BEFORE [T1, T2] ⇔ T1 ≥ T.Vet

[T.Vst, T.Vet] AFTER [T1, T2] ⇔ T2 ≤ T.Vst

[T.Vst, T.Vet] MEETS_BEFORE [T1, T2] ⇔ T1 = T.Vet + 1

[T.Vst, T.Vet] MEETS_AFTER [T1, T2] ⇔ T2 + 1 = T.Vst

Spatial Database Concepts

- Keep track of objects in a multi-dimensional space
  - Maps
  - Geographical Information Systems (**GIS**)
  - Weather
- In general spatial databases are n-dimensional
  - This discussion is limited to 2-dimensional spatial databases

Spatial Databases

- Typical Spatial Queries
  - **Range** query: Finds objects of a particular type within a particular distance from a given location
    - Example, find all hospitals within the M.A. city area, or find all ambulances within five miles of an accident location.
  - **Nearest Neighbor** query: Finds objects of a particular type that is nearest to a given location
    - Example, find the police car that is closest to the location of crime.
  - **Spatial joins** or overlays: Joins objects of two types based on some spatial condition (intersecting, overlapping, within certain distance, etc.)
    - Example, find all homes that are within two miles of a lake

Contents

1	Distributed Databases & Client-Server Architectures
2	Spatial and Temporal Database
3	**Multimedia Databases**
4	Geographic Information Systems
5	XML
6	Data Warehousing
7	Outsourcing database services
8	Big Data

Multimedia Databases

- In the years ahead multimedia information systems are expected to dominate our daily lives.
  - Our houses will be wired for bandwidth to handle interactive multimedia applications.
  - Our high-definition TV/computer workstations will have access to a large number of databases, including digital libraries, image and video databases that will distribute vast amounts of multisource multimedia content.

Multimedia Databases

- Types of multimedia data are available in current systems
  - **Text**: May be formatted or unformatted. For ease of parsing structured documents, standards like SGML and variations such as HTML are being used.
  - **Graphics**: Examples include drawings and illustrations that are encoded using some descriptive standards (e.g. CGM, PICT, postscript).

Multimedia Databases

- Types of multimedia data are available in current systems (cont.)
  - **Images**: Includes drawings, photographs, and so forth, encoded in standard formats such as bitmap, JPEG, and MPEG. Compression is built into JPEG and MPEG.
    - These images are not subdivided into components. Hence querying them by content (e.g., find all images containing circles) is nontrivial.
  - **Animations**: Temporal sequences of image or graphic data.

Multimedia Databases

- Types of multimedia data are available in current systems (cont.)
  - **Video**: A set of temporally sequenced photographic data for presentation at specified rates– for example, 30 frames per second.
  - **Structured audio**: A sequence of audio components comprising note, tone, duration, and so forth.

Multimedia Databases

- Types of multimedia data are available in current systems (cont.)
  - **Audio**: Sample data generated from aural recordings in a string of bits in digitized form. Analog recordings are typically converted into digital form before storage.

Multimedia Databases

- Types of multimedia data are available in current systems (cont.)
  - **Composite** or mixed multimedia data: A combination of multimedia data types such as audio and video which may be physically mixed to yield a new storage format or logically mixed while retaining original types and formats. Composite data also contains additional control information describing how the information should be rendered.

Multimedia Databases

- Multimedia applications dealing with thousands of images, documents, audio and video segments, and free text data depend critically on

  - Appropriate modeling of the structure and content of data
  - Designing appropriate database schemas for storing and retrieving multimedia information.

Contents

1 Distributed Databases & Client-Server Architectures

2 Spatial and Temporal Database

3 Multimedia Databases

4 Geographic Information Systems

5 XML

6 Data Warehousing

7 Outsourcing database services

8 Big Data

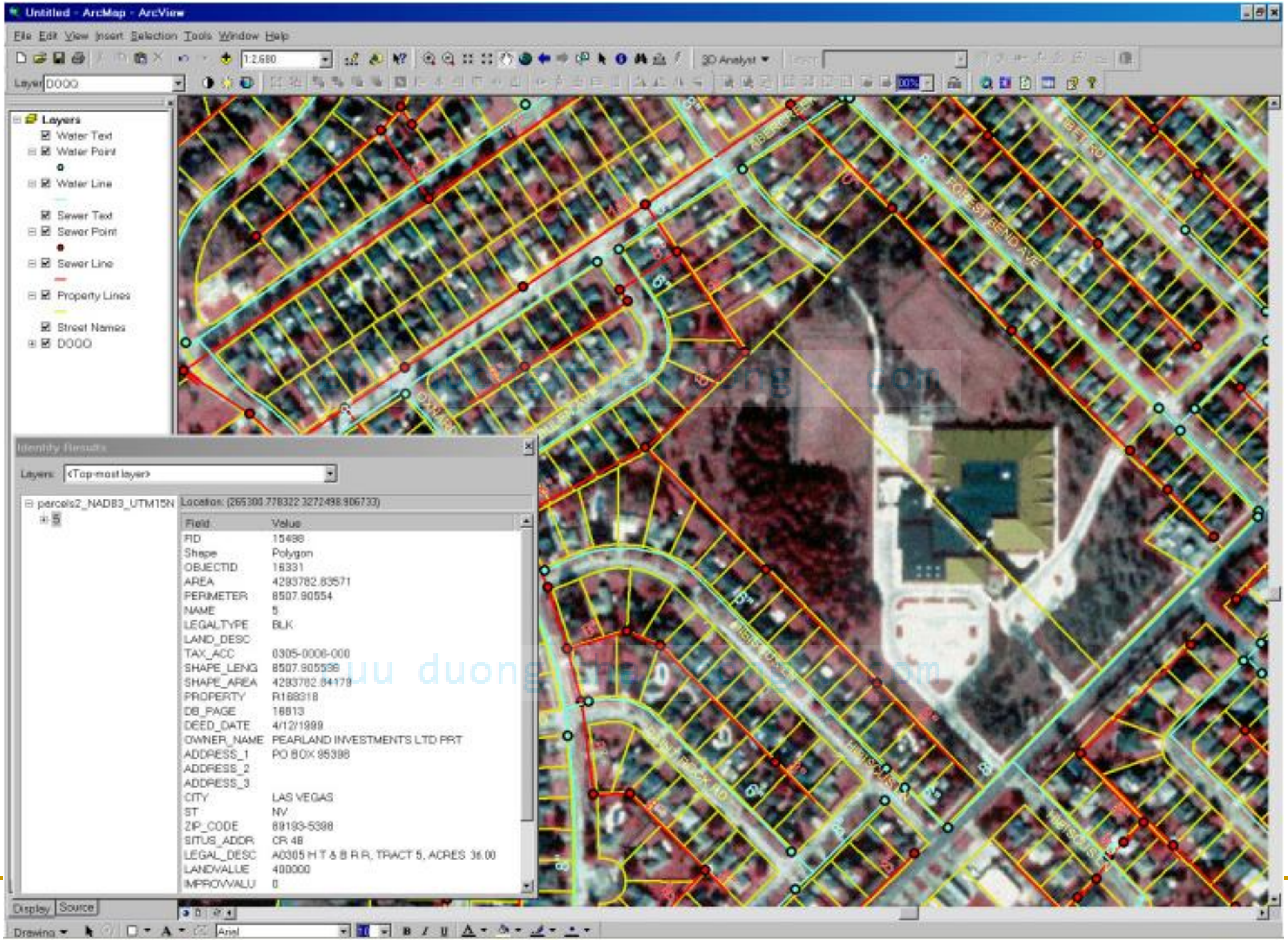
Geographic Information Systems

- Geographic information systems(GIS) are used to collect, model, and analyze information describing physical properties of the geographical world.

Geographic Information Systems

- The scope of GIS broadly encompasses two types of data:
  - **Spatial** data, originating from maps, digital images, administrative and political boundaries, roads, transportation networks, physical data, such as rivers, soil characteristics, climatic regions, land elevations, and
  - **Non-spatial** data, such as socio-economic data (like census counts), economic data, and sales or marketing information. GIS is a rapidly developing domain that offers highly innovative approaches to meet some challenging technical demands.

Geographic Information Systems



Identify Results

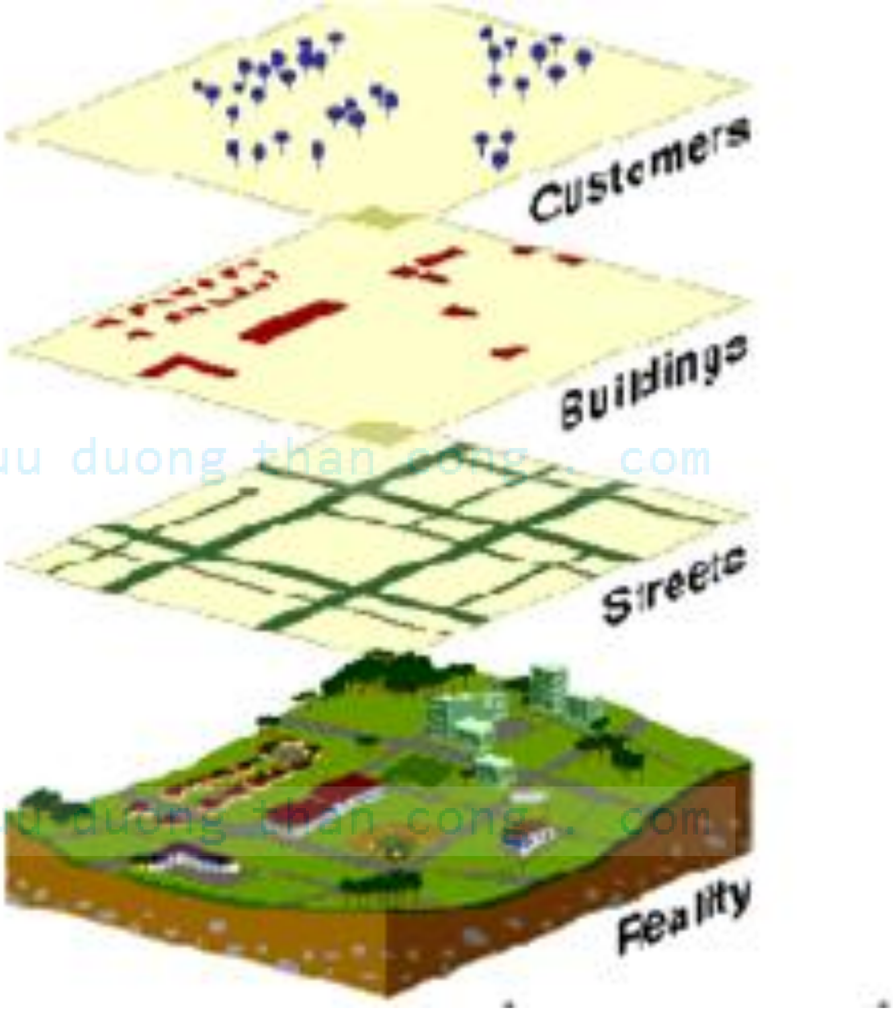
Layers: <Top-most layer>

parcels2_NAD83_UTM15N
5

Location: (265300.778322 3272498.906733)

Field	Value
FID	15498
Shape	Polygon
OBJECTID	16331
AREA	4283782.83571
PERIMETER	8507.90554
NAME	5
LEGALTYPE	BLK
LAND_DESC	
TAX_ACC	0305-0006-000
SHAPE_LENG	8507.905538
SHAPE_AREA	4283782.84178
PROPERTY	R168318
DB_PAGE	16813
DEED_DATE	4/12/1999
OWNER_NAME	PEARLAND INVESTMENTS LTD PRT
ADDRESS_1	PO BOX 85398
ADDRESS_2	
ADDRESS_3	
CITY	LAS VEGAS
ST	NV
ZIP_CODE	89193-5398
SITUS_ADDR	CR 48
LEGAL_DESC	A0305 H T & B R R, TRACT 5, ACRES 36.00
LANDVALUE	400000
IMPROVVALU	0

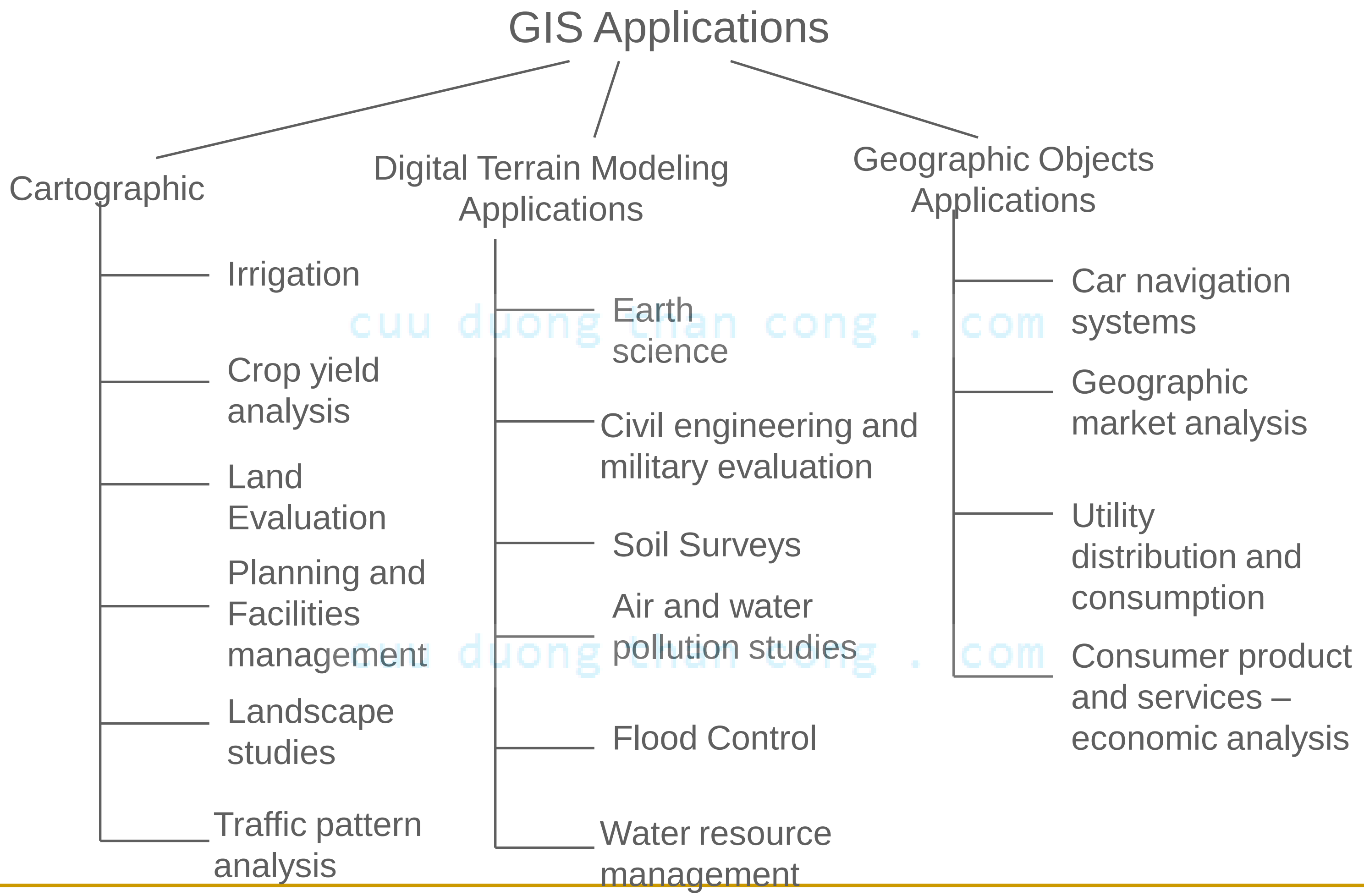
Spatial data



GIS Applications

- It is possible to divide GISs into three categories:
  - Cartographic applications
  - Digital terrain modeling applications
  - Geographic objects applications

GIS Applications(2)



GIS Applications

Cartographic
- Irrigation
- Crop yield analysis
- Land Evaluation
- Planning and Facilities management
- Landscape studies
- Traffic pattern analysis

Digital Terrain Modeling Applications
- Earth science
- Civil engineering and military evaluation
- Soil Surveys
- Air and water pollution studies
- Flood Control
- Water resource management

Geographic Objects Applications
- Car navigation systems
- Geographic market analysis
- Utility distribution and consumption
- Consumer product and services – economic analysis

Data Modeling and Representation

- GIS data can be broadly represented in two formats:
  - **Vector** data represents geometric objects such as points, lines, and polygons.
  - **Raster** data is characterized as an array of points, where each point represents the value of an attribute for a real-world location.

Specific GIS Data Operations

- The functionality of a GIS database is also subject to other considerations:
  - Extensibility
  - Data quality control
  - Visualization

- Such requirements clearly illustrate that standard RDBMSs or ODBMSs do not meet the special needs of GIS.
  - Therefore it is necessary to design systems that support the vector and raster representations and the spatial functionality as well as the required DBMS features.

Contents

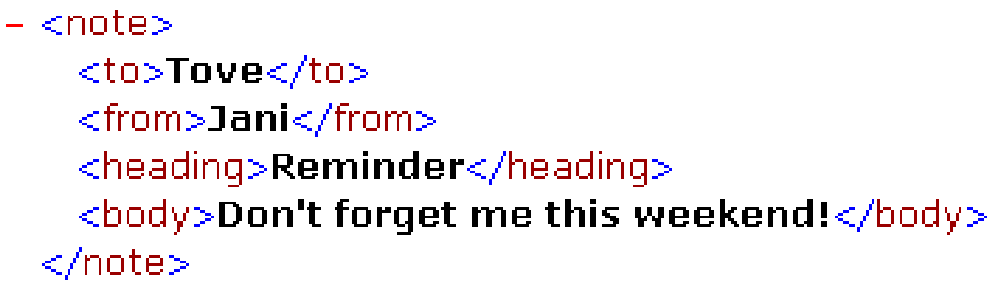
1	Distributed Databases & Client-Server Architectures
2	Spatial and Temporal Database
3	Multimedia Databases
4	Geographic Information Systems
5	**XML**
6	Data Warehousing
7	Outsourcing database services
8	Big Data

XML: Extensible Markup Language

- Although **HTML** is widely used for formatting and structuring *Web documents*, it is not suitable for specifying *structured data* that is extracted from databases.
- A new language—namely **XML** (eXtended Markup Language) has emerged as the standard for structuring and exchanging data over the Web.
  - XML can be used to provide more information about the structure and meaning of the data in the Web pages rather than just specifying how the Web pages are formatted for display on the screen.

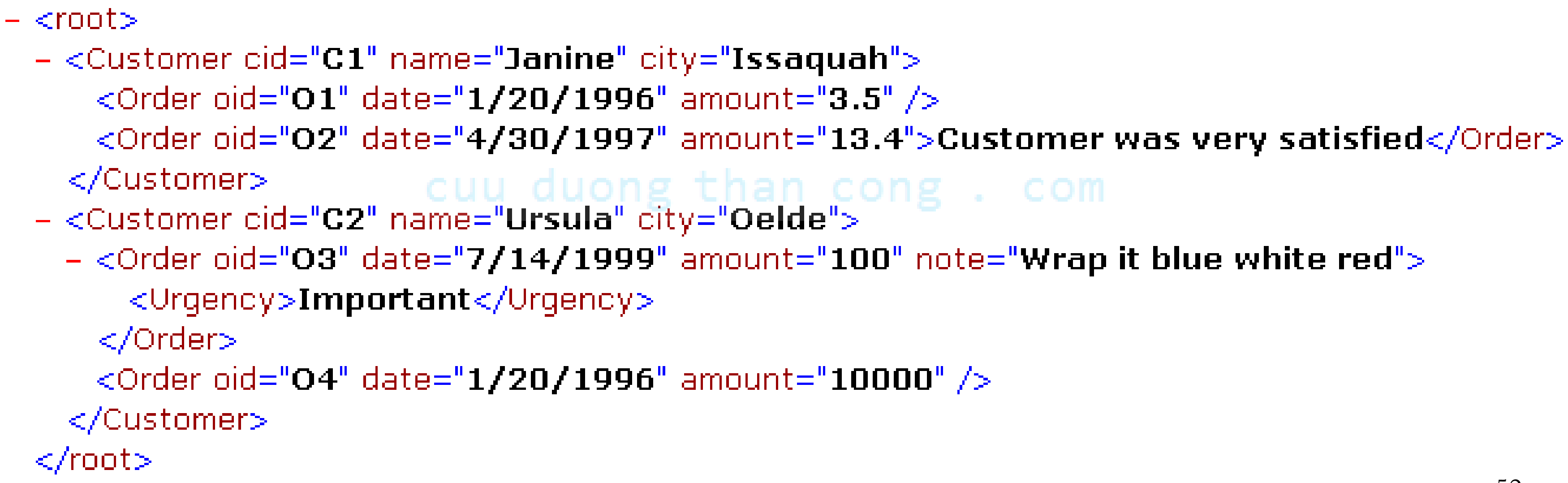
XML

- Example1:



```
- <note>
    <to>Tove</to>
    <from>Jani</from>
    <heading>Reminder</heading>
    <body>Don't forget me this weekend!</body>
  </note>
```

- Example2:



```
- <root>
  - <Customer cid="C1" name="Janine" city="Issaquah">
      <Order oid="O1" date="1/20/1996" amount="3.5" />
      <Order oid="O2" date="4/30/1997" amount="13.4">Customer was very satisfied</Order>
    </Customer>
  - <Customer cid="C2" name="Ursula" city="Oelde">
    - <Order oid="O3" date="7/14/1999" amount="100" note="Wrap it blue white red">
        <Urgency>Important</Urgency>
      </Order>
      <Order oid="O4" date="1/20/1996" amount="10000" />
    </Customer>
  </root>
```

XML

- The basic object is XML is the **XML document**.
- There are two main structuring concepts that are used to construct an XML document:
  - **Elements**
  - **Attributes**
- Attributes in XML provide additional information that describe elements.

XML

- Elements are identified in a document by their **start tag** and **end tag**.
  - The tag names are enclosed between angled brackets <…>, and end tags are further identified by a backslash </…>.
- **Complex** elements are constructed from other elements hierarchically, whereas **simple** elements contain data values.
- It is straightforward to see the correspondence between the XML textual representation and the tree structure.
  - In the tree representation, internal nodes represent complex elements, whereas leaf nodes represent simple elements.
  - That is why the XML model is called a **tree** model or a **hierarchical** model.

Contents

1	Distributed Databases & Client-Server Architectures
2	Spatial and Temporal Database
3	Multimedia Databases
4	Geographic Information Systems
5	XML
6	**Data Warehousing**
7	Outsourcing database services
8	Big Data

Data Warehousing

- The data warehouse is a historical database designed for decision support.
- Data mining can be applied to the data in a warehouse to help with certain types of decisions.
- Proper construction of a data warehouse is fundamental to the successful use of data mining.

Data Warehousing

- Purpose of Data Warehousing
  - Traditional databases are not optimized for data access only they have to balance the requirement of data access with the need to ensure integrity of data.
  - Most of the times the data warehouse users need only read access but, need the access to be fast over a large volume of data.
  - Most of the data required for **data warehouse** analysis comes from multiple databases and these analysis are recurrent and predictable to be able to design specific software to meet the requirements

Data Warehousing

- Applications that data warehouse supports are:
  - **OLAP** (Online Analytical Processing) is a term used to describe the analysis of complex data from the data warehouse.
  - **DSS** (Decision Support Systems) also known as EIS (Executive Information Systems) supports organization’s leading decision makers for making complex and important decisions.
  - **Data Mining** is used for knowledge discovery, the process of searching data for unanticipated new knowledge.

Definitions of Data Mining

- The discovery of new information in terms of patterns or rules from vast amounts of data.
- The process of finding interesting structure in data.

- The process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data.

Knowledge Discovery in Databases (KDD)

- Data mining is actually one step of a larger process known as **knowledge discovery in databases** (KDD).
- The KDD process model comprises six phases
  - Data selection
  - Data cleansing
  - Enrichment
  - Data transformation or encoding
  - Data mining
  - Reporting and displaying discovered knowledge

Comparison with Traditional Databases

- Data Warehouses are mainly optimized for appropriate data access.
  - Traditional databases are transactional and are optimized for both access mechanisms and integrity assurance measures.
- Data warehouses emphasize more on historical data as their main purpose is to support time-series and trend analysis.
- Compared with transactional databases, data warehouses are nonvolatile.
- In transactional databases transaction is the mechanism change to the database. By contrast information in data warehouse is relatively coarse grained and refresh policy is carefully chosen, usually incremental.

Contents

1 Distributed Databases & Client-Server Architectures

2 Spatial and Temporal Database

3 Multimedia Databases

4 Geographic Information Systems

5 XML

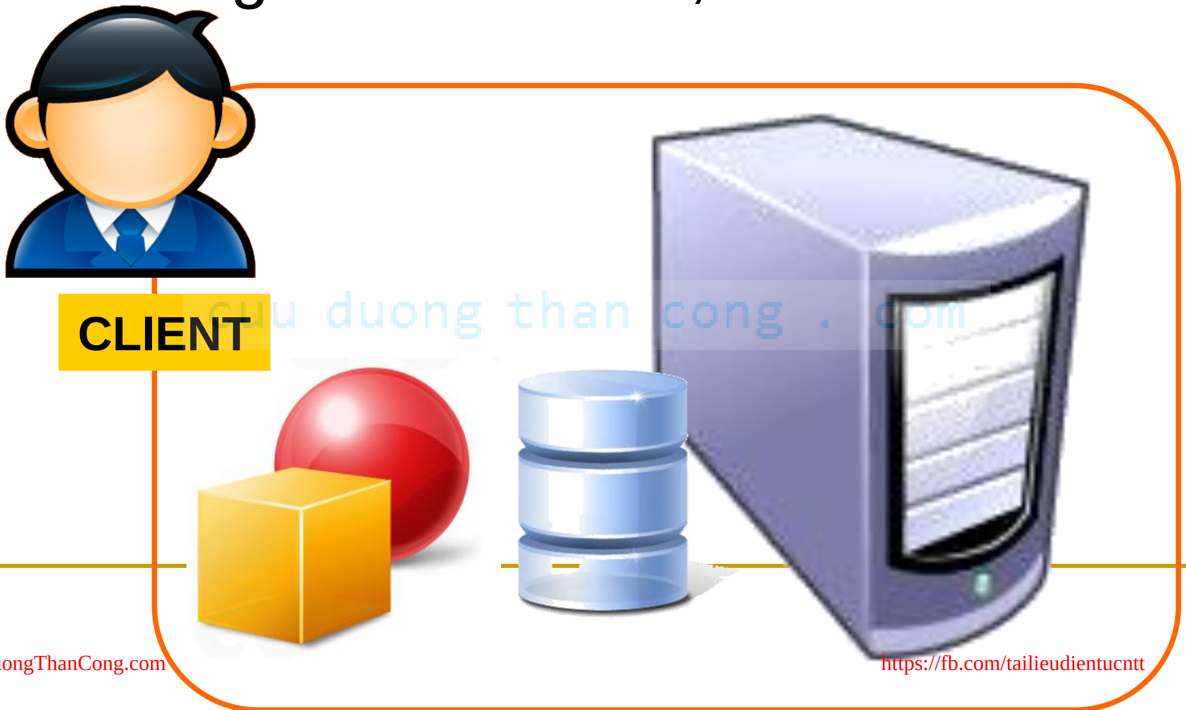
6 Data Warehousing

7 Outsourcing database services

8 Big Data

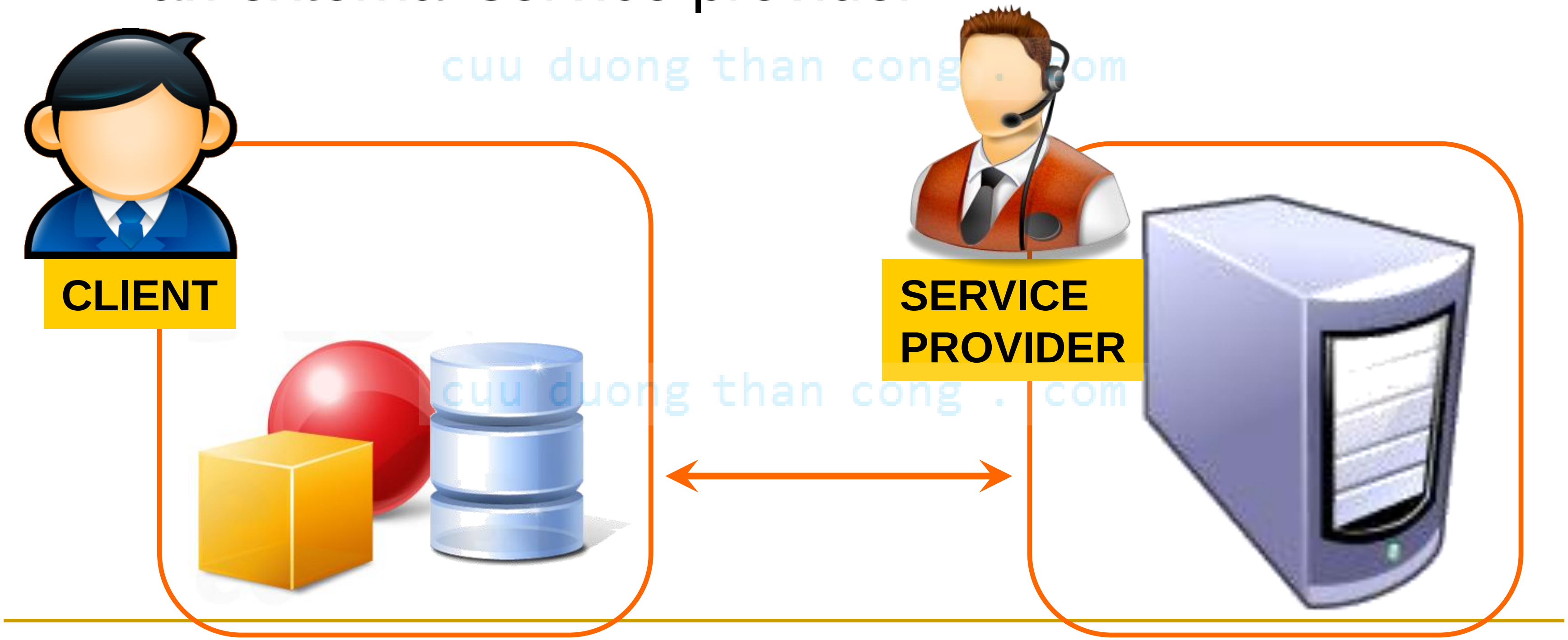
Introduction to Outsourcing Database Services (ODBS)

- Traditional model:
  - Client owns and manages database server
  - Benefits: Full access control
  - Disadvantages: Initial cost, maintenance cost



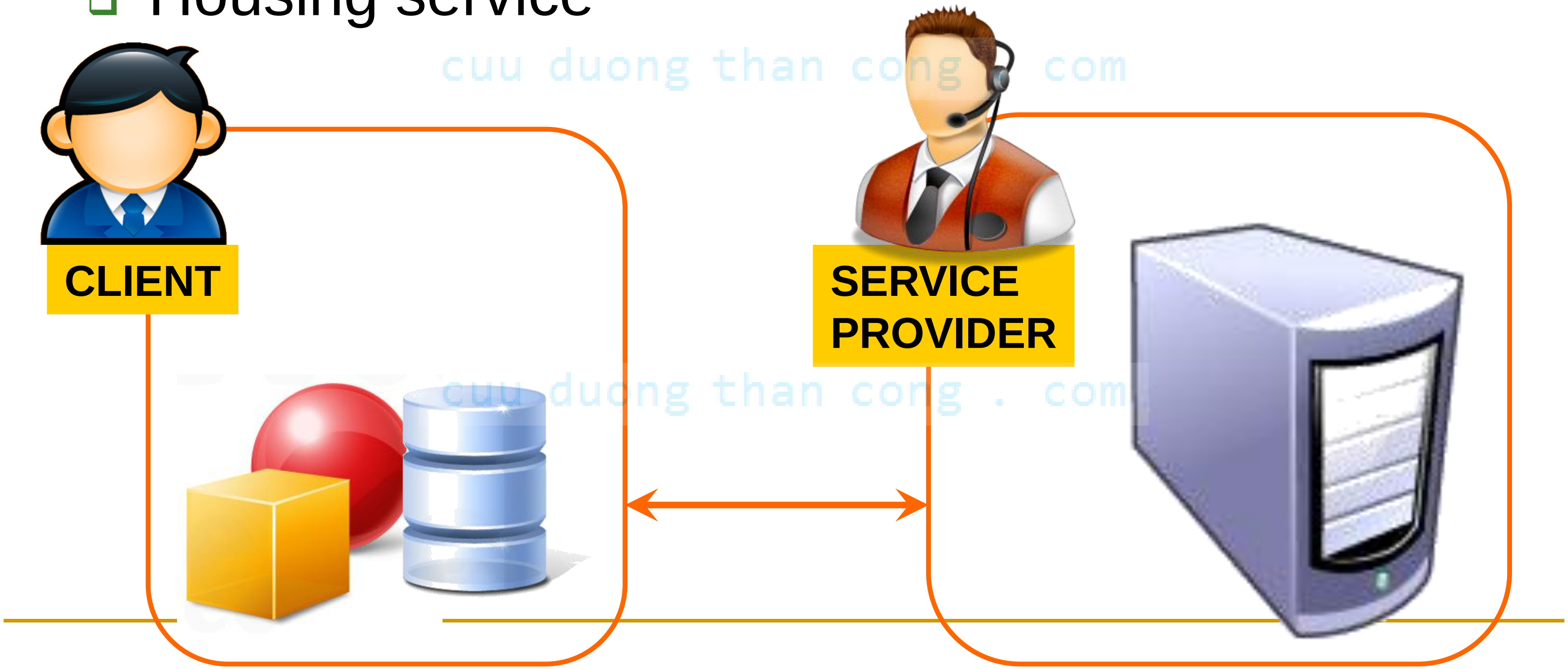
Introduction to Outsourcing Database Services (ODBS)

- Outsourcing database model
  - Client outsources his data management needs to an external service provider



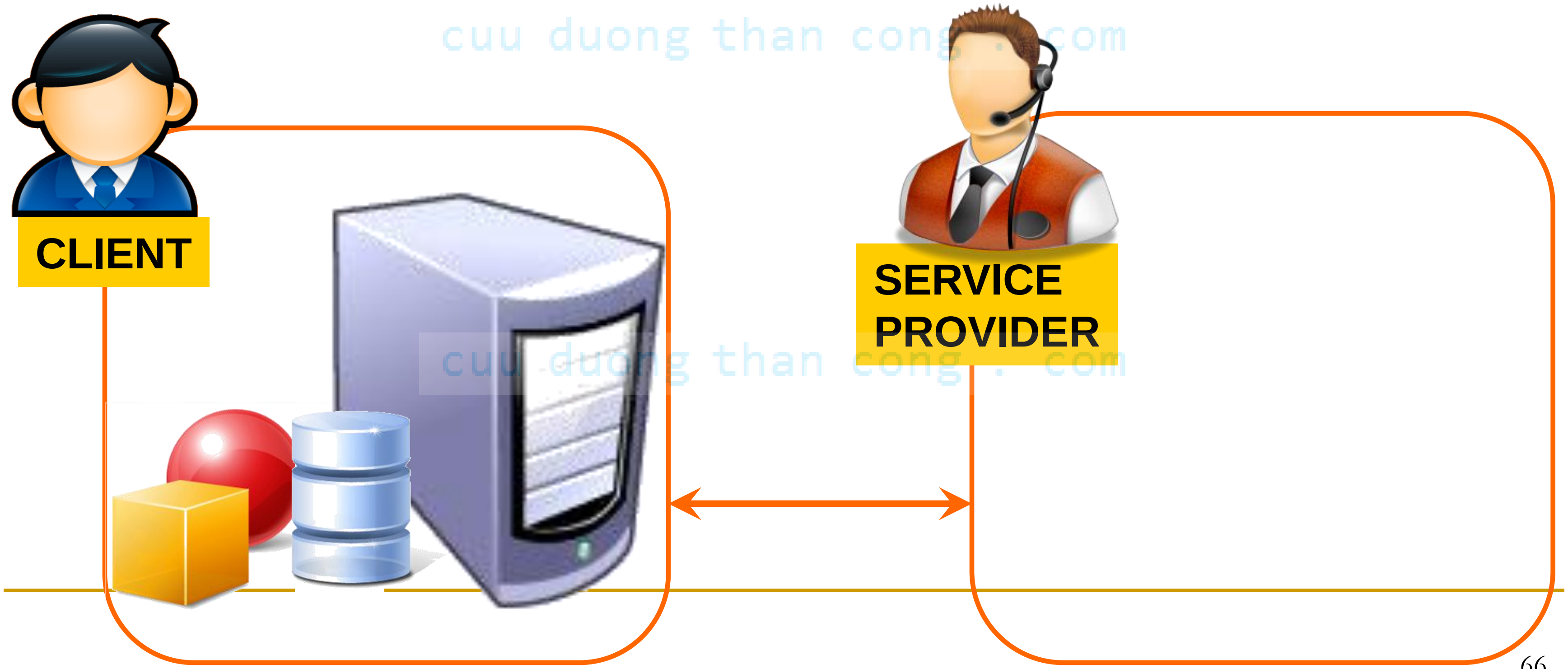
Introduction to Outsourcing Database Services (ODBS)

- Two categories:
  - **Hosting service**
  - Housing service



Introduction to Outsourcing Database Services (ODBS)

- Two categories:
  - Hosting service
  - **Housing service**



Some Database Outsourcing Vendors

- OBM
- Oracle
- EDS
- DbaDirect
- Ntirety

- Pythian
- TCS
- Satyam
- Wipro

Benefits of Outsourcing Database

- Save money:
  - Initial cost: hardware and software resources, facilities, technical staff
  - Maintenance cost
- Concentrate on core business
- Save time to set up the database system
- Share expertise
- Stable environments, with minimal changes
- Get resources that are not available internally
- …

… And Challenges

- Poor response time, poor turnaround time
- Hidden cost for advance services
- Quality of service
- Communication issues
- Lack of depth in troubleshooting
- Lack of full access control
- …

Contents

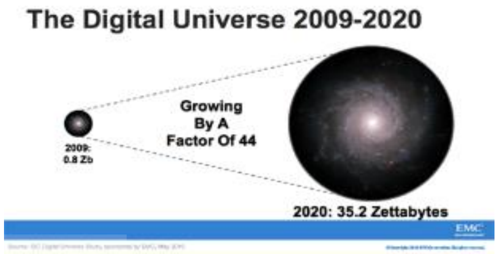
1	Distributed Databases & Client-Server Architectures
2	Spatial and Temporal Database
3	Multimedia Databases
4	Geographic Information Systems
5	XML
6	Data Warehousing
7	Outsourcing database services
8	**Big Data**

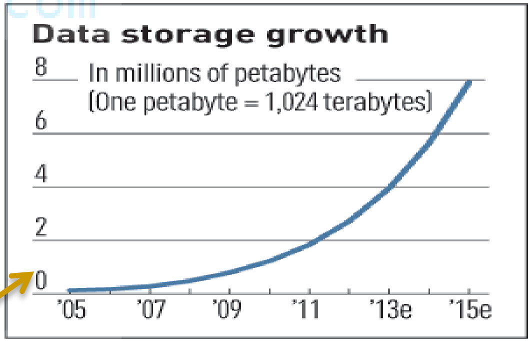
Big Data Definition

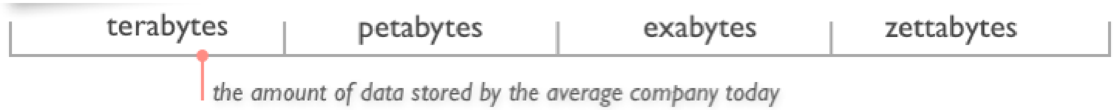
- Big data refers to large datasets that are challenging to store, search, share, visualize, and analyze.
- Big data is not a single technology but a combination of old and new technologies that helps companies gain actionable insight.
- “**Big data**” is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction.

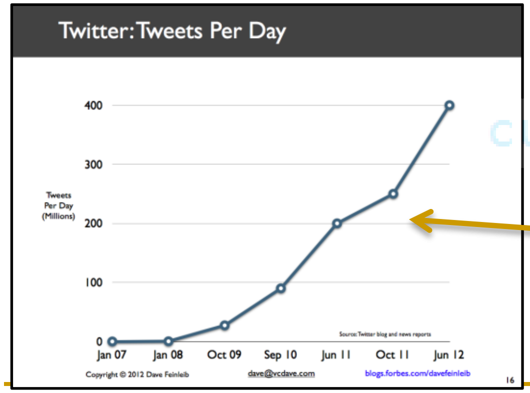
Characteristics of Big Data:
1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 to 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially







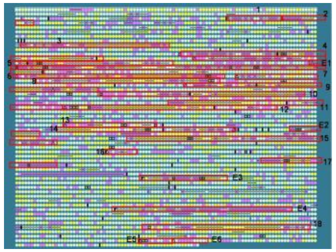


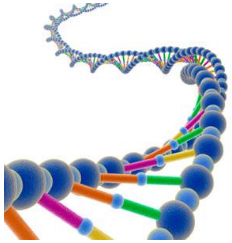
Exponential increase in collected/generated data

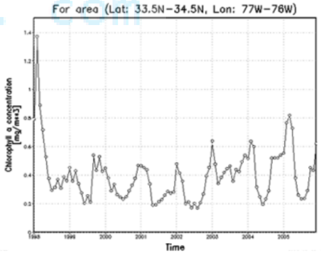
Characteristics of Big Data:
2-Complexity (Varity)

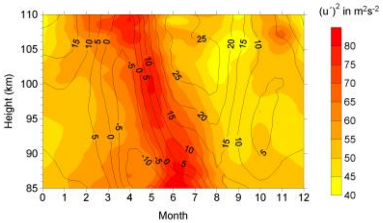
- Various formats, types, and structures.
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data.

To extract knowledge➔ all these types of data need to linked together






For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m**3]
Time

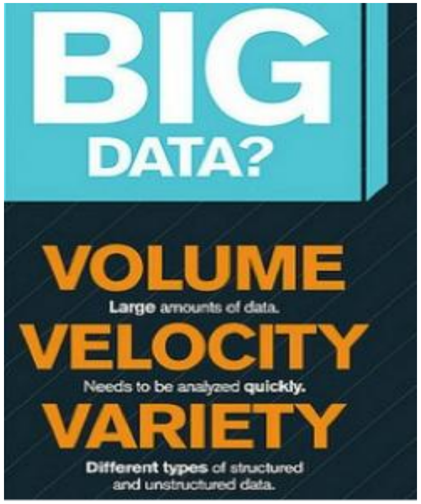


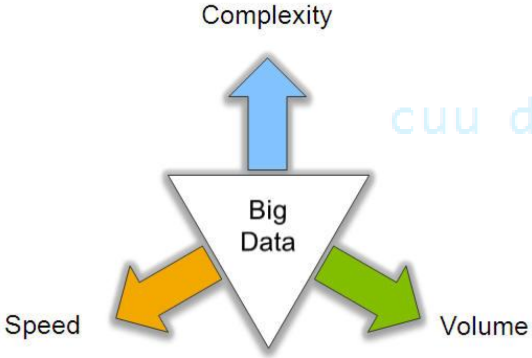
Characteristics of Big Data:
3-Speed (Velocity)

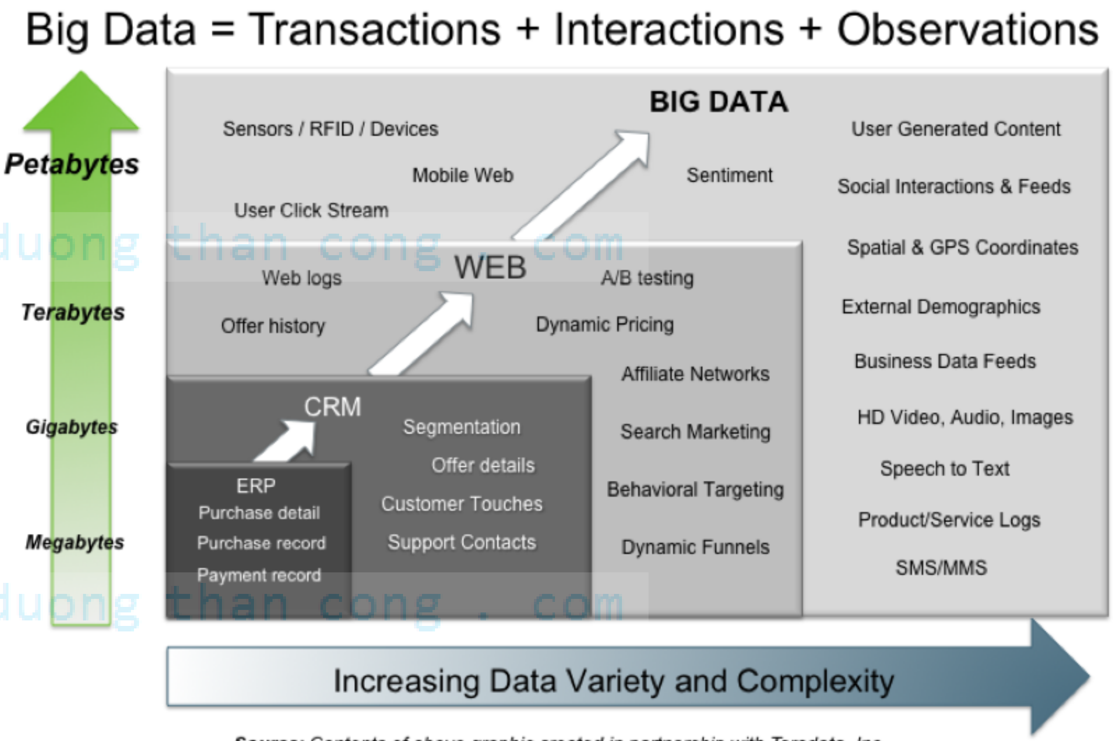
- Data is begin generated fast and need to be processed fast.

- Online Data Analytics.
- Late decisions ➔ missing opportunities.
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you.
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction.

Big Data: 3V's

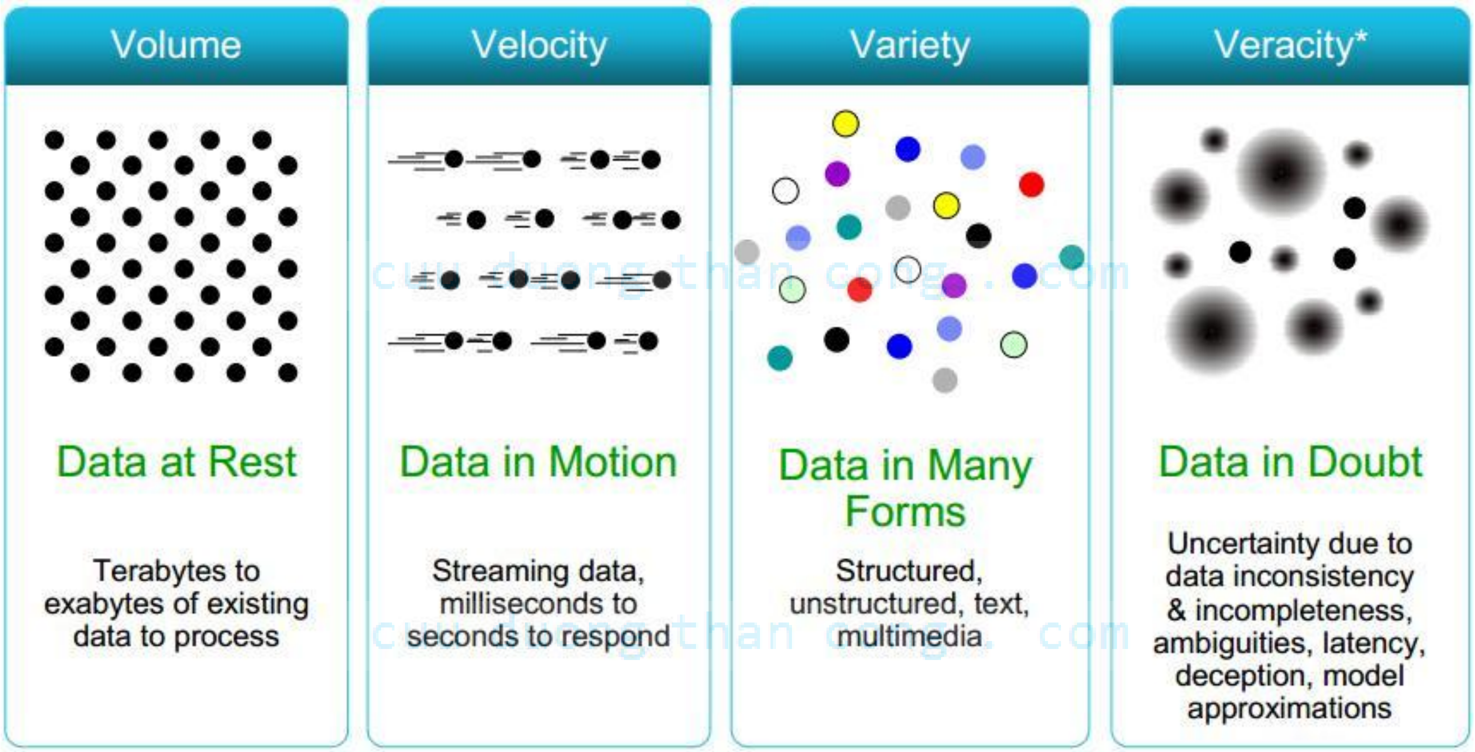

BIG DATA?
VOLUME
Large amounts of data.
VELOCITY
Needs to be analyzed quickly.
VARIETY
Different types of structured and unstructured data.





Source: *Contents of above graphic created in partnership with Teradata, Inc.*

Some Make it 4V's



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

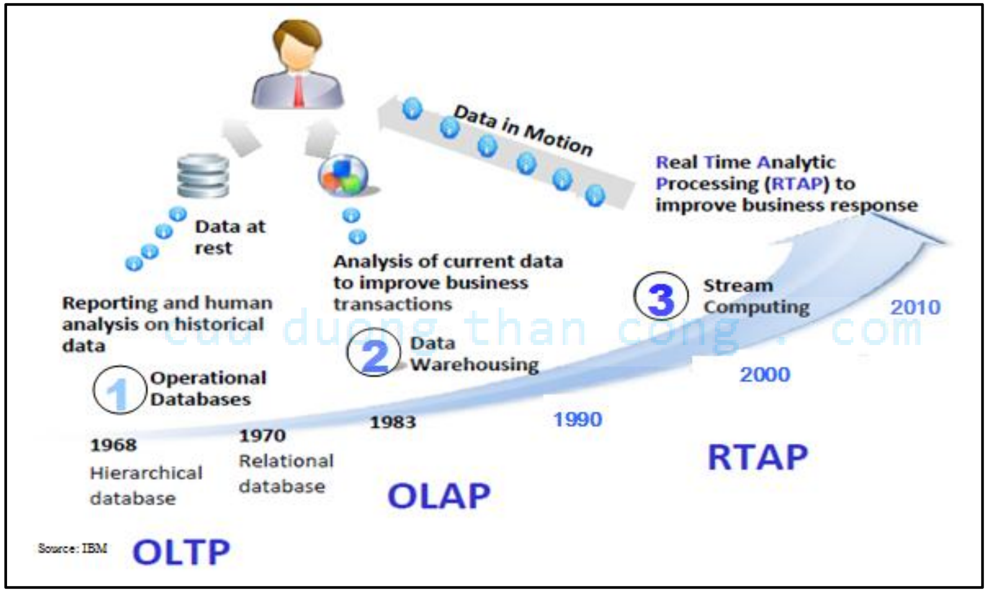
Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Harnessing Big Data


Data in Motion
Real Time Analytic Processing (RTAP) to improve business response
Data at rest
Analysis of current data to improve business transactions
Reporting and human analysis on historical data
3 Stream Computing
2010
2 Data Warehousing
2000
1 Operational Databases
1983
1990
1968 Hierarchical database
1970 Relational database
RTAP
OLAP
OLTP

Source: IBM

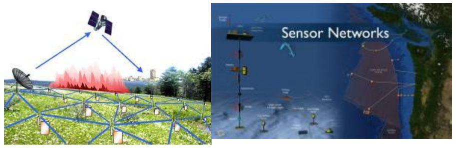
- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Who’s Generating Big Data?

Mobile devices
(tracking all objects all the time)

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

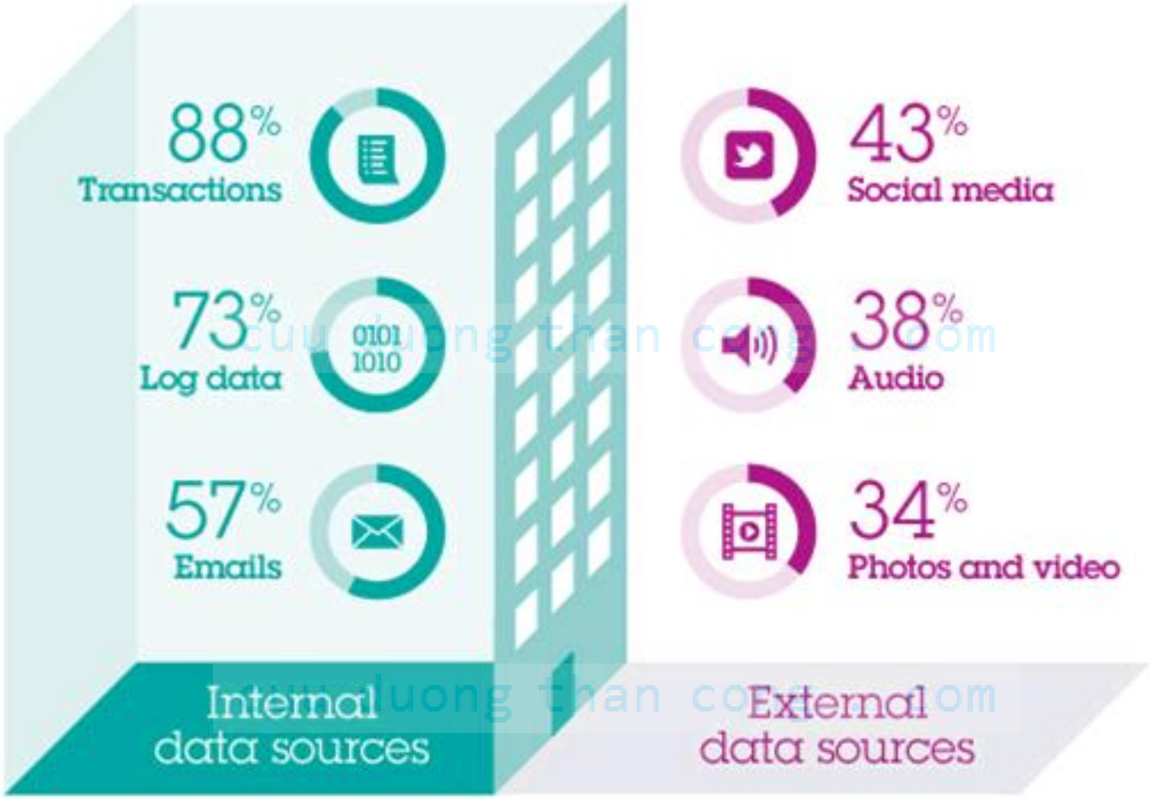


Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion.

Where does big data come from?

Most big data efforts are currently focused on analyzing internal data to extract insights. Fewer organizations are looking at data outside their firewalls, such as social media.



IBM

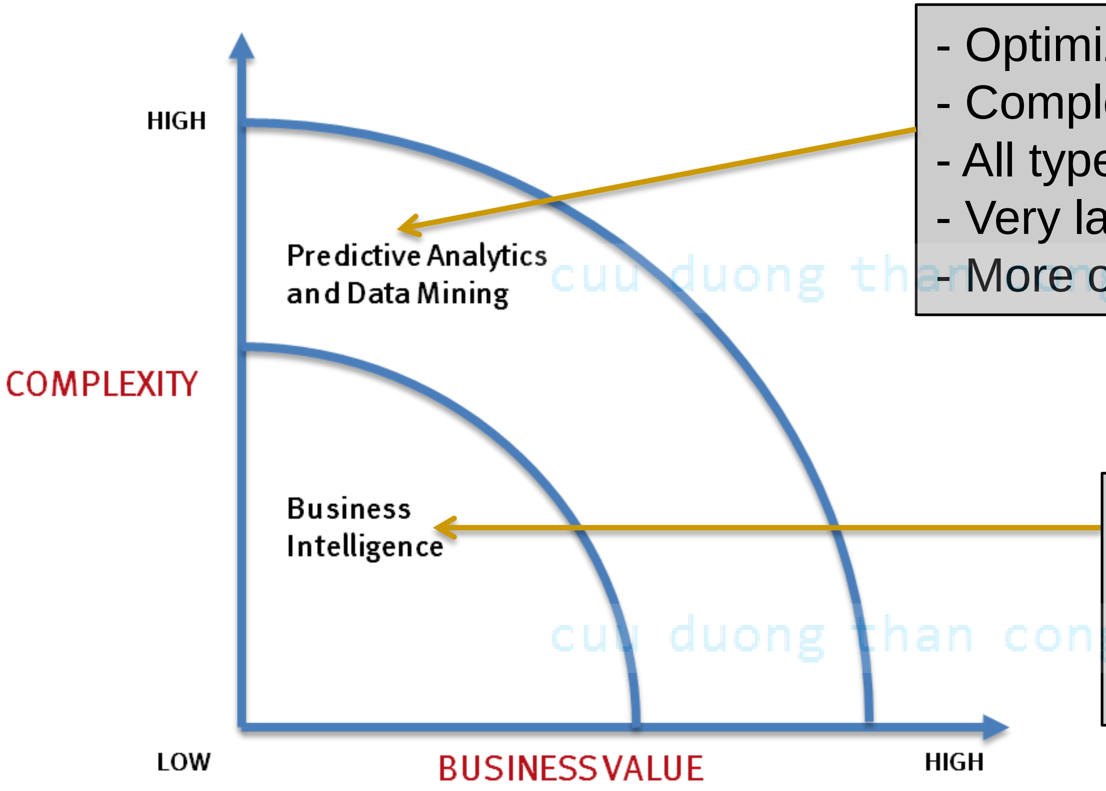
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

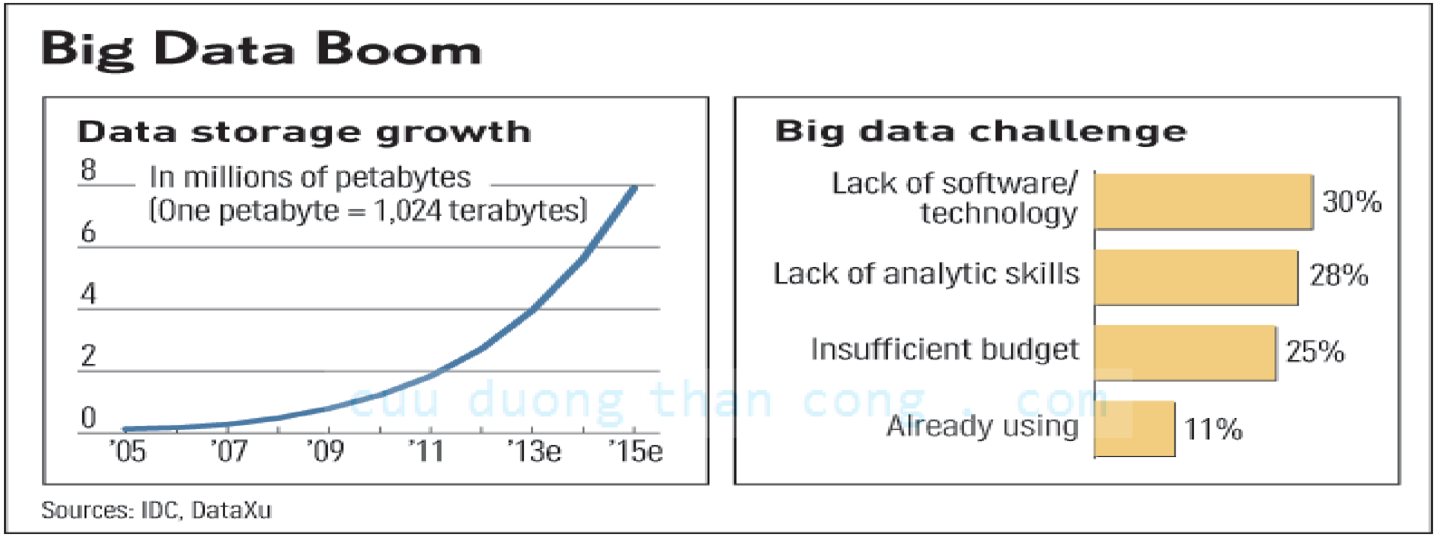
What’s driving Big Data?



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed.

- **Also in technical skills**
  - Experts in using the new technology and dealing with big data.

Big Data Platforms

- Data Integration
  - Informatica, Infosphere
  - talenD, Pentaho, Karmasphere, Apache Sqoop, Apache Flume
- Database Framework
  - Hadoop (Distributions: Cloudera, Hortonworks, MapR)
  - Hbase
  - Hive
- NoSQL Databases
  - MongoDB, CouchDB
- Machine Data Processing
  - Splunk, Mahout
- Text Analytics
  - Clarabridge, Lexanalytics

Big Data Landscape

Vertical Apps

- PREDICTIVE POLICING
- bloomreach
- MYRRIX

Log Data Apps

- splunk>
- loggly
- sumologic

Ad/Media Apps

- rocketfuel
- collective[i]
- bluefin
- Recorded Future
- LuckySort
- Media Science
- TURN
- DataXu
- Data. Insight. Action.

Business Intelligence

- ORACLE | Hyperion
- SAP
- Business Objects
- RJMetrics
- Microsoft | Business Intelligence
- IBM
- COGNOS
- birst
- MicroStrategy
- Autonomy
- QlikView
- bime
- DOMO
- Chart.io
- GoodData

Analytics and Visualization

- tableau
- Palantir
- OPERA
- metaLayer
- METAMARKETS
- dataspora
- centrifuge
- TERADATA ASTER
- sas
- TIBCO
- KARMASPHERE
- panopticon
- Real-Time Visual Data Analysis
- pentaho
- Datameer
- platfora
- ClearStory
- CIRRO
- alteryx
- visual.ly
- AYATA

Data As A Service

- factual.
- kaggle
- knoema beta
- GNIP
- DATASIFT
- Windows Azure Marketplace
- INRIX
- LexisNexis
- LOCATE Everything Location

Analytics Infrastructure

- Hortonworks
- VERTICA An HP Company
- MAPR TECHNOLOGIES
- cloudera
- INFOBRIGHT
- PARACCEL
- EMC²
- GREENPLUM.
- NETEZZA
- kognitio
- DATASTAX
- EXASOL
- calpont

Operational Infrastructure

- COUCHBASE
- 10gen | the MongoDB company
- TERADATA
- HADAPT
- TERRACOTTA
- VoltDB
- MarkLogic
- INFORMATICA

Infrastructure As A Service

- amazon web services
- Windows Azure
- infochimps
- Google BigQuery

Structured Databases

- ORACLE
- MySQL
- Microsoft SQL Server
- PostgreSQL
- IBM DB2.
- SYBASE
- memsql

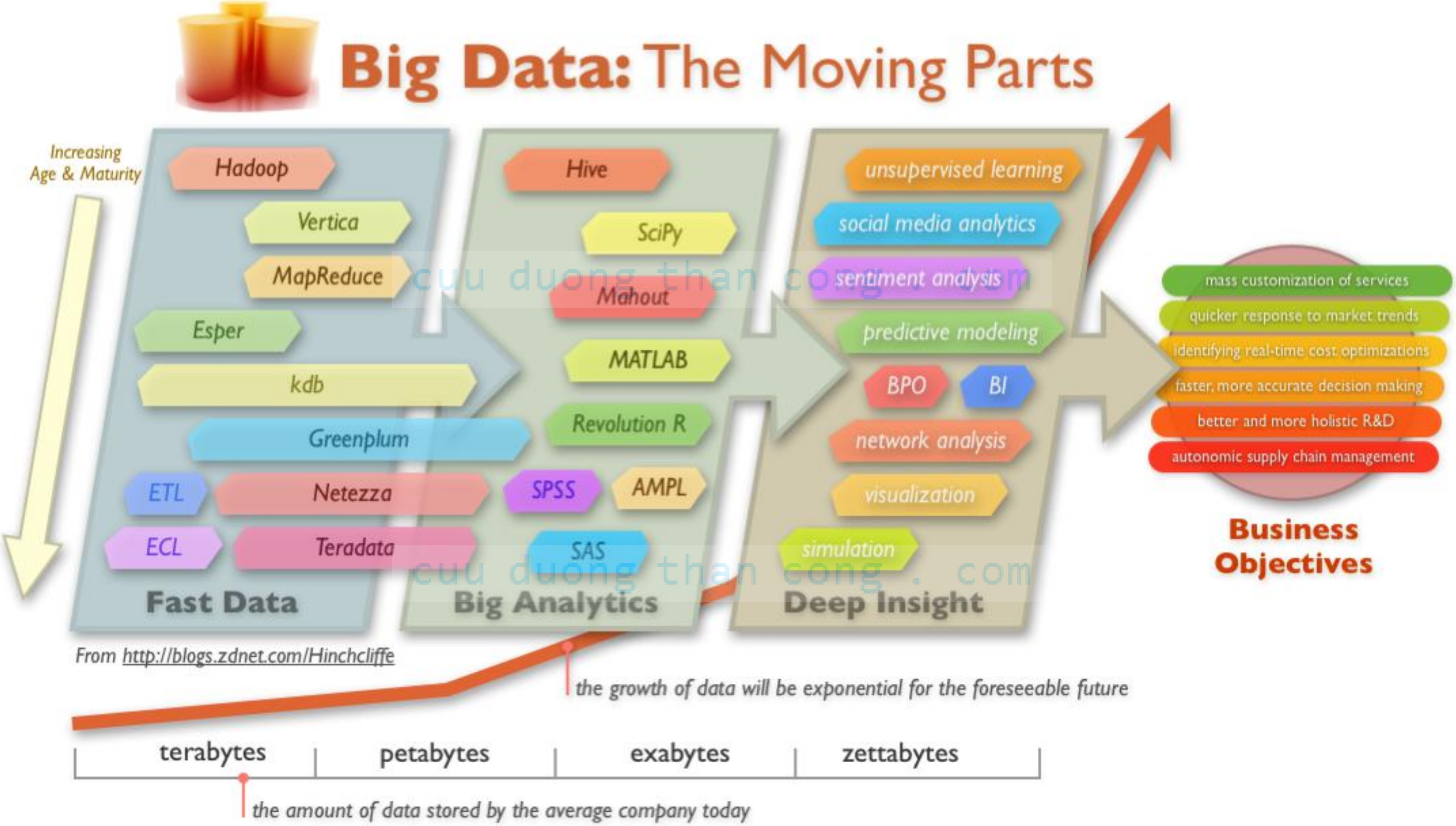
Technologies

- hadoop
- hadoop Map Reduce
- mahout
- APACHE HBASE
- Cassandra

dave@vcdave.com

blogs.forbes.com/davefeinleib

Big Data Technology


Big Data: The Moving Parts
Increasing Age & Maturity
Hadoop
Vertica
MapReduce
Esper
kdb
Greenplum
ETL
Netezza
ECL
Teradata
Fast Data
Hive
SciPy
Mahout
MATLAB
Revolution R
SPSS
AMPL
SAS
Big Analytics
unsupervised learning
social media analytics
sentiment analysis
predictive modeling
BPO
BI
network analysis
visualization
simulation
Deep Insight
mass customization of services
quicker response to market trends
identifying real-time cost optimizations
faster, more accurate decision making
better and more holistic R&D
autonomic supply chain management
Business Objectives
the growth of data will be exponential for the foreseeable future
terabytes
petabytes
exabytes
zettabytes
the amount of data stored by the average company today

From http://blogs.zdnet.com/Hinchcliffe

Summary

1	Distributed Databases & Client-Server Architectures
2	Spatial and Temporal Database
3	Multimedia Databases
4	Geographic Information Systems
5	XML
6	Data Warehousing
7	Outsourcing database services
8	Big Data

Q & A