Database Storage and Collection

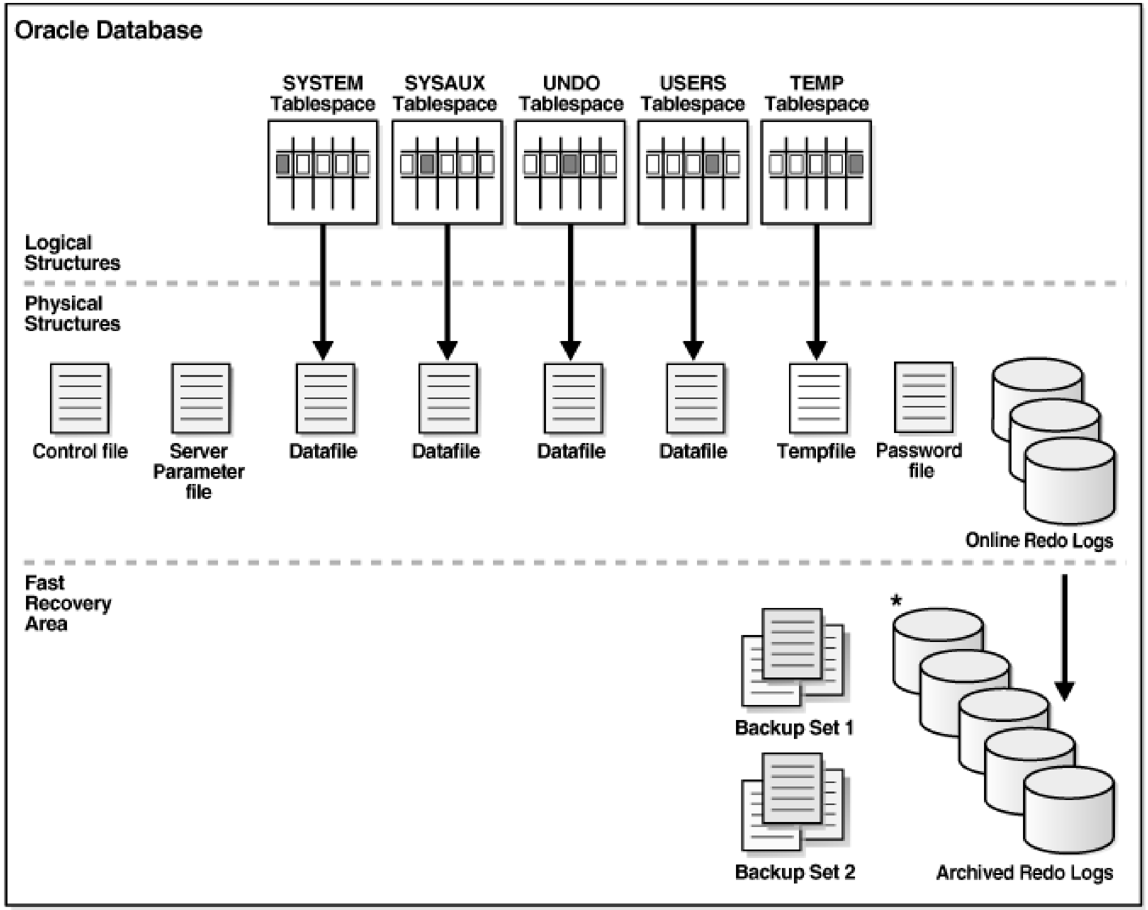
Tushar B. Kute,

http://tusharkute.com

Database Storage Structure

- A database is made up of physical and logical structures. Physical structures can be seen and operated on from the operating system, such as the physical files that store data on a disk.
- Logical structures are created and recognized by Database and are not known to the operating system. The primary logical structure in a database, a tablespace, contains physical files.
- The applications developer or user may be aware of the logical structure, but is not usually aware of this physical structure.
- The database administrator (DBA) must understand the relationship between the physical and logical structures of a database.

Database Storage Structure



* Archived Redo Logs present only after turning on log archiving (ARCHIVELOG mode)

Database Storage Structure

- Oracle Database can automate much of the management of its structure.
- Oracle Enterprise Manager Database Express (EM Express) provides a Web-based graphical user interface (GUI) to enable easier management and monitoring of your database.
- From a physical perspective, a multitenant container database (CDB) has basically the same structure as a non-CDB, except that each pluggable database (PDB) has its own set of tablespaces (including its own SYSTEM and SYSAUX tablespaces) and data files.

Database Storage Structure

- A CDB contains the following files:
  - One control file
  - One online redo log
  - One or more sets of temp files
  - One set of undo data files
  - A set of system data files for every container
  - Zero or more sets of user-created data files

Control Files

- A control file tracks the physical components of the database.
- It is the root file that the database uses to find all the other files used by the database.
- Because of the importance of the control file, Oracle recommends that the control file be multiplexed, or have multiple identical copies.
- For databases created with Oracle Database Configuration Assistant (DBCA), two copies of the control file are automatically created and kept synchronized with each other.

Control Files

- If any control file fails, then your database becomes unavailable.
- If you have a control file copy, however, you can shut down your database and re-create the failed control file from the copy, then restart your database.
- Another option is to delete the failed control file from the CONTROL_FILES initialization parameter and restart your database using the remaining control files.

Data Files

- Data files are the operating system files that store the data within the database.
- The data is written to these files in an Oracle proprietary format that cannot be read by other programs.
- Tempfiles are a special class of data files that are associated only with temporary tablespaces.

Data Files

- Data files can be broken down into the following components:
- Segment
  - A segment contains a specific type of database object. For example, a table is stored in a table segment, and an index is stored in an index segment. A data file can contain many segments.
- Extent
  - An extent is a contiguous set of data blocks within a segment. Oracle Database allocates space for segments in units of one extent. When the existing extents of a segment are full, the database allocates another extent for that segment.

Data Files

- Data block
  - A data block, also called a database block, is the smallest unit of I/O to database storage. An extent consists of several contiguous data blocks.
  - The database uses a default block size at database creation.
  - After the database has been created, it is not possible to change the default block size without re-creating the database.
  - It is possible, however, to create a tablespace with a block size different than the default block size.

Tablespace

- A database is divided into logical storage units called tablespaces, which group related logical structures (such as tables, views, and other database objects).
- For example, all application objects can be grouped into a single tablespace to simplify maintenance operations.
- A tablespace consists of one or more physical data files. Database objects assigned to a tablespace are stored in the physical data files of that tablespace.
- When you create an Oracle database, some tablespaces already exist, such as SYSTEM and SYSAUX.

Tablespace

- Tablespaces provide a means to physically locate data on storage. When you define the data files that comprise a tablespace, you specify a storage location for these files.
- For example, you might specify a data file location for a certain tablespace as a designated host directory (implying a certain disk volume) or designated Oracle Automatic Storage Management disk group.
- Any schema objects assigned to that tablespace then get located in the specified storage location. Tablespaces also provide a unit of backup and recovery.
- The backup and recovery features of Oracle Database enable you to back up or recover at the tablespace level.

Locally Managed Tablespace

- Space management within a tablespace involves keeping track of available (free) and used space, so that space is allocated efficiently during data insertion and deletion.
- Locally managed tablespaces keep the space allocation information within the tablespace, not in the data dictionary, thus offering better performance.
- By default, Oracle Database sets all newly created tablespaces to be locally managed with automatic segment management, a feature that further improves performance.

Tablespace Types

- There are three types of tablespaces. For example:
- Permanent
  - You use permanent tablespaces to store your user and application data.
  - Oracle Database uses permanent tablespaces to store permanent data, such as system data.
  - Each user is assigned a default permanent tablespace.

Tablespace Types

- Undo
  - A database running in automatic undo management mode transparently creates and manages undo data in the undo tablespace.
  - Oracle Database uses undo data to roll back transactions, to provide read consistency, to help with database recovery, and to enable features such as Oracle Flashback Query.
  - A database instance can have only one active undo tablespace.

Tablespace Types

- Temporary
  - Temporary tablespaces are used for storing temporary data, as would be created when SQL statements perform sort operations.
  - An Oracle database gets a temporary tablespace when the database is created. You would create another temporary tablespace if you were creating a temporary tablespace group.
  - Under typical circumstances, you do not have to create additional temporary tablespaces.
  - If you have an extremely large database, then you might configure additional temporary tablespaces.

Tablespace Status

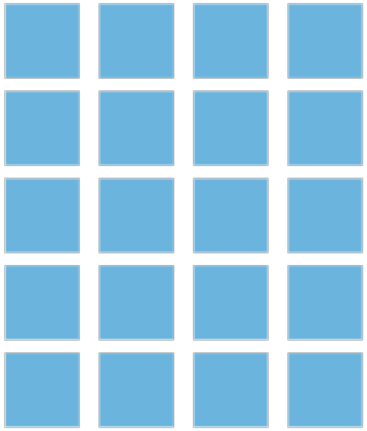
- You can set tablespace status. For example:
- Read Write
  - Users can read and write to the tablespace after it is created. This is the default.
- Read Only
  - If the tablespace is created Read Only, then the tablespace cannot be written to until its status is changed to Read Write.
  - It is unlikely that you would create a Read Only tablespace, but you might change it to that status after you have written data to it that you do not want modified.

Tablespace Status

- Offline
  - If the tablespace has a status of Offline, then no users can access it.
  - You might change the status of a tablespace to Offline before performing maintenance or recovery on the data files associated with that tablespace.

Structured Data

- The data which is to the point, factual, and highly organized is referred to as structured data.
- It is quantitative in nature, i.e., it is related to quantities that means it contains measurable numerical values like numbers, dates, and times.



Structured Data

- It is easy to search and analyze structured data. Structured data exists in a predefined format. Relational database consisting of tables with rows and columns is one of the best examples of structured data.
- Structured data generally exist in tables like excel files and Google Docs spreadsheets. The programming language SQL (structured query language) is used for managing the structured data.
- SQL is developed by IBM in the 1970s and majorly used to handle relational databases and warehouses.
- Structured data is highly organized and understandable for machine language. Common applications of relational databases with structured data include sales transactions, Airline reservation systems, inventory control, and others.

Unstructured Data

- All the unstructured files, log files, audio files, and image files are included in the unstructured data.
- Some organizations have much data available, but they did not know how to derive data value since the data is raw.

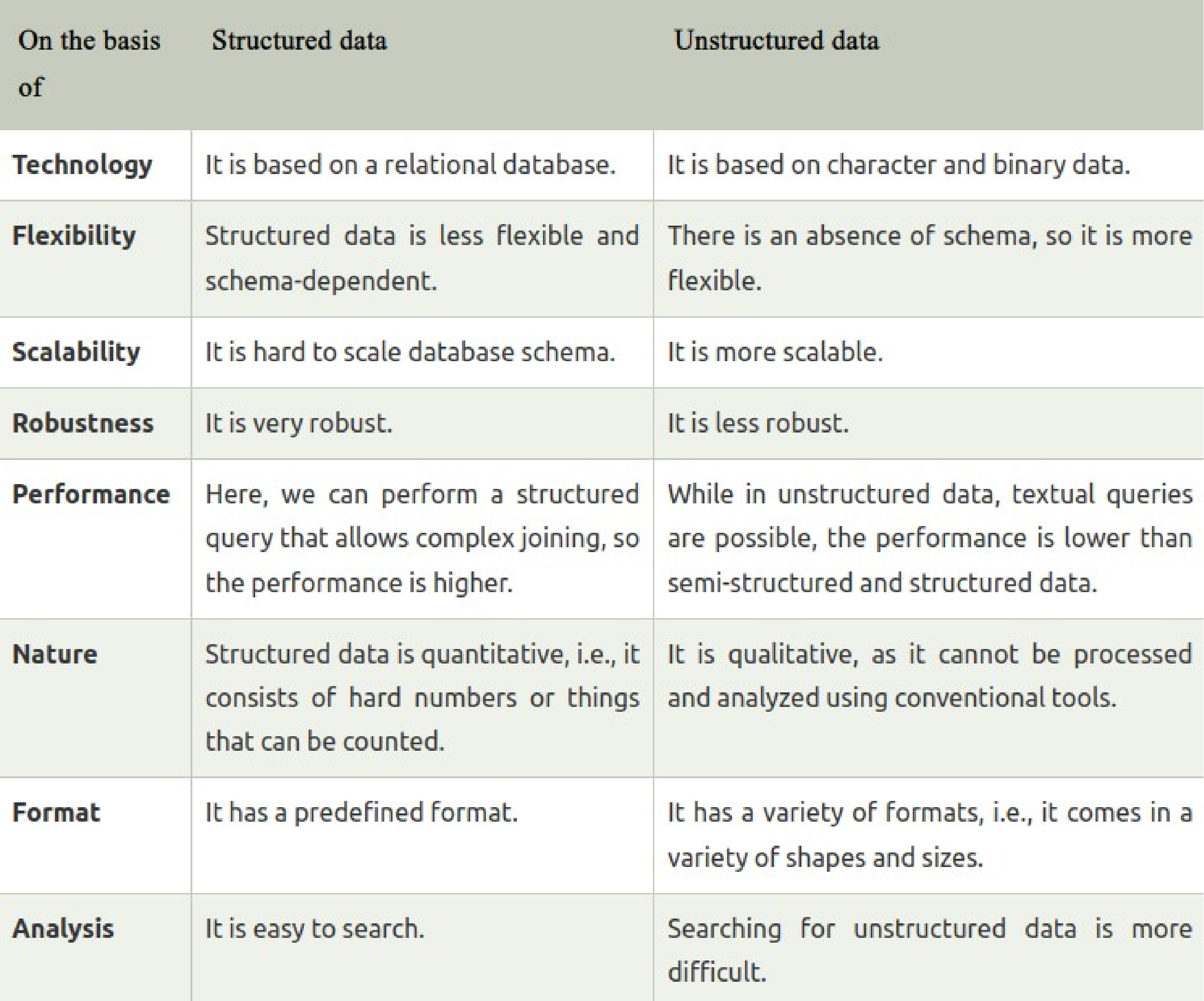
Unstructured Data

- Unstructured data is the data that lacks any predefined model or format. It requires a lot of storage space, and it is hard to maintain security in it.
- It cannot be presented in a data model or schema. That's why managing, analyzing, or searching for unstructured data is hard.
- It resides in various different formats like text, images, audio and video files, etc.
- It is qualitative in nature and sometimes stored in a non-relational database or NO-SQL.

Unstructured Data

- It is not stored in relational databases, so it is hard for computers and humans to interpret it.
- The limitations of unstructured data include the requirement of data science experts and specialized tools to manipulate the data.
- The amount of unstructured data is much more than the structured or semi-structured data.
- Examples of human-generated unstructured data are Text files, Email, social media, media, mobile data, business applications, and others.
- The machine-generated unstructured data includes satellite images, scientific data, sensor data, digital surveillance, and many more.

Comparison



On the basis of	Structured data	Unstructured data
Technology	It is based on a relational database.	It is based on character and binary data.
Flexibility	Structured data is less flexible and schema-dependent.	There is an absence of schema, so it is more flexible.
Scalability	It is hard to scale database schema.	It is more scalable.
Robustness	It is very robust.	It is less robust.
Performance	Here, we can perform a structured query that allows complex joining, so the performance is higher.	While in unstructured data, textual queries are possible, the performance is lower than semi-structured and structured data.
Nature	Structured data is quantitative, i.e., it consists of hard numbers or things that can be counted.	It is qualitative, as it cannot be processed and analyzed using conventional tools.
Format	It has a predefined format.	It has a variety of formats, i.e., it comes in a variety of shapes and sizes.
Analysis	It is easy to search.	Searching for unstructured data is more difficult.

Data Collection

- The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is Known as Data Collection.
- Knowledge is power, information is knowledge, and data is information in digitized form, at least as defined in IT.
- Hence, data is power. But before you can leverage that data into a successful strategy for your organization or business, you need to gather it. That’s your first step.

Data Collection

- Data is various kinds of information formatted in a particular way.
- Therefore, data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.
- Our society is highly dependent on data, which underscores the importance of collecting it.
- Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity.

Data Collection: Why?

- Before a judge makes a ruling in a court case or a general creates a plan of attack, they must have as many relevant facts as possible.
- The best courses of action come from informed decisions, and information and data are synonymous.
- There is far more data available today, and it exists in forms that were unheard of a century ago.
- The data collection process has had to change and grow with the times, keeping pace with technology.

Data Collection: Methods

- The following are seven primary methods of collecting data in business analytics.
  - Surveys
  - Transactional Tracking
  - Interviews and Focus Groups
  - Observation
  - Online Tracking
  - Forms
  - Social Media Monitoring

Data Collection: Methods

- Primary
  - As the name implies, this is original, first-hand data collected by the data researchers.
  - This process is the initial information gathering step, performed before anyone carries out any further or related research.
  - Primary data results are highly accurate provided the researcher collects the information.
  - However, there’s a downside, as first-hand research is potentially time-consuming and expensive.

Data Collection: Methods

- Secondary
  - Secondary data is second-hand data collected by other parties and already having undergone statistical analysis.
  - This data is either information that the researcher has tasked other people to collect or information the researcher has looked up.
  - Simply put, it's second-hand information. Although it's easier and cheaper to obtain than primary information, secondary information raises concerns regarding accuracy and authenticity.
  - Quantitative data makes up a majority of secondary data.

Data Collection Tools

- Word Association
  - The researcher gives the respondent a set of words and asks them what comes to mind when they hear each word.
- Sentence Completion
  - Researchers use sentence completion to understand what kind of ideas the respondent has.
  - This tool involves giving an incomplete sentence and seeing how the interviewee finishes it.

Data Collection Tools

- Role-Playing
  - Respondents are presented with an imaginary situation and asked how they would act or react if it was real.
- In-Person Surveys
  - The researcher asks questions in person.
- Online/Web Surveys
  - These surveys are easy to accomplish, but some users may be unwilling to answer truthfully, if at all.

Data Collection Tools

- Mobile Surveys
  - These surveys take advantage of the increasing proliferation of mobile technology. Mobile collection surveys rely on mobile devices like tablets or smartphones to conduct surveys via SMS or mobile apps.
- Phone Surveys
  - No researcher can call thousands of people at once, so they need a third party to handle the chore. However, many people have call screening and won’t answer.
- Observation
  - Sometimes, the simplest method is the best. Researchers who make direct observations collect data quickly and easily, with little intrusion or third-party bias. Naturally, it’s only effective in small-scale situations.

Data Collection : Key Steps

- 1. Decide What Data You Want to Gather
- 2. Establish a Deadline for Data Collection
- 3. Select a Data Collection Approach
- 4. Gather Information
- 5. Examine the Information and Apply Your Findings

Thank you

@mitu_skillologies

@mITuSkillologies

@mitu_group

@mitu-skillologies

@MITUSkillologies

kaggle

@mituskillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

@mituskillologies

contact@mitu.co.in

tushar@tusharkute.com