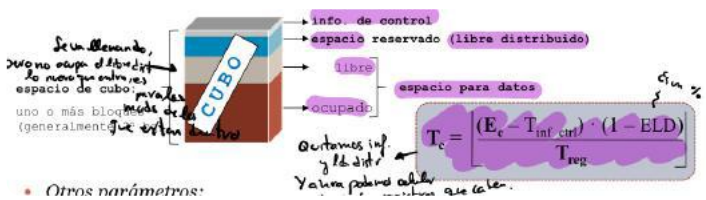
Tema 5. Ficheros: Introducción y Conceptos Básicos.

- **Punto de partida:**
  - **Enfoque lógico, eficacia**: El usuario ve archivos, que son colecciones de registros, que son agregaciones de datos.
  - **Enfoque físico, eficiencia**: La maquina acceder a ficheros, que son secuencias de bloques (unidad de acceso al soporte), que son conjuntos de bytes.
- **Estructura Física vs. Lógica:**
  - La unidad es el registro, y la unidad subatómica es el campo.
  - **Correspondencia físico-lógica a nivel de registro**: Consideración de tamaños.
    - **Registró expandido:** Registro lógico que abarca varios bloques(registros fisicos).
    - **Bloque**: Cuando en un registro físico, bloque, caben varios registros lógicos.
    - **Factor de Bloqueo:** Numero de registros lógicos que caben en un bloque. Solo se da en organización consecutiva, en no consecutiva se usa el tamaño de cubo.
  - **Correspondencia físico-lógica entre registros:** La organización de registros.
    - **Organización Consecutiva:** Los registros lógicos están uno tras otro, no espera al siguiente bloque si no cabe entero. Se dice de tamaño n bloques.
    - **Organización No consecutiva:** Cuando un registro lógico, si no cabe completo en el bloque, se pasa al siguiente bloque. Se dirá fichero de n cubos.
  - **Cubo:** Conjunto de bytes con una condición de acceso común. Para la no consecutiva es un conjunto de bloques que son utilizados como almacenamiento. Si se usan cubos pasaran a ser la unidad mínima, y se empezara a leer y escribir cubos.
    - **Espacio de cubo:** Información asignada a cubo.
    - **Tamaño de cubo:** Registros lógicos que caben en un cubo. Se redondea a la baja.
    - **Partes de un cubo:**
      - **Información de control:** Cabecera, directorio de cubo, puntero encadenamiento,..
      - **Espacio reservado(libre distribución):** Se reserva por si hay que modificar algún registro y aumenta el tamaño, así cabe y no hay que reescribirlo en un nuevo cubo
      - **Espacio para datos**: Hay espacio ocupado y espacio libre para añadir datos de los registros que hay dentro.



  - **Diseño de Ficheros:**
    - **Diseño Lógico:** Descripción y disposición de los elementos de datos de un registro, que en conjunto definen un individuo.
      - **Unidad subatómica**: El campo, unidad mínima e indivisible.
        - Notación: camp tipo(tamaño)
      - **Agregado de datos:** Colección de elementos. Elemento de datos de evaluación múltiple.
        - **Vector**: Numero fijo de elementos que definen un concepto. Ejem: Fecha
          - Notación: (elemento1; elemento2; elemento3; ...)
        - **Grupo repetitivo**: Compuesto por un numero fijo o variable de elementos cuya interpretación es común. Ej: Los hijos, puede haber 1 o x pero con la mima estructura.
          - Notación: (elemento)* desde 0 hasta N elementos.
          - Notación: (elemento)^+ desde 1 hasta N elementos.
    - **Diseño Físico-lógico (Físico del registro lógico):** La implementación de un registro lógico en secuencia de bytes que permiten su lectura y escritura. Descripción de las cadenas de bytes utilizadas para almacenar registros.
      - Se busca la eficiencia, reducir el espacio por lo tanto el numero de accesos.

- **Volumen y ocupación de un registro:**
  - **Volumen**: Numero de caracteres necesarios para almacenarlo.
  - **Ocupación útil:** Caracteres útiles del registro, los que usa de lo que le dan.
  - **Densidad ideal de un registro**: Relación entre cantidad de información útil y la cantidad de información almacenada. d= útil/real
- **Optimización**:
  - **Campos de control**: Marcas, mejoran el manejo que permiten ahorrar el relleno (padding) cuando ocupa menos de lo que se le da. La marca es un caracter, por lo que es un byte, 8 bits, que depende lo que almacene puede o no merecer la pena su uso.
    - **Elementos de datos:**
      - **Existencia**: Indica en un campo opcional, si esta o no.
      - **Longitud**: Indica la longitud en numero de caracteres.
      - **Reiteración**: Numero de ocurrencias en un grupo repetitivo.
      - **Fin de Campo:** Indica cuando acaba un campo, se usa para campos muy grandes. Su uso es peligroso.
    - **Registro**:
      - **Fin(inicio) de registro:** Separa registros consecutivos.
      - **Tipo**: Indica el tipo de registro a continuación.
      - **Mapa**: Indica los registros que se aplica, agrupación de bytes.
  - **Codificación de campos:** Consiste en sustituir una información por otra equivalente de menor tamaño. Algunos:
    - Utilizar codificación numérica.
    - Utilizar enumerados.
    - Fecha en formato Juliano (la de un numero).
    - Agrupación de varios campos.
- **Diseño Físico:** Disposición física de los registro en el soporte, para acceder a ellos con el menor coste y números de accesos que sea posible. El numero de accesos es el mas importante, por que estos conllevan también tiempo de acceso.
  - **Espacio de un fichero**: Se busca que ocupe lo mínimo. La densidad ideal es menor o igual que la real casi siempre, tenerlo en cuenta al obtener resultados.
    - **Densidad real(dr) de un fichero**: Relación entre cantidad de información útil y cantidad de información almacenada. Esta medida es mas global que la ideal, ya que la ideal se limita a un registro. La **formula** desglosada es: el numero de registros por lo que usamos de los registro para almacenar datos, sin información de control y libre, partido del numero de bloques por el tamaño de los bloques.

$$d_r = \frac{\text{ocupación fichero}}{\text{volumen fichero}} = \frac{\text{ocupación/reg} * \text{nºregistros}}{n\ \text{bloques} * T_{bq}}$$

    - **Densidad de ocupación(do) de un fichero no consecutivo**: Relación entre cantidad de registros almacenados y cantidad a de ellos que caben, el espacio potencialmente útil. **La formula** es el numero de registros que hay partido del numero de cubos por el numero de registros por cubo.

$$d_{oc} = \frac{\text{nºregistros}}{N\ \text{cubos} * T_c}$$

$$C(O_k,P) = \sum_{i=1..n} C_i \cdot f_i$$

  - **Coste Global:** Es la media ponderada de accesos lógicos(por unidad de tiempo o carga) de todos los procesos. Hay que hallar el coste para cada tipo de organización y observar cual es la que mejor se adapta. Una organización física del Sistema de Archivos (O) define todas las organizaciones base de los archivos que incluye. Cada proceso P, tendrá en O un coste C asociado, que se expresa en numero de accesos o tiempo.
    - Todos los sistemas de archivos están sometidos a un conjunto de procesos, y estos procesos tienen una frecuencia asociada referida a una unidad de tiempo(segundos, horas,...) La suma de todos los procesos de un sistema es 1.

- **Coste a bajo nivel:**
  - Determinados soportes mejoran el acceso serial.
  - Pueden almacenar bloques en memoria privilegiada (en paginas)
    - En memoria intermedia (buffer), es que es más rápida(ahorra accesos), pero mas costosa(es escasa)
  - **Hit ratio(hr|φ):** Porcentaje de accesos ahorrados por la memoria intermedia. Las veces que pide y esta en el buffer, y no tiene que traer.
    - **coste real(efectivo)= (1-hr)*coste global**
  - **PIO(Physical Input Output):** Numero de veces que físicamente se lee el bloque. Es el que se prioriza reducir, pero es mas difícil, por lo que se hace por medio del LIO. LIO y PIO son proporcionales.
  - **LIO(Logical Input Output)**: Numero de veces que pide una pagina, este o no en buffer. LIO y PIO se relacionan mediante el hit ratio.
- **Interacción con FF(Ficheros)**
  - **Clave**: Campo o secuencia de campos, en un orden, con una función especifica en la interacción de los usuarios con el fichero.Tipos:
    - **De identificación:** Campo o conjunto de campos que identifican unívocamente un registro. Sin valores repetidos.
    - **No identificativa**: Lo contrario. Presenta valores repetidos en el fichero.
      - **Coincidencia k**: Tasa media de registros que toman un determinado valor. k=r/ #valores. r numero de registros del fichero, que lo toman.
      - **Cardinalidad del dominio**: Numero de valores distintos. #valores.
    - **De búsqueda:** Campo o conjunto de campos que es frecuente para realizar una búsquedas. No solo son las del where si no también las de proyección.
    - **Privilegiada**: Campo sobre la que existe un mecanismo físico que hace la recuperación más eficiente.
      - **De direccionamiento**: Determina la ubicación del registro.
      - **De ordenación**: Criterio físico o lógico de ordenación.
      - **De indización:** Privilegiada mediante una estructura auxiliar.
      - **De agrupación**: Reúne registros con esa clave común.No Group By
- **Tipología de Procesos:**
  - **Diferenciar entre:**
    - **Actualización:** Implican escritura.
    - **Recuperación**: Solo implican lectura.
  - **Diferenciar entre:**
    - **Selectivo:** Aquel que impone un filtro o condición, solo a esos.
    - **Incondicional**: No impone condiciones, se refiere a al totalidad.
  - **Diferenciar entre:**
    - **Identificativa(exact match)**: Aquella cuya clave de búsqueda es identificativa, que encuentra a solo 1. De media recorre la mitad.
      - **Simple**: Si es la clave identificativa.
      - **Multiclave**: No tiene sentido, la identificativa ya lo encuentra.
    - **No identificativa:** Hay que recórrelos todos, ya que hay mas de de uno, pero no sabes o cuantos.
      - **Simple**: k registros (no identificativa). k registros en un rango.
      - **Multiclave**: k registros en un rango (window query). Proyección de pocos atributos del resultado de una WQ.
- **Conjunto de Direcciones Relevantes**:
  - Aquellas claves privilegiadas que permiten filtrar, y tras descartar las que no lo cumplan se recorren las restantes, a menos que se encuentre lo que se busca (de media (N+1)/2).
  - Se pueden utilizar booleanos para indicar que registros debemos recorrer.
  - Una de la estrategias de filtrados es, hacer un árbol de decisiones y la rama escogida es el plan de ejecución.