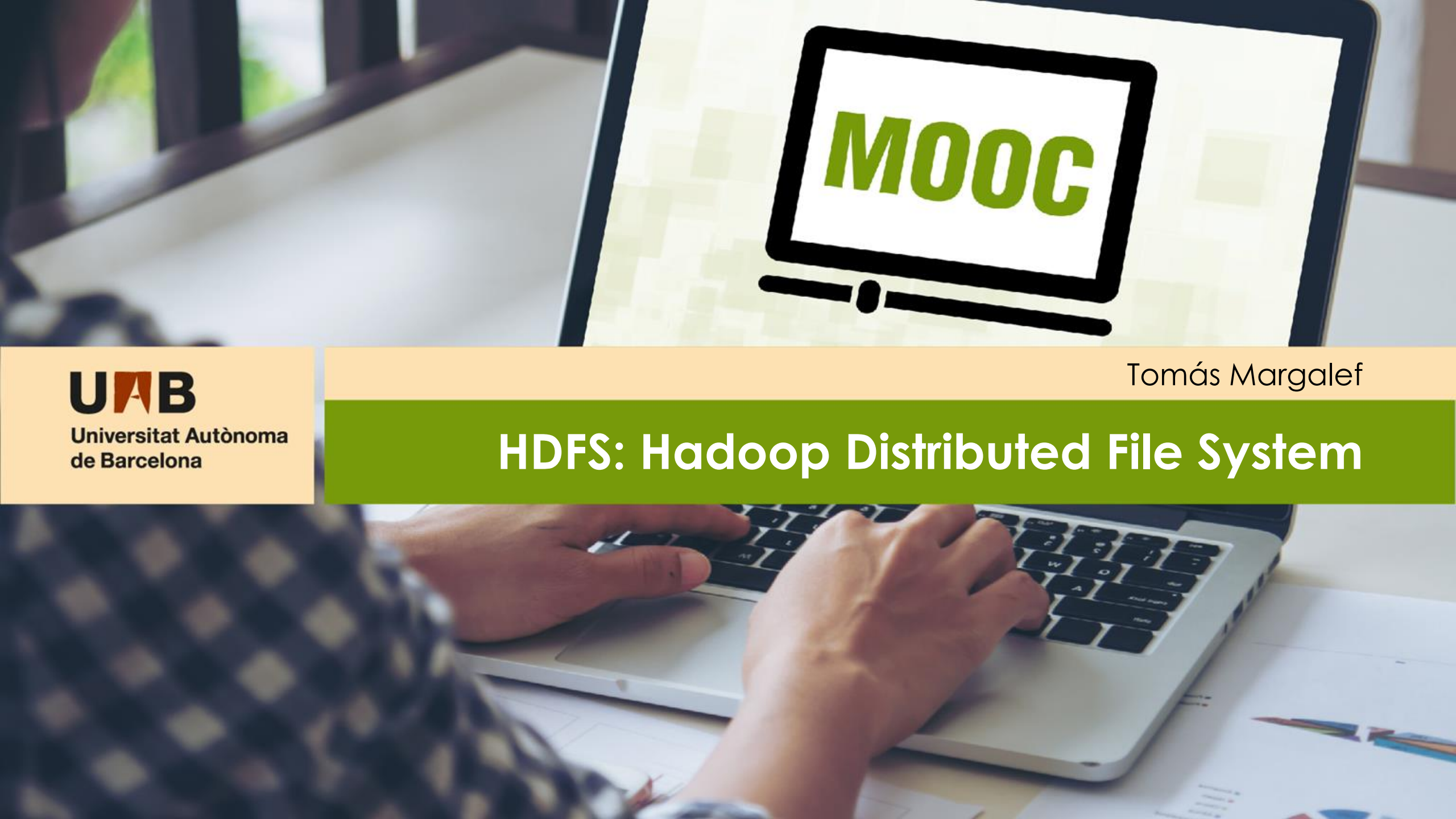

Tomás Margalef

UAB
Universitat Autònoma
de Barcelona

HDFS: Hadoop Distributed File System

HDFS: Hadoop Distributed File System

MapReduce 

YARN Yet Another Resource Negotiator 

HDFS Hadoop Distributed File System 

Hardware Básico

HDFS Hadoop Distributed File System

HDFS es un sistema de archivos distribuido con dos características fundamentales:

- Ha sido diseñado para trabajar en sistemas de cómputo de bajo coste (“*commodity hardware*”).
- Ha sido diseñado para ser altamente tolerante a fallos.

Originalmente fue diseñado como una infraestructura para el motor de búsqueda en la web de Apache Nutch, y actualmente es un subproyecto de Apache Hadoop.

HDFS Hadoop Distributed File System

Características de HDFS:

- Es adecuado para aplicaciones que manejan grandes volúmenes de datos.
- Una instancia HDFS puede consistir en cientos o incluso miles de nodos servidores, cada uno almacenando parte de los datos.
- HDFS proporciona acceso de alto rendimiento a los datos de las aplicaciones.

HDFS Hadoop Distributed File System

Características de HDFS (II):

Grandes volúmenes de datos

- Un archivo típico en HDFS tiene un tamaño de gigabytes o incluso terabytes.
- HDFS está diseñado para soportar grandes volúmenes de datos.
- HDFS proporciona alto ancho de banda agregado y escala a cientos de nodos en un clúster.
- Es capaz de gestionar millones de archivos en una sola instancia.

HDFS Hadoop Distributed File System

Características de HDFS (III):

Tolerancia a fallos

- El gran número de nodos comporta una alta probabilidad de fallo, de modo que es probable que haya nodos caídos.
- La detección de fallos y la recuperación rápida y automática de los mismos es un objetivo crucial.

HDFS Hadoop Distributed File System

Características de HDFS (IV):

Acceso a datos de "*streaming*"

- Las aplicaciones que se ejecutan sobre HDFS requieren acceso en "*streaming*" a los datos.
- HDFS ha sido diseñado más para procesamiento "batch" que para el uso interactivo de los datos por parte de los usuarios.
- El objetivo principal es conseguir un alto rendimiento ("*throughtput*") más que una baja latencia.

HDFS Hadoop Distributed File System

Características de HDFS (V):

Modelo de coherencia simple

- Las aplicaciones que se ejecutan sobre HDFS cumplen un modelo de una sola escritura y múltiple lectura ("write-once-read-many").
- Una vez que un archivo ha sido creado, escrito y cerrado no puede ser cambiado.
- Esta asunción simplifica las cuestiones relativas a la coherencia y capacita el acceso de alto rendimiento a los datos.

HDFS Hadoop Distributed File System

Características de HDFS (VI):

Movimiento de las aplicaciones

- El cómputo es mucho más eficiente si se ejecuta cerca de los datos sobre los que opera. Especialmente si los datos son enormes.
- Esto minimiza la congestión de la red e incrementa el rendimiento del sistema.
- HDFS proporciona interfaces para que las aplicaciones se muevan más cerca de donde los datos que van a necesitar están almacenados.

HDFS Hadoop Distributed File System

Características de HDFS (VII):

Portabilidad entre plataformas

- HDFS ha sido diseñado para ser portable entre plataformas hardware y software.
- Esto facilita la adopción generalizada de HDFS como el sistema de archivos para un gran conjunto de aplicaciones.

HDFS: Hadoop Distributed File System

UAB
Universitat Autònoma
de Barcelona