1. **“Tenemos un proyecto de Big Data, porque tenemos mucha información” (data minning, analytics, big data). Como se sabe que es un proyecto de big data.**

*3 V’s de big data. Depende del proyecto y estrcutura de la empresa*

Volumen: No entran en un disco/rack. Hay que partir los datos para poder analizarlos.

Velocidad: Un sistema convencional no llega a procesarlos/consumirlos. Hay que distribuir el procesamiento.

Variedad: Tipos de datos.

1. **Plataformas de Big Data**
2. **Que es Hadoop.**
3. **Componentes**
4. **High Availability – porque/como**
5. **Fault Tolerant – porque/como – HDFS & YARN– Como se detectan los fallos, como se recuperan**
6. **Es fault tolerant, pero tiene una gran contra… SPOF - NN**
7. **Standby NN, Secondary NN**
8. **Diferencias H1 vs H2**
9. **MR – pasos para crear un job**
   1. **Donde se guardan los resultados intermedios? En el FS local**
   2. **Que es un combiner? Cuand se usa?**
   3. **Como se elige el mapper? Data Locality**
10. **HDFS – data write (subir un archivo a HDFS)**
    1. **Que interfaz hay? CLI, simil POSIX**
    2. **Default block size**
    3. **Si archivo tiene tamaño más pequeño que el block size?**
    4. **Problema de (muchos) archivos pequeños? Overhead muy grande**
    5. **Data Read**
11. **YARN**
    1. **Funciona solo con MR?**
    2. **Como se ejecutan los Jobs en yarn?**
12. **Contras de Hadoop**
    1. **Archivos pequeños**
    2. **Datos relacionales – BD relacionales**
    3. **Ad hoc queries**
    4. **Joins**
    5. **Online analytics**
    6. **Random read, random update no se pueden 🡪 es append only**
13. **HBase, para accesos randoms a files**
14. **Para qué Hive/Pig?**