- pk tel archi ou tel algo ?

- si fichier petit, est ce qu’on aurait fait un hdfs put ? consommation reséau, limitation data manager qui est non scalable comme les data node,

- quelques techno de stockage ? hdfs, hbase, cassandra, mongodb,

- quelques techno de compute ? Spark, JAVA (mais pas distribué sauf utilisé avec Spark), Hive, Pig, Python

- Deseign by queri : crée une table pour répondre à une requête donnée dans des bdd opé (mongo, cassandra)

- la replication c’est quoi et sert à quoi ? securité et haute dispo des données pour parallélisation

- partitionning ? Découpage des données

- différence entre cassandra et mongodb ? mongodb modélise (et non stockage) en archi master-slave, en document et cassandra en colonne master-less

- différence entre ces 3 batch (spark), realtime (hadoop vs Glue), micro-batch (spark-streaming) : batch est le fait de lancer un ensemble d’opé pour mettre en exécution un ensemble de taches qui vont se lancer progressivement et donc pas de réactivité jusqu’à la fin, mais performante, et le realtime c’est une réactivité forte mais pas très rapide, micro-batch, performance bonne et réactivité bonne.

- c’est koi les element ELK et à quoi servent-ils ?

- quels sont les composants de HIVE ? Il contient le schéma des données

- composant de HDFS ? y’en a 2 principaux : data node (données) et name node (meta données qui donnent l’emplacement, leur replication)

- c’est quoi un RDD : Resilien (renconstituable en cas de perte de donnée) Distributed(dans le monde distribué) DataSet (des données)

- différence entre action et transformation dans une action ? une action (ex count) donne un résultat et transformation (ex map) donne un nouveau RDD

- Yarn : gestionnaire de ressources

- diff entre dataframe et dataset : dataframe (non typé) est un dataset(typé) de raw.

- les script hive sont convertis en quoi ? en mapreduce

- root dans unix et hdfs ? dans unix c’est root et dans hdfs c’est hdfs

- interet du partitionning dans hdfs ? Accès rapide et direct de la donnée sans passer dans les autres

- archi de base : dans l’archi, il faut toujours avoir au moins data node avec chacun leur node manager, et une machine avec le name node (seul sinon il sera trop sollicité), plus une machine avec le secondary name node et ressource manager.