I am working on a migration project where we are pull data from existing repository and loading that data in to Hadoop cluster have some what around 200+ tables in the database

--to do this we are using spark and Sqoop tools

--we are using sqoop tool – where we need to run incremental jobs on daily bases

--we use spark to pull the data and filter out what ever is required from database.

-- based on the requirement we may store processed data as AVRO or Parquet file/ or load it in to the database table

--in the same project we also get data of 2 to 3 GB which is semi structured data

--here we have to process the data

- by extracting some fields; and segregating it according to ad id – and load processed data back to bigdata cluster

To Know file format compression differences which one to use

<https://blog.cloudera.com/blog/2017/02/performance-comparing-of-different-file-formats-and-storage-engines-in-hadoop-file-system/>

using Parquet we can achieve max compression with Gzip/Bzip2

and some tables are stored in Avro format to have a room for modification in future(using schema evolution)