# Format Notes

* Several initial file formats: Avro, Parquet, JSON, CSV, ORC, XML
* Avro (binary data format, relies on schemas) :
  + JSON -> Avro:
    - Binary **.avsc** can be converted to JSON, and vice-versa using a jar file given in *avrotools* (package installed with avro)
  + CSV -> Avro:
    - Python open-source: <https://gist.github.com/lordjc/9694632#file-csv2avro-py>
    - “**FromTextTool**”.java in org.apache.avro.tool (*avrotools*)
  + XML -> Avro:
    - Java open-source: <https://github.com/stealthly/xml-avro>
* Parquet (columnar storage, nested data structure)



* + Avro -> Parquet (Non-hadoop writer): <http://blog.cloudera.com/blog/2014/05/how-to-convert-existing-data-into-parquet/>
  + Kite-CLI:
    - Kite-CLI?: infers CSV/JSON schema, and creates parquet format using Hive table
    - Might allow **non-hadoop installation**?
* JSON/CSV/XML -> CSV/XML/JSON (inter-format conversions)
  + **Several** open-source configurations in every language. Not an issue.
* CLI App w/ Scala:
  + <http://www.scala-sbt.org/0.13/docs/Command-Line-Applications.html>
* CLI App w/ Python:
  + <http://docs.python-guide.org/en/latest/scenarios/cli/>