Big Data con Hadoop y Spark

Módulo 02 – Casos de Uso

Hadoop Compression Types

- **gzip** - org.apache.hadoop.io.compress.GzipCodec
- **bzip2** - org.apache.hadoop.io.compress.BZip2Codec
- **LZO** - com.hadoop.compression.lzo.LzopCodec
- **Snappy** - org.apache.hadoop.io.compress.SnappyCodec
- **Deflate** - org.apache.hadoop.io.compress.DeflateCodec

https://docs.cloudera.com/documentation/enterprise/5-9-x/topics/introduction_compression.html

Hive SerDes

Acrónimo de Serializer/Deserializer. Permite interpretar diferentes formatos.

SerDes disponibles en Hive

- Avro (Hive 0.9.1 and later)
- ORC (Hive 0.11 and later)
- RegEx
- Thrift
- Parquet (Hive 0.13 and later)
- CSV (Hive 0.14 and later)
- JsonSerDe (Hive 0.12 and later in hcatalog-core)

https://cwiki.apache.org/confluence/display/Hive/SerDe

Ejemplo Parquet y Snappy

- In general LZO wins size benchmarks, Snappy good balance between size and CPU intensity.

```
led-zeppelin-albums.parquet/
•  _SUCCESS
•  _common_metadata
•  _metadata
•  Year=1969/
     – Part-r-00000-6d4d42e2-c13f-4bdf-917d-2152b24a0f24.snappy.parquet
     – Part-r-00001-6d4d42e2-c13f-4bdf-917d-2152b24a0f24.snappy.parquet
     – …
•  Year=1970/
     – Part-r-00000-35cb7ef4-6de6-4efa-9bc6-5286de520af7.snappy.parquet
     – ...
```

Casos de Uso


mongoDB


{JSON}
JavaScript Object Notation


kafka


AVRO


HIVE


Parquet

Revisión

- ¿Qué factores debemos evaluar para elegir un formato de almacenamiento?
- ¿Qué tipos de formatos se utilizan en un proyecto de Big Data?
- ¿Cuál es el formato mas optimo para realizar agregaciones?
- ¿Qué formato de compresión es el mas utilizado en Big Data?
- ¿Qué significa Schema Evolution?

Gracias