

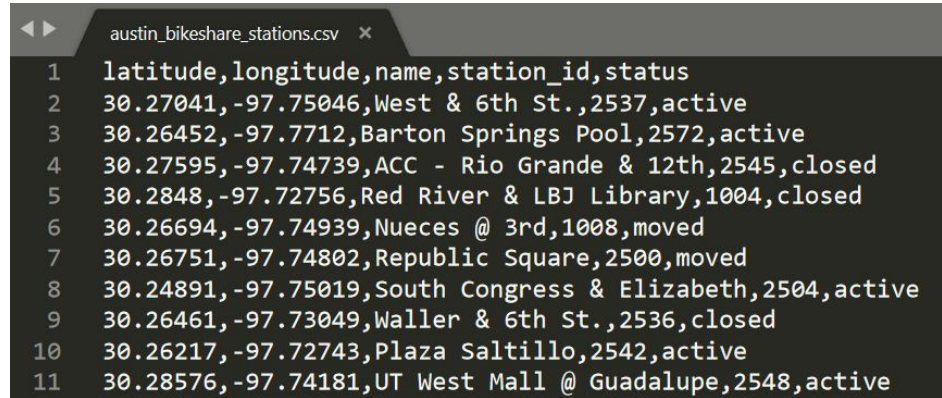
Big Data con Hadoop y Spark

Módulo 02 – Formatos de Almacenamiento

CSV

Texto plano delimitado por el carácter de la coma (pueden utilizarse otros separadores).

Generalmente este formato lo utilizan aplicaciones tradicionales para exportar datos hacia otros sistemas.



```
austin_bikeshare_stations.csv x
1 latitude,longitude,name,station_id,status
2 30.27041,-97.75046,West & 6th St.,2537,active
3 30.26452,-97.7712,Barton Springs Pool,2572,active
4 30.27595,-97.74739,ACC - Rio Grande & 12th,2545,closed
5 30.2848,-97.72756,Red River & LBJ Library,1004,closed
6 30.26694,-97.74939,Nueces @ 3rd,1008,moved
7 30.26751,-97.74802,Republic Square,2500,moved
8 30.24891,-97.75019,South Congress & Elizabeth,2504,active
9 30.26461,-97.73049,Waller & 6th St.,2536,closed
10 30.26217,-97.72743,Plaza Saltillo,2542,active
11 30.28576,-97.74181,UT West Mall @ Guadalupe,2548,active
```

JSON

Permite representar estructuras jerárquicas y relaciones en un solo documento (ejemplo MongoDB).

Es el formato estándar para comunicaciones vía HTTP (ejemplo lectura de datos de Twitter)

```
{
  "orders": [
    {
      "orderno": "748745375",
      "date": "June 30, 2088 1:54:23 AM",
      "trackingno": "TN0039291",
      "custid": "11045",
      "customer": [
        {
          "custid": "11045",
          "fname": "Sue",
          "lname": "Hatfield",
          "address": "1409 Silver Street",
          "city": "Ashland",
          "state": "NE",
          "zip": "68003"
        }
      ]
    }
  ]
}
```

Avro

Almacena los datos en formato binario para reducir el tamaño y mejorar la performance.

La definición de los datos (schema) se almacena en formato JSON.

Es recomendable utilizarlo para consultas de tipo SELECT *.



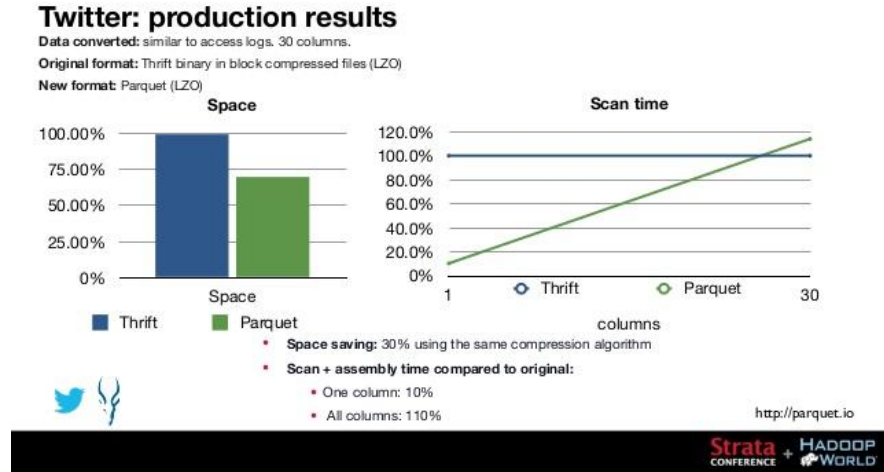
```
{
  "id": 123,
  "first": "ben",
  "last": "goldberg",
  "email": "ben@email.io",
  "phone": "1234567890",
  ...
} + {
  ...
  "fields": [
    { "name": "id", "type": "int" },
    { "name": "first", "type": "string" },
    { "name": "last", "type": "string" },
    { "name": "email", "type": "string" },
    { "name": "phone", "type": "string" },
  ]
}
```

Parquet

Es un formato de almacenamiento columnar que surge de la colaboración de Twitter y Cloudera.

Los datos se almacenan en formato binario y al final del archivo se agrega la metadata (schema).

Este formato es ideal para agregaciones AVG, SUM, etc.



<https://www.slideshare.net/julienledem/parquet-stratany-hadoopworld2013>

Resumen

Properties	CSV	JSON	Parquet	Avro
Columnar	X	X	✓	X
Compressable	✓	✓	✓	✓
Splittable	✓*	✓*	✓	✓
Readable	✓	✓	X	X
Complex data structure	X	✓	✓	✓
Schema evolution	X	X	✓	✓

Links de referencia

- **Parquet** <https://parquet.apache.org/documentation/latest/>
- **Avro** <https://avro.apache.org/docs/current/>
- **JSON** <https://www.json.org/json-es.html>

Gracias