Big Data con Hadoop y Spark

Módulo 02 - Almacenamiento

Objetivos

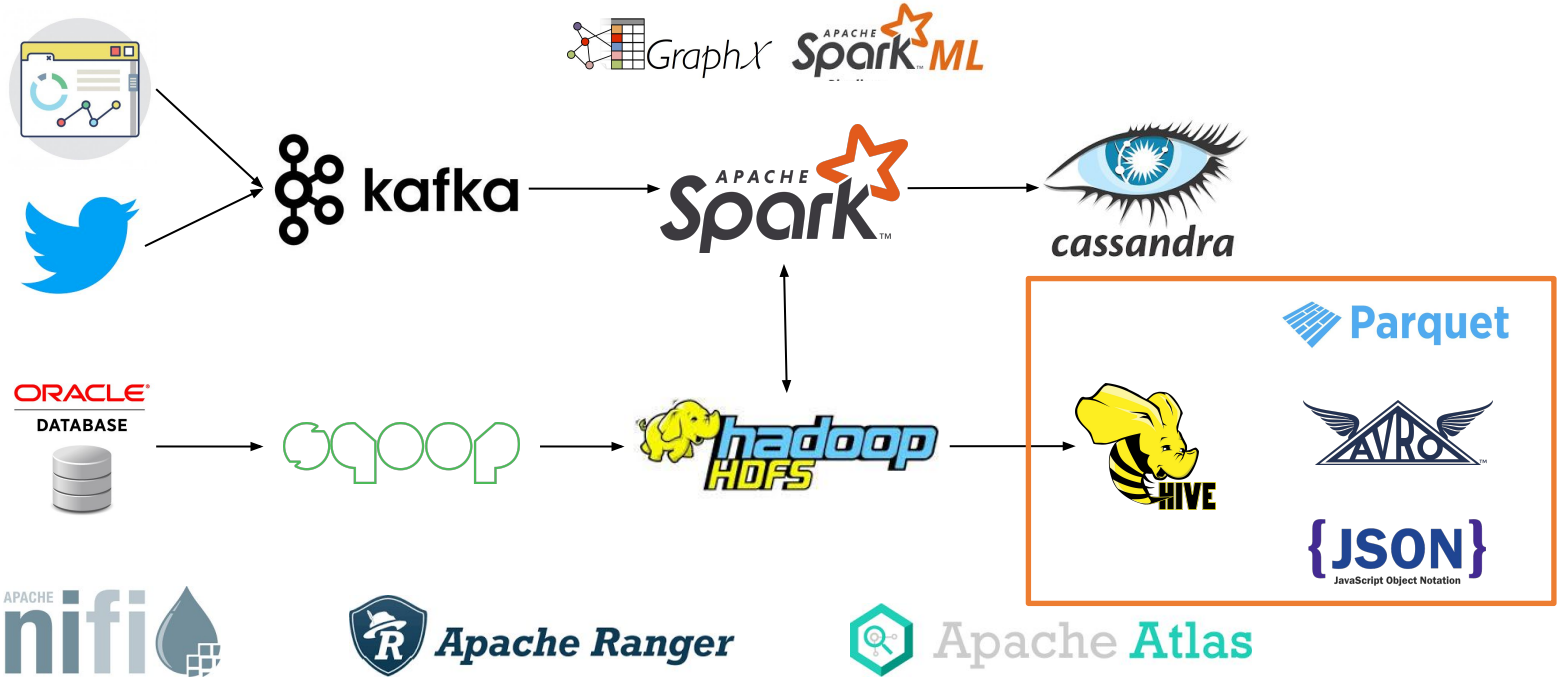
En este módulo verás:

- Introducción a JSON
- Introducción a Avro
- Introducción a Parquet

Al final de la clase serás capaz de:

- Seleccionar formatos de almacenamiento para Big Data
- Utilizar SerDes de Hive para interpretar JSON, Avro y Parquet

Frameworks

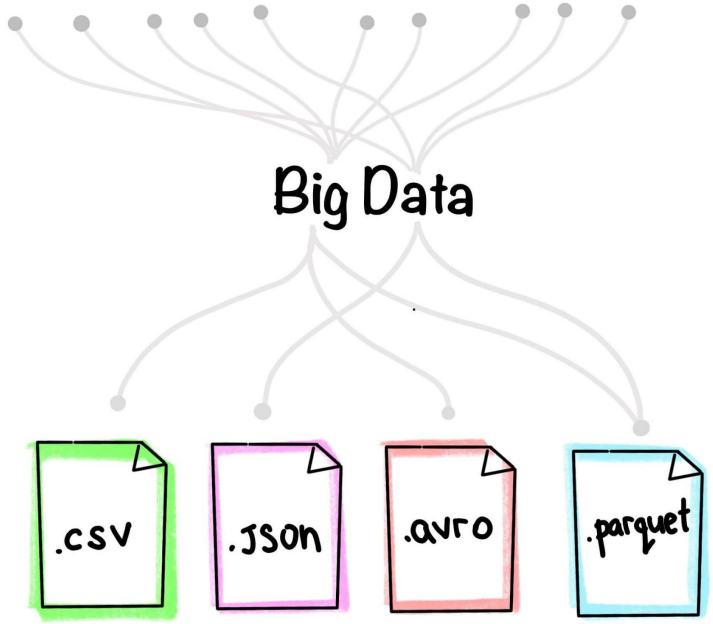

GraphX
APACHE Spark ML
kafka
APACHE Spark
cassandra
ORACLE DATABASE
sqoop
hadoop HDFS
HIVE
Parquet
AVRO
{JSON}
JavaScript Object Notation
APACHE nifi
Apache Ranger
Apache Atlas

Formatos de Almacenamiento

La elección de un formato correcto puede traducirse en mejoras de performance y reducción de costos.

Un ejemplo conocido es el uso de Apache Parquet en AWS Athena.

[AWS Athena - Columnar Storage]



Factores de Elección

Al momento de elegir un formato de almacenamiento, debemos considerar los siguientes puntos:

- **ROW vs COLUMN**: las consultas serán de tipo SELECT * o agregaciones AVG, SUM, etc
- **SCHEMA EVOLUTION**: que sucede si debemos agregar, eliminar o modificar un campo
- **COMPRESSION**: equilibrio entre espacio en disco utilizado y tiempo de procesamiento

Row vs Column

	day	location	product	sale
row 1	2017-01-01	l1	p1	300
row 2	2017-01-01	l1	p2	40
row 3	2017-01-01	l2	p1	44
row 4	2017-02-01	l1	p1	200

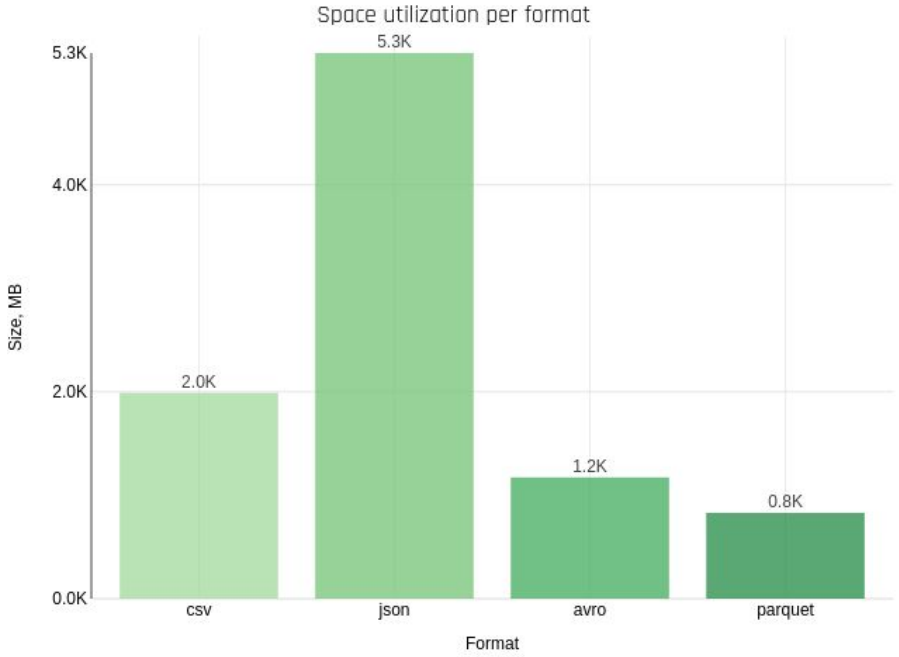
Traditional Memory Buffer	
row 1	2017-01-01
	l1
	p1
	300
row 2	2017-01-01
	l1
	p2
	40
row 3	2017-01-01
	l2
	p1
	44

Columnar Storage	
day	2017-01-01
	2017-01-01
	2017-01-01
	2017-01-02
location	l1
	l1
	l2
	l1
product	p1
	p2
	p1
	p1

Schema Evolution

```
{"namespace": "drwho.avro",
 "type": "record",
 "name": "drwho",
 "fields": [
    {"name": "drwho_season", "type": ["null","string"], "aliases": ["doctor_who_season"]},
    {"name": "drwho_actor", "type": ["null","string"], "aliases": ["doctor_actor"]},
    {"name": "episode_no", "type": ["null","string"]}, {"name": "episode_title", "type": ["null","string"]},
    {"name": "date_from", "type": ["null","string"]},  {"name": "date_to", "type": ["null","string"]},
    {"name": "estimated", "type": "string"}, {"name": "planet", "type": ["null","string"]},
    {"name": "sub_location", "type": ["null","string"]}, {"name": "main_location", "type": ["null","string"]},
    {"name": "hd", "type": "string", "default": "no"}
]}
```

Compression



https://luminousmen.com/post/big-data-file-formats

Gracias