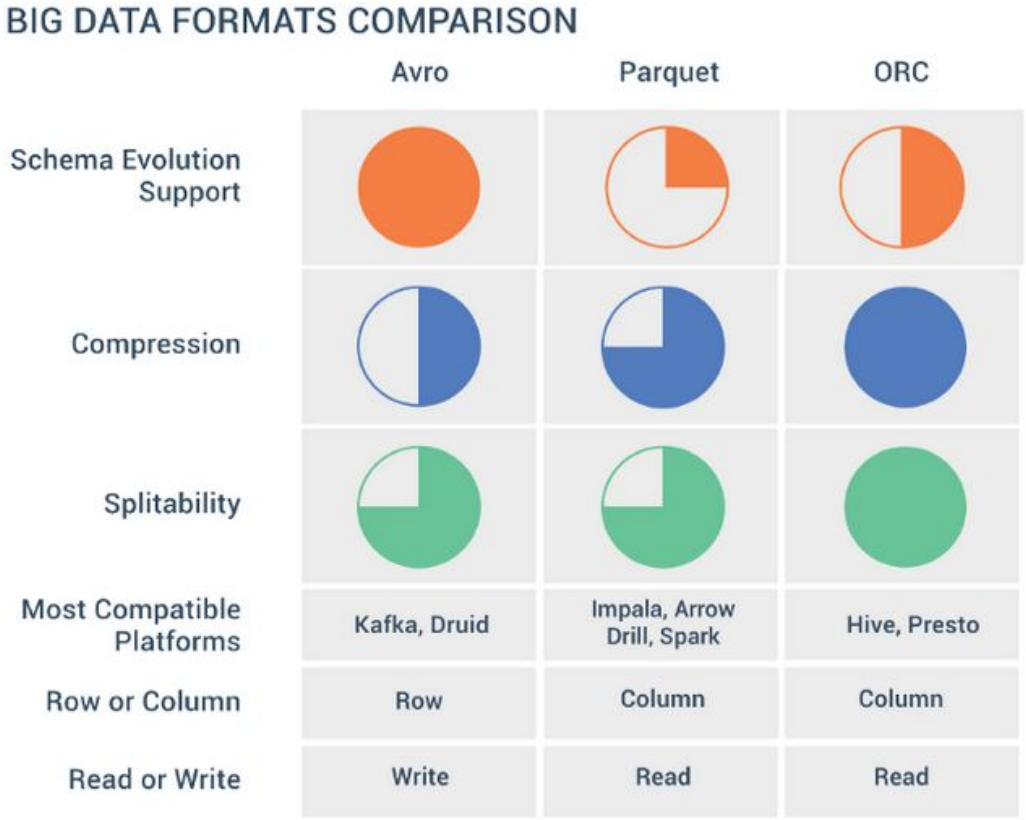
Formati datoteka na HDFS

Osobine, primeri

Formati datoteka dostupni na HDFS

- Text
  - strukturirani ili nestrukturirani tekst (CSV, JSON)
- SequenceFile
  - binarni sadržaj bez metapodataka
  - Podvarijanta MapFile
- Avro
  - Redovno-orijentisan
  - Proizvoljno kompleksna šema opisana u JSON formatu
- Parquet
  - Kolonski-orijentisan
  - Sadrži metapodatke
  - Podržava ugnježdene strukture podataka
- ORC (Optimized Row Columnar)
  - Kolonski-orijentisan za grupe redova
  - Sadrži metapodatke

HDFS formati - poređenje



Hive SerDe

- Serijalizacija/deserijalizacija podataka u Hive
- Hive podržava
  - TextInputFormat/HiveIgnoreKeyTextOutputFormat
  - SequenceFileInputFormat/SequenceFileOutputFormat
  - Thrift
  - Avro
  - ORC
  - Parquet
- Definiše se prilikom definicije tabele sa STORED AS
- Dodatni parametri u SERDEPROPERTIES

Text format

- Izvor: data.gov.rs
- CSV fajl sa svim registrovanim fondacijama u Srbiji
- Dodati fajl na HDFS
- Kreirati Hive eksternu tabelu sa CSV SerDe
- Proveriti podatke
- Kreirati tabelu particionisanu po godinama
- Učitati podatke iz eksterne tabele pomoću dinamičkog particionisanja
- Proveriti podatke i kreirane fajlove na HDFS

Avro format

- Kreirati Hive tabelu sa Avro formatom
  - AvroSerDe
  - Input i Output format
  - Avro schema u TBLPROPERTIES
- Učitati podatke iz eksterne u Avro tabelu
- Proveriti podatke i kreirane fajlove na HDFS

Parquet format

- Kreirati Hive tabelu u Parquet formatu
  - STORED AS PARQUET
- Podrazumevana kompresija
  - set parquet.compression=SNAPPY;
- Učitati podatke iz eksterne u Parquet tabelu
- Proveriti podatke i kreirane fajlove na HDFS

Zadatak

- Izvor: https://sdm.lbl.gov/fastbit/data/samples.html
- Koristeći zadati CSV fajl kreirati
  - Eksternu tabelu
  - CSV tabelu
  - Avro tabelu
  - Parquet tabelu
- Obratiti pažnju na veličinu fajlova u različitim formatima

Koji format odabrati?

- Avro je pogodan za *schema evolution* u slučaju česte promene strukture skupa podataka
- ORC je izvorno projektovan za Hive i nudi najefikasniju kompresiju podataka zbog kolonske ogranizacije
- Parquet takođe ima dobro kompresiju, podržava rad sa ugnježdenim kolonama, u praksi se izdvojio kao prvi izbor za kolonski format prilikom obrade podataka u Sparku