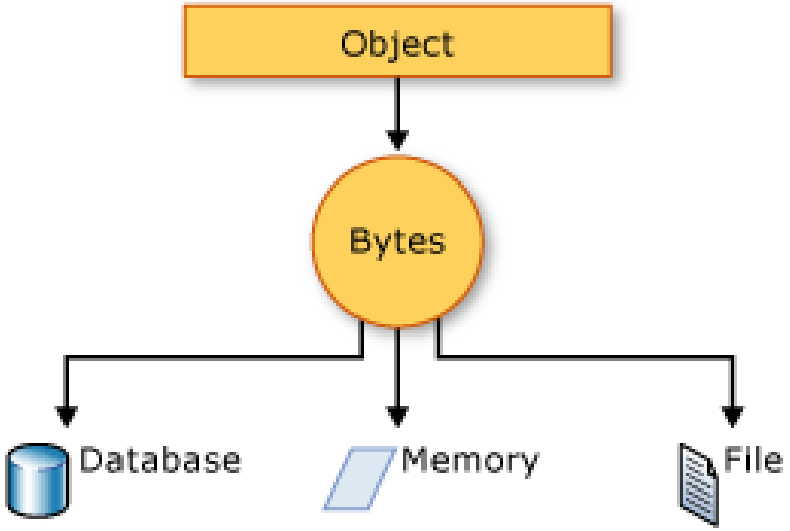
Hadoop i Spark

Mariusz Rafało
mrafalo@sgh.waw.pl
http://mariuszrafalo.pl

FORMATY PLIKÓW

Serializacja

Serializacja to proces konwersji obiektu w strumień bajtów. Serializacja pozwala na zapamiętanie poszczególnych stanów obiektu.



Podstawowe formaty danych

- Pliki tekstowe
- Avro
- Parquet
- ORC

Pliki tekstowe

- CSV
- JSON
- XML
- Logi

```
{
    id: "tt1231",
    title: "Pride and Prejudice"
    year: 2093,
    director: "Michael Bay",
    genres: [
        "Horror",
        "Comedy"
    ],
    stars: [
        {
            name: "Kiera Knightley",
            id: 9863
        },
        {
            name: "Danny DeVito",
            id: 2031
        }
    ]
}
```

Avro

- Serializowany format danych
- Wierszowe składowanie danych
- Schemat danych jest zakodowany wewnątrz pliku
- Umożliwia efektywną kompresję
- Umożliwia podział zbioru danych
- Umożliwia zmiany modelu danych

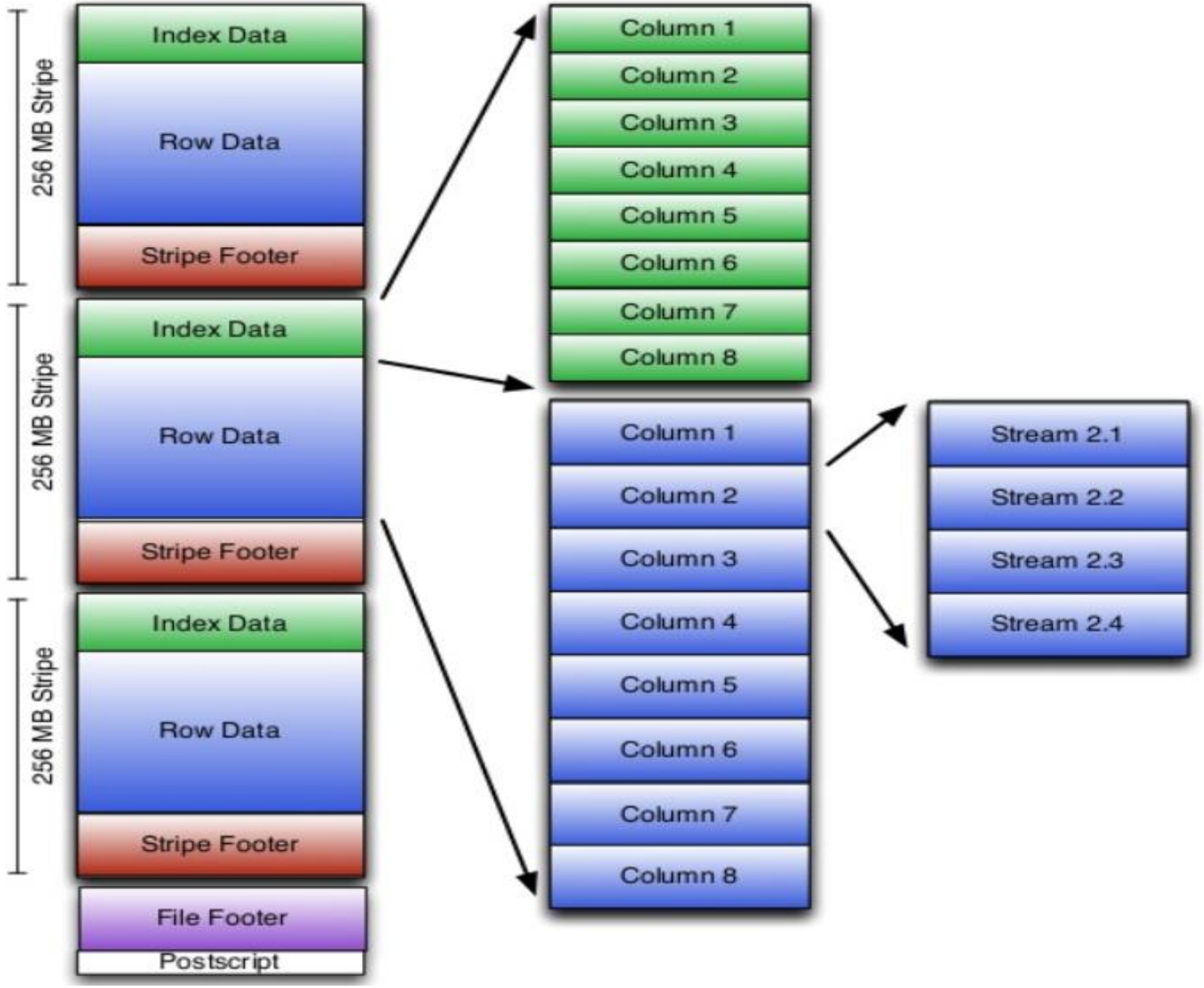
Parquet

- Serializowany format danych
- Kolumnowe składowanie danych
- Zastosowano algorytm składowania danych Dremel
- Każdy plik zawiera wartości dla wybranych wierszy
- Wydajny w zakresie operacji I/O

ORC (Optimized Row Columnar)

- Kolumnowy format pliku
- Struktura pliku posiada: *stripes, footer, postscript*
  - *Stripes*
    - Wiersze danych w układzie kolumnowym
  - *Footer*
    - Lokalizacje poszczególnych stripes
    - Typy danych
    - Podstawowe statystyki dla poszczególnych kolumn (min, max, count, sum)
  - *Postscript*
    - Parametry kompresji

ORC



Źródło: https://www.slideshare.net/oom65/orc-files

BIG DATA FORMATS COMPARISON

	Avro	Parquet	ORC
Schema Evolution Support			
Compression			
Splitability			
Most Compatible Platforms	Kafka, Druid	Impala, Arrow Drill, Spark	Hive, Presto
Row or Column	Row	Column	Column
Read or Write	Write	Read	Read

Source: Nexla analysis, April 2018

Dziękuję za uwagę