Formaty danych

Agenda

1. Partycjonowanie danych
2. Rodzaje formatów danych
3. Transformacja między formatami danych
4. Przetwarzanie formatów danych (część warsztatowa)

Jak to jest z danymi ?

- schema vs schema less
- ewolucja schematu
- wydajność
- integracja z ekosystemem
- struktury danych

Formaty danych

- Plain text (CSV, TSV)
- Sequence File
- JSON
- ORC
- **Avro**
- **Parquet**

Formaty kompresji

- GZip
- Bzip2
- LZO
- **Snappy**

Avro

- nie tylko format plików
- wymaga zdefiniowanego schematu
- JSON do opisu schematu
- duża ilość typów danych
- ewolucja oraz kompatybilność schematów

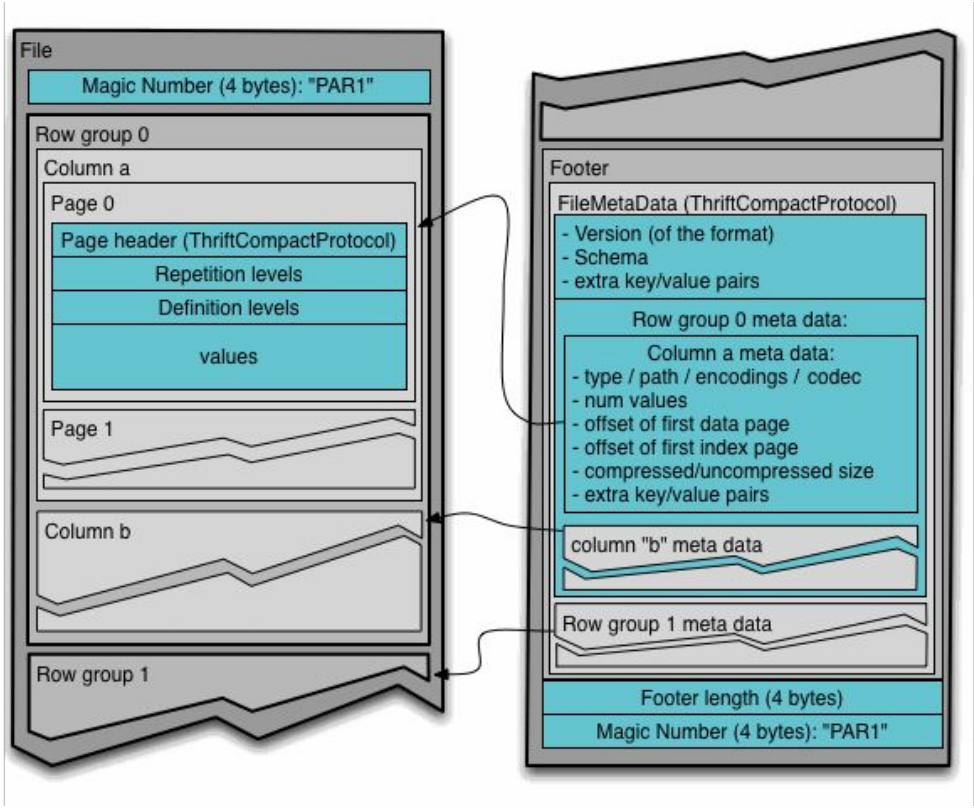
Avro - przykładowy schemat

```
{
    "namespace": "my.portal.users",
    "type": "record",
    "name": "User",
    "fields": [
     {"name": "name", "type": "string", "default": "NONAME"},
     {"name": "id",  "type": ["null", "int"], "default": NULL},
     {"name": "favorite_color", "type": ["null", "string"]}
    ]
}
```

Parquet

- kolumnowy format plików
- można używać dodatkowo kompresji
- obsługiwany przez wszystkie narzędzia ekosystemu
- wysoka wydajność

Parquet



Źródło: parquet.apache.org

Partycjonowanie danych

- push down predicate
- dane partycjonujemy po kluczu z którego najczęściej korzystamy
- nie bójmy się inaczej modelować zbiory oraz budować agregaty
- wykorzystaj ten format danych, który najbardziej odpowiada potrzebie