Semantix

Formato de Armazenamento

Aula 5

Formato de Armazenamento e Compressão

Semantix

Formatos de Arquivo

- Text File
- Sequence File
- RC File
- ORC File
- Avro
- Parquet

Text File

- Arquivos de texto
- Formato padrão
  - Hive
  - Sqoop
- Facilidade para compartilhar os dados do Hadoop com outros sistemas externos
- Facilidade de edição manualmente
- Menos eficiente que outros formatos
- Exemplo
  - txt
  - csv
  - Estruturas de texto
    - xml
    - json

Sequence File

- Arquivo de Sequência do Hadoop
- Formado por pares chave e valor
- Armazena em formato binário
- Mais eficiente que o arquivo de texto
- Facilidade para compartilhar os dados com outras ferramentas do Hadoop

RC File

- Record Colunnar File
- 1º formato de arquivo colunar do Hadoop
  - Formado por grupos de colunas
- Armazenamento horizontal dos dados
- Vantagem
  - Agilidade para carregamento de dados
  - Agilidade para processamento de consultas
  - Espaço de armazenamento eficiente
- Desvantagem
  - Utiliza mais memória e computação
  - Não suporta a evolução do esquema

ORC File

- Optimized Row Colunnar File
- Substituiu o formato RC File
  - Mesmas características
- Compacta melhor os arquivos RC
  - Consultas mais rápidas
- Projetado para otimizar o desempenho no Hive
  - Formado por faixas
    - Grupo de dados de linha
  - Não usado para MapReduce não-Hive
    - Pig
    - Impala
    - Java

Avro

- Formado por serialização de dados com neutralidade de linguagem
- Armazenamento dos dados e metadados juntos
- Vantagem
  - Suporta MapReduce
  - Suporta evolução do esquema

Parquet

- Formato colunar
  - Formado por grupos de colunas
- Vantagem
  - Suporta MapReduce
    - Pig
    - Hive
    - Java
  - processamento
    - Impala
    - SparK
  - Suporta evolução do esquema
    - Hive
    - Impala

Compressão de dados

- Armazenamento x Velocidade de leitura

- ZLIB (GZip)
  - Alta compressão, lento
- Snappy
  - Baixa compressão, rápido

Tipos de arquivo e compressão

Formato de armazenamento	Compressão	Tamanho do arquivo ou tabela
Texto	Uncompressed	
Texto	Snappy	
Texto	Gzip	
Avro	Uncompressed	
Avro	Snappy	
Avro	Gzip	
Parquet	Uncompressed	
Parquet	Snappy	
Parquet	Gzip	

Semantix

Obrigado!

Alguma pergunta?

Você pode me encontrar em:

rodrigo.augusto@semantix.com.br

GET SMARTER