

Data Sources

- Gravação de tabelas persistentes
- Bucketing
- Arquivos avro



Data Source

Local de onde se originam os dados que estão sendo usados. É a configuração do local onde os dados são armazenados



01

**Tabelas
persistentes**



02

Bucketing


Agrupamento por
classes



03

Arquivos Avro

Agrupamento serializado
com esquema de dados





01

Tabelas Persistentes

Tabelas persistentes

Salva o conteúdo do *DataFrame* como a tabela especificada.

Caso a tabela já exista, o comportamento desta função depende do modo de salvamento, especificado pela função mode. Quando o modo é Overwrite, o esquema do *DataFrame* não precisa ser o mesmo da tabela existente.

- **append**: Adiciona os dados do *Dataframe* aos dados existentes.
- **overwrite**: Substitui os dados existentes.
- **error ou errorifexists**: Lança uma exceção se os dados já existirem.
- **ignore**: Ignore silenciosamente esta operação se os dados já existirem.

Como usar?

- O comando *saveAsTable*

Esse comando materializa o conteúdo do DataFrame e cria um ponteiro para os dados no metastore



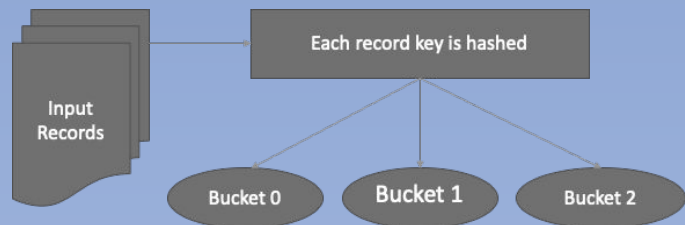
02



Bucketing

Agrupamento por classes

Bucketing



Agrupar os registros em buckets menores na tentativa de minimizar o risco de um invasor associar informações confidenciais a informações de identificação.

Assim, será possível reter o significado e a utilidade, mas os valores individuais que têm poucos participantes poderão ser ocultados.

Bucketing



Bucketing é uma técnica usada para otimizar o desempenho da tarefa. Com base no valor de uma ou mais colunas de buckets, os dados são alocados para um número predefinido de buckets.

O bucketing envolve classificar e embaralhar os dados antes da operação que precisa ser executada em dados como junções.



Bucketing

- Essa técnica beneficia as tabelas de dimensão, que são tabelas frequentemente usadas contendo chaves primárias.
- O bucketing é comumente usado para otimizar o desempenho de uma consulta de junção.

03

Arquivos AVRO

Agrupamento serializado com
esquema de dados





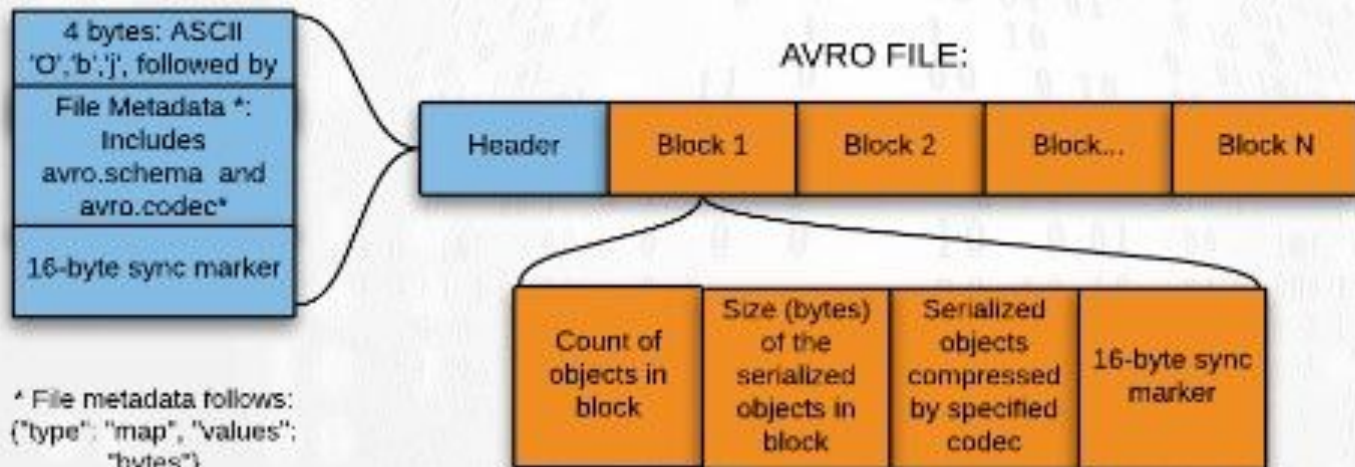
Avro

É um arquivo binário, que são compilados e precisam de bibliotecas para serem lidos.

Dentro do avro file existe um schema file, que fala os data types de cada campo, ou seja, nome/string, idade/number ou um campo booleano por exemplo. Isso no JSON não existe, ponto para o Avro.

O Avro funciona com uma estrutura de sub itens de um item assim como o JSON

File Structure - Avro



Avro



- Possui um sistema de particionamento o que acelera absurdamente na consulta de dados.
- Ponto negativo: quando uma informação é requerida ele lê todo o arquivo e não só a parte que precisa.



Obrigado!

Vinicius Carvalho
Big Data Analytics
11/2022



Referências



Rescue Point - Big Data

<https://medium.com/rescuepoint/tagged/big-data>

Bucketing no Spark

<https://blog.clairvoyantsoft.com/bucketing-in-spark-878d2e02140f>