K-Store / Parquet benchmarks

This benchmark tends to prove that K-Store has better performance than Parquet for read / write operations.

Introduction

All benchmarks have been run with 3different datasets. They contain 1 million, 10 million and 50 million lines. For each benchmark we change schema size with 8 columns, 24 columns, 48 columns. Finally, we never take the whole dataset, but select some columns which aren't ordered.

Data mapping between Parquet and K-Store

Parquet	K-Store
Binary	String
Int32	Int
Int64	Bigint
Int96	Bigint
Float	Float
Double	Double
Boolean	Tinyint
Fixed_len_byte_array	String

Hardware, parameters and compression

Data are stored in Amazon S3.

Benchmark have been run on a Amazon EC2 m4.xlarge virtual machine, including 4 vCPU and 16 GO RAM.

K-Store and Parquet use Snappy compression during all benchmarks.

Benchmark have been executed with java. Java VM options are: -Xms8000m -Xmx8000m

Benchmark results

Schema with 8 columns

We have accessed to 4 columns on position 1, 3, 5, 6. Those columns are int32, boolean, double, fixed_len_byte_array.

With 1M lines

Time (ms) / Storage	K-Store	Parquet
Average	1099.76	1919.76
Minimum	845.37	1734.65
Maximum	2802.21	3213.34

With 10M lines

Time (s) / Storage	K-Store	Parquet
Average	8.95	17.15
Minimum	8.16	16.58
Maximum	9.18	19.07

With 50M lines

Time (s) / Storage	K-Store	Parquet
Average	44.29	86.08
Minimum	41.69	83.09
Maximum	47.91	93.39

Schema with 24 columns

We have accessed to 6 columns on position 0, 4, 9, 15, 18, 22. Those columns are binary, float, int32, int96, int64, fixed_len_byte_array

With 1M lines

Time (s) / Storage	K-Store	Parquet
Average	2.01	4.08
Minimum	1.44	2.88
Maximum	2.78	7.22

With 10M lines

Time (s) / Storage	K-Store	Parquet
Average	15.56	32.34
Minimum	13.07	28.80
Maximum	19.91	36.53

With 50M lines

Time (s) / Storage	K-Store	Parquet
Average	78.10	136.19
Minimum	71.23	125.99

Maximum	95.95	161.46

Schema with 48 columns (part-1)

We have accessed to 8 columns on position 1, 8, 18, 21, 28, 33, 37, 40. Those columns are int32, binary, int64, double, float, int32, double, binary

With 1M lines

Time (s) / Storage	K-Store	Parquet
Average	1.85	3.98
Minimum	1.53	3.15
Maximum	2.82	6.70

With 10M lines

Time (s) / Storage	K-Store	Parquet
Average	15.61	35.61
Minimum	14.63	29.11
Maximum	16.75	55.29

With 50M lines

Time (s) / Storage	K-Store	Parquet
Average	74.27	206.12
Minimum	73.27	185.59
Maximum	75.13	247.62

Schema with 48 columns (part-2)

We have accessed to 9 columns on position 4, 8, 18, 22, 25, 28, 32, 38, 45. Those columns are float, binary, int64, fixed_len_byte_array, int32, float, binary, fixed_len_byte_array, double

With 1M lines

Time (s) / Storage	K-Store	Parquet
Average	3.01	6.56
Minimum	2.56	5.78
Maximum	4.49	9.24

With 10M lines

Time (s) / Storage	K-Store	Parquet
Average	25.71	66.09
Minimum	23.71	57.24
Maximum	27.88	82.07

With 50M lines

Time (s) / Storage	K-Store	Parquet
Average	130.49	367.40
Minimum	127.03	348.78
Maximum	132.77	394.06