Meeting 4/26/18

Wednesday, April 25, 2018 11:32 PM

Attendees: everyone

What did we learn about Spark, Parquet?

Parquet:

- Main goals:
  - Interoperability
    - Frameworks and libraries that are integrated with Parquet
      - Query Engines:
        - Hive, Impala, HAWQ, IBM Big SQL, Drill, Tajo, Pig, Presto
      - Frameworks:
        - Spark, MapReduce, Cascading, Crunch, Scalding, Kite
      - Data Models:
        - Avro, Thrift, ProtocolButters, POJOs (plain old java objects)
  - Space Efficiency
    - Column storage, store all data by column so they're all the same type, can use arrays versus
  - Query Efficiency
    - Skip all columns you don't need

The right encoding for the right job

Delta encodings:

For sorted datasets or signals where the variation is less important than the absolute value (timestamp, auto-generated ids, metrics,) Focuses on avoiding branching. (sorted, ints)

Prefix coding (delta encoding for strings)

When dictionary encoding does not work

Dictionary encoding

Small (60k) set of values (server IP, experiment ID)

Run Length Encoding:

Repetitive data

READINGS

https://research.google.com/pubs/pub36632.html -> Abstract. Actual link below
https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/36632.pdf

https://arxiv.org/abs/1209.2137

mapr = user
mapr = password + super user

V 8.2

Andy used this on the CLI:
Create table my_key as select mytab.marketing_info.keywords from dfs.tmp sampleparquet mytab;

Alter session set 'store_format'=parquet';

These scripts are in mapr demo

#quit

Steps to get the environment:
Get the OVA
Get Maven
Download Git repo for parquet tools (Parquet MR)

Future work:

- Configure environment
- Port CSV from previous tests to parquet file
- Experiment