Dremel: Interactive Analysis of Web-Scale Datasets

wbg231

January 2023

1 abstract

- dremel is a scalable interactive, ad-hco query system for analysis of read only nested data
- is fast as it combines multi level ecxeution trees with columnar data layout
- this scales both with data and cpus

introduction

- data used in science and the internet is often non relational
- hence a flexible data model is essential in these domains
- a lot of this data lends it's selt naturally to nested representations
- dremel is a system that supports interactive analysis of very large data sets over shared clusters of commodity machiens
- dremel works well with nested data and works well with map reduce
- the architecture of dremal uses serving tree structure used in web search
- the result of the auery is asembled by aggregatignng the replies recived from lower level of trees
- dremel proied a high level sql like language to express ad hoc queries
- it can execute queries with out map reduce jobs
- dremel is also column oriented which enables it to read less data from seconday storacges adn have chpeater compressin

background

- there needs to be interoperation between the query processor and ohter data manamgnet tools
- ie they need to be able to work togther as a combined system
- this reuqires a common storage later
- the next step of this is a shared storage format, coloumnar storage is good for relational data but dremel helps with nested data

data model

- i am going to go over the lecture and come back to this if there is time.