Distributed System MP4: MapleJuice + SQL

Group 41 Cheng-Hsuan Huang(chhuang5), Yen-Cheng Yeh(yy63)

Design

We have implemented a C++ based Distributed System: MapleJuice + SQL built on previous MPs.

MapleJuice

- Maple:

The Maple phase starts by having the leader divide all input files within the target directory evenly, based on the indicated number of maple tasks. Subsequently, the leader transmits information about these input files to each worker, specifying the file and starting file position they should process using maple.exe. The maple.exe process generates files according to different keys, allowing each server to write each respective key file to the SDFS system sequentially. When the leader detects server failure, it will resend the input information to other active servers and wait for the task to be completed.

- Juice:

The juice phase operates in a similar manner. The leader first examines the maple intermediate files within the SDFS system, divides them evenly, and dispatches information about these files to each worker. Each worker processes a group of key files and subsequently writes the key-value pair to the designated 'sdfs_dest_filename'. Similarly, multiple workers are prevented from simultaneously writing to these files. Additionally, if the delete parameter is set to 1, the maple intermediate files will be deleted. Like the maple phase, when the leader detects server failure, it will resend group of key information to other active servers and wait for the task to be completed.

SQL

The SQL layer is based on the MapleJuice system, we change the execution files in maple phase and juice phase and set the delete parameter to 1.

- Filter:

In the Filter function, the maple executable reads the input file and the specified regex condition. Maple takes in the regex condition as the key, and the entire line as value. Each worker generates its key file and subsequently writes these files into the same SDFS file (sequentially, not simultaneously). During the juice phase, no specific action is performed. The juice executable simply transfers the data from the key file to the destination file.

- Join:

In the Join function, the maple executable reads the input files of both dataset 1 and dataset 2, along with the specified field for each dataset. It generates the key file where the key aligns with the value of the specified field, then writes these files into the SDFS system. During the juice phase, the juice executable writes key-value pairs to the designated 'sdfs_dest_filename' and deletes the intermediate key files generated in the maple phase.

Measurements

Due to debugging and unforeseen code challenges, the experiment wasn't finished within the expected timeframe, therefore, were not able to complete the measurement section of the report. Despite setbacks, we *gained valuable insights into complexities involved. Our focus remains on resolving issues, refining code, and setting revised timelines for completion. Our commitment to achieving the experiment's objectives persists despite the challenges.*