Parallel Computing and I/O

HDF5-Zugriffsmusteranalyse zur Datenbankabstrahierung

- Author: Karl Lobedank
- Type: Bachelor's Thesis
- Date: 2021-05-12
- Reviewers: Jun.-Prof. Dr. Michael Kuhn, Dr. David Broneske
- Supervisors: Kira Duwe
- Download: PDF

HDF5 ist ein selbstbeschreibendes Datenformat, welches in vielen Forschungsgebieten Einsatz findet. Der Vorteil von HDF5 lieg in einem simplen Datenaustausch zwischen mehreren Institutionen. Analysen über die Dateien sind aber nur schwer zu realisieren, da die Dateien oft auf Bandarchiven gespeichert werden und jede Auswertung das Laden sämtlicher Dateien vom Bandarchiv erfordert. Diese Ladevorgänge hemmen die Analysen. Um dies zu vereinfachen, sollen HDF5-Dateien in eine Datenbank überführt werden. JULEA, ein flexibles Speicherframework, bietet bereits über ein VOL-Plugin die Möglichkeit zur Speicherung der Daten in einer relationalen Datenbank (für die Metadaten) und in einem Objektspeicher (für die reinen Daten) a Durch diese Abstrahierung lassen sich leichter diejenigen Dateien ausfindig machen, welche für eine gewählte Analyse erforderlich sind, wodurch Speicherzugriffe minimiert werden; des Weiteren können häufig verwendete Werte - Minima, Maxima oder Mittelwerte - gesondert in der Datenbank abgespeichert werden. In dieser Arbeit sollen die Zugriffsmuster auf HDF5-Dateie von HPC (High Performance Computing)-Anwendungen wie ENZO analysiert werden, um Aussagen über eventuell effizientere Datenbankmodelle zu treffen. Für diese Analyse wurde in dem bereits vorhandenen VOL-Plugin von JULEA ein Logging implementiert, welches alle Zugriffe dokumentiert. Die Zugriffe können dabei über die Konsole, eine Logdatei oder eine Datei zu Visualisierung über Graphviz ausgewertet werden. Hierbei hat der Anwender/die Anwenderin die Möglichkeit, selbst zu entscheiden, welche Dokumentationsvariante, welche Zugriffe oder welche Datentypen, auf die die Zugriffe erfolgen, dokumenti werden sollen. Mit Hilfe der Analyse konnten Aussagen über Anforderungen an eine Datenbank gemacht werden. Dabei stellte sich heraus, dass die aktuelle Speicherung in dem vorliegenden Schema im relationalen Datenmodell ineffizient ist. Es wurden daher zwei neue Strukturen entwickelt und entsprechenden Tests unterzogen, wobei gezeigt wurde, dass die neuen Strukturen die Daten effizienter speichern als die alte Struktur. Des Weiteren wurde eine theoretische Erörterung über weitere Datenbankmodelle durchgeführt, wobei das Wide-Column-Stores-Modell durch seine Flexibilität, die Skalierbarkeit und weitere Merkmale als das zur Abstrahierung der Daten effizienteste hervorging.