V236b　HSC 巨大天体カタログの高速検索に向けた次世代データベースの開発

古澤久徳, 高田唯史, 山田善彦, 大倉悠貴（国立天文台）, 鬼塚真（大阪大）, 須賀秀和, 黒澤亮二, 神林飛志（ノーチラス・テクノロジーズ）

近年データベース技術は、大規模データ高速処理への需要の高まりを受け、変革の時を迎えている。国内では産学連携の枠組みの中で、オープンソース基盤でのオンラインデータ登録と解析処理機能の開発、またそれらのハイブリッド化により、データベースを介した高速データ処理を実現するための新しいリレーショナルデータベースの開発が進められつつある。

一方、天文学におけるデータベース活用はこれまで小中規模の情報管理に限られ、比較的大きな天体カタログでも高々数億行程度の表を扱うのみであった。しかし、Hyper Suprime-Cam(HSC) の戦略枠観測では科学データアーカイブによる研究成果の推進を目標に掲げており、最終的には全積分ごとに 3000 パラメータ以上の測定情報を展開した 300 億行超えの表に対してより複雑な検索を扱わなければならない。HSC チームはデータベースの分散化などによる性能改善を試みているが、数百億行の表への応用を見越して技術開発を続ける必要がある。

我々は今回、上で述べた新しいデータベース開発の動作実証に参加する機会を得た。5 年計画の初年度は、HSC の全世界公開データをデータベースに格納し、典型的かつ負荷の高い検索例について Hadoop クラスター上での Spark SQL や Impala などの分散クエリ技術による高速化を試験した。この試験環境を用いて、クエリのワークロード情報による表設計の最適化を行い、さらに性能向上を図る予定である。また、時系列の天体測定情報を含むデータベースを作成し、様々な測定パラメータから外れ値を探索的に見つけ出す方法の開発も行う予定である。この技術を応用し、HSC 戦略枠の蓄積データから変動・突発天体候補を効率的に抽出することを目指している。