Leveraging Hadoop for preparing the NHCDC cost data

Andy Yang[1]

[1] IHPA

This paper will discuss executing traditional data preparation workload in the Hadoop environment at IHPA. Since 2011-12 financial year, IHPA has collected National Hospital Cost Data Collection (NHCDC) in-house on a yearly basis. IHPA conducts a practical business-context driven Extract, Transform and Load (ETL) process to prepare the data for National Efficient Price and National Efficient Cost development and benchmarking, which includes three main steps: 1) Extract activity data and link with costing files. DRG, URG, and SNAP are regrouped to derive classification under the same roof. A hash table algorithm in the URG grouping significantly boosts the processing efficiency gain by 30% in our tests. 2) For Acute and Subacute data stream – apply business rules for two-stage processing: Firstly, link by EpiNoMother, Linked by Date of Birth. Secondly, distribute the cost of unlinked unqualified Neonates (UQB) Costs to the unlinked Mother data. 3) Reconciliation and QA checks highlight anomalies by comparing the trends of activity and cost over last two years.

As technology advances, adopting Hadoop environment for data preparation equips us with 1) Computing power (Spark, NoSQL, MapReduce programming etc.) 2) User friendly query builder: The end users even with limited technical skills can create queries and interact with data on their own. 3) Scalability and cost saving. IHPA is able to cut monthly expenses by only increasing the nodes during the peak time. In the meantime, this paper explains our approach to tackle challenges that many companies face in the big data environment: 1) MapReduce programming is not a good match for solving all problems. 2) Data security. A demo will be provided to illustrate highly interactive dash boarding for monitoring the data quality during full life-cycle of analytical processes. This real-time analytical tool is developed based on our secure cloud computing platforms.

In summary, this paper explores the NHCDC data preparation on big data storage and visualization techniques by leveraging the power of hadoop.