MSE Database Seminar GPU Databases

By
Samuel Kurath

University of Applied Sciences, HSR

Autumn 2017

Samuel Kurath

Advisor:
Prof. Keller Stefan

MSE Database Seminar GPU Databases

Abstract

The content of this paper is splitted into three parts, it begins with an overview of data streams and their difficulties, followed by a part about Apache Storm a distributed stream processing framework. And finally a concrete implementation based on a given problem, solved with Apache Storm. The goal of the implementation is to analyze the minutely updated Augmented Diffs of OpenStreetMap and do a benchmark to test the performance of Apache Storm.

Contents

0 Introduction

Durring the Master of Science in Engineering the students have to participate at two seminars. The goal of these is to elaborate a theme on their own, discuss the result in group and write a paper about the topic.

The Databasesystems Seminar does a focus on GPU Database Systems. The students do have a closer look at a certain GPU Database product and have to do a benchmark.

The benchmark is based on the public NYC Taxi Rides dataset and the queries are predetermined by Prof. Stefan Keller.

With great power comes great responsibility.

Voltaire

1 GPU Databases

Initial example

General

There are some things you learn best in calm,
and some in storm.

Willa Cather

2 MapD

MapD is a GPU Database.

Figure 2.1: MapD

Overview

Basics

Installation, mapdql, similar to postgres \t \d etc.

A good idea is about ten percent and implementation and hard work, and luck is 90 percent.

Guy Kawasaki

3 Benchmark

Benchmark

NYC Taxi Rides

Figure 3.1: Taxi Dropoffs

Queries

The following list is about the queries

Results

The benchmark led to the following results.

References

[1] Breß, S., Heimel, M., Siegmund, N., Bellatreche, L., & Saake, G. (2014). Gpu-accelerated database systems: Survey and open challenges. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XV* (pp. 1–35). Springer.

A

Installation

Repositories

For the implementation part I used different Github repositories to store the code. And help the user with the deployment and installation of the applications.

MSE-Database-Seminar-GPU-Databases

The MSE-Database-Seminar-GPU-Databases* repository provides docker containers with installation for different GPU Database products. Furthermore the repository includes the queries for the benchmark. Docker containers:

- MapD
- PG-Strom
- BlazingDB

*https://github.com/geometalab/MSE-Database-Seminar-GPU-Databases