**Processing Large Raster and Vector Data in Apache Spark**

**ABSTRACT**

Spatial data processing frameworks in many cases are limited to vector data only. However, an important type of spatial data is raster data which is produced by sensors on satellites but also by high resolution cameras taking pictures of nano structures, such as chips on wafers. Often the raster data sets become large and need to be processed in parallel on a cluster environment. In this paper we demonstrate our STARK framework with its support for raster data and functionality to combine raster and vector data in filter and join operations. To save engineers from the burden of learning a programming language, queries can be formulated in SQL in a web interface. In the demonstration, users can use this web interface to inspect examples of raster data using our extended SQL queries on a Apache Spark cluster.

**Existing System**

Large raster data sets are produced by satellites observing the Earth and by high resolution cameras taking pictures of nano-sized structures – as performed in the Nano Positioning and Measurement Machines (NPMM)3[Ba14; Ha02]. Depending on camera resolution and magnification, the NPMM200 at TU Ilmenau produces data sets up to 17 TB per object, which cannot be handled by single node DBMSs anymore. For managing and processing raster data special purpose DBMSs have been proposed in the past, e. g. SciDB [Br10] or RasDaMan [Ba98]. Popular platforms for very large data sets are Hadoop MapReduce and Apache Spark. However, since their generic data model is not able to utilize characteristics of spatial objects (neighborhood, distances, etc. ), extensions that add spatial vector-only data support have been proposed [EM15; YWS16]. Rasterframes4 is the only system that supports raster data. The combination of raster and vector data, however, is limited to range queries only.

**Disadvantages**

1. Which cannot be handled by single node DBMSs anymore.
2. Generic data model is not able to utilize characteristics of spatial objects

**Proposed System**

In this paper we demonstrate our STARK framework with its support for raster data and functionality to combine raster and vector data in filter and join operations. To save engineers from the burden of learning a programming language, queries can be formulated in SQL in a web interface. In the demonstration, users can use this web interface to inspect examples of raster data using our extended SQL queries on a Apache Spark cluster.

**Advantages**

1. users can interact with the web application:
2. working with large spatial raster and vector data sets

**Hardware Requirements:**

# Processor - Pentium –IV

* Speed - 1.1 GHz
* Ram - 256 MB
* Hard Disk - 20 GB
* Key Board - Standard Windows Keyboard
* Mouse - Two or Three Button Mouse
* Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP
* Coding Language - java