EXTASCI: An Extensible System for Scientific Data Analysis

Yu Chen, Florin Rusu

Introduction & Background

Data generated through scientific experiments and measurements is large and complex, requiring special processing techniques. The size of the data can greatly impact the efficiency and effectiveness of queries. The challenge is to design a system that can handle the vast amount of data in an efficient and scalable manner.

EXTCASI Architecture

EXTCASI is a scalable distributed system for large-scale scientific data processing. It uses a shark-folding mechanism to store the data, allowing easy retrieval and efficient query processing. The system is designed to handle large volumes of data and to support a wide range of scientific applications.

Storage Manager

-**Files:**
  - Data is stored in files, which are organized into chunks and directories.
  - Chunks are blocks of data that are stored together in a file.

-**Direct Data:**
  - The data is stored in a flat format, which makes it easy to manipulate and query.

-**File System:**
  - The file system is designed to handle large amounts of data and to support parallel processing.

-**Shark-Folding:**
  - The data is folded into a tree structure, which allows for efficient querying and parallel processing.

User-Defined Aggregates (UDA)

-**UDAs:**
  - User-defined aggregates are created to handle complex data structures.
  - They can be used to perform specific calculations on the data.

-**UDAs in Practice:**
  - UDAs are used to perform calculations on the data, such as calculating averages or performing statistical tests.

SS-DB – Standard Benchmark for Scientific Databases

We evaluate the performance of our system by comparing it to other systems in the literature. Our system consistently outperforms other systems in terms of speed and efficiency.

• **Benchmark:**
  - The benchmark consists of various data sets and queries.
  - The queries are designed to test the system in different scenarios.

• **Results:**
  - Our system outperforms other systems in terms of speed and efficiency.

Experimental Results

We evaluated our system on a single node with 2 AMD Opteron 8-core processors for a total of 16GB of memory, 4TB of disk space, and 4TB of network bandwidth. We used various data sets and queries to test the system in different scenarios.

-**Data Sets:**
  - The data sets used in the experiments are generated using various scientific applications.

-**Queries:**
  - The queries are designed to test the system in different scenarios.

-**Results:**
  - Our system outperforms other systems in terms of speed and efficiency.

Conclusion

Our system, EXTASCI, is a scalable and efficient system for scientific data analysis. It allows researchers to easily handle large volumes of data and to perform complex queries. The system is designed to handle various scientific applications and can be used in different scenarios.

References

