Scalable Analytics Over Unstructured Multidimensional Time-series Data

Sangmi Pallickara
Computer Science Department
Colorado State University
sangmi@cs.colostate.edu
http://www.cs.colostate.edu/~sangmi

Challenges in large scale data analysis

- Voluminous data
  - Data accumulates fast because of the rates at which they arrive
- Multidimensional data
  - Spatial and chronological components
  - Other remotely-sensed and in situ observations
- Support high throughput data management
  - Efficient storage and retrievals of data
  - Expressive query evaluations
- Data movement
  - Network bandwidths = limited resource
- Support low-latency data analysis

Sangmi Pallickara, sangmi@cs.colostate.edu http://www.cs.colostate.edu/~sangmi

Galileo: Distributed Data Storage

- The ability to manage billions of small files.
- Support for multiple scientific data formats such as netCDF and HDF4 and 5 and data from the Defense Meteorological Satellite Program
- A scale-out architecture that enables the incremental assimilation of nodes in the system.
- Harnessing geospatial characteristics in the data.
- Support for queries encompassing multi-dimensional data over variable sized records
- Evaluate range queries over dimensions representing spatial, chronological, and numeric attributes.
- Support for a tunable replication framework
- Sponsored by DHS long range program, Amazon AWS, and Hewlett Packard

Sangmi Pallickara, sangmi@cs.colostate.edu http://www.cs.colostate.edu/~sangmi

Data Analytics using Galileo 1/2

- Project GLEAN (http://glean.cs.colostate.edu)

- Support for analytic, approximate, and fuzzy queries
- Significance evaluation and hypothesis testing
- Support for radial, proximity, and geometry constraints
- Support for Allen's Interval Algebra in time-series queries
- Scalable anomaly detection and autonomous adaptation to evolving data

Sangmi Pallickara, sangmi@cs.colostate.edu http://www.cs.colostate.edu/~sangmi

Data Analytics using Galileo 2/2

- Support for density (EM and GMMs) and distance based clustering
- Forecasts based on multiple linear regression, artificial neural networks, random forest, and use of ensemble methods
- Conditional probability and Naïve Bayes Classifications

Sangmi Pallickara, sangmi@cs.colostate.edu http://www.cs.colostate.edu/~sangmi

Visual Analytics using Galileo

- Project geoLens (http://www.cs.colostate.edu/geolens)
- Perceptual Scalability
  - Controlling data that is displayed onscreen
- Interactive Scalability
  - Latency to support interactive application
- Providing interactive visual analytics over multi-terabyte datasets without data pre-processing
  - Brushing and linking
  - Dynamic histogram

Sangmi Pallickara, sangmi@cs.colostate.edu http://www.cs.colostate.edu/~sangmi

Thank you!

- http://www.cs.colostate.edu/~sangmi
- http://galileo.cs.colostate.edu
- http://glean.cs.colostate.edu
- http://www.cs.colostate.edu/geolens

Sangmi Pallickara, sangmi@cs.colostate.edu http://www.cs.colostate.edu/~sangmi