IBM PureData System for Analytics

UF Research Computing

23 September 2014

Purpose-built analytic appliance

▸ Integration
  ◦ Database
  ◦ Server
  ◦ Storage
  ◦ Advanced analytic capabilities
▸ Designed for rapid and deep analysis of data volumes

The Hardware

- Disk Enclosures
  - 48 Terabytes of raw storage split into three sections:
    - User data
    - Mirrored user data
    - Swap space
    - 4-5x Compression ratio
  - Provides high-speed data streaming
- Hosts
  - Provide administration, SQL interface, optimization platform, and query planning
- Snippet Blades (S-Blades)
  - Processor and streaming database logic
  - Eight S-Blades in system

How it works

- Data is loaded into the appliance
  - The data is distributed across all of the SPU's (Snippet Processing Unit)
- Each SPU analyzes the data it has access to and creates descriptive statistics of the data
- When a query is made:
  - Only SPU's that contain appropriate data return information
  - Joins also are done locally on the SPU, then distributed
- Result: Less movement of data across the network

In-Database Analytics

Transformations

Mathematical

Geospatial

Predictive

Statistics

Time Series

Data Mining

3rd Party In-Database Analytics

- SPSS
- SAS 9.3+

Software Development Kit

- User-Defined Extensions
  - UDF, UDA, UDTF, UDAP
- Language Support
  - MapReduce, Java, Python, R, Lua, Perl, C, C++, Fortran, PMML

New Users

▸ Access to the system will be handled on a case-by-case basis

▸ Please contact support@rc.ufl.edu