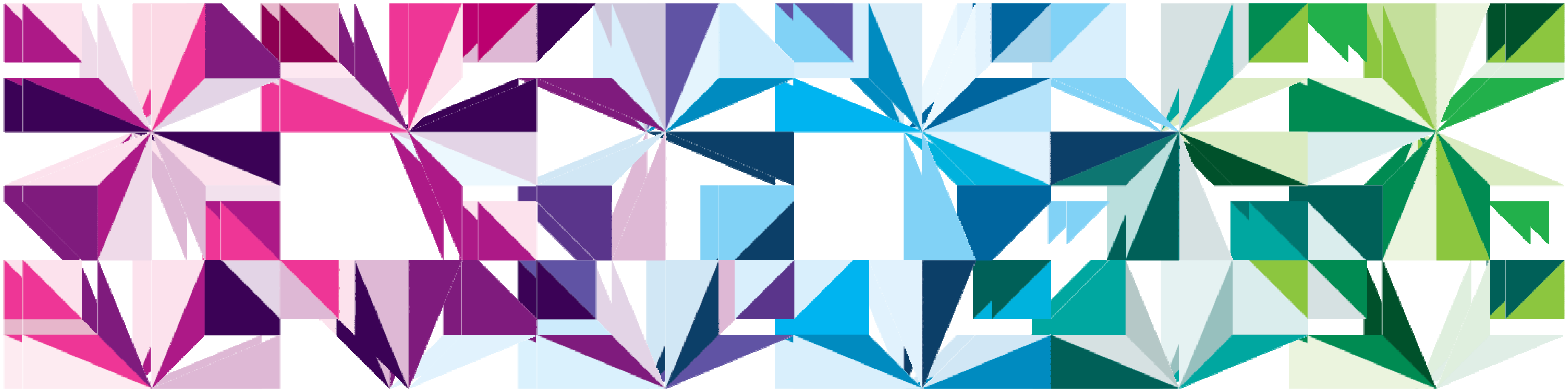
IBM® PureData™ System for Analytics A Technical Overview for University of Florida

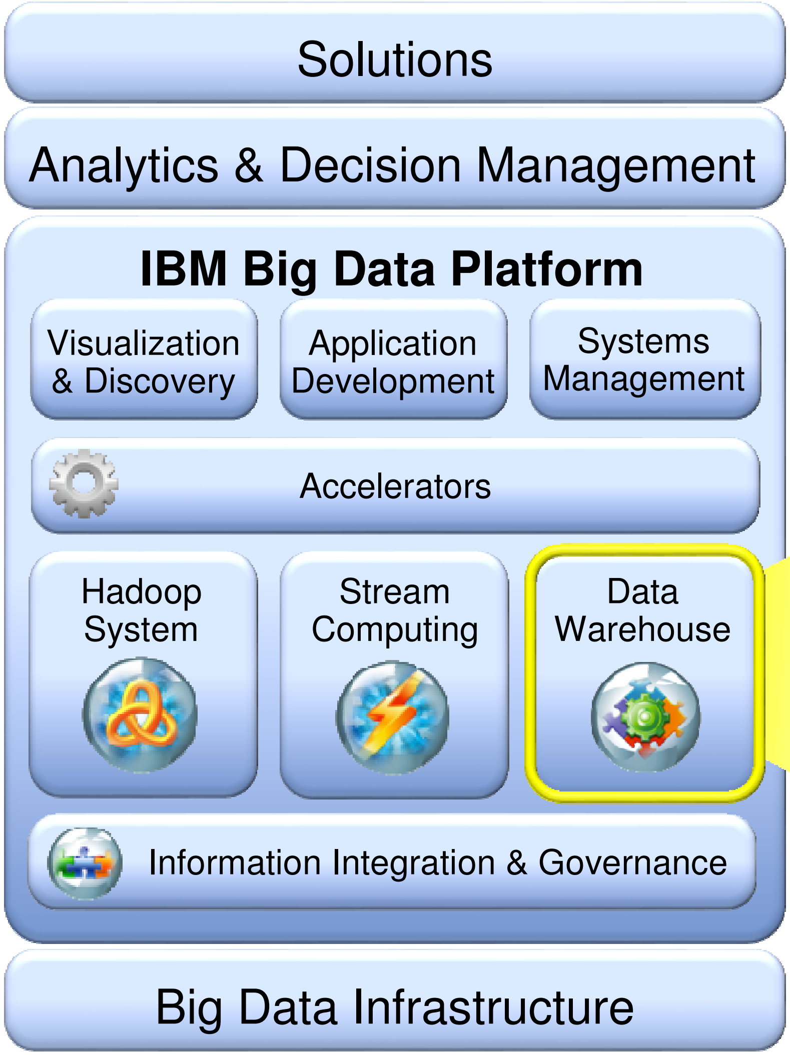
September 23, 2014



© 2013 IBM Corporation

Part of the IBM Big Data Platform

Workload Optimized Solutions for All Your Analytic Needs



Solutions

Analytics & Decision Management

IBM Big Data Platform

Visualization & Discovery

Application Development

Systems Management

Accelerators

Hadoop System

Stream Computing

Data Warehouse

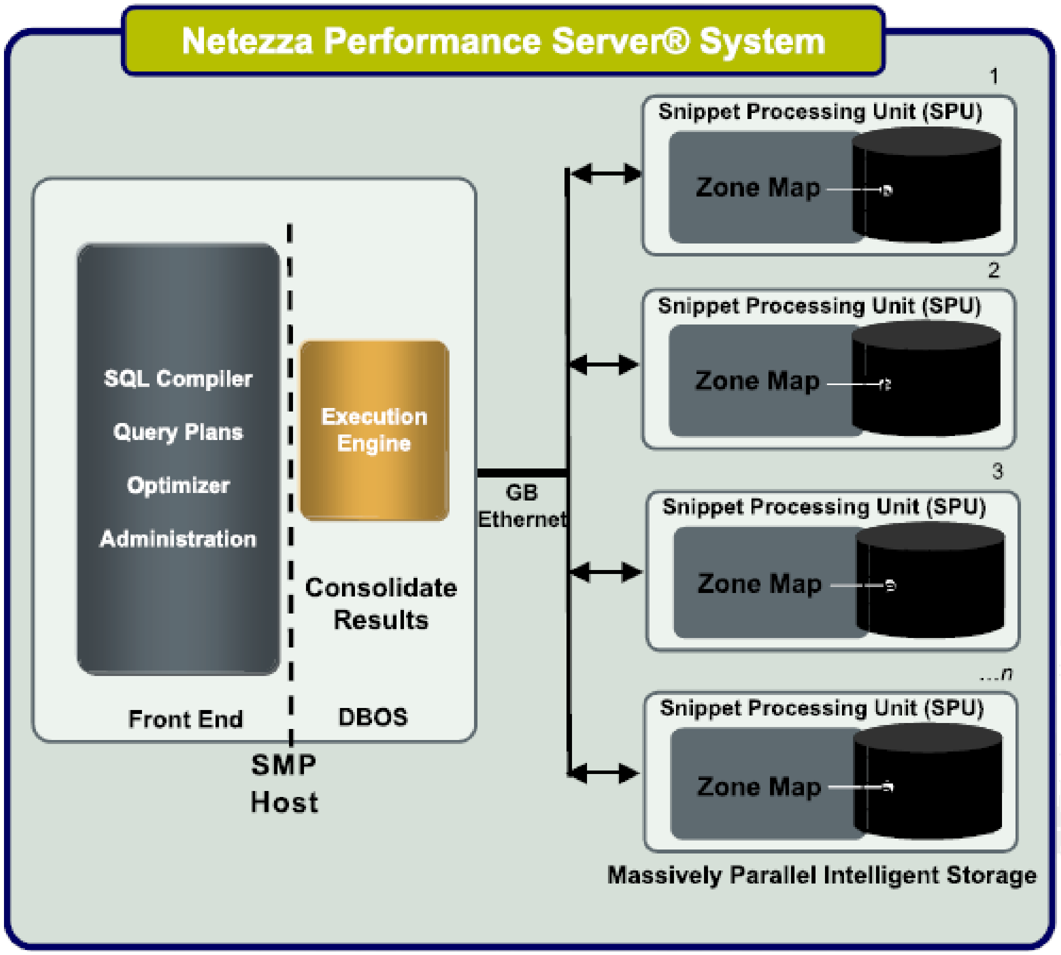
Information Integration & Governance

Big Data Infrastructure

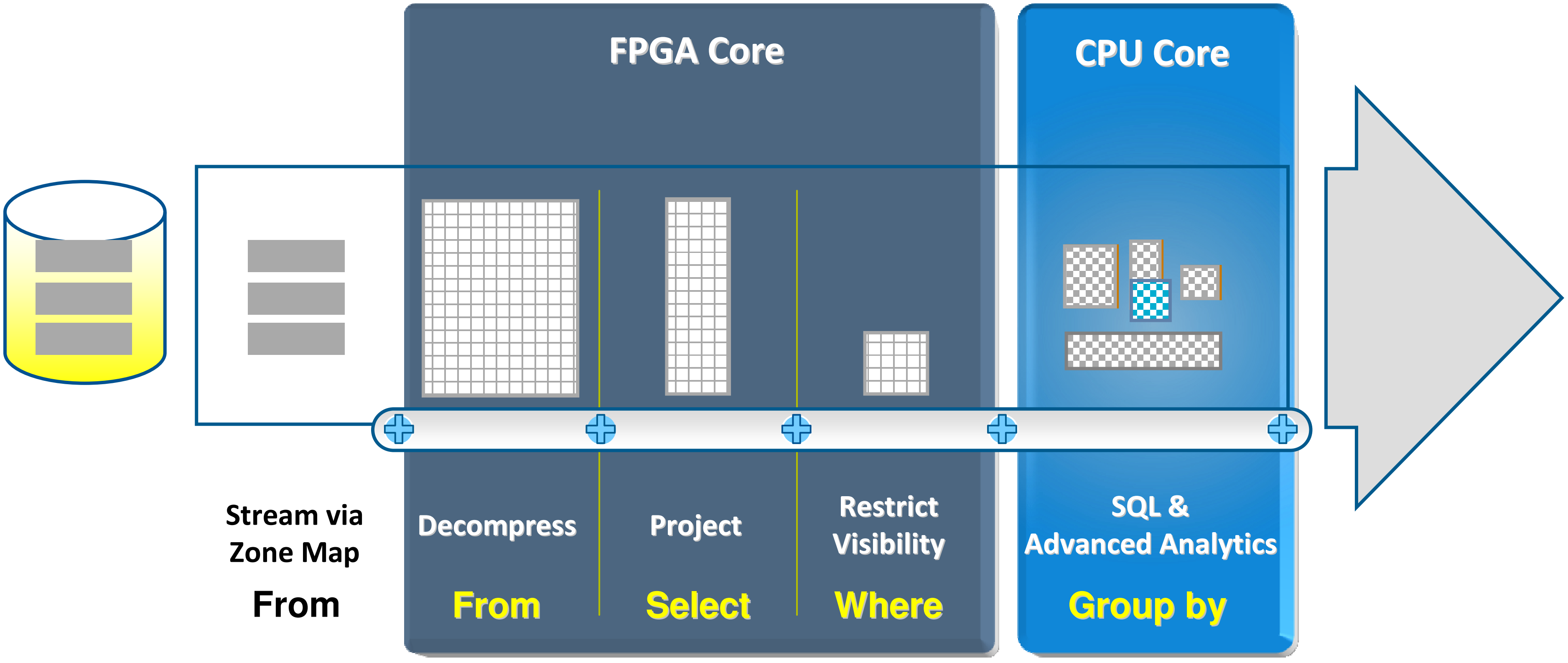
PureData System for Analytics

Page Level Zone Maps

- **An page is the smallest unit of disk allocation**
  - 128K of disk space
- **A zone map is an internal mapping structure to show the range (min and max) of values within each page**
- **During scans, zone maps are used to reduce IO by skipping pages that didn't qualify the query parameters**
- **Zone maps are internal to the system thus no administration involved**


Netezza Performance Server® System
SQL Compiler
Query Plans
Optimizer
Administration
Execution Engine
Consolidate Results
Front End
DBOS
SMP Host
GB Ethernet
1
Snippet Processing Unit (SPU)
Zone Map
2
Snippet Processing Unit (SPU)
Zone Map
3
Snippet Processing Unit (SPU)
Zone Map
...n
Snippet Processing Unit (SPU)
Zone Map
Massively Parallel Intelligent Storage

S-Blade Data Stream Processing



FPGA Core

CPU Core

Stream via Zone Map

From

Decompress

From

Project

Select

Restrict Visibility

Where

SQL & Advanced Analytics

Group by

Select State, Age, Gender, count(*) From MultiBillionRowCustomerTable Where BirthDate < '01/01/1960' **And State in ('FL', 'GA', 'SC', 'NC')** Group by State, Age, Gender Order by State, Age, Gender

Where BirthDate < '01/01/1960' Group by State, Age, Gender Order by State, Age, Gender

Loading the PureData System for Analytics

Data Integration

- IBM Information Server
- IBM InfoSphere Streams
- Oracle Data Integrator
- Oracle GoldenGate
- SAP Business Objects
- Composite Software
- IBM BigInsights

Data In

OLE-DB

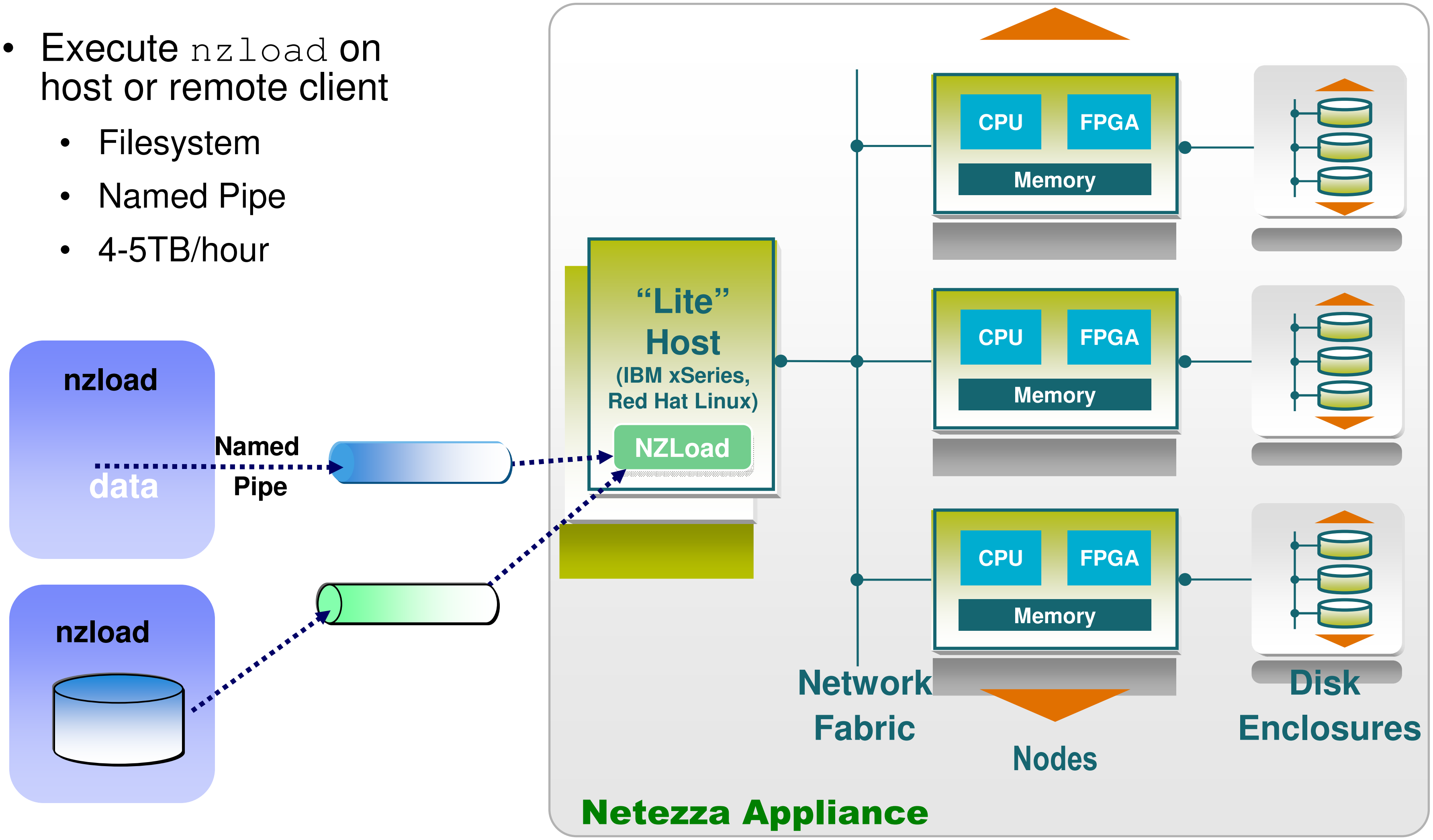
JDBC

ODBC

SQL

Netezza Data Load - nzload

- Execute `nzload` on host or remote client
  - Filesystem
  - Named Pipe
  - 4-5TB/hour



Querying the PureData System for Analytics

Reporting & Analysis

- IBM Cognos
- IBM SPSS
- IBM Unica
- Information Builders
- Kalido
- KXEN
- Microsoft Excel
- MicroStrategy
- Oracle OBIEE
- SAP Business Objects
- SAS
- Actuate

Data Out

OLE-DB

JDBC

ODBC

SQL

Workload Management

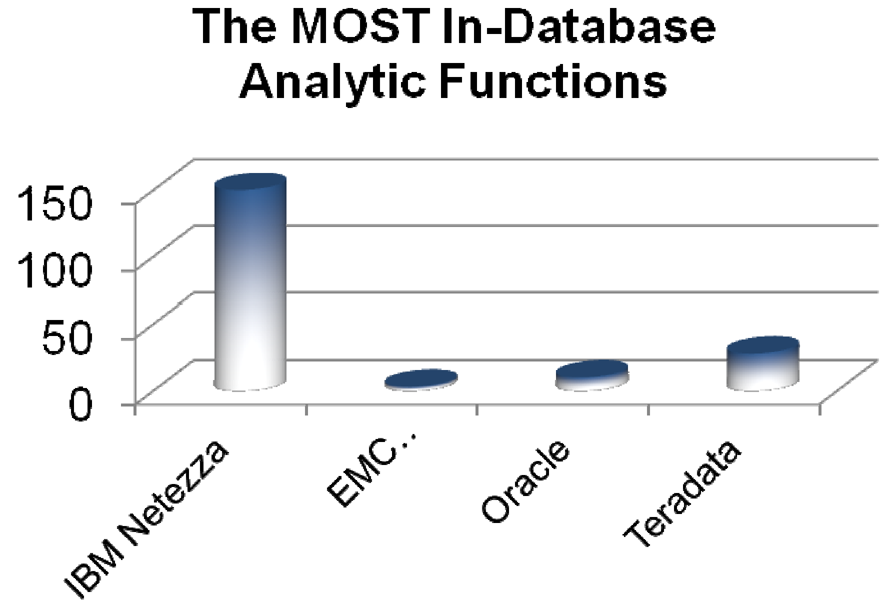
- **Workload Management (WLM)** provided optional functionality to manage resources and prioritize usage across a diverse multi-user environment to meet the need of mixed user workloads
- **Guaranteed Resource Allocation (GRA)**
  - Mechanism to **allocate system resources** among **groups** of **users** in a multi-user environment
- **Short Query Bias (SQB)**
  - Ensures users with **short queries** receive **faster**, **higher**, biased query **response time** under heavy system workloads
- **Prioritized Query Execution (PQE)**
  - Finer control over resource allocation by extending the notion of query priorities from **scheduling** to **execution.**
  - CRITICAL, HIGH, NORMAL, LOW
  - Can be set at the system, group, user or session level.

Integrated by Design

IBM Netezza In-Database Analytics Version 2.0

Netezza In-Database Analytics

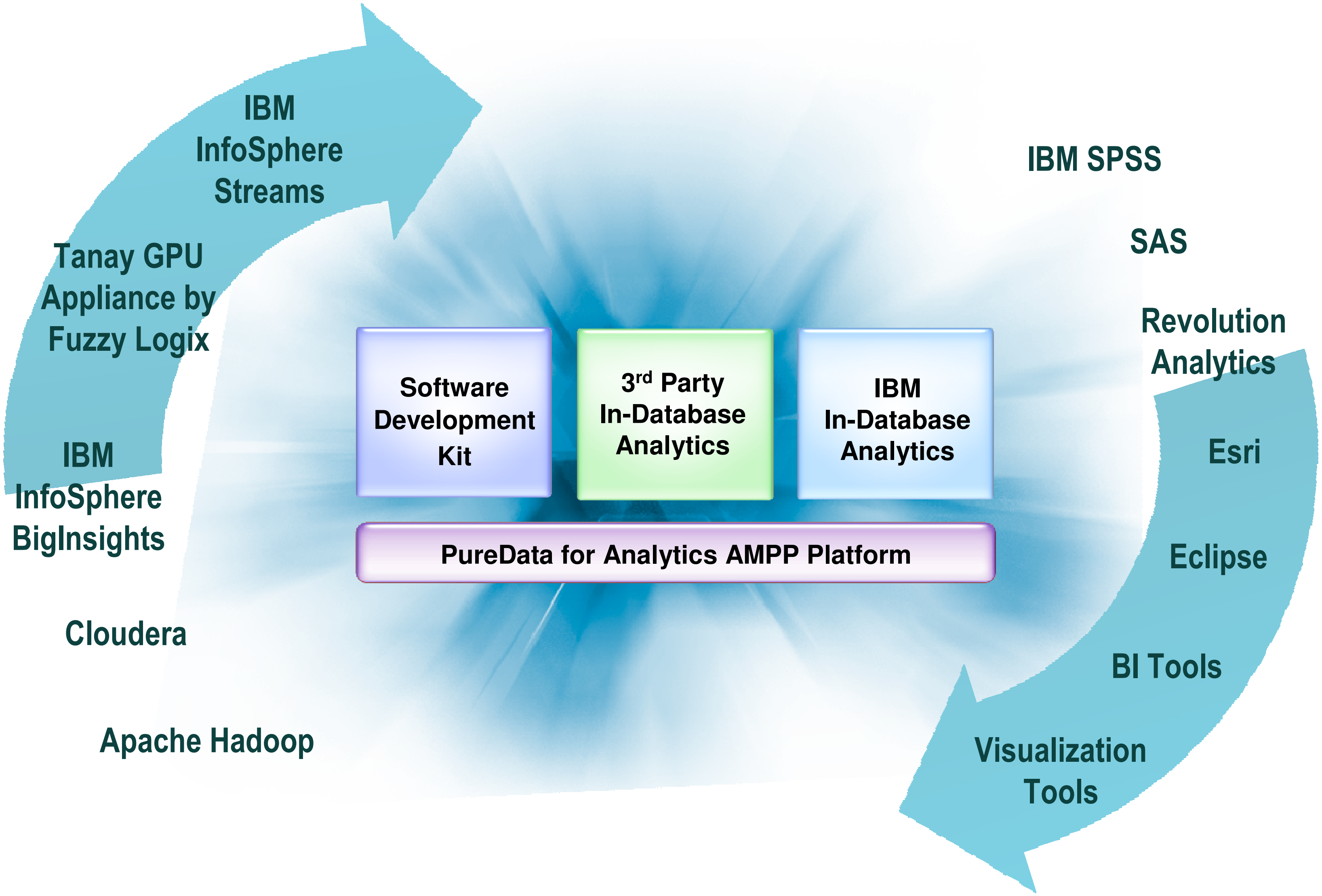
- Transformations
- Mathematical
- Geospatial
- Predictive
- Statistics
- Time Series
- Data Mining



The MOST In-Database Analytic Functions

- No data movement
- Analyze deep and wide data
- High performance, parallel computation

IBM Netezza Analytics: Built-In Features and Capabilities



IBM InfoSphere Streams

Tanay GPU Appliance by Fuzzy Logix

IBM InfoSphere BigInsights

Cloudera

Apache Hadoop

Software Development Kit

3rd Party In-Database Analytics

IBM In-Database Analytics

PureData for Analytics AMPP Platform

IBM SPSS

SAS

Revolution Analytics

Esri

Eclipse

BI Tools

Visualization Tools

SQL Analytics

- **What is the purpose?**
  - – Advanced uses of straight SQL leads to complex, powerful applications
- **What are we talking about?**
  - – SQL - Traditional SQL based analytics processing
  - – Stored Procedures - A registered set of SQL statements
- **How is this used?**
  - – SQL
    - Business Intelligence
    - “Monday Morning Reports”
  - – Stored Procedures
    - Many data mining tasks require multiple SQL statements E.g., K-means clustering
- **What is the API?**
  - – SQL
  - – Stored Procedures - NZPLSQL language, Registered with the database

User-Defined Extensions

- **What is the purpose?**
  - Extend SQL to enable needed features
- **What are we talking about?**
  - User-Defined Functions (UDFs)
    - Scoring, Data transformation, Data cleanliness
  - User-Defined Aggregates (UDAs)
    - Scoring, Domain-specific aggregates, Time-series analysis
  - User-Defined Table Functions (UDTFs)
    - Custom summaries/Windowed aggregates, “Unpivot” operations, Unstructured data parsing
- **What is the API?**
  - C/C++ classes
  - Run lock-step with the database
  - Called via SQL

SQL Analytics

- **What are some examples?**
  - SQL allows for many simple (and complex) analyses
    - Average stock price over last 6 months
    - Maximum (minimum) stock price
    - **`SELECT symbol, AVG(closing_price) FROM stocks GROUP BY symbol;`**
  - Both data movement and flow and analytic functions
    - GROUP BY, OVER(…), etc
    - MIN, MAX, STDDEV, RANK, etc
  - Stored procedures are called via SQL
    - **`SELECT kmeans(‘input=inTable, output=outTable, k=5, iter=10, id=idCol, distance=EuclideanUdf’);`**
    - **`Or: CALL kmeans(...); EXEC kmeans(...);`**

Questions and Answers

International Business Machines Corporation New Orchard Road Armonk, NY 10504