PureSystems: Changing The Economics And Experience Of IT

Accelerating Analytics – Faster Insight From Data Warehouses That Scale And Cost Less

Businesses Benefit By Using An Analytic Approach Over Intuition

600% increase in cross-sell campaign

$200 Million increase in cash flow

$13.8 Million in cost savings

80% decrease in reporting time on top of Oracle e-business suite

40% decline in homicide rates

The more analytics a business uses, the better it performs
Analytics Have Different Characteristics

*Different characteristics require different Workload Optimized Systems*

**Deep Analytics**
- Fewer users, more complex reports
- Emphasis on response time
- Efficiently applies all resources to a single task
- Dedicated Static Data Marts

**Operational Analytics**
- Many users, all types of reports
- Emphasis on throughput
- Effectively balances resources across all tasks
- Real Time Data Flow & Enterprise Data Warehouse
Optimizing both data processing and data access techniques

1. **Transaction Processing**
   PureData System For Transactions
   Many *data transactions* running in parallel for high throughput
   Operational Data

2. **Deep Analytics**
   PureData System For Analytics
   Very complex *data analysis* simplified and accelerated through massively parallel processing and SMP and hardware acceleration
   Dedicated Static Data Marts

3. **Operational Analytics**
   PureData System For Operational Analytics
   A mix of *data analysis* accelerated through massively parallel processing *and* many operational interactions running in parallel for high throughput
   Live Feeds and Data Warehouse
Built In Expertise Makes Deep Analytics As Simple As An Appliance

*PureData System for Analytics Is An Efficient Dedicated Appliance*

- **Simple**
  - Fast installation
  - No DB administrator tasks
  - Very easy operation

- **Speed**
  - Optimized for purpose

- **Smart**
PureData System for Analytics Is Workload Optimized For High Performance Analytics

Workload Optimized
- Massively parallel processing architecture
- Custom hardware acceleration
- Purpose-built for high performance analytics

Disk Enclosures
- User data, mirror, swap partitions
- High speed data streaming

SMP Hosts
- SQL Compiler
- Query Planner
- Optimizer
- Administration

Snippet Blades
- Hardware-based query acceleration with Field Programmable Gate Arrays
- Complex analytics executed as the data streams from disk
Innovative Streaming Technology Speeds Up Complex Queries

```
select DISTRICT, PRODUCTGRP, sum(NRX)
from MTHLY_RX_TERR_DATA
where MONTH = '20091201'
and MARKET = 509123
and SPECIALTY = 'GASTRO'
```

Slice of table MTHLY_RX_TERR_DATA (compressed)
PureData System For Analytics Provides Innovative Streaming Technology Breakthrough Simplicity

- No software installation
- No indexes and tuning
- No storage administration
  - No dbspace / tablespace sizing and configuration
  - No redo/physical/Logical log sizing and configuration
  - No page/block sizing and configuration for tables
  - No extent sizing and configuration for tables
  - No Temp space allocation and monitoring
  - No RAID level decisions for dbspaces
  - No logical volume creations of files
  - No integration of OS kernel recommendations
  - No maintenance of OS recommended patch levels
  - No JAD sessions to configure host/network/storage

Spend Less Time Managing and More Time Innovating

UP and Running:

PureData for Analytics
24 hours!

vs.

Exadata
2.5 weeks
PureData System For Analytics Is Simple To Set Up

- Create Database (nzsql commands)
- Create Tables (nzsql commands)
  - CREATE TABLE SALES_FACT (...)
    - distribute on random
    - organize on (order_day_key, retailer_site_key, product_key)
- Load Table (nzload)
- Groom tables (nzsql)
- Gather statistics (nzsql)
- Run test
- Done in 2.5 days – which includes 1 day of learning!
Exadata Requires Extensive Knowledge, Complex Configuration, And Time Consuming Tuning

- Create database using X3 data warehouse database template (dbca wizard)
- Create and configure multiple tablespaces and datafiles for non-partitioned and partitioned tables, DBFS staging filesystem, temp tables
- Configure ASM storage for datafile performance (Hot/Cold disk region settings)
- Create Tables

```sql
CREATE TABLE SALES_FACT (…)
PARTITION BY RANGE(ORDER_DAY_KEY)
SUBPARTITION BY HASH(SALES_ORDER_KEY, ORDER_DAY_KEY)
SUBPARTITIONS 32 STORE IN (BIDAYTS, BIDAYTS_1, BIDAYTS_2, BIDAYTS_3) (PARTITION ORDERS_P1 VALUES LESS THAN (20060501), PARTITION ORDERS_P2 VALUES LESS THAN (20060601), PARTITION ORDERS_P3 VALUES LESS THAN (20060701), PARTITION ORDERS_P4 VALUES LESS THAN (20060801), PARTITION ORDERS_P5 VALUES LESS THAN (20060901), PARTITION ORDERS_P6 VALUES LESS THAN (20061001), PARTITION ORDERS_P7 VALUES LESS THAN (20061101), PARTITION ORDERS_P8 VALUES LESS THAN (20061201), PARTITION ORDERS_P9 VALUES LESS THAN (MAXVALUE))
COMPRESS FOR QUERY HIGH STORAGE (CELL_FLASH_CACHE KEEP) STORAGE (INITIAL 8M NEXT 8M) PARALLEL;
```

- Prior to loading, run all queries on empty tables to populate column usage statistics
- Create DBFS filesystem, copy raw load files into DBFS, Load tables
- Gather statistics on tables, indexes (dimension PKs), fixed objects, dictionary,…
- Configure SGA/PGA memory sizes (Try to maximize without starving O/S)
- Run tests, **performance degraded** for some queries compared to V2
  - Access path (plan) changes in 11.2.0.3 causing less efficient plans to be used
  - Modify optimizer OFE from 11.2.0.3.1 to 11.2.0.2 to obtain better results
- **Done in 2.5 Weeks – by an experienced Exadata administrator**
BI Day Deep Analytics Test Measures Concurrent Multi-User Report Performance

4 Users doing **complex** reports

- 2 Users
- 2 Users

20 Users doing **intermediate** Reports

- 6 Users
- 8 Users
- 6 Users

Each report executes one or more queries

- 4 Connections
- 20 Connections

- 24 concurrent users
- Mix of complex and intermediate reports
- Focus on throughput

Note: Distribution of complex, intermediate, and simple workloads based on Forrester Research, Profiling the Analytic End User for Business Intelligence, 2004

05 PureData Analytics
PureData For Analytics Delivers 10x Better Price Performance On Deep Analytics Workload

IBM PureData for Analytics Full Rack (N2001)

System Cost
3 year TCA
$2.91 Million

1298 Intermediate Reports/Hour
19 Complex Reports/Hour

✔ 10x better price/performance
✔ 9x more intermediate reports
✔ 8x more complex reports

Based on IBM internal tests comparing PureData System For Analytics with a comparably priced, comparably tuned competitor configuration (version available as of 01/01/2013) executing an identical analytics workload in a controlled laboratory environment. Tests measure report throughput (reports per hour) over a 4 hour test period with 24 connected users concurrently executing mixed SQL query workload. More reports per hour indicates higher throughput. 3YR TCA means Total Cost of Acquisition for 3 years, based on list prices, including hardware, software, and maintenance. Competitor configuration: ¼ Unit (usable uncompressed capacity = 9.5TB) including competitor recommended software options and features. IBM configuration: PureData System for Analytics N2001-10 (usable uncompressed capacity = 46TB). Price performance calculated as 3YR TCA divided by test throughput. Results may not be typical and will vary based on actual workload, configuration, applications, queries and other variables in a production environment. Users of this document should verify the applicable data for their specific environment. Contact IBM and see what we can do for you.
Cores, SSDs And Infiniband Cannot Overcome Competitor’s Weaknesses

- Competitor database allocates resources on request arrival
  - If workload is light, lots of resource allocated
    - Request executes quickly
  - If workload is heavy, few resources allocated
    - Request executes very slowly

  **Result: Unpredictable performance**

- Competitor cannot reallocate resources as workload changes
  - If workload lightens and resources become available
    - In-flight requests continue with resources originally assigned
    - Available resources sit idle
    - In-flight requests execute very slowly

  **Result: Wasted resource and needless delay**
Netezza Outperforms Competitor Running Analytic Queries

Same queries, same clients, same data…
Different Results

See more like this on YouTube:
http://www.youtube.com/watch?v=T3O6yJ hdUU
Many Customers Have Had Similar Results With IBM PureData System For Analytics

<table>
<thead>
<tr>
<th>Digital Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACXIOM</td>
</tr>
<tr>
<td>Aol</td>
</tr>
<tr>
<td>bluekai</td>
</tr>
<tr>
<td>CBS Interactive</td>
</tr>
<tr>
<td>epsilon</td>
</tr>
<tr>
<td>Kelley Blue Book</td>
</tr>
<tr>
<td>MediaMath</td>
</tr>
<tr>
<td>Nielsen</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Financial Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>FICO</td>
</tr>
<tr>
<td>HBOS plc</td>
</tr>
<tr>
<td>ICE</td>
</tr>
<tr>
<td>KCB</td>
</tr>
<tr>
<td>Loan Performance</td>
</tr>
<tr>
<td>M&amp;I</td>
</tr>
<tr>
<td>NYSE Euronext</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Government</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lawrence Livermore National Lab</td>
</tr>
<tr>
<td>NEX</td>
</tr>
<tr>
<td>NSF</td>
</tr>
<tr>
<td>Pacific Northwest National Lab</td>
</tr>
<tr>
<td>Sandia National Laboratories</td>
</tr>
<tr>
<td>US Army Corps of Engineers</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Health &amp; Life Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Massachusetts</td>
</tr>
<tr>
<td>INGENIX</td>
</tr>
<tr>
<td>Marshfield Clinic</td>
</tr>
<tr>
<td>Premier</td>
</tr>
<tr>
<td>Shire</td>
</tr>
<tr>
<td>Wolters Kluwer Health</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Retail / Consumer Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahold</td>
</tr>
<tr>
<td>Burlington</td>
</tr>
<tr>
<td>CATALINA MARKETING</td>
</tr>
<tr>
<td>Guitar Center</td>
</tr>
<tr>
<td>Neiman Marcus</td>
</tr>
<tr>
<td>PACSUN</td>
</tr>
<tr>
<td>ROSS</td>
</tr>
<tr>
<td>SAPPORO</td>
</tr>
<tr>
<td>Yum</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Telecom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carphone Warehouse</td>
</tr>
<tr>
<td>Basis</td>
</tr>
<tr>
<td>docomo</td>
</tr>
<tr>
<td>orange</td>
</tr>
<tr>
<td>Virgin media</td>
</tr>
<tr>
<td>WIND</td>
</tr>
<tr>
<td>Communications</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Con-way</td>
</tr>
<tr>
<td>Marriott</td>
</tr>
<tr>
<td>Midwest ISO</td>
</tr>
<tr>
<td>OLYMPUS</td>
</tr>
<tr>
<td>Ryder</td>
</tr>
</tbody>
</table>
IBM PureData – Optimized For Data Workloads

Optimizing both data processing and data access techniques

1. **Transaction Processing**
   
   **PureData System For Transactions**
   
   Many *data transactions* running in parallel for high throughput
   
   Operational Data

2. **Deep Analytics**
   
   **PureData System For Analytics**
   
   Very complex *data analysis* simplified and accelerated through massively parallel processing and SMP and hardware acceleration
   
   Dedicated Static Data Marts

3. **Operational Analytics**
   
   **PureData System For Operational Analytics**
   
   A mix of *data analysis* accelerated through massively parallel processing *and* many operational interactions running in parallel for high throughput
   
   Live Feeds and Data Warehouse
Operational Analytics Mixes Analytic Queries With Operational Transactions

- Operational Analytic workload characteristics

  - Mix of deep analytic queries and short operational transactions
  
  - Continuous updates
    - Multiple ingest streams
  
  - Concurrency
    - Thousands of users
    - May incorporate normalized schemas
  
- Balance long running analytics with real time operational needs

- Receive data from a variety of sources

- Work with a wide variety of analytical tools

Real Time Fraud Detection

When your credit card company tells you that it denied a transaction with your card today – on the other side of the world – you probably are not thinking about all the transactions checked in just that second…

But the credit provider’s operational data warehouse is

Operational Analytics

Random and sequential reads & data loads + continuous ingest

Analytics split into many parts and narrow scope operations, all running in parallel

Partitioned data access
PureData System For Operational Analytics Can Collect And Analyze All Your Data Using The Tools That Fit Your Needs

- Ab Initio
- Cloudera
- Composite Software
- IBM Big Insights
- IBM Information Server
- IBM InfoSphere Streams
- IBM InfoSphere Change Data Capture
- Informatica
- Oracle Data Integrator
- Oracle GoldenGate
- SAP Business Objects

Reporting and Analysis

- IBM Cognos
- IBM SPSS
- IBM Unica
- Information Builders
- Kalido
- KXEN
- Microsoft Excel
- MicroStrategy
- Oracle OBIEE
- SAP Business Objects
- SAS
- Actuate
PureData System For Operational Analytics – Complete And Ready To Run

- **Hardware**
  - Power Systems servers
  - AIX v7.1
  - Storwize V7000 storage
  - EXP30 Ultra SSD
- **Software**
  - DB2
  - InfoSphere Warehouse
  - Tivoli Automation*
  - Optim Performance Manager
- **Analytics**
  - Cognos 10.1.1

* For Failover Orchestration
# A Size For Every Business

## IBM PureData System for Operational Analytics configurations*

<table>
<thead>
<tr>
<th></th>
<th>Extra Small</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Foundation rack 1 foundation module</td>
<td>Foundation + 1/3 rack 1 foundation module 1 data module</td>
<td>Foundation + 2/3 rack 1 foundation module 2 data modules</td>
<td>Foundation + full rack 1 foundation module 3 data modules</td>
</tr>
<tr>
<td>Cores</td>
<td>32</td>
<td>64</td>
<td>80</td>
<td>96</td>
</tr>
<tr>
<td>Memory</td>
<td>256 GB</td>
<td>512 GB</td>
<td>640 GB</td>
<td>768 GB</td>
</tr>
<tr>
<td>SSD Storage</td>
<td>4.8 TB</td>
<td>9.6TB</td>
<td>12 TB</td>
<td>14.4 TB</td>
</tr>
<tr>
<td>HDD Capacity (raw unformatted)</td>
<td>64.8 TB</td>
<td>151.2 TB</td>
<td>237.6 TB</td>
<td>234 TB</td>
</tr>
<tr>
<td>User Data Capacity (uncompressed)</td>
<td>29.7 TB</td>
<td>69.3 TB</td>
<td>108.9 TB</td>
<td>148.5 TB</td>
</tr>
</tbody>
</table>

*Scalable to 1 PB or more of user data
PureData System For Operational Analytics Has A Simple Integrated Single Management Console

- Single, integrated graphical console to manage all resources and work running on the system

- Role-based security and tasks
  - management
  - monitoring
  - maintenance

- Easy integration with broader enterprise monitoring tools and processes

- Same user interface as IBM PureApplication System for consistency across systems
BI Day Workload Measures High Levels Of Concurrently Executing Workloads

4 Users doing complex reports
- 2 Users doing complex report 1
- 2 Users doing complex report 3

20 Users doing intermediate Reports
- 6 Users doing intermediate report 9
- 8 Users doing intermediate report 10
- 6 Users doing intermediate report 11

56 Users doing simple reports
- 14 Users doing simple report 2
- 14 Users doing simple report 4
- 14 Users doing simple report 5
- 14 Users doing simple report 6

Each report executes one or more queries

Data Server

- 80 simultaneous users connected
- Measure concurrent throughput

Note: Distribution of complex, intermediate, and simple workloads based on Forrester Research, Profiling the Analytic End User for Business Intelligence, 2004
PureData System For Operational Analytics Delivers Better Price Performance On Operational Analytics Workload

IBM PureData System for Operational Analytics (Small Rack)

- 1 Foundation Node
- 1 Data Node
- System Cost
- 3 year TCA
- $2.98 Million

580,248 Simple Reports/Hour
532 Intermediate Reports/Hour
4.85 Complex Reports/Hour

✓ 17% lower system cost for more reports!
✓ 3.8x more intermediate reports
✓ 1.8x more complex reports
✓ 1.2x more simple reports

Based on IBM internal tests comparing PureData System For Operational Analytics with a comparably priced, comparably tuned competitor configuration (version available as of 01/01/2013) executing an identical analytics workload in a controlled laboratory environment. Tests measure report throughput (reports per hour) over a 4 hour test period with 80 connected users concurrently executing mixed SQL query workload. More reports per hour indicates higher throughput. 3YR TCA means Total Cost of Acquisition for 3 years, based on list prices, including hardware, software, and maintenance. Competitor configuration: ½ Unit (usable uncompressed capacity = 9.5TB) including competitor recommended software options and features. IBM configuration: PureData System for Operational Analytics A1791-02 (usable uncompressed capacity = 18.6TB). Price performance calculated as 3YR TCA divided by test throughput. Results may not be typical and will vary based on actual workload, configuration, applications, queries and other variables in a production environment. Users of this document should verify the applicable data for their specific environment. Contact IBM and see what we can do for you.
Summary - IBM Delivers Insight Faster And At A Lower Cost

For both Deep and Operational Analytics…

- Faster time to solution
- Faster performance
- Lower costs
- A broader spectrum of systems to fit every budget