Performance and Scalability Overview

This guide provides an overview of some of the performance and scalability capabilities of the Pentaho Business Analytics platform

PENTAHO PERFORMANCE ENGINEERING TEAM
Pentaho Scalability and High-Performance Architecture

Business Analytics solutions are only valuable when they can be accessed and used by anyone, from anywhere and at any time. When selecting a business analytics platform, it is critical to assess the underlying architecture of the platform to ensure that it not only scales to the number of users and amount of data organizations have today, but supports growing numbers of users and increased data sizes into the future.

By tightly coupling high-performance business intelligence with data integration in a single platform, Pentaho Business Analytics provides a scalable solution that can address enterprise requirements in organizations of all sizes. This guide provides an overview for just some of the performance tuning and scalability options available.

Pentaho Business Analytics Server is a Web application for creating, accessing and sharing reports, analysis and dashboards. The Pentaho Business Analytics Server can be deployed in different configurations, from a single server node, to a cluster of nodes distributed across multiple servers. There are a number of ways to increase performance and scalability:

- Deployment on 64-bit operating systems
- Clustering multiple server nodes
- Optimizing the configuration of the Reporting and Analysis engines

Pentaho Business Analytics Server
Deployment on 64-bit Operating Systems

The Pentaho Business Analytics Server supports 64-bit operating systems for larger amounts of server memory and vertical scalability for higher user and data volumes on a single server.

The Pentaho Business Analytics Server can effectively scale out to a cluster, or further to a cloud environment. Clusters are excellent for permanently expanding resources commensurate with increasing load; cloud computing is particularly useful if scaling out is only needed for specific periods of increased activity.

Optimizing the Configuration of the Reporting and Analysis Engines

Pentaho Reporting

The Pentaho Reporting engine enables the retrieval, formatting and processing of information from a data source, to generate user-readable output. One example for increasing the performance and scalability of the Pentaho Reporting solutions is to take advantage of result set caching. When rendered, a parameterized report must account for every dataset required for every parameter. Every time a parameter field changes, every dataset is recalculated. This can negatively impact performance. Caching parameterized report result sets creates improved performance for larger datasets.

Pentaho Analysis

The Pentaho Analysis engine (Mondrian) creates an analysis schema, and forms data sets from that schema by using an MDX query. Maximizing performance and scalability always begins with the proper design and tuning of source data. Once the database has been optimized, there are some additional areas within the Pentaho Analysis engine that can be tuned.

IN-MEMORY CACHING CAPABILITIES

Pentaho’s in-memory caching capability enables ad hoc analysis of millions of rows of data in seconds. Pentaho’s pluggable, in-memory architecture is integrated with popular open source caching platforms such as Infinispan and Memcached and is used by many of the world’s most popular social, ecommerce and multi-media websites.
In addition, Pentaho allows in-memory aggregation of data – where granular data can be rolled-up to higher-level summaries entirely in-memory, reducing the need to send new queries to the database. This will result in even faster performance for more complex analytic queries.

AGGREGATE TABLE SUPPORT
When working with large data sets, properly creating and using aggregate tables greatly improves performance. An aggregate table coexists with the base fact table, and contains pre-aggregated measures built from the fact table. Registered in the schema Pentaho Analysis can choose to use an aggregate table rather than the fact table, resulting in faster query performance.

“We have operational metrics for six different businesses running in each of our senior care facilities that need to be retrieved and accessed everyday by our corporate management, the individual facilities managers, as well as the line of business managers in a matter of seconds.

Now, with the high performance in-memory analysis capabilities in the latest release of Pentaho Business Analytics, we can be more aggressive in rollouts – adds more metrics to dashboards, giving dashboards and data analysis capabilities to more users, and see greater usage rates and more adoption of business analytics solutions.”

– BRANDON JACKSON, DIR. OF ANALYTICS AND FINANCE, STONEGATE SENIOR LIVING LLC.
PARTITIONING SUPPORT FOR HIGH CARDINALITY DIMENSIONALITY
Large, enterprise data warehouse deployments often contain attributes comprised of tens or hundreds of thousands of unique members. For these use cases, the Pentaho Analysis engine can be configured to properly address a (partitioned) high-cardinality dimension. This will streamline SQL generation for partitioned tables; ultimately, only the relevant partitions will be queried, which can greatly increase query performance.

Pentaho Data Integration
Pentaho Data Integration (PDI) is an extract, transform, and load (ETL) solution that uses an innovative metadata-driven approach. It includes an easy to use, graphical design environment for building ETL jobs and transformations, resulting in faster development, lower maintenance costs, interactive debugging, and simplified deployment. PDI’s multi-threaded, scale-out architecture provides performance tuning and scalability options for handling even the most demanding ETL workloads.

MULTI-THREADED ARCHITECTURE
PDI’s streaming engine architecture provides the ability to work with extremely large data volumes, and provides enterprise-class performance and scalability with a broad range of deployment options including dedicated, clustered, and/or cloud-based ETL servers.

The architecture allows both vertical and horizontal scaling. The engine executes tasks in parallel and across multiple CPUs on a single machine as well as across multiple servers via clustering and partitioning.

TRANSFORMATION PROCESSING ENGINE
Pentaho Data Integration’s transformation processing engine starts and executes all steps within a transformation in parallel (multi-threaded) allowing maximum usage of available CPU resources. Done by default this allows processing of an unlimited number of rows and columns in a streaming fashion. Furthermore, the engine is 100% metadata driven (no code generation) resulting in reduced deployment complexity. PDI also provides different processing engines that can be used to influence thread priority or limit execution to a single thread which is useful for parallel performance tuning of large transformations.

Additional tuning options include the ability to configure streaming buffer sizes, reduce internal data type conversions (lazy conversion), leverage high performance non-blocking I/O (NIO) for read large blocks at a time and parallel reading of files, and support for multiple step copies to allowing optimization of Java Virtual Machine multi-thread usage.
CLUSTERING AND PARTITIONING

Pentaho Data Integration provides advanced clustering and partitioning capabilities that allow organizations to scale out their data integration deployments. Pentaho Data Integration clusters are built for increasing performance and throughput of data transformations; in particular they are built to perform classic “divide and conquer” processing of data sets in parallel.

PDI clusters have a strong master/slave topology. There is one master in cluster but there can be many slaves. This cluster scheme can be used to distribute the ETL workload in parallel appropriately across these multiple systems. Transformations are broken into master/slaves topology and deployed to all servers in a cluster – where each server in the cluster is running a PDI engine to listen, receive, execute and monitor transformations.

It is also possible to define dynamic clusters where the Slave servers are only known at run-time. This is very useful in cloud computing scenarios where hosts are added or removed at will. More information on this topic including load statistics can be found in an independent consulting white paper created by Nick Goodman from Bayon Technologies, “Scaling Out Large Data Volume Processing in the Cloud or on Premise.”
**EXECUTING IN HADOOP (PENTAHO MAPREDUCE)**
Pentaho’s Java-based data integration engine integrates with the Hadoop cache for automatic deployment as a MapReduce task across every data node in a Hadoop cluster, leveraging the use of the massively parallel processing and high availability of Hadoop.

**NATIVE SUPPORT FOR BIG DATA SOURCES INCLUDING HADOOP, NOSQL AND HIGH-PERFORMANCE ANALYTICAL DATABASES**
Pentaho supports native access, bulk-loading and querying of a large number of databases including:

- NoSQL data sources such as:
  - MongoDB
  - Cassandra
  - HBase
  - HPCC Systems
  - ElasticSearch

- Analytic databases such as:
  - HP Vertica
  - EMC Greenplum
  - HP NonStop SQL/MX
  - IBM Netezza
  - Infobright
  - Actian Vectorwise
  - LucidDB
  - MonetDB
  - Teradata

- Transactional databases such as:
  - MySQL
  - Postgres
  - Oracle
  - DB2
  - SQL Server
  - Teradata
## Customer Examples and Use Cases

<table>
<thead>
<tr>
<th>INDUSTRY</th>
<th>USE CASE</th>
<th>DATA VOLUME AND TYPE</th>
<th># USERS (TOTAL)</th>
<th># USERS (CONCURRENT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retail</td>
<td>Store Operations Dashboard</td>
<td>5+ TB HP Neoview</td>
<td>1200</td>
<td>200</td>
</tr>
<tr>
<td>Telecom (B2C)</td>
<td>Customer Value Analysis</td>
<td>2+ TB in Greenplum</td>
<td>&lt;500</td>
<td>&lt;25</td>
</tr>
<tr>
<td>Social Networking</td>
<td>Website Activity Analysis</td>
<td>1 TB in Vectorwise 10+ TB in a 20-node Hadoop cluster Loading 200,000 rows per second 20 billion chat logs per month 240 million user profiles</td>
<td>Social Networking</td>
<td>Website Activity Analysis</td>
</tr>
<tr>
<td>System Integration (Global SI)</td>
<td>Business Performance Metrics Dashboard</td>
<td>500 GB to 1TB in an 8-node Greenplum cluster</td>
<td>&gt;100,000</td>
<td>3,000</td>
</tr>
<tr>
<td>High-tech Manufacturing</td>
<td>Customer Service Management</td>
<td>200 GB in Oracle Cloudera Hadoop Loading 10 million records per hour 650,000 XML documents per week (2 to 4 MB each) 100+ million devices dimension</td>
<td>High-tech Manufacturing</td>
<td>Customer Service Management</td>
</tr>
<tr>
<td>Stream Global Provider of Sales, Customer Service and Technical Support for the Fortune 1000</td>
<td>10 Operational Dashboards</td>
<td>Data from 28 switches around the world 12 source systems – e.g. Oracle HRMS, SAP, Salesforce.com 20 million records per hour</td>
<td>200+</td>
<td>Today 120-200 Will add 50-100 more 49 locations across 22 countries</td>
</tr>
<tr>
<td>Sheetz</td>
<td></td>
<td>2+ TB in Teradata</td>
<td>80</td>
<td>30</td>
</tr>
</tbody>
</table>
Learn more about Pentaho Business Analytics

pentaho.com/contact
+1 (866) 660-7555.

Global Headquarters
Citadel International - Suite 340
5950 Hazeltine National Drive
Orlando, FL 32822, USA
tel +1 407 812 6736
fax +1 407 517 4575

US & Worldwide Sales Office
353 Sacramento Street, Suite 1500
San Francisco, CA 94111, USA
tel +1 415 525 5540
toll free +1 866 660 7555

United Kingdom, Rest of Europe, Middle East, Africa
London, United Kingdom
tel +44 (0) 20 3574 4790
toll free (UK) 0 800 680 0693

FRANCE
Offices - Paris, France
tel +33 97 51 82 296
toll free (France) 0800 915343

GERMANY, AUSTRIA, SWITZERLAND
Offices - Munich, Germany
tel +49 (0) 322 2109 4279
toll free (Germany) 0800 186 0332

BELGIUM, NETHERLANDS,
LUXEMBOURG
Offices - Antwerp, Belgium
tel (Netherlands) +31 8 58 880 585
toll free (Belgium) 0800 773 83

ITALY, SPAIN, PORTUGAL
Offices - Valencia, Spain
toll free (Italy) 800 798 217
toll free (Portugal) 800 180 060