Delivering Extreme-Scale In-Memory Analytics with Pentaho Business Analytics 4.1

November 10th, 2011

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Pentaho Momentum

Over **9,000 production** deployments

Over 8 Million lifetime downloads

In more than **185 countries**

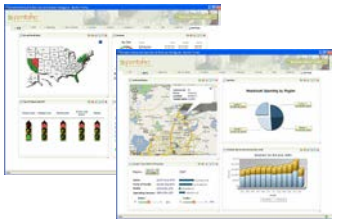
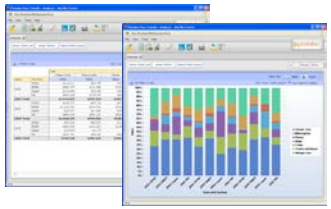
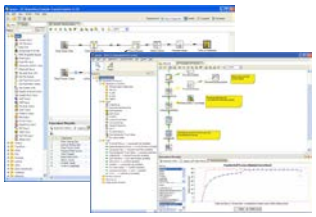
#1 In Web traffic of all independent BI vendors

Growing at over 100+%/Year

Average 1 download every 30 seconds

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Pentaho Business Analytics

Dashboards	Self-service Interactive KPI, Metrics, and Visualization		Business Users
Reporting	Ad hoc and Operational Reports		Business Users
Data Analysis	Self-service Interactive and Ad Hoc Analysis		Business Users
Data Integration	High Performance Data Integration and Cleansing		Developers & DBAs
Data Mining	Advanced Predictive Analysis		Advanced Power Users

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Pentaho Business Analytics 4.1

What's New

- Extreme scale in-memory data analysis
  - Predictable high-performance
  - Increased scalability
- In-memory aggregation
  - Faster performance
  - Reduced load on the database
- New supported data analysis data sources
  - Greenplum Database
  - Apache Hive for Hadoop
- Improved documentation

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Extreme Scale In-Memory Data Analysis

- Pentaho's data analysis capability now supports two external caching/data grid systems
  - Infinispan/JBoss Enterprise Data Grid
  - Memcached
- Extensible to other in-memory caching systems
- Access from tens of GB to potentially TBs of memory
  - RAM or flash memory
  - Single server or distributed

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

What does this mean to business users?

- Business users experience "in-memory performance"
  - Ask questions and <u>always</u> get answers at the "speed of thought"
  - Predictably fast response times
- Don't need to know or care if the data is in-memory

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

What does this mean to IT/developers?

- A sound, scalable and manageable analytics platform built on proven data warehousing and BI best practices
  - Easier to maintain and enhance
  - Having performance issues?
    - Throw memory at the problem vs. moving to a different database
  - Increased scalability
    - Data
    - Users

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Traditional In-Memory vs. Pentaho In-Memory

Traditional In-Memory	Pentaho In-Memory
Proprietary	Based on open source technology. Powers some of the world's biggest consumer web apps
Constrained to Java heap	Independent of the app's memory
Un-predictable performance	Predictable – always fast
Single-server	Distributed with failover
Specialized in-memory database skills	Standard DBA skills
Very expensive – need to add more BI servers or database servers	Cost-effective – scale-out by adding more memory or more commodity servers for memory

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

TECHNICAL DETAILS

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

External Memory Caches

- ~ 2001 – Shared caches to speed up large web sites
  - Store session state (e.g. shopping cart), profiles, content (e.g. images)
- Many providers:
  - memcached
  - Ehcache (Terracotta)
  - Infinispan (successor to JBoss Cache)
  - Gemcache (VMWare)
  - Coherence (Oracle)

- ~ 2005 – "Big data" problems popularize NoSQL architectures
  - Replicated data
  - Eventual consistency
  - Examples: BigTable, CouchDB, Cassandra, HBase

- JSR-107 JCACHE – Java Temporary Caching API
  - Specification process started in 2001
  - Will finally hit the streets in Java EE 7

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Infinispan / JBoss Enterprise Data Grid

- Peer to peer:
  - Infinispan instance lives within JVM
  - Write to your local Infinispan, and it propagates to other nodes
  - Nodes can join & leave system dynamically
- More than a cache
  - A data grid is an “authoritative source of data”
  - N+1 copies
  - Failover

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

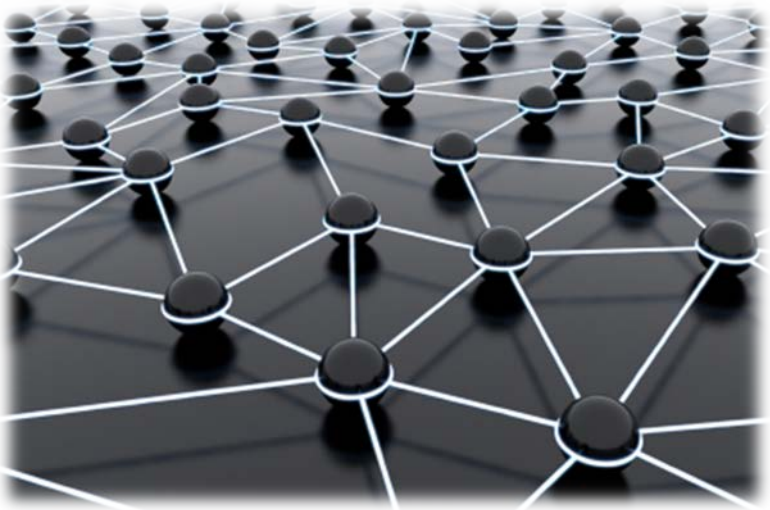
memcached

- Developed to improve performance of large web sites
- Used by YouTube, Reddit, Zynga, Facebook, Twitter, Wordpress
- Distributed hash table, flushing in least recently used (LRU) order
- Open source; commercial support from Couchbase & others

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Pentaho Analysis with Distributed Cache

- Patterns of usage:
  - One Analysis node with huge memory
  - Multiple Analysis nodes, sharing work
- Use the sum of all available memory across the cluster
  - E.g. 10 nodes with 32GB memory → 320GB effective memory
- High availability / failover
- Data grid lives longer than any server that belongs to it



© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Controlling the Contents of the Cache

- Cache is auto populated as a result of running queries
- You can prime the cache by running synthetic queries
- The cache grid lives longer than any constituent server
- Without distributed cache:
  - Usually sufficient to just restart Pentaho Analysis once a day (or whenever you reload the data warehouse)
- With distributed cache:
  - Explicit cache control is essential

- Pentaho Analysis has mature, sophisticated cache control
- Additions and deletions auto propagated through the cache
- Each cache access is effectively local

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Pluggable Caches

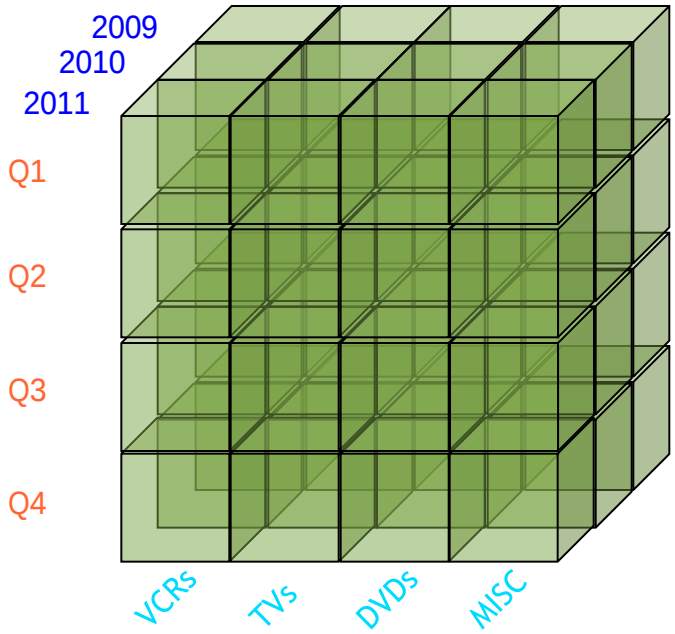
- Several cache providers and data grids available
  - Each has its own strengths
- Pentaho provides implementations for memcached, Infinispan
- Developers can use their own cache provider by implementing a Java SPI
  - interface mondrian.spi.SegmentCache

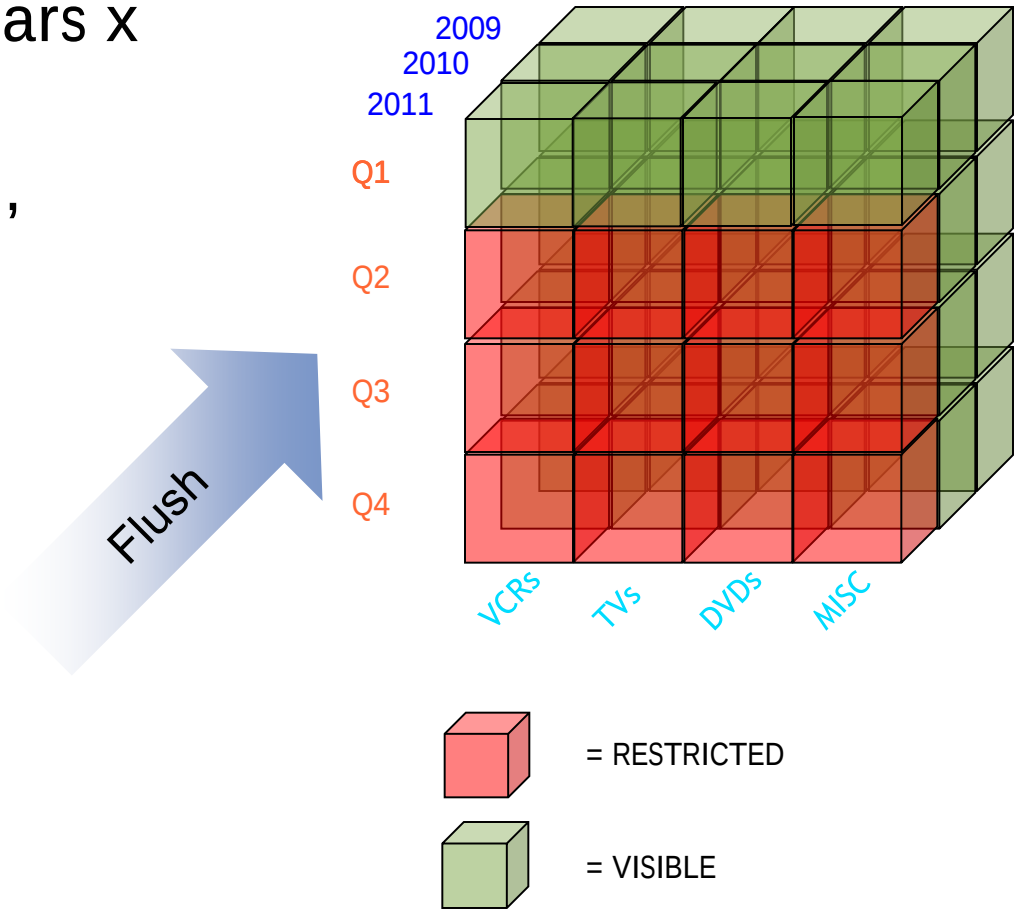
- Future directions:
  - Access non-relational data sources (e.g. HBase) via a “read-only cache”
  - Operational analytics by writing directly into cache

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Cache Management API

- Have a cell segment of years x quarters x products
- "Flush sales for year 2011, quarter Q2 onwards"
- "Flush whole sales cube"



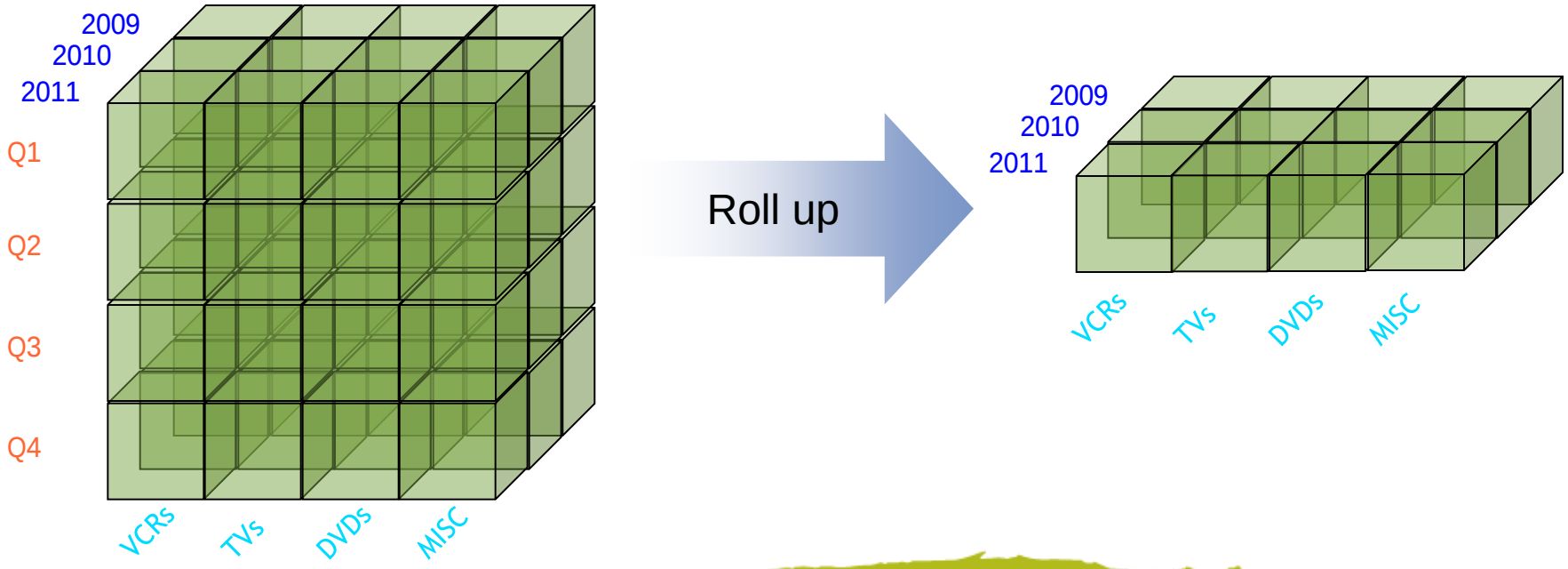


© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

In-Memory aggregation

- Have a cell segment of years x quarters x products
- Want years x products
- Roll up

- Very fast, due to compact in-memory data structure
- Can also combine multiple segments
- Segments can be in-memory or in external cache
- Offload the database



© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Case Study – StoneGate Senior Living

- Mission
  - Provide supported services in all areas of senior housing and long term care
- Business problem
  - Monitoring performance metrics across 6 different businesses
  - Speed of thought data access, analysis and dashboards by corporate management, facilities managers and line of business managers
- Solution
  - Pentaho's new in-memory analysis
  - Rolling out the system to more users
  - Adding more metrics to dashboards
- Results
  - More users, higher usage rates
  - Better, insight-driven business decisions

STONEGATE

REDEFINING ADULT LIVING

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Native Greenplum Support

- Greenplum Database
  - Next generation of “Big Data” warehousing and large-scale analytics processing
- Pentaho certified support
  - Pentaho analysis generates native SQL dialect for Greenplum
    - Instead of generic ANSI SQL
  - Optimizes performance when running on Greenplum
  - Leverages customer’s investment in the Greenplum Database

GREENPLUM
A DIVISION OF EMC

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Apache Hive Support

- Apache Hive
  - SQL access layer for Hadoop HDFS
  - Convenient generic data access and insert/update for Hadoop
- Pentaho support
  - Native SQL dialect support for Hive
  - Enables data analysis directly against Hadoop
    - But Hive's inherent high latency means this is more suitable for scheduled analytic reports and dashboards

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Pentaho BI 4 – Packages

Basic

Basic Edition Features
- Reporting
- Simple Analysis
- Self-Service Dashboards
- Data Integration
- Graphical ETL Designer
- Broad Connectivity

Standard Support
Support Portal
8x5 M-F

Professional

Includes All Basic Edition Features Plus:
- User-Driven Reporting
- **Powerful Analysis**
- **Extreme-Scale In-Memory**
- Data Discovery
- Guided Data Exploration
- Graphical Navigation
- Highly Interactive
- Rich Visualizations
- Powerful Charts
- Web-based
- Zero-Training

Standard Support
OR
Premium Support
- Phone & Support Portal
- 24X7X365 Coverage
- Consultative Support
- Consulting Checkpoints

Enterprise

Includes All Professional Edition Features Plus:
- Data Mining
- Predictive Analysis
- Hadoop Data Integration
- Hadoop Reporting

Premium Support
- Phone & Support Portal
- 24X7X365 Coverage
- Consultative Support
- Consulting Checkpoints

Red=External In-Memory Cache supported

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Next Steps

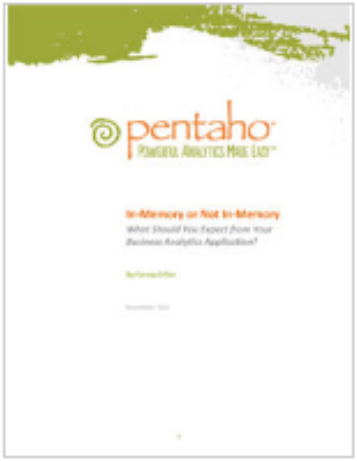
-

  - pentaho.com/get-started
  - Includes a FREE 30-day FULLY SUPPORTED evaluation
- Get More Information
  - pentaho.com/explore
- Developer resources:
  - Luc Boudreau's blog http://devdonkey.blogspot.com/2011/02/mondrian-spi-segmentcache.html
  - Segment Cache http://mondrian.pentaho.com/api/mondrian/spi/SegmentCache.html

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Resources

- **In-Memory or Not In-Memory:** What Should You Expect from your Business Analytics Application?



About this Document

Analytics today serve ever-growing volumes of data. While analyzing data from a few Excel spreadsheets or daily transactions may not require a lot of computing power, analysis of big data can be challenging.

In-memory analytics carry a great promise, but not all in-memory solutions perform the same. It is the characteristics of the in-memory solution that matters, not just having an in-memory option. The ideal solution is one that has in-memory capabilities, but does not limit you to an in-memory only architecture.

Download this white paper to learn:

- When you should use in-memory analytics
- When you should not
- What you should look for when evaluating an in-memory solution

Upload November 9th, 2011.

Read Now

http://www.pentaho.com/resources/whitepapers/

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.

Thank You!

http://blog.pentaho.com

@Pentaho

Facebook.com/Pentaho

Pentaho - Open Source Business Intelligence Group

© 2011, Pentaho. All Rights Reserved. www.pentaho.com.