Reducing the Implementation Efforts of Hadoop, NoSQL & Analytical Databases
### Goals for Today

**To understand:**

- Landscape of Big Data Databases Today
- Current Approaches to Implementation
- How Pentaho Reduces Time, Effort, and Complexity
Most businesses have 1 or more of these databases.

- Often used together to form a complete solution.

- Typically created / sold / supported by different vendor.

- Different programming interfaces for development and administration.
The Big Data Landscape

- Most businesses have 1 or more of these databases
- Often used together to form a complete solution
- Typically created / sold / supported by different vendor
- Different programming interfaces for development and administration
Transactional Databases

- Your everyday, garden-variety database
- Used to store data for line of business applications (ERP, CRM, HR)
- Serves as the “system of record” for business facts, e.g.
  - Inventory levels
  - Customer account status
  - Payroll data
- Offers rich sources of analytical information (potential facts and dimensions)
- Fast and reliable for INSERT|UPDATE|DELETE operations
Hadoop, HDFS and Hive

- Hadoop is a set of related open source Apache projects
- MapReduce is a programming framework
- HDFS is the Hadoop Distributed File System
  - Used to store structured (tabular) and unstructured files
- Structured files in HDFS can be queried via Hive
  - Hive provides a basic SQL interface on top of HDFS files
  - Hive generates MapReduce programs to fulfill the SQL requests
- Ideal for:
  - Unstructured data that doesn’t fit in a Transactional or NoSQL database
  - Email, twitter posts, images, voice logs, etc.
NoSQL Databases

• Non-relational databases
  – No (or very limited) support for the SQL language
    • Developers use programs, not queries to access data
  – Flexible definition of structure (“extend as you go”)
    • Limited metadata
  – No referential integrity / limited support for transactions

• But:
  – They can ingest (capture and persist) very large amounts of data with very low overhead, across multiple servers
    • 1000’s (to 100’s of thousands) of inserts per second
  – Excellent for large numbers of reads based on key values
    • Millions of reads per second
Analytical Databases

- Databases designed for Business Intelligence
  - Offer full support for SQL and relational metadata
  - Use specialized techniques to improve query performance
    - Compression
    - Column-based storage (e.g. “M” or “F”)
    - Clusters of independent servers
    - Advanced mathematics for index & query optimization
  - Support high-speed “bulk” inserts of structured data
  - May be offered as standalone software or via an appliance
- Ideal for:
  - Answering summary queries requiring complex filters, joins & groupings
  - Online Analytical Processing (OLAP) clients
Current Implementation Approaches
Complex and Time Consuming

Ingest data into big data store
Make data ready for analysis
- Transform
- Aggregate
- Structure-ize
Direct to big data store OR Move slice to DM/DW
Build metadata model:
- Dimensions
- Measures
- Hierarchies
Report
Dashboard
Analyze
Pivot
Drill
Act, then
iterate based on insights

Structured Data
CRM, POS, ERP, etc.

Unstructured Data
Forecasting
Clustering
Classification
Association
Regression

Current Developer Approach:
Coding
Coding
Coding
Coding
Coding

Current IT Approach:
Coding
Coding
ETL Tool
BI Tool Model Builder
BI Tool
Data Mining Tool
A Smarter Approach

Load/Ingest  Manipulate/Process  Access  Model  Visualize/Analyze  Predict  Decisions

- Ingest data into big data store
- Make data ready for analysis: Transform, Aggregate, Structure-ize
- Direct to big data store OR Move slice to DM/DW
- Build metadata model: Dimensions, Measures, Hierarchies
- Report, Dashboard, Analyze, Pivot, Drill
- Forecasting, Clustering, Classification, Association, Regression
- Act, then iterate based on insights

 STRUCTURED DATA
 CRM, POS, ERP, etc.

 UNSTRUCTURED DATA

- Forecasting
- Clustering
- Classification
- Association
- Regression

Current Developer Approach:
- ETL Tool
- BI Tool
- Model Builder
- BI Tool
- Data Mining Tool

Current IT Approach:
- Coding
- Coding
- Coding
- Coding
- Coding

Business Analytics for Big Data

pentaho®
Pentaho Visual Development for Big Data
15x Faster to Design & Deploy Big Data Analytics

Complete platform from data to analytics
- Continuity of one platform
- Eliminate disjointed steps
- Days instead of weeks

High productivity visual development
- Use existing skills for Hadoop
- Visual development for MapReduce jobs
- Visual development for **Sqoop, Oozie**

Fast performance within Hadoop
- Parallel execution of MapReduce jobs across the Hadoop cluster

Coding VS. Visual Data Preparation

Pentaho MapReduce
Sample Scenario

• A telecommunication company needs to:
  – Ingest, store and retrieve call detail records (CDRs) from multiple digital switches
  – Maintain a detailed profile of its commercial customers
  – Analyze the volume of calls over time by SIC code, region, state, county and city

• Which database technology would it use for each of these applications?
Pentaho Data Integration for Big Data

Leverage Native Capabilities of Big Data Vendors

- Transform the data for reporting *without coding*
- Execute in the Hadoop Cluster *without java coding*
- Visually orchestrate job workflows
- Visualize and analyze the data