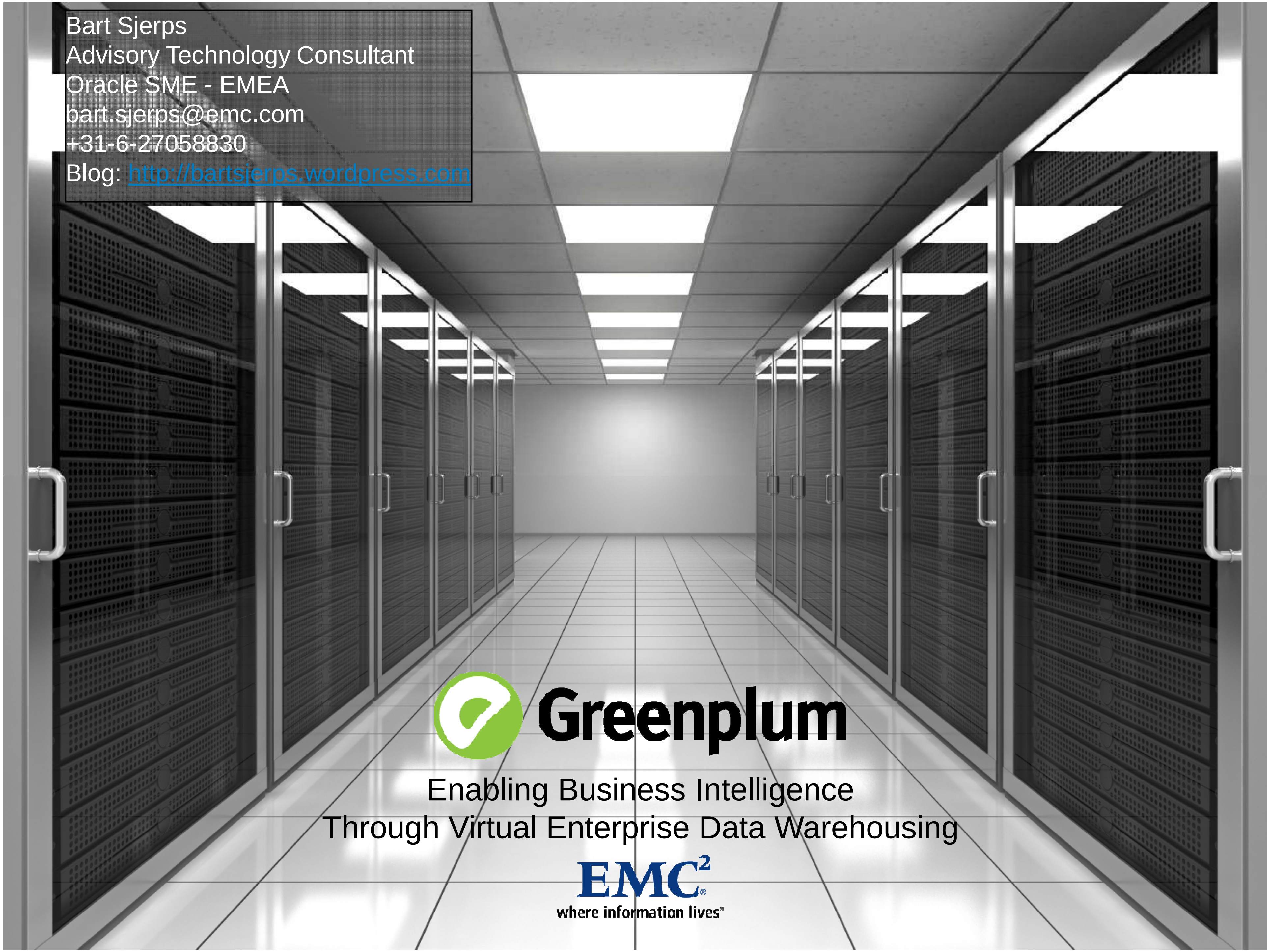


Bart Sjerps
Advisory Technology Consultant
Oracle SME - EMEA
bart.sjerps@emc.com
+31-6-27058830
Blog: http://bartsjerps.wordpress.com

Greenplum

Enabling Business Intelligence Through Virtual Enterprise Data Warehousing

EMC²
where information lives®

Introduction & Agenda

- What is Data warehousing?
  - And what's Business Intelligence?
  - Evolution in the Data Warehouse
  - Business purpose
  - Classic DWH architecture
- Present and future challenges
- EMC Solution
  - Greenplum

What is a Data Warehouse?

A Datawarehouse is not…

Vendor and consultant proclamations aside, a data warehouse is not:

- A project
  - With a specific end date
- A product you buy from a vendor
  - Like an ODS (such as SCT's)
  - A canned "warehouse" supplied by iStrategy
  - Cognos ReportNet
- A database schema or instance
  - Like Oracle
  - Or SQL Server
- A cut-down version of your live transactional database

According to Ralph Kimball and Joe Caserta, a data warehouse is:

> A system that extracts, cleans, conforms, and delivers source data into a *dimensional data store* and then supports and implements querying and analysis for the purpose of decision making.

Another def.: The union of all the enterprise's data marts

- Aside: The Kimball model is not without some critics:
  - E.g., Bill Inmon

Data Warehousing & Business Intelligence Definitions

Operational Data Store
An ODS is an environment that pulls together, validates, cleanses and integrates data from disparate source application systems.

Data Warehouse
A repository of an organization's electronically stored data, designed to facilitate reporting and analysis.

Data Mart
A smaller version of a data warehouse – typically targeted at a specific portion of an organization.

Business Intelligence
Aims to support better decision making by analyzing data contained in a data warehouse and data mart.

Extract Transform and Load (ETL)
A process where data is extracted from external sources, transformed to fit operational needs and then loaded into the database or data warehouse.

Scan Rate
How quickly a data warehouse or database can read and process data

Data Load Rate
How quickly a data warehouse or database can ingest data

History: From reports to advanced analytics

- Early days: run a simple report against the OLTP Database
- Run heavy batch reports against OLTP Database
  - Dayly, weekly, monthly, year-end, ad-hoc
- Run custom queries against OLTP Database (using standard reporting tools)
  - First use of what later became Business Intelligence, getting (market) knowledge from large amounts of information
- Note: Running Batch and reporting on OLTP kills OLTP response time and performance

- Offload databases for reporting and querying only
  - Implemented as 1:1 copies, or custom designed databases (the first pure Data Warehouses)
- Need for Extract, Transform, Load tools (ETL)
- Evolved into OLAP (Online Analytical Processing); specialized methods for running Analytics
- This required special reporting tools as well

Classic vs. Next-gen business intelligence

Old-style Datawarehousing:

- Frequently run reports/batch
  - Built by programmers, optimized for performance and minimizing resource usage, requires huge developer and DBA efforts
  - This is achieved by classic tuning such as using table indexes, partitioning, SQL optimization
  - Very efficient but only for predictable queries
- Ad-hoc queries against OLTP data
  - Can kill OLTP service levels, therefore this is often offloaded against prod database copy
  - Optimizing using “tricks” such as materialized views
  - Classic tuning fails (because it’s unpredictable)
- DWH misused for pieces of business process
  - Now mission-critical!
  - Consider HA / DR / Compliancy

New style Datawarehousing:

- Does not replace classic DWH!
- Get as much data from as many sources as possible
  - Web, data feeds, legacy systems, “smart” electronics, etc etc
- Clean it up and modify it for analytics using ETL tools
  - This is very resource intensive and typically requires long processing times
  - Loading in the DWH can be problematic
    - Classic DB systems again use workarounds for speeding it up
  - Data needs to be as up-to-date as possible (less than 24 hours old)
- Build multi-dimensional databases
  - That can have holes with “missing” data
- Build specialized data-marts
  - Optimized by purpose
  - Contains sub-set of all data

DW/BI purpose

- Risk Management
  - Credit risk, Operational risk, Market risk
- Financial Analysis
  - Customer credit, Cash inflow, Key financial ratios/performance
- Fraud Detection
  - Internal fraud, External fraud
- Compliance
  - Data integration, Reporting, Audit
- Customer Intelligence
  - Behavior analysis, Spend & Value analysis, Portfolio analysis, Scorecard and Rating applications
- Performance Management
  - Analysis of business methodologies, metrics, processes and systems

Real-world example of next-gen analytics: Customer relations in large financial

Initial findings from analytical reports:

- Most customers generate moderate to good profit
- Few customers generate large loss
- If this loss can be eliminated, net profits will be much higher

BI question: What is causing this loss?

- Customers cause loss due to heavy claim on involvement of financial experts, frequently change contracts (that only generate profit long-term), etc
- Based on such conditions the BI tools can identify customers that will likely cause more loss

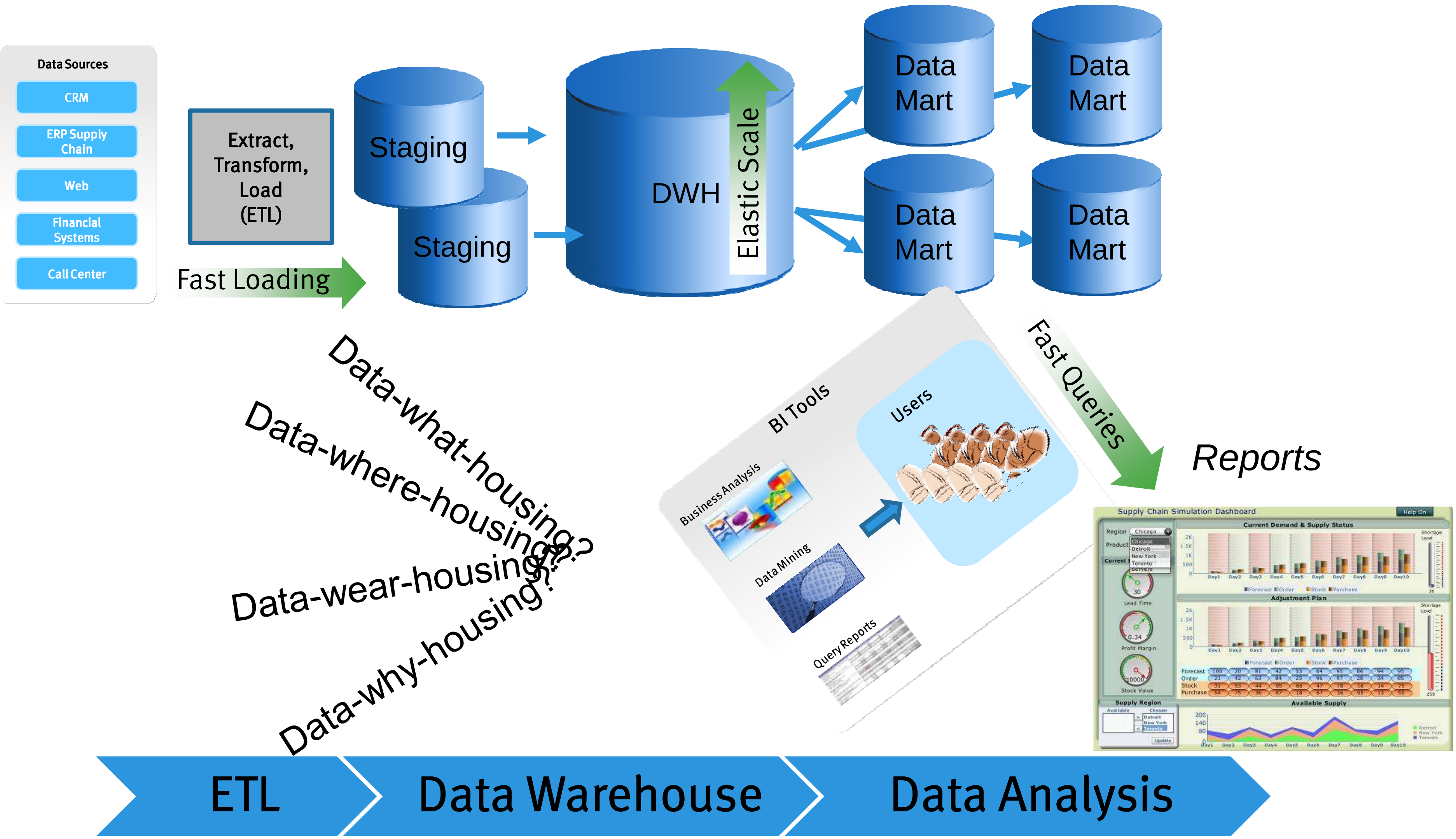
BI question: Why do customers leave?

- Unhappy with customer service (long time on-hold at call centre before being serviced)
- Wrong information and prices in their offers (sometimes offers had to be re-done 3-4 times)

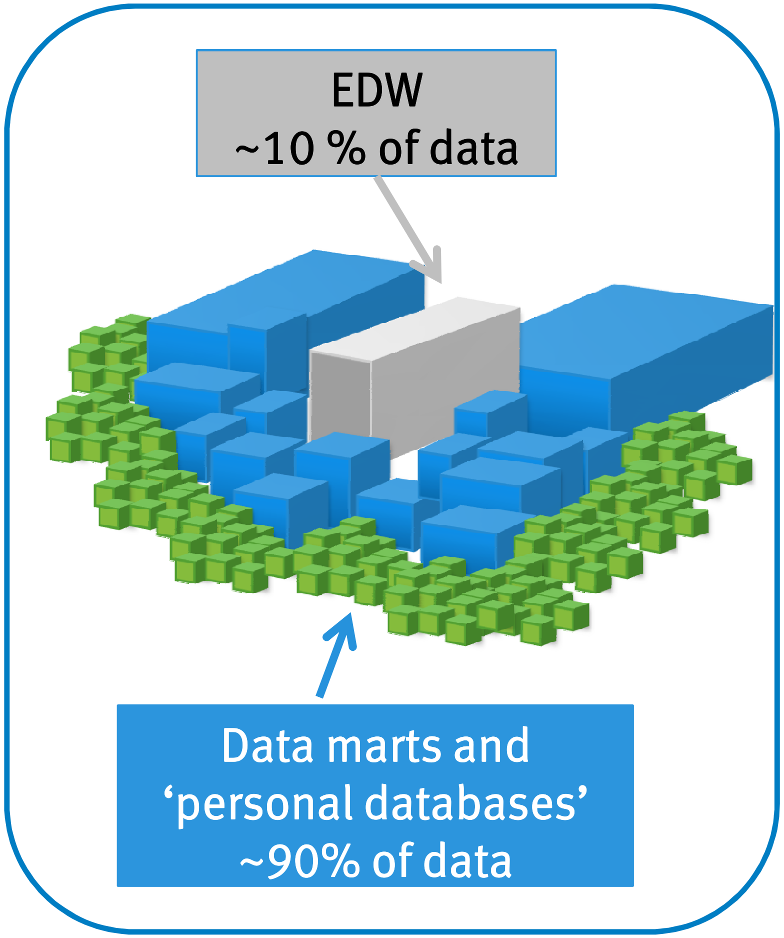
Decision (made by human analysist based on BI findings):

- Make loss-generating customers ***even more*** unhappy by ***deliberately*** annoying them so they will leave and go to a competitor (causing ***them*** to loose money)
  - (and improve service to “good” customers)
- Again, BI feedback into CRM systems can automate this process

Classic Data Warehouse Architecture



Datamart "Sprawl"



EDW
~10 % of data

Data marts and
'personal databases'
~90% of data

- Data is everywhere and growing
  - 44X data growth by 2020
  - 100s of data marts
  - 'Shadow' databases
- Critical business insight is outside EDW
- Centralized legacy systems are expensive
- System expansion is slow and process heavy
- Proprietary HW systems lag behind open systems innovation

Traditional solutions cannot scale to meet the DW/BI challenges

Business Intelligence Challenges (1)

Related to Infrastructure

- Higher service levels
  - DWH not allowed to be down for a few days
  - Need for backup/recovery/DR
  - No SPOF, high-availability architecture
  - Don’t forget security, auditing, compliancy, data leakage prevention, customer privacy considerations (think Facebook and Google)
- Massive growth
  - According to research firms, unstructured data will be biggest growth factor for companies
  - Business Intelligence is #2
  - Soon we will see datawarehouses 100’s of Terabytes in size (And the first Petabyte customers)
  - Business people want to store more and more in the DWH

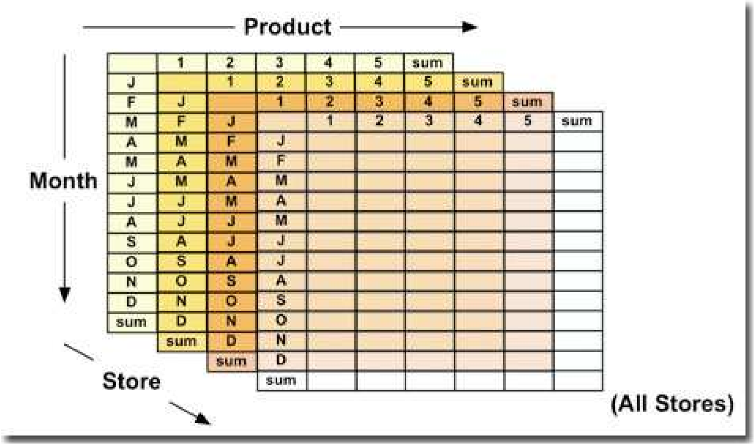
Business Intelligence Challenges (2)

Related to Infrastructure

- Loading time
  - DWH needs to have up-to-date info
  - Load times of multiple days is simply no longer acceptable
  - 24H is max (for the whole process, not just loading)
  - Long term, drive to real-time (ouch!)
- “Scan” time (how long does it take to run a query)
  - More data
  - More impatient end users
  - More ad-hoc queries
  - Cannot optimize this anymore with classic SQL tuning and database tricks & magic

Business Intelligence Challenges (3)

Related to Infrastructure



And finally… New paradigms

- Multi-dimensional OLAP databases
- In-memory statistical calculations
  - Needs to load a data subset in memory real quick
- Web users accessing BI data
  - Of course, through web applications
  - Massive scale-up in # of parallel transactions

Current State of the DW/BI Industry

- A separate market for data warehouse infrastructure exists because normal IT infrastructure does not meet customers' performance and scalability needs at acceptable cost.
- The storage piece of the DW/BI market has been defined by best performance at lowest cost per TB
  - This means direct attach JBOD and the lowest-end SAN storage are seen as the defacto standard
- Enterprise storage features for protecting the warehouse, such as SRDF or consistency technology, are sometimes a factor. But not often (Expected to change)
- A physical appliance market emerged due to the ease of deployment, and simplified sales model focusing directly on the customers' business unit.
- An existential battle is emerging between fully integrated vertical stack vendors, and horizontal infrastructure providers. Data warehousing technology is at the forefront of this battle.
- Concepts of virtualized data warehouse appliances, and cloud infrastructures optimized for data warehouse workloads are receiving attention.

EMC Focused Areas

- Data warehouse consolidation
  - Improve efficiency and data transparency
  - Reduce infrastructure redundancy
  - Improve Total-Cost-of-Ownership
- Deliver high performance & scalability for analytics workloads
  - Perform aggregation, reporting and translation much faster than conventional approaches
  - Improve reporting turn-around time to support better decision making process
- Manage historical raw data in the archive / cloud
  - Provide convenient access to historical raw data
  - Comply to new regulatory requirements
  - Maximize storage and retrieval efficiency

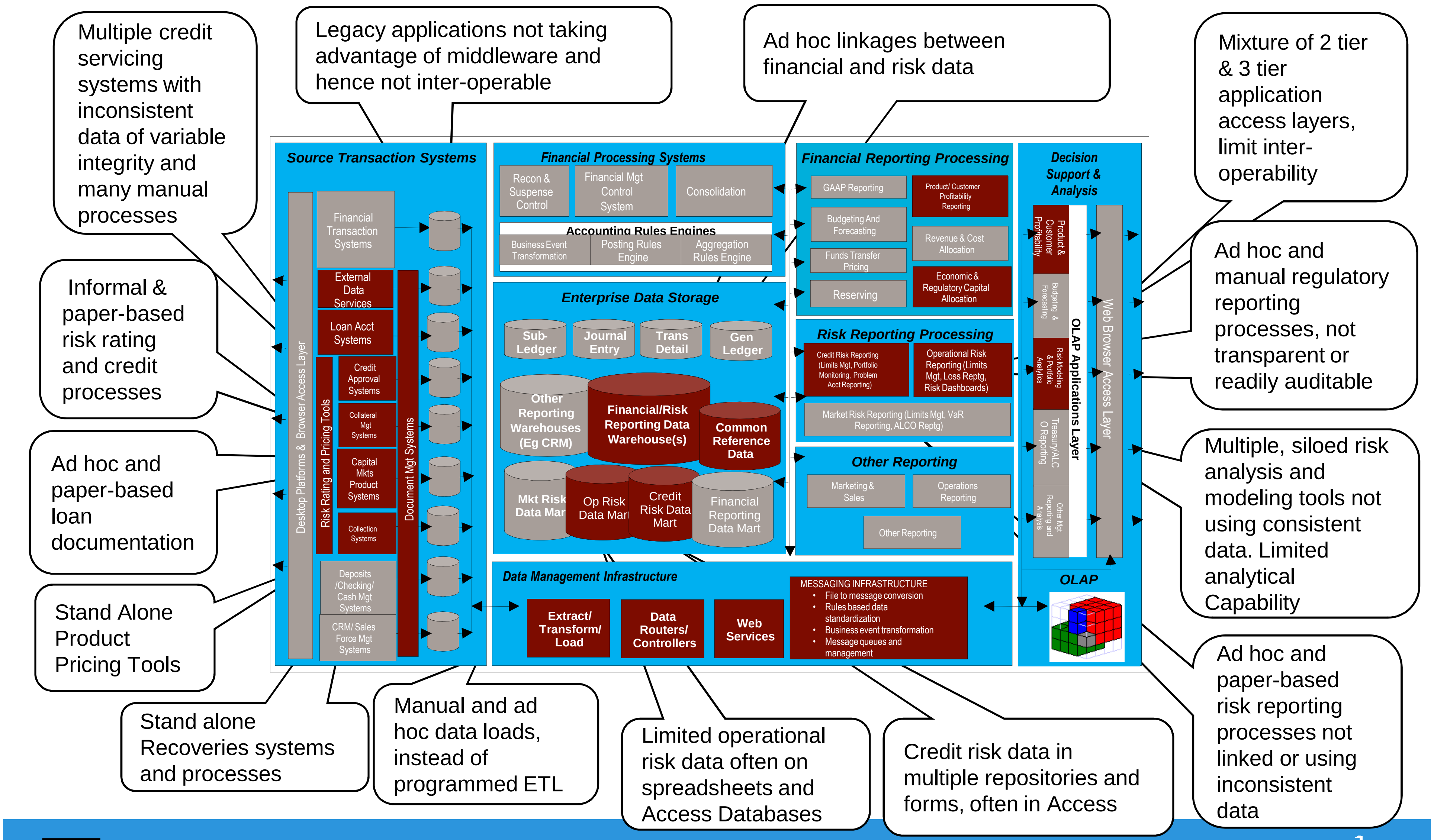
Business and technology challenges

- Increased regulatory scrutiny and business reporting requirements
  - Insufficient data transparency across all risk exposures
  - Processing cycle taking too long
  - Lengthy reporting turn-around time
  - Need to retain more data over extended period of time

- All these need to be enabled by IT and supporting Infrastructure
  - Maintain performance amid escalating data volumes
  - Aggregate data sets from many silos
  - Ad hoc analysis and reporting occurring more frequently
  - Enable accessibility to historical raw data
  - Enable easy provisioning and expansion

- Upgrading existing infrastructure is very expensive and in many cases is cost prohibitive

An illustration of the massive technical challenge



Multiple credit servicing systems with inconsistent data of variable integrity and many manual processes

Legacy applications not taking advantage of middleware and hence not inter-operable

Ad hoc linkages between financial and risk data

Mixture of 2 tier & 3 tier application access layers, limit inter-operability

Informal & paper-based risk rating and credit processes

Ad hoc and paper-based loan documentation

Stand Alone Product Pricing Tools

Ad hoc and manual regulatory reporting processes, not transparent or readily auditable

Multiple, siloed risk analysis and modeling tools not using consistent data. Limited analytical Capability

Ad hoc and paper-based risk reporting processes not linked or using inconsistent data

Stand alone Recoveries systems and processes

Manual and ad hoc data loads, instead of programmed ETL

Limited operational risk data often on spreadsheets and Access Databases

Credit risk data in multiple repositories and forms, often in Access

Source Transaction Systems

Desktop Platforms & Browser Access Layer

Financial Transaction Systems

External Data Services

Loan Acct Systems

Risk Rating and Pricing Tools

Credit Approval Systems

Collateral Mgt Systems

Capital Mkts Product Systems

Collection Systems

Deposits /Checking/ Cash Mgt Systems

CRM/ Sales Force Mgt Systems

Document Mgt Systems

Financial Processing Systems

Recon & Suspense Control

Financial Mgt Control System

Consolidation

Accounting Rules Engines

Business Event Transformation

Posting Rules Engine

Aggregation Rules Engine

Enterprise Data Storage

Sub-Ledger

Journal Entry

Trans Detail

Gen Ledger

Other Reporting Warehouses (Eg CRM)

Financial/Risk Reporting Data Warehouse(s)

Common Reference Data

Mkt Risk Data Mar

Op Risk Data Mart

Credit Risk Data Mart

Financial Reporting Data Mart

Financial Reporting Processing

GAAP Reporting

Product/ Customer Profitability Reporting

Budgeting And Forecasting

Revenue & Cost Allocation

Funds Transfer Pricing

Economic & Regulatory Capital Allocation

Reserving

Risk Reporting Processing

Credit Risk Reporting (Limits Mgt, Portfolio Monitoring, Problem Acct Reporting)

Operational Risk Reporting (Limits Mgt, Loss Reptg, Risk Dashboards)

Market Risk Reporting (Limits Mgt, VaR Reporting, ALCO Reptg)

Other Reporting

Marketing & Sales

Operations Reporting

Other Reporting

Decision Support & Analysis

Product & Customer Profitability

Budgeting & Forecasting

Risk Modeling & Portfolio Analytics

Treasury/ALCO Reporting

Other Mgt Reporting and Analysis

OLAP Applications Layer

Web Browser Access Layer

OLAP

Data Management Infrastructure

Extract/ Transform/ Load

Data Routers/ Controllers

Web Services

MESSAGING INFRASTRUCTURE
- File to message conversion
- Rules based data standardization
- Business event transformation
- Message queues and management

🟥 **Key Systems and Processes significantly impacted**

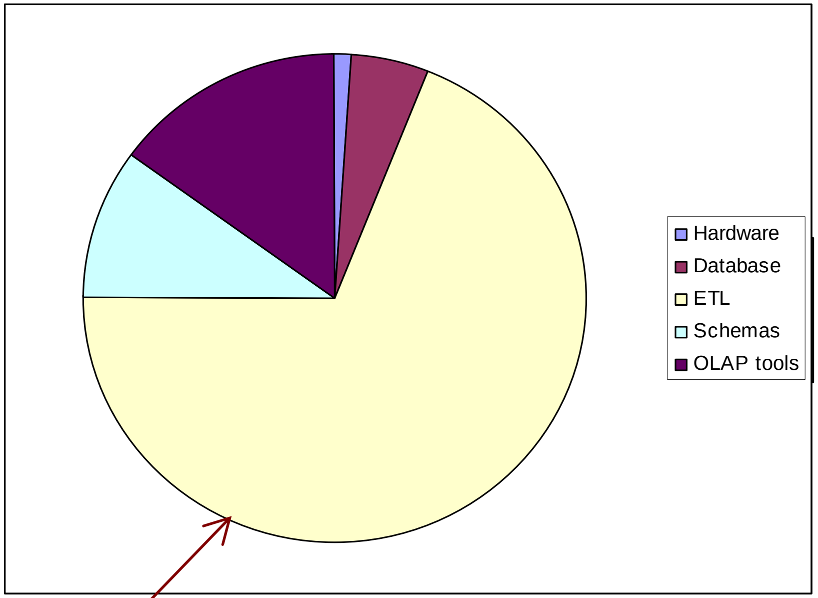
EMC² where information lives®

Implementing a Data Warehouse

- In many organizations IT people want to huddle and work out a warehousing plan, but in fact
  - The purpose of a DW is decision support
  - The primary audience of a DW is therefore College decision makers
  - It is College decision makers therefore who must determine
    - Scope
    - Priority
    - Resources
- Decision makers can’t make these determinations without an understanding of data warehouses
- It is therefore imperative that key decision makers first be educated about data warehouses
  - Once this occurs, it is possible to
    - Elicit requirements (a critical step that’s often skipped)
    - Determine priorities/scope
    - Formulate a budget
    - Create a plan and timeline, with real milestones and deliverables!

What Takes Up the Most Time?

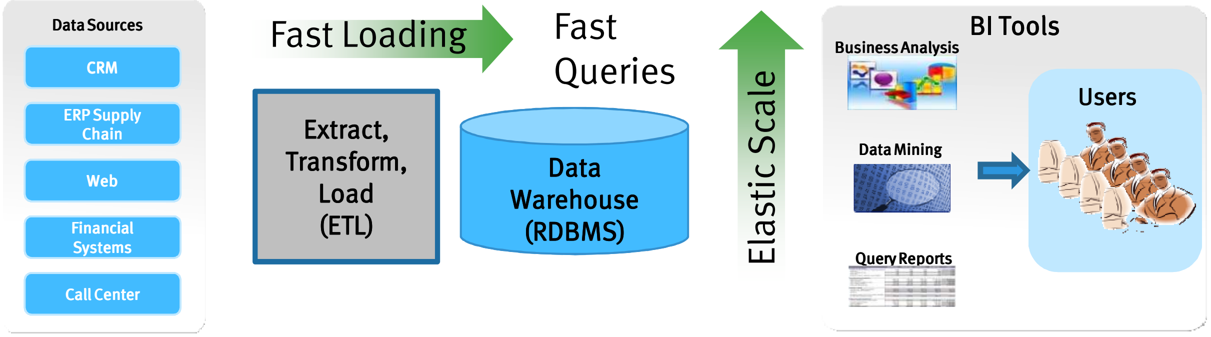
- You may be surprised to learn what DW step takes the most time
- Try guessing which:
  - Hardware
  - Physical database setup
  - Database design
  - ETL
  - OLAP setup



Acc. to Kimball & Caserta, **ETL** will eat up 70% of the time.
Other analysts give estimates ranging from 50% to 80%.

The most often underestimated part of the warehouse project!

Data Warehouse Requirements



Data Sources

CRM

ERP Supply Chain

Web

Financial Systems

Call Center

Fast Loading

Extract, Transform, Load (ETL)

Fast Queries

Data Warehouse (RDBMS)

Elastic Scale

BI Tools

Business Analysis

Data Mining

Query Reports

Users

Rapid Data Ingest From Many Sources

DATA PROTECTION
Self-healing and fault tolerance
Continuous remote replication
Fast and efficient backups

ELASTIC SCALE
Scale to Petabytes with automatic data distribution
Linear performance improvements
No manual partitioning

SIMPLE MANAGEMENT
User self service for agility
Real-time query optimization
Fair sharing workload management

MASSIVE PARALLEL PROCESSING ARCHITECTURE
“Shared Nothing” MPP scale-out architecture
Embedded, “in-database” analytics
Run compute near the storage
Distributed for scalability

Analysis in Parallel Across the Enterprise

EMC²
Greenplum