Zentrale hochverfügbare Konsolidierung von Unternehmensinformation

Holger Villringer NED EMEA PreSales
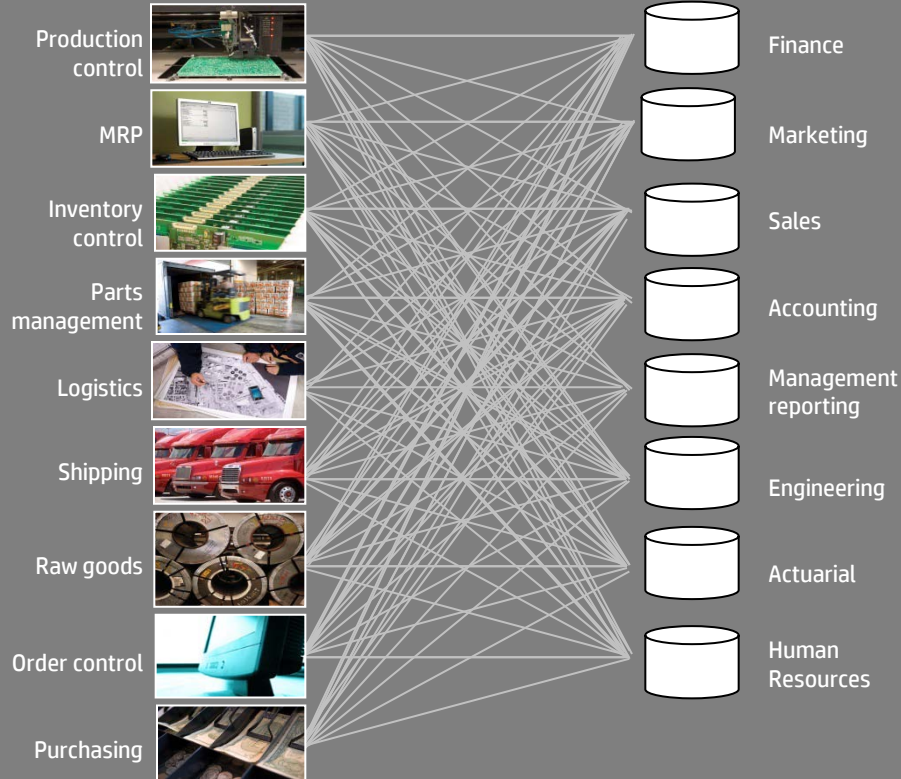
GTUG 2013

Agenda

- Data Mart chaos to Data Warehouse
- Big Data & Data Types
- Structured Data
- Semi- & Unstructured Data
- HP Logical Data Warehouse
- HP Analytic Cloud

History

Legacy applications + data marts = <u>chaos</u>

Production control · MRP · Inventory control · Parts management · Logistics · Shipping · Raw goods · Order control · Purchasing

Finance · Marketing · Sales · Accounting · Management reporting · Engineering · Actuarial · Human Resources

Data separated in
- Regions / Sub-Regions
- Business Unit
- Product Lines
- Administration functions
- Production plants

HP Example
- 762+ Data marts
  - ~ 10 feeds per mart
  - ~ 5 hops per mart
  - 7,620 feeds and 3,810 hops
- 20M commercial customers but 200M customer IDs
- 165 Countries but 695 Country IDs
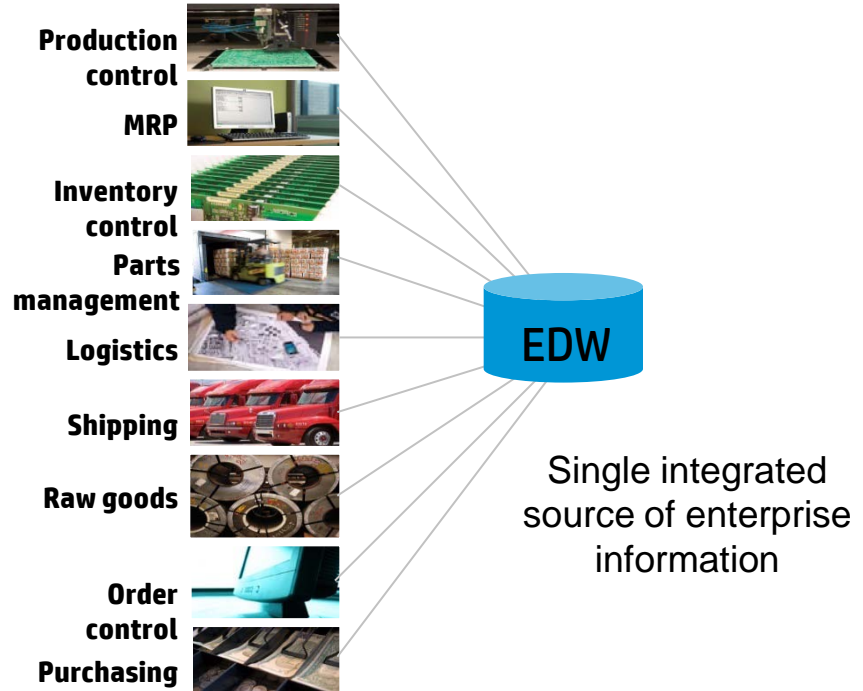- 85 Data Centers

Enterprise data warehouse

Simplification and less cost

EDW Principles

- Obtain data once and at the source
- Data is complete and detailed
- Define, model, and map all data
- Provide complete access
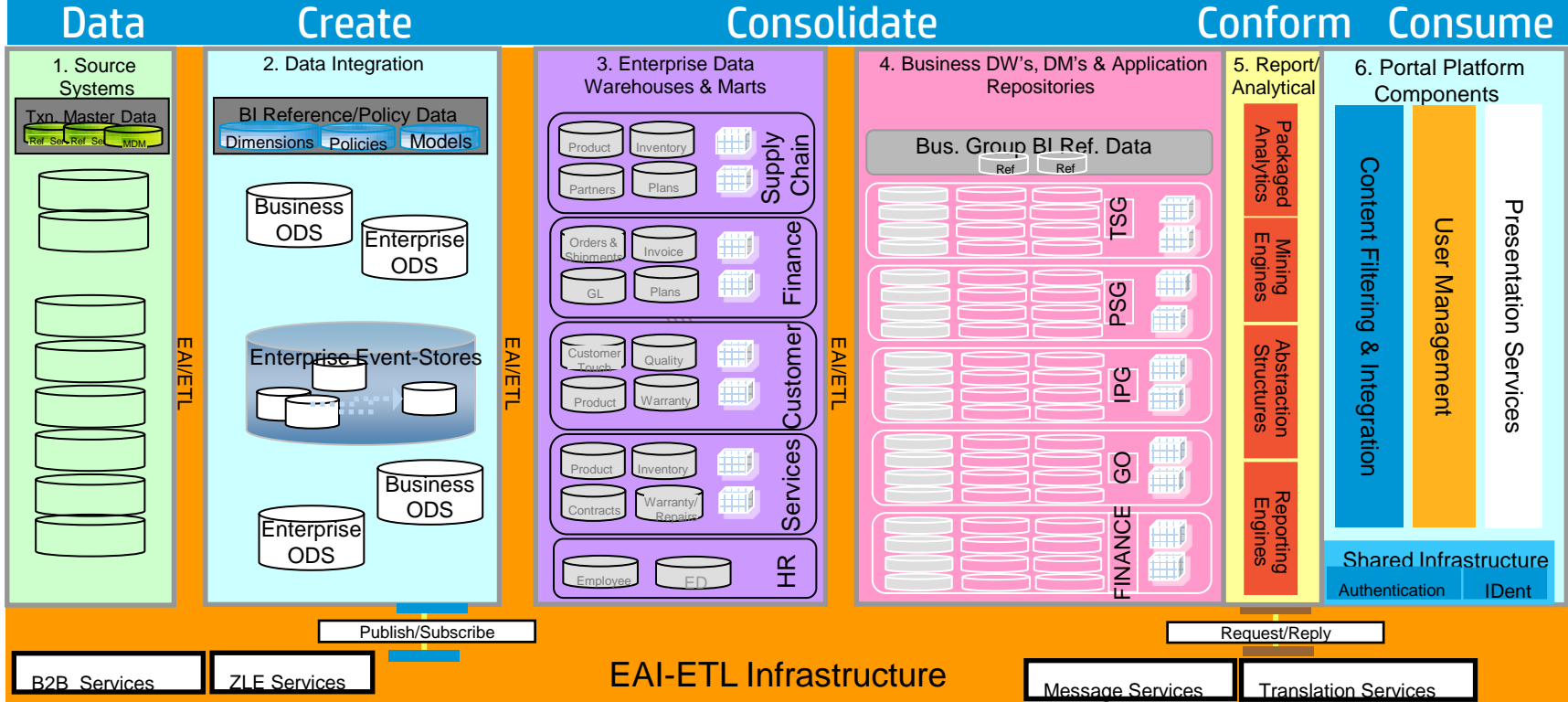- Flexibility and scale

EDW Benefit

- Greatly reduced integration and support
- Much less data movement
- Global view the Company
- Self-funding through DM retirement
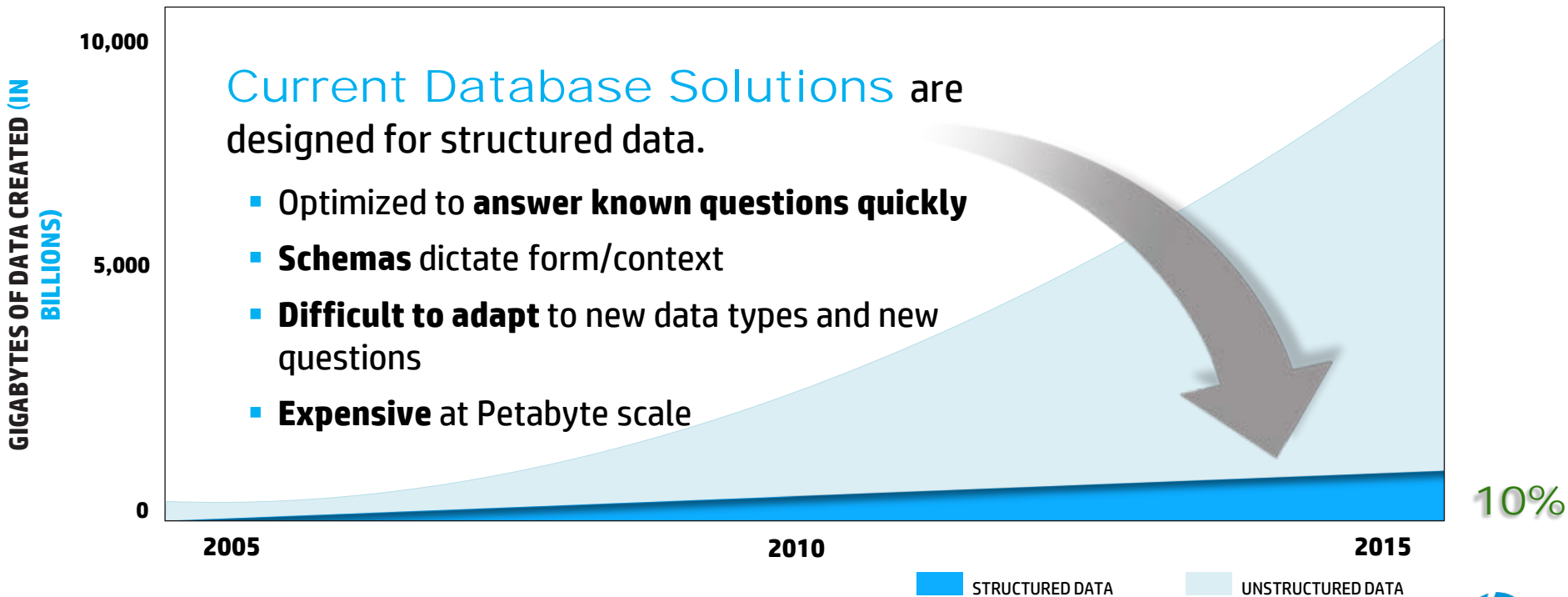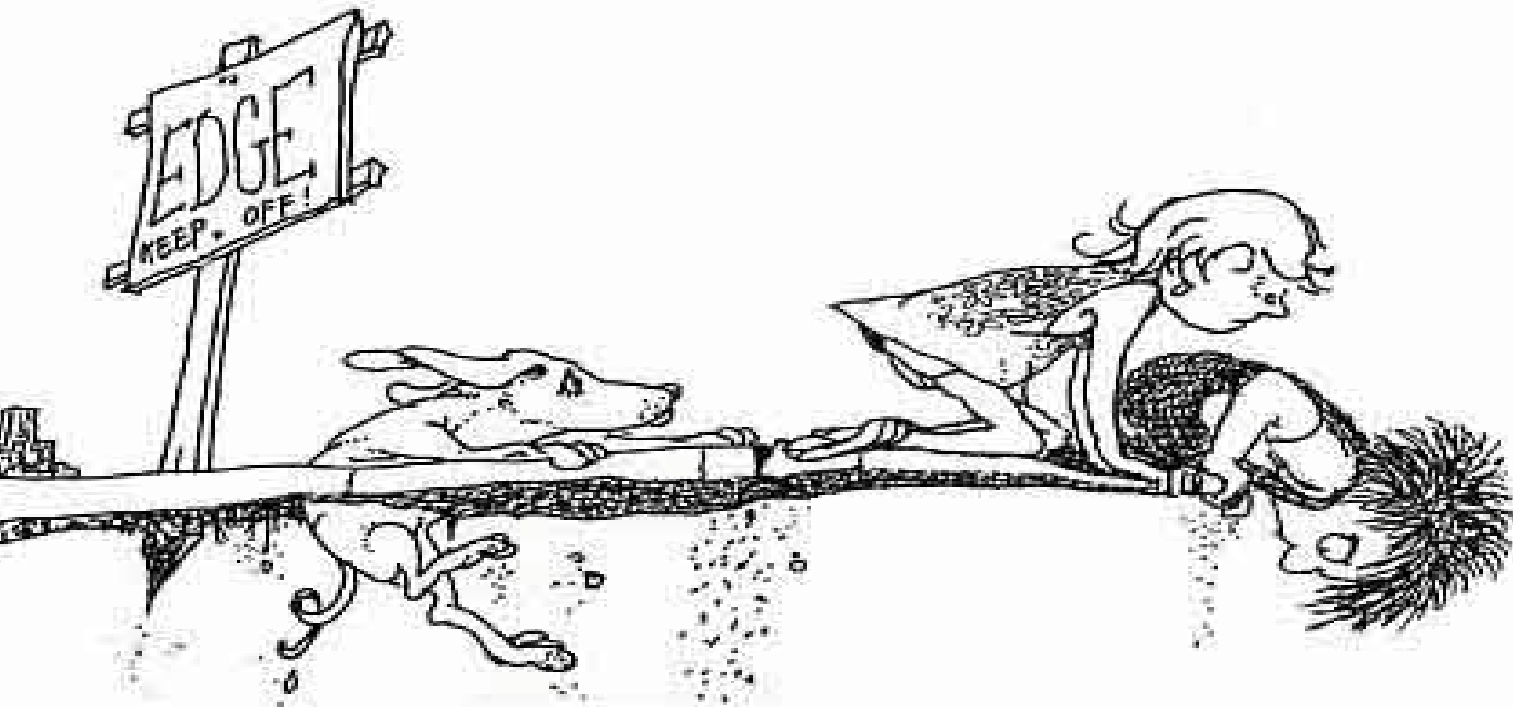 - system, support, networking, licensing, staff

Production control

MRP

Inventory control

Parts management

Logistics

Shipping

Raw goods

Order control

Purchasing

EDW

Single integrated source of enterprise information

Enterprise data warehouse

Enterprise Meta & Control Data Management

Data	Create	Consolidate	Conform	Consume

1. Source Systems
- Txn. Master Data
- Ref. Set, Ref. Set, MDM

2. Data Integration
- BI Reference/Policy Data
- Dimensions | Policies | Models
- Business ODS
- Enterprise ODS
- Enterprise Event-Stores
- Business ODS
- Enterprise ODS

3. Enterprise Data Warehouses & Marts

Supply Chain
- Product | Inventory
- Partners | Plans

Finance
- Orders & Shipments | Invoice
- GL | Plans

Customer
- Customer Touch | Quality
- Product | Warranty

Services
- Product | Inventory
- Contracts | Warranty/Repairs

HR
- Employee | ED

4. Business DW's, DM's & Application Repositories
- Bus. Group BI Ref. Data
  - Ref | Ref
- TSG
- PSG
- IPG
- GO
- FINANCE

5. Report/Analytical
- Packaged Analytics
- Mining Engines
- Abstraction Structures
- Reporting Engines

6. Portal Platform Components
- Content Filtering & Integration
- User Management
- Presentation Services
- Shared Infrastructure
  - Authentication | IDent

EAI/ETL (vertical bars)

Publish/Subscribe

Request/Reply

EAI-ETL Infrastructure

- B2B Services
- ZLE Services
- Message Services
- Translation Services

The Challenge: The Current Database Solutions

Current Database Solutions are designed for structured data.

- Optimized to **answer known questions quickly**
- **Schemas** dictate form/context
- **Difficult to adapt** to new data types and new questions
- **Expensive** at Petabyte scale

GIGABYTES OF DATA CREATED (IN BILLIONS)

10,000

5,000

0

2005   2010   2015

10%

STRUCTURED DATA   UNSTRUCTURED DATA

Where the Data Warehouse ends...
But not the end of the Data Warehouse!!!

Big Data

Expansion of the Digital Universe
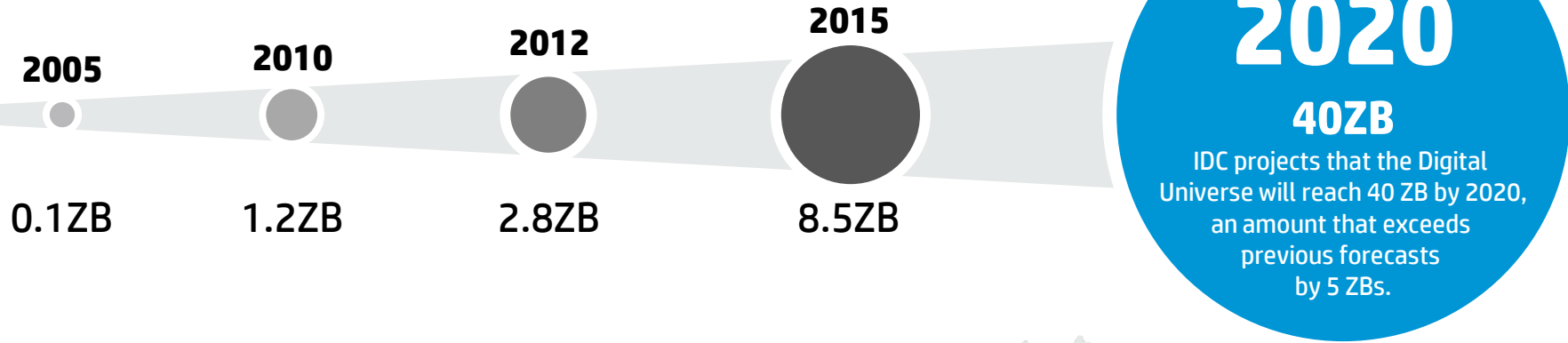
2005

2010

2012

2015

2020

40ZB

IDC projects that the Digital Universe will reach 40 ZB by 2020, an amount that exceeds previous forecasts by 5 ZBs.

0.1ZB

1.2ZB

2.8ZB

8.5ZB

40% by 2020

Machine-generated data is a key driver in the growth of the world's data – which is projected to increase 15x by 2020 (representing 40% of the digital universe)

Western Europe 19%

U.S. 32%

China 13%

rest of the world 32%

India 4%
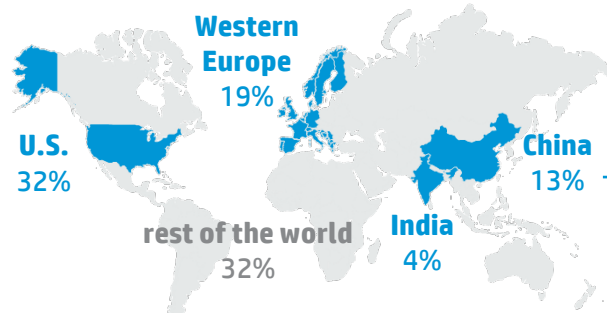
22% by 2020

Current breakdown of the Digital Universe

By 2020, China alone is expected to generate 22% of the world's data!

Todays Data by the numbers: challenges

Variety, velocity, volume, time to value

40% of digital content created **by 2020** will be from **sensor** data [1]

75% of currently deployed data warehouses will not scale sufficiently to meet new information **velocity** and complexity of demands **by 2016** [2]

**50**% Worldwide information **volume** growth of digital content in **2013** [3]

86% of corporations **cannot** deliver the right information, at the right **time** to support enterprise outcomes all of the time [4]

[1]Source: IDC Digital Universe in 2020, December 2012

[2]Source: Gartner - The State of Data Warehousing in 2012

[1]Source: IDC Predictions 2013: Competing on the 3rd Platform

[3]Source: Coleman Parkes Survey Nov 2012
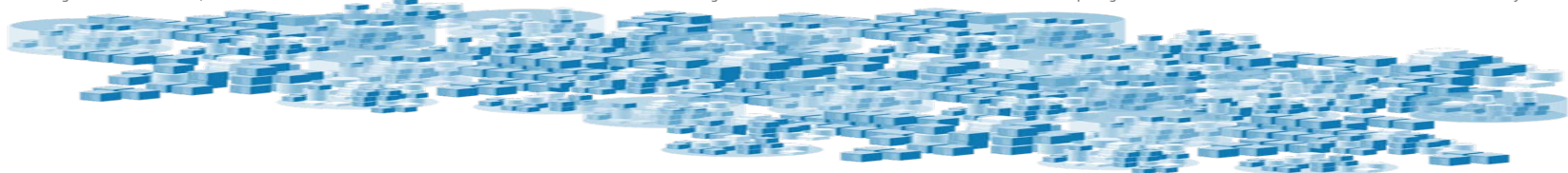
Todays Data opportunities—Won and lost

Competitive advantage in the digital universe in 2012
Massive amounts of useful data are getting lost

% of data that would be potentially useful IF tagged and analyzed

23%

3%

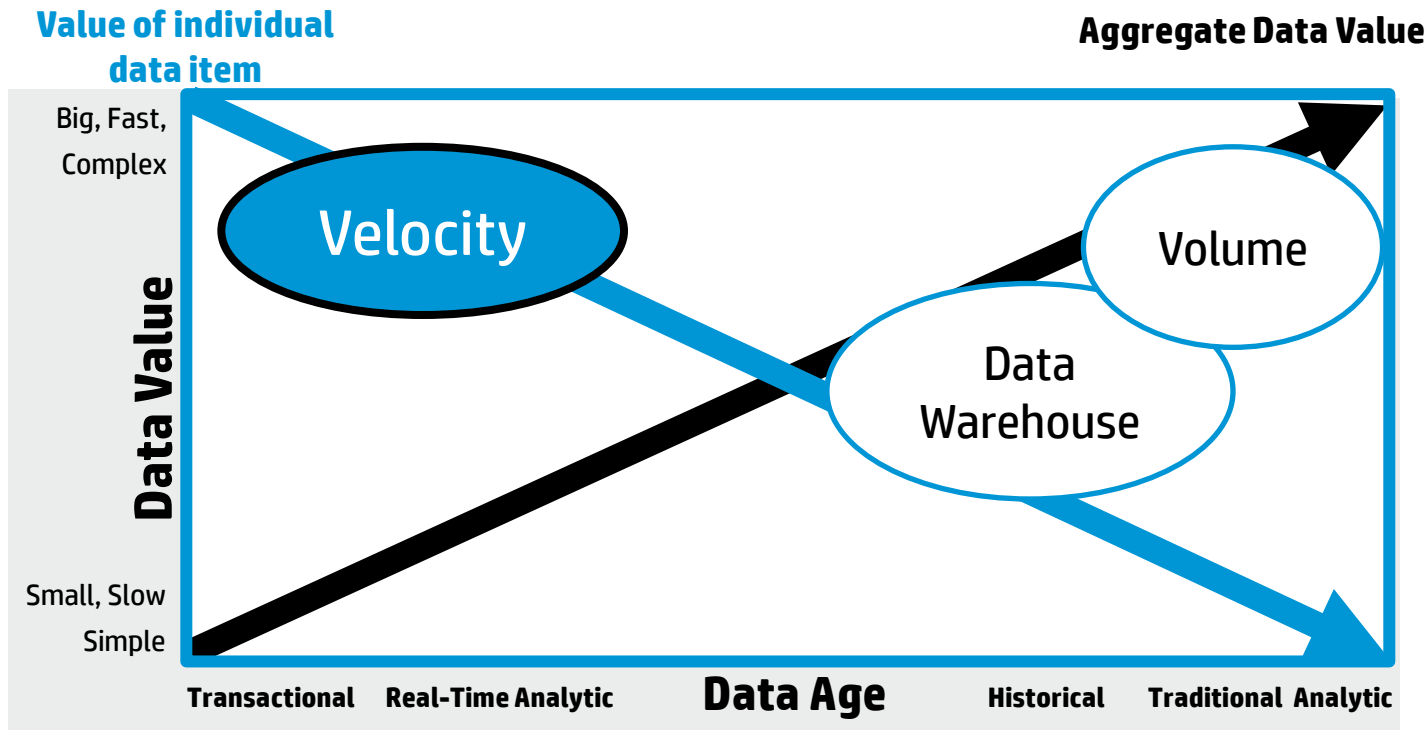% actually being tagged for Big Data Value (will grow to 33% by 2020)

0.5%

% of the Digital Universe that actually is being tagged and analyzed

Big Data Math – Quantification Characteristics

Value of individual data item

Aggregate Data Value

Big, Fast, Complex

Data Value

Velocity

Volume

Data Warehouse

Small, Slow Simple

Transactional Real-Time Analytic **Data Age** Historical Traditional Analytic

Courtesy: Dr. Michael Stonebraker, "Navigating the Database Universe"

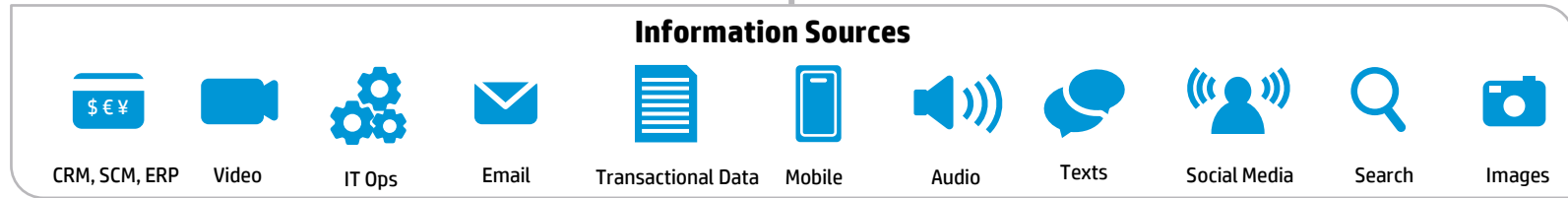# Big Data Math – Quantification Characteristics

Value of individual

Aggregate Data Value

Data Value

Big, Fast, Complex

Business Value

Business event

Business decision

Data analyzed

Information prepared

Decision latency

Business decision

Small, Slow Simple

Decision time ages data to the point where data shelf life has expired

| Transactional | Real-Time Analytic | **Data Age** | Historical | Traditional Analytic |
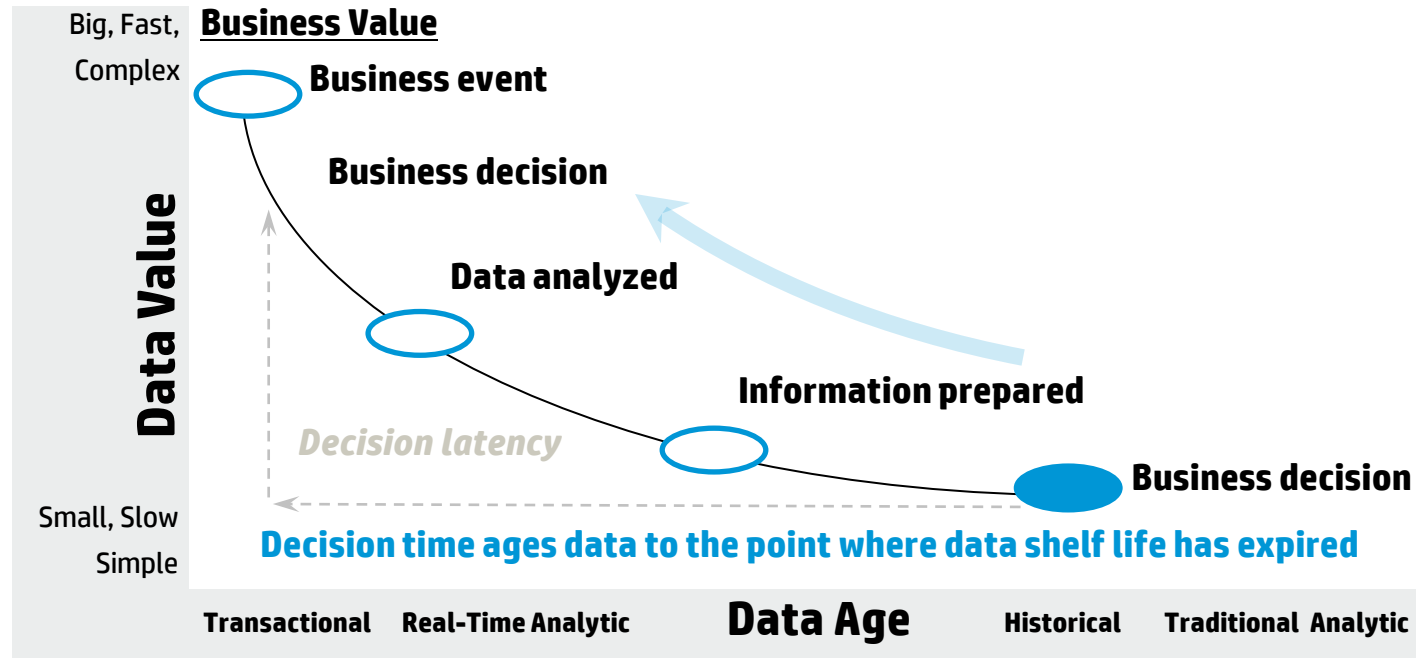
Big Data defined

Gartner: "Big data" is defined with the **3 "Vs"** as a **high-volume**, **-velocity and –variety** information assets that demand **cost-effective**, **innovative forms of information** processing **for enhanced insight and decision making**

IDC defines Big Data as being **greater than 100TB, growing at 60%+**

Velocity   **Volume**

Big Data

Variety   **Value**

Information Sources

CRM, SCM, ERP   Video   IT Ops   Email   Transactional Data   Mobile   Audio   Texts   Social Media   Search   Images

Big Data is no longer just a buzzword...

Bottom line

"Data" depends on your point of view

- **"Data Today"** will **vary from customer to customer, from application to application** and **from user to user**
- There are basically 3 types of "big data"
  - Structured
  - Semi-structured
  - Unstructured

Data

Structured + Semi-Structured + Unstructured Data

Information

Structured

| CRM |
| Order |
| ERP |
| Finance |

Semi-structured

RSS feeds	XML
Spreadsheets	Flat files in record format
Web logs	

Unstructured*

eMail text	Social networking
SMS	Audio
BLOGS	Video
Images	

90% of today's data volume
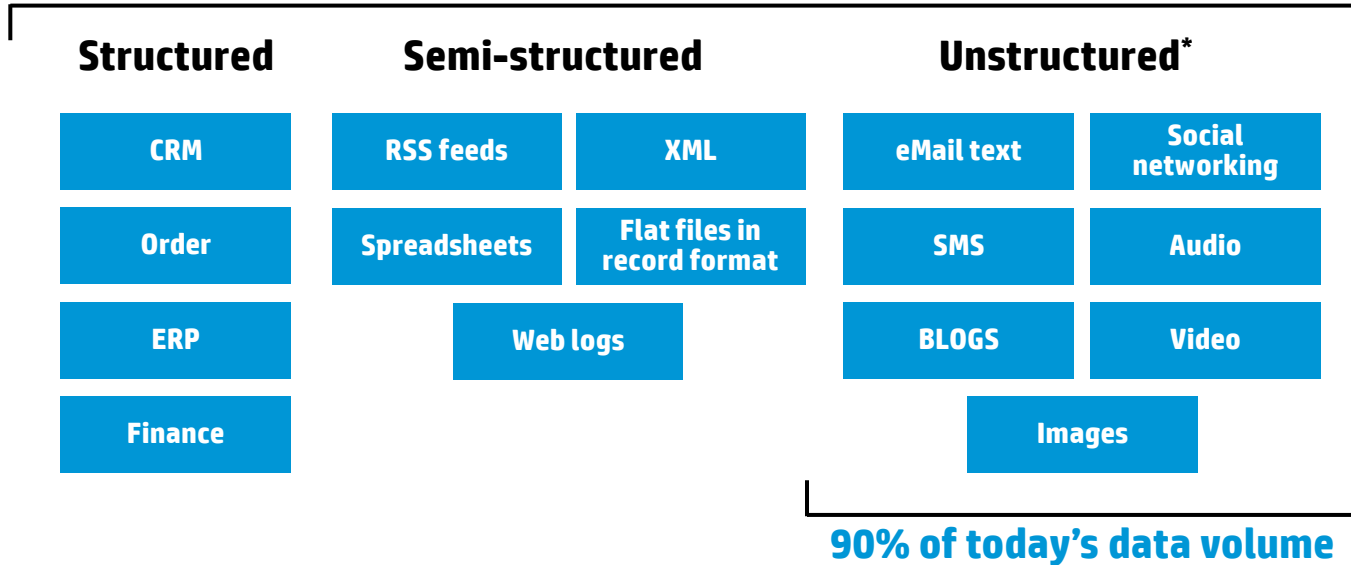
May require word-to-word and/or semantic analysis, metadata sometimes helps put into context.

The Model to Manage Extreme Information

OPERATIONAL ANALYTICS

- Real-time ODS for instant decision
- historical database for strategic analysis
- Capability to make rapid suggestions for operational actions

IN-MEMORY DBMS

- Massive data ingest
- Immediate reaction/distribution to events

NON-RELATIONAL DATA

- Sentiment analysis
- Image

CONTEXT & XML

- understanding tag context
- purpose of tag

COLUMN STORE DBMS

- Column-store DBMS is more efficient (smaller, faster and cheaper) for DW, BI, predictive and pattern-based analytics.
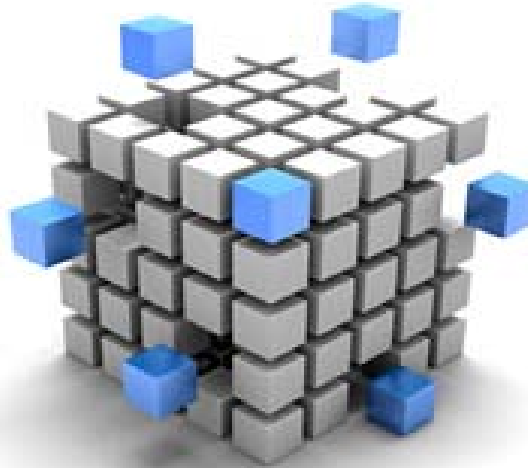
COMPLEX EVENT PROCESSING

- Aggregates information from distributed systems in real time
- Applies rules to discern patterns and trends that would otherwise go unnoticed
- React to events as they occur

Structured Data

Current Solution

Example

Accounts Payable

- Invoice receipt
- Control of mandatory elements
- Purchase order and invoice comparison
- Posting and approval
- Payment run

ERP

ERP

ERP

ERP

ERP
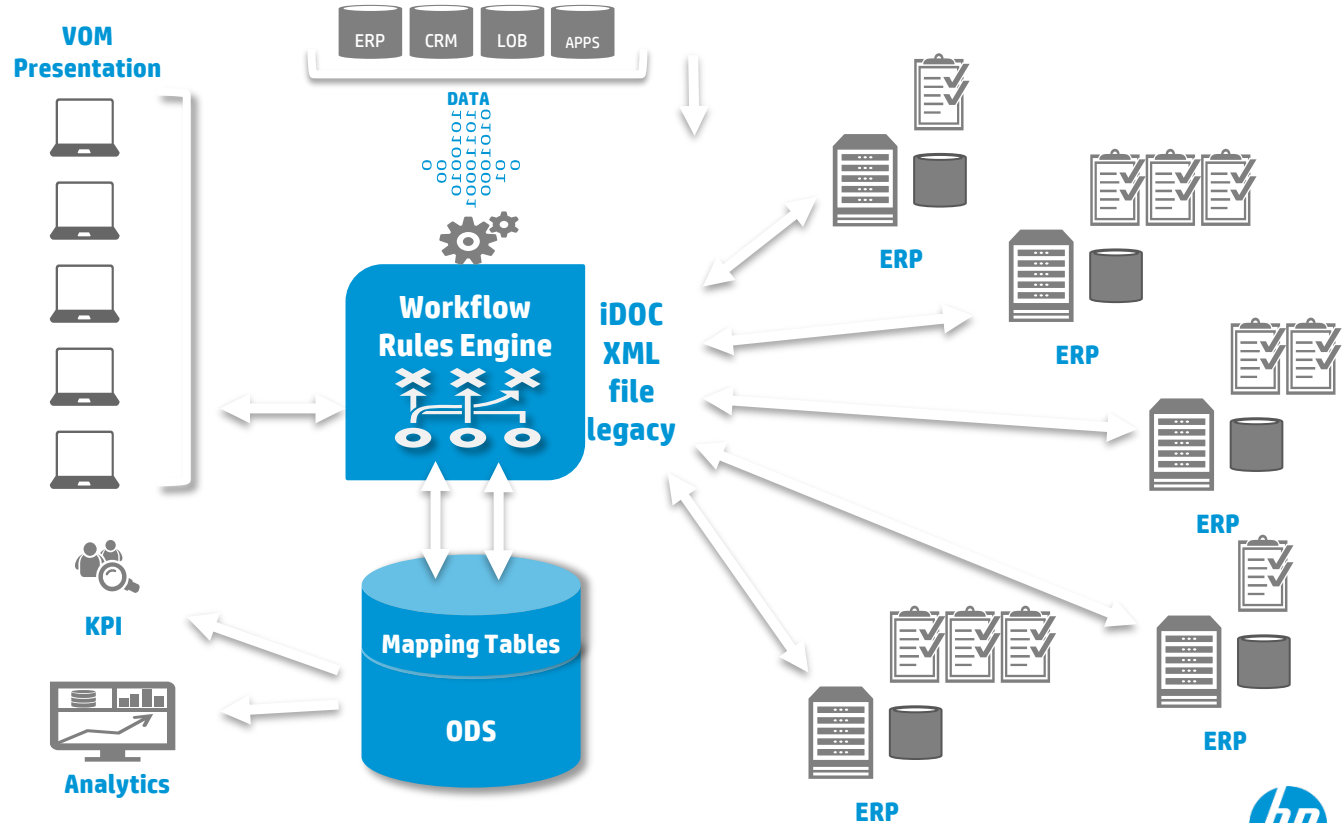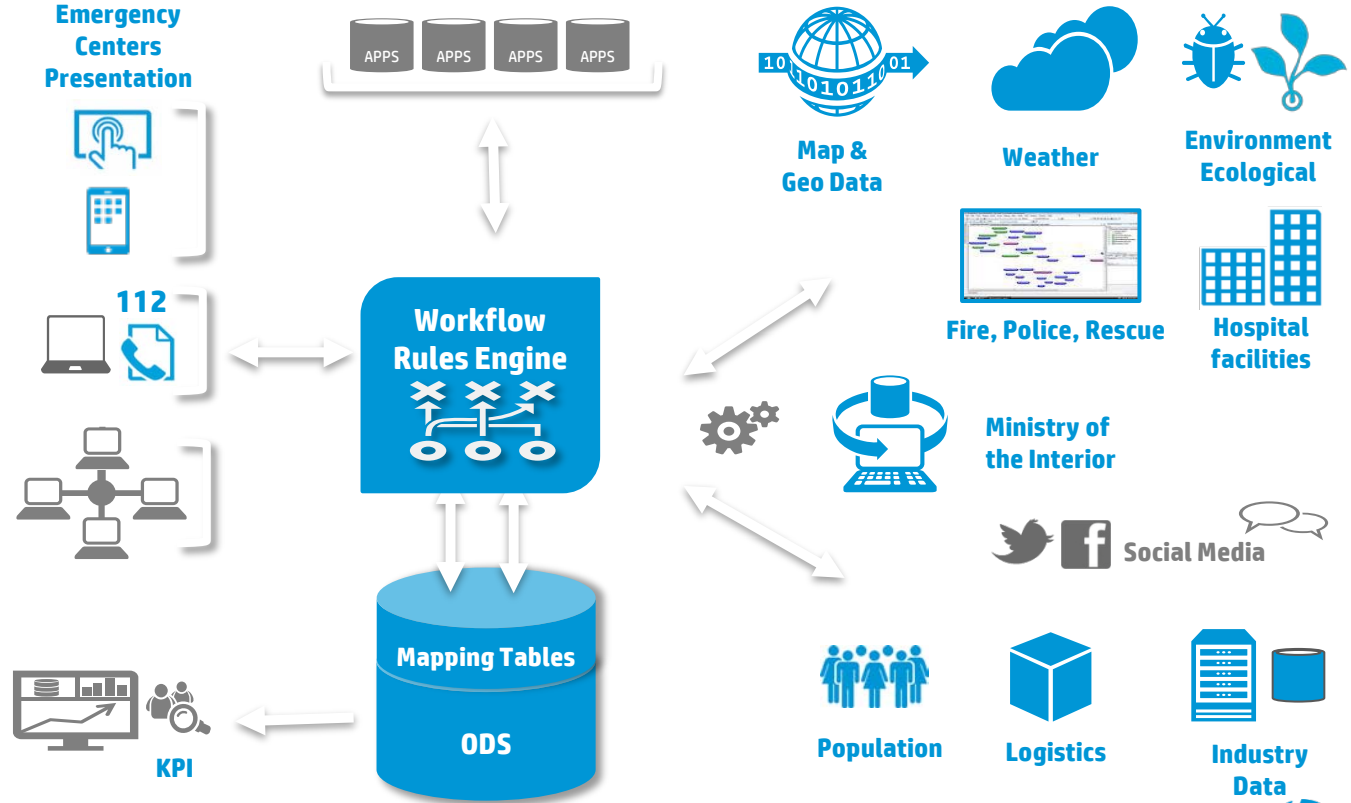
ERP
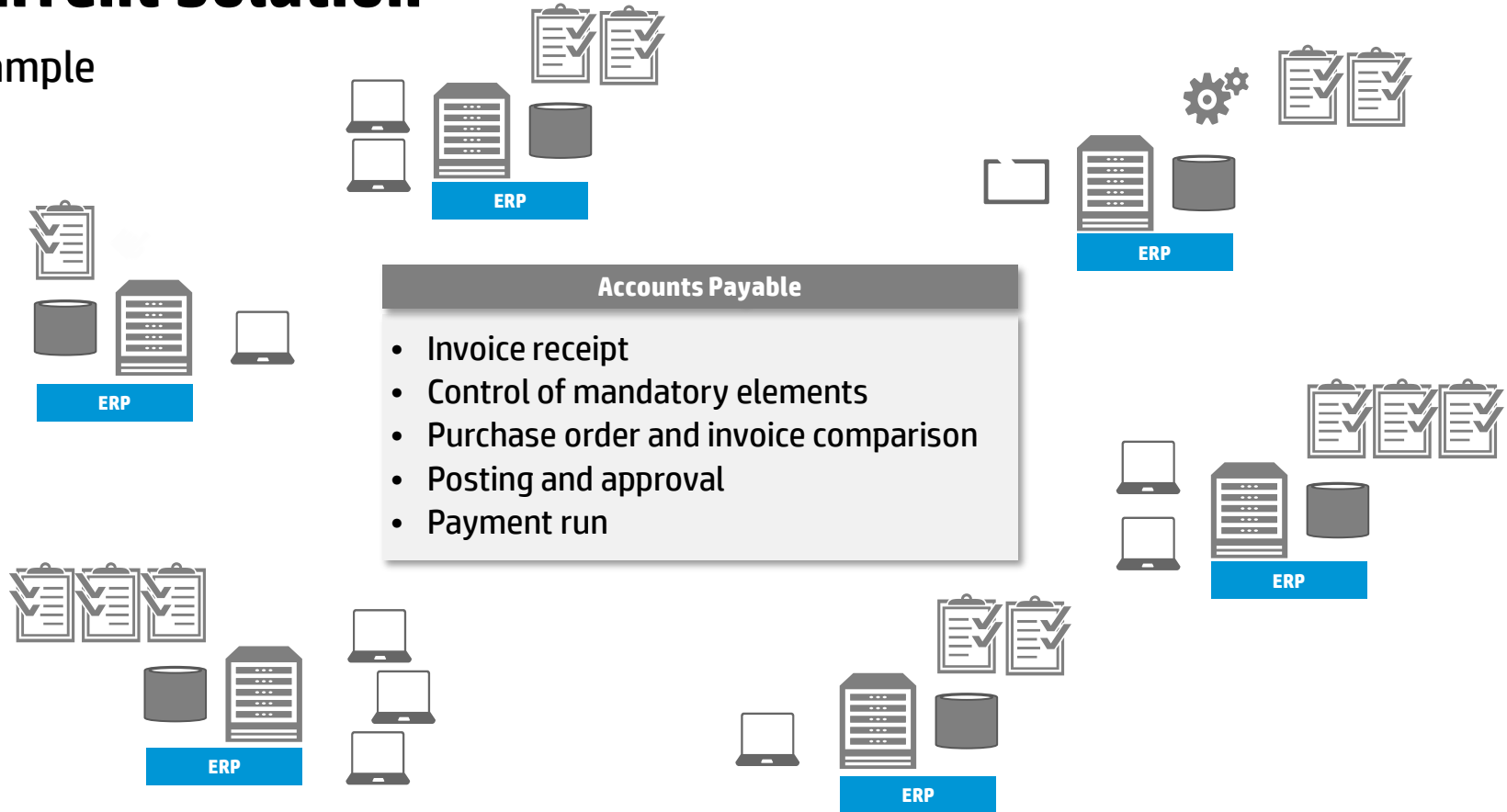
Consolidated Service Centre Data

- Integration between disparate systems
- Database caches relevant (all) information
- Maintains and delivers the information to the appropriate applications.
- Database act as a single ODS
- Eliminating the need to maintain multiple data stores
- EAI (middleware) eliminates writing custom interfaces between applications.
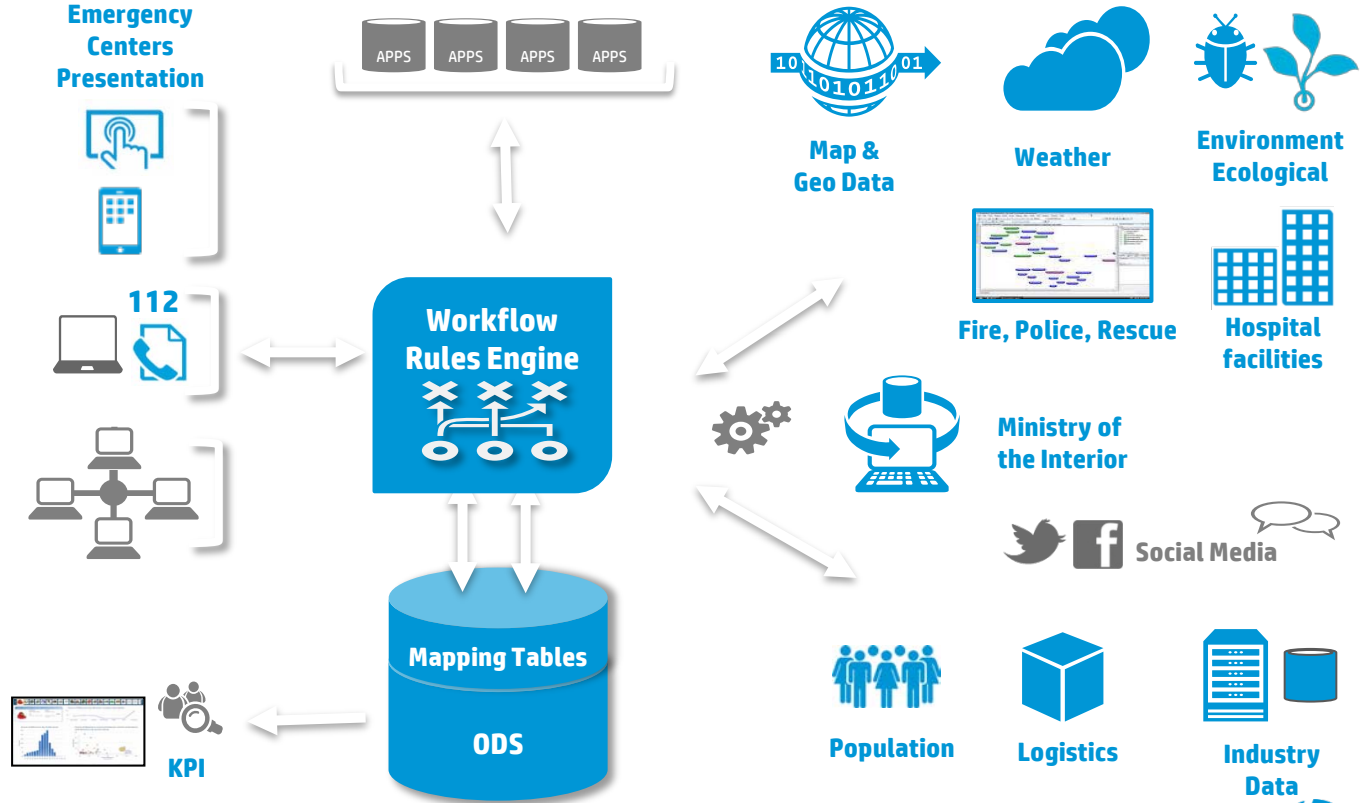
VOM Presentation

ERP CRM LOB APPS

DATA

Workflow Rules Engine

iDOC XML file legacy

Mapping Tables

ODS

KPI

Analytics

ERP

ERP

ERP

ERP

ERP

ERP

Consolidated Emergency Information

- Data from different Organisations
- Ad hoc Access to relevant information
- Customised view of information
- Database act as a single ODS
- EAI (middleware) eliminates
  - writing custom interfaces between applications
  - maintain multiple data stores
- Eliminating
  - time consuming information requests

Emergency Centers Presentation

APPS APPS APPS APPS

112

Workflow Rules Engine

Mapping Tables

ODS

KPI

Map & Geo Data

Weather

Environment Ecological

Fire, Police, Rescue

Hospital facilities

Ministry of the Interior

Social Media

Population

Logistics

Industry Data

10 1010110 01

Consolidated Emergency Information

- Data from different Organisations
- Ad hoc Access to relevant information
- Customised view of information
- Database act as a single ODS
- EAI (middleware) eliminates
  - writing custom interfaces between applications
  - maintain multiple data stores
- Eliminating
  - time consuming information requests

Emergency Centers Presentation

APPS APPS APPS APPS

112

KPI

Workflow Rules Engine

Mapping Tables

ODS

Map & Geo Data

Weather

Environment Ecological

Fire, Police, Rescue

Hospital facilities

Ministry of the Interior

Social Media

Population

Logistics

Industry Data

Operational Analytics architected for availability/scalability

HP NonStop Platform Advantages

Unrivaled Availability

Reliable and failure resilient hardware

Patented fault-tolerant software "process-pair" technology
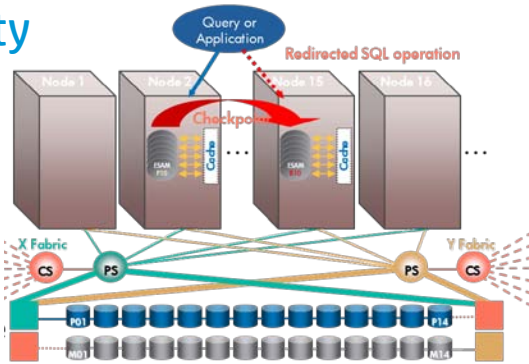
Shared-nothing MPP

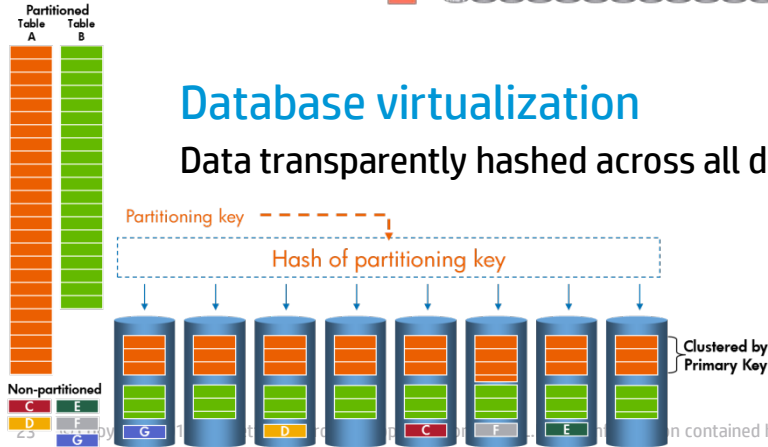16 nodes per segment

Dual active (X,Y) interconnect fabrics

Dual cluster switches for inter-segment I/O
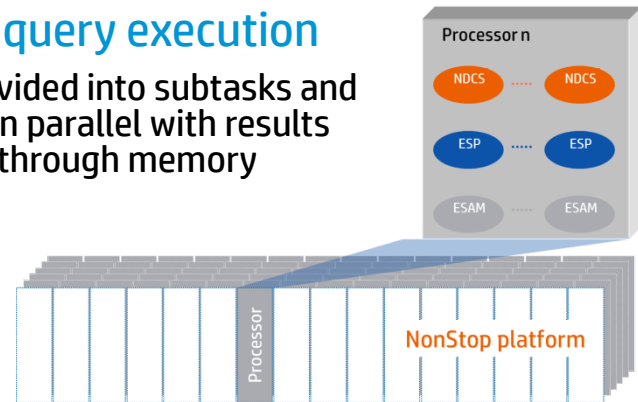
End-to-end disk checksum integrity
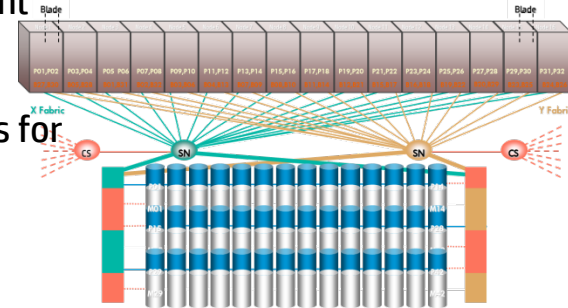
Database virtualization

Data transparently hashed across all disks

Parallel query execution

Queries divided into subtasks and executed in parallel with results streamed through memory

Unstructured Data

What is unstructured data?

Unlike structured data, which is organized, has clearly defined relationships, **unstructured data is free-flowing, disorganized**, and needs new technologies to run analytics on it

Structured Data: Example:*

Name	Title	Company	Business Unit
Dave Donatelli	EVP & GM	HP	ESSN
Bob Muglia	President	Microsoft	Server and Tools
Diane Bryant	CIO	Intel	Corporate
Jeff Benck	COO & EVP	Emulex	Corporate

Unstructured data Example:*

[…] last session for the day was the **HP ProLiant and BladeSystem Super Session**. The presenters were Bill Swales (Vice President, HP Americas Industry-Standard Servers & Software) and Mark Potter (Senior Vice President & General Manager, HP Industry-Standard Servers & Software).
Bill began the session by talking about the problems he is hearing from customers in America and how they can benefit […]

Characteristics:

- Organized in tables
- Stored as records in a database
- Well-defined relationships between data field
- **Typically used for data warehousing and analytics**
- Widespread in usage today

Characteristics:

- Free-flowing, unorganized, and unstructured
- Stored as file and blobs
- Little or no defined relationship
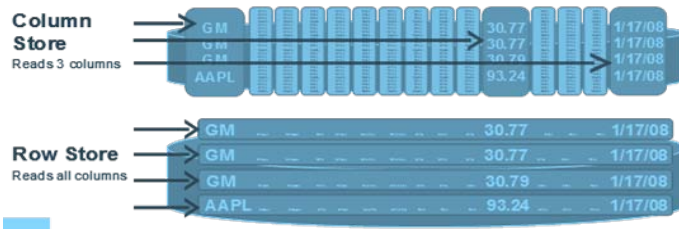- **Need unstructured data analytics platforms** Emerging and fast growing space
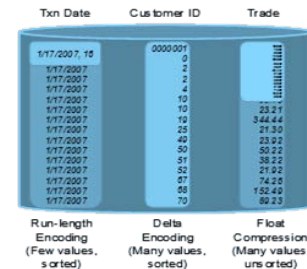
Vertica – Ad-Hoc Analytics

Column Storage

- Store data the way it's queried for the best performance

- Ideal for real-intensive workloads

- Dramatically reduces disk IO


Capacity Optimization

- Store more data, provide more views, use less hardware

- More compression options when similar data is grouped

- 12 compression schemes
  - Dependent upon data
  - System chooses which to apply
  - NULLs take virtually no space

- Typically 50 – 90% compression

- Queries data in encoded form

Vertica – Ad-Hoc Analytics

Clustering

- "scale out" infinitely by adding more hardware
- Columns are duplicated, if a node is down there is still a copy
- New hardware queries rest of system for data it needs
  - Rebuilds missing objects from other nodes
  - Benefit from multippe sort orders
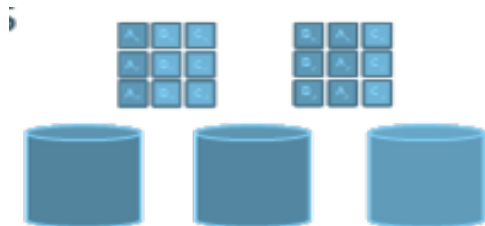
Continuous Performance

- Concurrent loading and querying
  - Get real-time views and eliminate nightly "load windows"
- On-the-fly scheme changes
  - Add columns, projections without downtime
- Automatic data replication, failover and recovery
  - Active redundancy increase performance
  - Nodes recover automatically by querying the system

HP Hadoop value proposition

Pan-HP portfolio of solutions, services, and tools for big data

Ease of acquisition and time to value

- Pre-tested, pre-configured reference architectures
- Factory integrated appliances
- HP Consulting Services
  - Roadmap Service for Hadoop
  - HP Big Data Discovery Workshop

Simplified management, risk-free scalability

- Insight Cluster Management Utility (CMU) systems management and monitoring
- Vertica connectors
- HP expertise with large scale-out clusters

Reliable solutions based on HP CI and open standards

- Choice of top Hadoop distributions
- Optimized ProLiant x86
  - Gen8 (iLO, SmartArray, DL300 series)
  - HPN A5830 switches
- End to end Vertica analytics
- HP open source commitment to Apache Hadoop for 4+ years

What is Apache Hadoop?

Apache Hadoop is an open source platform for data storage and processing that is...

- ✓ **Scalable**
- ✓ **Fault tolerant**
- ✓ **Distributed**

CORE HADOOP SYSTEM COMPONENTS

Hadoop Distributed File System (HDFS)

Self-Healing, High Bandwidth Clustered Storage

+

MapReduce

Distributed Computing Framework

Has the Flexibility to Store and Mine <u>Any</u> Type of Data

- Ask questions across structured and unstructured data that were previously impossible to ask or solve
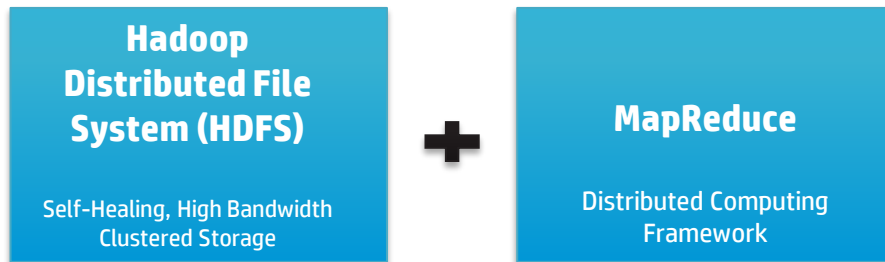
- Not bound by a single schema

Excels at Processing Complex Data

- Scale-out architecture divides workloads across multiple nodes

- Flexible file system eliminates ETL bottlenecks

Scales Economically

- Deployed on Readily Available, Industry Standard Servers

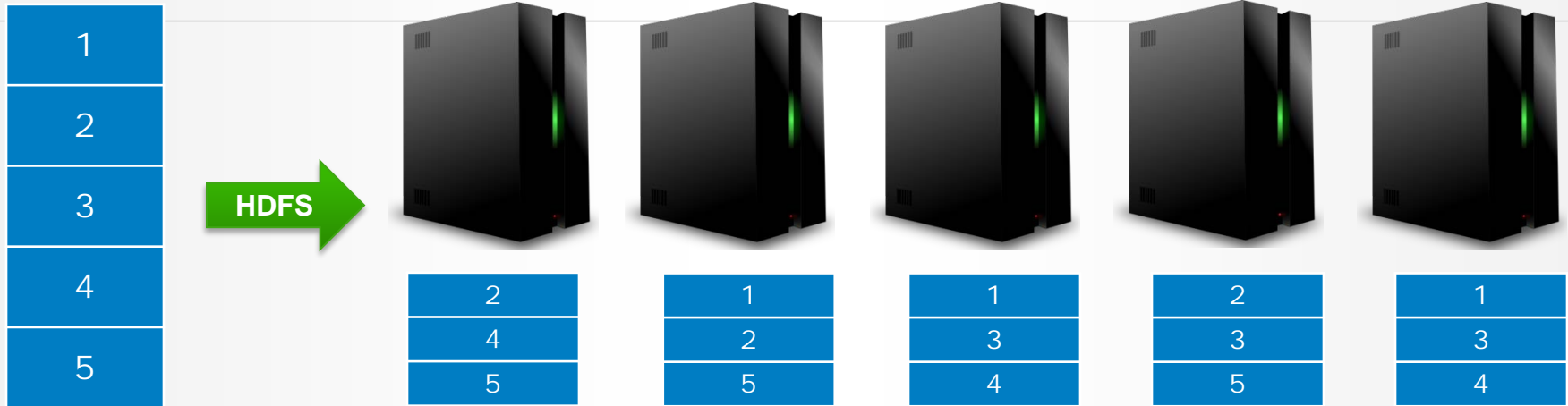- Open source platform guards against vendor lock

Core Hadoop: HDFS

Self-healing, high bandwidth **clustered storage.**

1		
2		
3	**HDFS** →	
4		
5		

2	1	1	2	1
4	2	3	3	3
5	5	4	5	4

HDFS breaks incoming files into blocks and stores them redundantly across the cluster
More customers run HP clusters than any other platform!

Core Hadoop: MapReduce

Distributed scalable computing framework (library and runtime) for analyzing data sets stored in HDFS
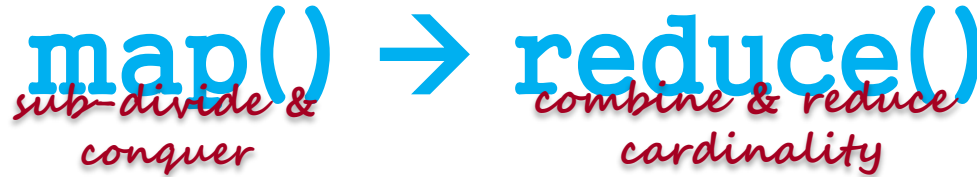

- **Model for Processing large jobs in parallel across many nodes and combines the results**
- **Provides all the "glue" and coordinates the execution of the Map and Reduce jobs on the cluster**

Hadoop MapReduce (MR) Jobs

- MapReduce jobs are composed of two functions:

$$map() \rightarrow reduce()$$

sub-divide & conquer   *combine & reduce cardinality*

- The map phase organize the data in preparation for the processing done in the reduce phase
  - Reads the input in parallel and distributes the data to the reducers
- MR framework provides all the "glue" and coordinates the execution of the Map and Reduce jobs on the cluster
- HDFS and MapReduce combined functionality:
  - Parses and indexes the full range of data types
  - Manages diverse types of data, documents, content, and schema
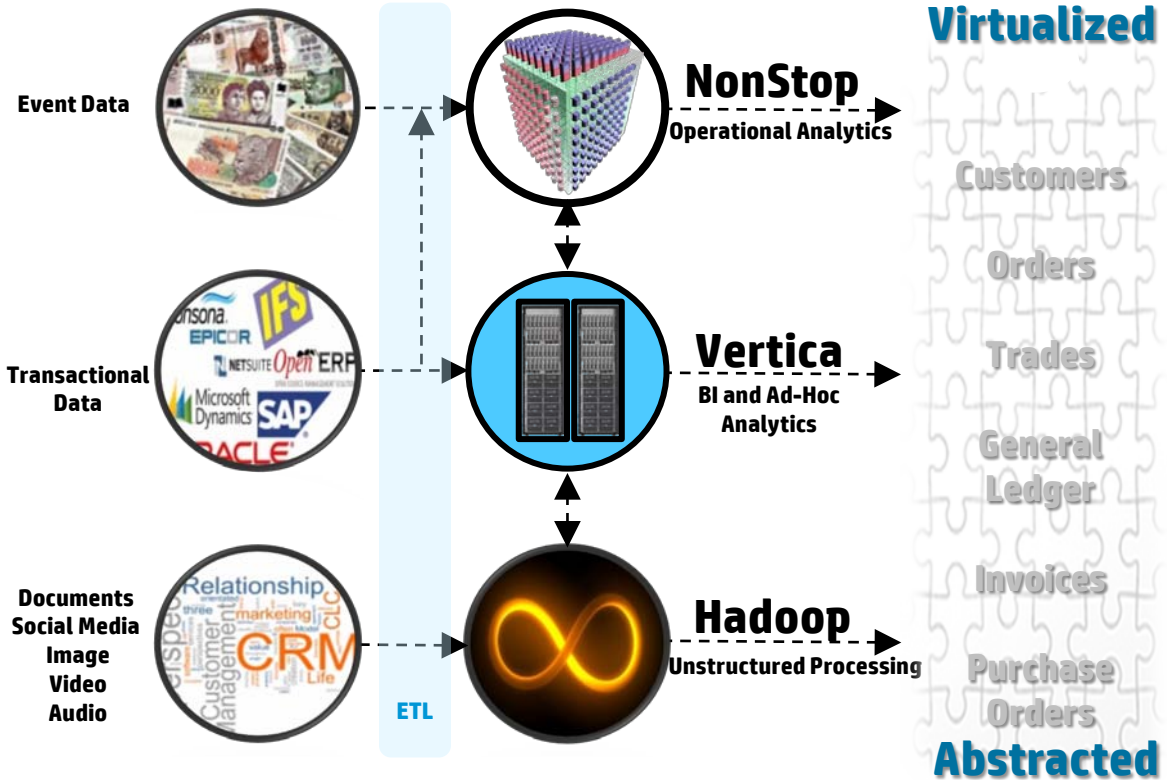- **Written in Java Language**

HP Logical Data Warehouse
HP Analytical Clooud

An HP Logical Data Warehouse

- **Distribute data and compute to the LDW node that best meets the requirements of the application or workload— price-performance, availability, data sensitivity, etc.**

- **Targeted Scalability**

- **Continuously available**

Event Data

Transactional Data

Documents
Social Media
Image
Video
Audio

ETL

Virtualized

NonStop
Operational Analytics

Vertica
BI and Ad-Hoc Analytics

Hadoop
Unstructured Processing

Customers

Orders

Trades

General Ledger

Invoices

Purchase Orders

Abstracted

Process   Enterprise Meta & Control Data Mgmt   Data Profiling   Validation & Cleansing   Interface Specifications & Mgmt   Data Quality & Stewardship Mgmt
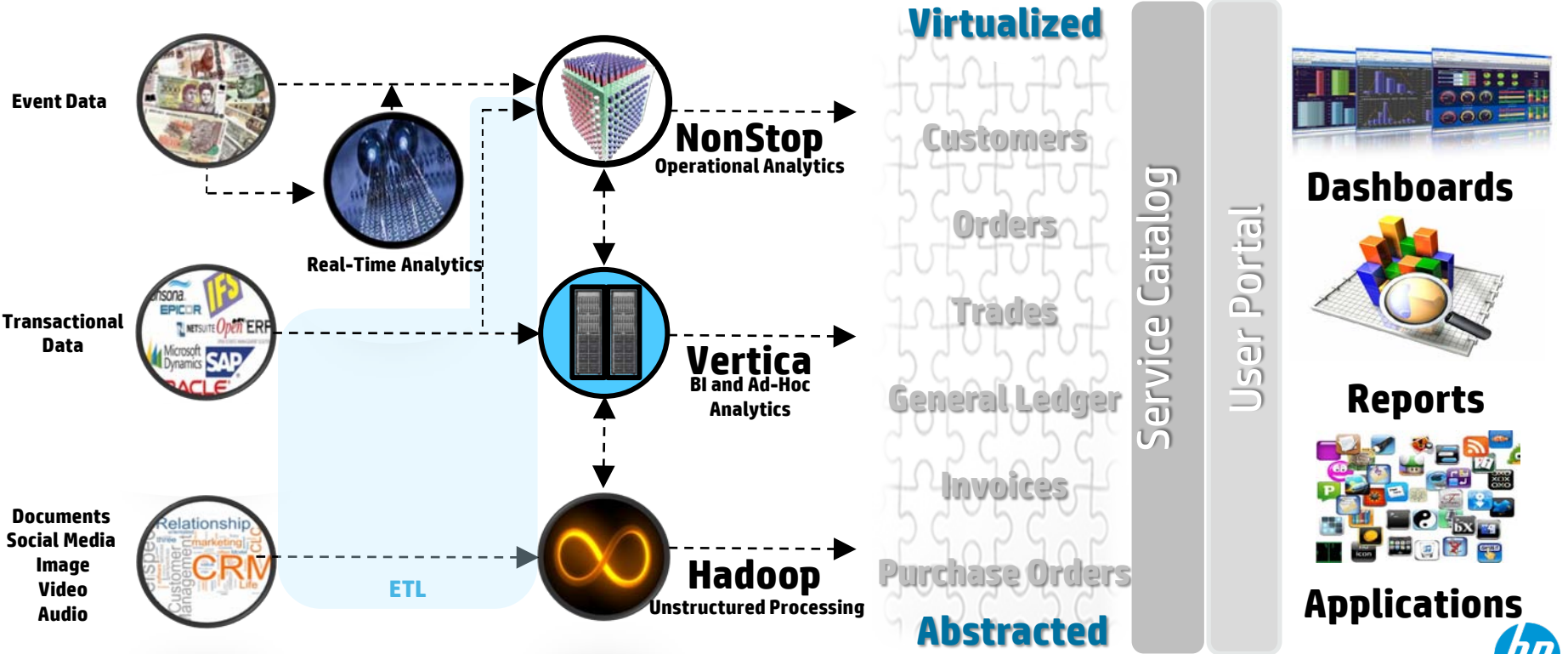
An HP Analytic Cloud

a Logical Data Warehouse based on a Time Continuum

Data | Create | Consolidate | Consume

Virtualized

Event Data

Real-Time Analytics

NonStop
Operational Analytics

Transactional Data

Vertica
BI and Ad-Hoc Analytics

Documents
Social Media
Image
Video
Audio

ETL

Hadoop
Unstructured Processing

Customers

Orders

Trades

General Ledger

Invoices

Purchase Orders

Abstracted

Service Catalog

User Portal

Dashboards

Reports

Applications

Process | Enterprise Meta & Control Data Mgmt | Data Profiling | Validation & Cleansing | Interface Specifications & Mgmt | Data Quality & Stewardship Mgmt

HP is leading in Big Data innovations

Fastest Time to Value; Purpose Built for Big Data Scale and Performance

HP solutions deliver more choice to meet specific workloads , data volumes and variety versus our competitors' "one-size-fits-all" approach

AppSystem for **Apache Hadoop**

AppSystem for **SAP HANA**

AppSystem for **Vertica**

AppSystem for **Autonomy**

NonStop System

HP approach helps deliver right-time intelligent business decisions

STORE

- **Intelligent Design** for performance and high availability

- **Intelligent Scaling by Workload** for high efficiency scale up or seamless scale out

- **Intelligent delivery models** for on-premise, cloud, or hybrid cloud

MANAGE

- **Intelligent Management** for a holistic management, governance and security

- **Intelligent Culling (Prioritization)** for separating signal from noise and optimizing resources to act on useful data

- **Intelligent Workload Characterization** for optimizing Fit-for-purpose platforms for transactional, warehouse or analytic workloads which can significantly vary in memory, processor, storage and networking requirements

UNDERSTAND and ACT

- **Intelligent pan-enterprise functionality** for multichannel management and delivery

- **Intelligent purpose-built scalable analytic platforms** for advanced, predictive and real-time analytics

HP solutions deliver more choice to meet specific workloads , data volumes and variety versus our competitors' "one-size-fits-all" mentality. NonStop is your best Operational Analytic choice.

Thank you