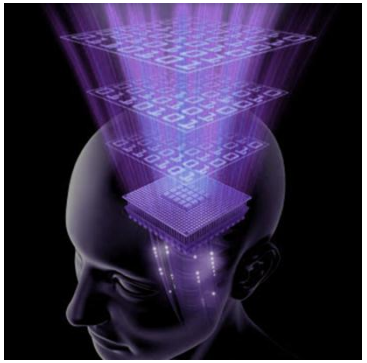
# Data Warehousing, Advanced Analytics and Big Data

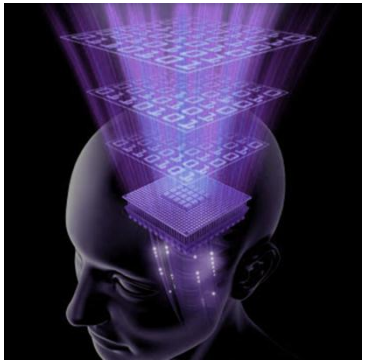
*What's changing and why it matters…*

***Kevin Dixon**, Vicom Infinity*
*kdixon@vicominfinity.com*

***Twitter:** KDixon_VI*



- **Big Data and Analytics**
- Infrastructure
- Architectural Strategies
- IBM Data Warehouse and Analytics Solutions

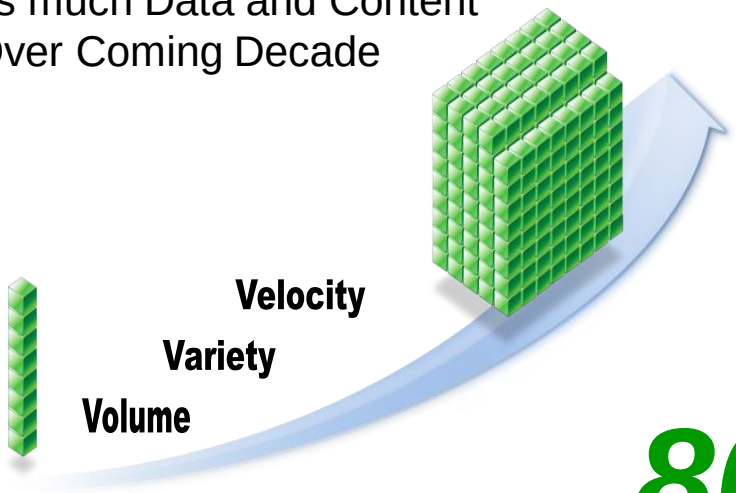


# Information is at the Center of a New Wave of Opportunity…

**44x**

as much Data and Content Over Coming Decade

**2020**
***35 zettabytes***




**2009**
***800,000 petabytes***

**80%**

Of world's data is unstructured

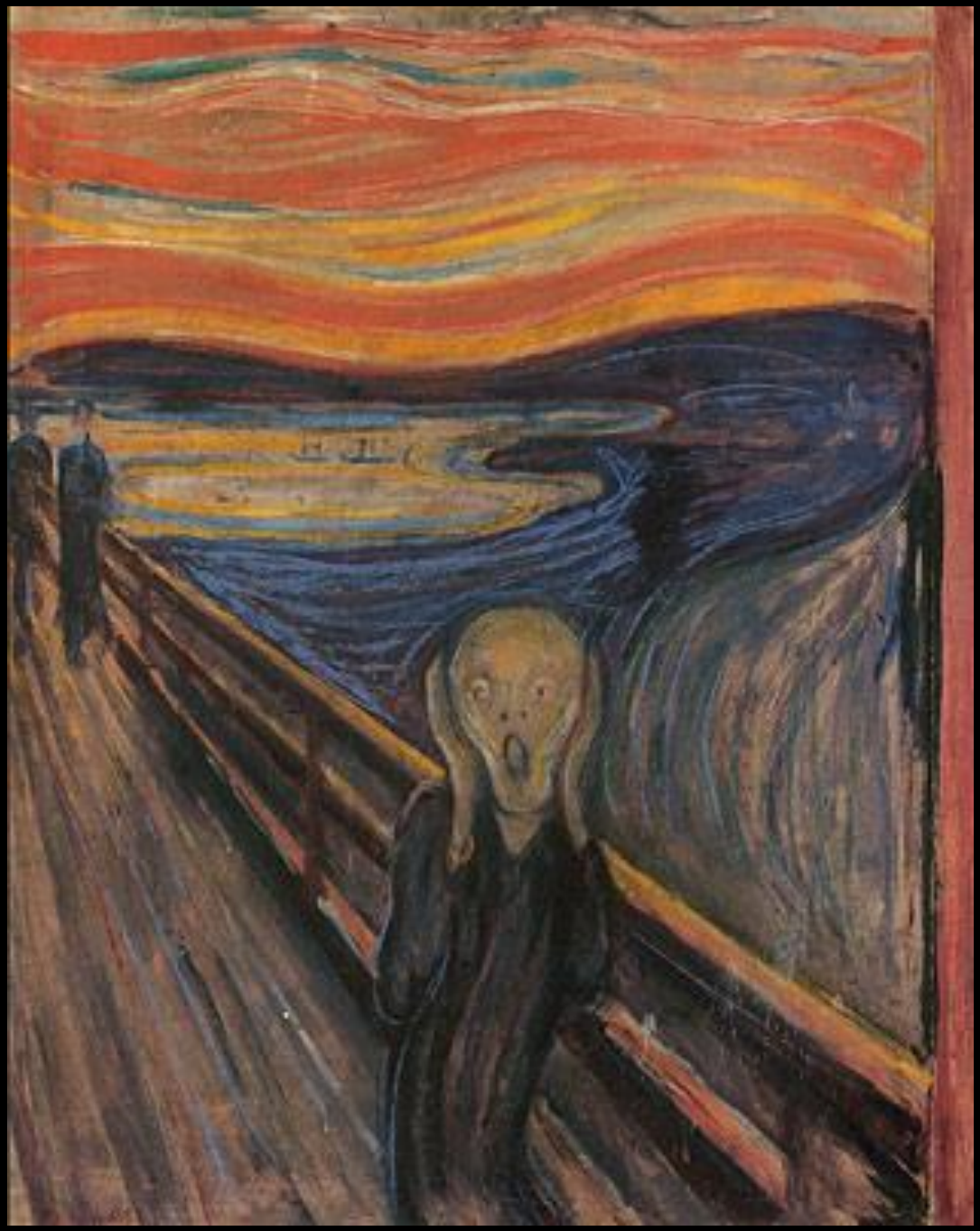
# … and Organizations need Deeper Insights

**1 in 3** Business leaders frequently make decisions based on information they don't trust, or don't have

**1 in 2** Business leaders say they don't have access to the information they need to do their jobs

**83%** of CIOs cited "Business intelligence and analytics" as part of their visionary plans to enhance competitiveness

**60%** of CEOs need to do a better job capturing and understanding information rapidly in order to make swift business decisions



**Big data** is data that exceeds the processing capacity of conventional database systems.

The data is too big, moves too fast, or doesn't fit the strictures of your database architectures.

To gain value from this data, you must choose an alternative way to process it.

**Edd Dumbill, O'Reilly** [1]

# The Big Data Opportunity

Variability


Value

| | |
|---|---|
| **Variety:** | Manage the complexity of multiple relational and non-relational data types and schemas |
| **Velocity:** | Streaming data and large volume data movement |
| **Volume:** | Scale from terabytes to zettabytes |

**Big data** used to be a technical problem.

Now it's a business opportunity.

***Philip Russom, TDWI report*** [2]

# Where to start?


**What is the problem?**

**What type of data?**

**How will it be used?**

# What is the Problem?

**Line of Business**

- I can't analyze **ALL my data** – I have to sample or summarize
- I have a report that takes **three days** to run
- My analyses are conducted on **stale and outdated** data
- I need to **involve IT** for every new report or query
- I need to analyze **new types of data**
- I need to use data for **competitive business advantage**

**IT**

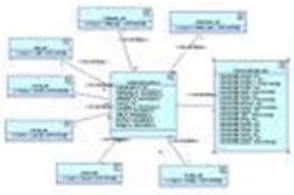
- I cannot keep up with **growing data, users** and applications
- We **regularly miss SLAs** for data freshness/availability
- Ad-hoc and analytic **queries take too long** or just not possible
- I have a **backlog of pending applications** projects
- I need to do more with less

# What types of Data?

**Poly-structured…**

**… multi-structured**

## Structured data



Data that resides in fixed fields within a record or file. *Relational* databases and spreadsheets are examples of structured data.

## Semi-Structured Data

Semi-structured data is a form of structured data that does not conform with the formal structure of tables and data models associated with relational databases but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.

- XML, other markup languages, (sometimes e-mail)
- Weblogs (e.g., page views, impressions, click-throughs, etc.)

## Unstructured Data


Data that does not necessarily following any format or sequence, does not follow any rules, is not predictable…

- Bitmap Objects: image, video or audio files
- Textual Objects: documents, e-mail, etc.

# How will it be used?

## Transactional processing

- *Traditional OLTP*
- *Large number of short, discrete, atomic transactions*
- *High throughput (transactions per second)*
- *Maintains data integrity in multi-user environments*

## Analytics processing

- *Fewer users, fewer requests*
- *Business analysts rather than customers and POS operators*
- *Queries can be very complex and resource-intensive.*
- *Response time is frequently measured in tens to hundreds of seconds.*

Structured | Semi-structured | Unstructured

What's the difference between **Business Analytics** and **Business Intelligence**?

The correct answer is: everybody has an opinion, but nobody knows, and you shouldn't care.

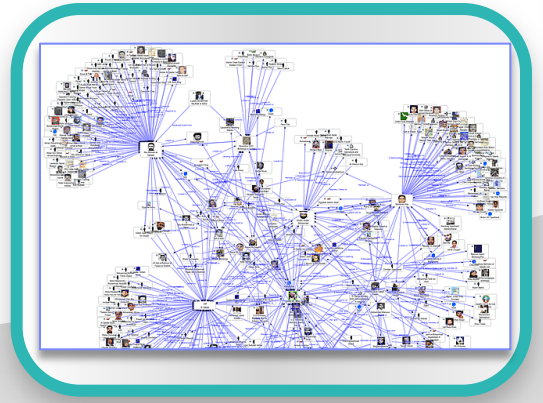
***Timo Elliot, SAP*** [3]

# The Spectrum of Analytics

## BI Reporting and Ad-Hoc Analysis

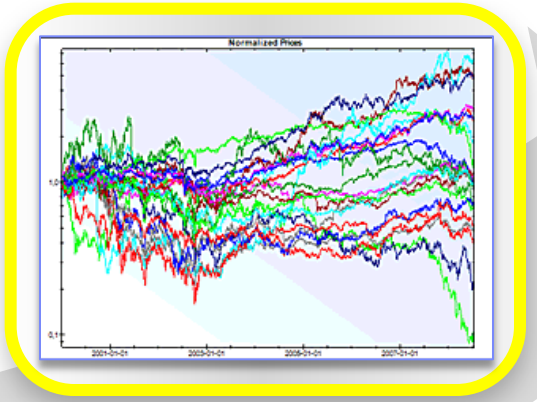

- What happened?
- When and where?
- How much?

## Predictive and/or Advanced Analytics



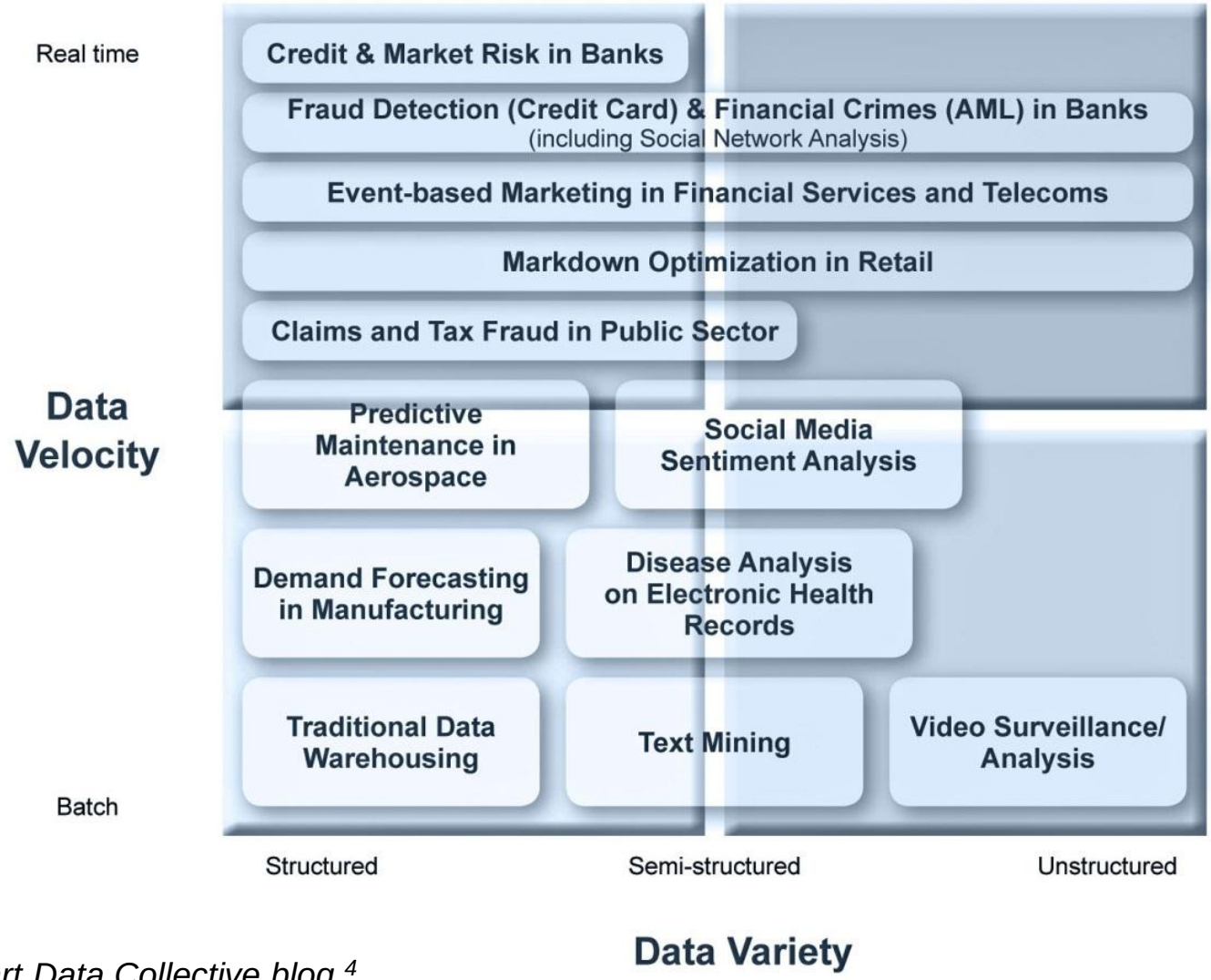
- What will happen?
- What will the impact be?

## Optimization




- What is the best choice?

# Sample Use Cases for Big Data Analytics



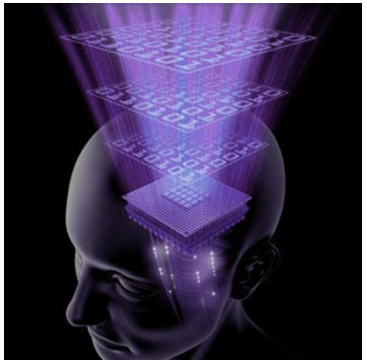

** from Smart Data Collective blog [4]*

Instead of “advanced analytics,” a better term would be **“discovery analytics,”** because that’s what users are trying to accomplish.

In other words… the user is typically a business analyst who is trying to discover new business facts that no one in the enterprise knew before.

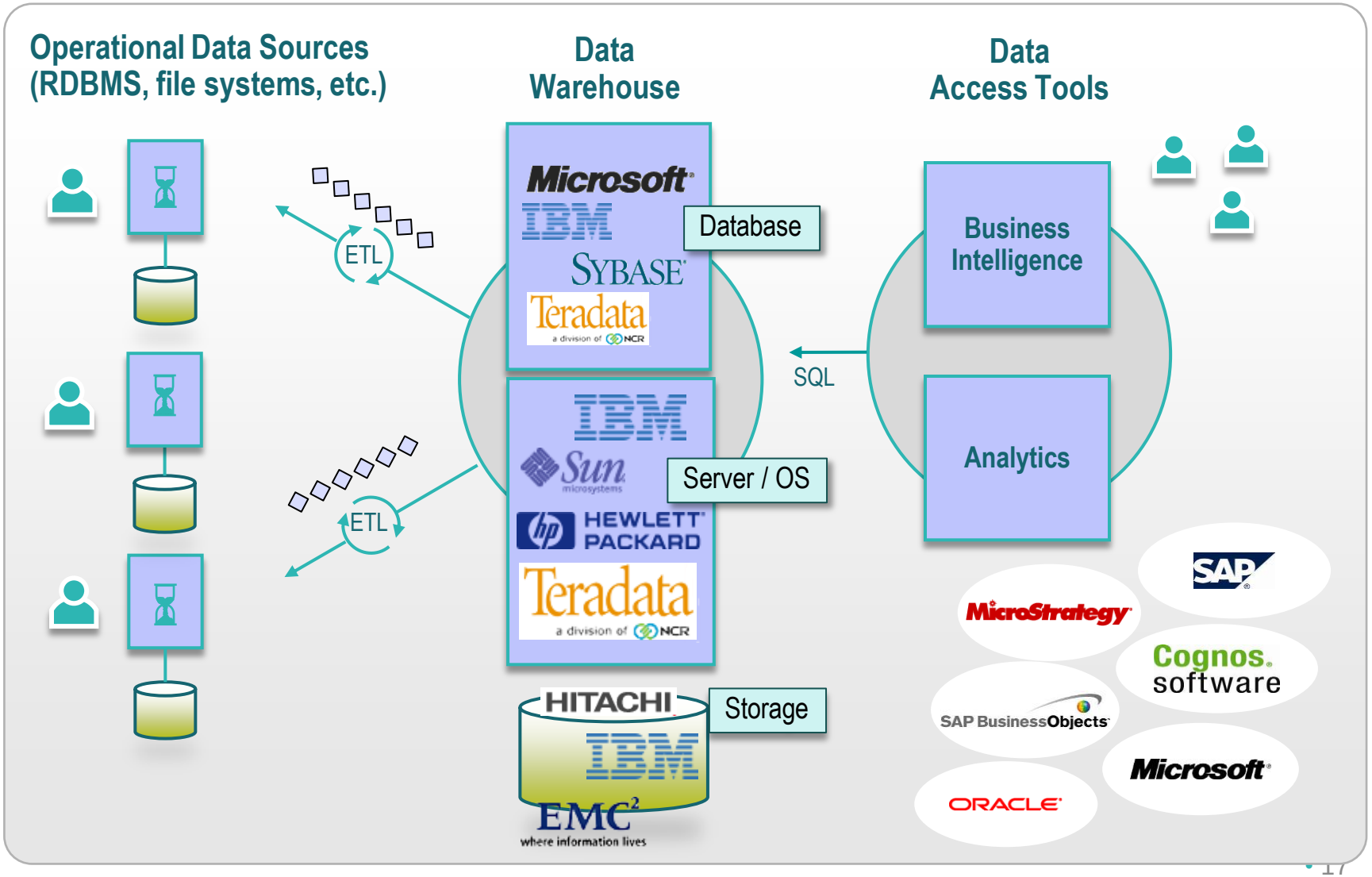
***Philip Russom, TDWI Report*** [5]

- Big Data and Analytics
- **Infrastructure**
- Architectural Strategies
- IBM Data Warehouse and Analytics Solutions



# Traditional Data Warehouse Architecture

***Move the Data to the Query, Resulting in Significant I/O Bottlenecks***

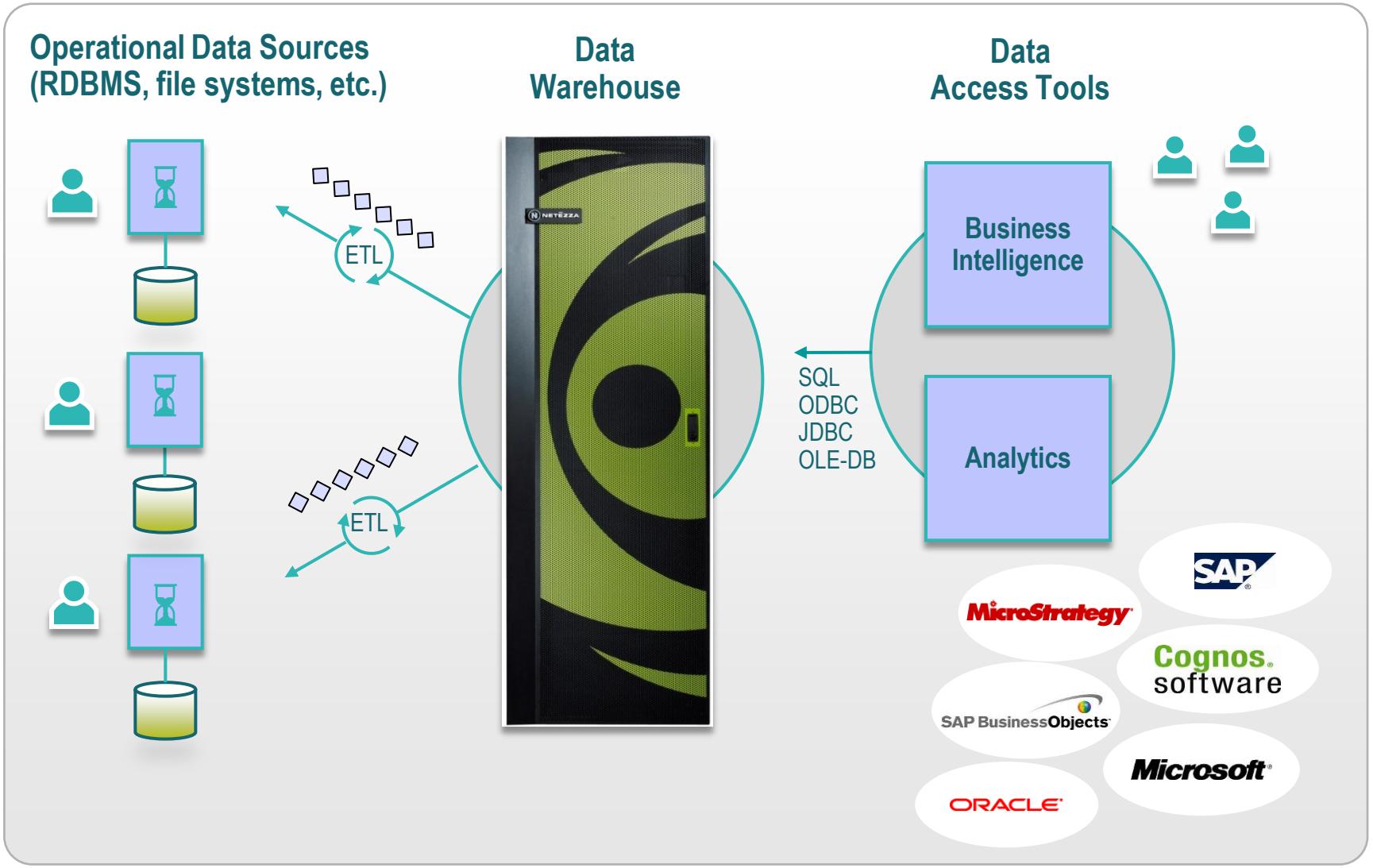


# Various Infrastructure Solutions

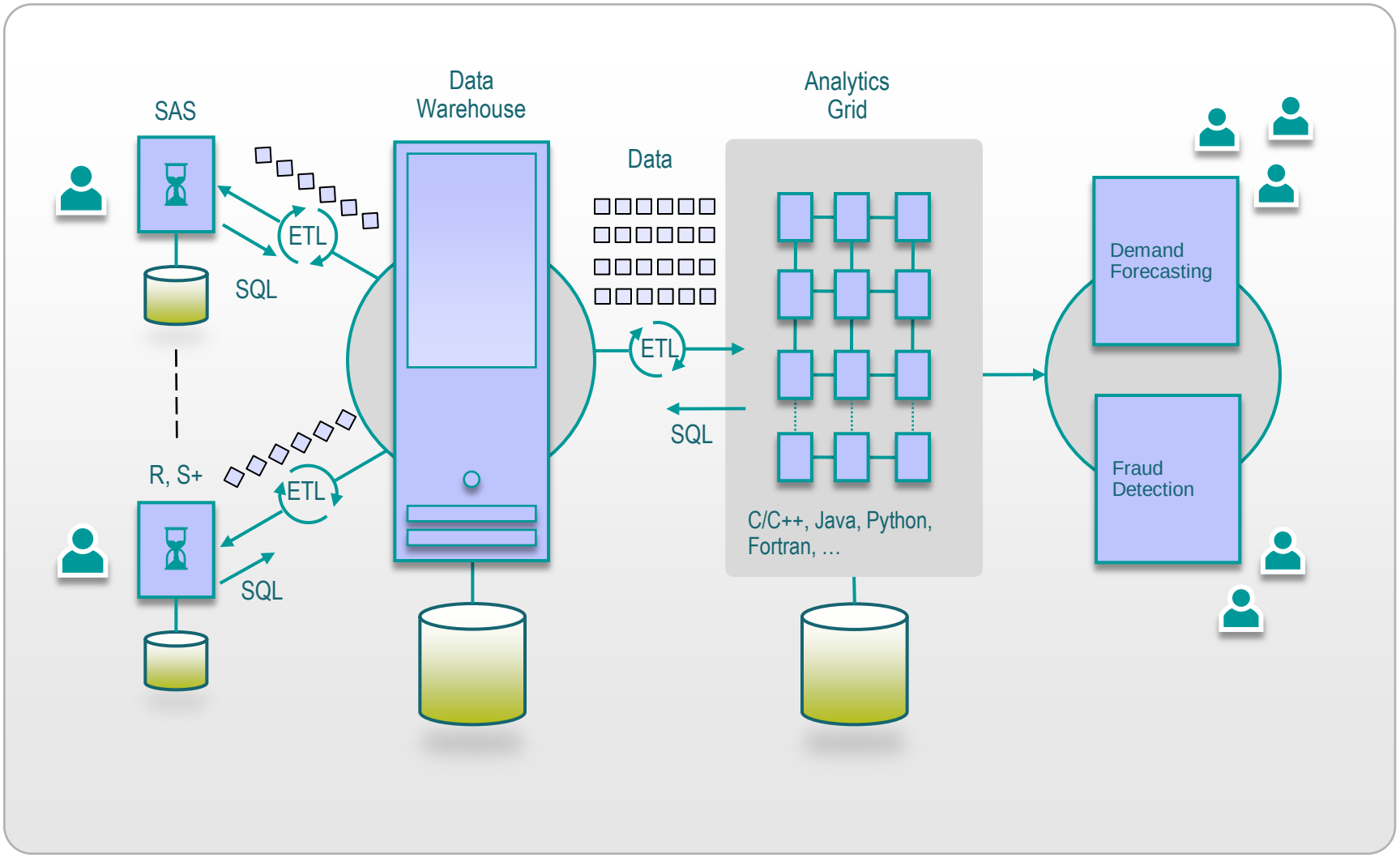
- RDBMS-based Software Data Warehouse
  - *Includes traditional as well as newer technologies*
- Integrated Optimized Systems
- Data Warehouse Appliances
  - *Emphasis on cost, ease of use, performance*
- Hadoop / MapReduce
- NoSQL and other Open Source solutions
- *Cloud / SaaS solutions*

# Data Warehouse with IBM Netezza Appliance

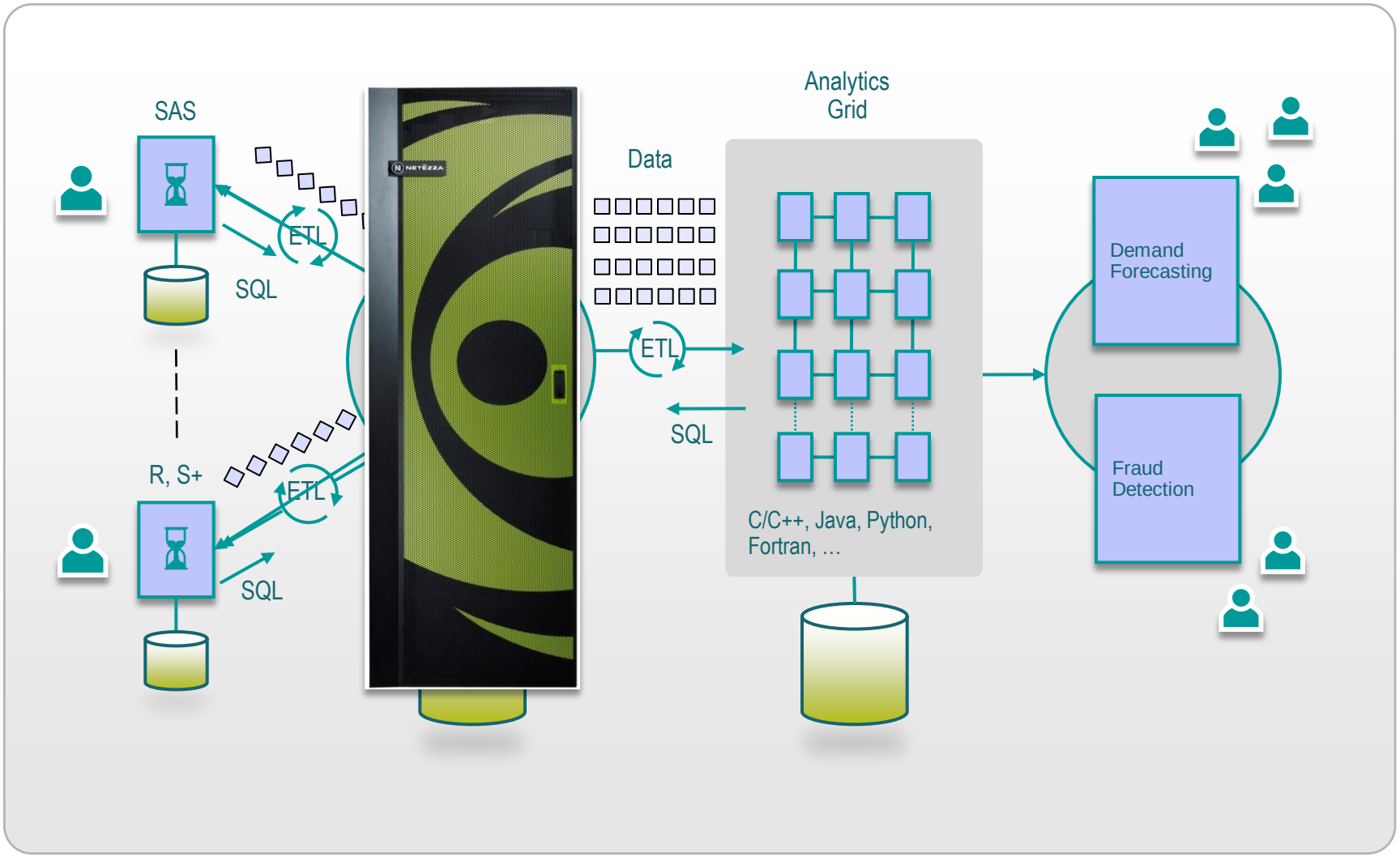
***Move the Query to the Data to eliminate I/O limitations***



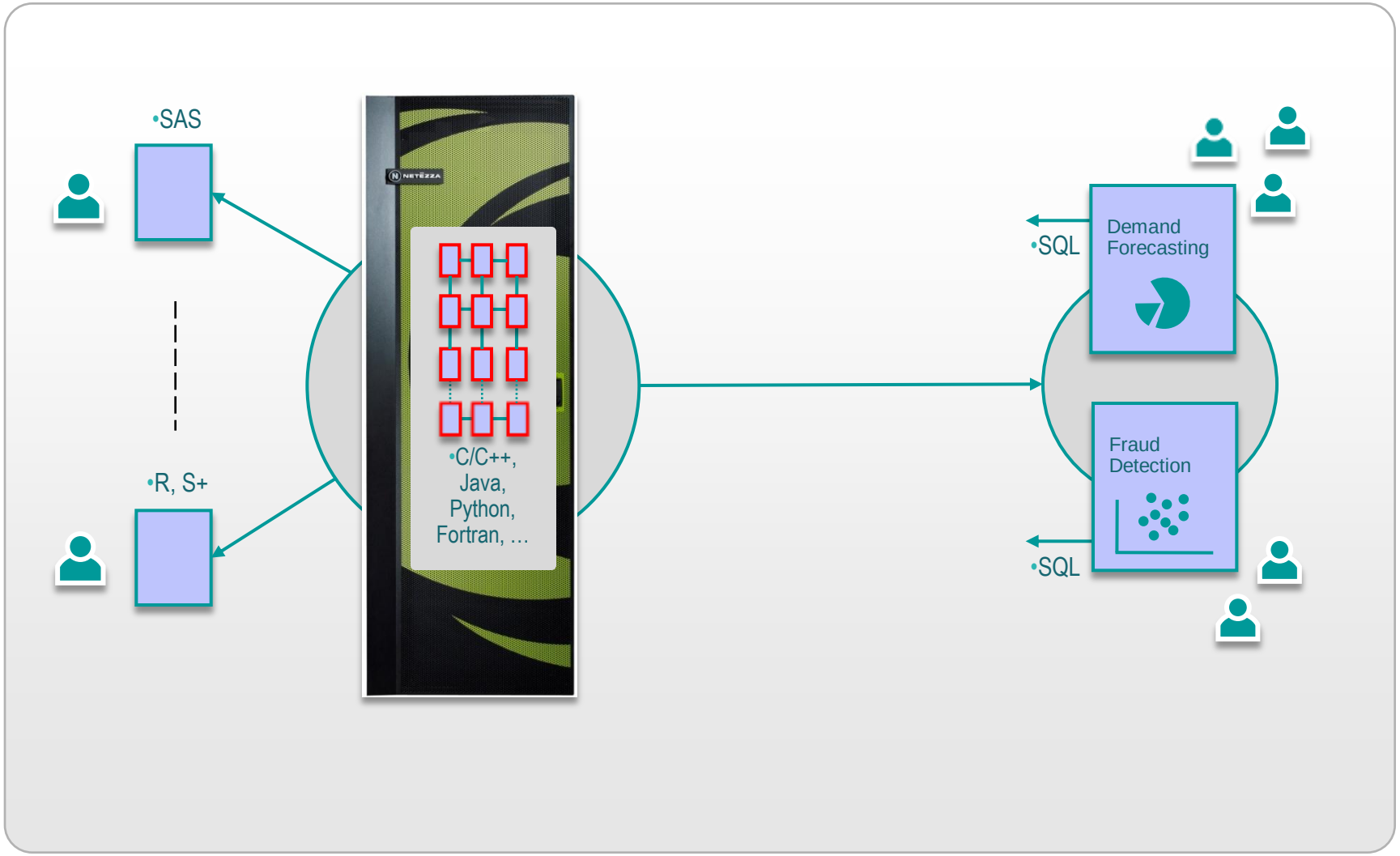
# Traditional Advanced Analytics Architecture



# Advanced Analytics with IBM Netezza Appliance



# Advanced Analytics with IBM Netezza Appliance



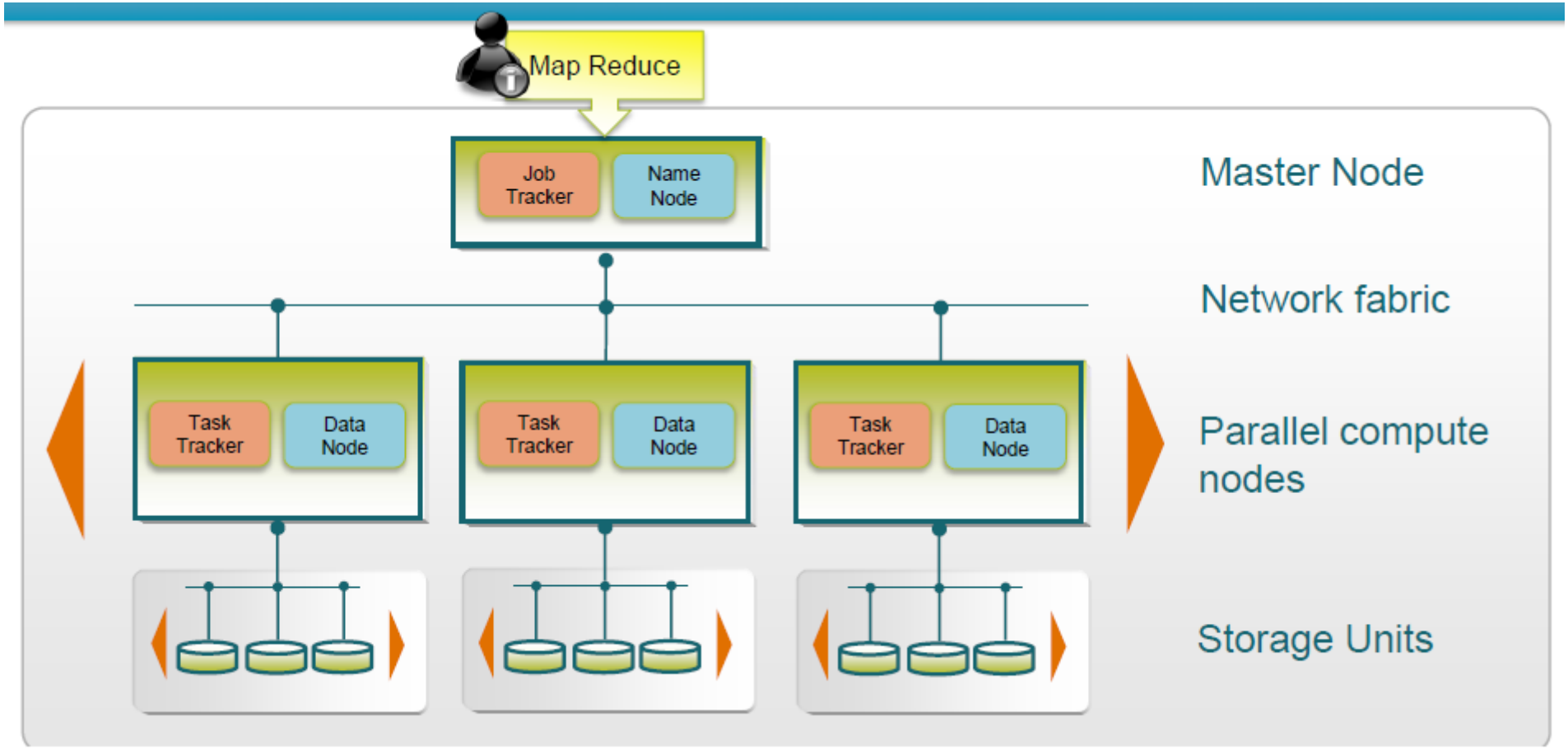
# Hadoop and MapReduce

At its core, Hadoop is a **platform for distributing computing problems across a number of servers.**

- First developed and released as open source by Yahoo, it implements the **MapReduce** approach pioneered by Google in compiling its search indexes. Hadoop's MapReduce involves distributing a dataset among multiple servers and operating on the data: the "map" stage. The partial results are then recombined: the "reduce" stage.
- To store data, Hadoop utilizes its own distributed filesystem, **HDFS**, which makes data available to multiple computing nodes.
- This process is by nature a batch operation, suited for analytical or non-interactive computing tasks. Because of this, Hadoop is not itself a database or data warehouse solution, but can act as an analytical adjunct to one. ***(Subject to opinion….!)***
- **Distributors**: Cloudera, HortonWorks, IBM BigInsights… and many more.

*From* http://radar.oreilly.com/2012/01/what-is-big-data.html

# Hadoop / MapReduce Architecture




***Sample use cases:*** *Repository and refinery for raw data, exploratory analysis, queryable archive, parallel ETL*

# RDBMS and Hadoop solutions

- **RDBMS-based EDW solutions** - such as Netezza appliances - enable low latency access to high volumes of data, provide data retrieval via SQL, integrate with a wide variety of enterprise BI and ETL tools and are optimized for price/performance across a diverse set of workloads.
- **Hadoop's** ability to run on commodity servers, store a broad range of data types, process analytic queries via MapReduce and predictably scale with increased data volumes are very attractive solution characteristics as it pertains to big data analytics.

*http://thinking.netezza.com/blog/architecting-big-data-solutions-part-i-hadoop-and-ibm-netezza*

# Mixed Use-cases (RDBMS DW and Hadoop)


**Unstructured data**

→ hadoop → NETEZZA

**Create context (classification, text mining)**

**Analyze**

---

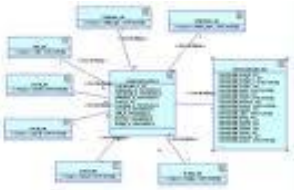
**Semi-structured data**

→ hadoop → NETEZZA

**Parse, aggregate**

**Analyze, report**

---



**Structured data**

→ NETEZZA → hadoop

**Analyze, report**

**Active archival**
**Long running queries**

***Organizations that architect their big data platforms integrating the two technologies have the ability to take advantage of the best of both worlds.***

# NoSQL

**Next Generation Databases** mostly addressing some of the points - being **non-relational**, distributed, open-source and horizontally scalable. The original intention has been modern web-scale databases.

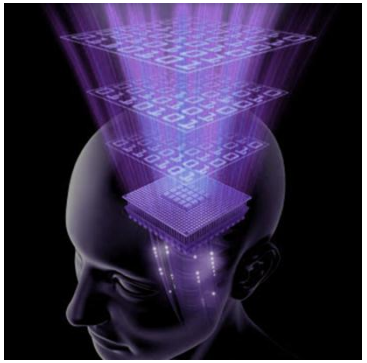
Often more characteristics apply as: **schema-free**, easy replication support, simple API, **eventually consistent / BASE** (not ACID*), a huge data amount, and more.

So the misleading term “NoSQL" (the community now translates it mostly with **"not only sql"**) should be seen as an alias to something like the definition above. [6]

**NoSQL databases:**

- Apache Cassandra, CouchDB, MongoDB, Hadoop Hbase, and many more
- Different types: Columnar stores, document stores, graph DBs, Key-Value stores, etc.

- Big Data and Analytics
- Infrastructure
- **Architectural Strategies**
- IBM Data Warehouse and Analytics Solutions

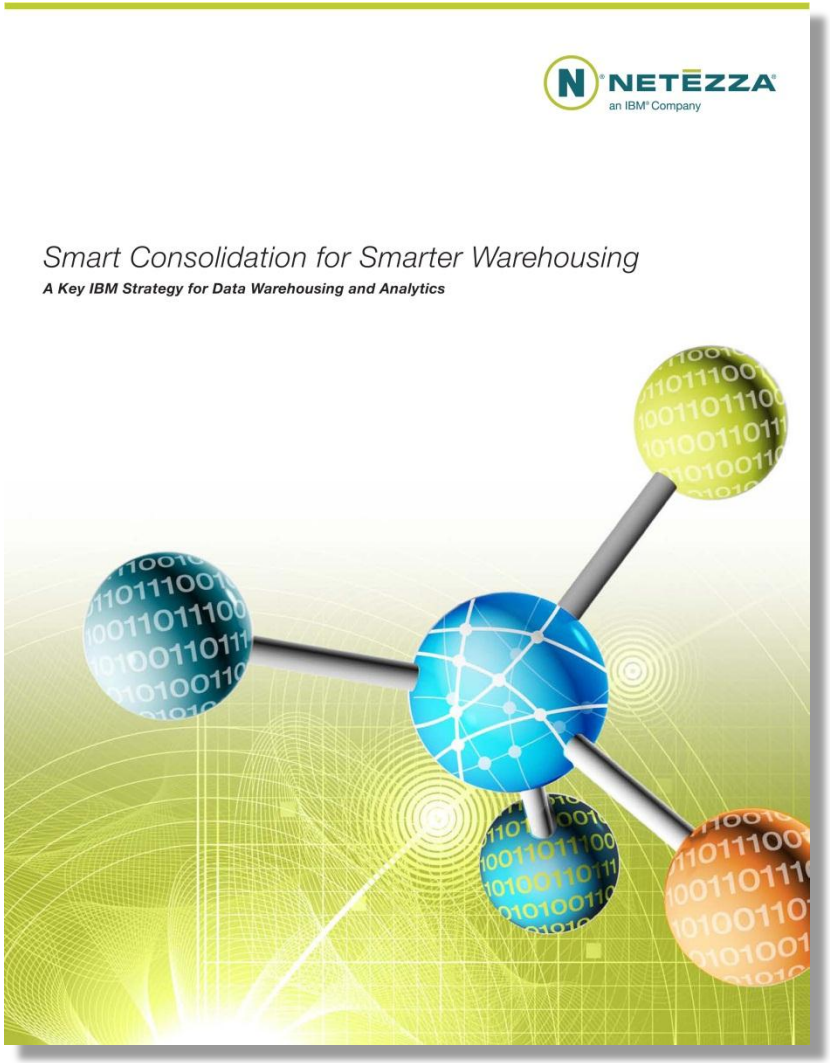


**There is no single technological silver bullet for all your analytic workloads.**

The successful corporate big data architectures will combine use case-driven delivery on an infrastructure that can put the right data, in the right place, in the right combinations to support those use cases.

***Dai Clegg, IBM Netezza*** [7]

# Smart Consolidation for Smarter Warehousing




At its BI Summit on May 2, 2011, Gartner observed that the traditional enterprise data warehouse vision has, in general, not been achieved. These industry analysts refer instead to a logical data warehouse.

In June, IBM announced a strategy, Smart Consolidation for Smarter Warehousing, that adopts Gartner's terminology and recommends an evolutionary change in direction from a single, centralized physical system to a distributed architecture, where computation is provided by individual systems, with each node optimized for specific workloads.

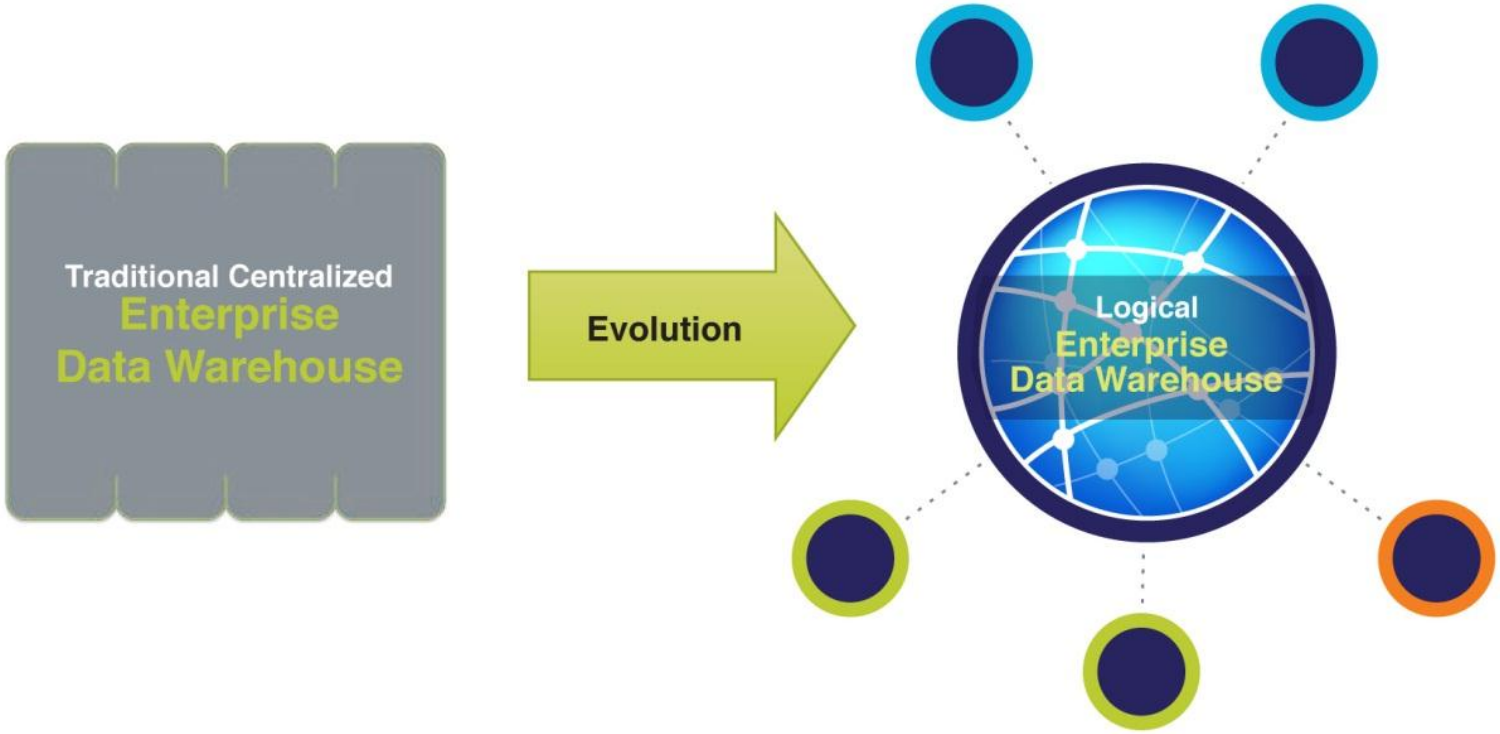
This paper elaborates on the Smart Consolidation strategy, emphasizing fresh use cases and proof points along the way.

*http://thinking.netezza.com/sites/default/files/document/IBM-Netezza-Smart-Consolidation-EDW-Strategy.pdf*

# Evolving to a Logical Data Warehouse

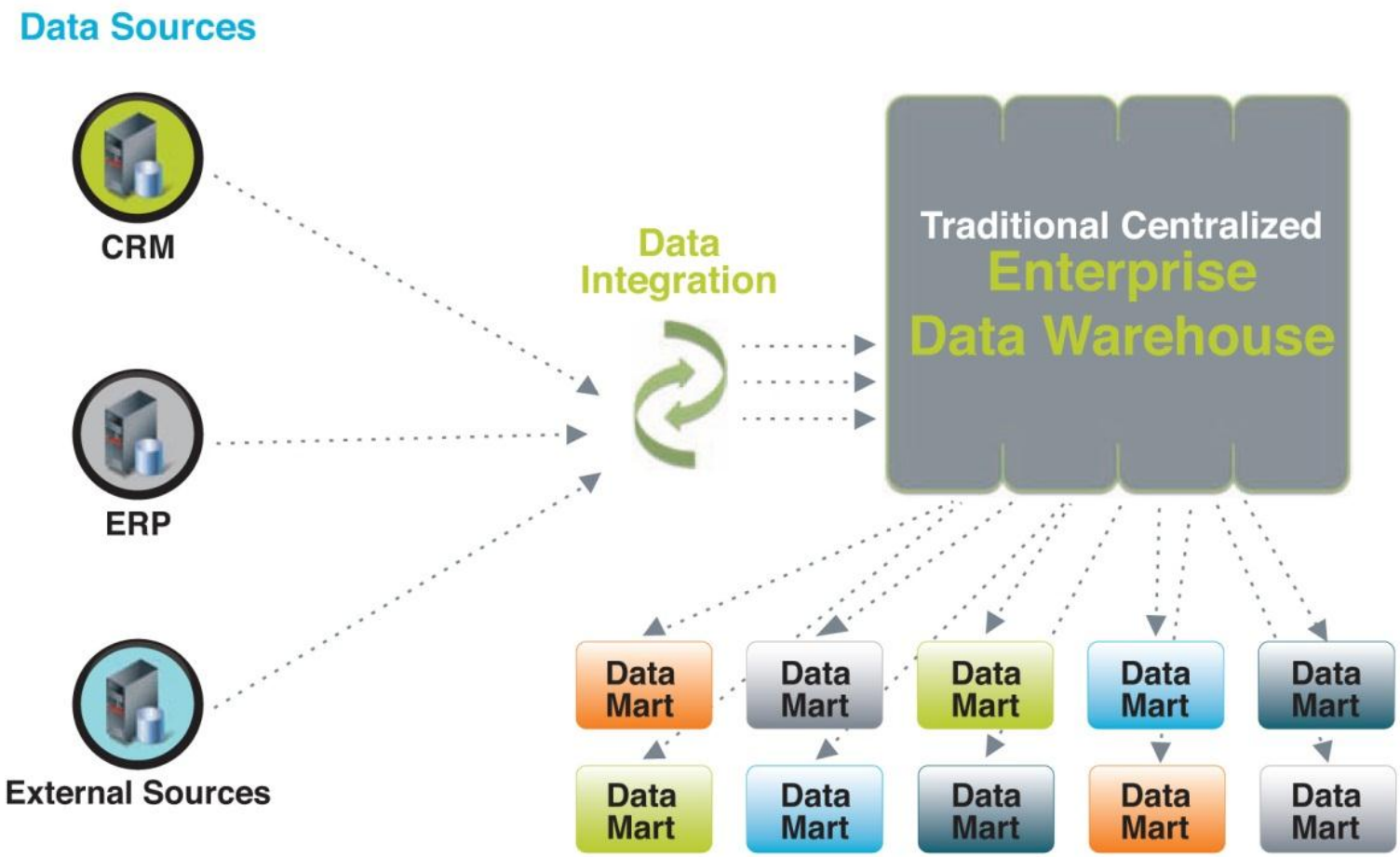
***Key Tenets***

1. Consolidate infrastructure to simplify analytics.
2. Process workloads on “fit for purpose” platforms.
3. Coordinate system management and data governance across the enterprise.



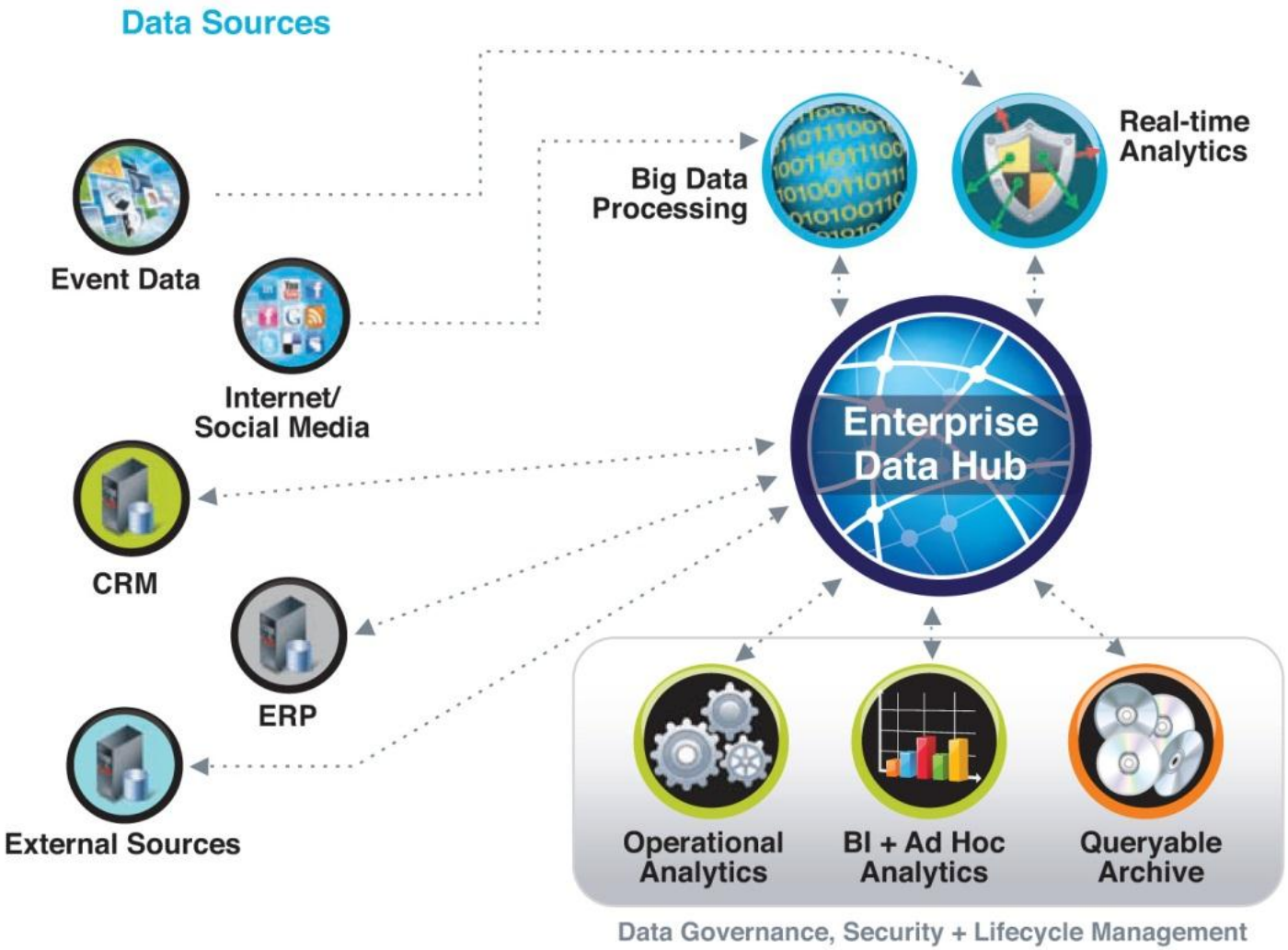
# Traditional Centralized EDW Model

*A single centralized EDW is simply unable to handle today's volume and variety of data. As a result, lines of business resort to ad hoc solutions, creating data mart sprawl, lack of true governance and escalating complexity and costs.*

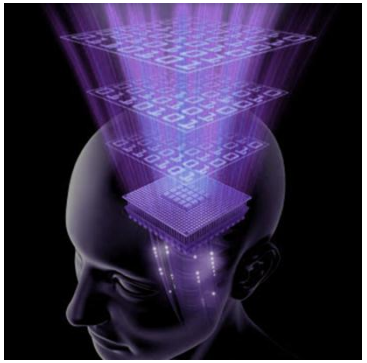


# Smart Consolidation: A New Approach

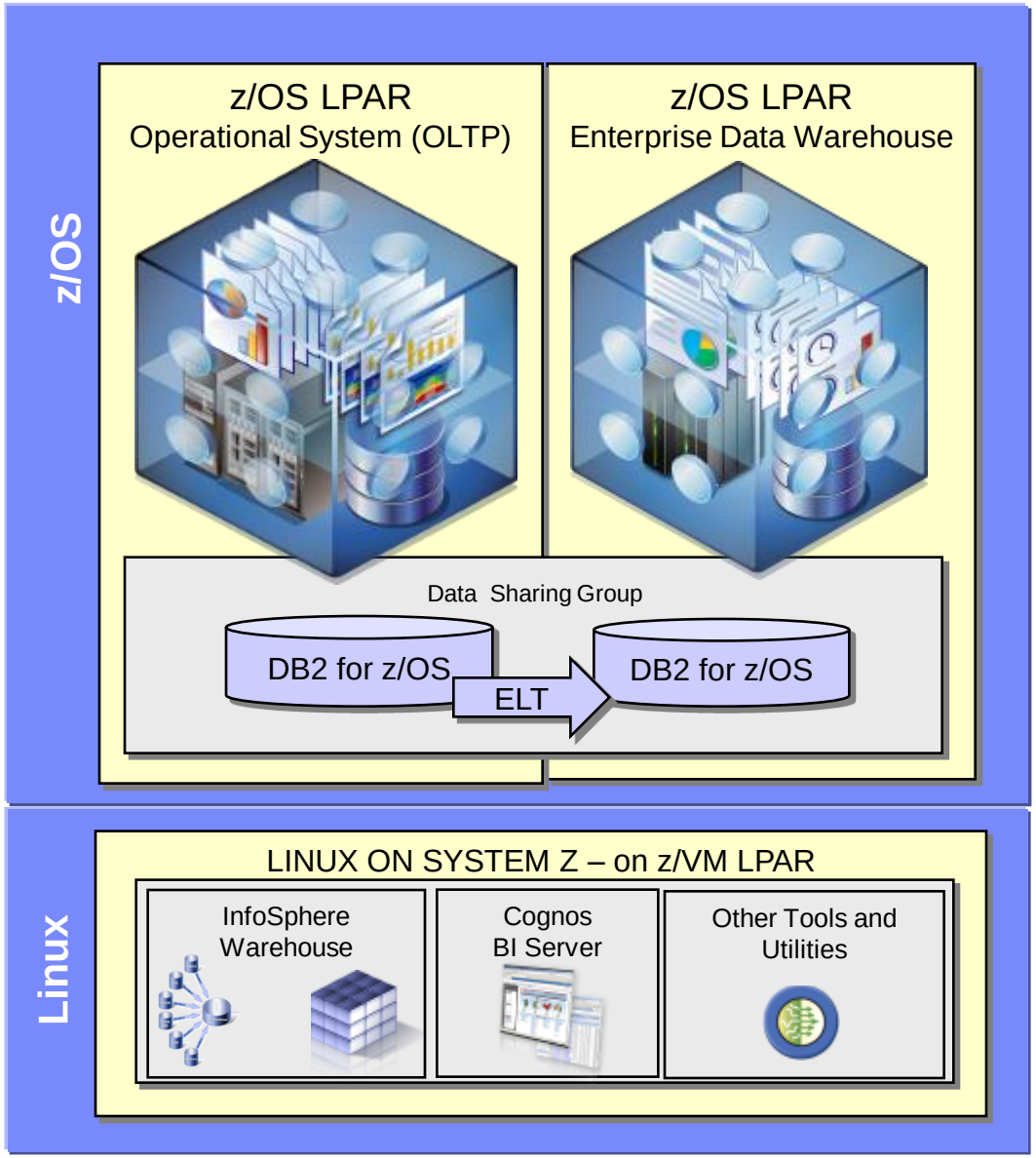
*An architecture, in which computation is managed centrally but executed by individual systems optimized for specific workloads, improves performance, governance, scalability and agility.*



- Big Data and Analytics
- Infrastructure
- Architectural Strategies
- **IBM Data Warehouse and Analytics Solutions**



# A data warehouse solution on a System z foundation

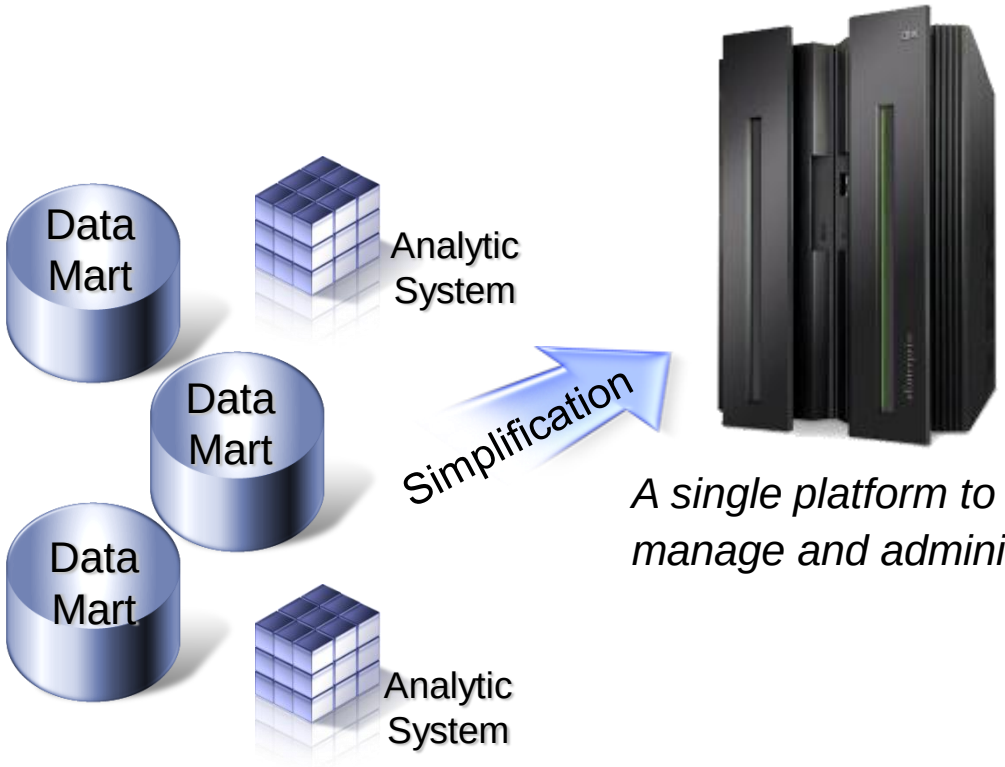



- Minimization of data movement between operational system and data warehouse
- Lowers data latency for time sensitive decisions
- Enables consolidation and simplification of data warehouse and data marts
- Leverages existing high availability, backup, disaster recovery, and security environments
- Greater scalability of multidimensional analysis through cubing services (data marts) and DB2 enhancements
- Complex transformations and data quality driven from Linux on System z with Data Studio

# Data mart consolidation

*Consolidation into a single footprint*

*Utilize virtualization to optimize the use of resources while reducing costs and gaining new agility*




*A single platform to manage and administer*

- Lower software costs
- Lower hardware costs
- Lower administration costs
- Lower environmental costs
- Greater flexibility
- Less complexity
- Fewer points for security intrusions

# Three ways to consolidate on System z

*Choose a platform to consolidate upon*

| DB2 for z/OS | DB2 for Linux | DB2 for Linux/AIX |
|---|---|---|
| zEnterprise | zEnterprise | zBX |
|  |  |  |
| **Robust Mixed Workload** | **Integrated System** | **Managed by System z** |

**Functionality** *The right mix of functionality and investment* **Investment**

# IBM Netezza
# Data Warehouse Appliance

**The true data warehousing appliance.**

- Purpose-built analytics engine
- Integrated database, server and storage
- Standard interfaces
- Low total cost of ownership

- Speed: 10-100x faster than traditional system
- Simplicity: Minimal administration and tuning
- Scalability: Peta-scale user data capacity
- Smart: High-performance advanced analytics

# DB2 Analytics Accelerator

*Accelerating decisions to the speed of business*

*Blending System z and Netezza technologies to deliver unparalleled, mixed workload performance for complex analytic business needs.*

## Get more insight from your data

- Fast, predictable response times for “right-time” analysis
- Accelerate analytic query response times
- Improve price/performance for analytic workloads
- Minimize the need to create data marts for performance
- Highly secure environment for sensitive data analysis
- Transparent to the application

# IBM Smart Analytics System 9700

*Mixed Workloads for Next Generation Business Analytics*

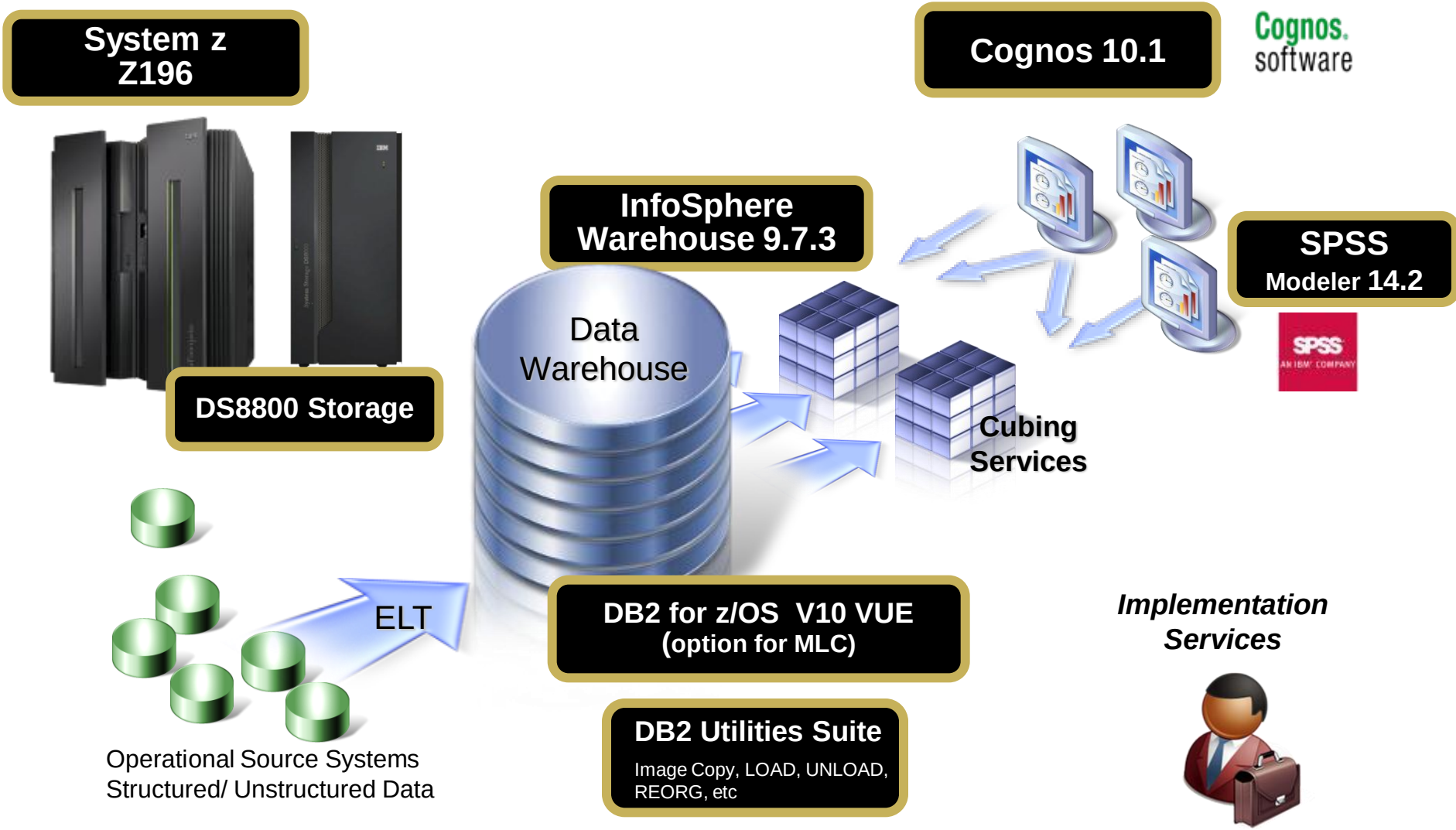
***The next generation of System z analytics; an integrated solution of hardware, software and services that enables customers to rapidly deploy cost effective game changing analytics across their business.***

- ***Secure, Available Business Analytics***
- ***Simplified administration***
- ***Proven Operational Characteristics***
- ***High Value Operational BI***

*Making every decision on facts, at the point of impact*

# IBM Smart Analytics System 9700

*High Value Data Warehousing – Standard Configuration*



# Simplicity, Flexibility, Choice

## IBM Data Warehouse & Analytics Solutions

**IBM Netezza**

**Appliances**

**IBM Smart Analytics System**

**Integrated Optimized Systems**

**IBM Warehouse Software**

**Custom Solutions**

**Warehouse Accelerators**

**Information Management Portfolio**

*(Information Server, MDM, Streams, etc)*

**Simplicity** — *The right mix of simplicity and flexibility* — **Flexibility**

# IBM Continues To Invest: *Your* Partner in Business Analytics and Data Warehousing Solutions

- More than $14B in Acquisitions Since 2005
- More than 10,000 Technical Professionals
- More than 7,500 Dedicated Consultants
- Largest Math Department in Private Industry
- More than 27,000 Business Partner Certifications

# Conclusion?

You need a comprehensive data strategy that addresses **“advanced analytics”** and **“big data”**.

**Vicom Infinity** can help!

## For More Information or follow up questions please contact:

**Len Santalucia**, CTO & Business Development Manager
One Penn Plaza – Suite 201
New York, NY 10119
212-799-9375
*lsantalucia@vicominfinity.com*

## About Vicom Infinity

- Account Presence Since Late 1990's
- IBM Premier Business Partner
- Reseller of IBM Hardware, Software, and Maintenance
- Vendor Source for the Last 4 Generations of Mainframes/IBM Storage
- Professional and IT Architectural Services
- Vicom Family of Companies Also Offer Leasing & Financing, Computer Services, and IT Staffing & IT Project Management

# Footnotes


1. http://radar.oreilly.com/2012/01/what-is-big-data.html

2. Russom, Phil, TDWI Best Practices Report, Big Data Analytics, 4Q 2011, http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx

3. http://timoelliott.com/blog/2011/03/business-analytics-vs-business-intelligence.html

4. Nawale, Tushar (January 2, 2012), "Big Data is the Answer - What was the Question?", http://smartdatacollective.com/tusharrn/44337/big-data-answer-what-was-question

5. Russom, ibid.

6. http://nosql-database.org/

7. Clegg, Dai (January 6, 2012). "Unstructured Data & Structured Data - Complementary, not Divisive", http://thinking.netezza.com/blog/unstructured-data-structured-data-complementary-not-divisive#en

# Recommended Reading

**Smart Consolidation for Smarter Warehousing**

http://thinking.netezza.com/information_resources/whitepapers

**Understanding Analytic Workloads: Meeting the complex processing demands of advanced analytics**

http://thinking.netezza.com/information_resources/whitepapers

**McKinsey Quarterly, Are you ready for the era of 'big data'?**

https://www.mckinseyquarterly.com/Are_you_ready_for_the_era_of_big_data_2864

**Curt Monash: Big data terminology and positioning**

http://www.dbms2.com/2012/01/08/big-data-terminology-and-positioning/

**Curt Monash: Terminology: poly-structured data, databases, and DBMS**

http://www.dbms2.com/2011/05/17/poly-structured-database/

# In the news...

**IBM's 2012 is All About Analytics**

http://www.trefis.com/stock/ibm/articles/96064/ibms-2012-is-all-about-analytics/2012-01-05?from=retweet&ashr=25673420

**Business Analytics at the Core of Recent IBM Moves**

http://www.theinfoboom.com/articles/business-analytics-at-the-core-of-recent-ibm-moves/

**Top BI Technology Trends in 2012**

http://smartdatacollective.com/brett-stupakevich/44205/top-bi-technology-trends-2012

**10 game-changing business innovations for 2012**

http://www.smartplanet.com/blog/business-brains/10-game-changing-business-innovations-for-2012/20656

**10 Business and IT Trends for 2012**

http://www.baselinemag.com/c/a/IT-Management/10-Business-and-IT-Trends-for-2012-512271/

**Cloud & Big Data are Two Big IT Trends of the Next 5 Years**

http://www.dataversity.net/archives/6019

**The "Big Five" IT trends of the next half decade: Mobile, social, cloud, Consumerization, and Big Data**

http://www.zdnet.com/blog/hinchcliffe/the-big-five-it-trends-of-the-next-half-decade-mobile-social-cloud-consumerization-and-big-data/1811

# Additional Backup Slides

# RDBMS & Hadoop – complementary, not competing

| RDBMS | Hadoop |
| --- | --- |
| Structured data with known schemas | Unstructured and structured |
| Records, long fields, objects, XML | Files |
| Updates allowed | Only inserts and deletes |
| SQL & XQuery | Hive, Pig, Jaql |
| Quick response, random access | Batch processing |
| Data loss is not acceptable | Data loss can happen sometimes |
| Security and auditing | Not yet |
| Encryption | Not yet |
| Sophisticated data compression | Simple file compression |
| Enterprise hardware | Commodity hardware |
| 30+ years of innovation | 2-3 years old technology |
| Random access (indexing) | Access files only (streaming) |
| Large DBA and Application development community, widely used | Small number of companies using it in production, many startups |

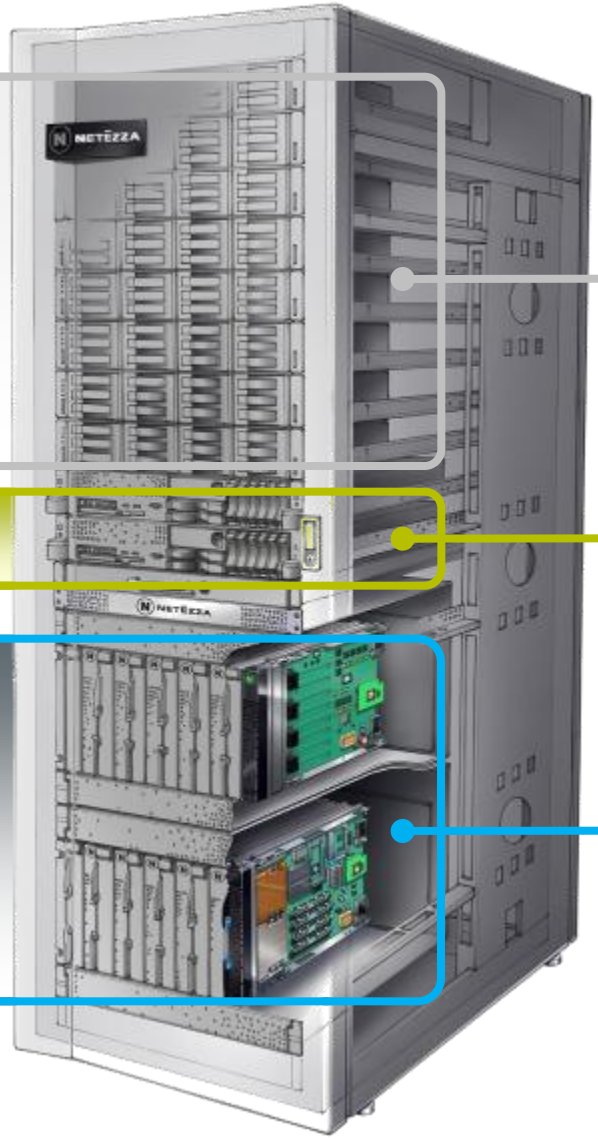
# DB2 Analytics Accelerator V2

*Powered by Netezza 1000 Appliance*

Disk Enclosures

SMP Hosts

Snippet Blades™
(S-Blades, SPUs)

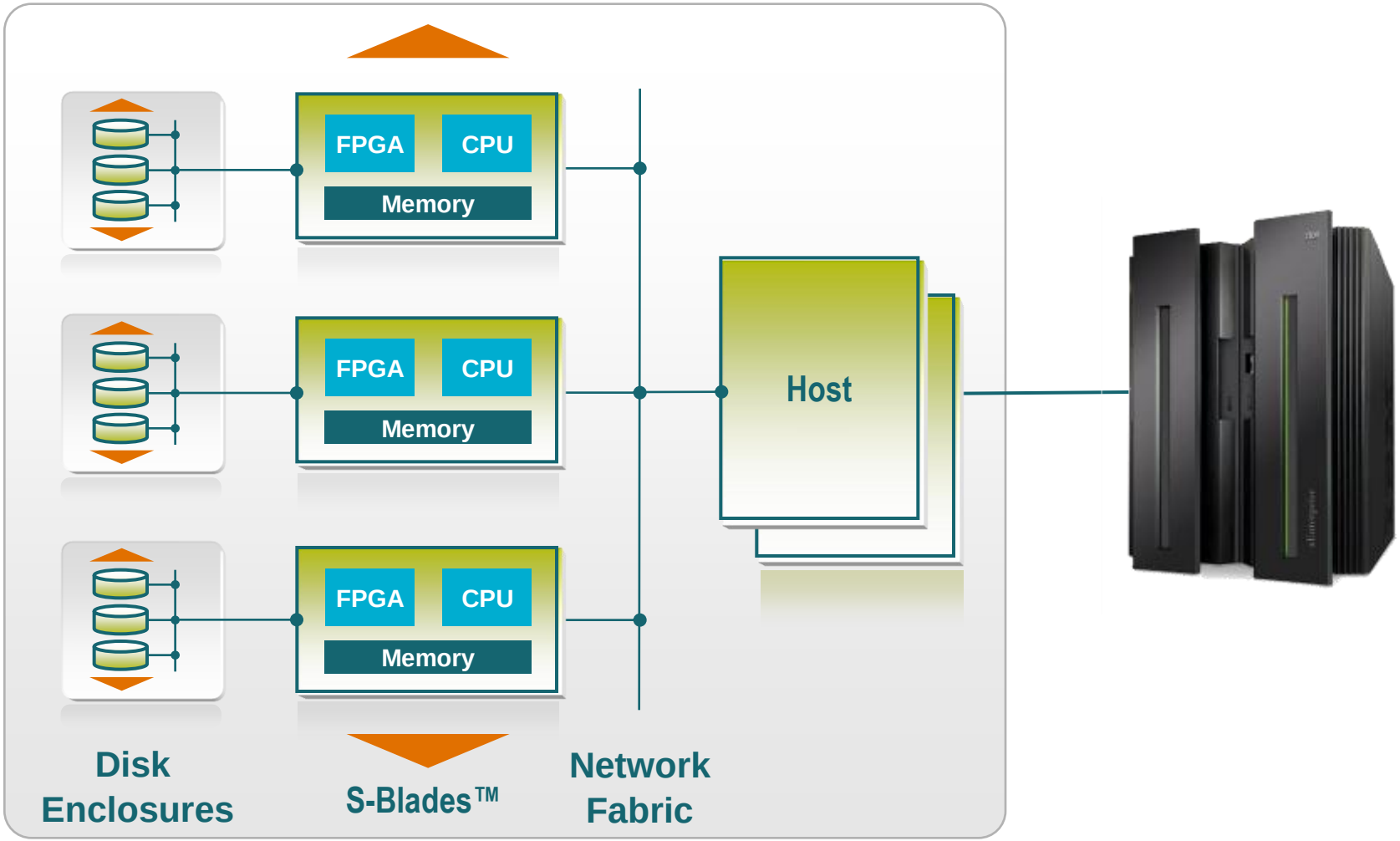



*Slice of User Data*
*Swap and Mirror partitions*
*High speed data streaming*
*High compression rate*
*EXP3000 JBOD Enclosures*
*12 x 3.5" 1TB, 7200RPM, SAS (3Gb/s)*
*max 116MB/s (200-500MB/s compressed data)*
*e.g. TF12:*
*8 enclosures → 96 HDDs*
*32TB uncompressed user data (→ 128TB)*

*IDAA Server*
*SQL Compiler, Query Plan, Optimize*
*Administration*
*2 front/end hosts, IBM 3650M3*
*clustered active-passive*
*2 Nehalem-EP Quad-core 2.4GHz per host*

*Processor &*
*streaming DB logic*
*High-performance database*
*engine streaming joins,*
*aggregations, sorts, etc.*
*e.g. TF12: 12 back/end SPUs*
*(more details on following charts)*

# The Netezza Appliance Connected to a System z




Netezza Appliance

# Netezza: The Key to the Speed