MICROSOFT BUSINESS INTELLIGENCE

James Serra – Sr BI Architect
JamesSerra3@gmail.com
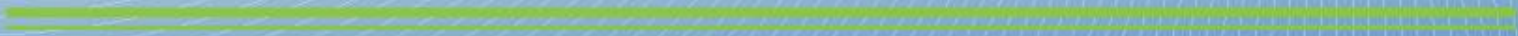http://JamesSerra.com/

About B.I. Voyage

Our Focus:

Microsoft SQL Server   Microsoft SQL Azure   Microsoft SQL SharePoint   Microsoft Office

Microsoft <u>Pure-Play</u> Data Warehousing & Business Intelligence Partner

Our Customers:

Time Warner Cable   TrueValue COMPANY   Lilly   NASDAQ   hp   DirectEdge   Microsoft   comcast
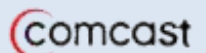
Our Reputation:

"B.I. Voyage came in and proved their subject matter expertise and worked well with our crew. Our CEO is now sold on the business value and corporate impact of having modern day business intelligence solutions!"
email from Steve Bonanno CIO, Direct Edge

"I do count you guys as **leading partners** and this view is represented throughout the PDW business. "
email from Christian Kleinerman, Head of PDW Engineering Microsoft

About me

- In IT for 28 years
- Worked as desktop/web/database developer, DBA, BI
- Been perm, contractor, consultant, business owner
- Six certifications in SQL Server 2008
- SME for SQL Server 2012 certs
- Currently a consultant working with PDW at DirectEdge and Nasdaq
- Blog at JamesSerra.com

Scaling SQL Server to HUNDREDS of Terabytes

Agenda

➢ Data Warehouse

➢ Fast Track Data Warehouse (FTDW)

➢ Business Data Warehouse Appliance (BDW)

➢ Business Decision Appliance (BDA)

➢ Database Consolidation Appliance (DBC)

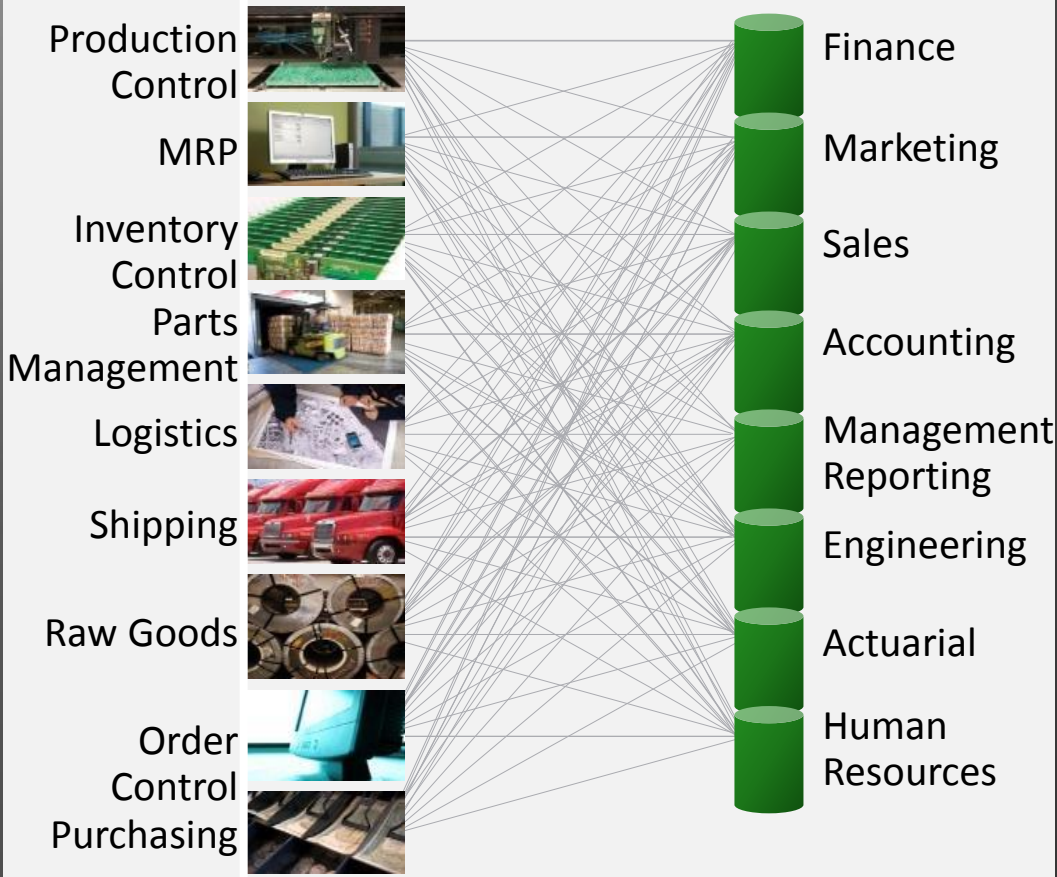➢ Parallel Data Warehouse (PDW)

Data Warehouse

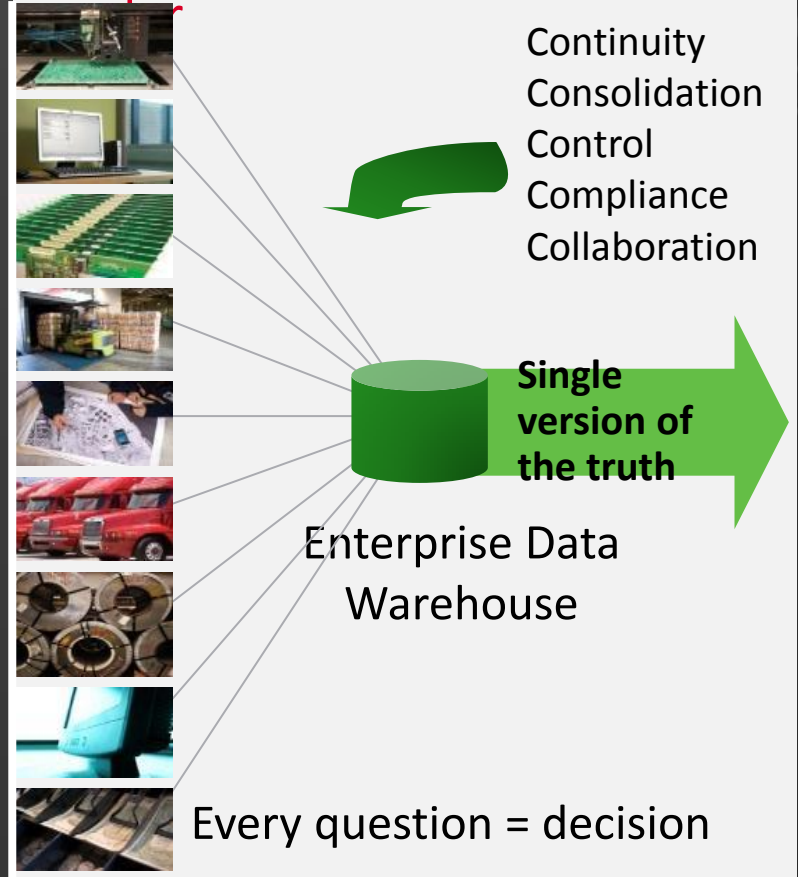All these solutions are for data warehouses only (not OLTP). Why use a data warehouse?

- Reduce stress on production system
- Optimized for read access, sequential disk scans
- Integrate many sources of data
- Keep historical records
- Restructure/rename tables and fields
- Use Master Data Management
- No IT involvement needed for users to create reports
- Improve data quality
- One version of the truth
- Easy to create BI solutions on top of it (SSAS cubes)

Data Warehouse

Legacy applications + data marts = chaos

Production Control
MRP
Inventory Control
Parts Management
Logistics
Shipping
Raw Goods
Order Control
Purchasing

Finance
Marketing
Sales
Accounting
Management Reporting
Engineering
Actuarial
Human Resources

Enterprise data warehouse =

Continuity
Consolidation
Control
Compliance
Collaboration

Single version of the truth
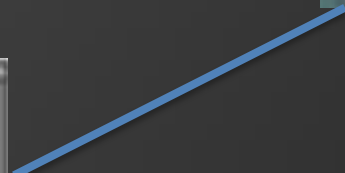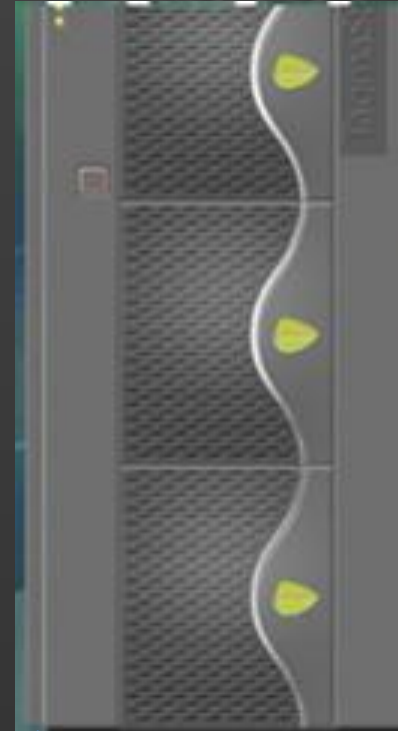
Enterprise Data Warehouse

Every question = decision

Some SQL Data Warehouses Today

What's wrong with
this picture???

Get a big SAN…

Connect it to the biggest server
you can get your hands on

Hope for the best!

System out of balance!!!

- This server CPUs can consume 16 GB/Sec of IO, but the SAN can only deliver 2 GB/Sec
  - Even when the SAN is dedicated to the SQL Data Warehouse, which it often isn't
  - Lots of disks for Random IOPS *BUT*
  - Limited controllers & Limited IO bandwidth
- System is typically IO bound and queries are slow
  - Despite *significant* investment in both Server and Storage
  - Result: Disappointed DBA turning to tuning to squeeze out a bit more performance

Potential Performance Bottlenecks
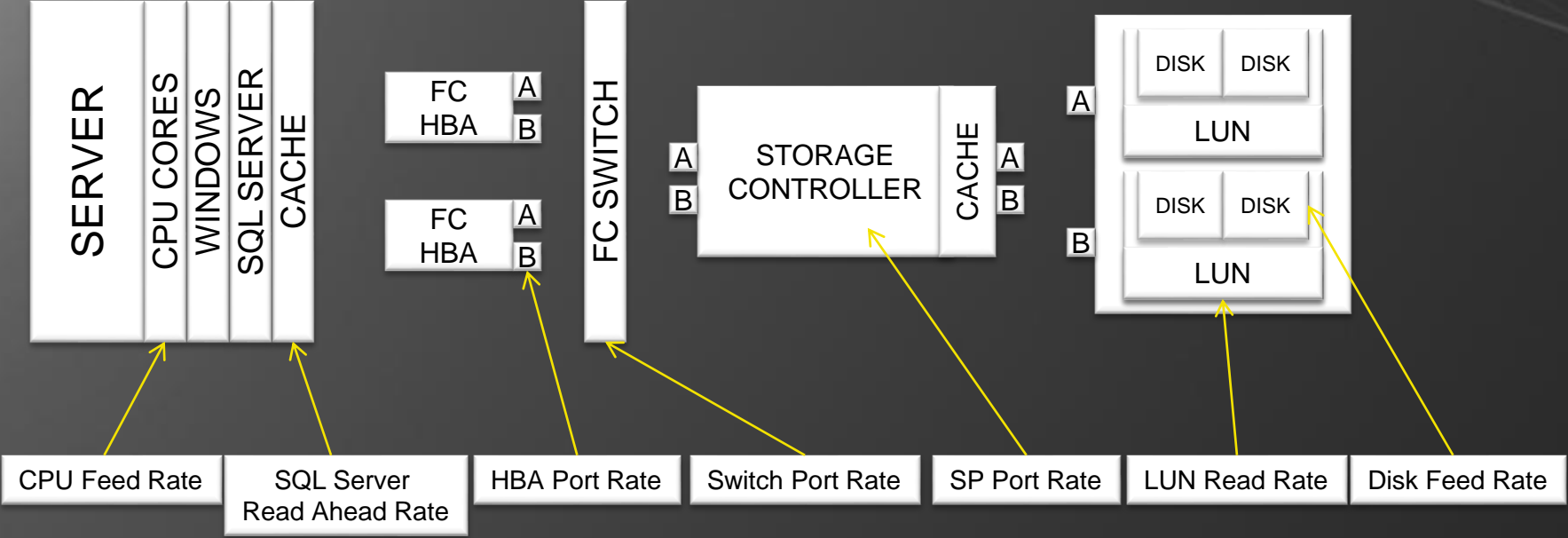
Solution

- Fast Track Data Warehouse - A reference configuration optimized for data warehousing. This saves an organization from having to commit resources to configure and build the server hardware. Fast Track Data Warehouse hardware is tested for data warehousing which eliminates guesswork and is designed to save you months of configuration, setup, testing and tuning. You just need to install the OS and SQL Server

- Appliances - Microsoft has made available SQL Server appliances that allow customers to deploy data warehouse (DW), business intelligence (BI) and database consolidation solutions in a very short time, with all the components pre-configured and pre-optimized. These appliances include all the hardware, software and services for a complete, ready-to-run, out-of-the-box, high performance, energy-efficient solutions

What are Microsoft 'Appliances'

- <u>Problem</u>: Large percentage of IT projects not successful. Too long/complex to install/deploy/configure/tune. Need too many experts.

- <u>Appliances</u> = HW + SW + Services
  - Hardware from vendor (buy from HP, Dell, etc)
  - Software from Microsoft: SQL/SharePoint VL (buy from Microsoft)
  - Services from vendor for entire solution
  - Optimized for the HW+SW: e.g. 3000+ SW parameters, and 500+ HW parts chosen for larger appliances

- <u>Marketing taxonomy (offer customer choice):</u>

Guidance	→	**Reference Architectures, "Fast Track" brand**	→	**Appliances**

- Build it yourself
- Custom configurations
- High IT expertise

- "Cooking recipe"
- Probably higher success
- Can be 'sold' to customers
- Tied to HW vendor

- Very fast time to value
- No options (besides 'size')

Data Warehouse

Microsoft 3 Data Warehousing offerings

Microsoft®
SQL Server 2008 R2
Parallel Data Warehouse

- DW Appliance
- DW Only
- MPP – Massive Parallel Processing
- Scales to 500 TB

Microsoft®
SQL Server 2008
Fast Track Data Warehouse

- SW and HW Reference
- DW Only
- SMP – runs on 1 Server
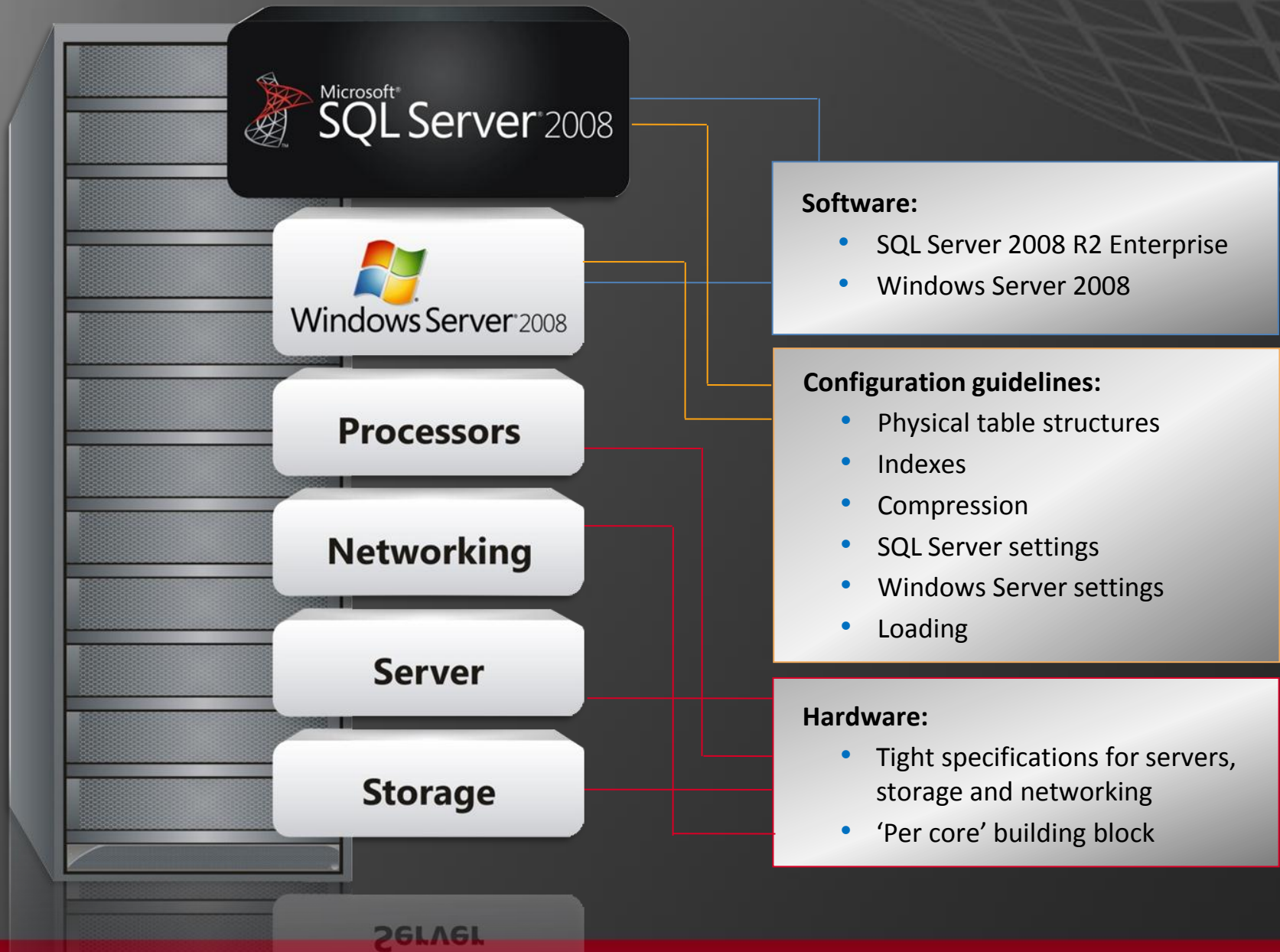- Scales > 40TB (in best of conditions)

Microsoft®
SQL Server 2008 R2

- Customer defines HW
- DW or OLTP
- SMP – runs on 1 Server
- Scales depending on HW, best for < 2 TB DW

Fast Track Data Warehouse

Fast Track Data Warehouse

Software:
- SQL Server 2008 R2 Enterprise
- Windows Server 2008

Configuration guidelines:
- Physical table structures
- Indexes
- Compression
- SQL Server settings
- Windows Server settings
- Loading

Hardware:
- Tight specifications for servers, storage and networking
- 'Per core' building block

Fast Track Data Warehouse

- Reference architecture
- Balanced hardware and database configuration
- Storage, server, application settings, configuration settings
- Predictable performance, scale from 3 to 80 TB
- Data warehouse workload-centric (not one-size-fits-all)
- Efficient disk scan (rather than seek) access
- Benchmarking procedures
- You put together after receiving all the hardware (you need to install the OS, the edition of SQL Server that you've purchased, and any other products such as SharePoint and PowerPivot for SharePoint)
- Eliminates guesswork and is designed to save you months of configuration, setup, testing and tuning

Fast Track Data Warehouse

- Reference architecture
- Balanced hardware and database configuration
- Storage, server, application settings, configuration settings
- Predictable performance, scale from 3 to 80 TB
- Data warehouse workload-centric (not one-size-fits-all)
- Efficient disk scan (rather than seek) access
- Benchmarking procedures
- You put together after receiving all the hardware (you need to install the OS, the edition of SQL Server that you've purchased, and any other products such as SharePoint and PowerPivot for SharePoint)
- Eliminates guesswork and is designed to save you months of configuration, setup, testing and tuning

Fast Track Data Warehouse

- SQL Server Best Practices
  - Data Architecture: Heap Tables, Clustered Index Tables, Table Partitioning
  - Indexing
  - Database Statistics
  - Compression
  - Managing Data Fragmentation
  - Loading Data methods

Fast Track Data Warehouse

Option	Pros	Cons
1. Basic Evaluation	• Very fast system set-up and procurement (days to weeks) • Minimize cost of design and evaluation • Lower infrastructure skill requirements	• Possibility of over-specified storage or under-specified CPU
2. Full Evaluation	• Predefined reference architecture tailored to expected workload • Potential for cost-saving on hardware • Increased confidence in solution	• Evaluation takes effort and time (weeks to months) • Requires detailed understanding of target workload
3. User-defined Reference Architecture	• Potential to reuse existing hardware • Potential to incorporate latest hardware • System highly tailored for your use-case	• Process takes several months • Requires significant infrastructure expertise • Requires significant SQL Server expertise

Fast Track Data Warehouse

- These metrics are use to both validate and position Fast Track RA's
    - **M**aximum **C**onsumption **R**ate (MCR) – Ability of SQL Server to process data for a specific CPU and Server combination and a standard SQL query
    - Benchmark Consumption **R**ate (BCR) – Ability of SQL Server to process data for a specific CPU and Server combination and a user workload or query
    - User Data Capacity (UDC) – Maximum available SQL Server storage for a specific Fast Track RA assuming 2.5:1 page compression factor and 300 GB 15K SAS. 30% of this storage should be reserved for DBA operations

Microsoft Appliances

Business Data Warehouse Appliance

Business Data Warehouse Appliance

- HP and Microsoft tuned and tested (Dell, SQL Server 2012)
- Optimized for SQL Server 2008 R2
- Data Warehouse up to 5TB
- Fast Track 3.0 compliant
- Windows Server 2008 R2 Enterprise and SQL Server 2008 R2 Enterprise already installed and configured
- Pre-tuned, pre-configured, pre-installed. Turn on and go!
- Single point of contact for support
- Quick Deployment Wizard and DDL & Data Loading Wizard
- Could be spoke in PDW hub and spoke architecture
- 2 CPU's (12 cores), 96GB memory, 2TB storage

Business Decision Appliance

Business Decision Appliance

- HP and Microsoft tuned and tested
- Made specifically for BI
- Optimized for SQL Server 2008 R2 and SharePoint 2010
- Windows Server 2008 R2 Enterprise, SQL Server 2008 R2 Enterprise, SharePoint 2010, PowerPivot already installed and configured
- Pre-tuned, pre-configured, pre-installed. Turn on and go!
- Single point of contact for support
- Quick Deployment Wizard
- 2 CPU's (12 cores), 96GB memory

Business Decision Appliance

Business Decision Appliance

Database Consolidation Appliance

Database Consolidation Appliance

- HP and Microsoft tuned and tested big Hyper-V environment
- Solves problem of SQL Server sprawl
- Virtual environment, private cloud, on-demand scalability
- New SQL Server databases provisioned in minutes
- Pre-installed Windows Server Datacenter 2008 R2, SQL Server 2008 R2 Enterprise, Hyper-V, System Center Suite
- Microsoft Database Consolidation 2012 software to manage the appliance
- Automatic load balancing, high availability
- Pre-tuned, pre-configured, pre-installed. Turn on and go!
- 192 cores, 400 disk drives, 2TB memory
- Offered as a reference architecture

Database Consolidation Appliance

Design, Build & Deploy in **weeks** rather than **months**

Custom-built solution

Phase	Task	Months
Design	Assess and understand workload	
Design	Define architecture	1
Design	Evaluate alternatives	
Design	Design specific implementation	
Build	Acquire HW & SW components	2
Build	Build solution	
Build	Load data	
Build	Proof Of Concept & Validation	3
Build	Tune & Balance HW & SW	
Deploy	Integrate in environment	
Deploy	Burn in & Stability	4
Deploy	Monitor and troubleshoot	
Use	Extract and manipulate data	5
Use	Generate reports	
Use	Make decisions	

Integrated & Optimized Appliance

Phase	Task	
Design	Choose appliance for workload	1
Build	Acquire appliance	
Deploy	Install appliance	2
Deploy	Extract & load data	
Deploy	Stand-up in production	
Use	Monitor & Manage	3
Use	Extract and manipulate data	
Use	Generate reports	4
Use	Make decisions	

Parallel Data Warehouse

Scale Out
for both Performance and Capacity simultaneously by adding racks

Control Rack

10 Node Data Rack

A prepackaged or pre-configured balanced set of hardware (servers, memory, storage and I/O channels), software (operating system, DBMS and management software), service and support, sold as a unit with built-in redundancy for high availability positioned as a platform for data warehousing.

HP PDW 1 Rack

17 Servers

22 Processors /132 Cores

125 TB

HP PDW 4 Rack:

47 Servers

82 Processors / 492 Cores

500 TB

SMP vs MPP

SMP	MPP with PDW
- HW advancements increasing ability to scale-up - Scaling is limited - High end SMP very expensive - Extremely high concurrency for some workloads - Less than 1-2 TB of data SMP will almost always be better. Usually <10TB - Full SQL Server functionality - HA must be architected in	- HW advancements increasing ability to scale-up & scale-out - Scaling to 1 PB+ - Scale out is relatively low cost - Relatively high concurrency for complex workloads > 10 TB (*typically*) up to 1 PB - Limited SQL Server functionality - HA is built in

PDW Benefits – Key components all in one package

Control Node (2)

Single connection point for SQL queries. Single touch point for DBAs.
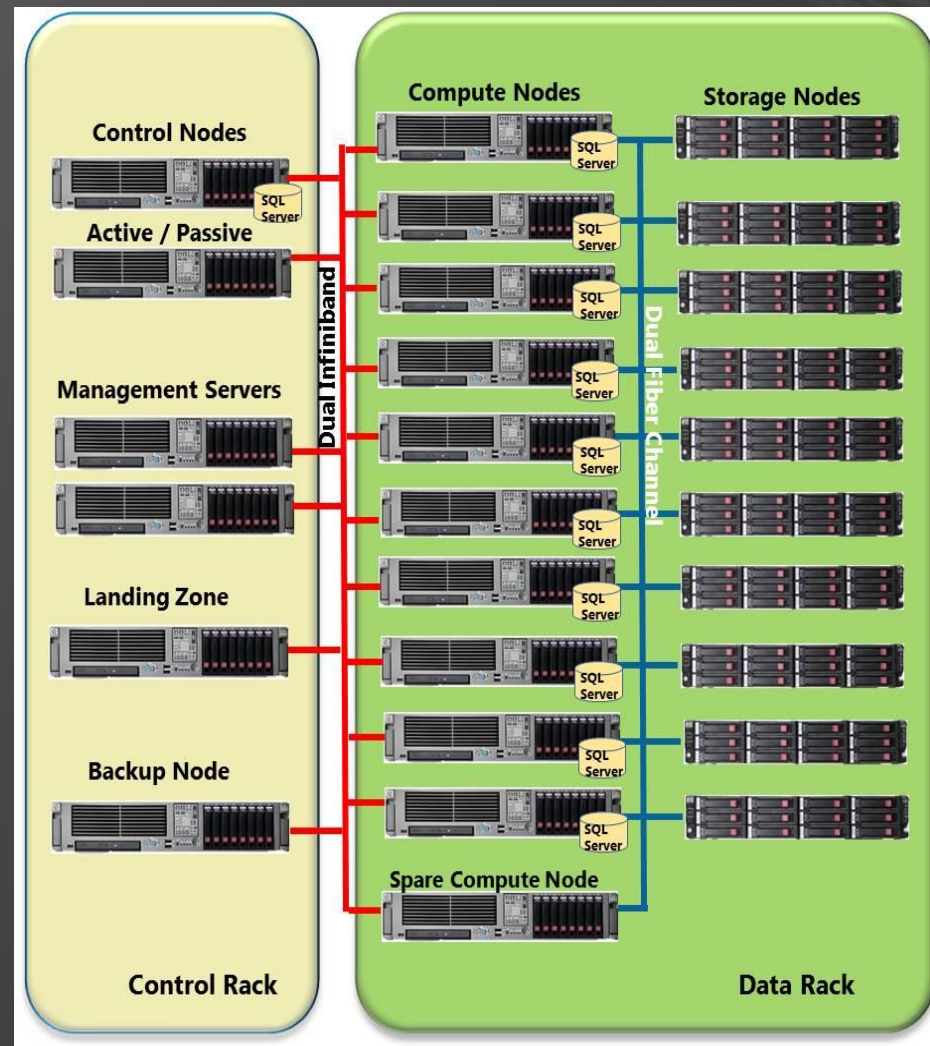
Management Servers (2)

Patch management. Active Directory. Node Image.

Landing Zone (1)

Server and storage dedicated to loading data.

Backup Node (1)

Server and storage dedicated to handling backup.

Control Nodes

Active / Passive

Management Servers

Landing Zone

Backup Node

Control Rack

Dual Infiniband

Compute Nodes

Storage Nodes

SQL Server

Dual Fiber Channel

Spare Compute Node

Data Rack

Failover Clusters
Dual networks
Mirrored drives
Hot swap drive
Dual power supplies
Dual cooling fans

Storage Node (10)

11 or 24 Disks each.
Dual network cards.
Dedicated SAN.

Compute Node (11)

A SQL Server 2008 Instance.
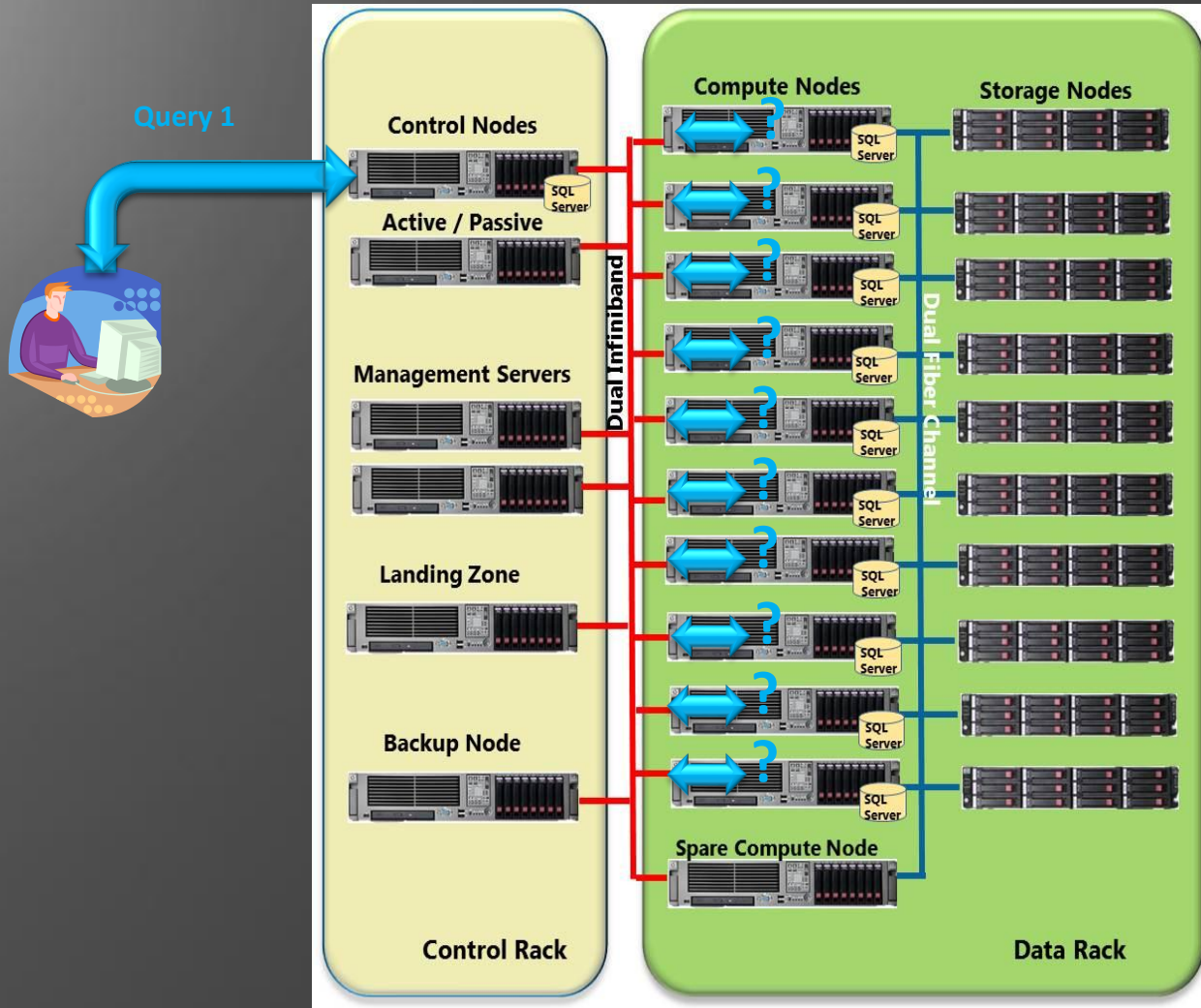Highly Tuned SMP.
8 Cores each.
8 Disks each (TempDB).

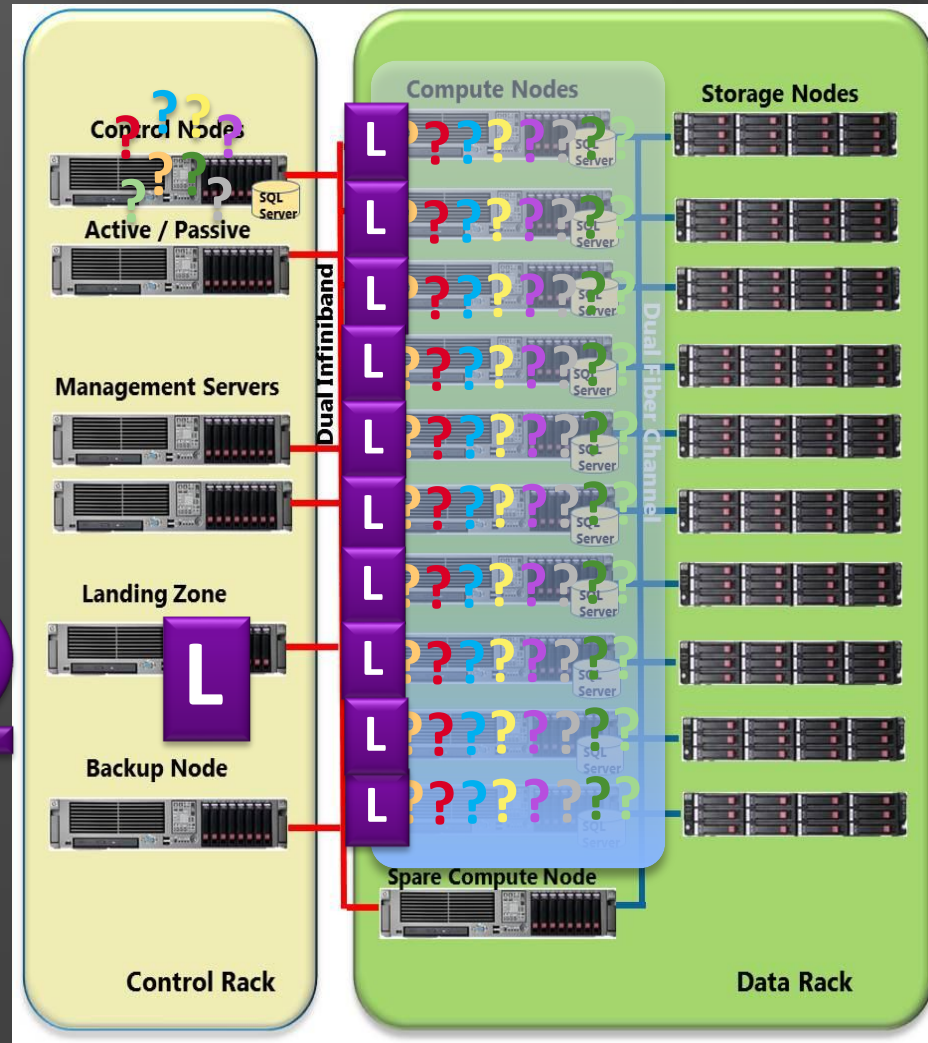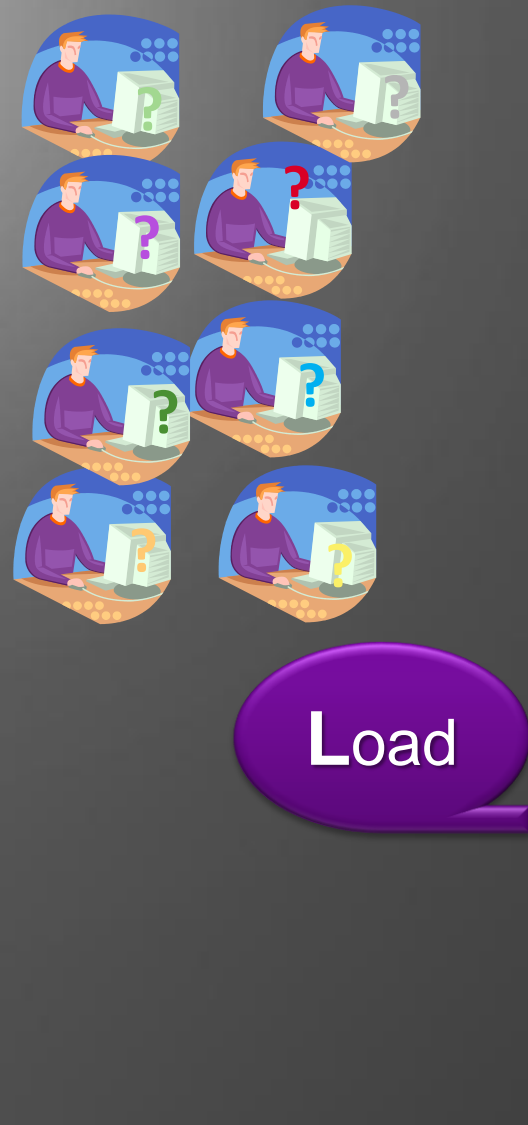# PDW Benefits – Massive Parallel Processing

Query 1 is standard T-SQL submitted to SQL Server on Control Node

Query is executed on all 10 Nodes

Results are sent back to client

PDW Benefits – Massive Parallel Processing

Multiple queries are simultaneously executed across all nodes.

PDW supports querying while data is loading.

PDW - Data Layout Options

Replicated

- A table structure that exists as A full copy within each discrete DBMS instance.

Distributed

- A table structure that is hashed on a single column and uniformly distributed across all nodes on the appliance. Each distribution is A separate physical table in the DBMS.

Ultra shared nothing

- The ability to design a schema of both distributed and replicated tables to minimize data movement
- Small sets of data can be more efficiently stored in full.
- Certain set operations are more efficient against full sets of data.

PDW – PASS Conference Demo

- Using TPC-H Data Model for
  Retail Store Analytics
  - PDW Database Size – 100+TB DW
  - Largest Table - Line_Item_Detail = 600B rows
  - Remaining Fact and Dimension Tables = 220B rows

- PDW Infrastructure – 4 Data Racks
  - Query 1 ran < 20 seconds

PDW – Demo Query Syntax

```
SELECT   n_name, r_name ,

        SUM(o_totalprice) AS totalprice,

        SUM (l_quantity)  AS totalqty

FROM     nation  ,

        region  ,

        customer,

        orders  ,

        lineitem

WHERE   r_regionkey = n_regionkey   AND n_nationkey = c_nationkey

        AND c_custkey   = o_custkey   AND o_orderkey  =l_orderkey

        AND     l_shipdate  BETWEEN '1997-12-01'  AND  '1997-12-07'

        AND     o_orderdate BETWEEN '1997-12-01'  AND  '1997-12-07'

GROUP BY n_name ,

        r_name ,

        o_orderstatus

HAVING   COUNT(l_partkey) > 4
```

PDW – Balanced across servers and within

Largest table	600,000,000,000
Randomly distributed across 40 SQL servers	15,000,000,000
In each server randomly distributed to 8 tables	1,875,000,000
Each partition - 2 years data partitioned by week	17,979,452

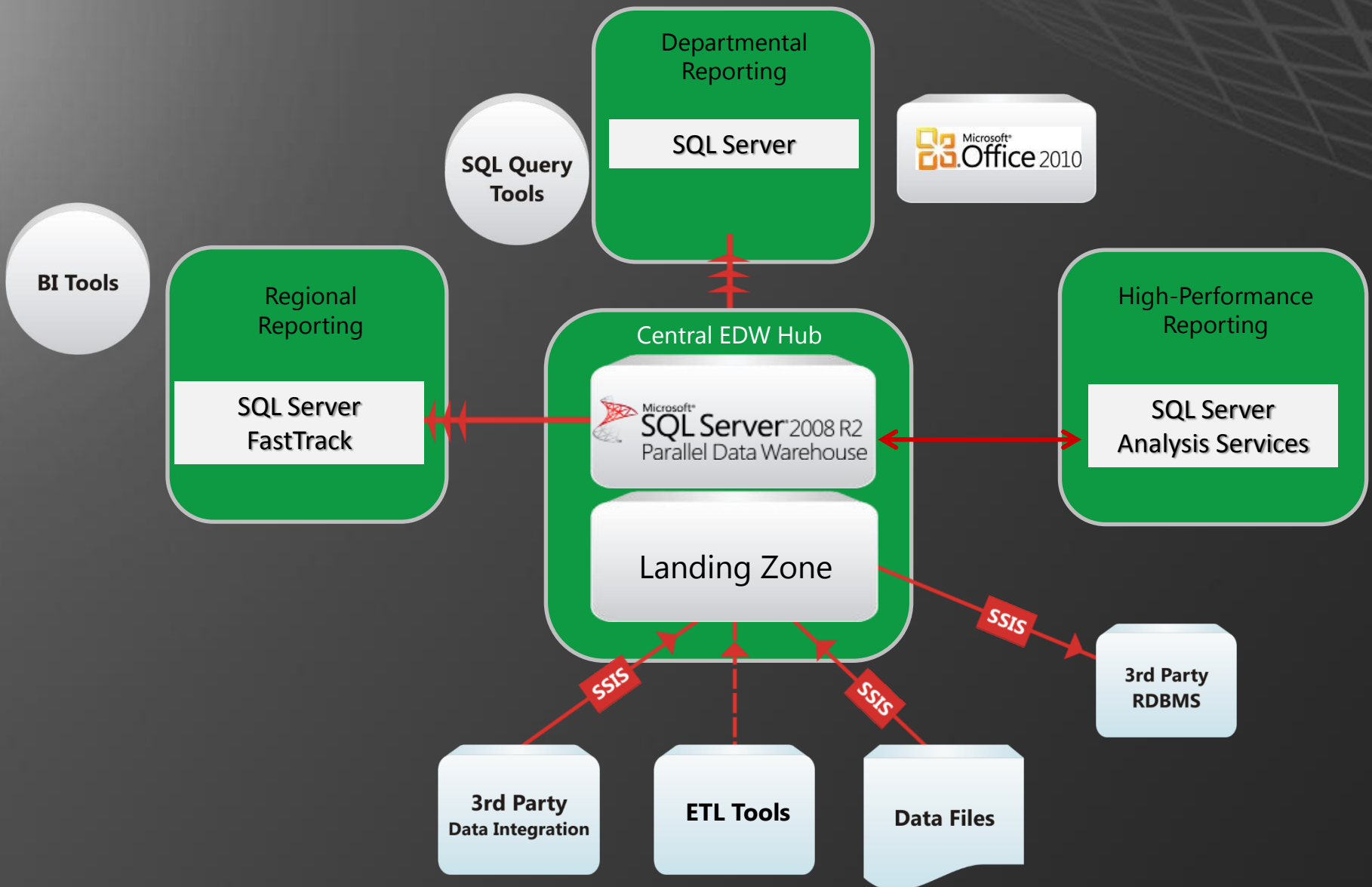As an end user or DBA you think about 1 table: LineItem.

You run "select * from LineItem"

PDW is an appliance, simple to use!

You don't care or need to know that there are actually 320 tables representing your 1 logical table.

That each of those 320 is using it own clustered index and has range partitioning.

PDW – Hub and spoke architecture

Parallel Data Warehouse

- Scale-out instead of scale-up
- MPP instead of SMP
- Ultra shared nothing architecture
- Infiniband
- Hub-and-spoke architecture with support for SMP spokes
- Hardware redundancy, failover clustering
- Parallel loading – 1.5TB per hour on 1 rack
- High speed scanning – 20 to 35GBps per rack
- All appliances can be part of this architecture
- SSIS data flow destination component, .net driver
- DWLoader.exe
- HP (EDW) and Dell
- Fills "Missing Piece" for Microsoft

Final thoughts and questions

- Fast Track for SQL Sever 2012
- Microsoft private cloud fast track reference architecture: http://bit.ly/w1cAh5
- OLTP reference architecture (HP Enterprise Transaction Processing Reference Architecture)
- OLTP reference appliances (Built on HP ProLiant DL980): http://bit.ly/yWEp9C

Resources:

- Microsoft SQL Server Parallel Data Warehouse (PDW) Explained: http://bit.ly/yyuEIC
- Microsoft SQL Server Reference Architecture and Appliances: http://bit.ly/y7bXY5
- Microsoft's Data Warehouse offerings: http://bit.ly/xAZy9h
- Microsoft and HP's Database Consolidation Appliance: http://bit.ly/yBz2qj
- Fast Track Data Warehouse 3.0 Reference Guide: http://bit.ly/AqmTm8
- 7 SQL Server Fast Track Data Warehouse FAQs http://bit.ly/ykS1PD

- HP Fast Track Solutions for Microsoft SQL Server http://bit.ly/wzSsdd
- IBM Reference Configurations for Microsoft SQL Server Fast Track Data Warehouse 3.0 http://ibm.co/zmYihU
- Dell | Microsoft SQL Server Fast Track v3.0 Data Warehouse http://bit.ly/y5DEUy
- Bull Fast Track http://bit.ly/xuTAnu
- Infrastructure Planning and Design Guides for SQL Server: http://bit.ly/AvZQ79