Delivering a Relational Data Warehouse

Week 1 – Introducing the Data Warehouse

Module 03
Exploring Data Warehouse Architectures

Microsoft

Module Outline
03 | Exploring Data Warehouse Architectures

Topic
► Hardware Considerations
► Microsoft Data Warehousing
► **Demo:** Provisioning Microsoft Azure Services

Module Outline
03 | Exploring Data Warehouse Architectures

Topic
Hardware Considerations
Microsoft Data Warehousing
Demo: Provisioning Microsoft Azure Services

Hardware Considerations

- If building the data warehouse on-premises, the project is responsible for technical architecture, product selection and installation
- Data warehouse deployments on the cloud, to a large extent, remove the need to consider or manage technical specifics
  - Cloud deployments can be:
    - **PaaS** (Platform as a Service)—hardware and software infrastructure is provided
    - **IaaS** (Infrastructure as a Service)—hosted virtual machines (VM), with the ability to configure a degree of scale

Hardware Considerations

- To build a high-performance data warehouse, careful hardware considerations must be made
- The goal is to achieve high scalability and performance, which is achieved through parallelism and with high I/O throughput

Hardware Considerations
(Continued)

- In addition, you need to understand:
  - Data size
    - Consider number of subject areas, fact granularity, and history, and also staging requirements and backups
  - Volatility of data
    - Consider how often dimensions or facts require updating
  - Number of concurrent users
  - Number of business processes
    - Consider separate hardware for each process
    - Consider also there may be a need for a centralized, integrated store

Hardware Considerations
(Continued)

- In addition, you need to understand:
  - Type of software
    - Relational databases, OLAP databases, etc.—and how they can be optimized
  - ETL workload
    - Consider data volumes, incremental processing requirements, dimension key lookups, fuzzy matching, etc.
  - Data model and machine learning processing workload
  - Query workload
    - Consider end user usage
    - Consider also data model and machine learning querying

Hardware Considerations
(Continued)

- In addition, you need to understand:
  - Existing platforms and software choices, and available skills
  - Disaster recovery and high availability requirements
  - Budget

Hardware Considerations
(Continued)

- In general, consideration needs to be given to:
  - Number, and placement, of servers
  - Server subsystems:
    - Processors
    - Memory
    - Disk
    - Networking

Hardware Considerations
Servers

- In large deployments, servers can be dedicated
to specific data warehouse roles:
  - Staging
  - Master Data Management
  - Data Quality Management
  - ETL processing
  - Data warehouse
  - Data mart
  - Data models and machine learning models
  - Reporting

Hardware Considerations
Servers (Continued)

- Consider co-locating services, especially where
workloads are compatible
  - For example, ETL processing and machine learning workloads can be
  balanced
- Servers can be located on-premises, or the cloud (IaaS), or a
hybrid topology spanning on-premises to the cloud
- Give consideration also to development and test environments

Hardware Considerations
Server Subsystems ► Processors

- Scale and performance is achieved with parallel processing architectures
  - SMP (Symmetric Multiprocessing)—scale within a single machine
  - MPP (Massively Parallel Processing)—scale across machines
    - Ensure software or design compatibility
- ETL processing, aggregation, index operations, data model and machine learning processing, and certain queries can be very processor-intensive

Hardware Considerations
Server Subsystems ► Memory

- Thanks to the decreased cost of memory, software is increasingly exploiting available memory
  - In general, memory is x100 faster than disk I/O
- For high performance, plan for sufficient memory to host entire datasets and models
- Ensure operating systems can use the installed memory

Hardware Considerations
Server Subsystems ► Disk

- I/O performance should be a key consideration
- The typical data warehouse workload is especially I/O intensive
  - Operations include large data loads and index builds, aggregation, and queries over large volumes of data
- The underlying I/O system for a data warehouse should be designed to meet these heavy requirements

Hardware Considerations
Server Subsystems ► Disk (Continued)

- Give consideration to appropriate storage solutions:
  - RAID (Redundant Array of Independent Disks)
    - High availability and performance
  - SSD (Solid State Storage)
    - Much higher I/O throughput than HDDs (Hard Disk Drive)
  - SAN (Storage Area Network)
    - Shared storage for multiple servers
  - NAS (Network-Attached Storage)
    - Network file server

Hardware Considerations
Server Subsystems ► Networking

- Ensure sufficient network bandwidth is available
  to transfer data volumes, especially during ETL processing and data model processing
- Consider also network throughput for SAN and NAS devices
- When leveraging cloud services, consider internet bandwidth and availability
  - PaaS or IaaS
  - Data sources and data services

Hardware Considerations
Recommended Practices

- Work closely with infrastructure experts
- Design for growth
  - Data warehouses grow rapidly, and tend do grow at accelerated rates
- Consider new in-memory technologies, available for relational databases, and also for data models
- Be aware hardware and software decisions, and design solutions to exploit them accordingly
- Consider cloud alternatives, including hybrid opportunities

Module Outline
03 | Exploring Data Warehouse Architectures

Topic
Hardware Considerations
Microsoft Data Warehousing
Demo: Provisioning Microsoft Azure Services

Microsoft Data Warehousing
Microsoft BI Strategy and Vision

To improve organizations
by providing business insights to all
employees, leading to better, faster,
more relevant decisions

- Microsoft has a long-term commitment to delivering a complete and integrated BI offering
- SQL Server continues to lead innovation in BI
- There is widespread delivery of intelligence through Office, SharePoint and Power BI
- The platforms are enterprise-grade and affordable, with on-premises or cloud deployment choices
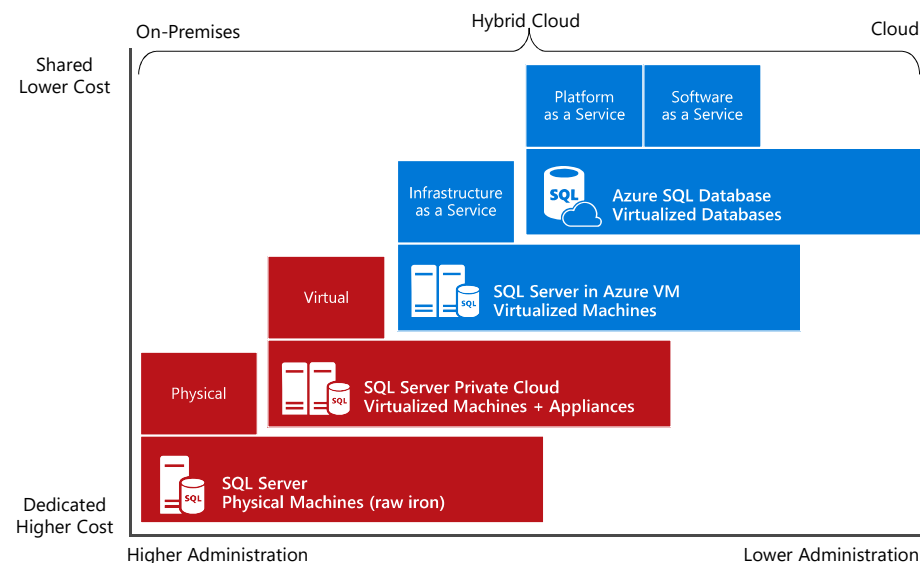
Microsoft Data Warehousing
Choice

▪ Specifically, to deliver a relational data warehouse, Microsoft provides deployment choices
  - On-premises:
    • SQL Server Database Engine, including BI services
    • Analytics Platform System (APS)—formerly the Parallel Data Warehouse
  - Cloud:
    • **IaaS:** Windows virtual machines, or SQL Server virtual machines
    • **PaaS:** Azure SQL Database, or Azure SQL Data Warehouse

Microsoft Data Warehousing
Data Platform Continuum

Microsoft Data Warehousing
On-Premises ► SQL Server

- **Microsoft SQL Server** is the foundation of Microsoft's data platform
- Central to the product is the Database Engine
- The on-premises Database Engine can be used for different relational workloads:
  - OLTP
  - Staging
  - Data warehouse

Microsoft Data Warehousing
On-Premises ► SQL Server (Continued)

- **Microsoft SQL Server** also delivers a comprehensive, enterprise-ready BI platform that helps transform complex data into actionable insights
  - Enterprise Information Management (EIM) suite:
    - Integration Services
    - Master Data Services
    - Data Quality Services
  - Analysis and Reporting:
    - Analysis Services (multidimensional and data mining mode, and tabular mode)
    - Reporting Services

Microsoft Data Warehousing
On-Premises ► Analytics Platform System

- The **Microsoft Analytics Platform System** is an appliance designed to meet the demands of an evolving data warehouse environment
- It is a scale-out, massively parallel processing (MPP) integrated system supporting hybrid data warehouse scenarios
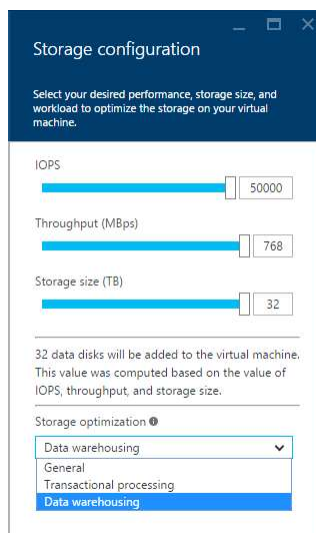
Microsoft Data Warehousing
On-Premises ► Analytics Platform System (Continued)

- Benefits:
  - Up to 100x performance gains over legacy data warehouses
  - Relational and non-relational data in one appliance
  - Seamless integration of the relational data warehouse and Hadoop with Microsoft PolyBase
  - Linear scale-out to 6 petabytes of user data capacity
  - The lowest price per terabyte for a data warehouse appliance in the industry

Microsoft Data Warehousing
Cloud ► IaaS ► Virtual Machines

Storage configuration

Select your desired performance, storage size, and workload to optimize the storage on your virtual machine.

IOPS

50000

Throughput (MBps)

768

Storage size (TB)

32

32 data disks will be added to the virtual machine. This value was computed based on the value of IOPS, throughput, and storage size.

Storage optimization ⓘ

Data warehousing
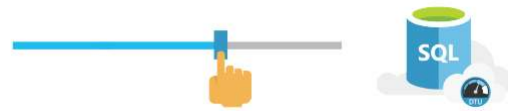General
Transactional processing
Data warehousing

- Azure Virtual Machines provide flexibility of virtualization for a wide range of computing solutions—including data warehousing
  - Provision a Windows machine and deploy your own SQL Server installation, or
  - Provision a SQL Server virtual machine
    - Storage can be optimized for a data warehousing workload
- Scale the VM resources to suit demand

Microsoft Data Warehousing
Cloud ► PaaS ► Azure SQL Database

- **Microsoft Azure SQL Database** is relational database-as-a-service, fully managed by Microsoft
- One of many advantages, is the ability to scale performance up or down, and on the fly to quickly adapt to changing workload
- It is also ideal for organizations looking to dramatically increase the DB:IT ratio

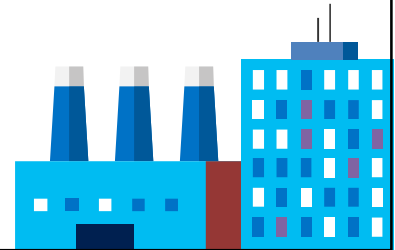# Microsoft Data Warehousing
Comparison of IaaS and PaaS

	IaaS: SQL Server in a VM	PaaS: Azure SQL Database
Best Fit	Existing applications which requires full-box product functionality	Applications that need elastic scale and/or reduced overhead
Resources	Business has ecosystem of IT resources for support and maintenance	Business does not want to add additional IT resources for support maintenance
TCO Benefits	Removing CAPEX	Avoiding CAPEX and OPEX
Scalability	Scale up to 20,000 IOPS	Scale out to thousands of databases, processing TBs of data

Microsoft Data Warehousing
Cloud ► PaaS ► Azure SQL Data Warehouse

- **Microsoft Azure SQL Data Warehouse** is an elastic data warehouse as a service with enterprise-class features
  - Industry's first enterprise-class cloud data warehouse that can grow, shrink, and pause in seconds
  - Full enterprise-class SQL Server experience
  - Petabyte scalability with massive parallel processing
  - Independent scale of compute and storage in seconds
  - Seamless compatibility with Power BI, Azure Machine Learning, HDInsight, and Azure Data Factory
  - Transaction of SQL queries across relational and non-relational data in Hadoop with PolyBase

Microsoft Data Warehousing
Cloud ► PaaS ► Complimentary Services

- **Microsoft Azure Data Factory**
  - Used to compose and orchestrate data services at scale
  - Can connect to on-premises and cloud data sources

- **Microsoft Azure Automation**
  - A highly available automation engine, suitable for ETL execution

- **Microsoft Azure HDInsight**
  - A managed Big Data service (Hadoop, Spark, Storm, HBase)

Microsoft Data Warehousing
Cloud ► PaaS ► Complimentary Services (Continued)

- **Microsoft Azure Machine Learning**
  - A manage cloud-based predictive analytics service

- **Microsoft Azure Data Catalog**
  - Used to register enterprise data assets, and enable their discovery

Microsoft Data Warehousing
Mapping Components to Products and Services

- **Extract, Transform and Load systems**
- Operational Data Stores (ODS)
- Staging systems
- Data warehouse, Data marts
- Master Data Management systems
- Data Quality Management systems
- Data models (OLAP)
- Prediction models
- Data dictionaries

Microsoft Product or Service

On-premises:
- SQL Server Integration Services
  - Automate with SQL Server Agent

Cloud:
- Azure Virtual Machine (IaaS)
  - SQL Server Integration Services
- Azure Data Factory (PaaS)
  - Automate with Azure Automation

Microsoft Data Warehousing
Mapping Components to Products and Services

- Extract, Transform and Load systems
- Operational Data Stores (ODS)
- Staging systems
- Data warehouse, Data marts
- Master Data Management systems
- Data Quality Management systems
- Data models (OLAP)
- Prediction models
- Data dictionaries

Microsoft Product or Service

On-premises:
- SQL Server Database Engine

Cloud:
- Azure Virtual Machine (IaaS)
  - SQL Server Database Engine
- Azure SQL Database (PaaS)

Microsoft Data Warehousing
Mapping Components to Products and Services

- Extract, Transform and Load systems
- Operational Data Stores (ODS)
- Staging systems
- Data warehouse, Data marts
- Master Data Management systems
- Data Quality Management systems
- Data models (OLAP)
- Prediction models
- Data dictionaries

Microsoft Product or Service

On-premises:
- SQL Server Master Data Services

Cloud:
- Azure Virtual Machine (IaaS)
  - SQL Server Master Data Services

Microsoft Data Warehousing
Mapping Components to Products and Services

- Extract, Transform and Load systems
- Operational Data Stores (ODS)
- Staging systems
- Data warehouse, Data marts
- Master Data Management systems
- **Data Quality Management systems**
- Data models (OLAP)
- Prediction models
- Data dictionaries

Microsoft Product or Service

On-premises:
- SQL Server Data Quality Services

Cloud:
- Azure Virtual Machine (IaaS)
  - SQL Server Data Quality Services

Microsoft Data Warehousing
Mapping Components to Products and Services

- Extract, Transform and Load systems
- Operational Data Stores (ODS)
- Staging systems
- Data warehouse, Data marts
- Master Data Management systems
- Data Quality Management systems
- **Data models (OLAP)**
- Prediction models
- Data dictionaries

Microsoft Product or Service

On-premises:
- SQL Server Analysis Services

Cloud:
- Azure Virtual Machine (IaaS)
  - SQL Server Analysis Services

Microsoft Data Warehousing
Mapping Components to Products and Services

- Extract, Transform and Load systems
- Operational Data Stores (ODS)
- Staging systems
- Data warehouse, Data marts
- Master Data Management systems
- Data Quality Management systems
- Data models (OLAP)
- **Prediction models**
- Data dictionaries

Microsoft Product or Service

On-premises:
- SQL Server Analysis Services

Cloud:
- Azure Machine Learning (PaaS)
- Azure Virtual Machine (IaaS)
  - SQL Server Analysis Services

Microsoft Data Warehousing
Mapping Components to Products and Services

- Extract, Transform and Load systems
- Operational Data Stores (ODS)
- Staging systems
- Data warehouse, Data marts
- Master Data Management systems
- Data Quality Management systems
- Data models (OLAP)
- Prediction models
- **Data dictionaries**

Microsoft Product or Service

On-premises:
- No product available

Cloud:
- Azure Data Catalog (PaaS)

Microsoft Data Warehousing
Course Focus

- The focus of this course will be to deliver relational data warehouse by using either:
  - SQL Server, on-premises
  - SQL Server, hosted in an Azure Virtual Machine
  - Azure SQL Database
- Other Microsoft products and services suitable for data warehousing application are covered in different course content

Microsoft

Module Outline
03 | Exploring Data Warehouse Architectures

Topic
Hardware Considerations
Microsoft Data Warehousing
Demo: Provisioning Microsoft Azure Services

Demo

Provisioning Microsoft Azure Services

Demo objectives:

1. Provision an Azure Virtual Machine
2. Provision an Azure SQL Database