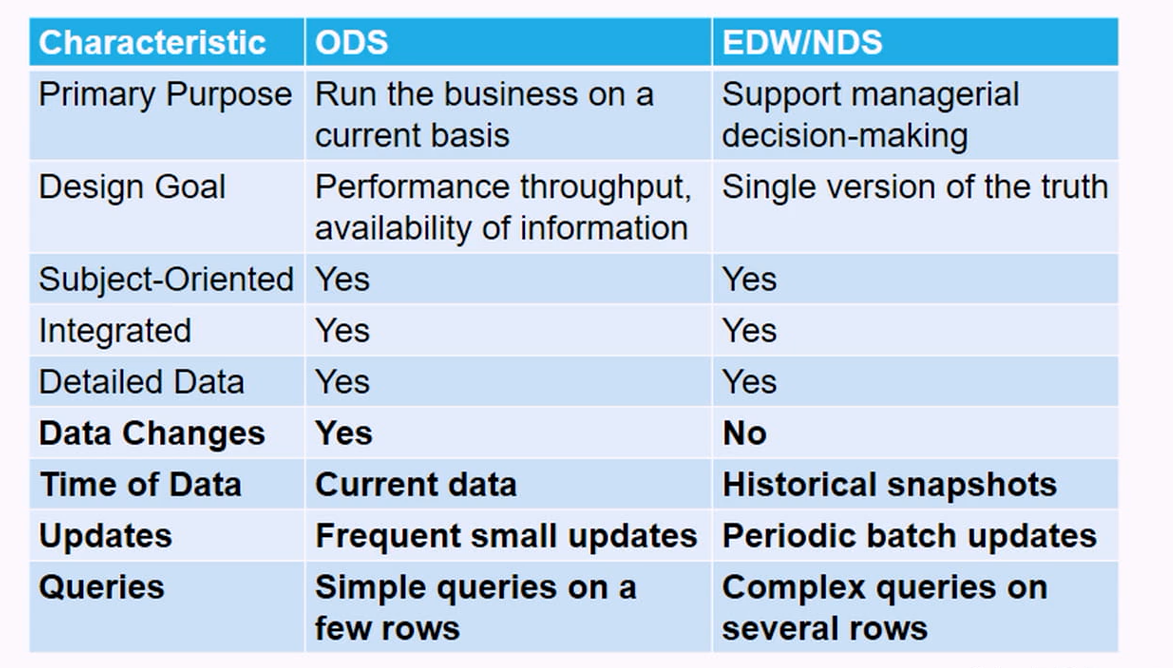
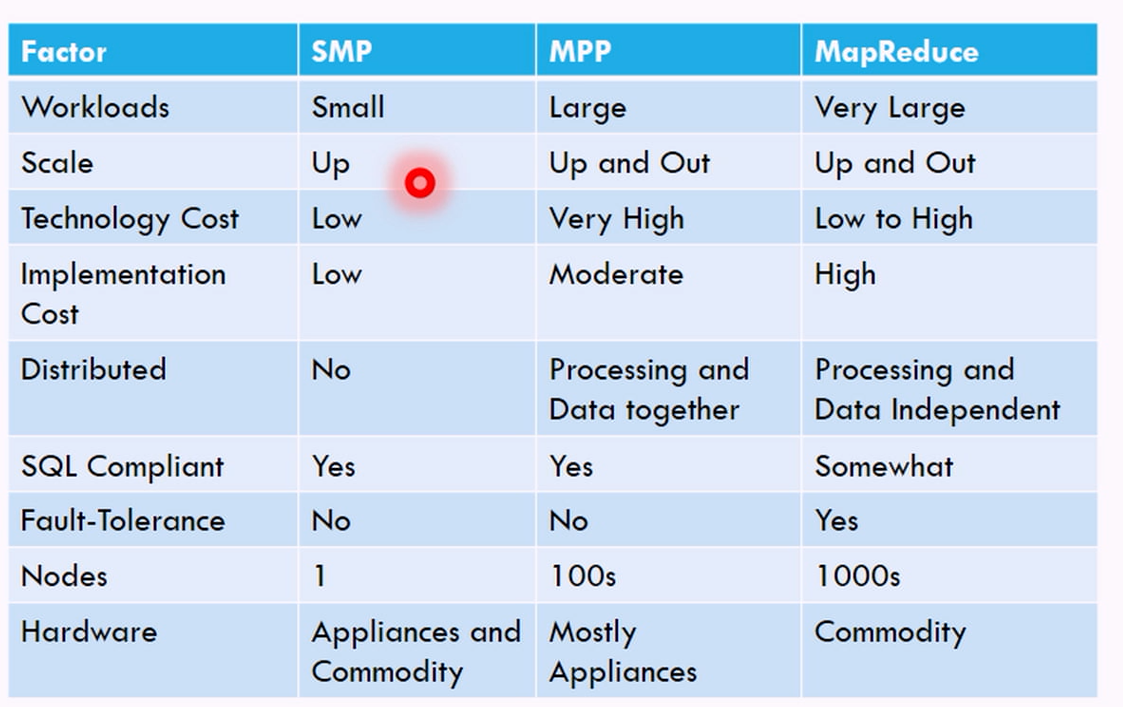
B - Data Warehouse Architectures

1. What is the difference between Systems Architecture and Technical Architecture? Video 2.2.1
   1. Answer:
      1. Systems Architecture: Physical configuration of technology and networks that support the technical architecture. Infrastructure that is a combination of hardware, software and networks.
      2. Technical Architecture: Logical architecture, discussing how data moves around. Movement of data from one data store ERP to other data store DW.
2. Components of Technical architecture. Video 2.3.1
   1. Answer:
      1. Data store: Typically stored in DMBS. Any place you can get data, HADOOP, multidimensional system. There are different types of data stores.
         1. User facing: available to end users for query.
         2. Internal: Used by the data warehouse only.
         3. Hybrid: combination of internal and user-facing
         4. External: not part of the warehouse
      2. ETL - Data in motion. Extract, transform and load. Process of moving data from one store to another.
3. What are the 4 types of data stores found in technical architectures? Video 2.3.5
   1. Answer: Staging is done first.
      1. Normalized data store: Internal data store, used as organization single version of the truth for the other systems. Subject oriented, integrated, non-volatile and time-variant. Stored in third normal form, which reduces redundancy. Grows in size due to historical data. Normalized data source is used as a source for data marts and the DDS data store. Common columns in NDS: Created on (when the data was added) and Last update (when the data was last changed).
      2. Operational data store: Hybrid data store, parts are internal and user-facing. Integrated, detailed, volatile and current data. Data is removed and updated to reflect current changes, consolidated from different sources. Does not grow over time, references a point in time, which is usually current. Should be stored as a separate DMBS as it is structured differently from NDS or DDS. Common columns in ODS: Last updated.
      3. Dimensional Data Store: User-facing data store. Subject-oriented, integrated, non-volatile and time-variant data from source systems. Grows in size over time due to historical data. Data are consolidated and denormalized. It is stored in dimensional format, uses a star-schema for a relational database and a cube for a multi-dimensional database. Common columns in DDS: Effective date, Expiration date and current Row. Reflects the changes to a product over time.
      4. MDS: MOLAP/Cube: When the DDS is multidimensional, it is called MOLAP. Multidimensional online analytical processing (MOLAP). Facts are pre-aggregated across all dimensions for improved performance. Is called a cube. Supports semantic meta-data (Sorts the day of the week, using hierarchy or a flow)
4. Review the 5 technical architectures discussed throughout the coursework. Video 2.4.1
   1. Answer: Complexity increases as we go down.
      1. Independent data marts: Least complex. Easy to get started with, difficult to scale. Departmentalized, and lacks enterprise focus. Has a source system with one or more DDS. Each DDS is its own system. No data consistency or data integration between data marts. Data marts do not share dimensions.
      2. Centralized: Data marts are consolidated into a single a DDS. Lack of integration among the dimensions, no data consistency between data marts. Has copies of the table.
      3. Enterprise bus architecture: All data marts in the DDS. Conformed dimensions, they are reused across the data marts. Single dimension for master data. Requires a holistic and enterprise focus to master data. Kimball technical architecture. Requires a lot of foresight.
         1. Enterprises with ODS – Includes an ODS for reporting on current, consolidated data. Conformed dimensions like enterprise bus.
      4. Hub and Spoke: Single version of the truth. Inmnon technical architecture. NDS is the single source and data gets send to dimensional data stores as needed. ODS can be added between stage and NDS like enterprise bus architecture. It can be used for other systems.
         1. Hub and Spoke with ODS – ODS is consolidated, current version of data. ODS is sourced from stage or NDS. ODS or stage can populate the NDS.
      5. Federated with ETL: Most complex architecture, cases where you have several data warehouse. Common during company takeover, where multiple data warehouses need to be merged. ETL unifies disparate sources into a single federated data warehouse. Used to integrate existing data marts, warehouses and legacy applications into a single federated data warehouse.
         1. Federated with EII – Enterprise application architecture. Software that can be purchased. Federated system is achieved through an EII software. Outputs are aggregated, on the fly.
5. Is the Corporate Information Factory a technical architecture? Explain. Video 2.4.5
   1. Answer: The corporate information factory (CIF) is an [enterprise data warehouse](https://www.sciencedirect.com/topics/computer-science/enterprise-data-warehouse) that follows a high-level data flow architecture advocated by Bill Inmon and Claudia. CIF is a hub and spoke architecture. Rectangle components – represent main components. Blobs components – represent applications, either external or decision support. Boxes components – represents business process. Cylinder components – represent data stores.
   2. External world – represent the applications in organization that are transactional in nature. They are outside the data warehouse and can be sourced from anywhere. ERP’s, business applications, internal data, logs and external data. These represents the data inputs/data sources for the CIF.
   3. Integration and Transformation layer – Sits between the external world and the enterprise data warehouse. Represents how the data is staged and the storage that would occur in the data warehouse. Takes unintegrated data from multiple sources and integrates and consolidates it. Computer programs are written to transform data from external world into the organizational data. Data comes from a variety of sources, need not be relational source, could be structured or unstructured. Adds structure to the unstructured data.
   4. ODS, EDW and Data marts – Inman only considers EDW to be the data warehouse. Inmon’s EDW is a DNS. His notion of data warehouse is just NDS. Most people consider the entire CIF as a data warehouse. EDW is the heart of the CIF. Grows in size over time due to historical data. Receives data from I&T layer and the ODS. ODS and EDW/NDS cannot share the same system as they are very different in processing. Data marts – an application of the data warehouse (Kimbel believes they are part of the data warehouse).



1. Systems architectures. Scalability increases as we go down:
   1. Answer:
      1. SMP – Symmetric multiprocessing. Single system with multiple CPU’s. Shared bus, memory and I/O. CPU’s share resources on a single system. Scales up but not out.
      2. MPP – Massively parallel processing. Nodes interconnected to form a single cluster. Multiple computers to appear as one. Single control node to orchestrate queries and data management. Processing and data partitions are tied together. Specialized hardware, difficult to scale out once configured.
      3. Hadoop Map Reduce/HDFS. Fault tolerant, used for Big data. Slower query execution than MPP. Runs on affordable commodity hardware.



1. Discuss the comparative success of the 5 technical architectures. Reading: “Which Data Warehouse Architecture Is Most Successful?”

The predominant architecture was the hub-and-spoke (39%), followed by the bus architecture (26%), centralized (17%), independent data marts (12%), and federated (4%). The most common platform for hosting the data warehouses was Oracle (41%), followed by Microsoft (19%) and IBM (18%). Independent data marts scored the lowest on all measures, which confirms the conventional wisdom that independent data marts are a poor architectural solution. The bus, hub-and-spoke, and centralized architectures earned similar scores on the success metrics. This finding helps explain why these competing architectures have survived over time—they are equally successful for their intended purposes. In terms of information and system quality and individual and organizational impacts, no single architecture is dominant. We found that the hub-and-spoke architecture is typically used with more enterprise wide implementations and larger warehouses. Overall, the hub-and-spoke architecture was the most expensive and time-consuming to implement. This is not surprising, however, because of the larger domain and size of these warehouses. The architecture also requires a considerable commitment to up-front planning, which takes time and money