Unit 02 Notes

Enterprise Information Landscape

Over the past decades, the explosive growth in storage and processing capabilities of computers, enterprise networks, as well as the internet, enabled businesses (and government organizations) to collect vast amounts of data. Historical data joined the operational transaction data at both summary and detailed granularity level. In the process, companies discovered that these data, augmented with appropriate analytical capabilities, can be a critical asset for decision making and other business processes and goals. As a result, they are taking increasing ownership of the quality and nature of their data, and work on leveraging it to their advantage.

However, the road to becoming a success story in the Big Data movement is not an easy one. Handling the data deluge to effectively support the business goals raises significant challenges. We will list such challenges under three categories: data-related, people-related, and analytics-related. One data-related challenge arises from the large number and variety of sources that produce the data: interactions between internal departments, and employees of the organization, the relations with customers, governments, and business partners, as well as track records of products and services offered by the company. This variety of **data sources** is mirrored by that of **data consumers**, thus generating many possible **data flows** in the system. In addition, data exist in various formats, both **structured** (database-like) and **unstructured** (documents, emails, spreadsheets, and embedded text).

Data producers and consumers add to the system their specific requirements related to local economy, politics, culture, and geography. Such people-related challenges are among the most difficult that confront large corporations today.

Many businesses were temporarily able to handle the ever-growing data with conventional data management systems. As a result, various parts of data ended up stored in separate systems and locations, with very little interlinking. Such data repositories were sometimes called “islands of information”, “pockets of information”, or “information silos”. In this infrastructure, answering different business questions often involves consulting different data repositories.

To address such challenges, the IT community developed new systems that bring together data from all over the enterprise. This type of system was termed **Data Warehouse (DW)**. The integrated data from data warehouses prompted new and creative **analytics** for discovering patterns and trends, generating reports, and using them to **improve decision making** (i.e., provide the right information to the right people at the right time). Improved decision making enabled businesses to better deal with growing consumer demand, competition, as well as operating complexities. These new analytical capabilities of DW-based systems were termed **Business Intelligence (BI)**. Recent **Commercial Off-the-Shelf (COTS)** applications have built-in analytic capabilities. The separate data repositories that are integrated in a DW, and augmented with analytic capabilities specific to their respective departments, were termed **Data Marts (DM)**.

We presented above the origins of three important keywords (Data Warehouse, Business Intelligence, and Data Mart). Today, these terms refer to components of information systems (sometimes called **information ecosystems**) that have complex and flexible **architectures**, with many components working together in various configurations. The components of the information ecosystem serve different departments or communities, while also working together for the integrated information ecosystem. One such architecture is the **Corporate Information Factory (CIF)** (Chapter 2, required reading).

A black-box representation of the CIF takes as input the raw data flowing into the company, and outputs information to consumers such as decision makers and Data Mart users. Raw data is collected by the **Applications** component (Chapter 4, optional reading). This component interacts directly with the data producers in the **External World** (Chapter 3, optional reading), gathers the data, audits them and prepares them for the **Integration and Transformation layer (I&T)**. This layer fundamentally integrates and transforms functional data into corporate data. The operational data passes then to the **Operational Data Store (ODS)** component, while the historical data passes to the **Data Warehouse** component. The latter can also receive data that passed through the ODS. The Data Warehouse data are ready to be delivered to **Data Marts** and **Decision-Support Systems (DSS)**. These components then deliver the information to their end users. Some complexities of the CIF design came from the necessity to handle the three types of data (**external data**, **reference data**, and **historical data),** and to support the four types of DSS users (**tourists**, **farmers**, **explorers**, and **miners**).

Architectures such as the CIF originated in different places and had different lifecycles over the years, and so did their names and definitions. Consequently, key terms such as DW, BI, DM, ODS, and DSS ended up having different definitions in different sources. Some examples follow:

- Inmon's definition of Data Warehouse imposes specific constraints on the nature of data stored in the system (i.e., subject-oriented, integrated, time-variant, and non-volatile). In contrast, Kimball's definition is purely functional (i.e., it only requires that the data can be queried).
- Inmon views the DW as the initial storage for data, which precedes and feeds the Data Marts customized to support individual business units. In contrast, Kimball views the Data Marts as the initial data storages, which precede and feed the DW that integrates them.
- Early data warehouse designs envisioned capabilities which later came to be termed Business Intelligence. In these designs, BI was a part of DW. In contrast, subsequent designs considered BI as an entity separate from DW. Today, some authors use the term use the DW/BI to refer to the entire system, whereas others use just BI for the same purpose.

Such architectural differences lead to differences in the lifecycles of the resulting systems. The data warehouse terminology features many terms whose definitions are in some other way or another less rigorous than we might expect. This should not come as a surprise in a relatively young area of IT.