The goal of the data warehouse is to use the data available in the enterprise and limited external data to yield business insight. In traditional settings, the volume of data is limited to a few terabytes, and the data is primarily operational and transactional in nature. Unstructured data is not available for analysis because of the limitations of the existing data warehouse. Consider by using big data technologies to incrementally augment with the existing data warehouse to gain better business insight from data that was previously unavailable for analysis.

View more content in this series
Many enterprises have information that is stored in various documents such as processes, certificates, legal information, contract documents, images, audio, and video that are stored in file systems or document management tools that are not part of the data warehouse data.

The data warehouse is limited by the amount of data it can handle. Often, because it cannot process unstructured data quickly, information other than transactional and operational data is underused. Content that is generated by Internet and mobile users is not available for analysis.

The existing data warehouse can be augmented by big data technology that can help get contextual information from massive volumes data, quickly, so that business decisions can be made in a timely manner. Big data technology can be used in many situations to augment the existing warehouse and make it possible to get better information to support more accurate decisions.

**Define initial data exploration with big data**

To get maximum value from big data, implement individual components of the big data platform. The existing reporting tools, extract, transform, and load (ETL) tools, and the data warehouse can be used along with the big data components. Figure 1 shows the traditional data warehouse architecture.

**Figure 1. Data warehouse architecture layers and tools**

Big data technology can improve the function of the data warehouse through *inbound augmentation* and *outbound augmentation*.

Inbound augmentation is the concept of receiving data from big data sources, digesting it, and passing it to the data source or staging area.

Outbound augmentation is the concept of using the data warehouse as one of the big data sources. The consumption layer can then connect to the big data platform to use the insight that is gleaned from big data.
Incorporate big data technology in the existing data warehouse

A big data infrastructure makes it possible to store and process massive volumes of data in a parallel and distributed way. The structured, semi-structured, or unstructured data can be at rest or in motion.

Data is stored in file systems such as Hadoop Distributed File System (HDFS), General Parallel File System (GPFS), document-oriented databases such as MongoDB, and other technologies. The data is processed by using a distributed processing model, such as MapReduce. Many tools are available to fetch the data from the distributed storage by using SQL-like programming models.

Identify and qualify the use cases for initial data exploration

The explosion of data as a result of Internet, sensors, and other advanced technologies offers many opportunities for businesses, industries, and scientific researchers to analyze the relevant data to help make better decisions.

Enterprises can begin to explore big data by augmenting the enterprise data warehouse to get more accurate business insight. They can start by augmenting ad hoc queries, standard reports, predictions, and perspective analytics. Figure 2 shows some of the sources of big data.

Figure 2. Sources of big data and the data warehouse for data exploration

In the following three situations, initial data exploration is performed by augmenting the existing data warehouse with big data technology.

Augment the data warehouse with social media data

Social media sites such as Facebook, LinkedIn, and Twitter have billions of users of all ages, millions of online transactions on e-commerce sites, various offers and retail discounts, and other types of information. This information is supplemented by channels such as smart devices,
including mobile phones that provide location-specific data. Some data providers such as gnip provide social media data (and other data) with required quality as a service.

This information can help businesses target the marketing and campaign efforts in the right direction by identifying the appropriate prospects. Using purchase history, prospect location, and customer profiles, businesses can identify groups or individuals who might be receptive to certain topics, products, or phrases. This group of prospects can provide a source of feedback for products and ideas.

The opportunity to gain valuable business insight is significant. Identifying the appropriate insight can lead to successful business decisions. To begin initial data exploration, follow these steps.

1. Acquire the data from social media sites, e-commerce sites, retail databases (for example, retail offers), and from smart devices, such as mobile phones.
2. Identify the entities from the data that is acquired in Step 1.
3. Augment these entities with the data marts in the data warehouse.
   a. Determine whether it is possible to use these entities as dimensions to the existing dimensions.
   b. Determine whether it is possible to add these entities as dimensions to the existing fact tables.
   c. Determine whether you need separate facts and dimensions to get the insight you need.
4. Identify the extra processing that is required, such as deduplication or analytic algorithms.
5. The existing reporting tools and products can still be used to present the insights.
6. Figure 3 shows the logical design for augmenting the data warehouse with social media.
Augment with environmental and geographical information

Combine environmental and geographical information with operational and transactional information to gain insight that can help predict crop yield, for example. Geographical and environmental information can be easily obtained from data providers.

Figure 4 shows the environmental and geographical data, the entities that you can derive from such sources, operational and business data from the data warehouse, and the benefits that can be derived from viewing the data in context. In this situation, data processing can be complex because of the velocity, volume, and variety of data. Big data tools are needed to perform the processing, which is contextual in nature, and to apply the required statistical algorithms.

Typically decisions about whether to import or export certain commodities are made based on ad hoc predictions gained from sample data. With the help of environmental, geographical, and smart devices, actual data can be collected so that the import and export decisions are more accurate. As shown in Figure 4, the environmental and geographical data can also help inform crop predictions that are based on the rainfall, temperature, and other environmental conditions.

Figure 4. Augment the data warehouse with environmental and geographical data
Combine the environmental and geographical data with the data warehouse data in two ways:

- **Augment the existing data warehouse with the environmental and geographical entities.** Only the entities that are resolved from the environmental and geographical data are augmented with the data warehouse schema. All the other tools such as reporting and custom logic is used as is. This approach yields limited insight because of the processing and presentation challenges.

- **Augment the environmental and geographical data with the transactional and operational data from the data warehouse.** Consider the set of operational and transactional data as one of the big data sources along with the set of environmental and geographical data. Use the big data platform and technologies to process and consume this data to yield more intuitive and advanced insight.

**Augment the data warehouse with unstructured data**

Consider the unstructured data, shown in Figure 5, that is not part of the warehouse schema.

**Figure 5. Augment the data warehouse with unstructured data**
Typically, organizations have a significant amount of data in the form of documents, images, audio, and video. Often, because this information is not available to the data warehouse, it is not available for analysis. Usually, these documents, such as images in insurance claims forms, must be scanned by humans. This unstructured information is scattered across document management systems, file systems, archived tapes, and sometimes, users' local hard disks. It gets purged because of data storage limitations.

This information can include valuable information that yields crucial insight. Using big data technology, you can store this information in a big data storage repository, as-is. Machine learning and MapReduce techniques can convert the data into structured data.

Combine the structured data with data warehouse data in two ways:

- Access the structured data and augment it with the data warehouse data at the data logic component in the step just before the data is consumed.
- Use data exploration to fetch the structured data and the data from the data warehouse for analysis.

**Summary**

Implementing a big data platform can be costly. From the perspective of the resources, skills, data, and tools that are required, many enterprises prefer incremental adoption of big data technology.
After you implement a single big data component, the organization can begin to take advantage of the benefit of better business insight.

One immediate application might be to begin to analyze the unused data inside the enterprise, data from social media and e-commerce sites, and environmental and geographical data.
### Resources

**Learn**

- **IBM Big data warehouse augmentation**: Video that describes various types of data warehouse augmentation scenarios.
- *Enterprise Information Protection - the Impact of Big Data* (Mike Ferguson, Intelligent Business Strategies): Provides the view on enterprise information security and privacy and the impact of big data on it.
- **IBM InfoSphere Data Explorer architecture**: IBM white paper provides an architectural view of the IBM® Data Explorer.
- **Why IBM for big data**: Describes the software that makes up the IBM big data platform.
- "**Big data architecture and patterns**" (Shweta Jain, Shrikant L Khupat, Divakar Mysore, IBM developerWorks, September 2013): Covers various architectural patterns with big data.
- Learn about **Elastic Storage**, a proven, scalable, high-performance data and file management solution (based on GPFS technology).

**Get products and technologies**

- Get geographical and weather information through an API from *World Weather Online*.

**Discuss**

- Ask questions and get answers in the *InfoSphere Data Explorer community*.
About the authors

Shrikant Khupat

Shrikant Khupat is an IBM application architect. He is experienced in defining enterprise class, distributed, disconnected, client/server architectures, and designs. He has exposure to various domains, such as insurance and energy and utilities. He worked on complex solutions that involve distributed data processing by using Apache Hadoop and unstructured data processing by using machine learning languages. His current interests include defining big data architecture and patterns.

Atul Ahire

Atul Ahire is an IBM Integration architect. He is TOGAF 8 certified. He is working as Integration Lead in Energy and Utilities sector with IBM and is extensively involved in Service Oriented Architecture [SOA] / Enterprise Application Design and Architecture solutions. He is interested in working with big data and defining big data architecture and patterns.

© Copyright IBM Corporation 2014
Trademark
(www.ibm.com/developerworks/ibm/trademarks/)