e-PGPathshala

Subject : Computer Science

Paper: Data Analytics

Module: Analytic Processes and its evolution

Module No: CS/DA/4

Quadrant 1 – e-text

1.1 Introduction

Methods used to perform data analytics is the process of examining data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. More than anything else, an analytical approach is the use of an appropriate process to break a problem down into the smaller pieces necessary to solve it. With big data analytics, data scientists and others can analyse huge volumes of data that conventional analytics and business intelligence solutions can't touch.

1.2 Learning Outcomes

- To understand the evolution of analytical processing
- To learn various configurations of analytic frameworks
- To know about types of analytic data sets

1.3 Evolution of analytic processing

Technology changes more rapidly than the processes in industry. Legacy processes for effective action of analytic routines aren’t able to take advantage of the current environment. If we change the key aspects of existing analytical processes, organizations will realize the gains in power and productivity that are possible with the new levels of scalability available today. It isn’t possible to tame big data using only traditional approaches to developing analytical processes.

Simpler and faster processing of relevant data, can use high-performance analytics. Using high-performance data mining, predictive analytics, text mining, forecasting and optimization on big data enables to continuously drive innovation and make the best possible decisions. In addition, organizations have started understanding machine learning techniques are ideally suited to address their fast-paced big data needs in new ways.

1.4 Analytic Framework

Analytic framework is a detailed sketch or outline of some social phenomenon, representing the initial idea of a scientist analyzing this phenomenon. Charles C. Ragin defines it as one of the four building blocks of social research (the other three being ideas (social theories), evidence (data) and images (new ideas synthetized from existing data). Thus, analytic frames are used to elaborate on starting ideas and usually consist of a list of some key elements found in most of the analysed phenomena (for example, social movements).

1.4.1The Analytic Framework configuration: Definition and Scope

An analytic framework configuration provides a set of resources which are useful to answer critical business questions. An analytic framework is ideal for data exploration, development of analytical processes, proof of concepts, and prototyping. Once things progress into ongoing, user-managed processes or production processes, then the configuration should not be involved. A configuration is going to be influenced by a fairly small set of users. There will be data created within the configuration that is segregated from the production database. Configuration users will also be allowed to load data of their own for brief time periods as part of a project. Data in a configuration will have a limited life. The data needed for the project will be built during the project. When that project is done, the data will be deleted. If used appropriately, a configuration has the capability to be a major driver of analytic value for an organization.

1.4.2 An Internal Configuration

For an internal configuration, a portion of an enterprise data warehouse or data mart is carved out to serve as the analytic configuration. In this case, the configuration is physically located on the production system. The configuration is a separate database container within the system as shown in figure 1.4.1.

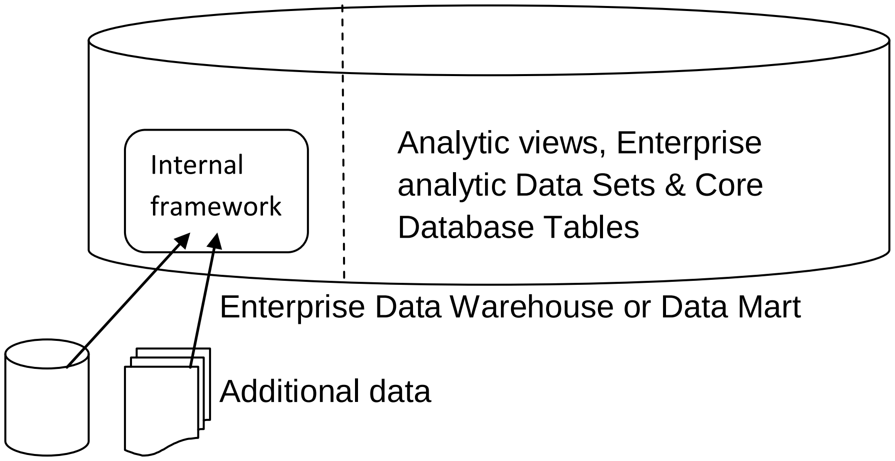
Internal framework
Analytic views, Enterprise analytic Data Sets & Core Database Tables
Enterprise Data Warehouse or Data Mart
Additional data

Figure 1.4.1 Detailed Internal Configuration View

Following are some of the pros and cons of internal configuration:

- Strength
  - Leverage existing hardware resources and infrastructure
  - Ability to directly join production data with analytical framework data
  - Cost-effective since no new hardware is needed
- Weaknesses
  - An additional load on the existing enterprise data warehouse or data mart
  - Can be constrained by production policies and procedures

1.4.3 An External Configuration

For an external configuration, as shown in figure 1.4.2, a physically separate analytic configuration is created for testing and development of analytic processes. It's relatively rare to have an environment that's purely external. Internal or hybrid configurations are more common. It is important to understand what the external configuration is, however, as it is a component of a hybrid configuration environment.

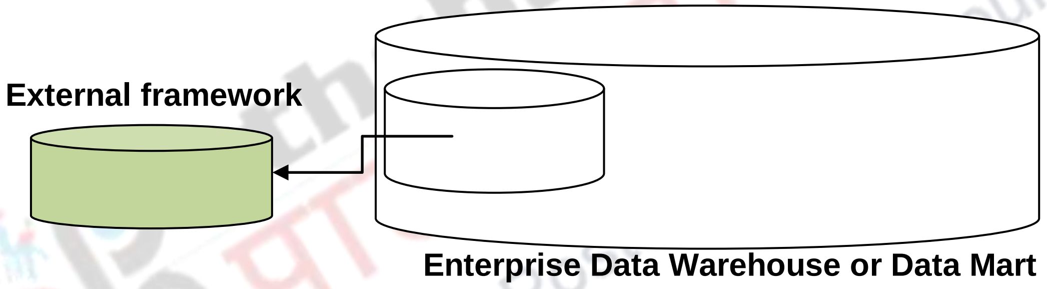

Figure 1.4.2 An External Configuration

Some of the pros and cons of external configuration are given below.

- Strength
  - A stand-alone environment, and it has no impact on other processes which is part of
  - Reduce workload management
- Weaknesses
  - The additional cost of the stand-alone system
  - Some data movement

1.4.4 A Hybrid Configuration

A hybrid configuration environment as given in figure 1.4.3 is the combination of an internal configuration and an external configuration. It allows analytic professionals the flexibility to use the power of the production system when needed, but also the flexibility of the external system for deep exploration or tasks that aren't as friendly to the database.

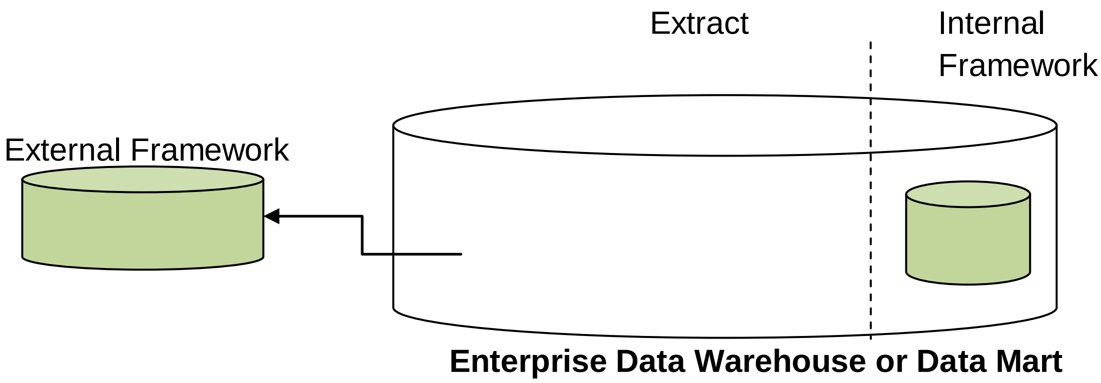

Figure 1.4.3 Hybrid Configuration

Some of the pros and cons of hybrid configuration are given below.

- Strength
  - Flexibility in the approach taken for an analysis
  - Can be run in a 'pseudo-production' mode temporarily
- Weaknesses
  - Maintain both an internal and external sandbox environment
  - Two-way data feeds may be required, which adds complexity

1.5 Analytic Data Set (ADS)

The data that is collected together to create an analysis or model is referred as an Analytic Data Set (ADS). It is data in the format required for the specific analysis work. ADS are generated by transforming, aggregating and reformatting the data. It may have a deformalized or flat file structure. That means there will be one record per every entity being analysed, that may be customer, location, product etc. The analytic data set helps to connect between efficient storage and ease of use.

1.5.1 Development versus Production Analytic Data Sets

There are two primary kinds of analytic data sets as shown in figure 1.5.1: Development ADS and Production ADS. To build an analytic process a development analytic data set is used. It will have all the unique variables that may be needed to solve a problem and will be very wide. A development analytic data set may have hundreds or thousands of variables or metrics. However, it is also fairly shallow, i.e., many modifications can be done on just a sample of data. This makes a development analytic data set very wide but not very deep.

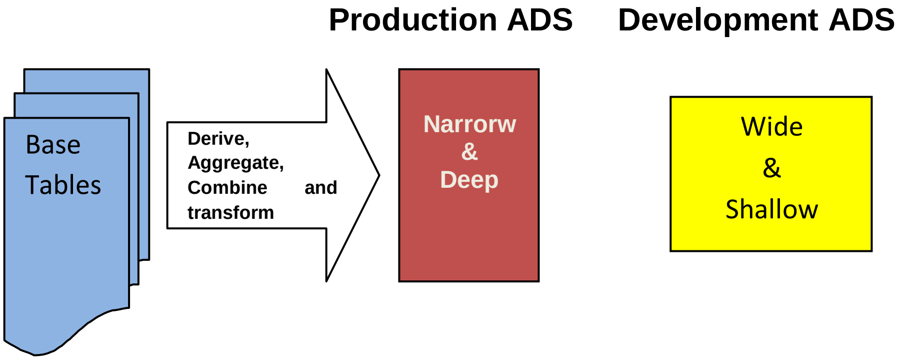

Figure 1.5.1 Development versus Production Analytic Data Sets

A production analytic data set, is useful for scoring and deployment. It has the specific metrics that were actually in the final solution. Most processes need only a small fraction of the metrics used during development. A big difference is that the scores need to be applied to every entity, not just for a sample. Every customer, every location, every product will need to be scored. Therefore, a production ADS is not going to be very wide, but it will be very deep.

1.5.2 Traditional Analytic Data Sets

In a traditional environment, as shown in figure 1.5.2, all analytic data sets are created outside of the database. Each analytic professional creates his or her own analytic data sets independently. This is done by every analytic professional, which means that there are possibly hundreds of people generating their own independent views of corporate data. An ADS is usually generated from scratch for each individual project.

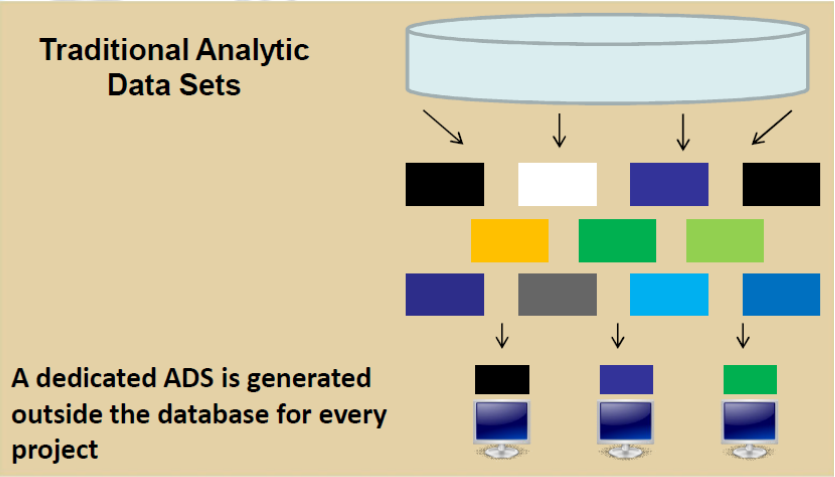
Traditional Analytic Data Sets
A dedicated ADS is generated outside the database for every project

Figure 1.5.2 Traditional analytic data set

The problem is not just that each analytic professional has a single copy the production data. Each analytic professional often makes new ADSs, and therefore a new copy of the data, for every project.

1.6 Enterprise Analytic Data Set

As models are built over time, certain standard metrics and manipulations become readily apparent. As an example, it's hard to imagine that total customer spending or number of customer transactions would not be of interest in most analysis efforts for a retailer. Similarly, it is hard to imagine that total product sales for recent periods would not be of interest to most product-level analytics. At the same time, any required cleansing or recoding of the detailed data required to facilitate such rollups will become constant once the right analytic procedures are established.

The various analytic processes can share the same and consistent set of metrics because an enterprise analytic data set is collaborative. It is going to simplify data access by making many metrics directly available to analytic professionals without further effort. They no longer have to go and navigate the raw third normal form tables and derive all the metrics themselves. It will greatly reduce time to results and it is a "build once, use many" useful achievement. As shown in figure 1.6.he centralized ADS tables (T1, T2, ..., Tn) and views are utilized across many projects (like p1, p2, p3, etc) as per the requirements and necessacity.

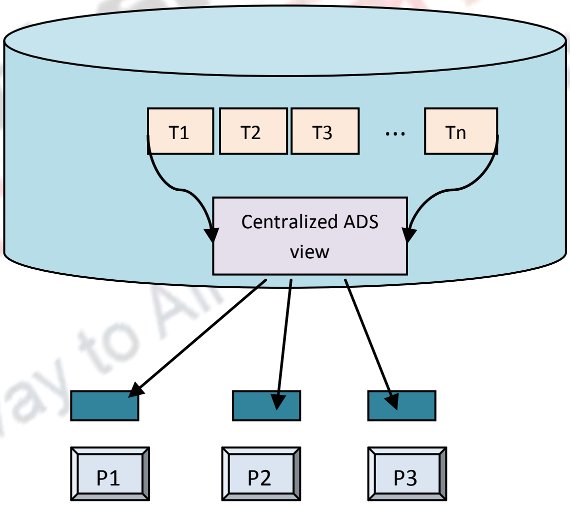

Figure 1.6.1 Enterprise Analytic Data Set

The consistency across analytic efforts is one of the most important benefits of an enterprise analytic data set. With greater consistency in the metrics feeding an organization's analytics, people can be comfortable that metrics feeding different processes were computed identically. Key features of an enterprise analytic data set include: A standardized view of data to support multiple analysis efforts. A method to greatly streamline the data preparation process. A way to provide greater consistency, accuracy, and visibility to analytics processes. A way to open new views of data to applications and users outside of the advanced analytics space.

1.6.1Updating an Enterprise Analytic Data Set

Enterprise analytic data set modification is one big driver. Different types of data should be updating with different frequencies. Sales metrics may be updated every day. Demographics may be updated quarterly. Survey information will not have any change. Based on that, it may be easier to have different types of data in different physical tables so that they can be updated independently. It will save system resources and reduce the overhead of a lot of extra metrics in a table when updating only a few. In addition, by having those separate tables or views, it makes it easy for analytic professionals to go and take the specific types of data that they want. Many databases have a limit to how many columns can be in one table. For a large EADS, multiple tables may be required to accommodate column limits if for no other reason. No matter how it is physically stored, views can be placed on top to gather together different data as needed.

1.7 Scoring

After a model creation based on historical data, it can then be applied to new data in order to make predictions about unseen behavior. This is what data mining (and more generally, predictive modeling) is all about. The process of using a model to make predictions about behavior that has yet to happen is called "scoring." The output of the model, the prediction, is called a score. Scores can take just about any form, from numbers to strings to entire data structures, but the most common scores are numbers (for example, the probability of responding to a particular promotional offer). Getting scoring processes embedded within a database environment has a number of benefits.

1. Scores run in batches will be available on demand.
2. Embedded scoring enables real-time scoring. This is especially important for situations such as web offers.
3. Embedded scoring will abstract complexity from users. It’s very easy for both individual users and applications to ask for a score. The system handles the heavy lifting. As a result, embedded scoring will make scores accessible to less technical people
4. A final benefit is having all the models contained in a centralized repository so they are all in one place.

There are four primary components required to effectively manage all of the analytic processes an enterprise develops. The components include analytic data set inputs, model definitions, model validation and reporting, and model scoring output. There are commercially available tools to help with model and score management, or a custom solution can be built to address an organization’s specific needs.

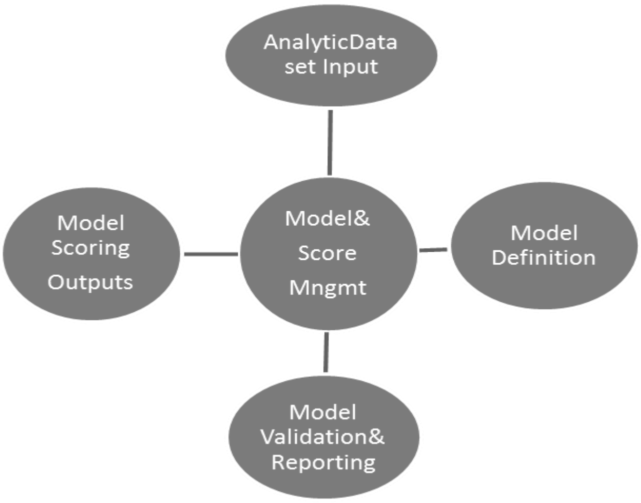

Figure 1.7.1 Model and Score Management Components

The reports will cover a range of topics and purposes. Information tracked includes:
- Reports that show how a specific run of scores compares to the development baselines.
- Specific summary statistics or validations, such as a lift or gains chart, that need to be reviewed after every scoring run.
- Model comparisons or variable distribution summaries.

When scores are updated, report output can be generated automatically. Such reports are often used for the critical step of monitoring the performance of a model over time. All models will degrade on time and business situations evolve. Reports will help identify when it's time to revisit a model.

1.7.1 Model Scoring Output

The last point is used to track model scores which is a result of scoring process. This is the real score score generated for every entity. The timestamp marking when a score was created and historical scores, as well as current scores are essential for many scenarios.

Case Studies

A) Teradata Enterprise Analytic Data Set (Source: www.teradata.com)

One major cellular company has created a 450-variable customer ADS in Teradata Database. By leveraging the standard data source, development of new models was cut from weeks to days.

One leading financial services company currently utilizes a Teradata ADS with 1,400 variables. This enabled them to complete an emergency analysis on exposure to Hurricane Katrina within hours, while their competition took weeks.

One well-known retailer has a Teradata ADS with 1,200 variables. The ADS was implemented as one component of an initiative that shortened model development from many weeks to days in most cases.

One major internet player maintains a large Enterprise ADS in their Teradata solution that allows them to access the data needed for new models within hours. They are now able to develop response models for new campaigns within days of execution.

Summary

- **D**ifferent configurations of analytical framework
- Analytical datasets are data that is pulled together in order to create an analysis or model
- Two types of analytical datasets