Notes of Data Science Unit III NEP 2023-24

Master of information technology (University of Mumbai)

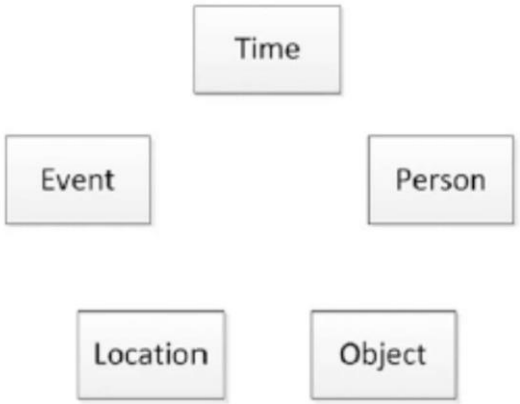
MODULE-3: Machine Learning for Data Science

Pushpa Mahapatro	**Do subscribe to my channel for useful educational contents: https://www.youtube.com/@mahapatrotutorials**

Process Superstep

The Process superstep adapts the assess results of the retrieve versions of the data sources into a highly structured data vault that will form the basic data structure for the rest of the data science steps. This data vault involves the formulation of a standard data amalgamation format across a range of projects.

The Process superstep is the amalgamation process that pipes your data sources into five main categories of data (below figure).



Five categories of data

Using only these five hubs in your data vault, and with good modeling, you can describe most activities of your customers. This enable you to then fine-tune your data science algorithms, to simply understand the five hubs' purpose and relationships that enable good data science.

1.0-Data Vault

Data vault modeling is a database modeling method designed by Dan Linstedt. The data structure is designed to be responsible for long-term historical storage of data from multiple operational systems. It supports chronological historical data tracking for full auditing and enables parallel loading of the structures.

1.1- Hubs

Data vault hubs contain a set of unique business keys that normally do not change over time and, therefore, are stable data entities to store in hubs. Hubs hold a surrogate key for each hub data entry and metadata labeling the source of the business key.

1.2-Links

Data vault links are associations between business keys. These links are essentially many-to-many joins, with additional metadata to enhance the particular link.

1.3-Satellites

Data vault satellites hold the chronological and descriptive characteristics for a specific section of business data. The hubs and links form the structure of the model but have no chronological characteristics and hold no descriptive characteristics. Satellites consist of characteristics and metadata linking them to their specific hub. Metadata labeling the origin of the association and characteristics, along with a time line with start and end dates for the characteristics, is put in safekeeping, for future use from the data section. Each satellite holds an entire chronological history of the data entities within the

specific satellite.

1.4-Reference Satellites

Reference satellites are referenced from satellites but under no circumstances bound with metadata for hub keys. They prevent redundant storage of reference characteristics that are used regularly by other satellites.

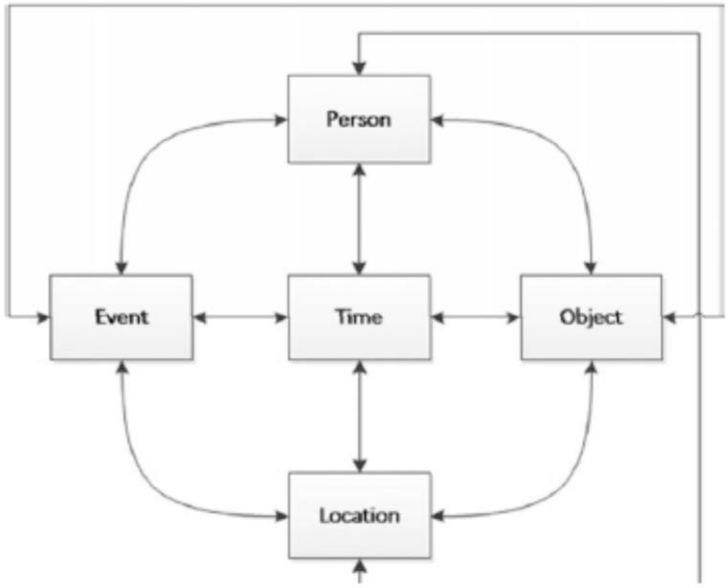
Typical reference satellites are

- *Standard codes:* These are codes such as ISO 3166 for country codes, ISO 4217 for currencies, and ISO 8601 for time zones.
- *Fixed lists for specific characteristics:* These can be standard lists that reduce other standard lists. For example, the list of countries your business has offices in may be a reduced fixed list from the ISO 3166 list. You can also generate your own list of, say, business regions, per your own reporting structures.
- *Conversion lookups:* Look at Global Positioning System (GPS) transformations, such as the Molodensky transformation, Helmert transformation, Molodensky-Badekas transformation, or Gauss- Kruger coordinate system.

  The most common is WGS84, the standard U.S. Department of Defense definition of a global location system for geospatial information, and is the reference system for the GPS.

2-Time-Person-Object-Location-Event Data Vault

The data vault we use is based on the Time-Person-Object-Location-Event (T-P-O-L-E) design principle



1-2. T-P-O-L-E—high-level design

2.1-Time Section

The time section contains the complete data structure for all data entities related to recording the time at which everything occurred.

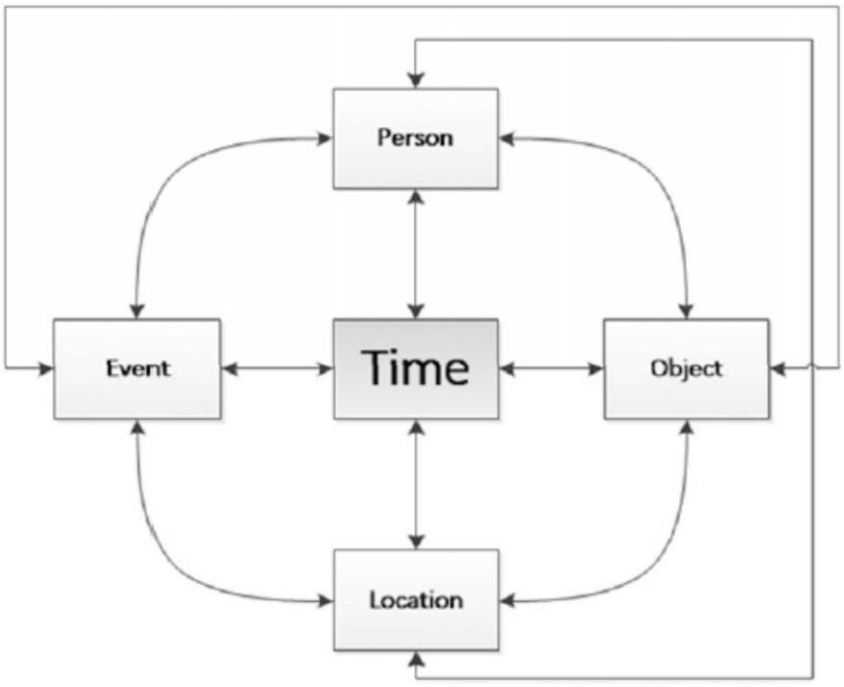
Time Hub

The time hub consists of the following fields:

```
CREATE TABLE [Hub-Time] (
    IDNumber      VARCHAR (100) PRIMARY KEY,
    IDTimeNumber  Integer,
    ZoneBaseKey   VARCHAR (100),
     DateTimeKey   VARCHAR (100),
     DateTimeValue DATETIME
 );
```

2.2-Time Links

The time links link the time hub to the other hubs.



9-3. Time links

The following links are supported.

Time-Person Link

This connects date-time values within the person hub to the time hub. The physical link structure is stored as a many-to-many relationship time within the data vault.

Dates such as birthdays, marriage anniversaries, and the date of reading this book can be recorded as separate links in the data vault. The normal format is BirthdayOn, MarriedOn, or ReadBookOn. The format is simply a pair of keys between the time and person hubs.

Time-Object Link
This connects date-time values within the object hub to the time hub. Dates such as those on which you bought a car, sold a car, and read this book can be recorded as separate links in the data vault. The normal format is BoughtCarOn, SoldCarOn, or ReadBookOn. The format is simply a pair of keys between the time and object hubs.

Time-Location Link
This connects date-time values in the location hub to the time hub. Dates such as moved to post code SW1, moved from post code SW1, and read book at post code SW1 can be recorded as separate links in the data vault. The normal format is MovedToPostCode, MovedFromPostCode, or ReadBookAtPostCode. The format is simply a pair of keys between the time and location hubs.

Time-Event Link
This connects date-time values in the event hub with the time hub. Dates such as those on which you have moved house and changed vehicles can be recorded as separate links in the data vault. The normal format is MoveHouse or ChangeVehicle. The format is simply a pair of keys between the time and event hubs.

Time Satellites
Time satellites are the part of the vault that stores the following fields.

2.3-Time Satellites

Time satellites are the part of the vault that stores the following fields.

```
CREATE TABLE [Satellite-Time-<Time Zone>] ( IDZoneNumber
                     VARCHAR (100) PRIMARY KEY,
     IDTimeNumber    INTEGER,
     ZoneBaseKey     VARCHAR (100),
     DateTimeKey     VARCHAR (100),
     UTCDateTimeValue  DATETIME,
     Zone            VARCHAR (100),
     DateTimeValue   DATETIME
);
```

Time satellites enable you to work more easily with international business patterns. You can move between time zones to look at such patterns as “In the morning . . . " or “At lunchtime . . . " These capabilities will be used during the Transform superstep, to discover patterns and behaviors around the world.

3- Person Section

The person section contains the complete data structure for all data entities related to recording the person involved.

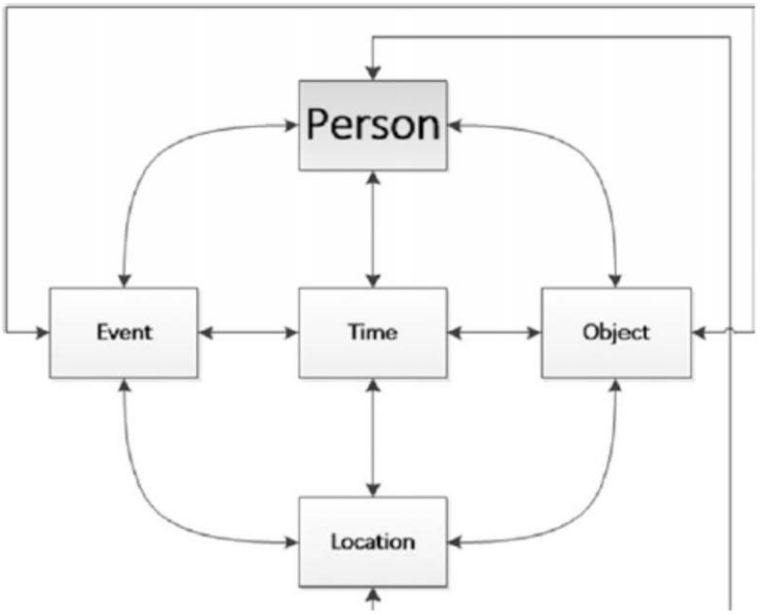
3.1-Person Hub

The person hub consists of a series of fields that supports a “real” person. The person hub consists of the following fields:

```
CREATE TABLE [Hub-Person] (
IDPersonNumber      INTEGER
FirstName           VARCHAR (200)
SecondName          VARCHAR (200)
LastName            VARCHAR (200)
Gender              VARCHAR (20),
TimeZone            VARCHAR (100)
BirthDateKey        VARCHAR (100)
BirthDate           DATETIME
);
```

3.2-Person Links

This links the person hub to the other hubs. (Figure 9-4).



9-4. Person links

The following links are supported in person links.

Person-Time Link

This link joins the person to the time hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Person-Time] (
    IDPersonNumber INTEGER,
    IDTimeNumber    INTEGER,
    ValidDate       DATETIME
);
```

Person-Object Link

This link joins the person to the object hub to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Person-Object] (
    IDPersonNumber   INTEGER,
    IDObjectNumber   INTEGER,
    ValidDate        DATETIME
);
```

Person-Location Link

This link joins the person to the location hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Person-Time] (
    IDPersonNumber       INTEGER,
    IDLocationNumber INTEGER,
    ValidDate            DATETIME
);
```

Person-Event Link

This link joins the person to the event hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Person-Time] (
    IDPersonNumber   INTEGER,
    IDEventNumber    INTEGER,
    ValidDate        DATETIME
);
```

3.3- Person Satellites

The person satellites are the part of the vault that stores the temporal attributes and descriptive attributes of the data. The satellite is of the following format:

```
CREATE TABLE [Satellite-Person-Gender] (
PersonSatelliteID VARCHAR (100),
IDPersonNumber INTEGER, FirstName VARCHAR (200), SecondName
VARCHAR (200), LastName VARCHAR (200), BirthDateKey VARCHAR
(20), Gender VARCHAR (10),
);
```

4- Object Section

The object section contains the complete data structure for all data entities related to recording the object involved.

4.1-Object Hub

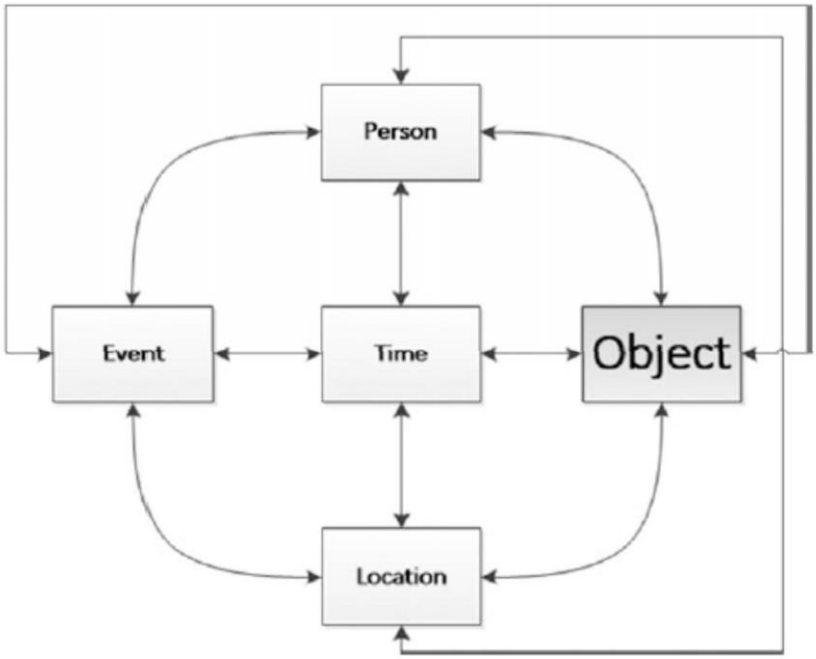
The object hub consists of a series of fields that supports a "real" object. The object hub consists of the following fields:

```
CREATE TABLE [Hub-Object-Species] (
IDObjectNumber INTEGER,
ObjectBaseKey VARCHAR (100),
ObjectNumber VARCHAR (100),
ObjectValue VARCHAR (200),
);
```

This structure enables you as a data scientist to categorize the objects in the business environment.

4.2-Object Links

These link the object hub to the other hubs (Figure 9-5)



9-5. Object links

The following links are supported:

Object-Time Link

This link joins the object to the time hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Object-Time] (
    IDObjectNumber INTEGER,
    IDTimeNumber INTEGER,
    ValidDate     DATETIME
);
```

Object-Person Link

This link joins the object to the person hub to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Object-Person] (
    IDObjectNumber INTEGER,
    IDPersonNumber INTEGER,
    ValidDate    DATETIME
);
```

Object-Location Link

This link joins the object to the location hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Object-Location] (
    IDObjectNumber INTEGER,
    IDLocationNumber INTEGER,
    ValidDate     DATETIME
);
```

Object-Event Link

This link joins the object to the event hub to describe the relationships between the two hubs.

4.3-Object Satellites

Object satellites are the part of the vault that stores and provisions the detailed characteristics of objects. The typical object satellite has the following data fields:

```
CREATE TABLE [Satellite-Object-Make-Model] (
IDObjectNumber INTEGER,
ObjectSatelliteID VARCHAR (200),
ObjectType VARCHAR (200),
ObjectKey VARCHAR (200),
ObjectUUID VARCHAR (200),
Make VARCHAR (200), Model VARCHAR (200)
);
```

The object satellites will hold additional characteristics, as each object requires additional information to describe the object. I keep each set separately, to ensure future expansion, as the objects are the one hub that evolves at a high rate of change, as new characteristics are discovered by your data science.

5- Location Section

The location section contains the complete data structure for all data entities related to recording the location involved.

5.1- Location Hub

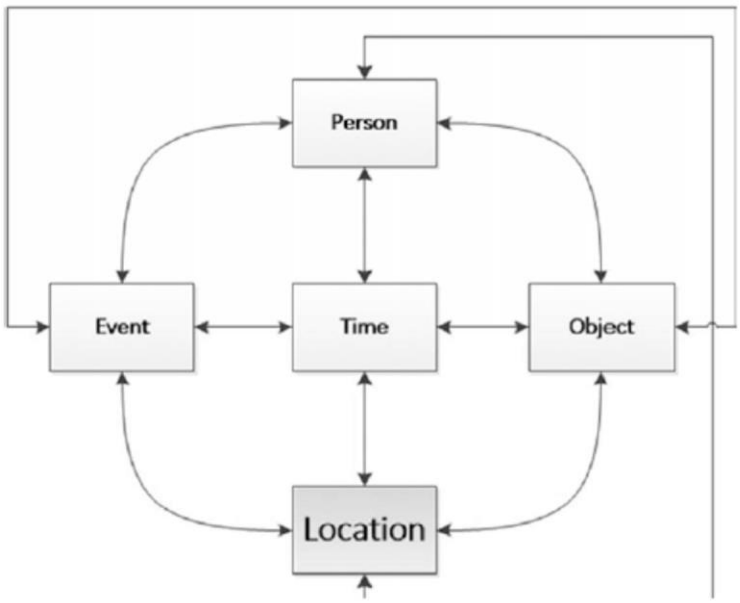
The location hub consists of a series of fields that supports a GPS location. The location hub consists of the following fields:

```
CREATE TABLE [Hub-Location] (
    IDLocationNumber INTEGER,
    ObjectBaseKey VARCHAR (200),
    LocationNumber INTEGER,
    LocationName VARCHAR (200),
    Longitude DECIMAL (9, 6),
    Latitude      DECIMAL (9, 6)
);
```

The location hub enables you to link any location, address, or geospatial information to the rest of the data vault.

5.2-Location Links

The location links join the location hub to the other hubs (Figure 9-6).



9-6. Location links

The following links are supported.

Location-Time Link

The link joins the location to the time hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Location-Time] (
    IDLocationNumber INTEGER,
    IDTimeNumber      INTEGER,
    ValidDate         DATETIME
);
```

These links support business actions such as **ArrivedAtShopAtDateTime** or **ShopOpensAtTime**.

Location-Person Link

This link joins the location to the person hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Location-Person] (
    IDLocationNumber INTEGER,
    IDPersonNumber    INTEGER,
    ValidDate         DATETIME
);
```

These links support such business actions as **`ManagerAtShop`** or **`SecurityAtShop`**.

Location-Object Link

This link joins the location to the object hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Location-Object] (
    IDLocationNumber INTEGER,
    IDObjectNumber    INTEGER,
    ValidDate         DATETIME
);
```

These links support such business actions as **`ShopDeliveryVan`** or **`RackAtShop`**.

Location-Event Link

This link joins the location to the event hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Location-Event] (
    IDLocationNumber INTEGER,
    IDEventNumber     INTEGER,
    ValidDate         DATETIME
);
```

These links support such business actions as **ShopOpened** or **PostCodeDeliveryStarted**.

5.3-Location Satellites

The location satellites are the part of the vault that stores and provisions the detailed characteristics of where entities are located. The typical location satellite has the following data fields:

```
CREATE TABLE [Satellite-Location-PostCode] (
IDLocationNumber  INTEGER,
LocationSatelliteID VARCHAR (200),
LocationType VARCHAR (200),
LocationKey VARCHAR (200),
LocationUUID VARCHAR (200),
CountryCode VARCHAR (20),
PostCode VARCHAR (200)
);
```

The location satellites will also hold additional characteristics that are related only to your specific customer. They may split their business areas into their own regions, e.g., **Europe**, **Middle-East**, and **China**. These can be added as a separate location satellite.

6-Event Section

The event section contains the complete data structure for all data entities related to recording the event that occurred.

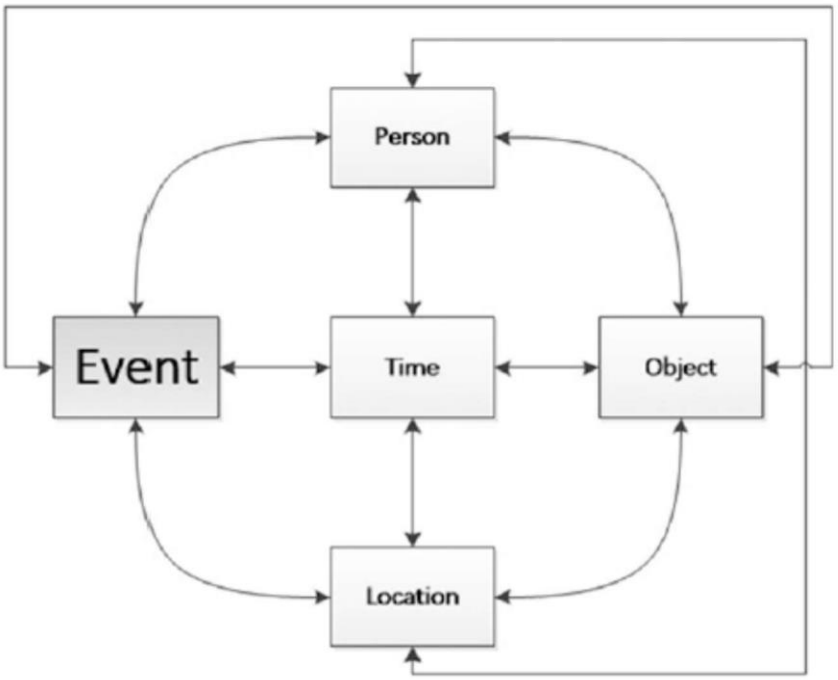
6.1-Event Hub

The event hub consists of a series of fields that supports events that happens in the real world. The event hub consists of the following fields:

```
CREATE TABLE [Hub-Event] (
    IDEventNumber INTEGER,
    EventType VARCHAR (200),
    EventDescription VARCHAR (200)
);
```

6.2-Event Links

Event links join the event hub to the other hubs (see Figure 9-7).



e 9-7. Event links

The following links are supported.

Event-Time Link

This link joins the event to the time hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Event-Time] (
    IDEventNumber INTEGER,
    IDTimeNumber    INTEGER,
    ValidDate       DATETIME
    );
```

These links support such business actions as **DeliveryDueAt** or **DeliveredAt**.

Event-Person Link

This link joins the event to the person hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Event-Person] (
    IDEventNumber INTEGER,
    IDPersonNumber INTEGER,
    ValidDate      DATETIME
);
```

These links support such business actions as **ManagerAppointAs** or **StaffMemberJoins**.

Event-Object Link

This link joins the event to the object hub, to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Event-Object] (
    IDEventNumber INTEGER,
    IDObjectNumber INTEGER,
    ValidDate      DATETIME
);
```

These links support such business actions as **VehicleBuy**, **VehicleSell**, or **ItemInStock**.

Event-Location Link

The link joins the event to the location hub to describe the relationships between the two hubs. The link consists of the following fields:

```
CREATE TABLE [Link-Event-Location] (
    IDEventNumber INTEGER,
    IDTimeNumber    INTEGER,
    ValidDate      DATETIME
);
```

These links support such business actions as **DeliveredAtPostCode** or **PickupFromGPS**.

6.3-Event Satellites

The event satellites are the part of the vault that stores the details related to all the events that occur within the systems you will analyze with your data science. I suggest that you keep to one type of event per satellite. This enables future expansion and easier longterm maintenance of the data vault.

Fishbone Diagrams

I have found this simple diagram is a useful tool to guide the questions that I need to resolve related to where each data source's data fits into the data vault. The diagram (Figure 9-10) is drawn up as you complete the 5 Whys process, as you will discover that there are normally many causes for why specific facts have been recorded.

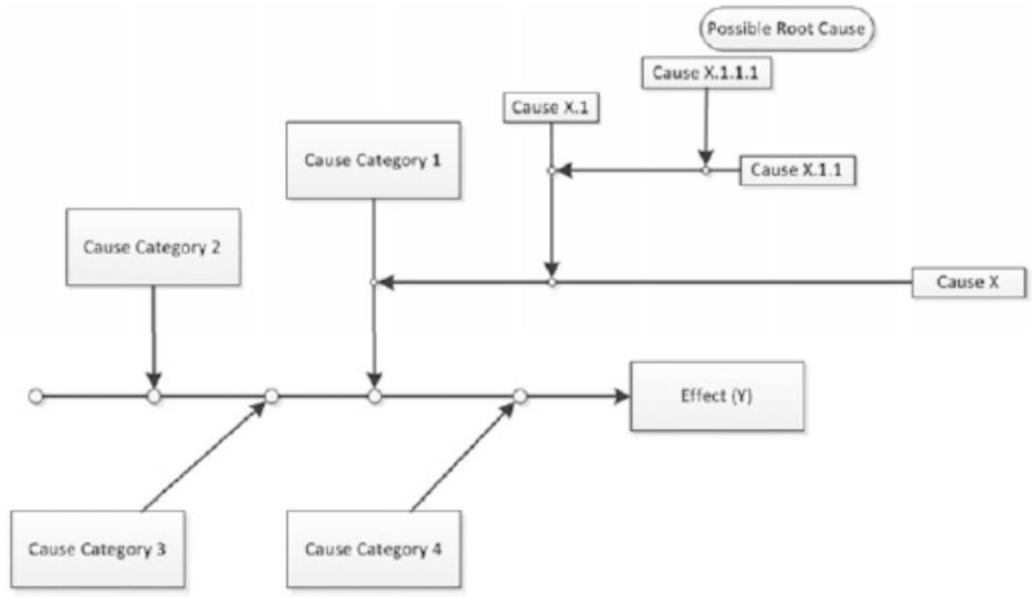


***Figure 9-10.** Fishbone diagram*

Let's return to my earlier event, the one in which I bought ten cans of beef curry. The ten cans are the effect (Y), but the four root causes of the purchase are

- I was hungry, so I bought ten tins. I did not like the brand of curry that I bought 10 cans of the previous week.
- My neighbor needed five cans, as she was no longer able to walk, and she requested the brand that I purchased, as it is the cheapest, at £1.10.
- I fed two cans to the dog, because I feel dog food is not nutritious, but I was not prepared to buy a more expensive brand of canned beef curry for the dog.
- I put three cans in the charity bin outside the local school, as it wanted that brand of curry for its soup kitchen and the tokens (vouchers) for its “Fund your School's new laptops” at £2.50 per can campaign.

This trivial sales disaster cost two data scientists and a buyer their jobs; a wholesaler was declared bankrupt; six people contracted food poisoning at the soup kitchen; and my dog had to be taken to the vet. The formula is simple: Dog + Curry = Sick puppy! I still place the liability on the shop; my wife disagrees.

5 Whys Example

So, let's look at the data science I performed with the retailer after the trouble with the cans of curry.

Problem Statement: Customers are unhappy because they are being shipped products that don't meet their specifications.

1. Why are customers being shipped bad products?

   • Because manufacturing built the products to a specification that is different from what the customer and the salesperson agreed to.

2. Why did manufacturing build the products to a different specification than that of sales?

   • Because the salesperson accelerates work on the shop floor by calling the head of manufacturing directly to begin work. An error occurred when the specifications were being communicated or written down.

3. Why does the salesperson call the head of manufacturing directly to start work instead of following the procedure established by the company?

   • Because the "start work" form requires the sales director's approval before work can begin and slows the manufacturing process (or stops it when the director is out of the office).

4. Why does the form contain an approval for the sales director?

   • Because the sales director must be continually updated on sales for discussions with the CEO, as my retailer customer was a top- ten key account.

In this case, only four whys were required to determine that a non-value-added signature authority helped to cause a process breakdown in the quality assurance for a key account! The rest was just criminal.

Monte Carlo Simulation

I want to introduce you to a powerful data science tool called a Monte Carlo simulation. This technique performs analysis by building models of possible results, by substituting a range of values—a probability distribution—for parameters that have inherent uncertainty. It then calculates results over and over, each time using a different set of random values from the probability functions. Depending on the number of uncertainties and the ranges specified for them, a Monte Carlo simulation can involve thousands or tens of thousands of recalculations before it is complete. Monte Carlo simulation produces distributions of possible outcome values. As a data scientist, this gives you an indication of how your model will react under real-life situations. It also gives the data scientist a tool to check complex systems, wherein the input parameters are high-volume or complex. I will show you a practical use of this tool in the next section.

Causal Loop Diagrams

A causal loop diagram (CLD) is a causal diagram that aids in visualizing how a number of variables in a system are interrelated and drive cause-and-effect processes. The diagram consists of a set of nodes and edges. Nodes represent the variables, and edges are the links that represent a connection or a relation between the two variables. I normally use my graph theory, covered in Chapter 8, to model the paths through the system. I simply create a directed graph and then step through it one step at a time.

Example: The challenge is to keep the "Number of Employees Available to Work and Productivity" as high as possible.

I modeled the following as a basic diagram, while I investigated the tricky process (Figure 9-11).

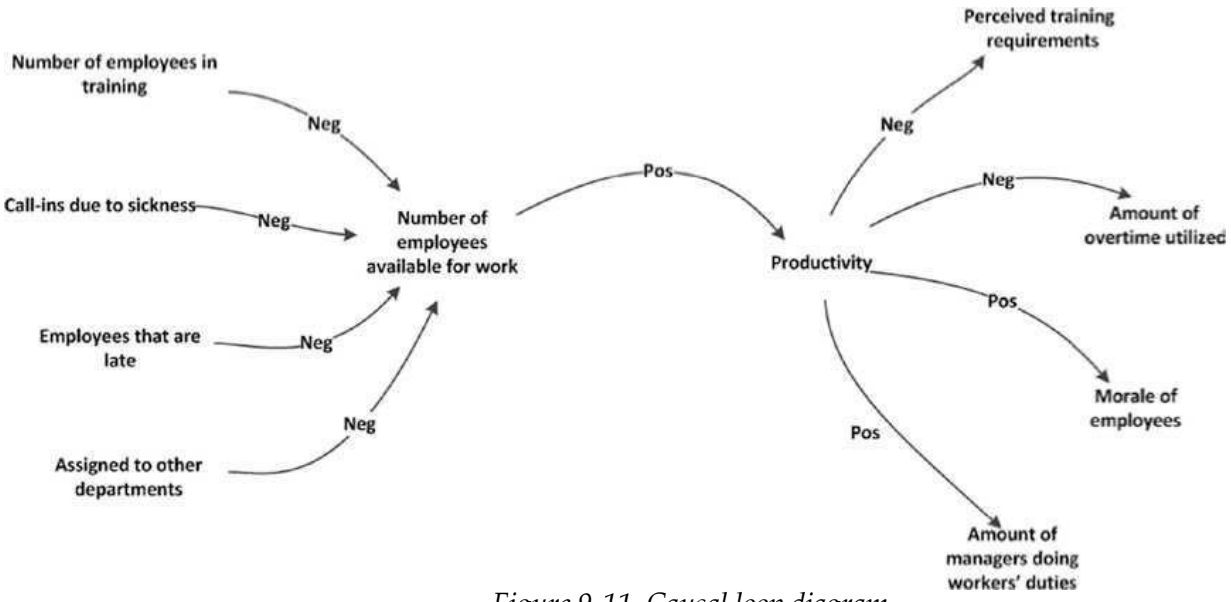


Figure 9-11. Causal loop diagram

Our first conclusion was "We need more staff." I then simulated the process, using a hybrid data science technique formulated from two other techniques—Monte Carlo simulation and Deep Reinforcement Learning—against a graph data model. The result was the discovery of several other additional pieces of the process that affected the outcome I was investigating. This was the result (Figure 9-12).

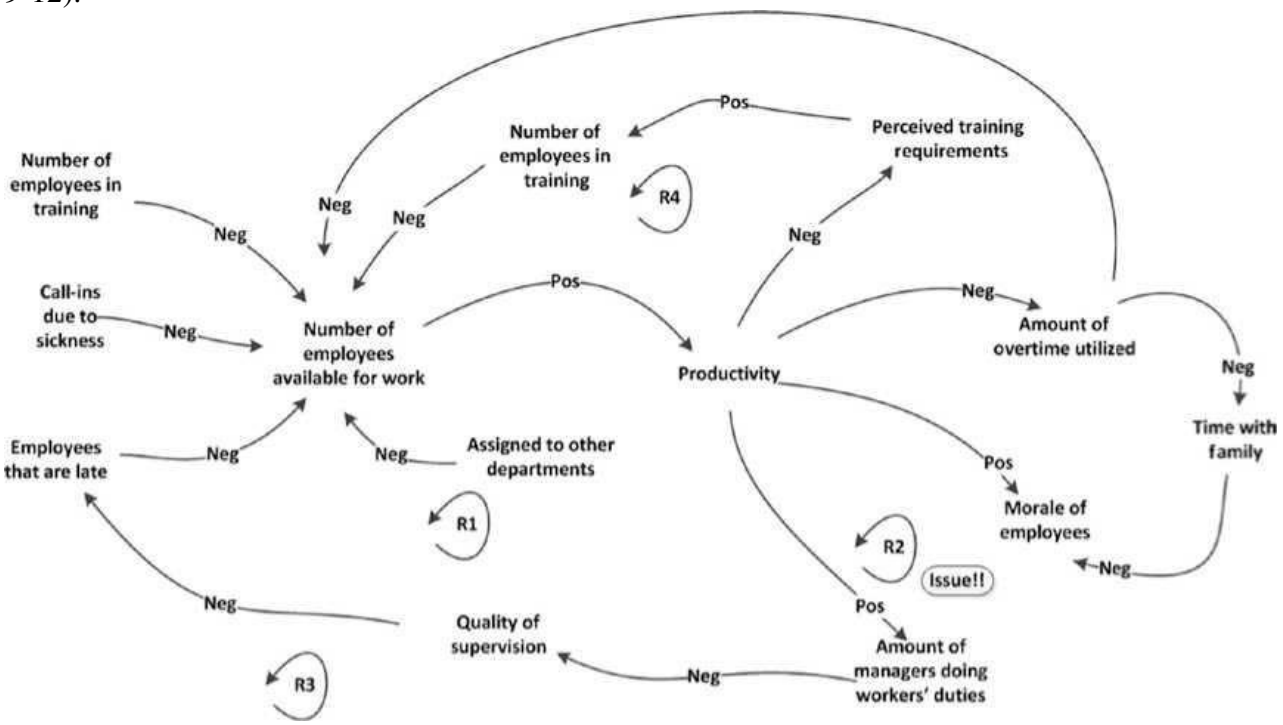


Figure 9-12. Monte Carlo result

The Monte Carlo simulation cycled through a cause-and-effect model by changing the values at points R1 through R4 for three hours, as a training set. The simulation was then run for three days with a reinforced deep learning model that adapted the key drivers in the system.

The result was “Managers need to manage not work." The R2—percentage of manage doing employees' duties—was the biggest cause and impact driver in the system.

Pareto Chart

Pareto charts are a technique I use to perform a rapid processing plan for the data science. Pareto charts can be constructed by segmenting the range of the data into groups (also called segments, bins, or categories). (See Figure 9-13.)

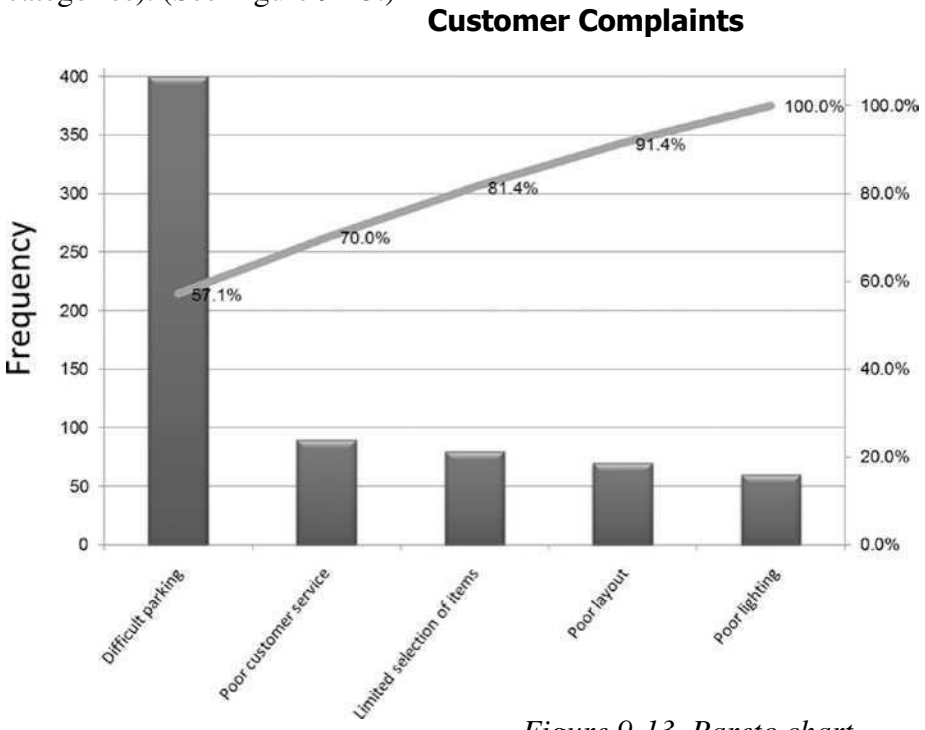


Figure 9-13. Pareto chart

Questions the Pareto chart answers:

- What are the largest issues facing our team or my customer's business?
- What 20% of sources are causing 80% of the problems (80/20 Rule)?
- Where should we focus our efforts to achieve the greatest improvements?

I perform a rapid assessment of the data science processes and determine what it will take to do 80% of the data science effectively and efficiently in the most rapid time frame. It is a maximum-gain technique.

Forecasting

Forecasting is the ability to project a possible future, by looking at historical data. The data vault enables these types of investigations, owing to the complete history it collects as it processes the source's systems data. You will perform many forecasting projects during your career as a data scientist and supply answers to such questions as the following:

- What should we buy?
- What should we sell?
- Where will our next business come from?

Data Science

You must understand that data science works best when you follow approved algorithms and techniques. You can experiment and wrangle the data, but in the end, you must verify and support your results. Applying good data science ensures that you can support your work with acceptable proofs and statistical tests. I believe that data science that works follows these basic steps:

1. Start with a question. Make sure you have fully addressed the 5 Whys.
2. Follow a good pattern to formulate a model. Formulate a model, guess a prototype for the data, and start a virtual simulation of the real-world parameters. If you have operational research or econometrics skills, this is where you will find use for them.

   Mix some mathematics and statistics into the solution, and you have the start of a data science model.

   Remember: All questions have to be related to the customer's business and must provide insights into the question that results in an actionable outcome and, normally, a quantifiable return-on- investment of the application of the data science processes you plan to deploy.
3. Gather observations and use them to generate a hypothesis. Start the investigation by collecting the required observations, as per your model. Process your model against the observations and prove your hypothesis to be true or false.
4. Use real-world evidence to judge the hypothesis. Relate the findings back to the real world and, through storytelling, convert the results into real-life business advice and insights.
5. Collaborate early and often with customers and with subject matter experts along the way.

Transform Superstep

The Transform superstep allows you, as a data scientist, to take data from the data vault and formulate answers to questions raised by your investigations. The transformation step is the data science process that converts results into insights.

It takes standard data science techniques and methods to attain insight and knowledge about the data that then can be transformed into actionable decisions, which, through storytelling, you can explain to non-data scientists what you have discovered in the data lake.

The Transform superstep uses the data vault from the process step as its source data. The transformations are tuned to work with the five dimensions of the data vault. As a reminder of what the structure looks like, see Figure 10-1.

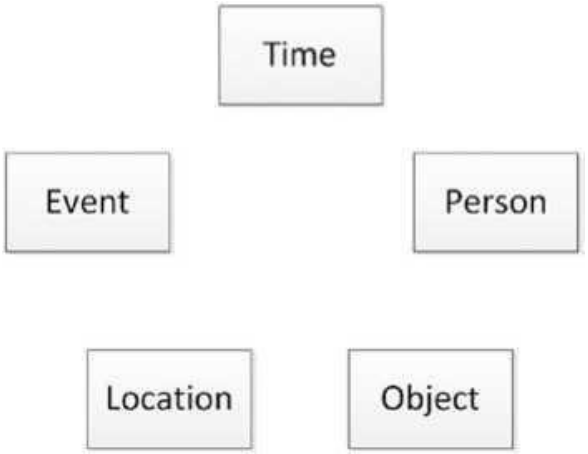


Figure 10-1. Five categories of data

1- Dimension Consolidation

The data vault consists of five categories of data, with linked relationships and additional characteristics in satellite hubs.

To perform dimension consolidation, you start with a given relationship in the data vault and construct a sun model for that relationship, as shown in Figure 10-2.

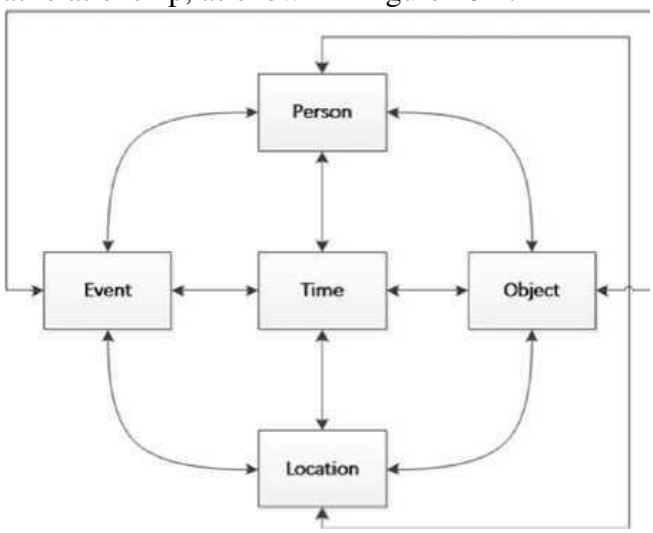


Figure 10-2. T-P-O-L-E hHigh-level design

2-Sun Model

The use of sun models is a technique that enables the data scientist to perform consistent dimension consolidation, by explaining the intended data relationship with the business, without exposing it to the technical details required to complete the transformation processing. So, let's revisit our business statement: GuSmundur Gunnarsson was born on December 20, 1960, at 9:15 in Landspitali, Hringbraut 101, 101 Reykjavik, Iceland.

2.1-Person-to-Time Sun Model

The following sun model in Figure 10-3 explains the relationship between the Time and Person categories in the data vault.

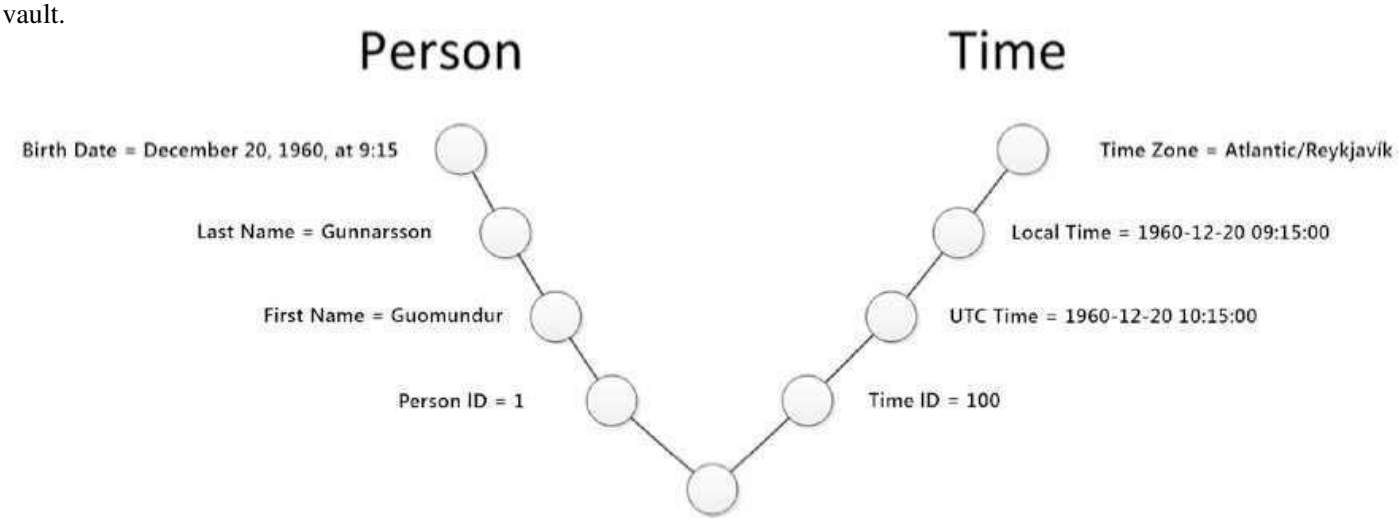


Figure 10-3. Person-to-Time sun model

The sun model is constructed to show all the characteristics from the two data vault hub categories you are planning to extract. It explains how you will create two dimensions and a fact via the Transform step from Figure 10-3. You will create two dimensions (Person and Time) with one fact (PersonBornAtTime), as shown in Figure 10-4.

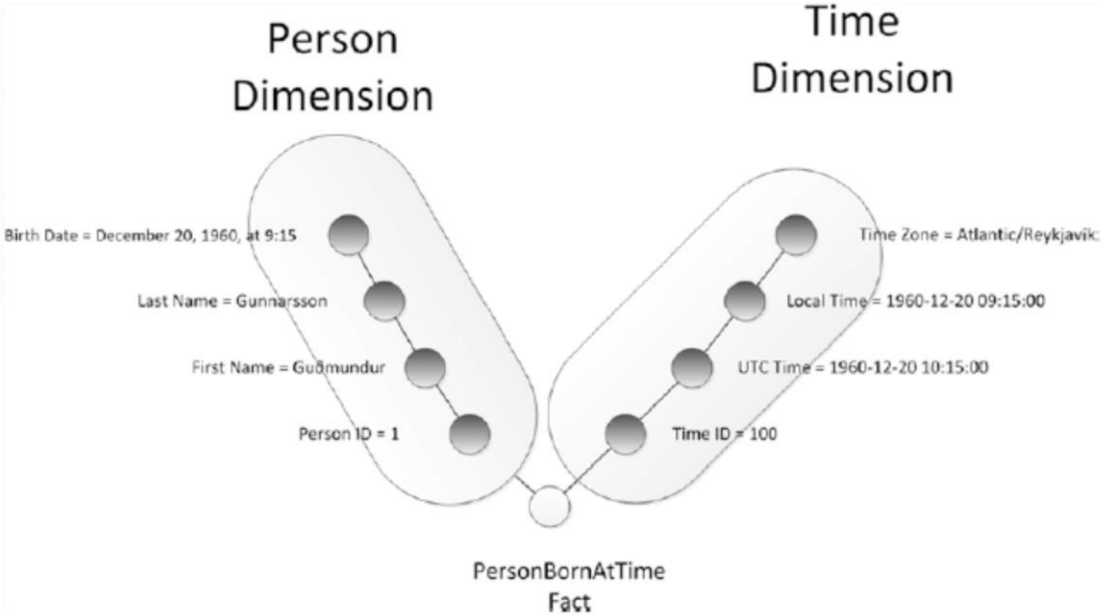


Figure 10-4. Person-to-Time sun model (explained)

2.2-Template Model

I also have several printed copies of sun models with blank entries (Figure 10-5), to quickly record relationships while I am meeting with clients.

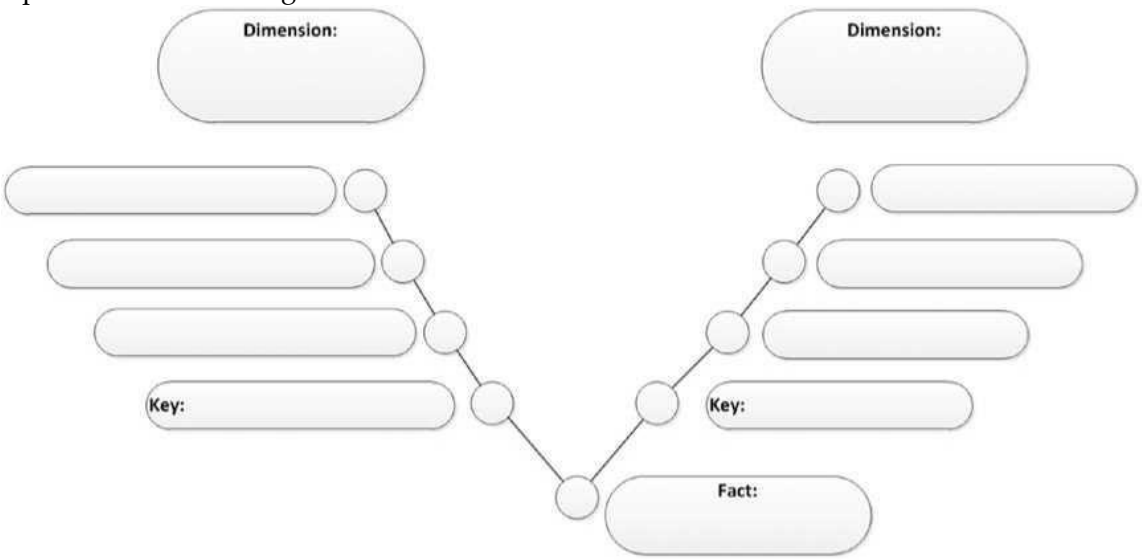


Figure 10-5. Template for a simple sun model

2.3-Person-to-Object Sun Model

Can you find the dimensions and facts on these already completed sun models? (See Figure 10-6.)

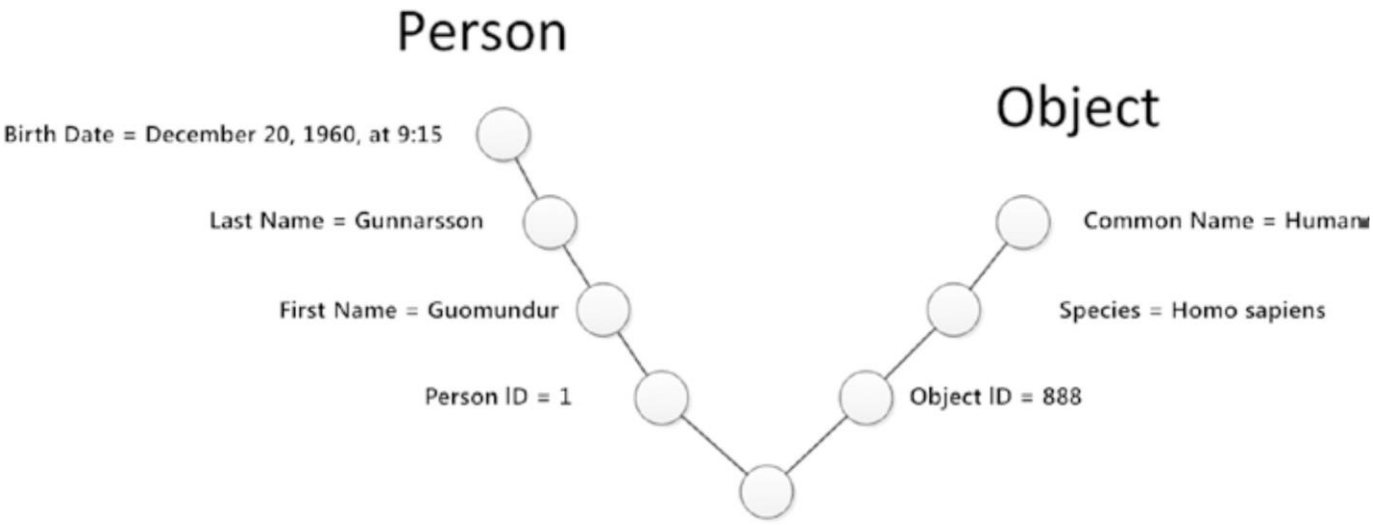


Figure 10-6. *Sun model for the PersonIsSpecies fact*

How did you progress? In Figure 10-6, dimensions are Person and Object. Fact is PersonIsSpecies. This describes Guðmundur Gunnarsson as a *Homo sapiens*.

2.4-Person-to-Location Sun Model

If you have dimensions Person and Location and fact PersonAtLocation, you have successfully read the sun model. (Figure 10-7).

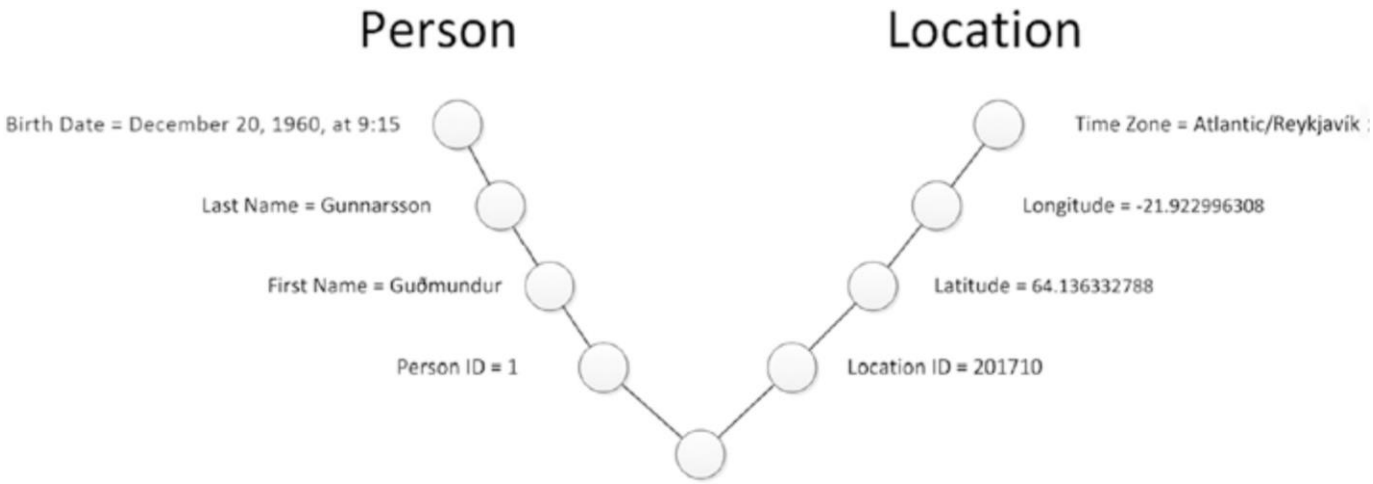


Figure 10-7. *Sun model for PersonAtLocation fact*

This describes that Guðmundur Gunnarsson was born at:

Latitude: 64° 08' 10.80" N or 64.136332788
Longitude: -21° 55' 22.79" W or -21.922996308

2.5-Person-to-Event Sun Model

And this event sun model? Can you extract the information in Figure 10-8?

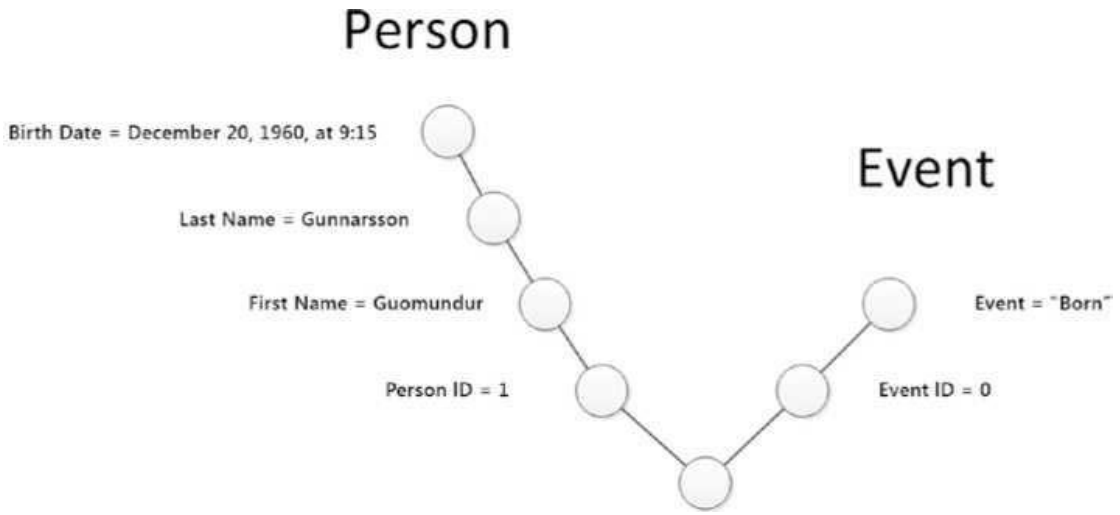


Figure 10-8. Sun model for PersonBorn fact

If you have dimensions Person and Event and fact PersonBorn, you have successfully read the sun model (Figure 10-8).

2.6-Sun Model to Transform Step

I will guide you through the creation of the Transform step, as shown by the sun model in Figure 10-9.

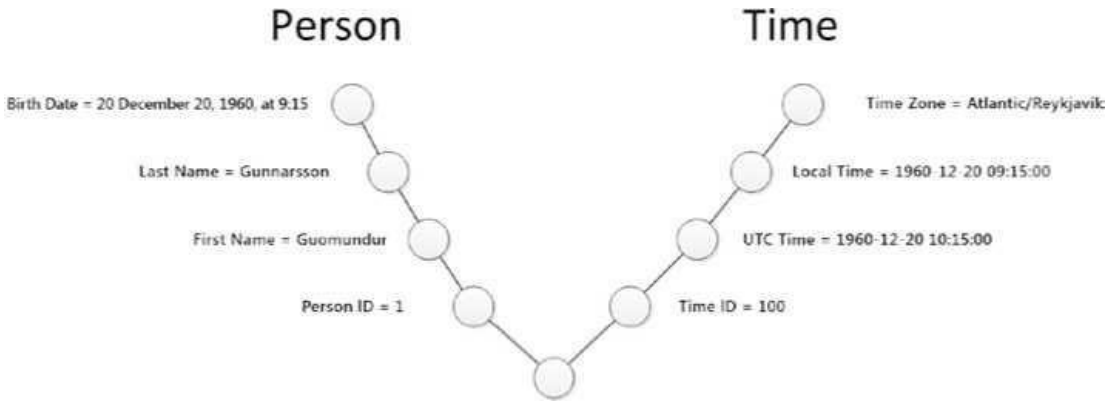


Figure 10-9. Sun model for PersonBornAtTime fact

You must build three items: dimension Person, dimension Time, and fact PersonBornAtTime.

3-Transforming with Data Science

You now have a good basis for data exploration and preparation from the data lake into data vault and from the data vault to the data warehouse. I will now introduce you to the basic data science to transform your data into insights. You must understand a selected set of basic investigation practices, to gain insights from your data.

3.1-Steps of Data Exploration and Preparation

You must keep detailed notes of what techniques you employed to prepare the data. Make sure you keep your data traceability matrix up to date after each data engineering step has completed. Update your Data Lineage and Data Providence, to ensure that you have both the technical and business details for the entire process. Now, I will take you through a small number of the standard transform checkpoints, to ensure that your data science is complete.

Missing Value Treatment

You must describe in detail what the missing value treatments are for the data lake transformation. Make sure you take your business community with you along the j ourney. At the end of the process, they must trust your techniques and results. If they trust the process, they will implement the business decisions that you, as a data scientist, aspire to achieve.

Why Missing Value Treatment Is Required

Explain with notes on the data traceability matrix why there is missing data in the data lake. Remember: Every inconsistency in the data lake is conceivably the missing insight your customer is seeking from you as a data scientist. So, find them and explain them. Your customer will exploit them for business value.
Why Data Has Missing Values

The 5 Whys is the technique that helps you to get to the root cause of your analysis. The use of cause-and-effect fishbone diagrams will assist you to resolve those questions.

I have found the following common reasons for missing data:

- Data fields renamed during upgrades

- Migration processes from old systems to new systems where mappings were incomplete
- Incorrect tables supplied in loading specifications by subject-matter expert
- Data simply not recorded, as it was not available
- Legal reasons, owing to data protection legislation, such as the General Data Protection Regulation (GDPR), resulting in a not-to- process tag on the data entry
- Someone else's "bad" data science. People and projects make mistakes, and you will have to fix their errors in your own data science.

What Methods Treat Missing Values?

During your progress through the supersteps, you have used many techniques to resolve missing data. Record them in your lineage, but also make sure you collect precisely how each technique applies to the processing flow.

3.2-Techniques of Outlier Detection and Treatment

During the processing, you will have detected several outliers that are not complying with your expected ranges, e.g., you expected "Yes" or "No" but found some "N/A"s, or you expected number ranges between 1 and 10 but got 11, 12, and 13 also. These out of order items are the outliers.

I suggest you treat them as you treat the missing data. Make sure that your customer agrees with the process, as it will affect the insights you will process and their decisions.

Elliptic Envelope

I will introduce a function called **EllipticEnvelope**. The basic idea is to assume that a data set is from a known distribution and then evaluate any entries not complying to that assumption. Fitting an elliptic envelope is one of the more common techniques used to detect outliers in a Gaussian distributed data set.

The **scikit-learn** package provides an object **covariance.EllipticEnvelope** that fits a robust covariance estimate to the data, and thus fits an ellipse to the central data points, ignoring points outside the central mode. For instance, if the inlier data are Gaussian distributed, it will estimate the inlier location and covariance in a robust way (i.e., without being influenced by outliers). The Mahalanobis distances obtained from this estimate are used to derive a measure of outlyingness.

Isolation Forest

One efficient way of performing outlier detection in high-dimensional data sets is to use random forests. The ensemble.IsolationForest tool "isolates" observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

Because recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of normality and our decision function. Random partitioning produces a noticeably shorter path for anomalies. Hence, when a forest of random trees collectively produces shorter path lengths for particular samples, they are highly likely to be anomalies.

Novelty Detection

Novelty detection simply performs an evaluation in which we add one more observation to a data set. Is the new observation so different from the others that we can doubt that it is regular? (I.e., does it come from the same distribution?) Or, on the contrary, is it so similar to the other that we cannot distinguish it from the original observations? This is the question addressed by the novelty detection tools and methods.

The sklearn.svm.OneClassSVM *tool is a good example of this unsupervised outlier detection technique.*

Local Outlier Factor

An efficient way to perform outlier detection on moderately high-dimensional data sets is to use the local outlier factor (LOF) algorithm. The **neighbors.LocalOutlierFactor** algorithm computes a score (called a local outlier factor) reflecting the degree of abnormality of the observations. It measures the local density deviation of a given data point with respect to its neighbors. The idea is to detect the samples that have a substantially lower density than their neighbors.

In practice, the local density is obtained from the k-nearest neighbors. The LOF score of an observation is equal to the ratio of the average local density of its k-nearest neighbors and its own local density. A normal instance is expected to have a local density like that of its neighbors, while abnormal data are expected to have a much smaller local density.

When the amount of contamination is known, this algorithm illustrates three different ways of performing-based on a robust estimator of covariance, which assumes that the data are Gaussian distributed-and performs better than the one-class SVM, in that case. The first is the one-class SVM, which has the ability to capture the shape of the data set and, hence, perform better when the data is strongly non-Gaussian, i.e., with two well-separated clusters. The second is the isolation forest algorithm, which is based on random forests and, hence, better adapted to large-dimensional settings, even if it performs quite well in the example you will perform next. Third is the local outlier factor to measure the local deviation of a given data point with respect to its neighbors, by comparing their local density.

Example:

Here, the underlying truth about inliers and outliers is given by the points' colors. The orange-filled area indicates which points are reported as inliers by each method.

3.4-What Is Feature Engineering?

Feature engineering is your core technique to determine the important data characteristics in the data lake and ensure they get the correct treatment through the steps of processing. Make sure that any featuring extraction process technique is documented in the data transformation matrix and the data lineage.

4- Common Feature Extraction Techniques

I will introduce you to several common feature extraction techniques that will help you to enhance any existing data warehouse, by applying data science to the data in the warehouse.

4.1-Binning

Binning is a technique that is used to reduce the complexity of data sets, to enable the data scientist to evaluate the data with an organized grouping technique. Binning is a good way for you to turn continuous data into a data set that has specific features that you can evaluate for patterns. A simple example is the

cost of candy in your local store, which might range anywhere from a penny to ten dollars, but if you subgroup the price into, say, a rounded-up value that then gives you a range of five values against five hundred, you have just reduced your processing complexity to 1/500th of what it was before.

4.2-Averaging

The use of averaging enables you to reduce the amount of records you require to report any activity that demands a more indicative, rather than a precise, total.

4.3-Latent Dirichlet Allocation (LDA)

A latent Dirichlet allocation (LDA) is a statistical model that allows sets of observations to be explained by unobserved groups that elucidates why they match or belong together within text documents. This technique is useful when investigating text from a collection of documents that are common in the data lake, as companies store all their correspondence in a data lake. This model is also useful for Twitter or e-mail analysis.

5-Hypothesis Testing

Hypothesis testing is not precisely an algorithm, but it's a must-know for any data scientist. You cannot progress until you have thoroughly mastered this technique.

Hypothesis testing is the process by which statistical tests are used to check if a hypothesis is true, by using data. Based on hypothetical testing, data scientists choose to accept or reject the hypothesis. When an event occurs, it can be a trend or happen by chance. To check whether the event is an important occurrence or just happenstance, hypothesis testing is necessary.

There are many tests for hypothesis testing, but the following two are most popular.

5.1-T-Test

A t-test is a popular statistical test to make inferences about single means or inferences about two means or variances, to check if the two groups' means are statistically different from each other, where $n < 30$ and standard deviation is unknown.

5.2-Chi-Square Test

A chi-square (or squared [x^2]) test is used to examine if two distributions of categorical variables are significantly different from each other.

6-Overfitting and Underfitting

Overfitting and underfitting are major problems when data scientists retrieve data insights from the data sets they are investigating. Overfitting is when the data scientist generates a model to fit a training set perfectly, but it does not generalize well against an unknown future real-world data set, as the data science is so tightly modeled against the known data set, the most minor outlier simply does not get classified correctly. The solution only works for the specific data set and no other data set. For example, if a person earns more than $150,000, that person is rich; otherwise, the person is poor. A binary classification of rich or poor will not work, as can a person earning about $145,000 be poor?

Underfitting the data scientist's results into the data insights has been so nonspecific that to some extent predictive models are inappropriately applied or questionable as regards to insights. For example, your person classifier has a 48% success rate to determine the sex of a person. That will never work, as with a binary guess, you could achieve a 50% rating by simply guessing.

Your data science must offer a significant level of insight for you to secure the trust of your customers, so they can confidently take business decisions, based on the insights you provide them.

6.1-Polynomial Features

The polynomic formula is the following: $(a_1 x + b_1)(a_2 x + b_2) = a_1 a_2 x^2 + (a_1 b_2 + a_2 b_1)x + b_r b_2$. The polynomial feature extraction can use a chain of polynomic formulas to create a hyperplane that will subdivide any data sets into the correct cluster groups. The higher the polynomic complexity, the more precise the result that can be achieved.

6.2-Common Data-Fitting Issue

These higher order polynomic formulas are, however, more prone to overfitting, while lower order formulas are more likely to underfit. It is a delicate balance between two extremes that support good data science.

7-Precision-Recall

Precision-recall is a useful measure for successfully predicting when classes are extremely imbalanced. In information retrieval,

- Precision is a measure of result relevancy.
- Recall is a measure of how many truly relevant results are returned.

7.1-Precision-Recall Curve

The precision-recall curve shows the trade-off between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both shows that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly.

Precision (P) is defined as the number of true positives (Tp) over the number of true positives (Tp) plus the number of false positives (Fp).

$$P = \frac{\mathrm{Tp}}{Tp + Fp}$$

Recall (R) is defined as the number of true positives (Tp) over the number of true positives (Tp) plus the number of false negatives (Fn).

$$R = \frac{\mathrm{Tp}}{Tp + Fn}$$

The true negative rate (TNR) is the rate that indicates the recall of the negative items.

$$TNR = \frac{\text{Tn}}{Tn + Fp}$$

Accuracy (A) is defined as

$$A = \frac{\text{Tp} + \text{Tn}}{Tp + Fp + Tn + Fn}$$

7.2-Sensitivity and Specificity

Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as a classification function. Sensitivity (also called the true positive rate, the recall, or probability of detection) measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition). Specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

7.3-F1-Measure

The F1-score is a measure that combines precision and recall in the harmonic mean of precision and recall.

$$F1 = 2 * \frac{\text{P} * \text{R}}{P + R}$$

7.4-Receiver Operating Characteristic (ROC) Analysis Curves

A receiver operating characteristic (ROC) analysis curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true positive rate is also known as sensitivity, recall, or probability of detection.

You will find the ROC analysis curves useful for evaluating whether your classification or feature engineering is good enough to determine the value of the insights you are finding. This helps with repeatable results against a real-world data set. So, if you suggest that your customers should take a specific action as a result of your findings, ROC analysis curves will support your advice and insights but also relay the quality of the insights at given parameters.