Data Warehouse Project

YAO   ZELAING

Part1: Drivers of the project

Purpose of the data warehouse :

The purpose is to help a property insurance company a toll motorways management company to design Data Warehouse with the aim of **managing its customers' insurance policies (contracts) as well as the accidents declared by these customers.**

Main actors :

Customers and stakeholders :

1. The sponsors of the Insurance Company

2. The Customers who buy insurance in this company

3. The Design team are external consultants design the data warehouse

4 The information system administrators who manage database and maintenance the data warehouse.

5 All the stuff in the company

Users of the product :

There are three main users of this data warehouse :

• The insurance agents. They can have all the information about their customers' insurance policies also the accidents declared by their customers

• The decision-maker or the manager of the company, who intend to request, manage or analyze surfing the data warehouse by the simple analysis.

• The RD department of the company, which is committed to make advanced display, segmentation, correlation, prediction and simulation by statistical and complex analysis and support decisions

Analyze the need :

We have to analyze needs first before get started

The general need of this case is to help the insurance company manage the policy of the customer well , also for the accident declared by his customer, we can analyze the need from different sides of the Users:

Insurance agents: they create, update or delete an insurance policy, so the policy's attribute must be well designed for them , easy to operate.

The decision-maker or manager: they will look at the general graph to make the essential decision for the company, the need the accurate and presize data of the policy and transaction, (the number total , amount total) …

The RD team: they do the research or improvement depends on the data involved of policies and accidents, so we have to design these two parts well fonctioned

Main actors :

Contracting authority :

1.<u>Operational Department</u> controls working order and supervise the coherence of the data, their evolution, and their quality

2.<u>The Technical Department</u> of the data warehouse will be members of the IT service. The IT service will have to plan these resources to be able to use them at the appropriate moment

3.<u>The meta-data Administration</u> guarantee the evolution of the data warehouse for the users.

4.<u>User Support develop</u> by the customer service. An impact study on the advantages and the inconveniences of each solution is to be launched.

Governance of the project

1. Initial stage :

From the defined objectives, we should focus on the feasibility, the estimated cost compared with the expected gainsThen we need to make the identification of the concerned actors. At last we should make the determination of the action plan.

The data warehouse is building in successive lots with the initiatives defined first.

Governance of the project

2. Design and development ：

*Functional part： Design of the data warehouse.Specifications and model making.

*Technical part： technical specifications and prototyping (for each of the phases), implementation of the technical environments, development of tools, technical and integration acceptance tests, design of the administration and exploitation tools.

*Cross-functional part： coherence of the repository coherence of the architecture and the flows, management of development environments.

Main identified risks

Description	possibility	severity	Mitigating actions
Lack of people skills or experience	Medium	High	Outsource or recruit consultant
Lack of strong sponsor or sponsor not at high enough level to get people motivated	Medium	High	Find widely cooperation or sponsor,Strengthen partnerships with existing partners.
Poor project management	Medium	High	Strengthen the management team，establisha reasonable plan
Lack of a data quality project as an integral part of the data warehouse implementation.	Medium	High	Strengthen survey data, strengthen the management data warehouse, employment a strong data team

Part2:Contraints on the project

Conventions of naming and definitions :

The naming convention is a key for a practical and effective Data Warehouse design and must be established early in the life-cycle of a project.

Each object has one name.

Object names are descriptive.

Abbreviations and acronyms are discouraged, The most effective approach is to maintain a list of approved abbreviations, and try not to add to them without a good reason.

Object names are unique.

Object names are not too long.
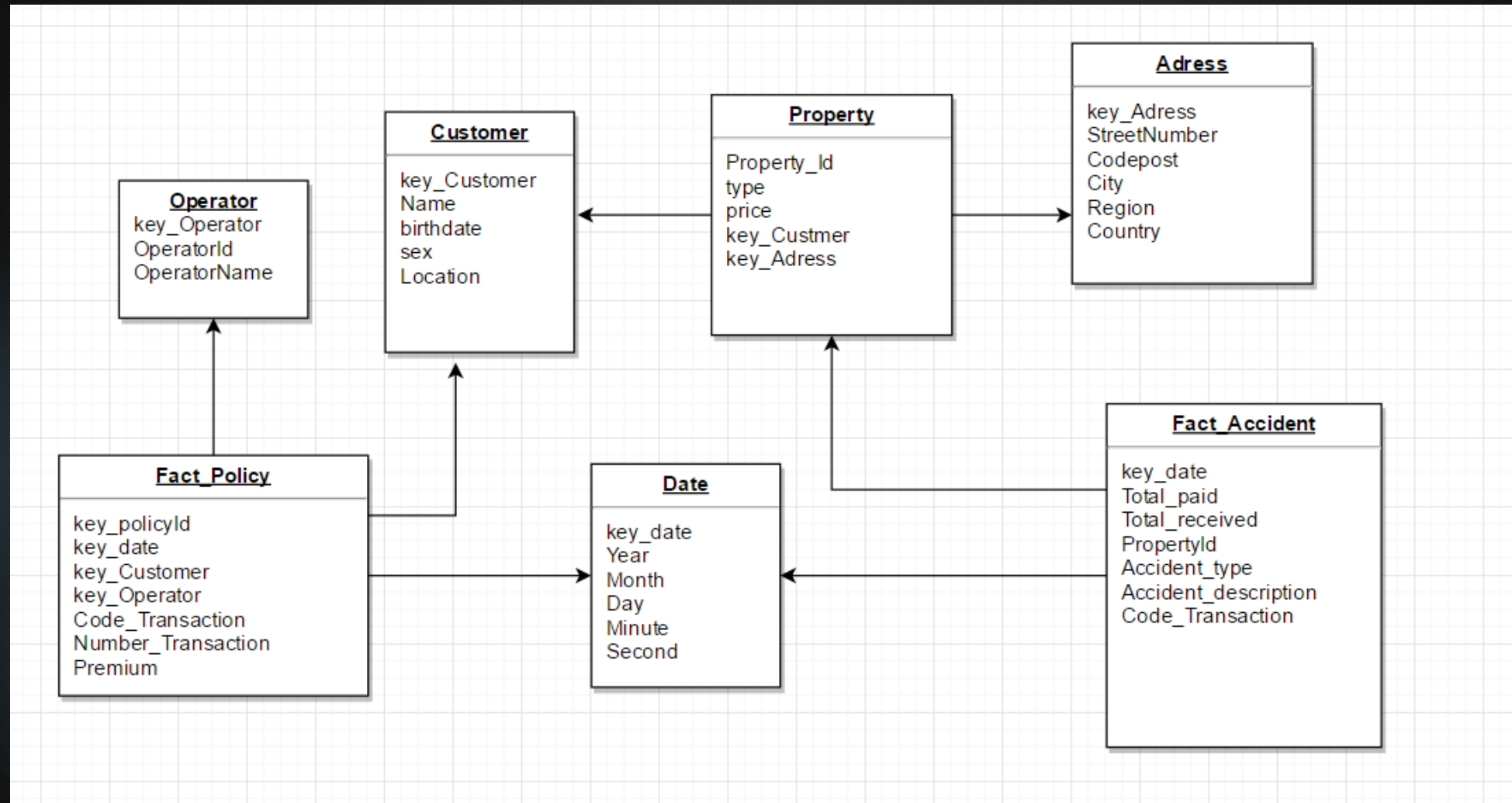
Facts and useful hypotheses :

Facts:

For the policy management:
The date of transaction , the effective date, the number and description of the policy, the transaction code

For the accident management:
The date of transaction ,effective date, the amount of payment, the accident type, location, the transaction code

Data model proposed

Part 3: Functional requirements

Fact Table :Fact_Policy

- Fact Table :Fact_Policy
- Number of customers: 30 millions
- Number of insured properties per customer(average)=3
- **Time : 3 years * 12 months = 36 month**
- **Number of transactions during a month for one properties(average)=2**
- **Number of fields of key type : 3**
- **Size of the digital attributes= 4 bytes**
- **Size of the table of the fact: 30,000,000*3*36*2*12B=77760000000B=0.78TB**

Part 3: Functional requirements

Fact Table :Fact_Accident

- Fact Table :Fact_Accident
- Number of customers: 30 millions
- Number of insured properties per customer(average)=3
- **Time : 3 years * 12 months = 36 month**
- **Number of transactions during a month for one properties(average)=2**
- **Number of fields of key type : 2**
- **Size of the digital attributes= 4 bytes**
- **Size of the table of the fact: 30,000,000*3*36*2*8B=51840000000B=0.52TB**

Functional requirements and requirements on the data

Doubtful dimension: Dates, sometimes date for a record can be doubtful as it can be record in different format: mm/yy, mm/yyyy or yyyy/mm. So conversion of the date to the same syntax is needed. We use the Europe naming convention: mm/yyyy.

Degenerate dimension: Since our dimensions are all explicitly defined, there is no degenerate dimensions.

Functional requirements and requirements on the data

Engine of the data warehouse:

ROLAP stands for *Relational OLAP*, an implementation based on relational DBMSs.

MOLAP stands for *Multidimensional OLAP*, an implementation based on multidimensional DBMSs.

HOLAP stands for *Hybrid OLAP*, an implementation using both relational and multidimensional techniques.

We choose to use OLAP as engine, OLAP can satisfy our needs for data and other technical demands

Applicative architecture proposed

There are two types of applicative architecture
1. EIS

EIS has a more stable dashboard， very friendly interface.But the
limit is the end user can display the information initially planned by the designer of the
dashboard,it's a no "custom-made product"

2. DSS

DSS has a less friendly interface, but it has fewer limits than EIS, like possibility of building its
own models and requests

According to the differences and our purpose of our dwh,the
dashboard is stable and the user don't need to do extensional
requests, after considered all sides, we choose the EIS

Technical architecture proposed
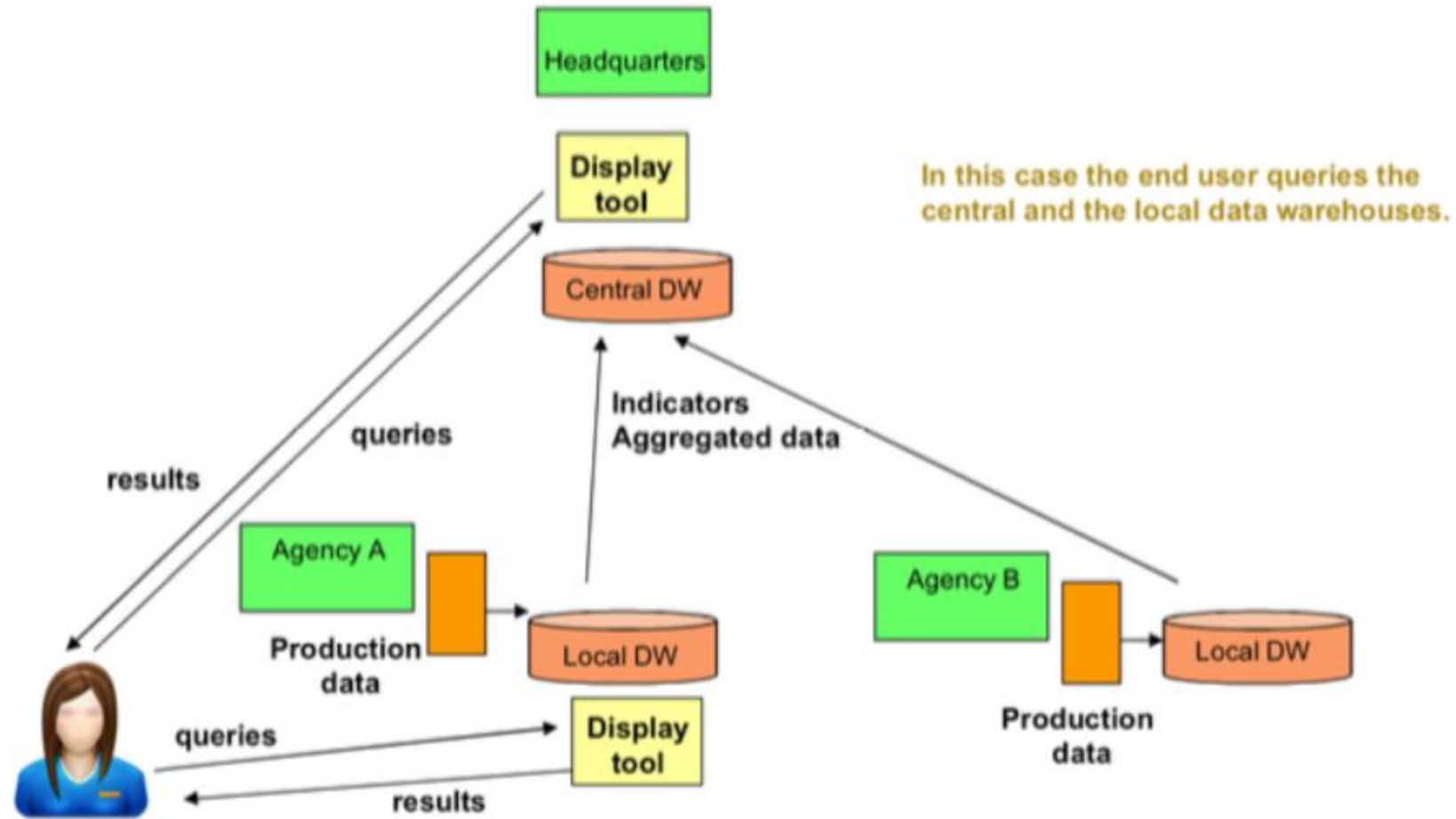
Main tools of solutions

The main tools of solutions are showed below:

ETL Tools

OLAP Tools

Presentation layer

Main tools of solutions

Our choise for ETL tools: **Talend**

Talend's flagship product, Talend Open Studio, is the most open, innovative and powerful data integration solution on the market today. Talend Open Studio provides advanced capabilities that dramatically improve the productivity of data integration job design; and proven scalability to ensure optimal execution.

Main tools of solutions

Our choise for OLAP tools:IBM Cognos

IBM Cognos is the predictive analytics software to search
through and perform analytics on big data from a variety of
sources, including data warehouses, Excel files, and Hadoop
distributions

Main tools of solutions

Our choise for presentation layer: IBM Congnos

We have to test our OLAP tools through the following standards:

- Data sources connections capabilities

- Security Features

- Customization

- Export capabilities

Performance

The performance of data warehouse is very important.If there are to many performance problems in the running of the Data Warehouse the viability of the project becomes marginal or questionable. It is important to remember the success of a Data Warehouse can only be measured once the data is loaded and users are able to make business level decisions by extracting data from the Data Warehouse.

Users now demand not only more complex and flexible analysis but also more timely information. Data must be available 24x7, and many business users demand that data supporting decision making be accessible within hours—in some cases, minutes or even seconds—of when an event occurs. Organizations also realize that the same data needs to be utilized by many different processes and thus many different workload profiles.

Security of the system

A. **Harden the Server** – No changes here. The software, operating system, network, and physical environment supporting SQL Server requires the same protection as before.

B. **Regulate Network Connectivity** – The guidance provided here is intended to make SQL Server harder to locate on the network by all but the intended applications. As the data warehouse is expected to provide flexible access to a variety of applications, hiding the SQL Server instance, disabling the SQL Browser, and employing application-specific endpoints doesn't exactly support this objective. Configuring non-standard TCP ports and named pipes some albeit little additional security to the data warehouse.

C. **Secure the Authentication Process** – The exclusive use of Windows authentication is strongly encouraged so that you may more accurately monitor end-user activity.

Merci pour votre attention