## 1

1. **Data warehouse** – BI system which stores information’s about events –
   * Bill Inmon - defines data warehouse as a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management’s decision making process
   * Ralph Kimball - A copy of transaction data specifically structured for query and analysis
2. **ETL process** – Extract, Transform, Load
3. **DSS** – decision support system :
4. **BI** – business intelligence – powerful analytical tool
5. Goals of system supporting decisions for the company
   * Easy access to the information
   * Process vast amount of data
   * Readable and easy to interpret
   * Ready for changes
6. **Analytical tools:**
   * **Checking for existence**
   * **Comparison**
   * **Trend analysis**
   * **Ranking**
   * **Statistical analysis**
7. **OLAP** – Online analytical processing
   * Trend analysis
   * Cross- sectional analysis
   * Strategic analysis
8. **KPI** – Key performance indicators
9. BI Dashboard
   * Single screen
   * Status of business analytics metrics
   * KPI
   * Important data for an organization
10. Business process – is a collection of related, structured activities or tasks that produce a specific service or product
    * Usually expressed in verbs
    * Often supported by systems
    * Generate numeric metrics
11. **OLTP** – Online transactional process
12. **OLAP features**
    * **Volume** – data of what size?
    * **History** – are historical data processed?
    * **Number of users** – how many users process data?
    * **Query complexity** – how complex queries are?
    * N**ormalization** – are data normalized?
    * **Addition** – are data added?
    * **Removal** – Are data removed?
    * **Updates** – Are data updated?
    * **Transactions and ACID** – are data processed with transaction?

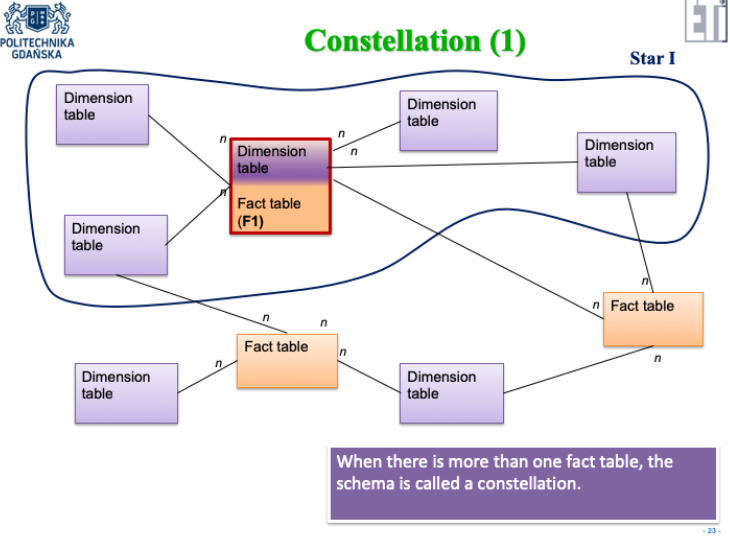
|  |  |  |
| --- | --- | --- |
|  | OLTP | OLAP |
| Volume | From small to big | big |
| History | Sometimes | Always |
| Numbers of users | Big | Small |
| Querry complexity | Small or medium | Big |
| Normalization | Yes | No |
| Addition | Yes | Yes ( BULK ) |
| Removal | Yes | No |
| Updates | Yes | Rarely |
| Transactions | Yes | No |

## 2

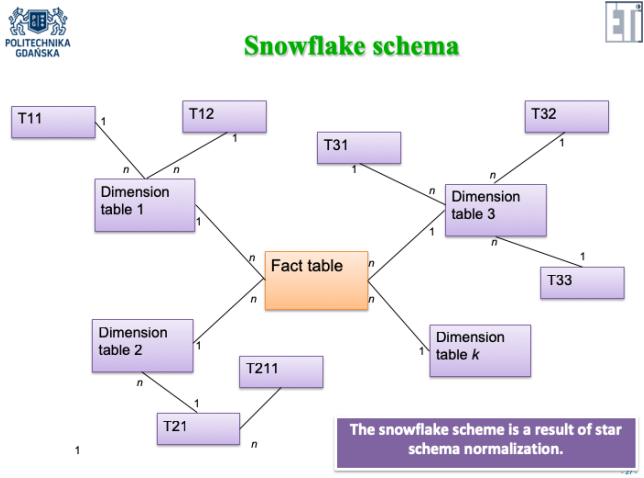
1. Features of a data warehouse
   * **Subject-oriented** – A data warehouse can be used to analyse a particular subject area, for example “sales”
   * **Time-variant** - historical data is kept in a data warehouse
   * **Non-volatile** – Once data is in the data warehouse, it will not change
2. Structures
   * **RDBMS**
   * **Multidimensional model**
3. **Multidimensional model**
   * **Fact** – single event associated with process and contains the measurement data associated with the event. Each fact is uniquely identified by the dimension members.
   * **Dimension** – is a collection of reference information about measurable event (fact). One dimension contains descriptive attributes i.e. **dimension attributes. Dimension attributes**  are text labels describing facts.
4. Consequences of this rule “. Each fact is uniquely identified by the dimension members”
   * **The number of facts in the presented cube is not greater then** [number of the store dimension members] \* [number of the product dimension members] \* [number of the time dimension members] = 5 \* 6 \* 4 = 120
   * **Ther is exactly one fact with the same all dimensions**
   * **Knowing all the dimension members, we can unambiguously identify the fact that these dimension members describe**
5. **Aggregation** is aggregated value of the measure for given slice of the cube
6. **Aggregation functions**
   * SUM
   * Average
   * MAX
   * MIN
   * ….
7. **DW Schemas**
   * **Star**
   * **Snowflake**
   * **Constellation**
8. **Hierarchical dimensions** – dimension which attributes form a hierarchy ex. E.g. time, date, location
   * Product Department -> Product Category -> Product Name
   * Product Category -> Product Department -> Product Name
   * Each of them is correct and its usefulness depends only on business needs.
9. **Data mart** – is a repository of data that is designed to serve a particular community of knowledge workers
10. **Inmon Approach:**
    * Enterprise-wide centralized data warehouse: All company data is first integrated into one central, normalized data warehouse.
    * Subject-area focused: It’s designed around major subject areas (like customer, product, vendor).
    * Department-specific data marts are then created from the central warehouse, each department only accessing what is relevant to them.
    * Emphasizes "single version of the truth" — consistency and integration of all data across the organization.
11. **Kimball Approach:**
    * Business process-focused: Starts by identifying key business processes and builds dimensional (star schema) data marts for each.
    * These data marts can be built independently and integrated through conformed dimensions (e.g., shared Customer or Product tables).
    * No separate, centralized data warehouse — the data marts together form the data warehouse.
    * Prioritizes fast delivery and ease of use for reporting and analysis

## 3

1. **Types of events (facts)**
   * **Transaction event** – one time
   * **Periodic event** – Every specific time interval
   * **Accumulating event** – event occurring In a longer period of time
2. **Granularity** – The granularity is nothing more than an unambiguous definition of the dimensions of given fact. **Atomic granularity** refers to the lowest level at which data are generated by given business process. It is not allowed to mix different granularities within the same type
3. **Types of measures**
   * **Additive measures** – summed across any of the dimensions associated with the fact
   * **Semi-additive measures** – summed across some dimensions, but not all
   * **Non-additive measures** - cannot be summed across any dimension
4. **Fact table properties**
   * Row of fact table contains measures an FK
   * Null not allowed
   * A lot of facts rows
5. **Dimension properties**
   * Lack of FK
   * Allowed null
   * Not to many of dimensions rows
   * Descriptive values
   * Contains surrogate key
   * Can have business key
   * No coding answers
6. **Slowly changing dimensions** – they are changing over time
   * **SCD 1** – values are overwritten with new data values. Historical value is lost and historical analysis can be incorrect
   * **SCD 2** – a new record encompassing the change is added, the old record is marked as inactive. For the new record a new surrogate key is generated and the BK is unchanged. Additionally the insertion date and deactivation date are settled
   * **SCD 3** – a second column to store most recent past value of the column is added
7. **Junk dimension** – we put here attributes which are not functionally depended
   * We only insert these rows that actually exist
   * We can create more than one junk
8. **Degenerated dimension**
   * Is a dimension key in the fact table that does not have its own dimension table, because the only dimension attribute is in the dimension key
   * short sting of characters or numbers, and this value is repeated only for a few facts. In such situation the value is stored directly in the fact table. In the presence of degenerate dimensions, the fact is identified by the composition of foreign keys and values of degenerate dimensions
9. **Constellation**



1. **Snowflake**



1. Role playing dimensions – multiple connection of fact with the dimension
2. Events types comparison

|  |  |  |  |
| --- | --- | --- | --- |
|  | Transaction | Periodic | Accumulating |
| Periodicity | Discrete point in time | Repeated every defined period of time | An indefinite period of time for given flow of events |
| Granularity | One row for one transaction event | One row for one periodic event | One row for the whole event flow |
| Date dimensions | Transaction date | The summary creation date | Many dates, one for each milestone of given flow |
| Measures | Transaction | Cumulative for a given time interval | Related to milestones |
| The density of the rows in the fact table | Dense or rare depending on the activity | Dense | Dense or rare depending on the event flow |
| Fact table update | Update only incorrect data | Update only incorrect data | Update when flow activity occurs |

## 4

1. **Calculated measure** – measure composed of other measures and constants

## 5

1. **Accessing models**
   * **RDBMS – SQL language**
   * **OLAP – MDX language**
2. **Mapping** - are associations between concepts from the multidimensional model and the relational model
3. **Aggregates** – precalculated aggregated measure values
4. **Processing a cube** - is creating a multidimensional model in OLAP server it is build basing on the relational schema
5. **Possible models**
   * **MOLAP** – all data is stored in multidimensional model, means that OLAP server stores the cube schema, mapping, all data and precalculated aggregations. It works without RDBMS
   * **ROLAP** – only cube metadate and mappings are stored in OLAP server. The MDX query issued to OLAP server, executes as follow
     + The MDX is formulated in terms of multidimensional model
     + Basing on the metadata, OLAP server transforms MDX query to SQL query and sends this SQL query to RDBMS
     + The result of the SQL query is returned to OLAP server and then transformed to multidimensional terms
     + The result is returned to the end user
   * **HOLAP** – Combination of both
6. **OLAP server is column oriented-database.** Why? Because in relational database we can access by index or by scanning and In analytical queries we usually must scan entire content of the table
   * Row oriented must read the entire table
   * Column oriented must only read the needed columns so it is faster

## 6

1. MDX instruction elements
   * Number
   * String
   * Dimension member
   * Tuple
   * Set – set of tuples
2. Commands
   * Select {x} on Columns {y} on Rows From Measure
   * Where
   * .Members .Children - same
   * COLUMNS, ROWS, PAGES, SECTIONS, CHAPTERS
   * WITH
   * Order
   * TopCount , BottomCount
   * TopSum , BottomSum
   * Filter
   * Union, Except , CrossJoin , NonEmptyCrossJoin
   * Sum, Avg, Median, Min, Max, Var, Stdev
   * Count

## 7

1. Helping the perception summary:
   * Highlight schemes;
   * Highlight outliers;
   * Pay attention to the contrasts (use it or avoid it);
   * You can add supporting measure;
   * You can add the grid....... but too much grid is too much;
   * Automatic and static lines can be helpful;
   * Sorting can be helpful;
   * Additional labels can give you a lot of additional information at a small cost;
   * As well as plotting data aggregates with their components;
   * Filtering extremes is very informative;
2. Data Visualization Do's and Don'ts - A General Conclusion
   * Time axis. When using time in charts, set it on the horizontal axis. Time should run from left to right. Do not skip values (time periods), even if there are no values.
   * Proportional values. The numbers in a chart (displayed as bar, area, bubble, or other physically measured element in the chart) should be directly proportional to the numerical quantities presented.
   * Data-Ink Ratio. Remove any excess information, lines, colours, and text from a chart that does not add value. More about data-Ink ratio
   * Sorting. For column and bar charts, to enable easier comparison, sort your data in ascending or descending order by the value, not alphabetically. This applies also to pie charts.
   * Legend. You don't need a legend if you have only one data category.
   * Labels. Use labels directly on the line, column, bar, pie, etc., whenever possible, to avoid indirect look-up.
   * Inflation adjustment. When using $ monetary values in a long-term series, make sure to adjust for inflation.
   * Colours. In any chart, don't use more than six colours.
   * Colours. For comparing the same value at different time periods, use the same colour in a different intensity (from light to dark).
   * Colours. For different categories, use different colours. The most widely used colours are black, white, red, green, blue, and yellow.
   * Colores. Keep the same colour palette or style for all charts in the series, and same axes and labels for similar charts to make your charts consistent and easy to compare.
   * Colores. Check how your charts would look when printed out in grayscale. If you cannot distinguish colour differences, you should change hue and saturation of colours.
   * Colores. Seven to 10 percent of men have colour deficiency. Keep that in mind when creating charts, ensuring they are readable for colour-blind people. Use Vis check to test your images. Or, try to use colour palettes that are friendly to colour-blind people.

Exam questions

1. Density of fact table
   * Marketing Campaign – it has big density because each of the campaign correspond to the one marketing campaign conduced on the property which is the most common event in our company
   * Campaign Methods -it has very big density because for each Marketing campaign there exist there is 3 to 6 more entries of campaign method and marketing campaign itself if very common event
2. Measures types
   * Additive – Total number of views can be added across all dimensions
   * Semi additive – if we stored number of employees we couldn’t add it across the date but we could add it across the team dimension
   * Non-additive – Views per dollar spend cannot be added because adding ratio don’t make sense
3. Fact description
   * Marketing Campaign – describes marketing campaign fact which has started at given date ended at given date, concern property which was bought at given date and sold at given date also it is connected with only one specific property. It describes accumulative event because it has significant milestones. Sale date, and Campaign end date
   * Campaign Methods – describes Campaign method of promotion used for given marketing campaign and it is connected with only one marketing method. It is transaction event as it occurs one if it was inserted there is no need to change it unless it was entered incorrectly.
4. Conformed dimension
   * If we would have a property sale fact the date dimension would be conformed dimension as it would be connected to 2 fact tables Property sale and Marketing campaign.
5. Actors using BI system
   * CEO – needs to make important decisions based on the KPI and other metrics like total sales, total cost and so on
   * CFO / Accounting – can check how much money was spend on the proportion of the given property
   * Marketing department – Can check success rate of the given campaigns and observe the trends with possible division of the method used
6. SCD 2 type
   * This could be dimension property as it can change its sale range attribute in time and it can be useful to know in what types of house those changes occur how often they occur, and what is their scale
7. Fact tables corresponding to the same event but with various grains.
   * Different grains gives us different level of detail
   * Example
     1. Fact 1 – grain – by each individual Marketing campaign event it gives as detailed analysis of each campaign and can be used in various scenarios like analysis across type and time at the same time
     2. Fact 2 – grain by Marketing Campaigns per region I limits our possibility of searching across date and type of apartment but increase the speed of query because we have less entries
8. In the process of data warehouse design, analyse the laboratory subject and define atomic fact, then define and design calculated measure which calculates the average number of … (whatever you want and has a business sense for your project). Explain the process of calculation of this measure
   * Marketing Campaign – describes marketing campaign fact which has started at given date ended at given date, concern property which was bought at given date and sold at given date also it is connected with only one specific property. Data is drawn from database as well Marketing excel
   * Average cost per Marketing Campaign SUM ( Total cost ) / COUNT(ID)
9. MDX queries
   * To have the most optimal time of response for query resolution I would decide to implement MOLAP server, because all data, mappings, schemas and precalculated measures are stored in multidimensional model which makes query resolution faster and does not rely on RDBMS. I would also design aggregations to increase query resolution as they can increase query resolution.
10. Surogate key – SCD 2 I do identyfikacji
11. BP można zmienić
12. MOLAP na kolumnach
13. Fact może zawierac descriptiv np. numer lotu – i count
14. Measure and agregation – Flight duarion : SUM()\
15. Liczba factow Count(1)
16. Dashboard – jednen ekran duzo wizualizacji