# (Big) Data Engineering In Depth

## From Beginner to Professional

Mostafa Alaa Mohamed
Senior Big Data Engineer
MoustafaAlaa Moustafa Alaa @Moustafa_alaa22
mustafa.alaa.mohamed@gmail.com

[1]Big Data & Analytics Department, Epam Systems

The Definitive Guide to Big Data Engineering Tasks

# Videos classification

| Watching Method / Audience | Computer | Mobile/Tablet | Just listening |
|---|---|---|---|
| **Developer** | ● | | |
| **DevOps** | ● | | |
| **Business** | ● | | |

Table: Video classification
The green circle ● means short video.
The blue circle ● means medium video.
The red circle ● means long video

## Sub-Section: Fact Table

# Fact Table Recap

What is the fact table?

- It is the foundation of the data warehouse.
- It consists of facts and measurements of a particular business aspect and processes ex: daily revenue for a product.
- It is the target of queries in most of DWH analysis and reports.
- It contains measurements/facts and foreign keys to *dimensions table*.
- It located at the center of the schema and surrounded by dimension tables.

# Fact Table Recap

> *"There is no point in hoisting fact tables up the flagpole unless they have been chosen to reflect urgent business priorities"*

Ralph Kimball, *kimballgroup.com*

# How to design a fact table?

# How to design a fact table?

- Choose the business process.
- Identify the grain.
- Identify the dimensions.
- Identify the fact.

# Fact Granularity

- The grain is the definition of what a single row in the fact table will represent or contains.
- The grain describes the physical event which needs to be measured.
- Grain controls the dimensions which are available in fact.
- Grain represents the level of information we need to represent. It is not always time; it could be the physical business measurement level.
- Design from the lowest possible grain.

# Sub-Section: Fact Table Types

# Fact Types

There are three types of fact tables:

- Transaction.
- Periodic.
- Accumulated Snapshot.

# Fact Types: Transaction Fact Table

- Fact grain set at a single transaction
- It has one row per transaction.
- For each transaction, we add a new single record.
- The transaction fact table is known to grow very fast as the number of transactions increases.

# Fact Types:Transaction Example

| customer_id | trns_date | trns_time | call_type | duration |
|---|---|---|---|---|
| 1234 | 2020-01-01 | 12:22:45.9 | Incoming | 29 |
| 1234 | 2020-01-01 | 12:22:45.9 | Incoming | 3134 |
| 1234 | 2020-01-02 | 15:22:45.0 | Outgoing | 890 |
| 1234 | 2020-01-02 | 15:22:45.0 | International | 119 |
| 1234 | 2020-01-03 | 23:22:45.0 | Incoming | 145 |
| 1234 | 2020-01-03 | 23:22:45.0 | Outgoing | 124 |
| 1234 | 2020-01-03 | 23:22:45.0 | Outgoing | 1200 |

Table: Transaction fact example of telecom calls data.

# Fact Types: Periodic Fact Table

- A periodic fact table contains one row for a *group* of transactions over a period.
- It must be from lower granularity to higher granularity hourly, daily, monthly, and quertrly, then yearly.

# Fact Types: Periodic Fact Table Example

| cust_id | month_id | incoming | outgoing | international |
|---|---|---|---|---|
| 1234 | 20200131 | 3308 | 2124 | 119 |

Table: Periodic fact example of telecom calls data.

# Fact Types: Accumulated Snapshot Fact Table

- An accumulating fact table stores one row for the entire process.
- It does not accumulate time it accumulates business process.
- A row in an accumulating snapshot fact table summarizes the measurement events occurring at predictable steps between the beginning and the end of a process
- Accumulating Fact tables are used to show the activity of progress through a well-defined process and are most often used to research the time between milestones.
- These fact tables are updated as the business process unfolds, and each milestone is completed.

# Fact Types: Accumulated Snapshot Fact Table Example

- Accumulated Snapshot use cases are engaged when we need to report the entire process life-cycle.Fact Types: Accumulated Snapshot Use Cases.
- It also uses to measure the process performance life-cycle.
  - Order life-cycle.
  - Insurance processing.
  - Hiring process.

# Fact Types: Accumulated Snapshot Fact Table Example

An insurance company

- It has a fact table named: *fact_claim_processing*.
- This fact represents the claim life-cycle inside the company.
- It contains detail related to claim.
- This fact update after each stage finished.

# Fact Types: Accumulated Snapshot Fact Table Example

Example of Accumlated Snapshot: An insurance company

- It fact table named: fact_claim_processing.
- This fact represents the claim life-cycle inside the company.
- It contains detail related to claim.
- This fact update after each stage finished.
- The requirement it to report the number of days (lag) between stages (milestone) and the claim data (starting).


Figure: Claim Life-Cycle

# Fact Types:Accumulated Snapshot Example

- One solution to implement the requirement is to use SCD.
- In this case, we will have stages and dates, and we will calculate the difference between stages and dates using complex sub-query.
- Another solution is to implement an accumulated snapshot fact.

FACT_CLAIM_PROCESSING

| |
|---|
| CLAIM_KEY |
| CUSTOMER_KEY |
| POLICY_KEY |
| CLAIM_DATE |
| INVESTIGATION_DATE |
| REVIEW_DATE |
| DECISION_DATE |
| PAYMENT_DATE |

# Fact Types:Accumulated Snapshot Example

FACT_CLAIM_PROCESSING

| CLAIM_KEY |
| --- |
| CUSTOMER_KEY |
| POLICY_KEY |
| CLAIM_DATE |
| INVESTIGATION_DATE |
| REVIEW_DATE |
| DECISION_DATE |
| PAYMENT_DATE |

FACT_CLAIM_PROCESSING_ACCUM

| CLAIM_KEY |
| --- |
| CUSTOMER_KEY |
| POLICY_KEY |
| CLAIM_DATE |
| INVESTIGATION_DATE |
| DAY_TO_INVESTIGATE |
| REVIEW_DATE |
| DAY_TO_REVIEW |
| DECISION_DATE |
| DAY_TO_DECISION |
| PAYMENT_DATE |
| DAY_TO_PAYMENT |

# Fact Types: Accumulated Snapshot Table Example

| column_name | column_value |
|---|---|
| claim_key | 123 |
| customer_key | 5235326 |
| policy_key | 23632623 |
| claim_date | 2020-01-01 |
| investigation_date | 2020-01-03 |
| day_to_investigate | 2 |
| review_date | 2020-01-07 |
| day_to_review | 6 |
| decision_date | 2020-01-08 |
| day_to_decision | 7 |
| payment_date | 2020-01-11 |
| day_to_payment | 10 |
| process_completed_flag | 10 |

Table: Accumulated Snapshot Fact Example on Claim Process Data.

# Fact Table Types: Comparison

| Feature | Transaction | Periodic | Accumulating |
|---|---|---|---|
| Grain | 1 row/transaction | 1 row/time-period | 1 row/entire event stages |
| Date Dimension | Lowest granularity | End-of-period granularity | Multiple date |
| Facts | Transaction activities | Periodic activities | Defined lifetime activities |
| Size | Largest | Medium | Smallest |
| Update | No | No | Yes, after stage finished |

Table: Fact tables types comparison.

# Fact types

# Fact types

Each fact table includes facts and it has different types:

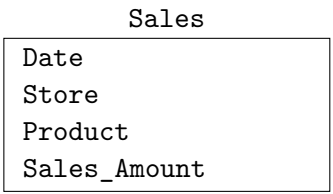
- Additive facts.
- Semi-additive facts.
- Non-additive facts.
- Derived facts.
- Textual facts.
- Factless fact.

# Additive facts

- It is the most flexible and useful facts.
- Its measures can be summed across any of the dimensions associated with the fact table.

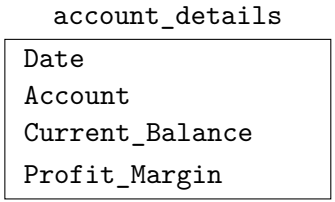
# Additive facts

- It is the most flexible and useful facts.
- It can be summed across any of the dimensions associated with the fact table.



# Semi-additive facts

- It can be added across some dimensions but not all also known as (partially-additive).




- what's the total current balance for all accounts in the bank?
- What's the current balances for a given account for each day of the month does not give us any useful information?

# Non-additive facts

- It can't be added for any of the dimensions.
- Non-additive facts are usually the result of ratios (percentage) or other mathematical calculations.
- **Profit_Margin** is an example non-additive.

account_details

| Date |
| --- |
| Account |
| Current_Balance |
| Profit_Margin |

# Derived facts

- Derived facts are created by performing a mathematical calculation on a number of other facts, and are sometimes referred to as calculated facts. Derived facts may or may not be stored inside the fact table.
- Total_sales = Qty_Sold * ( Unit_price - Discount)

# Textual facts

- A textual fact consists of one or more characters such as flags and indicators.
- It should be avoided in the fact table.

# Factless fact

- A fact table with only foreign keys and no facts is called a factless fact table.

# References


- https://www.nuwavesolutions.com/accumulating-snapshot-fact-tables/
- https://www.kimballgroup.com/2008/11/fact-tables/
- https://www.1keydata.com/datawarehousing/fact-table-types.html