Source Details

Database Details

Here are the details to connect to retail db running under MySQL Database Server.

* Hostname
* Port Number
* Database Name: retail\_db
* User Name
* Password

Here are the details related to tables from which we need to pull data.

* Transactional
* Master Data Tables

Data Pipeline

Read Data

* Frequency: DAILY
* Schedule: 2 AM Every Day
* Dimensions
  + Extract data from products, categories and departments
* Facts
  + Extract data from orders and order\_items based on order\_date

Process Data

Mapping Document - Master Data Tables to Dimensions

* Type 1 dimension
* Apply transformation rules based on the mapping.
* Approach 1 - Get join results by running a join query between 3 tables.
* Approach 2 - Get data from all the master data tables, apply processing (join) as part of the pipeline.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target Table** | **Target Column** | **Source Table** | **Source Column** | **Description** |
| dim\_customers |  | customers |  | Get all columns from source |
| dim\_customers | batch\_id |  |  |  |
| dim\_customers | batch\_date |  |  | Date on which data is loaded |
| dim\_products | product\_id | products | product\_id |  |
| dim\_products | product\_name | products | product\_name |  |
| dim\_products | product\_price | products | product\_price |  |
| dim\_products | category\_id | categories | category\_id |  |
| dim\_products | category\_name | categories | category\_name |  |
| dim\_products | department\_id | departments | department\_id |  |
| dim\_products | department\_name | departments | department\_name |  |
| dim\_products | batch\_id |  |  |  |
| dim\_products | batch\_name |  |  |  |

Mapping Document - Transaction Tables to Facts

* Get all the required fields from orders and order\_items by passing a date as argument.
* Filtering and Aggregations should be done using Pandas as part of the pipeline
* Data should be filtered and pre-aggregated as per the transformation logic and then should be stored into the fact.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target Table** | **Target Column** | **Source Table** | **Source Column** | **Description** |
| fact\_product\_revenue\_dly | date\_id (int) | orders | order\_date (date) | Just store the date part of it, ignoring timestamp.  (e. g: 20140101) |
| fact\_product\_revenue\_dly | product\_id | order\_items | order\_item\_product\_id | Get the product ids as is from source to target as per this mapping. |
| fact\_product\_revenue\_dly | product\_revenue | order\_items | order\_item\_subtotal | * Filtering - COMPLETE or CLOSED orders (orders.order\_status) * Aggregation - sum(order\_item\_subtotal) group by orders.order\_date and order\_items.order\_item\_product\_id |
| fact\_product\_revenue\_dly | outstanding\_revenue | order\_items | order\_item\_subtotal | * Filtering - PENDING\_PAYMENT or PROCESSING or PENDING orders (orders.order\_status) * Aggregation - sum(order\_item\_subtotal) group by orders.order\_date and order\_items.order\_item\_product\_id |
| fact\_revenue\_dly | date\_id (int) | orders | order\_date (date) | Just store the date part of it, ignoring timestamp.  (e. g: 20140101) |
| fact\_revenue\_dly | revenue | order\_items | order\_item\_subtotal | * Filtering - COMPLETE or CLOSED orders (orders.order\_status) * Aggregation - sum(order\_item\_subtotal) group by orders.order\_date |
| fact\_revenue\_dly | total\_order\_cnt | orders |  | * Aggregation - count(1) group by orders.order\_date |
| fact\_revenue\_dly | revenue\_order\_cnt | orders |  | * Filtering - COMPLETE or CLOSED orders (orders.order\_status) * Aggregation - count(1) group by orders.order\_date |
| fact\_revenue\_dly | canceled\_order\_cnt | orders |  | * Filtering - CANCELED orders (orders.order\_status) * Aggregation - count(1) group by orders.order\_date |
| fact\_revenue\_dly | outstanding\_order\_cnt | orders |  | * Filtering - PENDING\_PAYMENT or PROCESSING or PENDING orders (orders.order\_status) * Aggregation - count(1) group by orders.order\_date |

Load Data

* Frequency: DAILY
* SLA: By 8 AM every day
* Dimensions
  + Delete the existing dimension data
  + Load the snapshot from the source
* Facts
  + Facts should be loaded after dimensions are loaded

Target Details

Database Details

Here are the details related to the data warehouse into which data should be loaded.

* Hostname
* Port Number
* Database Name: retail\_dw
* User Name
* Password

Here are the details about Facts and Dimensions.

* Facts
  + fact\_product\_revenue\_dly
* Dimensions
  + dim\_product
  + dim\_customer