ACCT / MIS 6309

Instructor: Kevin R. Crook

Fall 2016

Data Warehouse Design Project

Last Name: Dhavale

First Name: Gaurav

Middle Name: Dilip

Comet Creed - *"As a Comet, I pledge honesty, integrity, and service in all that I do."*

Electronic Signature: Gaurav

Contents

Data Warehousing of GDD Gas Supplier

Description:

GDD Company is $9 billion company founded in the year 2001 by one person and currently, holds more than 40 thousand employees. The company deals with end to end oil and gas business solutions which mainly focuses on following business units. Production of oil, gas and other products
Distribution of products to customers such as gas stations, storage tanks, refineries, etc. using different transportation channels like pipelines, vehicle, and ships Purchase, and sales of products, Forecasting of the various business products in order to define their production, Inventory management, Planning, and HR.
Every unit of business itself is very vast. Let's discuss every unit in brief.
This business is divided into upstream, midstream and downstream sector. Upstream deals with extraction of raw material such as crude oil, natural gas, etc. and processing it into a fine product. Production belongs to the upstream sector. The distribution and inventory management belong to midstream sector. Distribution business unit is also a part of downstream sector along with sales and purchase of products where it reaches to every customer to supply the products. Forecasting defines the production by considering sales histories for different products based on regions, customers, their categories, and seasons. Planning and HR units deal with resource planning, pipeline planning, and berth planning.
All these business units are interconnected to each other for their operations. Degradation of the performance of any unit will hamper the entire business. So it's very necessary to streamline the connectivity between these business units, resolved the business issues.
Different of business problems:

- Forecasting of production of gas based on sales
- Define and improve different distribution channels
- Maintain customer's product interest
- Inventory management
- Profit and loss tracking of purchase and sales of products
- Financial planning

Let's discuss these problems in brief.

- Forecasting of production of products based on sales: In order to forecast sales of any product, It's essential to consider different factors from the market. We need to consider trends in markets, seasonal product demand, regular product demand, production rate, generated production, the market value of products over the past years. Considering all these factors, forecast generation becomes a very tedious process. There are different ways of generating forecast. A forecast can be system generated using different forecasting methods such as trend, seasonality, the trend with seasonality. This forecast can be termed as statistical forecast. The forecast can also be defined by customers as well as the business expert. Forecast defined by the customer can be considered for production of the product.
  Now the issue arises while obtaining data required for forecasting. A product travel from extraction process till it reaches to the final consumer. Data for all these operations is scattered in the different business process. It's very necessary to capture this raw data and convert it to

required information. Using data warehousing, information can be captured in a structural manner. This will also help to analyze the data and ultimately to do forecasting of products.

- Maintain customer's interest: Tracking of demand and supply of products to customer is essential in defining customers need in future. We can see the product suggestions whenever we try to purchase any product online. Data warehousing and business intelligence tools play a very important role in defining customers interest. By analyzing customers purchase trend, we can predict his demand and can plan supply or production accordingly.

- Product pricing: product pricing plays a very important role in competition in the market. The amount of profit in product pricing can be defined by analyzing pricing of a different vendor. There are certain products like Jet fuel which are required by airline customers on daily basis fetches more or regular pricing due to constant demand and supply. They also need dedicated transport mechanisms which may affect products costing. There are many government norms that also affects products pricing. Such products can fetch Data warehousing can help a lot in this analysis.

- Define and improve different distribution channels: Product can be delivered to different locations by different means of transportation such as train, truck, vessel, and pipeline. Various channels in product distribution can help in defining most optimal and cheapest mode for delivering products. The pipeline requires initial heavy infrastructure setup but in later stages, it can be very effective and fast and continuous mode of transport. Data warehousing can help to analyze all these mechanisms.

Let's Analyze different data warehousing methods in order to define which methods suits well to my company.

Data Warehousing Methodologies

Inmon's Methodology:

Inmon's uses normalized data model in 3NF and consume less space as compared to Kimball's. It uses enterprise design approach which means the top down design approach where we design normalized database first and then based on different business units, we design dimensional data marts.

Kimball's Methodologies:

Kimball has de-normalized Dimensional Data Model. It's a dimensional design approach which is also known as a bottom-up design approach. In this approach, we don't perform data normalization rather we design data marts which resolve problems of different business process and then all these data marts are combined to build a data warehousing.

Stand-alone Data Marts:

stand-alone data marts are not at the enterprise level. It skips a data warehousing design. It collects data from different transactional systems based on business needs and supplies it to business analytical tools. This methodology requires less space, cost and time as well. However, it's not advisable to use this approach as it skips data warehouse and for a longer run, it can cause a lot of problems in data analysis and business intelligence.

For oil and gas industry, having a tremendous amount of data can lead to more and more space for warehousing. It also requires enterprise data handling structure. Kimball requires a lot of space for warehousing. A famous company Walmart follows Kimball's methodologies. Their warehouse size is in

petabytes and it increases hugely every year. A stand-alone data mart is not suitable for my business as I don't have timings constraints and financial issues. Moreover, I have structured normalized database infrastructure set up already with excellent employees who will work on warehouse design. This will further reduce my implementation time for Inmon's methodology. Hence it's the best deal to go ahead with Inmon's methodology.

Data warehouse Method that can be applied: Inmon

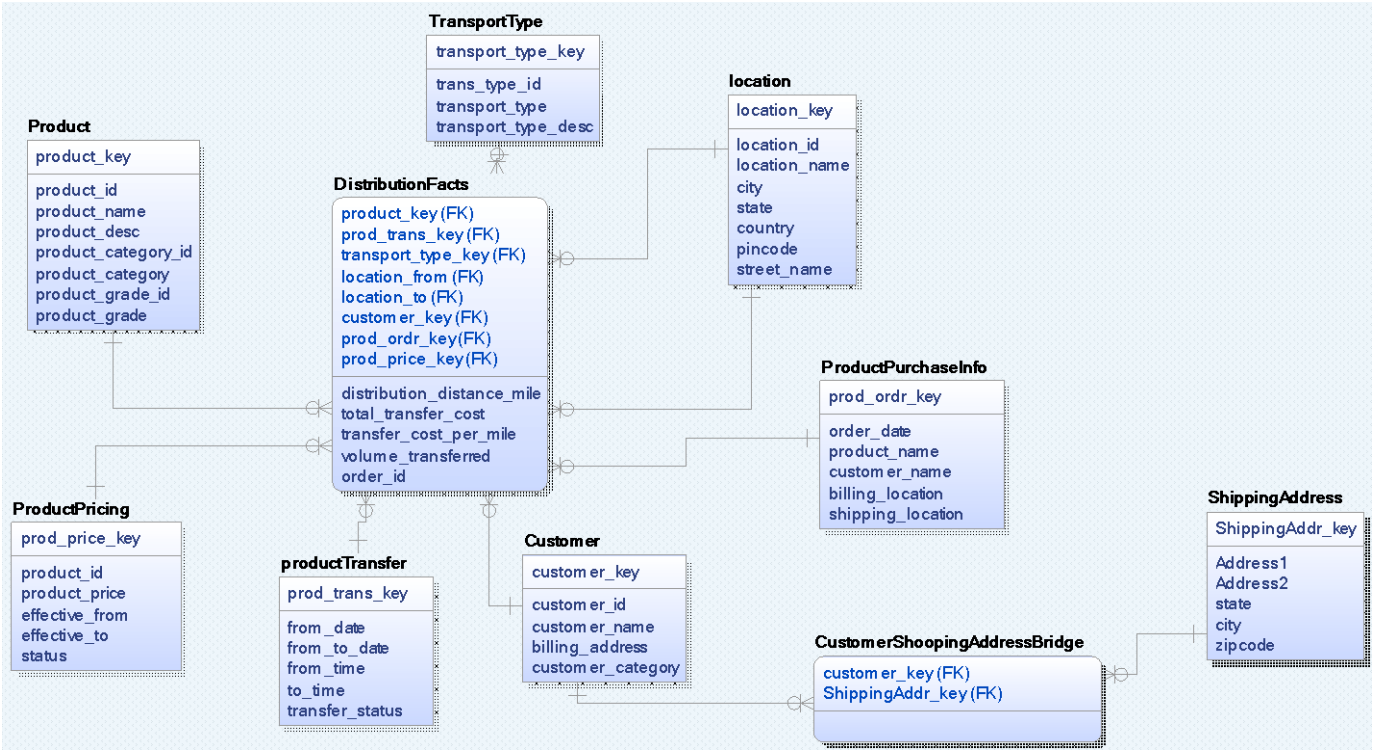
Out of the above-specified business issues, let's resolve five specific business problems in "Define and improve different distribution channels" section.

Specific business Questions:

1. Amount of barrel of oil transfers from one location to other
2. Financial costing involved in product transfer
3. Costing difference for different transport mechanism and thus define accurate transport mode selection based on available resources.
4. Pipeline infrastructure across different location
5. Average and lowest oil distribution in a month by pipeline

This can be achieved by analyzing the requirements, defects for every unit which can be done significantly using business intelligence techniques. A data warehouse solution can be designed in order to tackle various business problems.

Star Schema (with bridge)



The above star schema will help us to resolve most of the dedicated 5 questions. The design information goes as follows:

Fact Table: DistributionFacts: It holds information about barrels of customer's product transferred from one location to another location by a particular transport type. It also gives the transfer cost involved in this operation.

Grain: Amount of volume transferred

Additive facts: distribution_distance_mile, transfer_cost_per_mile, total_transfer_cost, volume_transferred

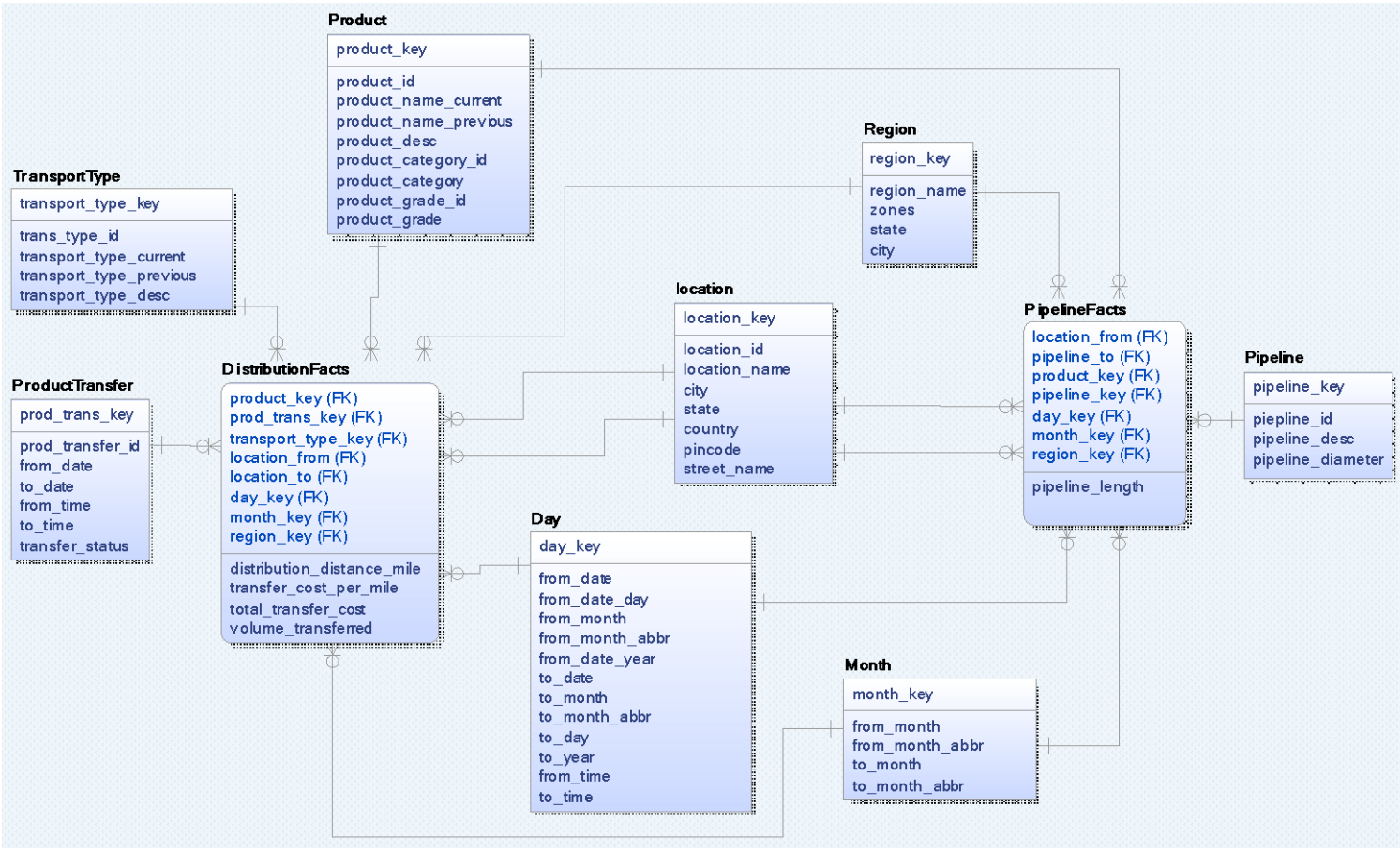
- **Affinity Dimensions:**
  - Product: Captures products information: => type 2
  - ProductPricing: Define pricing of product for a particular period: => Timestamp
  - TransportType: transport mode which can be vehicle, pipeline => Type 1
  - ProductTransfer: Define product transferred timing period along with progress status. =>Timestamp
  - Location: product transferred to and from which location => Type 2
  - Customer: customer to whom the product is delivered. =>Type 1/2 hybrid
- **Junk Dimension:** ProductPurchaseInfo: It captures overall products purchase data. It's a set of entities which are not logically kept but rather to capture an overview of a product such as purchased on which date, by which customer and billing, shipping details.
- **Degenerate Dimension:** The attribute order_id is added in fact table DistributionFacts instead of storing it in the junk dimension ProductPurchaseInfo.
- **Bridge:** CustomerShippingAddressBridge. The customer can have multiple shipping addresses. A customer can have multiple shipping addresses. This business requirement can be handled using the bridge.

- Browsable dimension: Product

The above star schema can explain the following set of questions

1. Amount of barrel of oil transfers from one location to other
   Distribution facts captured volume details, and location details as well. We can join the fact table with product, location and product transfer dimensions to retrieve the amount of barrel of oil transferred from one location to other.
2. Financial costing involved in product transfer
   Transfer cost defined financial costing involved in product transfer with the help of product pricing, volume transferred distribution distance in miles.
3. Costing difference for different transport mechanism and thus define accurate transport mode selection based on available resources.
   For every transport mode, we can obtain transfer cost and use this cost we can also define costing difference. This information can be used to analyze transport mechanism for other different locations in order to deduce accurate transport mode.

Start Schema with conformed dimensions



The design for above star schema goes as follows:

- Facts Tables
    - DistributionsFacts: It holds information about barrels of customer's product transferred from one location to another location by a particular transport type. It also gives the transfer cost involved in this operation.
    - PipelineFacts: pipeline facts cover pipeline related information such as its starting and end node, product delivered via pipeline to different locations.
- Additive facts: distribution_distance_mile, transfer_cost_per_mile, total_transfer_cost, volume_transferred
- Non-Additive facts: pipeline diameter
- Affinity Dimensions: Product, TransportType, ProductTransfer, Location, Pipeline, Month, Day
- TransportType is a Type 3 dimension.
- Conformed Dimensions:
    - Shared dimension: Product
    - Overlapping dimension: Region, Location
    - Conformed rollups: day, month
- **Rich dimension**: Day
- Rich attributes from Day dimensions are derived mostly from "from_date" and "to_date"
- Timestamp dimension: ProductTransfer
- Highly Browsable dimension: ProductTransfer: gives details about which product transfer on which period with its progress status.

This star schema is a combination of two-star schemas that can help us to understand distributions of products over the pipeline. We can determine a product ‘A’ transfer on a particular date from location 1 to location 2. Also, we can determine transferred amount, the cost to transfer the product, time is taken to transfer the product, total distance of pipeline covered.

Thus we will be able to answer the following question

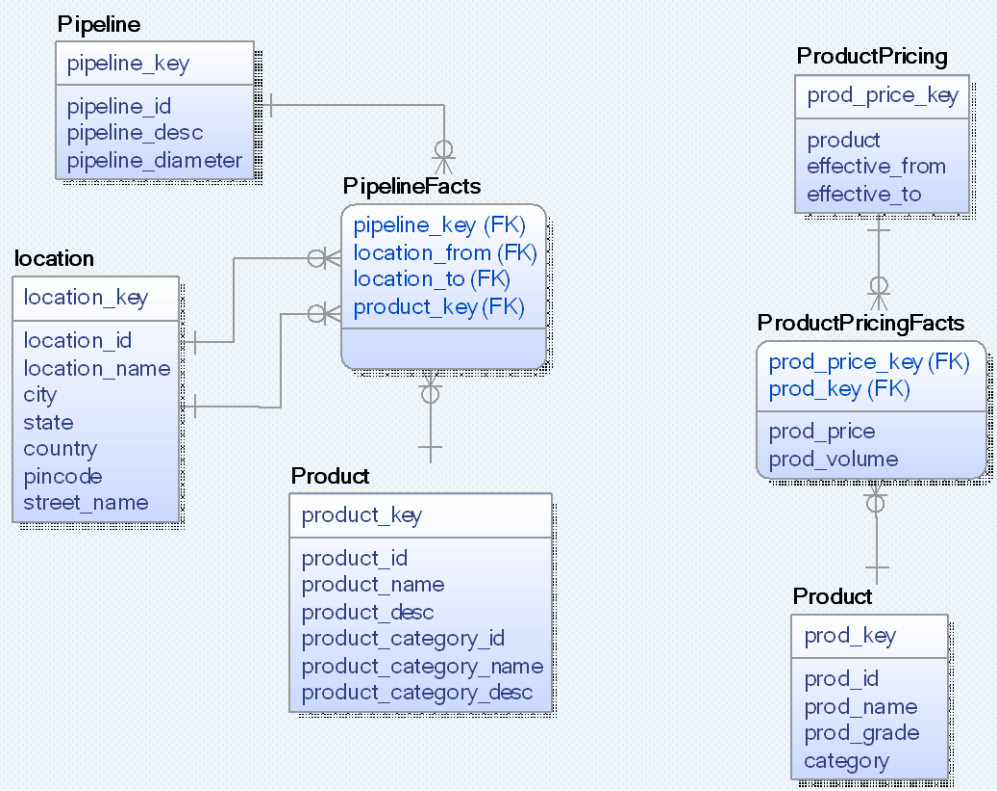
1. Average and lowest oil distribution in a month by pipeline: we can determine these values using volume_transferred field by pipeline on monthly basis. We need to do a drill across the two-star schemas to obtain this information.
2. Pipeline infrastructure across different location
   We can use schemas with pipeline facts and distribution facts to get following details
   - The pipeline structure spread across which locations, regions
   - Kind of product transferred from pipeline
   - Product delivery timings

We can define a query to fetch a kind of product transferred for TransferType = pipeline along with a transfer to and from location, transfer period from star schema with DistributionFacts.

We can define another query to get a pipeline location, a product that flows in it on any date period.

By combining this two query, we will be able to define pipeline infrastructure. These queries can also help to define average and lowest oil distribution.

Stovepipe example:



The star schema with ProductPricingFacts gives pricing details for different products and the star schema with PipelineFacts is a **Factless fact table** that gives detail of pipeline structure. Now the issue with this design is that there are no conformed dimensions which will help us to drill across. If we want to find total cost for the products transfer in the pipeline, then we may not be able to find this.

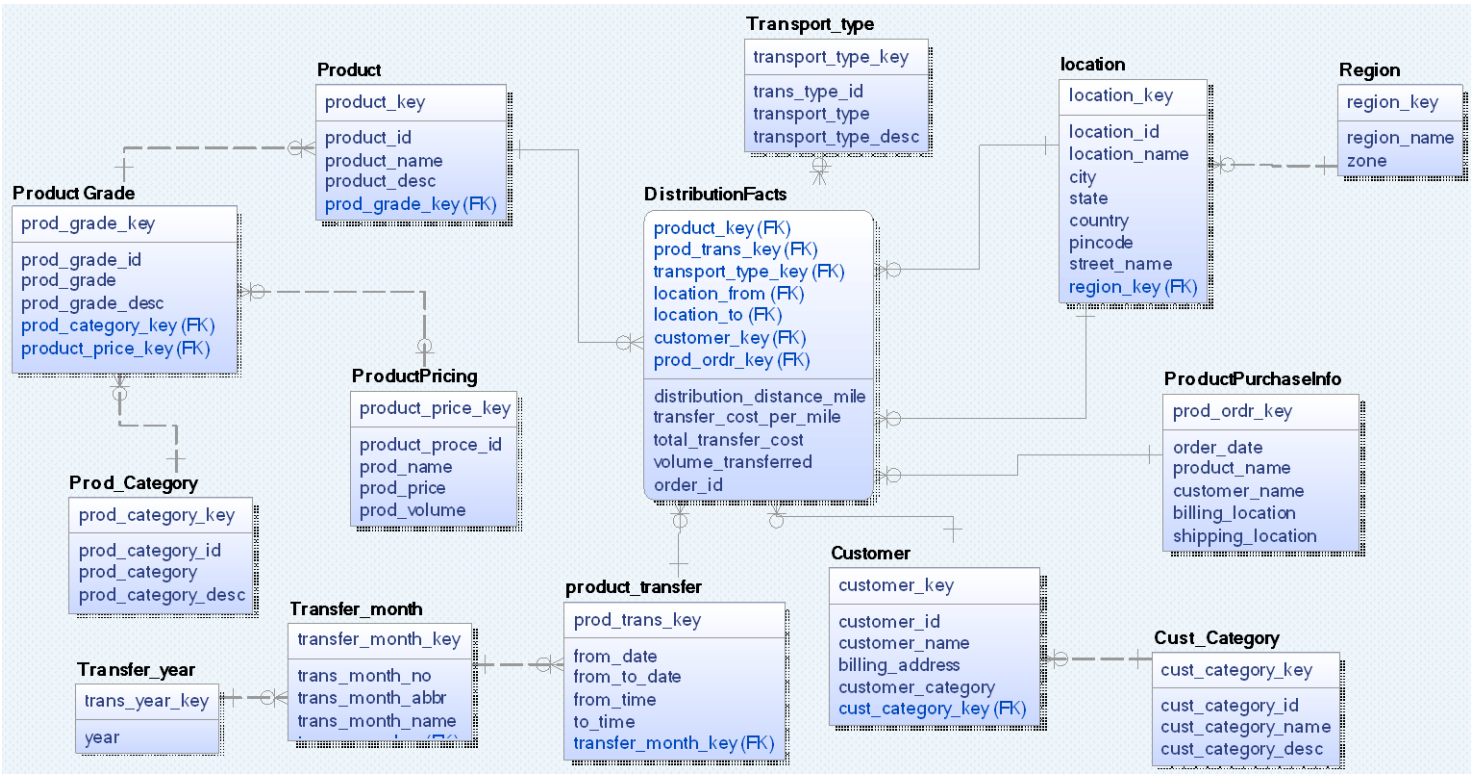
Design information:

- Fact table
  - PipelineFacts: Grain: pipeline location
  - ProductPricingFacts: Grain: product price; Additive Facts: prod_price, prod_volume
- Dimensions:
  - Pipeline: Affinity dimension, holds pipeline information: => type 2
  - Location: Affinity dimension, holds location details: => type 2
  - Product: Affinity dimension, capture product details: => type 2
  - ProductPricing: Affinity dimension, pricing information for products: => timestamp
  - Product: Affinity dimension, holds product details: => type 2

Timestamp to Type 3 conversion

To convert ProductPricing from timestamp dimension to type 3, we can add prod_proce_current and prod_price_previous columns instead of effective_from and effective_to.

Snowflake Schema



Core structure: A fact table DistributionFacts surrounded by other dimensions such as transport_type, location, customer, product_transfer, product. Dimensions are de-normalized.

Outriggers:

Outriggers for product dimensions: product Grade, prod_category, productPricing
Outriggers for product_transfer: transfer_month, transfer_year
Outriggers for customer: cust_category
Outriggers for location: region

References:

https://www.chevron.com/?utm_term=chevron&utm_campaign=[campaign]&utm_medium=cpc

http://www.psac.ca/business/industry-overview/

https://en.wikipedia.org/wiki/Midstream

http://www.yourarticlelibrary.com/marketing/marketing-mix/pricing-of-products-definition-factors-and-other-details-marketing-mix/41099/

http://www.computerweekly.com/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse

http://www.zentut.com/data-warehouse/data-mart/