The Data Warehouse Toolkit

Cap. 9 – Serviços Financeiros

Vinícius Pereira

CONTEÚDO

- Estudo de caso
- Triagem de dimensões
- Dimensão Unidade Domiciliar
- Dimensão Multivalor
- Minidimensões Revisitadas
- Associação de fatos com valores arbitrários
- Saldo Atual
- Esquemas de produtos heterogêneos

Estudo de caso bancário

Estudo de caso

- Meta inicial do banco: criar a capacidade de melhor analisar suas contas
- Usuários desejam poder separar e combinar (*slice and dice*) contas individuais, assim como os agrupamentos domiciliares residenciais a que pertencem
- Um dos principais objetivos do banco: negociar com mais eficácia, oferecendo produtos adicionais para unidades domiciliares

Estudo de caso

Conjunto de requisitos:

1. Os usuários desejam analisar 5 anos de dados de instantâneos mensais históricos em cada conta

2. Toda conta possui um saldo primário. A empresa deseja agrupar diferentes tipos de contas na mesma análise e comparar tais saldos

3. Todo tipo de conta (conhecido como *produtos* dentro do banco) possui um conjunto de atributos de dimensão personalizados e de fatos numéricos que tendem a ser muito diferentes entre os produtos

Estudo de caso

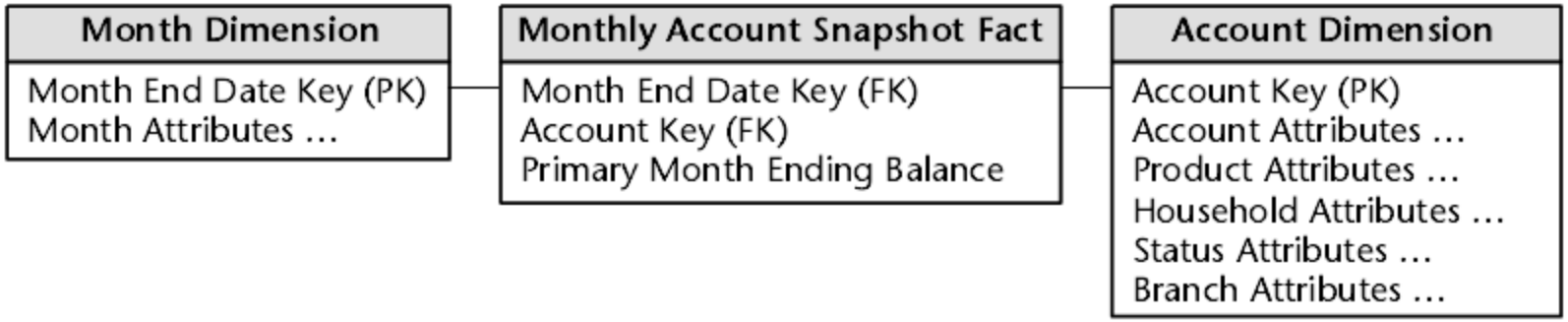
4. Toda conta pertence, teoricamente, a uma unidade domiciliar. Há uma quantidade surpreendente de transitoriedade em relacionamentos de conta/unidade domiciliar em virtude das alterações no status marital e de outros fatores de etapas da vida

5. Os usuários estão interessados em informações demográficas no que diz respeito a pertencerem a clientes individuais e unidades domiciliares. Além disso, o banco captura e armazena escores de comportamento relacionados à atividades ou às características de cada conta e unidade domiciliar.

Triagem de dimensões

Triagem de dimensões

- Iniciamos com uma tabela de fatos básica que registra os saldos primários de cada conta ao final de cada mês
- O grão é uma linha para cada conta ao final de cada mês
- Com base nisso, imaginamos, inicialmente, um projeto com duas dimensões: mês e conta



Triagem de dimensões

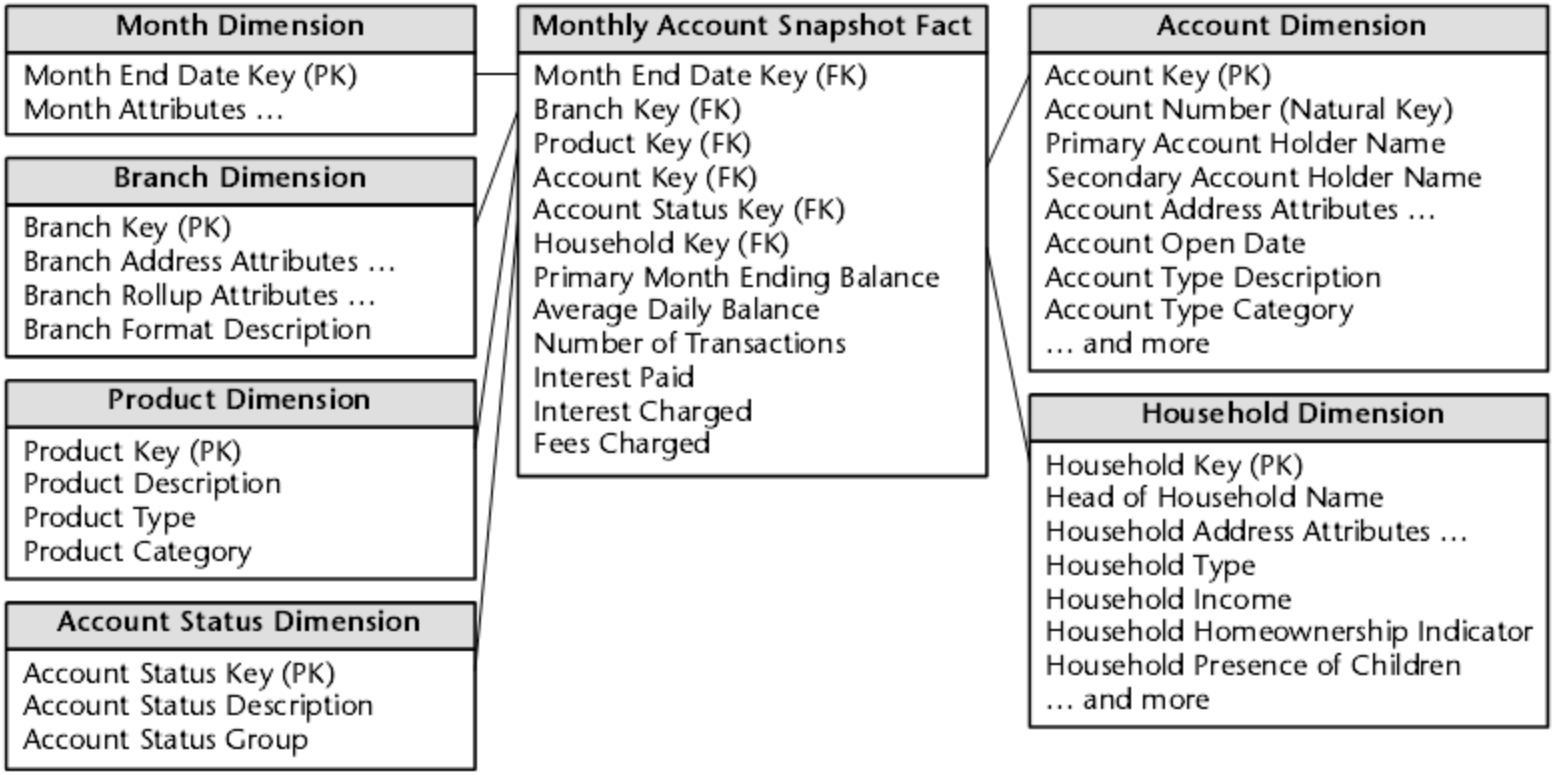
- Porém, tal esquema não reflete adequadamente as dimensões de negócios naturais

- No lugar de uma enorme dimensão Conta, dimensões analíticas adicionais (como Produto e Agência), espelham melhor como os usuários do banco pensam em seus negócios
  - Assim, elas lidam com os objetivos de desempenho e uso de um modelo dimensional

- Por fim, uma dimensão Conta em um banco de grande porte pode abranger milhões de linhas.
  - Pode gerar SCD tipo 2 (dimensão que muda lentamente)

Triagem de dimensões

- Com base nos requisitos do banco, por fim escolhemos as dimensões a seguir para o nosso esquema:



Triagem de dimensões

- Conforme visto na figura, na interseção das 6 dimensões, existe um instantâneo mensal e o registro do saldo primário de outras métricas que fazem sentido no contexto (juros pagos, juros cobrados, por exemplo)

- Os saldos não são aditivos em nenhuma medida de tempo. (Pensar em saldos de conta como saldos de estoque)

- Dimensão Produto – hierarquia de produtos simples que descreve os produtos do banco
  - Nome do produto, tipo e categoria

- A necessidade dessa categorização é a mesma que a de mercadorias genéricas em um supermercado

Triagem de dimensões

- Diferença: o banco desenvolve um número grande de atributos de produtos personalizados para cada tipo de produto (veremos mais a frente)

- Dimensão Agência: semelhante as dimensões Loja ou Local vistas em outros capítulos (loja de varejo ou armazém central de distribuição)

- Dimensão Status da Conta: útil para registar a condição da conta ao final de cada mês (ativa, inativa, alterações durante o mês, como uma nova abertura de conta)

- De certa forma, a dimensão Status é um outro exemplo de minidimensão

Dimensão
Unidade Domiciliar

Dimensão Unidade Domiciliar

- Os usuários querem poder analisar o relacionamento do banco com uma unidade econômica inteira (entender o perfil e vender produtos adicionais)

- Capturar atributos demográficos:
  - Receita da unidade
  - A casa é própria ou não
  - Se há crianças

- Atributos mudam com o passar do tempo

- Se for comercial ao invés de clientes, teríamos uma Unidade Corporativa (atributos semelhantes)

Dimensão Unidade Domiciliar

- Pode ser composta por várias contas e correntistas individuais. Ex:
  - John e Mary Smith formam uma unidade
  - John tem conta corrente e Mary poupança
  - Possuem conta-conjunta, cartão de crédito e hipoteca
  - Esses 5 tipos de conta são consideradas como parte da Unidade Domiciliar Smith

- Como visto no Cap 6, há produtos e serviços especializados para fazerem a correspondência necessária e determinarem atribuições de unidade domiciliar

- A decisão de tratar contas e unidades domiciliares separamente é uma simples questão de projeto
  - Apesar de serem correlacionadas, estão separadas no esquema devido ao tamanho da dimensão Conta e das mudanças que uma Unidade sofre (evitar SCD tipo 2 na dimensão Conta)

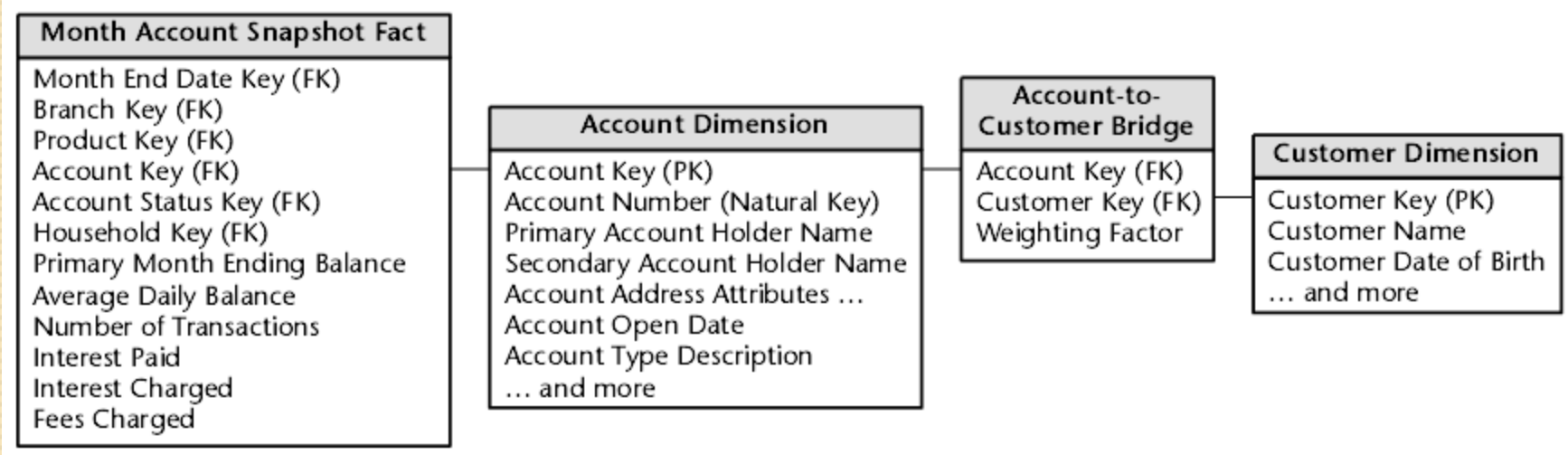
Dimensões Multivalor

Dimensões Multivalor

- Como acabamos de ver, uma conta pode ter um ou mais correntistas (clientes associados a conta)
- Não podemos incluir o cliente como atributo de conta – viola a granularidade da dimensão (N indivíduos podem ser associados a uma conta)
- Não podemos incluir o cliente como uma dimensão na tabela de fatos – viola a granularidade dessa tabela (uma linha por conta por mês, pelo mesmo motivo acima)
- Esse caso é um exemplo clássico de dimensão multivalor (detalhes no Cap 13)

Dimensões Multivalor

- Por ora, basta dizer que é preciso usar uma tabela de ponte de conta para cliente, dessa forma vinculando uma dimensão Cliente individual a uma tabela de fatos de granularidade de conta



Dimensões Multivalor

- **DICA:** um atributo multivalor ilimitado pode ser associado a uma linha de dimensão usando-se uma tabela de ponte para associar os atributos multivalor com a dimensão

- Em algumas empresas, o cliente individual é identificado e associado a cada transação
  - Cartão de crédito com números exclusivos por cliente
  - Não há necessidade de ponte entre conta e cliente porque os fatos da transação atômica estão no grão do cliente
  - Conta e Cliente seriam chaves externas nessa tabela de fatos

Minidimensões revisitadas

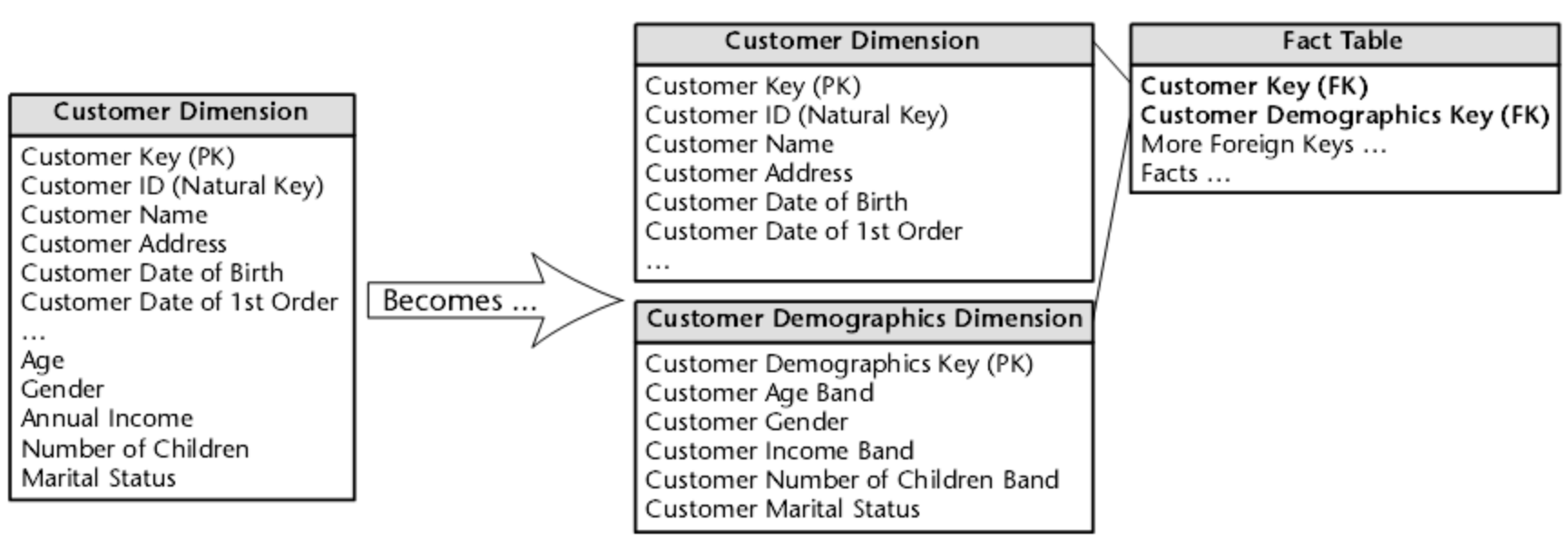
Minidimensões revisitadas

- Como visto no Cap 6, há na dimensão Cliente uma ampla variedade de atributos para descrever as contas, cliente e unidades domiciliares
  - Atributos relativos a crédito mensal
  - Dados demográficos externos
  - Escores calculados para identificar características de comportamento, lucratividade, entre outras coisas
- Interesse em compreender as alterações nesses atributos e responder a elas com o passador do tempo
- Evitar SCD tipo 2 para controlar essas alterações na dimensão Conta

Minidimensões revisitadas

- Separar os atributos navegáveis e alteráveis em múltiplas minidimensões, como áreas de crédito e minidimensões demográficas (chaves incluídas na tabela de fatos)

- Exemplo da recomendação, figura 6.4



Minidimensões revisitadas

- As minidimensões nos permitem separar e combinar os dados com base em uma extensa lista de atributos, enquanto controlam as alterações de atributo com o passar do tempo
  - Apesar de serem extremamente eficiente, evitar o uso em demasia
- Como visto no Cap 6, as minidimensões geralmente usam faixas de valores ao invés de valores distintos
  - US$31.257,98 estaria em uma faixa como US$30.000-US$34.999
- Existe duas situações em que essa idéia pode ser inadequada:
  - A análise de exploração de dados muitas vezes requer valores distintos em vez de associações fixas para serem mais eficazes
  - Necessidade de analisar os valores distintos para determinar se as associações selecionadas são apropriadas
  - Nesses casos, mantemos as associações nas minidimensões, mas também armazenamos os lavores distintos como fatos na tabela de fatos

Associação de fatos com valores arbitrários

Associação de fatos com valores arbitrários

- Vamos supor a necessidade de gerar relatórios com associação de valores em um fato numérico padrão, como saldo em conta, mas sem lidar com associações predefinidas

- Um relatório, baseado no instantâneo do saldo em conta, parecido com este:

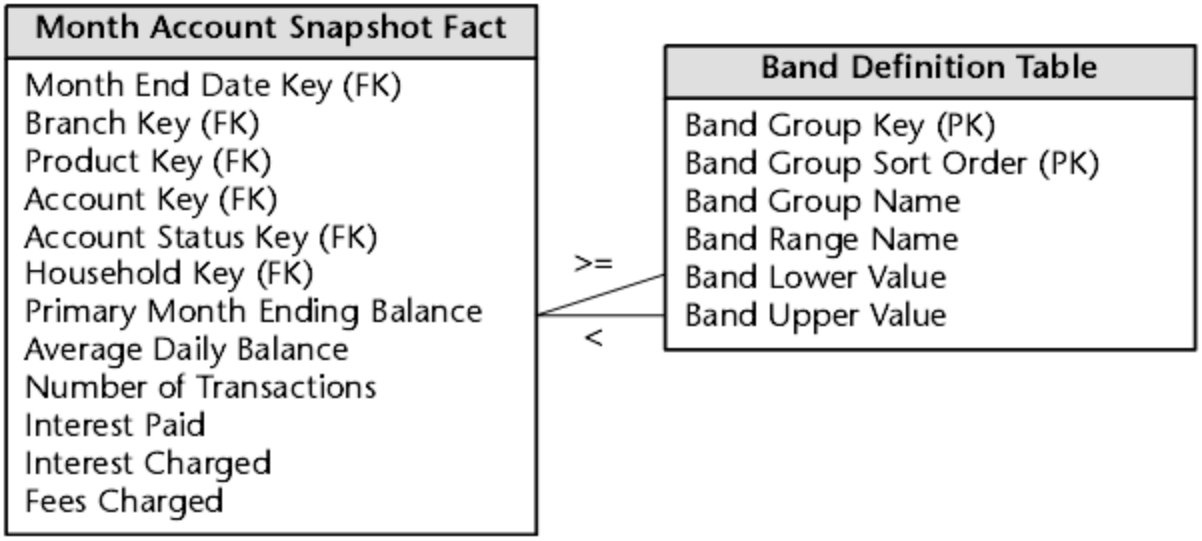
Balance Range	Number of Accounts	Total of Balances
0-1,000	45,678	$10,222,543
1,001-2,000	36,788	$45,777,216
2,001-5,000	11,775	$31,553,884
5,001-10,000	2,566	$22,438,287
10,001 and up	477	$8,336,728

- Usando a figura como base, é difícil criar esse relatório diretamente da tabela de fatos

Associação de fatos com valores arbitrários

- A SQL não tem generalização da cláusula GROUP BY que separa valores aditivos em faixas
- Complicação: as faixas não são de tamanho igual e tem nomes textuais como "10,001 and up"
- Os usuários precisam de flexibilidade para redefinir as associações no momento da consulta com diferentes limites e níveis de precisão
- O esquema a seguir nos permite criar relatórios fléxiveis com associação de valores
- O problema, atualmente, é melhor tratado com o comando CASE da SQL

Associação de fatos com valores arbitrários



- A Tabela de Definição da Associação (direita) pode conter quantos conjuntos de associações de relatório forem necessários

- Essa tabela é associada à de fatos de saldo usando-se um par de junções "menor que" e "maior que"

Associação de fatos com valores arbitrários

- O relatório utiliza o nome da faixa com associações como o cabeçalho da linha e classifica o relatório pela coluna de classificação com associações

- Controlar o desempenho dessa consulta pode ser um desafio
  - No exemplo são mais de 90.000 contas
  - Tudo que essa junção faz é agrupar os 90.000 saldos
  - Necessidade de colocar um índice diretamente no fato saldo
  - Se o SGBD puder classificar e comprimir o fato individual de modo eficiente, o desempenho será melhorado consideravelmente

- Considerando o quão barato está o Gb atualmente, será que isso é um problema? E se for, de tão grande porte assim?

Saldo Atual

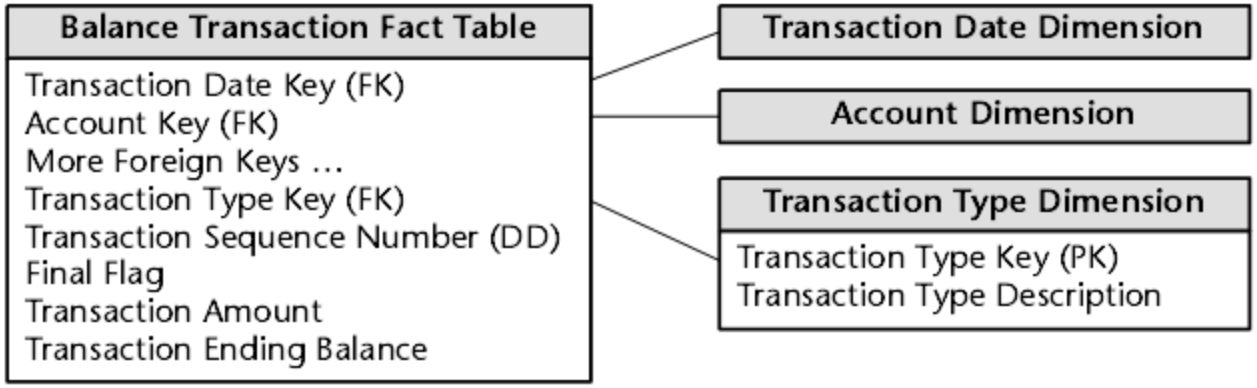
Saldo Atual

- Até agora analisamos instantâneos de saldo de final de mês
  - Se necessário, podemos suplementar a tabela de fatos com uma segunda tabela de fatos que forneça o instantâneo mais atual (atualização noturna)
  - Ou uma tabela de fatos que forneça instantâneos de saldo diário na última semana ou mês

- No entanto, o que ocorrerá se tivermos a necessidade de relatar o saldo de uma conta em qualquer ponto histórico selecionado arbitrariamente?
  - Criar instantâneos diários – 10 milhões de contas = 3,65 bilhões de linhas de fatos por ano

- Para simplificar, reduziremos a tabela de fatos de transações de conta a um projeto extremamente simples, como mostrar a imagem a seguir

Saldo Atual



- A chave de tipo de transação se junta a uma pequena tabela de dimensão de tipos de transação permitidos.
- O número da sequência de transações aumentará continuamente enquanto a conta existir
- O sinalizador final indicará se é a última transação de uma conta em um dado dia
- Total de transações é autoexplicativo
- O saldo é o saldo final na conta após o evento de transação

- Adicionamos um linha à tabela de fatos na figura anterior somente quando ocorre uma transação

Esquemas de produtos heterogêneos

Esquemas de produtos heterogêneos

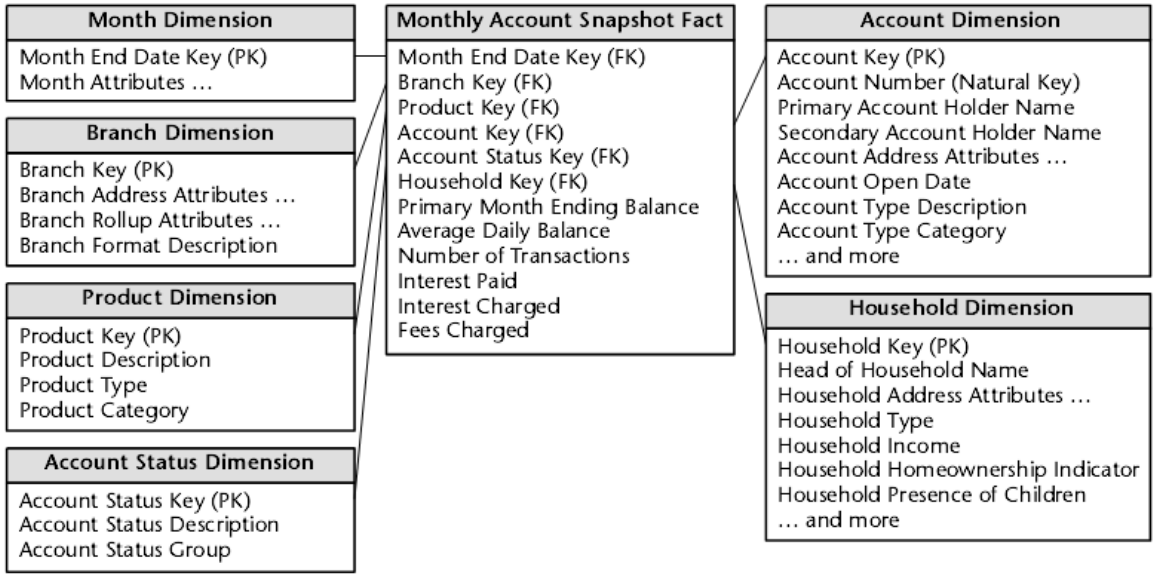
- Como mencionado anteriormente, um banco de varejo oferece uma variedade de produtos, desde conta correntes a hipotecas, aos mesmos clientes

- Toda conta tem um saldo primário e um total de juros associado a ela, mas cada tipo de produto possui diversos atributos especiais e fatos medidos que não são compartilhados
  - Ex: conta corrente tem saldo mínimo, limite de saque e encargos sobre serviços
  - Depósitos a prazo tem alguns atributos de conta corrente, além de datas de vencimento, frequências compostas e taxa de juros annual

Esquemas de produtos heterogêneos

Duas perspectivas normalmente exigidas pelos usuários:

- Visão global, incluindo a capacidade de separar e combinar todas as contas de forma simultânea, independente de seu tipo
  - Necessário para planejar as estratégias de venda cruzada e gerenciamento apropriado de relacionamento com o cliente
  - Precisamos da tabela de fatos básica única cruzando todas as linhas de negócio para dar uma idéia do portfólio de contas completo

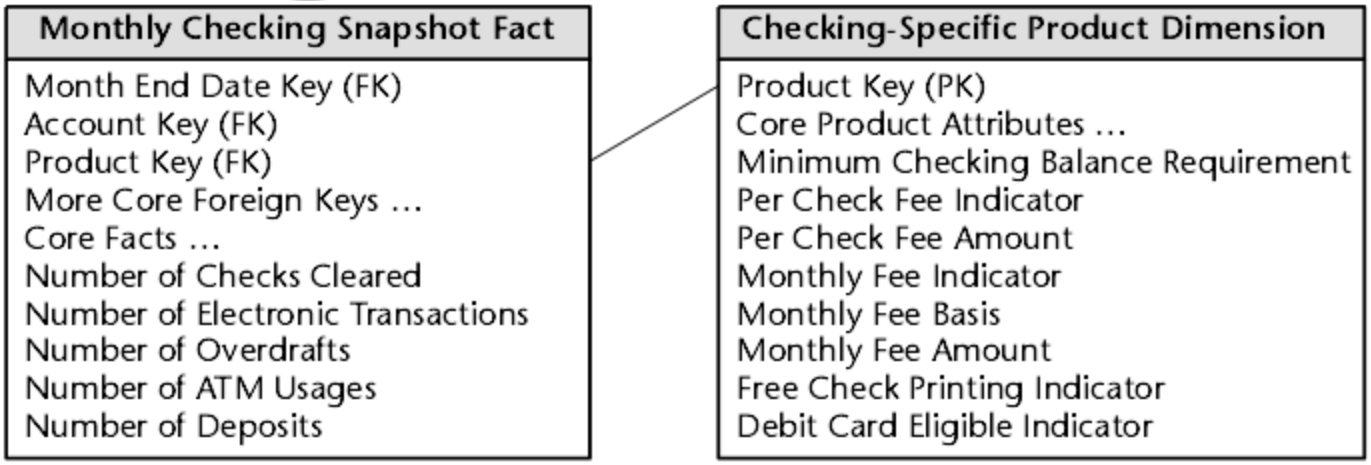


Esquemas de produtos heterogêneos

- A segunda perspectiva é o modo de visão de linha de negócio específico, que se concentra nos detalhes profundos de um negócio (como as contas correntes)
  - Longa lista de fatos e atributos especiais
  - Não podem ser incluídos na tabela de fatos básica. Caso contrário, acabaríamos com uma centena de fatos especiais, a maioria dos quais com valores nulos em uma linha específica
  - O mesmo ocorreria na dimensão Produto, se incluíssemos atributos de linha de negócio específicos
  - Qualquer das duas opções resultaria em um "queijo suiço"

- A solução é criar um esquema personalizado para a linha de contas correntes do negócio, como mostra a figura a seguir

Esquemas de produtos heterogêneos



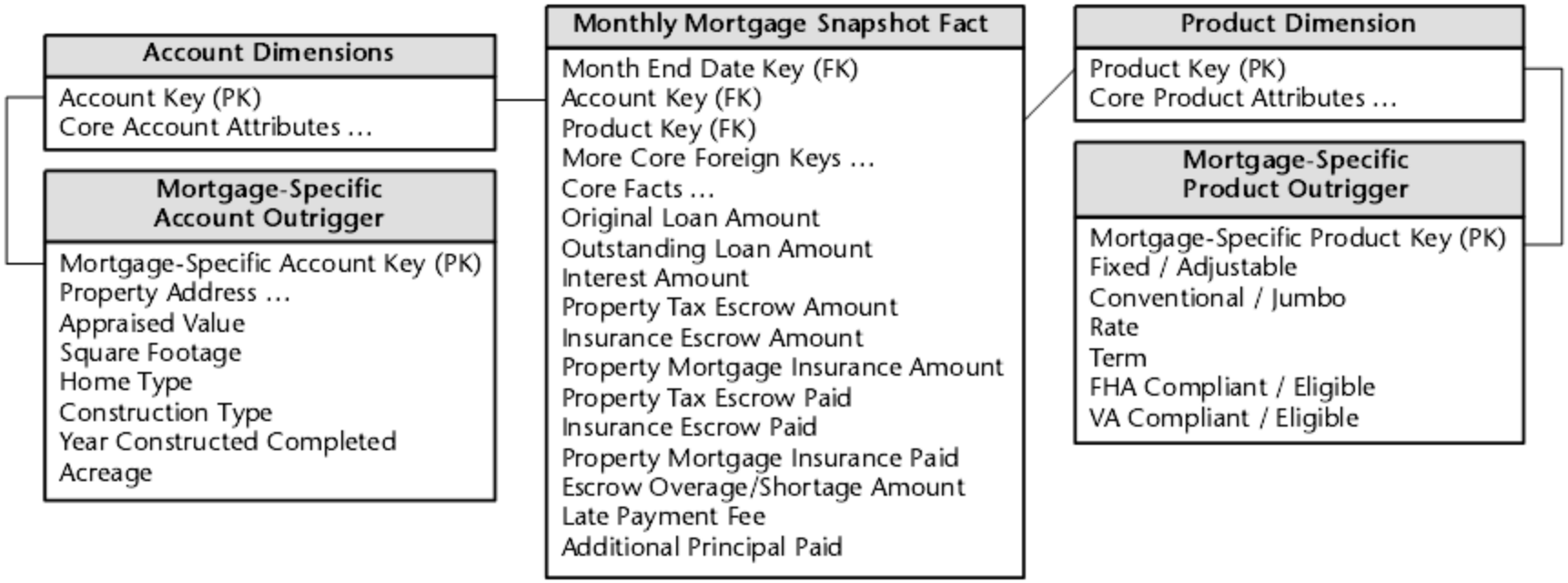
- Tabela de fatos de contas correntes personalizada e a dimensão Produto da conta corrente correspondente são ampliadas
  - descrevem os fatos e atributos específicos que só fazem sentido para produtos de conta corrente
  - Contém, também, os fatos e atributos principais, o que evita juntar tabelas dos esquemas básicos e personalizado para obter o conjunto completo de fatos e atributos

Esquemas de produtos heterogêneos

- Da mesma forma, criaríamos tabelas de fatos e de produtos personalizadas para as outras linhas de negócio
- Apesar de parecer complexo, apenas o DBA vê todas as tabelas de uma vez
- As chaves das dimensões Produto personalizadas são as mesmas usadas na dimensão Produto básica
  - Cada dimensão Produto personalizada é um subconjunto de linhas da tabela de dimensão Produto básica
- Cada dimensão Produto personalizada contém atributos específicos a um tipo de produto

Esquemas de produtos heterogêneos

- Podemos considerar os atributos específicos como um *outrigger* dependente do contexto para a dimensão Produto



- Isolamos os atributos básicos na tabela dimensão Produto base e podemos incluir uma chave de *snowflake* e, cada registro base que aponte para seu *outrigger* de produto estendido apropriado

Esquemas de produtos heterogêneos

- Se as linhas de negócio em nosso banco são separadas fisicamente (cada uma com seu *data mart*)
  - as tabelas de fatos e de dimensão personalizadas provavelmente não residem no mesmo espaço das tabelas de fatos e dimensões básicas
  - Os dados na tabela de fatos básica seriam duplicados uma única vez para implementar todas as tabelas personalizadas

- Se as linhas de negócio compartilharem o mesmo espaço na tabela física, poderemos evitar tal duplicação
  - Isso é feito atribuindo uma chave de junção especial a cada linha da tabela de fatos básica que identifique uma única conta em um único mês
  - Usando essa chave de junção, vinculamos fisicamente os fatos personalizados estendidos à tabela de fatos básica
  - Exemplo na próxima imagem

Esquemas de produtos heterogêneos