UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Tecnologias de Banco de Dados (TI-BD)

Bancos de Dados Distribuídos

Nome: Ícaro Pires de Souza Aragão

Matrícula: 15/0129815

Brasília, DF

2019

1. INTRODUÇÃO

Existem diversas tecnologias diferentes para armazenamento persistente de dados, cada uma atende melhor um tipo de necessidade de diferentes tipos de usuários e/ou organizações. Neste trabalho será tratado sobre os Bancos de Dados Distribuídos, que podem ser utilizados principalmente para resistência a falhas e escalabilidade de uso.

Muitos dos conceitos, vantagens e desvantagens desse tema são, naturalmente, os mesmos ou altamente inspirados na questão mais geral, que são Sistemas Distribuídos. Por isso estes serão abordados primeiro, e então aspectos mais específicos das bases de dados serão abordadas depois.

Além dos conceitos, serão discutidas algumas vantagens e desvantagens de seu uso, assim como também serão citados algumas das tecnologias mais utilizadas e alguns casos de grandes empresas que as utilizam.

2. SISTEMAS DISTRIBUÍDOS

Antes de entender o que são os Bancos de Dados Distribuídos é importante entender o que são Sistemas Distribuídos, sendo os bancos de dados apenas uma aplicação desse conceito num contexto mais específico. Segundo Tanenbaum, Sistemas Distribuídos são uma “Coleção de elementos computacionais independentes que mostra-se aos usuários do sistema como um único e coerente sistema”’.

Para justificar isso, Tanenbaum apresenta dois pontos principais nessa definição, o primeiro refere-se aos Sistemas Distribuídos como um conjunto de elementos computacionais que conseguem se comportar independentes uns dos outros, os nós. Já a segunda parte da definição se refere ao da complexidade de uso, que é um dos elementos chave na decisão em se usar ou não uma tecnologia distribuída.

Nesse sentido, defende que a interface com o sistema, ou seja, como o usuário vai interagir com ele, deve ser de forma coerente (o que torna o comportamento mais previsível) e que toda a comunicação entre nós e funcionamento interno deve ser transparente ao usuário, no sentido de proteger o usuário da complexidade.

3. BANCOS DE DADOS DISTRIBUÍDOS

3.1 Conceito

Entendendo-se uma base de dados como um conjunto de dados armazenados eletronicamente em algum lugar e tendo em vista que um sistema que lida com isso é um Sistema Gerenciador de Banco de Dados, ao se incluir também o conceito de Sistemas Distribuídos, discutido na seção anterior, pode-se extrapolar o conceito de Bancos de Dados Distribuídos para: um Sistema Distribuído que é responsável pelo gerenciamento de mais de uma base de dados que estarão distribuídas em processos do sistema operacional que podem ou não estar no mesmo dispositivo físico.

De maneira mais formal, David Bell [2] define esse tipo de tecnologia como a união dos conceitos de integração (no sentido da base de dados) e de distribuição (no sentido dos elementos de rede), dois conceitos divergentes à primeira vista mas que a compõe uma base de dados distribuída.

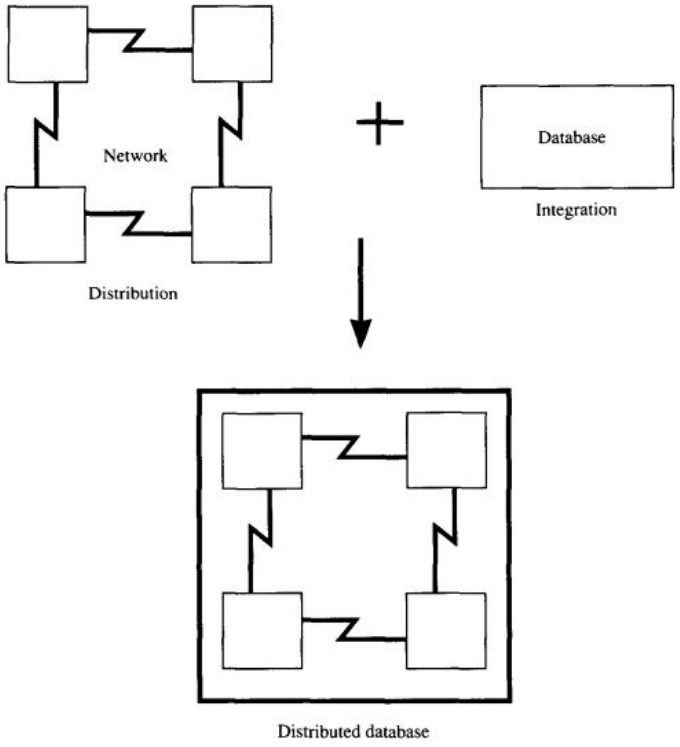

Figure 2.1 The distributed database concept.

Figura 1: O conceito de base de dados distribuída
Fonte: Bell, 1992

3.2 Objetivos gerais

David Bell aponta que muitos tipos de companhias veem em bases de dados distribuídas uma oportunidade de integração dentre várias bases de dados diferentes de diferentes setores, derivando disso uma estrutura com uma única interface (apesar de atribuir interesses também de usuários em pequena escala). Mas esse não é o único motivo pelo qual se utiliza bases de dados distribuídas, elas podem oferecer diversas vantagens se comparadas a bases de dados relacionais simples.

3.2.1 Vantagens

- Integrar bases de dados de diferentes tipos como por exemplo: SQL, NoSQL, tecnologias de empresas diferentes, etc;

- Prover autonomia local numa base de dados sem afetar as outras, como no caso de vários setores de uma empresa, por exemplo.

- Prover uma estrutura tolerante a falhas. Um nó pode cobrir a falha de outros, por exemplo;

- Suportar escalabilidade de performance em caso de muitos acessos de grandes empresas.

3.2.2 Desvantagens

- Aumento de custos para se manter a estrutura de forma que mantenha um funcionamento satisfatório. Pode ser necessário mais equipamentos de rede, conexão mais veloz de rede, etc;

- Aumento da complexidade arquitetural, dificultando aspectos como: estimar capacidade de processamento, estimar fluxo de dados entre os nós, controlar a competição por recursos entre áreas e nós, etc;

- Dificuldade em manter a integridade dos nós de forma distribuída. Da mesma forma que transações locais mantêm a integridade dos dados numa base local, é preciso que elas mantenham a integridade dos de forma distribuída, mas agora tendo em vista várias localizações diferentes.

3.3 Motivação

A necessidade de se utilizar Banco de Dados de forma distribuída, segundo David Bell [1], vem principalmente de duas pressões principais, a do usuário e das próprias tecnologias. Segundo ele, conforme as companhias crescem, cresce também a frustração dos usuários se os dados se mantém centralizados, principalmente devido a sensação de falta de controle do que se está acontecendo nela, por isso, acabou-se tornando um padrão *de facto* descentralizar a base de

dados conforme a companhia cresce, e assim, cada departamento da empresa tem sua própria base.

Já a pressão da tecnologia se refere ao barateamento de componentes de rede e modernização das tecnologias de comunicação entre computadores em meados dos anos 80. Um outro aspecto é permitir integrar bases que já estava distribuídas mas não tinham interação entre sí. Esses aspectos trouxeram a viabilidade para se usufruir das vantagens das bases de dados distribuídas.

3.4 Estratégias para manutenção da integridade das bases

Existem vários modelos que representam estratégias diferentes de como manter os dados consistentes entre diversas bases de dados, como pode ser visto com mais detalhes em [1]. O que essas estratégias têm em comum é que todas elas incluem uma combinação de diferentes usos dos conceitos: replicação e duplicação.

O processo de replicação é o mais complicado, nele as diferenças entre as bases de dados são identificadas e então os bancos são sincronizados. Esse processo pode ser longo e demorado dependendo da região geográfica em que as bases de dados se encontram e também de qual alteração foi feita. Lembrando que isso tem que ser feito através de transações para evitar condições de corrida.

Já no processo de duplicação esse mecanismo se torna mais simples, nele se define uma base de dados como *master* e outra como secundária, na *master* pode acontecer todos os tipos de operações comuns, já a secundária, em que é permitida apenas operações de leitura, é periodicamente atualizada.

A duplicação é geralmente um procedimento bem mais barato do que o anterior. Ele pode ser útil para situações de tolerância à falhas, quando por exemplo o *master* se torna indisponível, a base secundária pode então se tornar o *master* nesse período, mantendo também a transparência de acesso ao usuário, que nem precisa saber que está utilizando agora uma base substituta. A duplicação também pode ser útil

em casos para diminuir a latência de acesso causada pela localização geográfica dos servidores.

3.5 Principais tipos de arquitetura

De maneira mais abrangente, existem dois principais tipos de arquitetura: arquiteturas homogêneas e heterogêneas. O mais simples são as arquiteturas homogêneas, nelas pode-se assumir que que o ambiente e ferramentas em que todas as bases de dados estão localizadas são os mesmos (mesmo sistema operacional, mesmo gerenciador do banco, mesmas ferramentas auxiliares, etc), se não forem os mesmos, serão, pelo menos, oficialmente compatíveis.

Em contrapartida, as arquiteturas heterogêneas podem ter hardware, software e até modelos de dados diferentes responsáveis pelas bases. Como consequência, elas se tornam mais difíceis de se gerenciar e mais custosas para se manter, apesar disso, trazem a vantagem da integração de ambientes diferentes e gerando uma mesma interface.

3.6 Tecnologias e exemplos de uso

Como um grande exemplo de projeto de Banco de Dados Distribuído pode-se citar o Voldemort, esse banco foi projeto para ser escalável horizontalmente num grande número de servidores, provendo assim várias das vantagens discutidas anteriormente. Além disso, ele também é um dos principais bancos utilizados pelo LinkedIn [3].

Uma aplicação característica de bancos de dados distribuídos é em contextos de big data, nesses cenários, além de serem utilizados bancos distribuídos, também são bastante utilizados os bancos NoSQL. Um grande exemplo desse caso é o HBase, geralmente utilizado junto ao sistema de arquivos para Big Data: HDFS (que faz parte do ecossistema Hadoop).

Esse é uma dos motivos pelo qual é utilizado por grandes companhias que agem no contexto de big data. Uma dessas companhias é o Netflix, segundo ele próprio, o HBase facilita o crescimento do *cluster* e a redistribuição da carga entre os nós inclusive em quando em tempo de execução.

4. BIBLIOGRAFIA

[1] Tanenbaum, Andrew S.; Steen, Maarten van (2002). *Distributed systems: principles and paradigms*. Upper Saddle River, NJ: Pearson Prentice Hall. ISBN 0-13-088893-1.

[2] BELL, David; GRIMSON, Jane (1992). Distributed Database Systems. Addison-Wesley. ISBN 0-201-54400-8

[3] MATELJAN, Vladimir; CISIC, D.; OGRIZOVIC, D. Cloud database-as-a-service (DaaS)-ROI. In: **The 33rd International Convention MIPRO**. IEEE, 2010. p. 1185-1188.

[4] IZRAILEVSKY, Yuri. **Medium:** NoSQL at Netflix, 2011. Disponível em: https://medium.com/netflix-techblog/nosql-at-netflix-e937b660b4c. Acesso em: 9 set. 2019.